

## Part XVII

### **Supplementary Contributed Papers**





# Clustering of Waveforms-Data Based on FPCA Direction

Giada Adelfio<sup>1</sup>, Marcello Chiodi<sup>1</sup>, Antonino D'Alessandro<sup>2</sup> and Dario Luzio<sup>3</sup>

<sup>1</sup> Dip. di Scienze Statistiche e Matematiche “S. Vianelli”, University of Palermo, Italy, *adelfio@unipa.it*, *chiodi@unipa.it*

<sup>2</sup> Centro Nazionale Terremoti OBS Lab. Gibilmanna, INGV, Italy, *antonino.dalessandro@ingv.it*

<sup>3</sup> Dip. di Chimica e Fisica della Terra, University of Palermo, Italy, *luzio@unipa.it*

**Abstract.** The necessity of finding similar features of waveforms data recorded for earthquakes at different time instants is here considered, since eventual similarity between these functions could suggest similar behavior of the source process of the corresponding earthquakes. In this paper we develop a clustering algorithm for curves based on directions defined by an application of PCA to functional data.

**Keywords:** FPCA, clustering of curves, waveforms

## 1 Introduction

In this paper we combine the aim of finding clusters from a set of individual curves with the functional nature of data, applying a variant of a  $k$ -means algorithm based on the principal component rotation of data.

Indeed, looking for curves similarity could be a complex issue characterized by subjective choices related to the continuous transformation of observed discrete data (Chiodi (1989)). Here we apply a classical clustering method to rotated data, according to the direction of maximum variance.

A  $k$ -means clustering algorithm based on PCA rotation of data is proposed, as an alternative to methods that require previous interpolation of data based on splines or linear fitting (García-Escudero and Gordaliza (2005), Tarpey (2007), Sangalli *et al.* (2008)).

## 2 Dealing with functional data

When data are observed as functions of time (such as financial time series, temperature recorded by some central source, etc.), we refer to as functional data. Since in many statistical applications realizations of continuous time series are available as observations of a process recorded in discrete time intervals, one crucial point is to convert discrete data to continuous functions,

that is from vectors to curves or more generally functions in  $\mathbb{R}^d, d \geq 1$ . When we talk about functional data, then we refer to  $n$  pairs  $(t_i, y_i)$  where  $y_i$  is the value of an observable variable  $x$  at time  $t_i$ , and we focus on a set of functions defined on  $[0, T]$ , such that:

$$\{y_i = x_i(t); i = 1, 2, \dots, I; 0 \leq t \leq T\}$$

Therefore, assuming that a functional datum for replication  $i$  arrives as a set of discrete measured values  $y_{i1}, y_{i2}, \dots, y_{in}$  the first task is to convert these values to a function  $x_i$  with values  $x_i(t)$  computable for any  $t$ , called functional objects. The conversion from discrete data to functions may involve smoothing (Ramsay and Silverman (2006)). One smoothing procedure often used is obtained representing each function as the linear combination of  $K$  base functions  $\phi_k$ :

$$x_i(t) = \sum_{k=1}^K c_{ik} \phi_k(t)$$

The conversion of functional data to functional objects requires to store the coefficients  $\{c_{ik}\}$  in a  $I \times K$  matrix; a main issue is the choice of the basis, such as Fourier basis useful for periodic data, B-splines, exponential basis.

Let  $\{y_{ij}\}$  be the observed value of the  $i^{th}$  function at time  $t_j$ , the basis representation of  $x(t)$  can be obtained by a least squares criterion, such that:

$$\min_{c_{ik}} \sum_{j=1}^J \left( y_{ij} - \sum_{k=1}^K c_{ik} \phi_k(t) \right)^2 \quad (1)$$

The smoothness degree depends on  $K$ , since small (large) values of  $K$  induce more (less) smoothed curves.

The best known basis expansion is obtained by the Fourier series, although it is useful when the observed functions are periodic and do not fluctuate in any particular interval more rapidly than the basis elsewhere. Therefore sometimes a roughness penalty approach can be used, that retains the advantages of the basis function and local expansion smoothing (like kernel and local polynomial fitting techniques), but overcome some of their limitations.

The spline smoothing method estimates a curve from observations with the aim to ensure both regularity and goodness of fit to the data, that is between variance and bias. For this purpose penalty terms can be added to the residual sum of squares.

## 2.1 Principal Components Analysis for Functional Data

Principal Components Analysis (PCA) is aimed to the reduction of the original set of variables  $X_1, X_2, \dots, X_k$  to a smaller set  $f_1, f_2, \dots, f_p (p < k)$  of linear orthogonal combinations

$$f_i = \sum_{j=1}^k \beta_{ij} x_j \quad i = 1, 2, \dots, p,$$

and able to display types of variations that are strongly represented in data. The corresponding PCA for functional data (FPCA) is now reviewed to introduce some notation used throughout the paper. In the functional context the value of the  $k^{th}$  variable on the  $i^{th}$  unit is now a function of the time  $x_i(t), i = 1, \dots, p$ ; therefore the principal components are now function values and the discrete index  $j$  is now replaced by the continuous index  $s$ , such that:

$$f_i = \int \beta(s)x_i(s)ds \quad (2)$$

with  $\beta(s)$  weight functions. Each functional principal component is obtained by maximizing:

$$\frac{1}{p} \sum_i f_i^2$$

and satisfying orthogonal constraints (Ramsay and Silverman (2006)).

### 3 Some analysis for waveforms data

Earthquake is an oscillatory motion of the ground, identified by a source point in the Earth (hypocenter), from which seismic waves spread out.

The instrumental observation of earthquakes is realized by seismographs, constituted by transducers, that convert the oscillatory energy of the ground involved by the seismic waves generated by earthquakes, nuclear explosions, and other seismic sources, in electric signals proportional to a kinematic motion parameter, by electronic machineries that amplify, filtrate, digitalize and send the electric signal and by displaying and recording systems.

Seismographs, then, transform the complex motion of the ground during an earthquake occurrence in a permanent record, named seismogram.

A seismogram is a record of the ground motion at a measuring station as a function of time. Seismograms typically record motions in three cartesian axes ( $x$ ,  $y$ , and  $z$ ), with the  $z$  axis perpendicular to the Earth's surface and the  $x$  and  $y$  axes parallel to the surface.

The elastic waves, also called seismic waves, that are spread out from the source point, are basically discriminated in P and S waves. The name P-wave stands for primary wave, as the P-wave is the fastest among the elastic waves, compared to the S-waves and surface waves (Rylyleigh and Love waves).

Usually seismograms show a fundamental structure in the recorded oscillations: the starting ones correspond to P-waves; in the central part S-waves overlap; in the final part surface waves can be observed. This structure is more complex if we are far away from the hypocenter, because of the reflection and refraction effects of the waves inside the Earth.

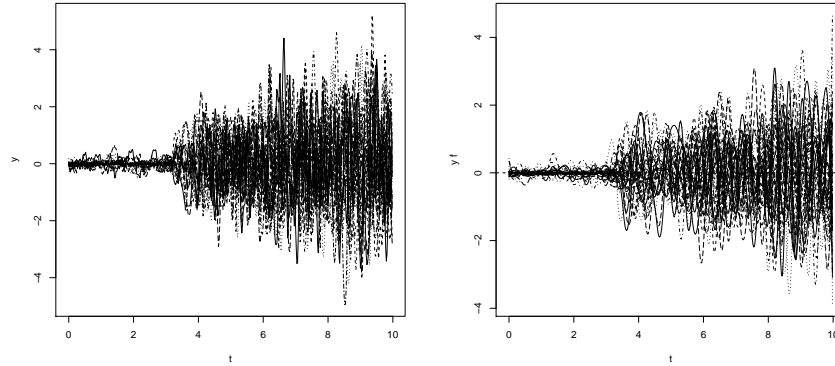
Waveforms correlation techniques have been introduced to characterize the degree of event similarity (Mezcua and Rueda (1994), Menke (1999)) and in facilitating more accurate relative locations within similar event clusters

by providing more precise timing of P and S arrivals (Gillard et al. (1996), Phillips et al. (1997)).

In this paper, FPCA together with a clustering approach are used to highlight common characteristics of data and to summarize these characteristics by few components. We analyze the waveforms similarity on over 32 earthquakes recorded by the station of Augusta (South East Sicily) and observed for 10 seconds, by intervals of 0.02 seconds.

### 3.1 Clusters of waveforms

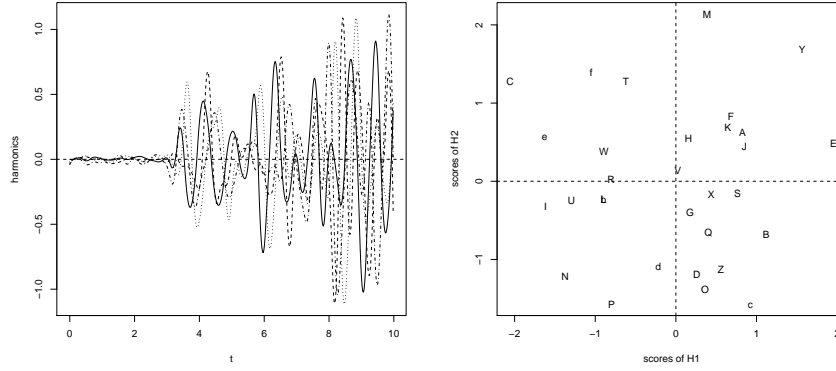
To construct functional data objects we specify a set of basis functions and a set of coefficients defining a linear combination of these basis functions (see fig. 1). We use B-spline basis (Hastie (1997)), since they are often used for non-periodic functions. B-spline basis functions are polynomial segments jointed end-to-end at argument values called knots, breaks or join points; the segments have specifiable smoothness across these breaks. B-spline basis functions have the advantages of very fast computation and great flexibility.



**Fig. 1.** Signal interpolation and B-spline-basis function for each event

Through this analysis we want to summarize the information given by all the 32 curves in few curves, explaining a large part of the variability related to each event.

We found that the first four harmonics explain almost the 60% (20%, 16%, 12%, 10%) of the variance of data. The average of the eigenvalues (mean scores) of the first four harmonics  $H_1$ ,  $H_2$ ,  $H_3$  and  $H_4$  are reported in table 1; these scores indicate the relative importance of each mode of variation. Since the remaining components are unnecessary they are not included in the model to avoid any over-fitting problems.



**Fig. 2.** Mean Scores of the First four harmonics and Scores scatterplot on the first principal plane

H1	H2	H3	H4
0.8839	0.7816	0.6201	0.5948

**Table 1.** Mean scores (absolute value) of the First four harmonics

The first four harmonics and the scores relative to the first principal plane, indicated with letters from **A** to **f** each corresponding to a single event and then to each curve, are reported in fig. 2, showing a possible clustering of curves according to these directions.

#### 4 Clustering of curves

In this section we use a variation of the trimmed  $k$ -means clustering algorithm proposed by García-Escudero and Gordaliza (2005). Their algorithm is a kind of robust version of  $k$ -means methodology through a trimming procedure. In few words, given a  $q$ -variate data sample  $\mathbf{X}_1, \dots, \mathbf{X}_n$  with  $\mathbf{X}_i = X_{i1}, \dots, X_{iq}$ , and fixed the number of clusters  $k$ , the trimmed  $k$ -means clustering algorithm looks for the  $k$  centers  $C_1, \dots, C_k$  that are solution of the minimization problem:

$$O_k(\alpha) = \min_Y \min_{C_1, \dots, C_k} \frac{1}{[n(1-\alpha)]} \sum_{X_i \in Y} \inf_{1 \leq j \leq k} \|X_i - C_j\|^2 \quad (3)$$

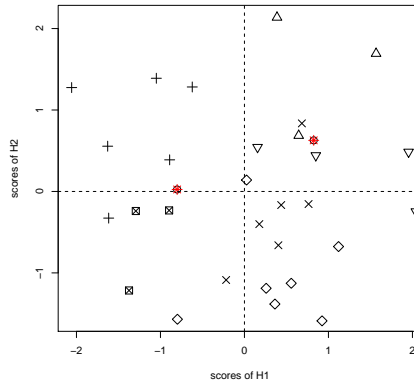
with  $\alpha$  the trimming size and  $Y$  is the set of subsets of  $X_1, \dots, X_n$  containing  $[n(1-\alpha)]$  data points, where  $[x]$  is the integer part of  $x$ . This method allocates each non-trimmed observation to the cluster identified by its closest center

$C_j$ , dealing with possible outliers by the given proportion of observations to be discarded  $\alpha$ . Their curve clustering procedure is based on a least-squares fit to cubic B-spline  $q$ -dimensional functions bases, applying the trimmed  $k$ -means clustering (3) on the resulting coefficients.

Here we use the PCA functions defined in (2) to get a linear approximation of each curve by a finite  $p$  dimensional vector of coefficients defined by the PCA scores. The number of starting clusters  $k$  is determined on the basis of a simple procedure based on the scores volume, such that we assign events to the clusters defined by events that have a distance less than a fixed threshold in the space of PCA scores. Once  $k$  is obtained we use the modified trimmed  $k$ -means algorithm, that consider the matrix of PCA scores instead of the coefficients of a linear fitting to B-spline bases. In other words we look for clusters in the direction of data with largest variance.

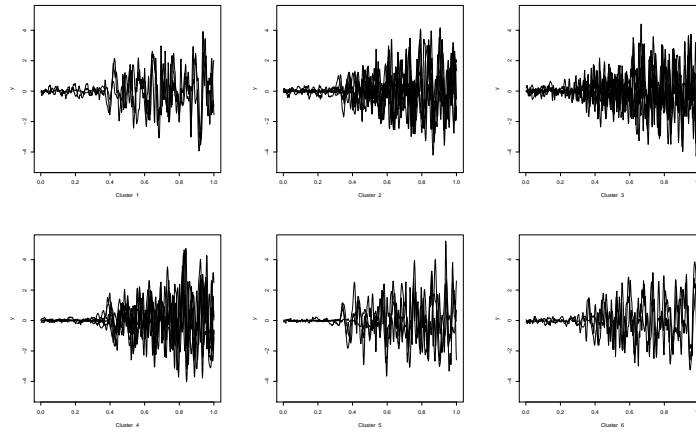
In particular for a better fitting we consider the first four harmonics, that explain almost the 60% of the global variance of data. Actually from empirical results, we observed that the choice of the number of harmonics does not influence significantly neither the value of the object function nor in terms of clustering results.

Fixing  $\alpha = 0.05$ , our procedure defines six clusters of earthquakes with common features and two outlier curves. The clusters defined on the first principal plane are reported in figure 3; the waveforms assigned to the six clusters and the two curves with outlying behavior are showed in figure 4 and 5, respectively.

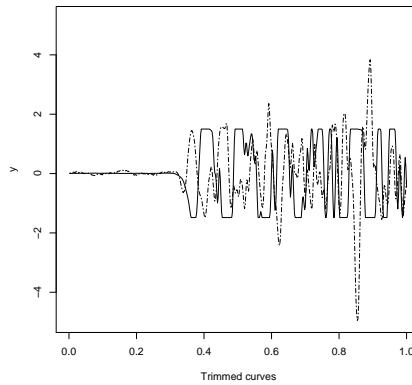


**Fig. 3.** Clusters of events reported on the first principal plane; \* symbol is used to identify the trimmed curves

Finally it seems interesting to note that by using the FPCA clustering approach, the value of the object function defined in (3) is 2.24, while if we



**Fig. 4.** Waveforms for each cluster



**Fig. 5.** Trimmed waveforms

use the clustering algorithm introduced by García-Escudero and Gordaliza (2005), that is based on the coefficients of B-splines fitting, we have  $O_k(\alpha) = 4.8072$ .

## 5 Conclusion

Functional data are a very convenient approach to deal with data depending on time, providing theoretical tools necessary to analyze observations of the

process recorded in discrete time intervals and convert them to continuous functions.

In particular we focus on the necessity of finding similar features of waveforms data recorded for earthquakes at different time instants. Indeed we think that eventual similarity between these functions would suggest similar behavior of the source process of the corresponding earthquakes.

Therefore we propose a variation of the algorithm for curves clustering proposed by García-Escudero and Gordaliza (2005), considering the directions defined by an application of PCA to functional data.

We think that the advantage of the procedure is related to an immediate use of PCA for functional data avoiding some objective choices related to the splines fitting, since cluster results for the different fits based on B-splines, Fourier or power basis, can differ considerably (Tarpey (2007)).

## References

- CHIODI, M. (1989): The clustering of longitudinal data when time series are short, *Multivariate data analysis*, 445–453.
- GARCÍA-ESCUADERO, L. A. and GORDALIZA, A. (2005): A proposal for robust curve clustering. *Journal of classification*, 22, 185–201.
- GILLARD, D., RUBIN, A.M. and OKUBOM, P. (1996): Highly concentrated seismicity caused by deformation of Kilauea’s deep magma system. *Nature*, 384, 343–346.
- HASTIE, T. J. (1997): *Generalized additive models*. Chapman and Hall, London.
- MENKE, W. (1999): Using waveform similarity to constrain earthquake locations. *Bull. Seismol. Soc. Am.*, 89, 1143–1146.
- MEZCUA, J. and RUEDA, J. (1994): Earthquake relative location based on waveform similarity. *Tectonophysics*, 233, 253–263.
- PHILLIPS, W.S., HOUSE, L.S., FEHELER, J. (1997): Detailed joint structure in a geothermal reservoir from studies of induced microearthquake studies. *Journal of Geophysical Research*, 102, 745–763.
- RAMSEY, J. O. and SILVERMAN, B. W. (2006): *Functional Data Analysis*. Springer, New York.
- SANGALLI, L. M., SECCHI, P., VANTINI, S., and VITELLI, V. (2008): K-means alignment for curve clustering. *MOX (Modeling and Scientific Computing)-Report*, 13.
- TARPEY, T. (2007): Linear transformations and the  $k$ -means clustering algorithm: applications to clustering curves. *American statistician*, 61(1), 34–40.



# Symbolic Data Analysis of Complex Data: Application to nuclear power plant

Filipe Afonso<sup>1</sup>, Edwin Diday<sup>2</sup>, Norbert Badez<sup>3</sup>, and Yves Genest<sup>3</sup>

<sup>1</sup> SYROKKO

5 rue de Copenhague BP13918, 95731 Roissy CDG, France, *afonso@syrokko.com*

<sup>2</sup> CEREMADE, Universite Paris Dauphine

Pce du M. Lattre de Tassigny 75775 Paris, France *diday@ceremade.dauphine.fr*

<sup>3</sup> EDF DTG-CEAN

12 rue saint Sidoine 69003 Lyon, France *norbert.badez,yves.genest@edf.fr*

**Abstract.** Complex data are here composed by several data tables describing different kinds of observations. The fusion of these data in order to get new knowledge requires to use symbolic data which are an extension of standard numerical or categorical data in order to loose less information than means by using intervals, distributions, sets of categories and the like. Symbolic Data Analysis (SDA) have been studied in recent books as Billard and Diday (2006), Diday and Noirhomme (2008), and enables to build and analyze such data. SDA of complex data is illustrated by the study of the degradation problems occurring on nuclear power plant cooling towers. Different kinds of measures have been collected by the French energy company EDF since the construction of each cooling tower. Several data tables describe cracks, corruptions, subsidence taking care of the shape. The fusion of these heterogeneous measures results in a symbolic data table containing in each cell histograms and intervals. SDA has shown to be suitable in order to study data on buildings degradation, performing the combination of the measures, discovering the correlations between them, highlighting and analyzing different problems of degradation, ordering and classifying the deteriorations of the towers.

**Keywords:** symbolic data analysis, complex data, data visualization, industrial application

## 1 Introduction

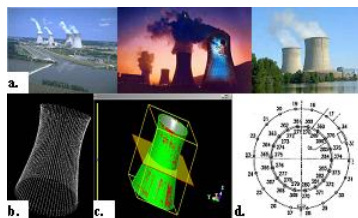
On the web site of ECML/PKDD 2007 on Mining Complex Data, "complex data" have been defined in the following way: "in contrast to the typical tabular data, complex data can consist of heterogeneous data types, can come from different sources, or live in high dimensional spaces. All these specificities call for new data mining strategies". In practice, sometimes "complex data" refer to complex objects like images, video, audio or text documents. Sometimes, it refers to distributed data or structured data or more specifically: spatial-temporal data or heterogeneous data as a mixture of data, for example, a medical patient described by images, text documents and socio-demographic information. In practice, complex data are more or less based

on several kinds of observations described by standard numerical or (and) categorical data contained in several related data tables. This is the case of the data on power plants cooling towers described in this paper. When it is needed to describe classes or categories of observations in order to merge such data for knowledge discovery, symbolic data (described hereunder) naturally appears. The usual data mining model is based on two parts: the first concerns the observations, the second, contains their description by several standard variables including numerical or categorical. The Symbolic Data Analysis (SDA) model (see Billard and Diday (2006), Diday and Noirhomme (2008)) needs two more parts: the first concerns units called concepts and the second concerns their description by symbolic data. The concepts are characterized by a set of properties called intent and by an extent defined by the set of observations which satisfy these properties. In order to take care of the variation of their extent, these concepts are described by symbolic data which are standard categorical or numerical data but moreover intervals, histograms, sequences of weighted values and the like. These new kinds of data are called symbolic as they cannot be manipulated as numbers. Then, based on this model, new knowledge can be extracted by new tools of data mining extended to concepts considered as new kinds of observations. The case study on complex data enters in the framework of building surveying (see for example Kavanagh (2009)) and aims at putting in evidence the added value of SDA in a nuclear power plant data base (see for example Nag (2007)). In this paper, we first describe the power plant cooling towers data: several data tables which describe cracks, corruptions, subsidence taking care of the shape. Then we give the objective which is mainly to compare the towers and to study the correlations between the measures of the different data tables. Then we present our strategy based on the fusion of the complex initial data by symbolic data and finally we give some of the obtained results by applying SDA methods in order to extract new knowledge.

## 2 Application : the degradation of buildings

### 2.1 The Data

We study data on degradation of nuclear power plant cooling towers provided by the French energy company EDF (See examples fig. 1 a.). EDF Company collects surveying data since the construction of each tower in order to analyze their degradations. The twenty-one cooling towers are described by eight different files corresponding to four different categories of data: i) Geometric distance: EDF has collected, with scanning methods, clouds of points describing the external hyperbolic shapes of the cooling towers (See fig. 1 b.). For each tower, we obtain data files where the observations are three-dimensional points (height, radius, and angle) and the variable describing each observation is the difference between the measured point and the theoretical point or between two measures at two different years (See fig. 1 c.); (ii) Data on



**Fig. 1.** **a.** Nuclear power plant cooling towers. **b.** Cloud of points describing the cooling tower. **c.** Calculation of the geometric distances between two states of the cooling tower. **d.** Measures of subsidence for different vertical sections (angles).

cracks : each crack is described with the extremity points of the crack (three-dimensional points), its length, its width, its orientation (horizontal, vertical or oblique line). We have two files describing the cracks for two different years; (iii) Data on corrosion : each corrosion is described with the extremity points of the corrosion (three-dimensional points), its length, its width. We have two files describing the corruptions for two different years; (iv) Data on subsidence : the subsidence is measured every year for different sections (different angles) of the towers (see fig. 1 d.) so that we have, for each tower, a data file with the angles as observations (from 20 to 50 measures depending on the tower) described by several measures of subsidence over time.

## 2.2 The Objectives

To study the degradation of the cooling towers, we need to study simultaneously in a same data analysis the four categories of information (geometric distance, cracks, corruptions, subsidence) stored in eight files, that means  $21 \times 8 = 168$  files for the 21 cooling towers. These files are totally heterogeneous in their format. Above all, the observations are not the same in the different files. We have here complex data composed by several data tables with different kinds of observations described by different variables.

Moreover, "surveying data" need to be handled at different levels of generality to compare the towers and to study in detail each tower from different angles and heights in order to answer questions as : which towers are degraded? What are the damages? Which sections of the towers are degraded? SDA is able to easily switch from one concept to another which represent different kinds of generalization: the towers, the horizontal sections of a same tower, the vertical sections of a same tower, any divisions of a same tower.

## 2.3 From complex data to symbolic data

Merging these data tables in a unique one requires to use symbolic data which extend standard numerical or categorical data to data which keep more information than means for instance. In order to compare the degradation of

the respective cooling towers, we consider the cooling towers as concepts. In fig. 2, for each cooling tower, data on geometric distance are aggregated up to the respective tower. Quantitative variables (height, radius, distance) can be aggregated up into intervals (height, radius) or into quantitative histograms (distance). In the same way, fig. 3, data on cracks are aggregated to the respective tower. Quantitative variables can be aggregated into intervals or into histograms (lengthCracks) and qualitative variables are aggregated into qualitative histograms (orientation). We note that we can add classical variables describing specifically each tower as the number of cracks by tower. Finally, the data on corruptions and the data on subsidence are added to the previous data table in the same way too. The result is a unique data table describing the concepts "towers" with all the surveying measures aggregated up to the level of the towers.

Nevertheless, discovering problems of deterioration in some towers and studying the correlations between the different measures at the level of the concepts "towers" are necessary to understand physical phenomena but it is not sufficient for the engineers to bring solutions to the problems occurring over time in cooling towers. Engineers need to know which sections to repair so that they have to know the most damaged sections and they have to understand the correlations between the different kinds of damages in the different sections of a same construction. It is then possible to merge all the surveying data files describing the damages of a given tower into a unique symbolic data table describing the different sections of a tower. Fig. 4, we merge all the measures (geometric distance, cracks, corruptions, subsidence) into a unique symbolic data table describing the concept "vertical sections of the tower". This is possible by discretization of the variable "Angle" which is common to all files. For each vertical section, the surveying measures are aggregated up to the corresponding vertical section using symbolic descriptions.

## 2.4 Some results

These tables can be now analyzed by SDA methods taking into account all the measures, thanks to the academic Sodas Software and the professional SYR software. The reader may refer to Diday and Noirhomme (2008) for a complete presentation of SDA methods developed by two European projects until 2003. More recently, the SYR software has been developed by SYROKKO company. The first aim of SYR is to extract, from a data base of several millions of observations a reduced number of units which are concepts summarizing and merging the initial data tables by symbolic data. Then SYR software (2009) create, handle (select, cut, move rows or columns) and visualizes a symbolic data file thanks to user-friendly graphical output and produces new knowledge by SDA tools. For instance, the SYR software offers a symbolic data spreadsheet called TabSyr, where each row corresponds to a concept and each column corresponds to a classical numerical or categorical variable or to an interval or an histogram-valued variable describing

**Concept « Cooling Tower »****First category of data files: the geometric distances**

For each tower, we have

- File 1 : geometric distances between two years (called Gdist\_1\_2)
- File 2 : geometric distances between the measured points at a first year and the theoretical points (Gdist\_t\_1)
- File 3 : geometric distances between the measured points at a second year and the theoretical points (Gdist\_t\_2)

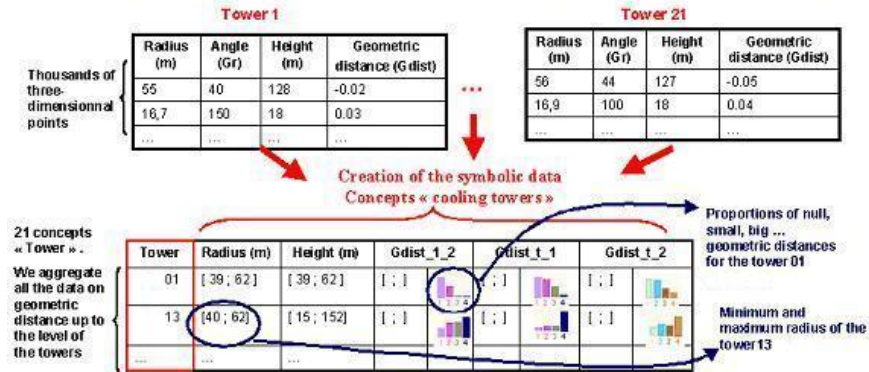


Fig. 2. Creation of the concept "cooling tower" described by symbolic data.

**Concept « Cooling Tower »****Second category of data files: the cracks**

For each tower, we have:

- 1 File: data describing the cracks at a first year
- 1 File: data describing the cracks at a second year

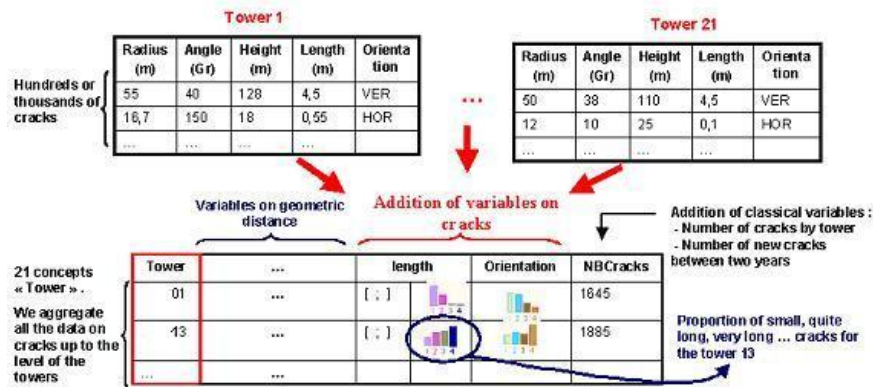
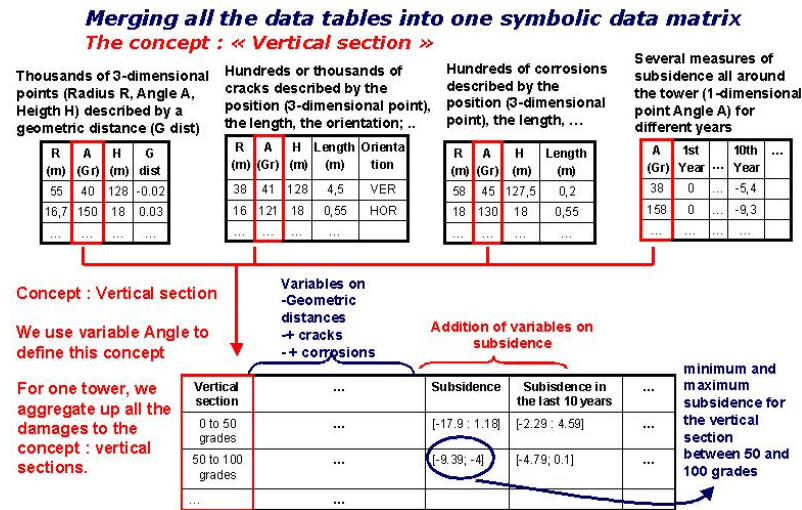


Fig. 3. At the level of the concept "cooling tower", we merge all the data files in an unique table thanks to symbolic data.

the concepts (see fig. 5). The software is efficient to compare the different descriptions of the concepts cooling towers, to synthesize the data in a relevant and comprehensive form, to see the more or less deteriorated towers

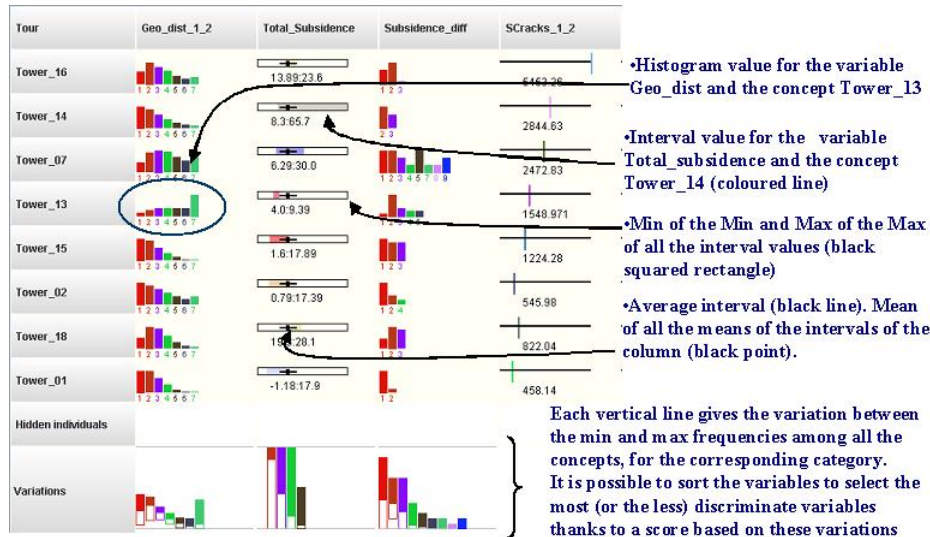


**Fig. 4.** Thanks to the common variable "Angle", we merge all the data files in an unique symbolic data table .

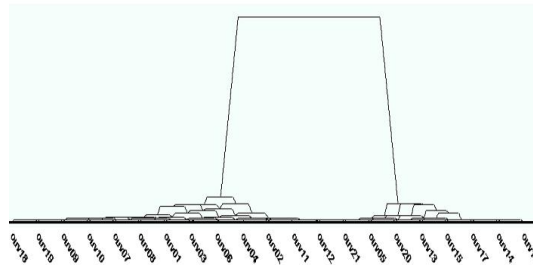
thanks to scoring methods, to see unusual or extreme values, to select the most discriminate variables.

We give some examples of methods applied to the symbolic data table describing the towers. It is for instance interesting to compare the different measures. fig. 6, we perform a pyramid on the symbolic data table describing the towers thanks to the Sodas software. The symbolic pyramid method generalizes the hierarchical classification method to overlapping clusters and symbolic data (see Brito (2002)). We can see here two clusters standing out. These two clusters are described by vectors of symbolic data and displayed together by the View module of the same software, see fig. 7. This module provides a visualization as a star where each branch, associated with a symbolic histogram or interval variable, superimposes the values of the two clusters. It is then easy to compare them and we can observe that the pink cluster is the cluster with more subsidence (up than 12.8 cm), more geometric distances (*geo.dist* between 0.05m and 0.1m). The geometric differences are positive (*relative.diff* between 0 and 0.10 m). In the green cluster, the geometric differences are smaller and they are both positive and negative (*relative.diff* between -0.05 and 0.05 m). Moreover, we observe more big cracks (*cracks.length* between 0.5m and 3m) in the pink cluster.

Finally, a comprehensive study of these correlations has been achieved and a linear statistical model combining several weighted degradation variables in order to estimate the degradation of the towers and anticipate possible problems was performed. This model is a combination of histogram (*geo.dist*, *difference.diff*), interval (*Total.subsidence*) and classical quantitative (*SCracks*, sum of the lengths of the cracks) variables. In fig. 5, the



**Fig. 5.** Visualization of the symbolic data table by TABSYR. The rows are sorted from the most to the less damaged tower thanks to a model, combination of different kinds of symbolic variables (histograms, intervals, standard variables...)



**Fig. 6.** Pyramidal classification applied to a symbolic data table, concept tower

rows of the symbolic data table are sorted from the most to the less damaged tower thanks to this model.

### 3 Conclusion

The study of physical phenomena as deteriorations of buildings presented in this paper requires multiple related data tables describing different kind of observations. We can consider these data as complex data because of the difficulty to link the tables. It is this ability to rationally merge multiple data tables into a single symbolic data table by the emergence of "concepts" as new observations that makes the SDA approach relevant to the study of these complex data. The Sodas and SYR softwares are able to study these



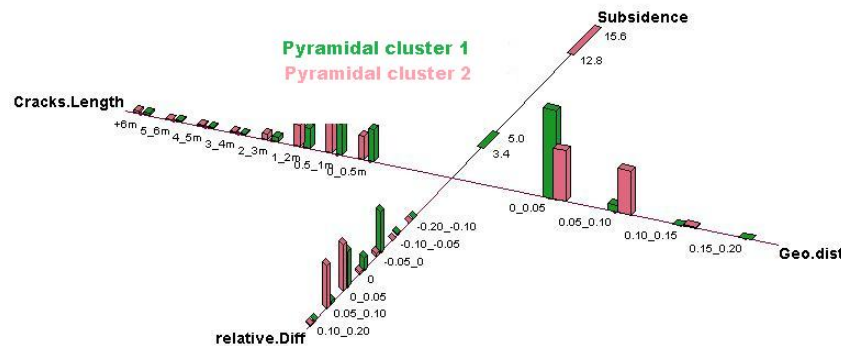


Fig. 7. Comparison of two symbolic descriptions with the Sodas software

concepts described by numerical, interval-valued and histogram-valued variables. SDA has shown to be suitable in order to study data on buildings degradation by performing the combination of the measures, discovering the correlation between them, highlighting and analyzing different problems of degradation as they yield to detect, classify and sort the deteriorations of the towers or their sections. Finally, theoretical and practical SDA tools are helpful to choose the appropriate concepts, to switch from a level of generality to another (damage  $\rightarrow$  section of tower  $\rightarrow$  tower), to synthesize the data in a relevant and comprehensive way, to apply many well known statistical methods already extended to symbolic data.

## 4 Acknowledgments

The authors would like to thanks the region Ile de France and its service de valorisation de la recherche, for its support of the S3 project directed by C. Cremona (LCPC, Paris) which the authors want to thank also for his advices.

## References

- BILLARD, L. and DIDAY, E. (2006): *Symbolic Data Analysis: conceptual statistics and data Mining*. Wiley, 330 p., ISBN 0-470-09016-2
- BRITO, P. (2002): *Hierarchical and Pyramidal Clustering for Symbolic Data*, *Journal of the Japanese Society of Computational Statistics*, Vol. 15, 2, 231-244.
- BRITO, P., BERTRAND, P., CUCUMEL, G. and DE CARVALHO, F. (2007): *Selected contributions in Data Analysis and Classification*. Springer, 634 p.
- DIDAY, E. and NOIRHOMME, M. (eds) (2008): *Symbolic Data Analysis and the SODAS software*. Wiley, 457 p., ISBN 9780-470-01883-5.
- KAVANAGH, B.F. (2009): *Surveying with Construction Applications*. 7th Edition, Prentice Hall, 704p., ISBN 978-0135000519
- NAG, P. K. (2007): *Power Plant Engineering*. 3rd Edition, McGraw-Hill, 992p., ISBN 978-0070648159



# Different P-spline Approaches for Smoothed Functional Principal Component Analysis

Ana M. Aguilera, M. Carmen Aguilera-Morillo, Manuel Escabias and  
Mariano J. Valderrama

Department of Statistics and O.R. University of Granada.  
Campus de Fuentenueva, 18071-Granada, Spain, *aaguiler@ugr.es*

**Abstract.** In order to reduce the dimension and to explain the dependence structure of a functional data set in terms of uncorrelated variables, it is usual to use Functional Principal Component Analysis (FPCA). When the sample curves are not smooth enough, the principal component curves have a lot of variability and are difficult to interpret. Regularized FPCA continuously controls the degree of smoothness by introducing a roughness penalty in his own formulation. In this paper we consider two different forms of smoothed FPCA, both of them based on penalized splines (P-splines) smoothing with B-splines basis. They differ in that the first applies the roughness penalty in the construction of principal components whereas the second incorporates it in the approximation of sample curves and then carries out an unsmoothed FPCA.

**Keywords:** functional data, principal component analysis, B-spline expansion, P-splines

## 1 Introduction

Functional data analysis (FDA) methods have received much attention in the last years generalizing a lot of multivariate techniques to the functional field where data are a sample of curves instead of vectors as in classic multivariate data analysis. A detailed study and interesting applications related with this topic can be seen in Ramsay and Silverman (2002 and 2005).

FPCA is a reduction dimension technique very useful for exploring the main features of a set of sample curves. One usual form of estimating FPCA from discrete observations of the sample curves is based on basis expansion approximation (Ramsay, 2005). This way, FPCA of a set of curves is reduced to multivariate PCA of a transformation of the matrix of basis coefficients (Ocaña et al., 2007). Cubic spline interpolation with B-splines basis was considered in Aguilera et al. (1996). Quasi-natural cubic spline interpolation has been applied to estimate the risk of drought from FPCA of annual temperature curves (Escabias et al., 2005). On the other hand, least squares approximation of stress level curves has been recently used to estimate FPCA and predict the occurrence of a lupus flare in patients suffering this disease (Aguilera et al., 2008).

Usually, observed records are not very smooth and, consequently the principal component curves show substantial variability. In this case, there is a clear need of smoothing of the estimated principal component curves. The smoothed or regularized FPCA contemplates the application of smoothing to functional principal components analysis. To control the degree of smoothness, Silverman (1996) considered as roughness penalty the integrated squared second derivative of the principal component curves. An application of this method with regularized Gaussian basis expansions is available in Kayano and Konishi (2008).

In this paper we provide two different versions of smoothed FPCA, both of them based on penalized splines (P-splines) smoothing (Eilers and Marx, 1996) and B-splines basis expansion of sample curves. First, we propose to carry out an unsmoothed FPCA on the P-spline smoothing of sample curves. Second, we incorporate a discrete penalty in the formulation of the FPCA based on the order two differences of the basis coefficients of principal components. The accuracy of the estimates obtained with both methods has been tested in a simulation study.

## 2 P-spline smoothing of functional data

Let  $x_1(t), x_2(t), \dots, x_n(t)$  be a sample of functions, which are the sample information related to a functional variable. We can consider them like observations of a stochastic process of second order  $X = \{X(t) : t \in T\}$ , continuous in quadratic mean whose sample functions belong to Hilbert space  $L^2(T)$  of square integrable functions with the usual inner product

$$\langle f, g \rangle_u = \int_T f(t) g(t) dt, \quad \forall f, g \in L^2(T).$$

It is difficult to observe functions in continuous time so that we have observations of this functions in a finite set of times  $\{t_{i0}, t_{i1}, \dots, t_{im_i} \in T, i = 1, \dots, n\}$ , that can be different for each individual. Then, the sample information is given by the vectors  $x_i = (x_{i0}, \dots, x_{im_i})'$ , with  $x_{ik}$  the observed value for the sample path,  $x_i(t)$ , in the time  $t_{ik}$  ( $k = 0, \dots, m_i$ ).

To reconstruct the functional form of sample paths from the discrete observed data, we can use several methods depending on how the functional data has been observed and the main characteristics of curves.

One usual solution is to assume that sample paths belong to a finite-dimension space generate by a basis  $\{\phi_1(t), \dots, \phi_p(t)\}$ , so that they are expressed as

$$x_i(t) = \sum_{j=1}^p a_{ij} \phi_j(t), \quad i = 1, \dots, n.$$

In this paper we work with B-splines bases (De Boor, 1977) which are appropriate for smooth curves. However, there are other useful basis systems

like Fourier bases for periodic data, piecewise constant bases for counting processes or wavelets bases for curves with strong local behavior.

The most used method to estimate the basis coefficients when data are observed with error is least squares approximation. The regression splines so obtained are not very smooth. In order to control the degree of smoothness it is successful to introduce a roughness penalty in the least squares function. A natural measure of the roughness of a function is its integrated squared second derivative (Ramsay, 2005). Following the work of Eilers and Marx (1996), we use a penalty based on differences of order  $d$  between the adjacent B-splines basis coefficients. The functions estimated in this way are called P-splines

P-splines are innovative in the sense that they penalize directly the basis coefficients of the curves instead of penalizing the curve, which reduce the problem dimension. This kind of penalty is more flexible so that it doesn't depend of the degree of the B-splines used in the approximation. In addition, it is a discrete approximation of the integrated squared second derivative over the interval of interest.

P-spline smoothing of each sample curve is obtained by minimizing the following penalized least squares function:

$$S(x_i, \lambda_i | a_i) = (x_i - \Phi_i a_i)' (x_i - \Phi_i a_i) + \lambda_i a_i' P_d' P_d a_i$$

whose solution is

$$\hat{a}_i = (\Phi_i' \Phi_i + \lambda_i P_d' P_d)^{-1} \Phi_i' x_i,$$

with  $P_d = (\Delta^d)$ ,  $\Phi_i = (\phi_j(t_{ik}))_{m_i \times p}$  and  $a_i = (a_{i1}, \dots, a_{ip})'$ .

### 3 Functional Principal Components Analysis

In order to reduce the infinite dimension of a functional predictor and to explain its dependence structure by a reduced set of uncorrelated variables, multivariate PCA is extended to the functional case.

When sample curves are not very smooth, the principal components curves can present a lot of variability and have difficult interpretation. To solve this problem, in this work we carry out two different ways of introducing smoothing in FPCA. The first consists of FPCA of the P-spline smoothing of sample curves.

Let us assume that the observed curves,  $x_i(t)$ , result from subtraction the mean function value, so that the sample mean,  $n^{-1} \sum_i x_i(t)$ , is zero.

In general, the  $j$ -th principal component is given by

$$\xi_{ij} = \int_T x_i(t) f_j(t) dt, \quad i = 1, \dots, n$$

where the weight function or loading  $f_j$  is obtained by solving

$$\begin{aligned} & \text{Max}_f \text{Var} \left[ \int_T x_i(t) f(t) dt \right] \\ & \text{r.t. } \left\{ \|f\|^2 = 1 \text{ and } \int f_\ell(t) f(t) dt = 0 \ell = 1, \dots, j-1 \right\}. \end{aligned}$$

The weight functions are the solutions to the eigenequation

$$Cf = \lambda f,$$

with  $C$  being the *Covariance Operator* defined by

$$Cf = \int \hat{c}(\cdot, t) f(t) dt,$$

in terms of the sample covariance function  $c(s, t) = \frac{1}{n} \sum_{i=1}^n x_i(s) x_i(t)$ .

Then, curves are expressed in base to functional principal components as

$$x_i(t) = \sum_{j=1}^{n-1} \xi_{ij} f_j(t).$$

Let us now suppose that the sample paths are expressed in terms of basis functions as  $x = A\phi$ , with  $A = (a_{ij})$  the coefficients matrix,  $\phi = (\phi_1, \dots, \phi_p)'$  and  $x = (x_1, \dots, x_n)'$ . Then, we can see in Ocaña et al. (2007) that FPCA is equivalent to the multivariate PCA of  $A\Psi^{\frac{1}{2}}$  matrix, where  $\Psi = (\Psi_{ij})_{n \times p} = \int \phi_i(t) \phi_j(t) dt$ . In this paper matrix  $A$  will be the matrix of basis coefficients of the P-spline approximations of sample curves presented in previous section.

## 4 Smoothing Functional Principal Components Analysis via P-splines

The second way of smoothing in FPCA which is proposed in this work is a version of regularized PCA carried out in Silverman (1996). Instead of applying a continuous penalty in the construction of principal components, we propose a discrete penalty based on P-splines whit B-splines basis. An application of regularized FPCA with continuous penalty is applied to actuarial science in Segovia-Gonzales et al. (2009).

Let us consider  $f(t) = \sum_{k=1}^p b_k \phi_k(t) = \phi(t)' b$ , where  $\phi(t)$  is a B-splines basis. The *Roughness Penalty Function* is defined by

$$PEN_d(f) = b' P_d' P_d b,$$

with  $P_d = (\Delta^d)$  and  $b = (b_1, \dots, b_p)$ . In the example studied in this paper we will use  $PEN_2(f)$  that is a good discrete approximation of the integrated squared second derivative of  $f$ .

Given  $\xi = \int x(t) f(t) dt$ , its sample penalized variance is

$$PCAPSV(f) = \frac{\text{var}[\int x(t) f(t) dt]}{\|f\|^2 + \lambda PEN_d(f)},$$

whit  $\lambda$  a smoothing parameter which regulates the importance of the roughness penalty term.

If we consider the basis expansion of each  $x_i(t)$  and  $f(t)$  in terms of B-splines functions, then

$$PCAPSV(f) = \frac{b' \Psi V \Psi b}{b' (\Psi + \lambda P_d' P_d) b},$$

where  $V = n^{-1} A' A$  and  $A = (a_{ij})_{n \times p}$ .

Then, the  $j$ -th principal component is given by

$$\xi_{ij} = \int_T x_i(t) f_j(t) dt, \quad i = 1, \dots, n$$

and we must estimate  $f_j$  so that

$$\text{Max}_f [PCAPSV(f)]$$

$$\text{r.t. } \{\|f\|^2 = b' \Psi b = 1 \text{ and } b' \Psi b_l + b' P_d' P_d b_l = 0, \quad l = 1, \dots, j-1\}.$$

This variance maximization problem is then converted into an eigenvalues problem

$$\Psi V \Psi b = \beta (\Psi + \lambda P_d' P_d) b.$$

By applying the SVD or Choleski factorization  $LL' = \Psi + \lambda P_d' P_d$ , we have to solve the eigenvalue problem

$$(L^{-1} \Psi V \Psi L^{-1'}) u = \beta u,$$

with  $L'b = u$ . This way, P-spline smoothed PCA is reduced to classic PCA of the vectors of coefficients  $L^{-1} \Psi a_i$ .

A very important thing when we use an smoothed PCA is to select the suitable smoothing parameter  $\lambda$ . There are different model selection criteria which can be used to select  $\lambda$ . The most known are CV (*cross validation*), GCV (*genelized cross validation*) and AIC (*Akaike information criterium*).

In this paper, we have adapted the generalized cross validation method (Ramsay and Silverman, 2005) by considering the discrete penalty based on differences between adjacent coefficients.

## 5 Simulated example

The aim of this study is to check the good behavior of the proposed penalized PCAs to rebuild the original principal components curves from noisy data.

The stochastic process considered for simulating the sample curves is

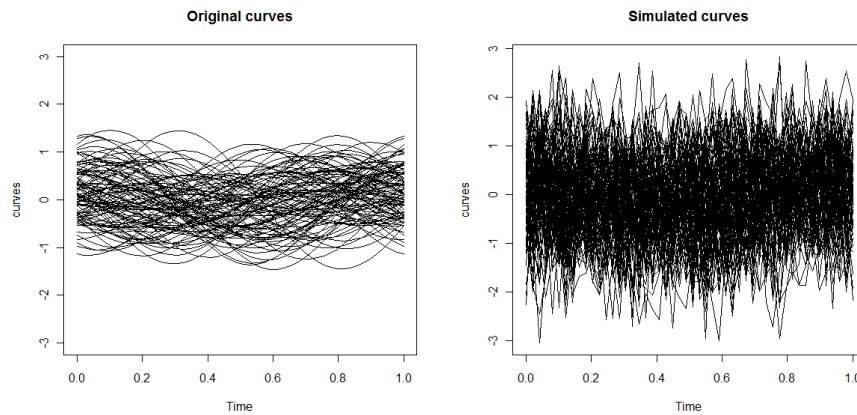
$$X(t) = R \cos(2\pi t + \theta),$$

where  $R$  and  $\theta$  are independent random variables with *Raileight*(0.5) and *Uniform*(0,  $2\pi$ ) distributions, respectively.

In this example we have simulated  $n = 100$  sample paths on  $m = 50$  equally spaced time points at the interval  $[0, 1]$ . In the simulation we have added at every time point a random error generated from a normal distribution with mean zero and variance 0.5.

The covariance function of the stochastic process  $X$  has a unique eigenvalue with multiplicity two,  $\lambda = 0.25$ , with associated eigenvectors given by  $f_1(t) = \sqrt{2} \sin(2\pi t)$  and  $f_2(t) = \sqrt{2} \cos(2\pi t)$ .

As we can see in Figure 1 simulated data are very noisy so that a roughness penalty approach is necessary to estimate FPCA.

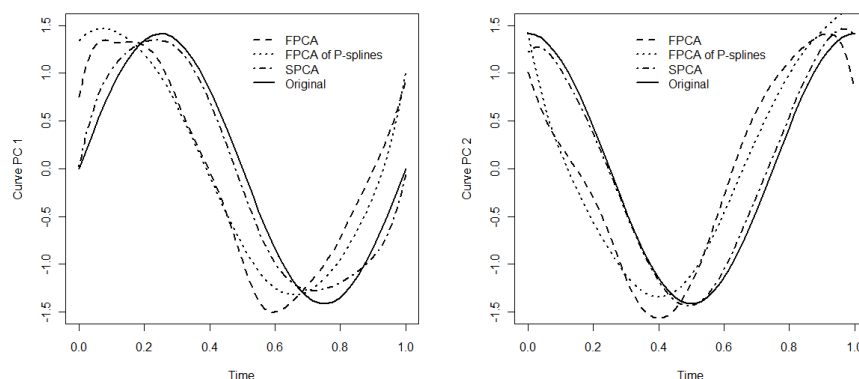


**Fig. 1.** Original sample curves of  $X$  (left) and simulated noisy sample curves (right).

In Figure 2 we can see the estimations of the principal component curves with the different FPCA approaches considered in this paper. The smoothing parameters used in the approximation have been  $\lambda = 4.5$  for the P-spline approximation of sample curves and  $\lambda = 0.005$  for the P-spline smoothing of PCA. In order to compute all the approximations we have considered ten

equally spaced knots in the interval  $[0,1]$  to define the cubic B-spline functions ( $p=12$ ).

Let us observe in Figure 2, the good behavior of Smoothed PCA has been shown. We can see as SPCA provides principal component curves (dashed lines with dots) closer to the original ones (continuous lines) than others used methodologies.



**Fig. 2.** First and second principal component (PC) curves of  $X$ . Original PC curves (continuous lines) and estimated PC curves from FPCA of B-spline approximation (dashed lines), FPCA of P-spline approximation with  $\lambda = 4.5$  (dotted lines) and P-spline smoothed functional PCA (SPCA) with  $\lambda = 0.005$  (dashed lines with dots).

## 6 Acknowledgements

This research has been funded by project P06-FQM-01470 from "Consejería de Innovación, Ciencia y Empresa. Junta de Andalucía, Spain" and project MTM2007-63793 from Dirección General de Investigación, Ministerio de Educación y Ciencia, Spain.

## References

- AGUILERA, A. M., GUTIÉRREZ, R. and VALDERRAMA, M. J. (1996): Approximation of estimators of the PCA of a stochastic process using B-splines. *Communications in Statistics. Simulation and Computation* 25 (3), 671-691.
- AGUILERA, A. M., ESCABIAS, M. and VALDERRAMA, M. J. (2008): Discussion of different logistic models with functional data. Application to Systemic Lupus Erythematosus. *Computational Statistics and Data Analysis*, 53, 151-163.

- CRAVEN, P. and WAHBA, G. (1979): Smoothing noisy data with splines functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* 31, 377-403.
- DE BOOR, C. (1977): Package for calculating with B-splines. *Journal of Numerical Analysis* 14, 441-472.
- DURBAN, M. and DAE-JIN, L. (2008): *Splines con penalizaciones (P-Splines). Teoría y aplicaciones*. Universidad Pública de Navarra.
- EILERS, P. and MARX, B. (1996): Flexible smoothing with B-splines and penalties. *Statistical Science* 11, 89-121.
- ESCABIAS, M., AGUILERA, A.M. and VALDERRAMA, M.J. (2005): Modelling environmental data by functional principal component logistic regression. *Environmetrics* 16, 95-107.
- KAYANO, M. and KONISHI, S. (2008): Functional principal component analysis via regularized Gaussian basis expansions and its application to unbalanced data. *Journal of Statistical Planning and Inference* 139 (2009), 2388-2398.
- OCAÑA, F.A., AGUILERA, A.M. and ESCABIAS, M. (2007): Computational considerations in functional principal component analysis. *Computational Statistics* 22, 449-465.
- RAMSAY, J. O. and SILVERMAN, B. W. (1997, 2005): *Functional data analysis*. Springer-Verlag.
- RAMSAY, J. O. and SILVERMAN, B. W. (2002): *Applied functional data analysis: Methods and case studies*. Springer-Verlag.
- SEGOVIA-GONZALES, M.M., GUERRERO, F.M. and HERRANZ, P. (2009): Explaining functional principal component analysis to actuarial science whit an example on vehicle insurance. *Elsevier: Mathematic and Economics* 45 (2009), 278-285.
- SILVERMAN, BW. (1996): Smoothed functional principal component analysis by choice of norm. *Annals of Statistics* 24 (1), 1-24.



# Peak Detection in Mass Spectrometry Data Using Sparse Coding

Theodore Alexandrov<sup>1</sup>, Klaus Steinhorst<sup>1</sup>, Oliver Keszöcze<sup>1</sup>, and Stefan Schiffler<sup>1</sup>

University of Bremen, Center for Industrial Mathematics  
Bibliothekstr. 1, D-28334 Bremen, Germany *theodore@math.uni-bremen.de*

**Abstract.** Mass spectrometry is an important tool in the analysis of chemical compounds. A crucial step of mass spectrometry data processing is the peak detection which selects peaks corresponding to molecules with high concentrations. We present a new procedure of the peak detection based on a sparse coding algorithm, for which we propose an elastic-net modification. The evaluation with simulated data shows that using the sparse coding prototype spectra gives improvement over using per-class mean spectra, although the former ones are extracted in an unsupervised manner. Finally, we apply the procedure to a colorectal cancer and to a liver diseases datasets.

**Keywords:** mass spectrometry, peak picking, sparse coding

## 1 Introduction

Mass spectrometry is an important chemical technique and is a major tool in proteomics, a discipline interested in large-scale studies of proteins. Given a sample of serum or urine, a mass spectrometer produces a high dimensional histogram-like spectrum. The high sharp peaks of this spectrum contain information about molecules with high concentrations. A peak detection procedure calculates for a raw spectrum the so-called line spectrum which is a list of positions and intensities of the peaks. The peak detection is usually a starting and a crucial step of mass spectrometry data analysis. This paper presents a new procedure of the peak detection based on the sparse coding algorithm Lee et al. (2006). This procedure is motivated by recent successes of sparsity-enforcing algorithms in mass spectrometry. The key idea of the sparse coding Lee et al. (2006) is to represent vectors of a matrix in a linear subspace where not only the coefficients but also the basis vectors of the subspace are optimized. The optimization is performed minimizing (i) the distance between the original data and the found approximation and (ii)  $l_1$ -norm of the coefficients used in the representation.

Let us consider a set of spectra containing several classes, where one class is characterized by class-specific peaks with the same positions and similar heights taking place in each spectrum of this class. Our procedure of peak detection is as follows. First, the sparse coding algorithm finds a sparse

approximation of the dataset. Each spectrum is represented using a few basis vectors (of the same dimension as the spectra) which capture the common features of the spectra. Then a simple peak detection method is applied on each basis vector. Finally, we do an additional check which reduces the number of false positives.

Section 2 presents the peak detection procedure. Section 3 evaluates the procedure with simulated datasets. Additionally, a real-life colorectal cancer dataset is processed. Section 4 concludes the paper.

## 2 Proposed procedure of peak detection

### 2.1 Feature extraction using sparse coding

Let us consider a dataset of  $R$  spectra of length  $L$ , which belong to  $D$  classes ( $D \ll R$ ). Each class is characterized by peaks at the same positions and with similar heights. Given the matrix  $\mathbf{X} \in \mathbb{R}^{L \times R}$  with spectra in columns, the sparse coding algorithm Lee et al. (2006) represents each spectrum in terms of some basis where the coefficients are optimized to produce a sparse representation using

$$\begin{aligned} \min_{\mathbf{B}, \mathbf{S}} \quad & \frac{1}{2} \|\mathbf{X} - \mathbf{BS}\|_{\text{F}}^2 + \alpha \sum_j \|S_j\|_1, \\ \text{subject to} \quad & \|B_j\|_2^2 \leq C, \end{aligned} \quad (1)$$

with respect to a matrix of basis vectors  $\mathbf{B} \in \mathbb{R}^{L \times L}$  and a matrix of the corresponding coefficients  $\mathbf{S} \in \mathbb{R}^{L \times R}$ , where  $\|\cdot\|_{\text{F}}$ ,  $\|\cdot\|_1$ , and  $\|\cdot\|_2$  are the matrix Frobenius and the vector  $l_1$ - and Euclidean norms, respectively;  $S_j$  and  $B_j$  denote the  $j$ -th column of  $\mathbf{S}$  and  $\mathbf{B}$ , respectively. The hyperparameters are the regularization parameter  $\alpha$  and the boundary on the norm of a basis vector  $C$ .

The minimization problem (1) is solved in two steps. First, we learn the coefficients  $\mathbf{S}$  using the feature-sign search algorithm minimizing (1) for a fixed  $\mathbf{B}$ , then for the learned coefficients we optimize the basis  $\mathbf{B}$  using the Lagrange dual. For more details see Lee et al. (2006). At the end, each column  $X_j$  of  $\mathbf{X}$  is sparsely represented using only a few basis vectors  $B_j$  ( $j \in \mathcal{I}$ ), where  $\mathcal{I}$  are indices of non-zero rows of the matrix  $\mathbf{S}$ . We call these basis vectors prototype spectra.

The sparse coding algorithm inverts the matrix  $\mathbf{B}^T \mathbf{B}$  which is close to singular in our case. We propose approximating the matrix  $\mathbf{B}^T \mathbf{B}$  by  $\mathbf{B}^T \mathbf{B} + \beta \mathbf{I}$  for a small  $\beta$ , where  $\mathbf{I}$  is the identity matrix. In Alexandrov et al. (2009) we show that the algorithm Lee et al. (2006) with this modification solves not (1) but the following optimization elastic-net problem:

$$\begin{aligned} \min_{\mathbf{B}, \mathbf{S}} \quad & \frac{1}{2} \|\mathbf{X} - \mathbf{BS}\|_{\text{F}}^2 + \alpha \sum_j \|S_j\|_1 + \beta \sum_j \|S_j\|_2^2, \\ \text{subject to} \quad & \|B_j\|_2^2 \leq C. \end{aligned} \quad (2)$$

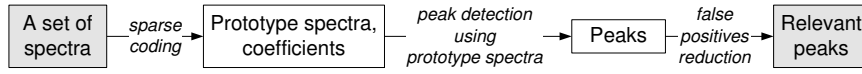


Fig. 1. The general scheme of the proposed peak detection procedure.

## 2.2 Peak detection in the features extracted

In general, the basis  $\{B_j\}_{j \in \mathcal{I}}$  represents a compressed version of  $\mathbf{X}$ . In our case, the peaks have large values compared to other regions of spectra and the class-specific peaks take place in each spectrum of a class. For these reasons, it is natural to expect the class-specific peaks to be presented in  $\{B_j\}_{j \in \mathcal{I}}$  and this is what we observe in our applications (Section 3). Moreover, the provided representation provides a sort of denoising, since the noise contribution is spectrum-specific and is left out for the sake of the dataset compression.

Taking this into account, we propose to do peak detection considering not raw spectra, as usual, but the prototype spectra (the basis vectors  $\{B_j\}_{j \in \mathcal{I}}$ ).

To show the potential of this approach, we applied to prototype spectra a simple peak detection method (the Matlab R2008a function *findpeaks* from the SP toolbox). Given a prototype spectrum, this algorithm looks for local maxima, whose height is two times larger than the mean value of the spectrum.

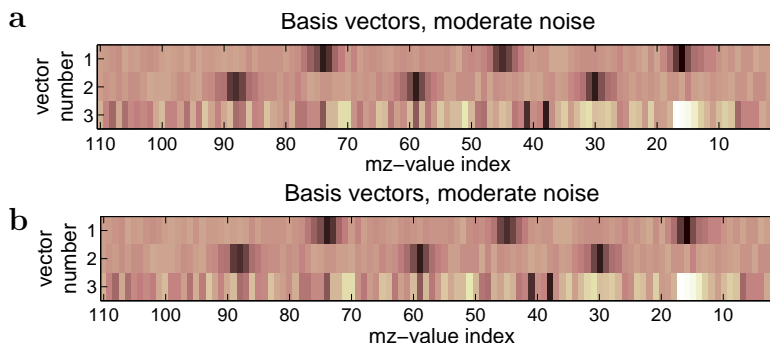
One could expect, that the sparse coding algorithm provides only one basis vector  $B_j$  for each class. However, for the simulated data (Section 3) we found out that in many cases more vectors  $B_j$  are found. Sometimes the redundant vectors mostly contain noise. This can be probably because of insufficient  $l_1$ -penalization or, although optimization of  $\mathbf{S}$  or  $\mathbf{B}$  is a convex problem with an unique solution, the problem (2) has no globally unique solution in general. To reduce the number of false positives, we check for each peak, whether the area under it is large enough. For this we specify the minimal possible peak width, that can be done for a real-life dataset examining selected spectra.

The scheme of our procedure of the peak detection is presented in Fig. 1.

## 3 Evaluation

### 3.1 Simulated data

We evaluated the proposed peak detection procedure using simulated data, where a spectrum was simulated as follows. Several class-specific Gaussian peaks were generated on pre-specified positions with slightly variable heights. Then peaks of smaller heights were added at random positions. Finally, a white Gaussian noise was added throughout a spectrum. We considered a high and a moderate noise level. The length of spectra is  $L = 110$ , that represents only a part of a real-life spectrum; the number of spectra is  $R = 50$ ; the number of classes is  $D = 2$ ; three class-specific peaks are given for each



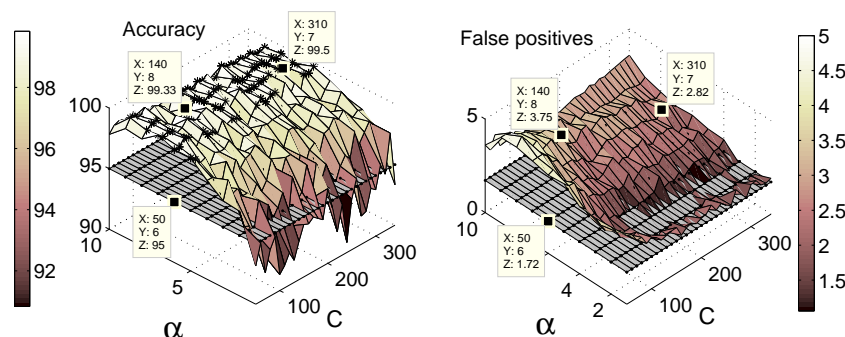
**Fig. 2.** (a) An example of generated spectra dataset (two classes of spectra each with three peaks), with moderate noise as well as (b) the learned prototype spectra (basis vectors plotted along  $mz$ -value index).

class. An example of the simulated spectra set (moderate noise) is depicted in Fig. 2. The proposed procedure has three parameters besides the minimal possible peak width, namely, the regularization parameters  $\alpha$  and  $\beta$ , and the basis vector norm constraint  $C$ . In this paper we study the parameters  $\alpha$  and  $C$ . The parameter  $\beta$  has been investigated in Alexandrov et al. (2009). Note that even without a strategy for their choice one can apply the procedure, for example, when the peak detection is combined with classification of spectra. Then the parameters can be optimized using e.g. cross-validation minimizing the classification accuracy.

The investigation of the parameters has been done using a grid search through ten values of  $\alpha$  (1:1:10) (this denotes values from one to ten taken with a step one) and 26 values of  $C$  (50:10:300) for the following values of  $\beta$  ( $1, 10^{-1}, 10^{-5}, 10^{-10}$ ). For the grid search we simulated 100 replicates of spectra sets both with the high and moderate noise level. Then we calculated the mean values of (i) the accuracy of detection of class-specific peaks (the ratio of between the number of correctly detected peaks and six, which is the number of all ground-truth peaks in a dataset) and (ii) the number of false positives (the peaks which are detected by the procedure but are not the ground-truth peaks) for one spectrum. Fig. 3 shows these measures for  $\beta = 10^{-10}$  for the moderate noise level.

One can see that, as expected, the accuracy mostly depends not on  $C$  but on the regularization parameter  $\alpha$ . The same, but to a lesser extent, is true for the number of false positives. Moreover, the procedure provides almost perfect accuracy (more than 99%) in a large region of parameters. However, in the same region the number of false positives increases.

In order to obtain reference values, we considered the mean spectrum-based peak detection procedure, when the simple peak detection method from Section 2.2 is applied to the mean spectrum of the dataset  $\mathbf{X}$ . This allows us to establish whether the prototype spectra are more suitable for the

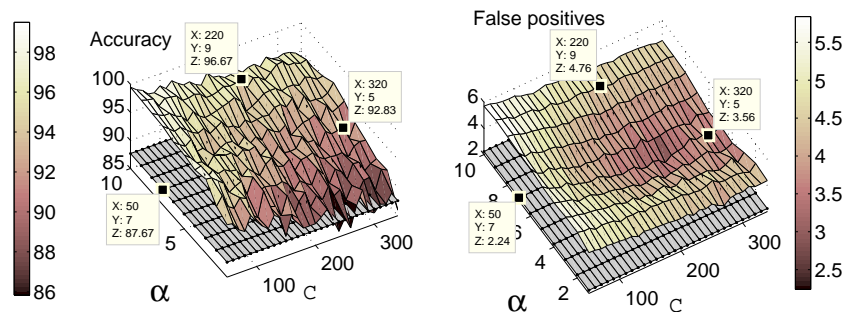


**Fig. 3.** The mean values of the accuracy of determining the peak positions (left) and the number of false positives (right), moderate noise,  $\beta = 10^{-10}$ . For better illustration the values at two arbitrary points are given. The grey horizontal planes correspond to the reference results of the mean spectrum-based procedure. The accuracy values higher than 99% are marked with stars.

peak detection than with the standardly used mean spectrum. For the same replicates with a moderate noise level, the mean spectrum-based procedure provides the mean accuracy of 95% and the number of false positives per spectrum of 1.72. These results are plotted in Fig. 3 as grey horizontal planes. Comparing the results of our procedure with the reference values, one can see that the sparse coding-based procedure provides better accuracy in a large region of parameters (98%–99% vs. 95%) but results in a larger number of false positives (around 3 vs. 1.72). Thus, we conclude that our procedure leads to more accurate peak detection than the mean spectrum-based procedure, although is less specific. This could be because we have two classes and a class-specific peak in the mean spectrum has only a half of its real intensity. This guess was confirmed by experiments with one-class dataset, where the difference between our procedure and the reference values was less than in the two-classes dataset (results are not shown).

Let us consider the results for the high noise level, see Fig. 4. Again, as for the moderate noise level, our procedure is more accurate than the reference procedure, although is less specific. Note that the difference between the reference procedure and our procedure is increased. The accuracy of the mean spectrum-based procedure is equal to 88% (vs. 95% for a moderate noise), whereas is 93%–98% (vs. 98%–99% for a moderate noise) for our procedure in the region of high values. We conclude that our procedure is not only more accurate than the mean spectrum-based procedure but is also more robust to a high noise.

Based on the simulation results, we suggest the following strategy for the choice of parameters. The parameter  $\alpha$  should be selected as large as possible (which leads to a high accuracy), while the procedure returns reasonable amount of basis vectors. For instance, under assumption of two classes, one



**Fig. 4.** The mean values of the accuracy of determining the peak positions (left) and the number of false positives (right), high noise,  $\beta = 10^{-10}$ . For better illustration the values at two arbitrary points are given. The grey horizontal planes correspond to the reference results of the mean spectrum-based procedure.

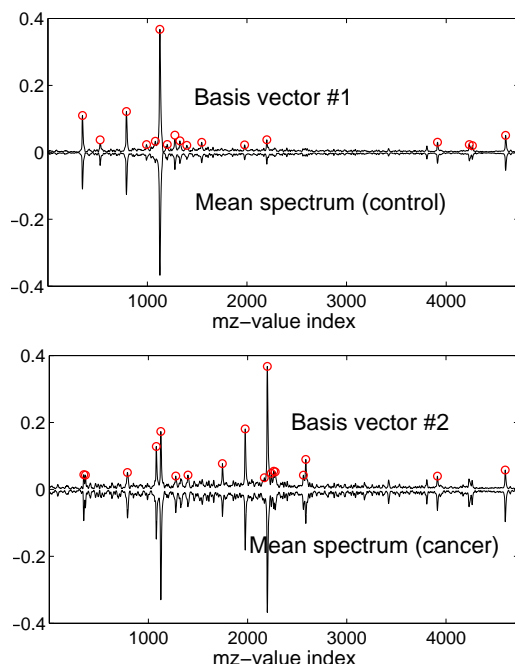
should increase  $\alpha$  as long as not less than two vectors are returned. The choice of  $C$  is not crucial. The parameter  $\beta$  should be as small as possible as proposed in Alexandrov et al. (2009).

### 3.2 Real-life dataset: colorectal cancer

Finally, we considered a real-life dataset de Noo et al. (2006) of 64 colorectal cancer and 48 control mass spectra in the interval  $[1100, 3000]$  Da containing 4731 bins ( $L = 4731$ ). The parameter  $\alpha = 120$  was selected as proposed above (under assumption of two classes), five different values of  $C$  (500:500:2500) have been tested producing similar results,  $\beta$  was selected equal to  $10^{-10}$ . The prototype spectra for  $C = 500$  are shown in Fig. 5. Note their similarity to the per-class mean spectra.

### 3.3 Real-life dataset: liver diseases

In order to illustrate application of our procedure to a set of spectra of more than two classes, we examined a liver diseases dataset of Resson et al. (2007) with mass spectra corresponding to 78 hepatocellular carcinoma (HCC), 51 cirrhosis, and 72 control serum samples. For our purpose we considered only the region  $[1500, 2500]$  Da with 5108  $mz$ -values (spectrum length  $L = 5108$ ). This region was selected since the mean spectra of the three classes in this region are quite different (see Fig. 6). In Resson et al. (2007) significant peaks were found in the region  $[7500, 8500]$  Da. However, in Resson et al. (2007) another two-classes classification problem of comparison disease spectra (HCC and cirrhosis jointly) with control spectra was considered. The parameter  $\alpha = 94.5$  was selected as proposed above (under assumption of three classes), five different values of  $C$  (500:500:2500) have been tested producing similar results,  $\beta = 10^{-10}$ . Again, as for two-classes real-life dataset, the basis vectors are very similar to the per-class mean spectra.



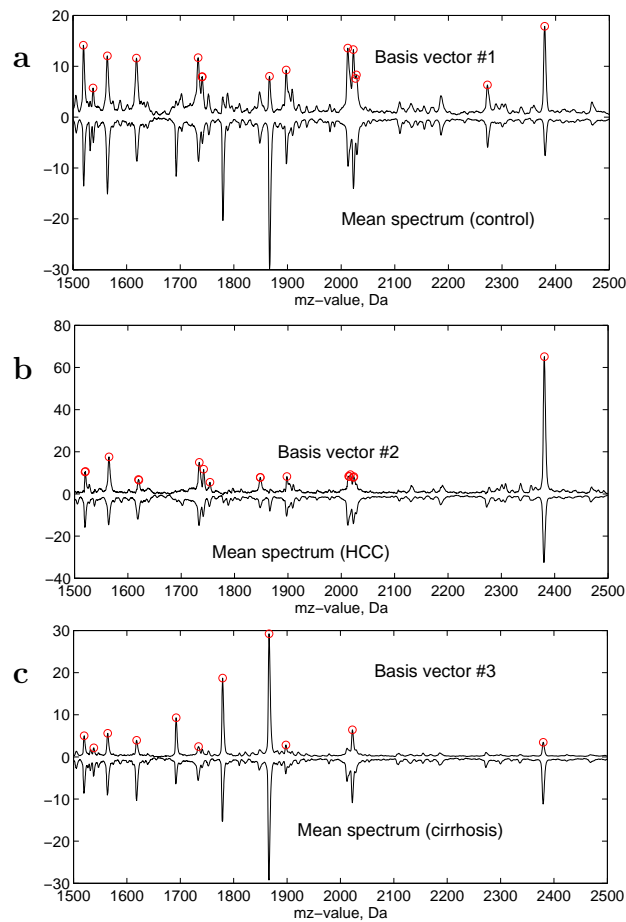
**Fig. 5.** Colorectal cancer results. First (top) and the second (bottom) prototype spectra plotted with the mean control and cancer spectra. Found peaks are marked with circles. Mean spectra are attributed to corresponding basis vectors manually, then scaled and plotted with negative sign.

## 4 Conclusions

The main contribution of this paper is a way of sparse representation of mass spectra, which we propose to use for the peak detection. Interestingly, the prototype spectra are similar to per-class mean spectra, although are obtained in an unsupervised manner. Hence, our peak detection procedure can be applied when assignments of spectra to classes are unknown. We have detected peaks in prototype spectra with a simple method to demonstrate the potential of our approach and expect that application of a more advanced peak detection method can improve the results.

## References

- ALEXANDROV, T., KESZÖCZE, O., LORENZ, D. A., SCHIFFLER, S. and STEINHORST, K. (2009): An active set approach to the elastic-net and its applications in mass spectrometry. In *Proc. Signal Processing with Adaptive Sparse Structured Representations (SPARS'09)*. Available at <http://hal.inria.fr/inria-00369397>.



**Fig. 6.** Liver diseases results. First (a), second (b) and third (c) basis vectors plotted with mean spectra of control, HCC and cirrhosis classes. Found peaks are marked with circles. Mean spectra are attributed to the basis vectors manually, then scaled and plotted with negative sign.

- DE NOO, M., MERTENS, B., OZALP, A., BLADERGROEN, M., VAN DER WERFF, M., VAN DE VELDE, C., DEELDER, A. and TOLLENAAR, R. (2006): Detection of colorectal cancer using MALDI-TOF serum protein profiling. *Eur. J. Cancer* 42(8):1068-76.
- LEE, H., BATTLE, A., RAINA, R. and NG A. Y.. (2006): Efficient sparse coding algorithms. In *Proc. Neural Information Processing Systems (NIPS'06)*, 801-8.
- RESSOM, H. W., VARGHESE, R. S., DRAKE, S., HORTIN, G. L., ABDELHAMID, M., LOFFREDO, C. A. and GOLDMAN, R. (2007): Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics* 23(5):619-26.



# A Comparison between Beale Test and Some Heuristic Criteria to Establish Clusters Number

Angela Alibrandi<sup>1</sup> and Massimiliano Giacalone<sup>2</sup>

<sup>1</sup> Department of Economical, Financial, Social, Environmental, Statistical and Territorial Sciences (S.E.F.I.S.A.S.T.), University of Messina, Via dei Verdi 75, 98122 Messina, *aalibrandi@unime.it*

<sup>2</sup> Department of Public Organization Law, Economy and Society (D.O.P.E.S), University of Catanzaro "Magna Graecia", Campus of Germaneto, 88100 Catanzaro, *maxgiacit@yahoo.it*

**Abstract.** In cluster analysis the individualization of the adequate clusters number represents a fundamental decision to be taken into correctly assigning the units to not-previously defined groups of observations. Many criteria have been proposed in literature in order to establish the best clusters number. Purpose of this paper is to examine the theoretical bases of some most common criteria: Beale test, based on the significance logic and two heuristic methods as Pseudo  $T^2$  Hotelling and Cubic Clustering Criterion. Moreover, we want to compare them in terms of flexibility and applicability, taking in account the assumptions on which they are based; finally we apply all these criteria on real data and we compare the obtained results.

**Keywords:** clusters number, grouping structure, Beale test,  $T^2$ Hotelling, Cubic Clustering Criterion

## 1 Introduction

As it is known in literature, cluster analysis (Johnson, 1998) is a multivariate technique that aims to assign statistical units in not-previously defined categories, creating groups of observations in order to be homogeneous within them and heterogeneous among them. So, purpose of the clustering is to synthesize statistical units in an inferior number of clusters, maximizing the infra-groups distance and minimizing, in the same time, the intra-groups variability. Cluster analysis represents an ideal data-mining tool because the classes or groups that the data form are unknown, especially as the state definition is expanded to include an increasing number of variables. Cluster analysis uncovers these underlying patterns in the data and assigns each case to a cluster. Unlike the discriminant analysis, in the cluster analysis there isn't information about the number of cluster and the characteristics of the groups in the population. The individualization of the grouping structure constitutes, therefore, a fundamental decision to be taken. In literature various criteria (Milligan and Cooper, 1985) have been proposed to individualize

the best structure and to assess the cluster validity ((Xu Rui et al., 2008; Halkidi, 2002). In this context our paper aims to examine three of the most utilized criteria: Beale test (Gordon, 1999), based on the significance logic and two heuristic methods as Pseudo  $T^2$  Hotelling (Halkidi, 2002) and Cubic Clustering Criterion (Sarle, 1983). Moreover, we want to compare them in terms of flexibility and applicability, taking in account the assumptions on which they are based; finally we apply all these criteria on real data and we compare the obtained results. In particular, the paper is so structured:

- in paragraph 2 the theoretical bases of the three considered criteria are exposed;
- in paragraph 3 the application of the examined criteria is shown and a comparison between the obtained results is performed;
- in paragraph 4 some final remarks and a discussion conclude the paper.

## 2 Theoretical bases of Beale test, CCC and PST2

### 2.1 Beale's probabilistic algorithm

Beale's probabilistic algorithm (Beale, 1969) replies to the exigency of choosing the suitable number of clusters, allowing to verify the significance of grouping. Such as it is reported in Gordon (1999), Beale test is based on a F-type statistic and allows to compare goodness of clustering with  $r$  clusters compared to  $r - 1$  clusters, capturing the tightness of clusters. The criterion refers to a matrix of Euclidean distances. Let's suppose to have  $k$  quantitative modalities on each of  $n$  statistic units of a population and we need to individualize a grouping of  $n$  units in  $r$  groups, with  $r < n$ . Let's indicate by  $W(r)$  the residual sum of squares (within group), relative to a partition in  $r$  clusters. Beale test allows to verify the hypothesis according to which, proceeding from  $r - 1$  to  $r$  clusters, there is a significant reduction of within-groups deviance. In order to decide if a partition with  $r$  clusters has to be preferred to another with  $r - 1$ , we can employ:

$$F = \frac{W_{r-1} - W_r}{W_r} \quad (1)$$

whose asymptotic critical region is the right tail of F distribution, with  $\nu_{num} = k$  and  $\nu_{den} = k(n - r)$  degrees of freedom.

Beale proposed a correction factor, indicated with  $C$  that, for large samples, is function of the attended diminution of  $W_r$  to the increasing of  $r$

$$C = \frac{n - (r - 1)}{n - r} \left( \frac{r}{r - 1} \right)^{\frac{2}{k}} - 1 \quad (2)$$

The adjusted statistic test is given by the ratio between  $F$  and the correction factor  $C$ :

$$F' = \frac{F}{C} \quad (3)$$

$F'$  keeps the same degrees of freedom [ $\nu_{num} = k$  and  $\nu_{den} = k(n - r)$ ]. This test has to be calculated for every couple  $(r - 1)$  and  $r$ , until we reach the significance of the test. In order to verify the above-mentioned significance, the test has to be compared with the critical value of the Snedecor -Fisher  $F$  test with  $\nu_{num} = k$  and  $\nu_{den} = k(n - r)$  degrees of freedom, at a fixed  $\alpha$  level of significance.

Large values of the test indicate a better clustering solution. So, if the empirical  $F$  is greater than the critical  $F$  (i.e. the associated p-value is less than  $\alpha = 0.05$ ), we can say that the change from  $r - 1$  to  $r$  clusters yields the reduction of a significant quantity of within-groups deviation and so  $r$  can be considered the optimal number of groups; this result indicates a stopping point. Otherwise, the solution with the smaller number of clusters has to be preferred.

## 2.2 The Pseudo $T^2$ statistics

Another method of judging the number of clusters is the *Pseudo- $T^2$*  statistic ( $PST^2$ ), that is a variant of Hotelling's  $T^2$  (Halkidi, 2002), based on the assumption that two clusters are drawn from two independent multivariate normal distributions with the same mean and covariance.  $PST^2$  computed to compare the means of two aggregated clusters in hierarchical models and is expressed as follows:

$$PST^2 = \frac{n_1 - n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^T W^{-1} (\bar{x}_1 - \bar{x}_2) \quad (4)$$

where  $n_1$  and  $n_2$  indicate the number of observations into the two clusters,  $\bar{x}_1$  and  $\bar{x}_2$  are the mean values, respectively, and  $W$  is the unbiased pooled covariance matrix estimate. In particular,  $PST^2$  measures the degree of separation between the two last aggregated clusters; it can't be considered as significance test because it isn't distributed exactly as  $t$  random variable. So, for its interpretation we have to examine the values that it assumes: elevated values suggest to arrest the clustering to the previous level. If the *Pseudo- $T^2$*  statistic value is large, the means are significantly different and so the considered clusters should not be combined; if the value is small, instead, the clusters can safely be combined. A general rule for interpreting the  $PST^2$  is to observe all values of statistics, calculated for each number of clusters,

until a value results markedly larger than the previous one, establishing the acceptability of the partition. The choice of groups is based on the analysis of the peaks achieved by that index, typically we have to prefer a group with  $k + 1$  classes if at the  $k$  classes, the index assumes high values, followed by a sharp fall.

### 2.3 The Cubic Clustering Criterion

Another criterion proposed in literature in the choice of the optimal number of cluster is the Cubic Clustering Criterion or *CCC* (Sarle, 1983). It's based on the assumption that an uniform distribution on a hyperrectangle will be divided into clusters shaped roughly like hypercubes. The *CCC* aims to verify the null hypothesis according to which the clusters are hypercubes (obtained from a uniform distribution on a hyperbox) against the alternative one for which the data have been sampled from a mixture of spherical multivariate normal distributions, with equal variances and sampling probabilities. In other words, *CCC* is a comparative measure of the deviation of the clusters from the expected distribution, if data points were obtained from an uniform distribution. The criterion is calculated as:

$$CCC = \ln \left( \frac{1 - E(R^2)}{E(R^2)} \right) k \quad (5)$$

where  $E(R^2)$  is the expected  $R^2$  value,  $R^2$  is the observed  $R^2$  (ratio of "between group variance" on "within group variance", that furnishes a measure of the clustering quality) and  $k$  is the variance-stabilizing transformation.

For its interpretation, we have to consider that:

- large positive values of *CCC* ( $>3$ ) indicate good clustering, showing a larger difference from an uniform (no clusters) distribution;
- if *CCC* continues to increase with the number of clusters, it may be an indication of formation of pockets of sub-clusters in more clusters;
- values between 0 and 2 indicate potential clusters; negative values indicate that grouping structure isn't appropriate;
- large negative values can indicate outliers.

This criterion seems to give good results especially on samples of high abundance, while its strength could be lower if the number of observations in each group is low. In any case, it may not be used in probabilistic terms, choosing the number of groups in correspondence to an absolute maximum or relative maximum possible. For these reasons, it's helpful to plot the *CCC* values calculated for each number of clusters and to look for the peaks where *CCC*  $>3$ . However, the *CCC* may be incorrect if variables are highly correlated.

For the last two criteria, which are heuristic, the information deducible from the construction of these indicators should not be interpreted in probabilistic terms, they may be used, noting the trend, so as to identify potential "natural" clusters of the considered units.

## 2.4 Main differences among the three examined criteria

From a methodological point of view, we can compare the examined criteria. Beale test is based on the significance logic and it is characterized by large flexibility and applicability because it's released by restrictions on assumption about the distribution of the studied variables. PST<sup>2</sup> statistic is based on the assumption of normality and CCC index is based on the assumption of uniformity. For both criteria these assumptions are often hardly realizable.

Moreover, either PST<sup>2</sup> or CCC must be calculated to every hierarchical level, whilst Beale test has to be carried until the reaching of the significance.

Finally, CCC and PST<sup>2</sup> are heuristic methods and they can't be analytically considered because the sampling distributions for these indexes are unknown; on the contrary, Beale test is based on F test and follows the same distribution.

## 3 An application to real data

In order to illustrate the utility of the above - mentioned criteria we have applied them on a real dataset. We have examined monthly data, referred to building abusiveness phenomenon in Messina, noticed by the department of Environmental Police of Messina in the year 2007, for each of fourteen districts in which the city was divided.

Our variables are represented by the count of violations to some articles of the Regional Law 10 August 1985, n. 37 "New norms in subject of urbanistic control of the activity - house building, rearranges urbanistic and confirmation of the unauthorized works":

- the first concerned the articles 5 and 9 (Administrative Sanctions);
- the second referred to article 20 (Urbanistic Law);
- the third related to sequestration (Building Sequestrations).

Before data analysis (carried on using the centroid-method, the Euclidean metric and the hierarchical classification) each variable has been standardized.

Tables 1 and 2 report the results of Beale test and the application of the two heuristic methods PST<sup>2</sup> and CCC, respectively.

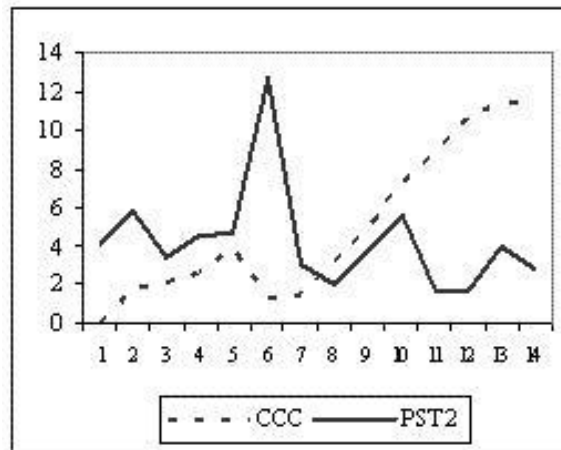
Comparison	F	$\nu_{num}$	$\nu_{den}$	p-value
Cluster 3 vs 2	F=0.143	36	396	0.984
Cluster 4 vs 3	F=0.647	36	360	0.973
Cluster 5 vs 4	F=1.384	36	324	0.022

**Table 1.** Results of Application of Beale test

Clusters	$PST^2$	$CCC$
1	4.10	0
2	5.8	1.86
3	3.5	2.16
4	4.5	2.66
5	4.7	3.9
6	12.7	1.36
7	3.1	1.52
8	2.0	3.06
9	3.7	4.74
10	5.6	7.32
11	1.7	8.62
12	1.7	10.72
13	4.0	11.42
14	2.8	11.52

**Table 2.** Results of Application of  $CCC$  and  $PST^2$  statistics

It's evident that the suitable number of clusters in this framework can be determined by the comparison between five and four clusters, as we can see in the last column of Table 1. Finally, in the last row of last column in Table 1 we note a significant p-value ( $\alpha = 0.05$ ), that suggests us to choose five clusters as optimal partition.



**Fig. 1.** Results of  $CCC$  and  $PST^2$  statistics application

Examining Table 2 and Figure 1 we can note that the largest value of  $PST^2$  is reached for six clusters and this criterion suggests that the optimal choice is the previous level of clustering. Also  $CCC$  indicate five cluster as

best partition, because of the presence of a peach. The fourteen districts are grouped in five clusters as it is illustrated in Table 3.

Clusters	Districts
1	III - XIV - II - IV - XIII
2	V - X - VII
3	I - VIII - XII
4	VI - XI
5	IX

**Table 3.** Allocation of the fourteen Districts in the five Clusters

We are aware that the potential corrected solutions could be more than one. In our application, on the bases of the obtained results, we considered appropriate to use the partition into five groups, because all the applied criteria lead to the choice of this optimal clusters number.

#### 4 Final Remarks and discussion

The ability to identify the appropriate number of clusters for a given set of data is one of the most fundamental shortcomings of non-hierarchical techniques. While local knowledge and experience can play a role in data analysis, user defined parameters such as  $k$  groups builds significant subjectivity into analysis. Furthermore, implicit to most discussions of the k-means approach there are no established methods for determining the optimal number of clusters (Levine, 1999). In fact, there are many and many methods outlined in the statistics literature detailing potential methods for detecting the appropriate number of clusters (Everitt 1979; Gordon 1998; Gubresic, 2006; Lozano J.A. et al (1986), Milligan and Cooper, 1985). Three of the more effective procedures for determining the number of clusters are examined in our data set: the CCC (test statistic provided by the SAS package), the  $PST^2$  statistic and the Beale test. The CCC column of SAS output has been analyzed from  $n$  groups to 1 group. These inflection points are indicative of appropriate cluster groupings for the data. Moreover, we observed more than a single inflection point. Alternatively, graphic plots of CCC values have been utilized for our analysis. The CCC values were also used in conjunction with pseudo F (PSF) and  $t^2$  statistics in our application. Both measures provide additional information, with large PSF values suggesting a good stopping point. Inflections in the  $t^2$  statistic also suggest possible cluster stops. On the bases of the jointly use of the three examined criteria (CCC,  $PST^2$  statistics, and Beale test), we can note that all of them confirm that the best partition of our observations can be reached by grouping in five clusters.

Comparing the three criteria, we can retain that Beale test, based on the significance logic, represents the most flexible and applicable compared to the

others: in applications the assumptions of normality of  $PST^2$  statistic and uniformity of CCC index are often hardly realizable, so Beale test is more extendible because it's released by restrictions on assumption about the distribution of the studied variables. Both  $PST^2$  and CCC must be calculated to every hierarchical level, while Beale test has to be carried on until the reaching of the significance, that represents the stopping point.

Moreover, CCC and  $PST^2$  are heuristic methods and they are rather lacking because the sampling distribution for these indexes are unknown; on the contrary, Beale test is based on F test, assumes the same hypothesis and follows the same distribution, so it's a significance test that inferentially has to be preferred than the other criteria of choice.

## References

- BEALE E. M. L.(1969): Euclidean Cluster Analysis. Bulletin of International Statistical Institute. In: *Proceedings j of the 37th Session*.
- EVERITT B.S.(1979): Unresolved problems in cluster analysis. *Biometrics*, 35 (1), 169-181.
- GORDON A.D. (1998): *How Many Clusters? An Investigation of Five Procedures for Detecting Nested Cluster Structure*. Data Science, Classification and Related Methods. Springer-Verlag.
- GORDON A.D. (1999): *Classification*. 2nd edition. Chapman and Hall, 60-65.
- GRUBESIC T.H. (2006): On the application of fuzzy clustering for crime hot spot detection. *Journal of Quantitative Criminology Springer Netherlands* 22(1),77-105.
- HALKIDI M., BATISTAKIS Y. and VAZIRGIANNIS M.(2002): On clustering validation techniques. *Journal of Intelligent Information System*, 17, (2), 107-145.
- JOHNSON D.E.(1998): Cluster analysis. *Applied Multivariate Methods for Data Analysis*, Duxbury Press, 319-396.
- LEE K. M., HERRMANA T. J., LINGENFELSERA J. and JACKSON D. S. (2005): Classification and prediction of maize hardness-associated properties using multivariate statistical analyses. *Journal of Cereal Science* (41),85-93.
- LEVINE N. (1999): *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations*. Washington DC: Ned Levine and Associates; National Institute of Justice.
- LOZANO J.A., LARRANAGA P. and GRANA M. (1996): Partitional cluster analysis with genetic algorithms: searching for the number of clusters. In: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. Bock, and Y. Baba: *Data Science, Classification and Related Methods*, Tokyo, Springer-Verlag.
- MILLIGAN G.W. and COOPER M.C. (1985): An examination of procedures for determining the number of clusters in a data set. *Psychometrika* (50), 159-179.
- SARLE W.S. (1983): Cubic Clustering Criterion. *SAS Technical Report*, (1), p.108.
- XU RUI, II DONALD C. WUNSCH (2008): Recent advances in cluster analysis, *International Journal of Intelligent Computing and Cybernetics*, Emerald Group Publishing Limited.



# Estimating Population Proportions in Presence of Missing Data

Encarnaci3n lvarez-Verdejo<sup>1</sup>, Antonio Arcos<sup>1</sup>, Silvia Gonzlez<sup>2</sup>, Juan Francisco Muoz<sup>1</sup> and Maria Rueda<sup>1</sup>

<sup>1</sup> University of Granada, Spain, *encarniav@ugr.es*, *arcos@ugr.es*, *jfmunoz@ugr.es*, *mrueda@ugr.es*

<sup>2</sup> University of Jan, Spain *sgonza@ujaen.es*

**Abstract.** This paper discusses the estimation of a population proportion in the presence of missing data and using auxiliary information at the estimation stage. A general class of estimators, which makes an efficient uses of the available information, is proposed. Some theoretical properties of the proposed estimators are analyzed, and they allow us to find the optimum values in each proposed class. Optimum estimators are thus more efficient than the customary estimator. In particular, the estimator based on the difference method is an optimal estimator in the sense it has minimal variance into the class. Results derived from a simulation study show that the proposed optimum estimators give desirable results in comparison to alternative estimators.

**Keywords:** auxiliary information, ratio and difference estimators

## 1 Introduction

The problem of missing data is a common aspect in many surveys and other practical situations. The treatment of missing data in survey research is not a simple issue. A variety of methods have been developed to attempt to compensate for missing data in a general purpose way that enables the survey's data file to be analyzed without regard for the missing data.

There are many options for dealing with missing data. The simplest solution is to do nothing, that is, missing values are flagged on the output data file, leaving it up to the data user or analyst to deal with them. This solution to the problem of missing data is usually adopted when it is too difficult to impute values with sufficient accuracy. An obvious consequence is that the actual sample size is less than the planned one, which can produce important biases in the point estimation and larger sampling variances.

The preferred option for survey users is to impute missing data within individual records (see, for example, Little and Rubin (1987)), but treating these imputed values as true observations, one may apply without any missing observation. Such a practice may tend to invalidate the inferences and may often have serious consequences. Imputation procedures should be based on the Fellegi-Holt principles.

Contending that the deleted observations may contain valuable information, a third option is to try to improve the precision of the estimators by including the available information at the estimation stage. Some authors have defined indirect estimators for the population mean when some observations are missing, see, e.g., Tracy and Osahan (1994), Toutenburg and Srivastava (1998), Rueda and Gonzalez (2004), Rueda et al. (2007). However, the problem of the estimation of a population proportion in presence of missing data is a problem which has not been discussed. This is the main concern in this paper. We thus define a general class of estimators of a population proportion on the basis of a random sample drawn according to any sampling design.

## 2 Indirect estimators of proportion with missing values

Let  $U = \{1, 2, \dots, N\}$  be a population of  $N$  identifiable elements. In this paper, we consider the problem of estimating the population proportion  $P_A = N^{-1} \sum_{i \in U} A_i$ , where  $A_i$  is an attribute indicator for unit  $i$ , i.e.,  $A_i = 1$  if unit  $i$  has the attribute of interest  $A$ , and  $A_i = 0$  otherwise.  $P_A$  is the parameter of interest and needs to be estimated. For this purpose, a random sample  $s$ , of size  $n$ , is selected from  $U$  according to a given sampling design. The first- and the second-order inclusion probabilities associated to the sampling design are denoted, respectively, as  $\pi_i$  and  $\pi_{ij}$ , and which are assumed to be strictly positive. The design weight associated to the unit  $i$  is given by  $d_i = \pi_i^{-1}$ .

$P_A$  can be estimated by using the well-known Hájek type expansion estimator, which is given by

$$\hat{p}_A = \frac{1}{\hat{N}} \sum_{i \in s} d_i A_i, \quad (1)$$

where  $\hat{N} = \sum_{i \in s} d_i$ . Estimator (1) makes no use of auxiliary information. However, it is common to assume that there exists an auxiliary variable which can be used at the estimation stage to improve the estimation of the parameter of interest. For this reason, we assume an auxiliary attribute  $B$ , whose population proportion,  $P_B$  is known from a census or estimated without sampling errors. This assumption is commonly used in the survey sampling context. On the other hand, a problem of missing data can occur in the sample  $s$  and the estimator (1) can not be calculated in this situation.

Throughout this paper, we assume missing data on the sample  $s$ , which can be divided into the disjoint sets

$$\begin{aligned} s_1 &= \{i \in s / A_i, B_i \text{ are non-missing}\} \\ s_2 &= \{i \in s / A_i \text{ are missing, and } B_i \text{ are non-missing}\} \\ s_3 &= \{i \in s / B_i \text{ are missing, and } A_i \text{ are non-missing}\}, \end{aligned}$$

with  $s_1$  of size  $n - p - q$ ,  $s_2$  of size  $p$  and  $s_3$  of size  $q$ . We assume that  $p$  and  $q$  are integer numbers verifying  $0 < p, q < n/2$ .

Hájek type estimators can be computed from the samples  $s_1$ ,  $s_2$  and  $s_3$  as follows

$$\begin{aligned}\hat{p}_{AH}^{(1)} &= \frac{1}{\hat{N}_1} \sum_{i \in s_1} d_i A_i, & \hat{p}_{AH}^{(3)} &= \frac{1}{\hat{N}_3} \sum_{i \in s_3} d_i A_i \\ \hat{p}_{BH}^{(1)} &= \frac{1}{\hat{N}_1} \sum_{i \in s_1} d_i B_i, & \hat{p}_{BH}^{(2)} &= \frac{1}{\hat{N}_2} \sum_{i \in s_2} d_i B_i,\end{aligned}$$

where  $\hat{N}_j = \sum_{i \in s_j} d_i$ ,  $j = 1, 2, 3$ .

For the case of complete observations (data on sample  $s_1$ ), Hájek type estimators can be used to derive ratio, difference and regression type estimators as follows:

$$\hat{p}_r^{(1)} = \frac{\hat{p}_{AH}^{(1)}}{\hat{p}_{BH}^{(1)}} P_B \quad (2)$$

$$\hat{p}_d^{(1)} = \hat{p}_{AH}^{(1)} + (P_B - \hat{p}_{BH}^{(1)}) \quad (3)$$

$$\hat{p}_{reg}^{(1)} = \hat{p}_{AH}^{(1)} + b(P_B - \hat{p}_{BH}^{(1)}) \quad (4)$$

Motivated by Srivastava and Jhajj (1981), we suggest a class of estimators of the population proportion  $P_A$  as:

$$g_A = G(\hat{p}_{AH}^{(1,3)}, u_1, u_2) \quad (5)$$

where

$$\hat{p}_{AH}^{(1,3)} = \frac{1}{\hat{N}_{13}} \sum_{i \in s_1 \cup s_3} d_i A_i, \quad u_1 = \frac{\hat{p}_{BH}^{(1)}}{P_B}, \quad u_2 = \frac{\hat{p}_{BH}^{(2)}}{P_B},$$

$\hat{N}_{13} = \sum_{i \in s_1 \cup s_3} d_i$  and  $G(\cdot)$  is continuous in a closed convex sub-space,  $L$ , containing the point  $(P_A, 1, 1)$ , and such that:

- $G(P_A, 1, 1) = P_A$ ,
- $G_1(P_A, 1, 1) = 1$ , where  $G_1(\cdot)$  is the first partial derivative of  $G(\cdot)$  with respect to  $\hat{p}_{AH}^{(1,3)}$ , and
- the first and second order partial derivatives of  $G$  exist and are also continuous in  $L$ .

Any parametric function  $G$  satisfying these conditions can generate an asymptotically acceptable estimator. Note that the ratio, difference and regression estimators defined by (2), (3) and (4) are included in (5).

*Theorem 1*

*Up to terms of order  $n^{-1}$  for all estimators in class (5)*

$$V(g_A) \geq V(\hat{p}_{AH}^{(1,3)}) - \sigma' \Sigma^{-1} \sigma$$

where

$$\Sigma = \begin{pmatrix} V(\hat{p}_{BH}^{(1)}) & cov(\hat{p}_{BH}^{(1)}, \hat{p}_{BH}^{(2)}) \\ cov(\hat{p}_{BH}^{(1)}, \hat{p}_{BH}^{(2)}) & V(\hat{p}_{BH}^{(2)}) \end{pmatrix}$$

and

$$\sigma = \left( cov(\hat{p}_{AH}^{(1,3)}, \hat{p}_{BH}^{(1)}), cov(\hat{p}_{AH}^{(1,3)}, \hat{p}_{BH}^{(2)}) \right)'.$$

Proof of Theorem 1.

By expanding  $G$  about the point  $(P_A, 1, 1)$  in a second order Taylor series, it is found that:

$$\begin{aligned} g_A = G(P_A, 1, 1) + (\hat{p}_{AH}^{(1,3)} - P_A)G_1(P_A, 1, 1) + \frac{\partial G}{\partial u_1}|_{(P_A, 1, 1)}(u_1 - 1) + \\ + \frac{\partial G}{\partial u_2}|_{(P_A, 1, 1)}(u_2 - 1) + O(n^{-1}). \end{aligned}$$

It is obvious that the bias of  $g_A$  is of the order  $n^{-1}$ . By squaring both sides in the last expression, taking expectations and neglecting higher order terms we obtain the approximation

$$V(g_A) = V(\hat{p}_{AH}^{(1,3)}) + H'\Sigma H + 2H'\sigma,$$

where

$$H' = \left( \frac{\partial G}{\partial u_1}|_{(P_A, 1, 1)}, \frac{\partial G}{\partial u_2}|_{(P_A, 1, 1)} \right).$$

The optimum value of  $H$  can be obtained by differentiating the above expression and equating to zero, thus obtaining  $H = -\Sigma^{-1}\sigma$ . On substituting the optimum value of  $H$  we find that the minimum value of the variance in the class is  $V_{min} = V(\hat{p}_{AH}^{(1,3)}) - \sigma'\Sigma^{-1}\sigma$ .  $\square$

### 3 The difference estimator

The traditional difference method can be easily applied to the estimation of a population proportion. However, it requires that all units in the samples are known, which is not the scenario in this paper. A solution is to use the difference estimator on the sample  $s_1$ , which contains information on both variables and indirect estimators such as the difference estimator can be applied without any complication. This difference estimator based on complete observations is defined by (3). We observe that (3) makes no use of the information provided by the samples  $s_2$  and  $s_3$ .

We now consider a choice within the class  $G$  of the type

$$\hat{p}_{gd} = \hat{p}_{AH}^{(1,3)} + c(P_B - \hat{p}_{BH}^{(1)}) + d(P_B - \hat{p}_{BH}^{(2)}), \quad (6)$$

which incorporates the information of the samples  $s_2$  and  $s_3$  at the estimation stage.  $c$  and  $d$  are regression coefficient which need to be estimated. The asymptotic unbiasedness of the proposed class of estimators (6) is easily derived from their linear expression, and using that  $\hat{p}_{AH}^{(1,3)}$ ,  $\hat{p}_{BH}^{(1)}$  and  $\hat{p}_{BH}^{(2)}$  are asymptotically unbiased estimators.

It is possible to obtain the values of  $c$  and  $d$  that provide the minimum variance of the estimator (6). It can be seen that  $(c_{opt}, d_{opt})' = \Sigma^{-1}\sigma$  and  $V_{\min}(\hat{p}_{gd}) = V(\hat{p}_{AH}^{(1,3)}) - \sigma'\Sigma^{-1}\sigma$ .

It is interesting to note that the lower bound of the asymptotic variance of  $g_A$  is the variance of the difference estimator  $\hat{p}_{gd}$  with the optimum  $c$  and  $d$  values. Thus, asymptotically,  $\hat{p}_{gd}$  is an optimal estimator in this class, in the sense that it has a lower asymptotic variance and, moreover, it is better than any estimator of the class  $g_A$ .

## 4 Sample-based counterparts

The population-based quantities  $c_{opt}$  and  $d_{opt}$  have sample-based analogues,  $\hat{c}_{opt}$  and  $\hat{d}_{opt}$ , which are now defined.

The optimum values  $c_{opt}$  and  $d_{opt}$  depend on the unknown parameters  $\Sigma$  and  $\sigma$ , and so the optimal difference estimator cannot be used in practice.  $\Sigma$  and  $\sigma$  can be estimated by using estimators of variances and covariances of the Hájek estimators, which can be calculated because we have assumed that the  $\pi_{ij}$ 's are strictly positive  $\forall i, j \in U$ . Now, the sample-based difference estimator is as follows:

$$\hat{p}_{gd}^* = \hat{p}_{AH}^{(1,3)} + \hat{c}(P_B - \hat{p}_{BH}^{(1)}) + \hat{d}(P_B - \hat{p}_{BH}^{(2)}), \quad (7)$$

where  $(\hat{c}, \hat{d})' = \hat{\Sigma}^{-1}\hat{\sigma}$ , and  $\hat{\Sigma}$  and  $\hat{\sigma}$  are the sample-based estimators of  $\Sigma$  and  $\sigma$ .

Assuming simple random sampling without replacement (SRS), the Hájek estimators coincide with the customary sample means, and the optimal difference estimator takes the form

$$\hat{p}_{gdSRS}^* = \hat{p}_A^{(1,3)} + \hat{c}(P_B - \hat{p}_B^{(1)}) + \hat{d}(P_B - \hat{p}_B^{(2)}) \quad (8)$$

where  $\hat{p}_A^{(1,3)}$ ,  $\hat{p}_B^{(1)}$  and  $\hat{p}_B^{(2)}$  are the customary sample proportions based on samples  $s_1 \cup s_3$ ,  $s_1$  and  $s_2$ , and the expressions for the variances and covariances of the estimators can be easily obtained.

Assuming SRS, estimators in the proposed class (6) and the estimator (8) have asymptotically the same distribution. Results derived from Randles (1982) can be applied to show that (8) has asymptotically the same distribution than the optimum estimator

$$\hat{p}_{gdopt} = \hat{p}_A^{(1,3)} + c_{opt}(P_B - \hat{p}_B^{(1)}) + d_{opt}(P_B - \hat{p}_B^{(2)}).$$

## 5 Simulation study

We now compare empirically the performance of the proposed estimator (7) with the existing estimators (1) and (3). For this reason, we carried out a simulation study, which uses as population a data set from the Spanish National Health Survey (NHS) in 2006. The first stage units are the census sections (2236 census sections are selected in the sample), which are grouped into strata in according with the size of the municipality. The second stage units are main family dwellings. Within these units, no sub-sampling is carried out. Within each household, an adult (16 years old or over) is selected in order to complete the questionnaire sets for adults, and if there are children (0 to 15 years old) in the household, one of them is selected in order to complete the children's questionnaire. Our data refers to this children's questionnaire, which are composed by 9063 children from 31300 households.

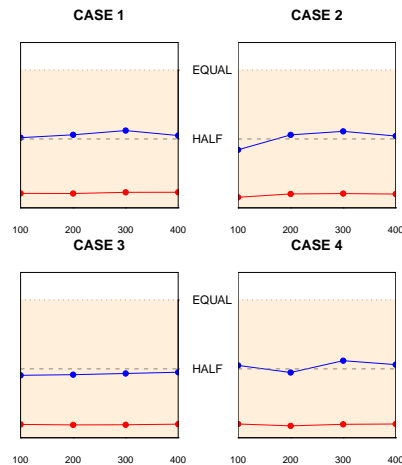
The variable of interest indicates whether or not the child has asthma, whereas the auxiliary variable indicates whether or not the child has taken any treatment for asthma in the last two weeks.

We considered samples draw under simple random sampling with sample sizes  $n = \{100, 200, 300, 400\}$ . Four different cases were considered, which are described in Table 1. Let  $r_2 = p/n$  and  $r_3 = q/n$  denote, respectively, the rates of missing data in samples  $s_2$  and  $s_3$ .

**Table 1.** Definition of the various cases considered in the simulation study.  $r_2$  and  $r_3$  are, respectively, the rates of missing data in samples  $s_2$  and  $s_3$ .

	$r_2\%$	$r_3\%$
CASE 1	5	5
CASE 2	10	5
CASE 3	5	10
CASE 4	10	10

Proposed estimator (7) is analyzed in terms of  $RE_d$  and  $RE_A$ , where  $RE_d = MSE(\hat{p}_{gd}^*)/MSE(\hat{p}_d^{(1)})$  and  $RE_A = MSE(\hat{p}_{gd}^*)/MSE(\hat{p}_A)$  are the relative efficiencies of the proposed estimator  $\hat{p}_{gd}^*$  compared to the estimators (3) and (1), and  $MSE$  denotes the empirical mean square error. Figure 1 shows the values of  $RE_d$  and  $RE_A$ . Note that lines situated on the shaded area indicate that the proposed estimator is more efficient than their competitors. We observe that the proposed estimator is always the most efficient estimator. The gain in efficiency in comparison to the direct estimator  $\hat{p}_A$  is considerable. These results indicate that the proposed estimator  $\hat{p}_{gd}^*$  can provide desirable estimates in the presence of missing data, as it makes a good use of the available information at the estimation stage.



**Fig. 1.** Values of  $RE_d$  (top line) and  $RE_A$  (bottom line). Samples are extracted under SRS with sizes from  $n = 100$  to  $n = 400$ .

## References

- LITTLE, R.J.A. and RUBIN, D.B. (1987): *Statistical analysis with missing data*. John Wiley, New York.
- RANDLES, R.H. (1982): On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics* 10, 462-474.
- RUEDA, M. and GONZÁLEZ, S. (2004): Missing data and auxiliary information in surveys. *Computational Statistics* 19 (4), 555-567.
- RUEDA, M., MUÑOZ, J.F., BERGER, Y.G., ARCOS, A. and MARTÍNEZ, S. (2007): Pseudo empirical likelihood method in the presence of missing data. *Metrika* 65, 349-346.

- SRIVASTAVA, S.K. and JHAJJ, H.S. (1981): A class of estimators of the population mean in surve sampling using auxiliary information. *Biometrika* 68, 341–343.
- TOUTENBURG, H. and SRIVASTAVA, V.K. (1998): Estimation of ratio of population means in survey sampling when some observations are missing. *Metrika* 48, 177–187.
- TRACY, D.S. and OSAHAN, S.S. (1994): Random nonresponse on study variable versus on study as well as auxiliary variables. *Statistica*, 54, 163–168.



# Sub-quadratic Markov Tree Mixture Models for Probability Density Estimation

Sourour Ammar<sup>1</sup>, Philippe Leray<sup>1</sup>, and Louis Wehenkel<sup>2</sup>

<sup>1</sup> Knowledge and Decision Team

Laboratoire d'Informatique de Nantes Atlantique (LINA) UMR 6241

Ecole Polytechnique de l'Université de Nantes, France,

*sourour.ammar@univ-nantes.fr, philippe.leray@univ-nantes.fr*

<sup>2</sup> Department of Electrical Engineering and Computer Science & GIGA-Research,  
University of Liège, Belgium, *L.Wehenkel@ulg.ac.be*

**Abstract.** To explore the “Perturb and Combine” idea for estimating probability densities, we study mixtures of tree structured Markov networks derived by bagging combined with the Chow and Liu maximum weight spanning tree algorithm and we try to accelerate the research procedure by reducing its computation complexity below the quadratic and keeping similar accuracy.

We empirically assess the performances of these heuristics in terms of accuracy and computation complexity, with respect to mixtures of bagged Markov trees, and single Markov tree *CL* built using the Chow and Liu algorithm.

**Keywords:** density estimation, mixture of trees, Perturb and Combine

## 1 Introduction

Learning of graphical probabilistic models essentially aims at discovering a maximal factorization of the joint density of a set of random variables according to a graph structure, based on a random sample of joint observations of these variables. Such a graphical probabilistic model may be used for elucidating the conditional independencies holding in the data-generating distribution, for automatic reasoning under uncertainties, and for Monte-Carlo simulations. Unfortunately, currently available optimization algorithms for graphical model structure learning are either restrictive in the kind of distributions they search for, or of too high computational complexity to be applicable in very high dimensional spaces. Moreover, not much is known about the behavior of these methods in small sample conditions and, as a matter of fact, one may suspect that they will suffer from overfitting when the number of variables is very large and the sample size is comparatively very small.

In the context of supervised learning, a generic framework which has led to many fruitful innovations is called “Perturb and Combine”. Its main idea is to on the one hand perturb in different ways the optimization algorithm used to derive a predictor from a dataset and on the other hand to combine in some

appropriate fashion a set of predictors obtained by multiple iterations of the perturbed algorithm over the dataset. In this framework, ensembles of weakly fitted randomized models have been studied intensively and used successfully during the last two decades. Among the advantages of these methods, let us quote the improved predictive accuracy of their models, and the potentially improved scalability of their learning algorithms (e.g. Geurts et al. (2006)).

In the context of density estimation, bagging and boosting of normal distributions has been proposed by Ridgeway (2002). The link between mixture of models and bayesian modelling framework has been described in Ammar et al. (2009). In Ammar et al. (2008) the Perturb and Combine idea for probability density estimation with probabilistic graphical models was first explored by comparing large ensembles of randomly generated (directed) poly-trees and randomly generated undirected trees. In Ammar et al. (2009) other comparisons were made essentially between ensembles of optimal trees derived from bootstrap copies of the dataset by the Chow and Liu algorithm (Chow and Liu (1968)), which is of quadratic complexity with respect to the number of variables (called bagging of Markov trees), and mixtures of tree structures generated in a totally randomized fashion with linear complexity in the number of variables. This work proved that *Bagged* ensembles of Markov trees significantly outperform totally randomized ensembles of Markov trees, both in terms of accuracy and speed of convergence when the number of mixture components is increased. Thus, in the present paper we focus on our methods with bagging and we study various manners to improve them by forcing the complexity of the optimization level in the *Chow Liu MWST* algorithm to come down below the quadratic and keeping the same accuracy. The main idea of this work is to weaken the Chow Liu algorithm search procedure, which is the more expensive step, by considering only reduced ensemble of mutual information terms chosen at random. We consider two ways to randomize which terms will be considered in the search procedure. The first one consist on choosing totally at random these terms, the second one exploits the result of previous iterations to choose part of considering terms. We assess the accuracy of these two methods empirically on a set of synthetic test problems in comparison to methods described in Ammar et al. (2009).

The rest of this paper is organized as follows. Section 2 recalls briefly the principle of learning random mixtures of models and Section 3 describes the proposed research heuristics. Section 4 collects our simulation results and Section 5 briefly concludes and highlights some directions for further research.

## 2 Randomized Markov tree mixtures

Randomized Markov tree mixtures was studied in Ammar et al. (2009) and Ammar et al. (2008) and applied to be an alternative to classical methods of density estimation in the context of high-dimensional spaces and small datasets.

In the space of Markov tree structures probabilistic inference (Pearl (1986)) and parameter learning are of linear complexity in the number of variables  $n$ . Importantly, Markov tree models may be learned efficiently by the Chow and Liu algorithm which is only quadratic in the number of vertices (variables).

Let  $X = \{X_1, \dots, X_n\}$  be a finite set of discrete random variables, and  $D = (x^1, \dots, x^d)$  be a dataset (sample) of joint observations  $x^i = (x_1^i, \dots, x_n^i)$  independently drawn from some data-generating density  $\mathbb{P}_G(X)$ .

A mixture distribution  $\mathbb{P}_{\hat{T}}(X_1, \dots, X_n)$  over a set  $\hat{T} = \{T_1, \dots, T_m\}$  of  $m$  Markov trees is defined as a convex combination of elementary Markov tree densities, ie.

$$\mathbb{P}_{\hat{T}}(X) = \sum_{i=1}^m \mu_i \mathbb{P}_{T_i}(X), \quad (1)$$

where  $\mu_i \in [0, 1]$  and  $\sum_{i=1}^m \mu_i = 1$ , and where we leave for the sake of simplicity implicit the values of the parameter sets  $\tilde{\theta}_i$  of the individual Markov tree densities.

Our generic procedure for learning a random Markov tree mixture distribution from a dataset  $D$  is described by algorithm 1 (Ammar et al. (2009)). This algorithm returns the  $m$  tree-models, along with their parameters  $\theta_{T_i}$  and the weights of the trees  $\mu_i$ .

**Algorithm 1 (Learning a Markov tree mixture)**

- a. Repeat for  $i = 1, \dots, m$ :
  - (a) Draw random number  $\rho_i$ ,
  - (b)  $T_i = \text{DrawMarkovtree}(D, \rho_i)$ ,
  - (c)  $\tilde{\theta}_{T_i} = \text{LearnPars}(T_i, D, \rho_i)$
- b.  $(\mu)_{i=1}^m = \text{CompWeights}((T_i, \theta_{T_i}, \rho_i)_{i=1}^m, D)$
- c. Return  $(\mu_i, T_i, \tilde{\theta}_{T_i})_{i=1}^m$ .

Some versions of this algorithm procedures used in our experiments are further discussed in Ammar et al. (2009). We concentrate in this work on *DrawMarkovtree* procedure and we describe in the following section new heuristics based on the Perturb and Combine principle in order to reduce the complexity of previous proposed approaches.

### 3 *DrawMarkovtree* procedure sub-quadratic heuristics

We proposed in Ammar et al. (2009) variants for *DrawMarkovtree* procedure. The first one randomly generates unconstrained Markov trees (by sampling from a uniform density over the set of all Markov tree models). The second one builds optimal tree structures by applying the MWST (Maximum Weight Spanning Tree) structure learning algorithm (Chow and Liu (1968)) on a random bootstrap replica of the initial learning set. We demonstrated in this previous work that the consistently best method is the method which uses bagging of tree structures running with a quadratic complexity.

Thus, we propose to study this method and try to accelerate the learning procedure by keeping similar accuracy. This procedure can be decomposed in three steps : the first one consists on computing the mutual information between each pair of variables to fill an  $n \times n$  symmetrical mutual information matrix called  $MI$ , the second one consists on finding the maximum weight spanning tree by applying the MWST algorithm (we use the Kruskal algorithm), and finally the third consists on learning structure parameters.

The first step is quadratic on the number of variables ( $n^2$  terms to compute) while the second step has a complexity of  $E \log(E)$  where  $E$  is the number of considered edges. If the  $MI$  matrix is totally filled ( $E = n^2$  terms), the complexity of the second step is  $2 n^2 \log(n)$ . The third step is linear on the number of variables.

As we combine weak models from random trees to optimal ones learnt by applying a maximum weight spanning tree research with the  $MI$  matrix, in order to obtain a good density estimation, we propose here to apply again the Perturbe and Combine principle by using intermediate models learnt with an incomplete  $MI$  matrix. The number of terms considered  $K$  is a key parameter to estimate the total procedure complexity.

This procedure is then described by the algorithm 2.

**Algorithm 2 (Naive DrawMarkovTree Subquadratic procedure)**

- a.  $MI_i = [ ]_{n \times n}$
- b.  $D_i = GenSamples(D, i)$ ,
- c. Repeat for  $k = 1, \dots, K$ :
  - (a) Draw random pair of number  $(i_1, i_2)$ ,
  - (b)  $MI_i[i_1, i_2] = ComputeMI(X_{i_1}, X_{i_2})$
- d.  $T_i = CompKruskal(MI_i)$ ,
- e. Return  $T_i$ .

$\{(i_1, i_2)\}$  represents a set of pair indices which are generated randomly according to a uniform distribution and will be used to partially fill the  $MI_i$  matrix by *ComputeMI*. *CompKruskal* takes as input the partially filled  $MI_i$  and builds the corresponding maximum weight spanning tree which will be returned by the algorithm.

We propose to consider different values of this parameter and study his impact on the accuracy of the result model. We report in this paper simulations and results for one value of the parameter  $K : n \log(n)$ .

If  $K = n \log(n)$ , then the first step complexity will be  $n \log(n)$ . In the second step, we will consider  $E = n \log(n)$  edges to compute the corresponding maximum weight spanning tree, and then the complexity of this step will be :  $E \log(E) = n \log(n) \log(n \log(n))$ , which is sub-quadratic and very close to the quasi-linear.

An other idea is considered to compute sub-optimal maximum weight spanning tree by *DrawMarkovtree* procedure. This idea consists in taking advantage of the resulting Markov tree built in the previous iteration to

compute the next one. Edges indices of the Markov tree built at iteration  $i$  will be used first to fill the  $MI_{i+1}$  matrix of next iteration  $i + 1$ , then we complete the  $K$  terms indices by generating them at random. This idea can be described by algorithm 3.

**Algorithm 3 (Inertial DrawMarkovTree Subquadratic procedure)**

- a.  $MI_i = [\ ]_{n \times n}$
- b.  $D_i = \text{GenSamples}(D, i)$ ,
- c. Repeat for  $k = 1, \dots, \text{nbEdges}(T_{i-1})$ :
  - (a)  $(i_1, i_2) = \text{GetIndices}(\text{GetEdge}(T_{i-1}, k))$ ,
  - (b)  $MI_i[i_1, i_2] = \text{ComputeMI}(X_{i_1}, X_{i_2})$
- d. Repeat for  $k = 1, \dots, K - \text{nbEdges}(T_{i-1})$ :
  - (a) Draw random pair of number  $(i_1, i_2)$ ,
  - (b)  $MI_i[i_1, i_2] = \text{ComputeMI}(X_{i_1}, X_{i_2})$
- e.  $T_i = \text{CompKruskal}(MI_i)$ ,
- f. Return  $T_i$ .

We consider 2 variants of the *GenSample* function used in step 2. of algorithm 2 and 3. The first one uses the same original dataset in each iteration. The second one generates a bootstrap replica of the initial dataset.

Finally, we consider two variants for the *CompWeights* function proposed in Ammar et al. (2009), namely uniform weighting and Bayesian averaging.

## 4 Empirical simulations

### 4.1 Protocol

In order to evaluate the different heuristics proposed to ameliorate the reserach procedure complexity, we carried out repetitive experiments for different data-generating (or target) densities, by proceeding in the following way.

**Choice of target density** All our experiments were carried out with models for a set of  $n = 1000$  binary random variables. To choose a target density  $\mathbb{P}_G(X)$ , we first decide whether it will factorize according to a general directed acyclic graph structure. Then we use the appropriate random structure and parameter generation algorithm described in Ammar et al. (2008) to draw a structure and their parameters.

**Generation of datasets** For each target density and dataset size, we generated 10 different datasets by sampling values of the random variables using the Monte-Carlo method with the target structure and parameter values. We carried out simulations with dataset sizes of  $N = 1000$  elements. Given the total number of  $2^n$  possible configurations of our  $n$  random variables, we thus look at rather small datasets.

**Learning of mixtures** For a given dataset and for a given variant of the mixture learning algorithm we generate ensemble models of growing sizes, respectively  $m = 1$ ,  $m = 10$ , and then up to  $m = 150$  by increments of 10.

This allows us to appraise the effect of the ensemble size on the quality of the resulting model.

**Accuracy evaluation** The quality of any density inferred from a dataset is evaluated by the approached Kullback-Leibler divergence between this density and the data-generating density  $\mathbb{P}_G(X)$  used to generate the dataset.

**Software implementation** Our various algorithms were implemented in C++ with the Boost library (<http://www.boost.org/>) and APIs provided by the ProBT© platform (<http://bayesian-programming.org>).

Our generic algorithm can be declined by varying the tree generation function (random, algorithms 2 or 3), the learning dataset (initial data  $D$  or bootstrap replica  $B$ ) and the weighting coefficient (uniform or BDeu).

Table 3 summarizes the name of the different variants we will compare in the next section.

Variants name	Tree Generation	Dataset	Coefficients	Complexity
<b>MTU</b>	Random	D	Uniform	Linear
<b>MTBDeu</b>	Random	D	BDeu	Linear
<b>FBU</b>	Algo 2	B	Uniform	Sub-quadratic
<b>FBBDeu</b>	Algo 2	B	BDeu	Sub-quadratic
<b>FDU</b>	Algo 2	D	Uniform	Sub-quadratic
<b>FDBDeu</b>	Algo 2	D	Uniform	Sub-quadratic
<b>FRBU</b>	Algo 3	B	Uniform	Sub-quadratic
<b>FRBBDeu</b>	Algo 3	B	BDeu	Sub-quadratic
<b>FRDU</b>	Algo 3	D	Uniform	Sub-quadratic
<b>FRDBDeu</b>	Algo 3	D	BDeu	Sub-quadratic
<b>CL</b>	MWST	D	Uniform	quadratic

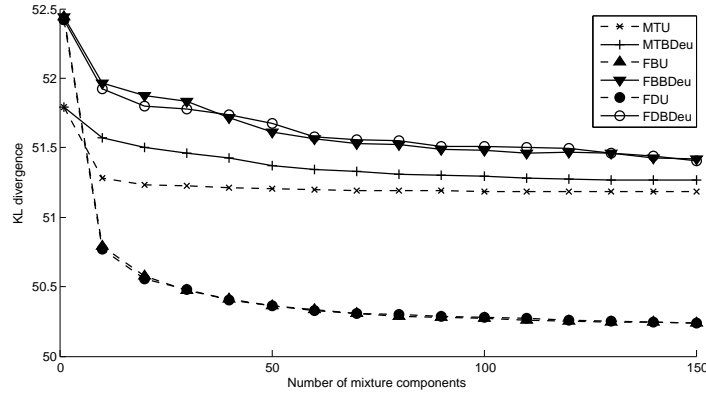
**Table 1.** Algorithms' variants name

## 4.2 Results

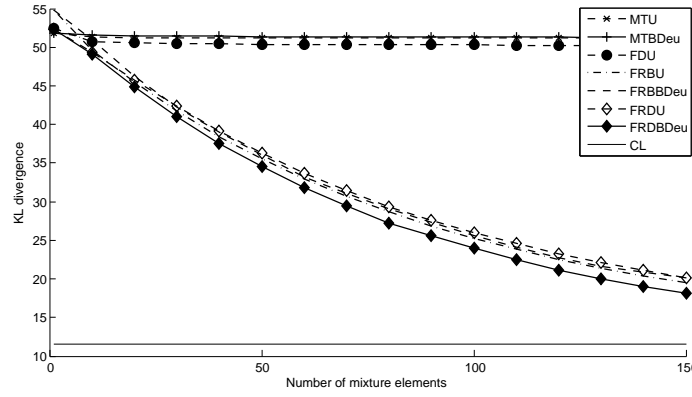
Figure 1 provides a representative set of learning curves corresponding to our simulations. The horizontal axis corresponds to the number  $m$  of mixture terms, whereas the vertical axis corresponds to the  $KL$  measures with respect to the target density. All the curves represent average results obtained over ten different datasets of 1000 learning samples and seven target distributions.

We thus observe in Figure 1 that our two sub-optimal tree mixture methods are clearly outperforming the random Markov tree mixture methods  $MTU$  and  $MTBDeu$  in terms of accuracy when we use uniform weighting schema ( $FDU$  and  $FBU$ ), but are slightly worst when we use non uniform weights.

Concerning our second proposed method, we observe from Figure 2 that all variants outperform well the random Markov tree mixture methods (which is linear in the number of variables) in terms of accuracy and are approaching the baseline  $CL$  (which is quadratic in the number of variables) when the number of mixture components grows. With all this sub-quadratic approaches, we also notice that the uniform weighting procedure is actually



**Fig. 1.** Naive sub-quadratic mixtures of trees for density estimation with a DAG target distribution. 10 experiments with a sample size of 1000 for 7 random target distributions of 1000 variables. (lower is better).



**Fig. 2.** Inertial sub-quadratic mixtures of trees for density estimation with a DAG target distribution. 10 experiments with a sample size of 1000 for 7 random target distributions of 1000 variables. (lower is better).

better than the one using weights based on the posterior probabilities given the dataset. Finally, we note that bagging principle do not provide better results than using the original dataset in this context of high dimensional problems and small datasets. All in all, the consistently best method in these trials is the method which uses uniform mixtures of sub-optimal trees built using the Chow and Liu algorithm on a partially fillet mutual information matrix whose terms are not generated at random and using the original dataset.

From a computational point of view, our proposed methods, whose complexity is  $n \log(n) \log(n \log(n))$  (sub-quadratic) provide better results than the linear methods and approach the single *CL* method which is quadratic.

## 5 Summary and future works

We have proposed in this paper to apply the Perturb and Combine principle in the context of unsupervised density estimation with graphical probabilistic models by using sub-optimal models learnt with an incomplete MI matrix and the Chow and Liu algorithm. We have presented two research heuristics for doing this and provide sub-quadratic computation complexity. The perturbation was done by partially fill the MI matrix by generating at random part of this matrix terms and use it to optimize the structure component, or by bootstrapping data before filling the MI matrix.

The most interesting result is that our second proposed method with a complexity very close to the quasi-linear, provides much better results than linear randomized mixtures of Markov trees and approaches the *CL* method which is quadratic in the number of variables.

As future work, complexity of our methods can be further improved by using some linear approximation of spanning tree algorithm (Chazelle (2000)) in order to obtain a lower complexity. We also believe that the combination of our approaches with sequential methods such as Boosting or Markov-Chain Monte-Carlo which have already been applied in the context of graphical probabilistic models, might provide a very rich avenue for the design of novel density estimation algorithms.

## References

- AMMAR, S., LERAY, Ph., DEFOURNY, B. and WEHENKEL, L. (2008): High-dimensional probability density estimation with randomized ensembles of tree structured bayesian networks. In: *Proceedings of the fourth European Workshop on Probabilistic Graphical Models (PGM08)*. 9–16.
- AMMAR, S., LERAY, Ph., DEFOURNY, B. and WEHENKEL, L. (2009): Probability Density Estimation by Perturbing and Combining Tree Structured Markov Networks. In: *ECSQARU '09: Proceedings of the 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer-Verlag, 156–167.
- CHAZELLE, B. (2000): A minimum spanning tree algorithm with inverse-Ackermann type complexity. *ACM 47 (6)*, 1028-1047.
- CHOW, C.K. and LIU, C.N. (1968): Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory 14 (3)*, 462-467.
- GEURTS, P., ERNST, D. and WEHENKEL, L. (2006): Extremely Randomized Trees. *Journal of Machine Learning 63 (1)*, 3-42.
- KULLBACK, S. and LEIBLER, R. (1951): On Information and Sufficiency. *Annals of Mathematical Statistics 22 (1)*, 79-86.
- PEARL, J. (1986): Fusion, Propagation, and Structuring in Belief Networks. *Artificial Intelligence 29*, 241-288.
- RIDGEWAY, G. (2002): Looking for lumps: boosting and bagging for density estimation. *Journal of Computational Statistics & Data Analysis 38 (4)*, 379-392.



# Data Management in Symbolic Data Analysis

Teh Amouh<sup>1</sup>, Monique Noirhomme-Fraiture<sup>1</sup>, and Benoit Macq<sup>2</sup>

<sup>1</sup> Faculté d’informatique, FUNDP

21 rue Grandgagnage, 5000 Namur, Belgique, {*tam, mno*}@info.fundp.ac.be

<sup>2</sup> Université catholique de Louvain, UCL

2 Place Stevin, 1348 Louvain-la-neuve, Belgique, *benoit.macq@uclouvain.be*

**Abstract.** Current data management process in the Symbolic Data Analysis framework totally relies on flat files for symbolic data persistency. This leads to some difficulties that we propose to solve using a database management system in place of flat files. Using our approach, symbolic data analysis and visualization algorithms will benefit effective and efficient data services that are currently difficult to implement. Our approach will also help in applications like time series analysis in the Symbolic Data Analysis framework.

**Keywords:** symbolic data management, database

## 1 Introduction

Symbolic data types include classical quantitative and categorical data types as well as structured data types. Most often used symbolic data types are single-valued quantitative data (e.g., numerical value 23.4), single-valued categorical data (e.g., value “French”), multivalued categorical data (e.g., set value {Brazilian, Spanish, French}), interval data (e.g., interval value [23, 29]) and modal data (e.g., modal value {(0.8)Spanish, (0.1)Brazilian, (0.1)French}). A way to obtain symbolic data is by aggregating classical data. This aggregation (or generalisation) process is illustrated below by an example based on Diday (2008). Let’s suppose that we have some classical data (see Table 1 and Table 2) about football players and teams that were involved in the World Cup 1998. Players are described by variables such as the team in which they play, their age, weight, height, nationality, place of

Player	Team	Age	Weight	Height	Nationality	...
Fernández	Spain	29	85	1.84	Spanish	...
Rodríguez	Spain	23	90	1.92	Brazilian	...
Mballe	France	25	82	1.82	Senegalese	...
Zidane	France	27	78	1.80	Senegalese	...

**Table 1.** Initial classical data table describing football players by three single-valued quantitative and two single-valued categorical variables.

Team	Goals	Coach	Experience	Captain	Skills	...
Spain	18	Clemente	10	Fernández	medium	...
France	24	Jacquet	21	Zidane	high	...

**Table 2.** Initial classical data table describing football teams by two single-valued quantitative and three single-valued categorical variables.

birth, etc... Teams are described by variables such as number of goals, coach's name, how many years the coach has been coaching football teams, captain's name, football skills of the captain and so on. If the purpose of our data analysis is to find an explanation for the number of goals scored by a team during the World Cup 1998, a natural way to proceed is to look for the answer by analyzing the variables describing the teams. For example, one could find that the number of goals scored by a team could be explained both by the experience of the coach and the skills of captain. However, a more elaborated explanation could be found if the variables describing football players are aggregated and included in the analysis. An aggregation process is applied to a set of players belonging to the same team and the result of this process is a statement of the internal variation of the description of the players inside each team (note that the set of players belonging to a given team is called the *extent* of the team). Table 3 is called a symbolic data table (SDT) and is the resulting table from the aggregation process. In this table, some of the initial classical variables describing teams are hidden so that there be enough place to show the aggregated variables whose names are written in capital letters in order to distinguish them from the underlying classical variables describing players. For example  $\text{AGE}(\text{Spain})=[23, 29]$  means that according to the

Team	Goals	AGE	WEIGHT	HEIGHT	NATIONALITY	...
Spain	18	[23, 29]	[85, 90]	[1.84, 1.92]	{(0.5) Sp, (0.5) Br}	...
France	24	[21, 28]	[78, 82]	[1.85, 1.90]	{(0.5) Fr, (0.5) Se}	...

**Table 3.** Symbolic data table (describing football teams) obtained by the generalization of classical variables that describe football players.

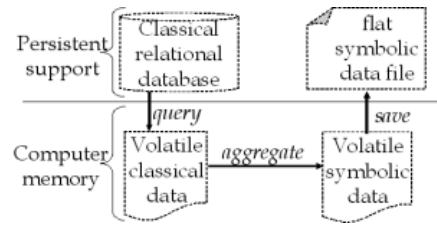
available initial data, the players with the team of Spain vary in age between 23 and 29. We can also read  $\text{NATIONALITY}(\text{France})=\{(0.5) \text{ Fr}, (0.5) \text{ Se}\}$  which means that according to the available data, 50% of the players with the team of France are French and the other 50% are Senegalese. Just like classical variables that describe the teams, symbolic variables generated by the aggregation process can also occur among the factors that explain the number of goals scored by a team. For example, one could find that in addition to the experience of coach and the skills of captain, teams with lighter and smaller players tend to score more goals.

The above example is a rather simple scenario where the aggregation process is guided by the values of only one variable (“Team”) and there is no other constraints weighing on the process. In real applications, the aggregation process can be guided by the Cartesian product of any number of variables, and there can exist logical and/or hierarchical dependency between variables. These are addressed by a research field called Symbolic Data Analysis which has attained a stature attested to by many publications and conferences (see Diday (2008) for a detailed state of the art). Currently, works in this research field concentrate on analysis and visualization techniques and there is no publication dealing with symbolic data storage and retrieval using database management technologies. In fact two relevant works relating to databases can be mentioned. Stéphan et al. (2000) have developed an approach that can extract classical data from databases and build symbolic data by aggregation, whereas Malherba et al. (2008) have developed the inverse process which consists in searching in databases the extent of some given symbolic data. In both cases however, no database technology is used for symbolic data management.

In Section 2 current data management process in Symbolic Data Analysis framework is recalled and the difficulties it poses are explained. We briefly present our proposition to use a DBMS (abbreviation for *database management system*) at the heart of the symbolic data management process rather than using flat files as persistent support for symbolic data. In Section 3 we propose a database schema and give details of our approach.

## 2 Data management process in Symbolic Data Analysis

We consider the case where symbolic data are computed from classical data stored in a relational database. Current data management actions in this case are shown on Figure 1 along with their input and output. The different



**Fig. 1.** Current data management process: from persistent classical data to persistent symbolic data.

actions (*query*, *aggregate* and *save*) are represented as arrows connecting an input to an output. First of all an SQL query is executed on the relational database in order to get initial quantitative and categorical data into the

computer memory. These data are aggregated and a SDT is constructed in the computer memory. Symbolic data analysis and visualization algorithms can then be applied on this SDT. The SDT along with analysis results can be saved onto persistent media which are currently flat files. Although many works have been done pertaining to the format of flat symbolic data files, current data management process in symbolic data analysis suffers from at least four problems.

One of these problems is the fact that flat files need to be manually managed (choice of file names, directories and physical storage units) and they can not be accessed simultaneously (or transversally). For example whenever complementary data are stored in symbolic variables spread throughout many data files, these files must be merged in order to get complete information. Thereby we obtain an additional file that is also manually maintained.

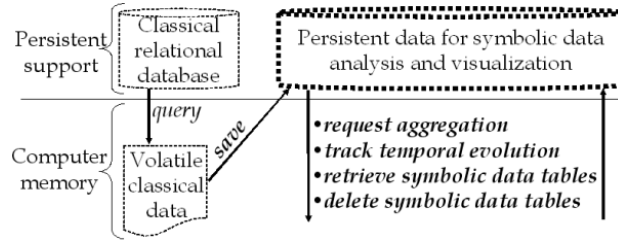
Another problem is that once a SDT is saved as a flat file and the base volatile classical data lost, the content of the saved flat file can not be used to derive other symbolic data using, for example, other aggregation criteria. Currently, if other aggregation criteria are to be used, the base volatile classical data must be reconstructed in the computer memory by issuing the corresponding query to the relational database where classical data are stored. An illustration of this problem is given by the breakdown process developed in Noirhomme-Fraiture and Nahimana (2008). This process constructs new symbolic descriptions, using a database search, through interaction with a visual representation.

Traditional tabular representation of symbolic data usually requires exponential computation time for analysis algorithms whenever hierarchical and/or logical dependency rules apply on the symbolic variables. However, the computation can be done in polynomial time if symbolic descriptions are decomposed in a way (called *normal symbolic form*) proposed by Csernel and de Carvalho (2008). There currently exists no flat file format for storing symbolic data in normal symbolic form.

Another problem relates to tracking temporal evolution of the base classical data provided by a relational database. It is currently fairly difficult to adjust symbolic data in accordance with modifications that occur on the base classical data.

When dealing with the management of large amounts of data, the superiority of a DBMS over flat files is obvious. To address the problems cited above, we propose to place a DBMS at the heart of the symbolic data management process as shown on Figure 2. We design a database schema that is able to support two main functionalities (see details in Section 3):

- Constructing different SDTs using different aggregation criteria on the same initial data without going back to query the relational database;
- Adjusting SDTs in accordance with changes in initial classical data.



**Fig. 2.** Our proposition to use a database management system at the heart of the symbolic data management process.

### 3 Details of the proposed approach

In this paper, we deliberately ignore eventual dependency rules among variables (variables are columns in a SDT). Metadata are also not considered. Dependency rules and metadata will be later integrated into our approach.

The aggregation process handles two kinds of variables:

**Description variables.** In Table 3, all columns except from the first one are said to be *description variables* because if we consider any row in this table, the set of values of these columns give a description of the team referred to by the row.

**Dimension variables.** Table 3 is one dimensional SDT because the aggregation process in our introductory example is guided by the values of only one variable (“Team”). Variable “Team” acts like a dimension for Table 3 in the sense that any row of this table is identified by one value of “Team”. Whenever the aggregation process is guided by a Cartesian product of two or more variables, the resulting SDT is multi-dimensional and any row in the SDT is identified by one value taken from the Cartesian product of these variables (the dimensions).

In our approach, any dimension variable is considered as a taxonomic variable which means that subsets of the set of values of a dimension variable are organised as a hierarchical class system. In our introductory example, we are interested in singleton subsets of the set of values of “Team”, so there is only one level in the hierarchy. Generally, viewing dimension variables as taxonomic variables provides the user the ability to define one hierarchy for each dimension variable and specify the level at which the aggregation process should operate. This information is stored in the database.

#### 3.1 Database schema and functionalities

The entity-relationship schema proposed on Figure 3 shows two entities (“QUERY” and “AGGREGATIONCRITERION”) connected by a relationship (“QUE\_AGG”). This conceptual database schema is designed to meet two requirements: the

first one is the ability to track temporal evolution of classical data, and the second one is to provide effective data services to analysis and visualization algorithms.



**Fig. 3.** Entity-relationship schema for data management in the Symbolic Data Analysis framework.

Tracking temporal evolution of initial data : Initial data provided by relational databases may change over time. There is currently no way to adjust SDTs basing on changes in the underlying initial data. In our approach, as illustrated on Figure 3, regardless of the relational database where initial data reside, user SQL commands stored into the “QUERY” table (along with the corresponding connection string in XML format), can be (re)executed at any time. The execution timestamp along with the response from the relational database can be appended to the XML content of the “Resultset” attribute corresponding to the executed query. The content of the “Resultset” attribute can be used to track changes overtime in initial data.

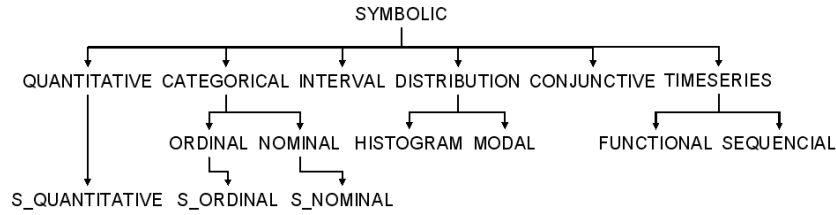
Providing effective data services to users : Figure 2 lists some data services our approach is able to provide. Users can at any time request data aggregation based on different criteria (as it is the case with the breakdown process mentioned in Section 2). The specification of these criteria (stored in XML format into the “AGGREGATIONCRITERION” table) includes a list of description variables and their symbolic types, and a specification of dimension variables and the hierarchies of their values. The generated symbolic data table is automatically stored into the “Data” attribute of the association between tables “QUERY” and “AGGREGATIONCRITERION”. The “Data” attribute is of a complex type called “SDT”. In Section 3.2 we consider implementation issues related to data type “SDT”.

A powerful feature of our approach lies in the fact that it gives users the ability to build symbolic data tables from the merge of two or more initial data tables extracted from different relational databases. This feature is enabled by the “[1-N]” cardinality at the side of table “AGGREGATIONCRITERION”.

### 3.2 Implementation issues

Regarding the aggregation process, different algorithms are available in literature (see Stéphan et al. (2000) and Lechevallier et al. (2008)). Our approach simply relies on these algorithms. We also refer to literature for the inverse process (Malherba et al. (2008)).

Regarding the symbolic database, current DBMSs are mostly based on the relational database model. The fundamental data representation used by relational DBMSs is a table of rows and columns where each column has a scalar (or atomic) data type (such as `INTEGER`, `DECIMAL`, `CHARACTER`, `DATE`, etc...). Since symbolic data types are mostly complex types (intervals, distributions and so on), mapping the content of a symbolic data table to a relational database table is not straightforward. Eventhough the database schema we propose in Figure 3 does not directly map the content of symbolic data tables to relational database tables, it makes use of a complex user-defined type namely “SDT” type that can hardly be dealt with using a relational DBMS. A more appropriate database model is the object-relational model which results from the convergence of the relational and object-oriented database worlds. In addition to the features of relational DBMSs (access control, transaction management, etc.), object-relational DBMSs provide among other object orientation functionalities the ability to create and manipulate instances of complex (or structured) data types defined by the user. These complex data types are called *User-Defined Types* (UDTs) and can include user-defined functions that applications can use to invoke appropriate behaviors of UDT instances (see Melton (2003)). Data type “SDT” as well as all symbolic data types can then be implemented as user-defined types hierarchies in an object-relational database management system (ORDBMS). We choose to implement our approach using the PostgreSQL environment which is an open-source ORDBMS. Figure 4 illustrates the type hierarchy we propose for symbolic data management with PostgreSQL DBMS. Data type “SDT”



**Fig. 4.** Symbolic data type hierarchy.

does not appear in this hierarchy because it is the data type of a symbolic data table. Each cell of a symbolic data table is of one of the symbolic types listed below type “SYMBOLIC” in the hierarchy. In the introduction of this paper, we discussed the following symbolic data types: single-valued quantitative, single-valued nominal, multivalued nominal, interval and modal. On Figure 4, these types are respectively referred to as “S\_QUANTITATIVE”, “S\_NOMINAL”, “NOMINAL”, “INTERVAL” and “MODAL”. Multivalued quantitative data are of type “QUANTITATIVE”. Both nominal and ordinal data are categorical data. Multivalued ordinal categories are of type “OR-

DINAL” and single-valued ordinal categories are of type “S\_ORDINAL”. “MODAL” data and “HISTOGRAM” data are both distributions. They differ from one another in the fact that for histograms, the probabilities (or weights) relate to intervals rather than categories. “FUNCTIONAL” data and “SEQUENTIAL” data are both time series. They differ from one another in the fact that for sequential data, values in the series are categories rather than quantities. “CONJUNCTIVE” data designate the case where several categories appear simultaneously (Diday (2008)). For example the color of an apple can be red or green or yellow but it can also be ‘red and yellow’.

## 4 Conclusion

The aim of Symbolic Data Analysis research field is to generalize data mining and statistics to higher level units described by symbolic data. Unfortunately, symbolic data are currently merely stored into flat files and researchers concentrate their efforts on analysis and visualization algorithms which operate on these flat files. An elaborated symbolic data management approach is proposed in this paper in order to take advantage of interesting functionalities (such as search facilities, changes tracking, etc.) provided by database management systems. This approach will help in applications like time series analysis.

## References

- CSERNEL, M. and DE CARVALHO, F.A.T (2008): The normal symbolic form. In: E. Diday and M. Noirhomme-Fraiture (Eds.): *Symbolic Data Analysis and the SODAS Software*. Wiley, New York, 93–107.
- DIDAY, E. (2008): The state of the art in symbolic data analysis: overview and future. In: E. Diday and M. Noirhomme-Fraiture (Eds.): *Symbolic Data Analysis and the SODAS Software*. Wiley, New York, 3–41.
- LECHEVALLIER, Y., EL GOLLI, A. and HEBRAIL, G. (2008): Improved generation of symbolic objects from relational databases. In: E. Diday and M. Noirhomme-Fraiture (Eds.): *Symbolic Data Analysis and the SODAS Software*. Wiley, New York, 45–59.
- MALERBA, D., ESPOSITO, F. and APPICE, A. (2008): Exporting symbolic objects to databases. In: E. Diday and M. Noirhomme-Fraiture (Eds.): *Symbolic Data Analysis and the SODAS Software*. Wiley, New York, 62–66.
- MELTON, J. (2003): *Advanced SQL:1999, Understanding Object-Relational and Other Advanced Features*. Elsevier Science, San Francisco.
- NOIRHOMME-FRAITURE, M. and NAHIMANA, A. (2008): Visualization. In: E. Diday and M. Noirhomme-Fraiture (Eds.): *Symbolic Data Analysis and the SODAS Software*. Wiley, New York, 109–120.
- STEPHAN, V., HEBRAIL, G. and LECHEVALLIER, Y. (2000): Generation of symbolic objects from relational databases. In: H.-H. Boch and E. Diday (Eds.): *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin, 78–105.



# Variable Selection for Semi-Functional Partial Linear Regression Models

Germán Aneiros<sup>1</sup>, Frédéric Ferraty<sup>2</sup> and Philippe Vieu<sup>2</sup>

<sup>1</sup> Departamento de Matemáticas, Universidade da Coruña  
Campus de Elviña s/n, 15071 A Coruña, Spain, *ganeiros@udc.es*

<sup>2</sup> Institut de Mathématiques, Université Paul Sabatier  
31062 Toulouse cedex, France, *ferraty@cict.fr*, *vieu@cict.fr*

**Abstract.** We consider a regression model where the regression function is the sum of a linear and a nonparametric component (that is, a partial linear regression model). More specifically, we focus on the case where the covariate that enters in a nonparametric way is of functional nature (see Aneiros-Pérez and Vieu (2006) for a first paper), the number of covariates in the linear part is divergent, and the corresponding vector of regression coefficients is sparse. The aim of this work is variable selection and estimation in such a model.

A penalized-least-squares based procedure to simultaneously select variables and estimate coefficients of variables is proposed, and a guideline is given for indicating how to select the various tuning parameters corresponding to our estimator. Finally, in order to illustrate the practical interest of the proposed procedure, a simulation study is reported

This work is related with those of Liang and Li (2009), Ni et al. (2009) and Xie and Huang (2009), who studied estimation and variable selection in partial linear regression models where all the covariates were scalar. Our main contribution is the introduction of a functional covariate in the model.

**Keywords:** functional data, semi-parametric regression, variable selection

## 1 Introduction

Modeling the relationship between a response and a set of predictors is of main interest in order to predict values of the response given the predictors. The larger the number of predictors, better fitted the model will be, but, if some predictors included in the model really do not influence the response, the model will not be good for predicting. Thus, in practice, it is needed some kind of methodology for selecting the significant covariates. Recent statistical literature has attacked this topic by proposing nonconcave penalized procedures.

In a setting of linear regression with sparse regression coefficients, Tibshirani (1996) proposed the LASSO method, a version of Ordinary Least Squares (OLS) that constrains the sum of the absolute regression coefficients, and Efron et al. (2004) gave the LARS algorithm for model selection (a modification of this algorithm implements the LASSO). Fan and Li (2001) proposed

and studied the use of nonconcave penalized likelihood for variable selection and estimation of coefficients simultaneously. Fan and Peng (2004) generalized the paper of Fan and Li (2001) to the case where a diverging number  $p_n < n$  of parameters is considered, and they noted that the prize to pay is a slower rate of convergence. Huang et al. (2008a) and Huang et al. (2008b) focused on particular classes of penalty functions (giving marginal bridge and adaptive LASSO estimators, respectively). Under a partial orthogonality condition on the covariance matrix, they obtained that their procedure can consistently identify the covariates with zero coefficients even when  $p_n > n$ .

Other authors dealt this topic in the setting where the regression function is the sum of a linear and a nonparametric component (that is, in Partial Linear Regression (PLR) models). Liang and Li (2009) considered a PLR model with fixed number  $p$  of covariates in the linear part, and measurement errors. In order to extend the procedure of Fan and Li (2001) to the new semi-parametric setting, they used local linear regression ideas. Ni et al. (2009) allowed a diverging number  $p_n < n$  of parameters and studied a double-penalized least squares. These authors used spline smoothing to estimate the nonparametric part of the model, and penalized both the roughness of the nonparametric fit and the lack of parsimony. Xie and Huang (2009) used a penalized least squares function based on polynomial splines, and also considered the case of a diverging number  $p_n < n$  of parameters. Their main contribution is in the fact that the proposed estimator was obtained as a global minimum of a penalized least squares function (in general, the estimators proposed in the statistical literature are obtained as local minimum).

In this paper we focus on a PLR model where the covariate that enters in a nonlinear way is of functional nature, such as a curve, an image, ... (see Aneiros-Pérez and Vieu (2006) for a first paper). In addition, the number of covariates in the linear part is divergent, and the corresponding vector of regression coefficients is sparse. The topic we deal is that of variable selection and estimation of coefficients simultaneously. We extend the methodology proposed in scalar models to this new functional setting, and deeply indicate how to select the various tuning parameters corresponding to our estimator. Finally, in order to illustrate the practical interest of our procedure, a modest simulation study is reported. As far as we know, this is the first paper attacking the problem of variable selection in a semi-functional PLR model.

## 2 The methodology

### 2.1 The model

We are concerned with the semi-functional PLR model

$$Y_i = \mathbf{X}_i' \boldsymbol{\beta}_0 + m(T_i) + \varepsilon_i, \forall i = 1, \dots, n, \quad (1)$$

where  $\beta_0 = (\beta_{01}, \dots, \beta_{0p_n})'$  is a vector of unknown sparse real parameters,  $m$  is an unknown smooth real function and  $\varepsilon_i$  are i.i.d. random errors satisfying

$$\mathbb{E}(\varepsilon_i \mid \mathbf{X}_i, T_i) = 0. \quad (2)$$

The covariates  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip_n})'$  and  $T_i$  take values in  $\mathbb{R}^{p_n}$  and some abstract semi-metric space  $\mathcal{H}$ , respectively.

## 2.2 The penalized least squares estimator

The regression function in (1) has a parametric and a nonparametric component. Thus, we need to simultaneously use parametric and nonparametric techniques in order to construct good estimators. On the one hand, the nonparametric approach that we consider is that of kernel estimation. More specifically, a Nadaraya-Watson type estimator is constructed by using the weight function

$$w_{n,h}(t, T_i) = \frac{K(d(t, T_i)/h)}{\sum_{j=1}^n K(d(t, T_j)/h)}, \quad (3)$$

where  $d(\cdot, \cdot)$  is the semi-metric associated to  $\mathcal{H}$ ,  $h > 0$  is a smoothing parameter and  $K : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  is a kernel function. On the other hand, the parametric procedure that we use is that of penalized least squares.

The steps to construct our estimator are, first, using kernel regression to transform the semi-parametric model (1) into a parametric model; then, apply to the transformed model the penalized-least-squared procedure in order to estimate  $\beta_0$ . To show this procedure clearer, let us denote  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)'$ ,  $\mathbf{Y} = (Y_1, \dots, Y_n)'$  and, for any  $(n \times q)$ -matrix  $\mathbf{A}$  ( $q \geq 1$ ),  $\tilde{\mathbf{A}}_h = (\mathbf{I} - \mathbf{W}_h)\mathbf{A}$ , where  $\mathbf{W}_h = (w_{n,h}(T_i, T_j))_{i,j}$ . Because

$$Y_i - \mathbb{E}(Y_i \mid T_i) = (\mathbf{X}_i - \mathbb{E}(\mathbf{X}_i \mid T_i))' \beta_0 + \varepsilon_i, \forall i = 1, \dots, n$$

(see (1) and (2)), we consider the approximate model

$$\tilde{\mathbf{Y}}_h \approx \tilde{\mathbf{X}}_h' \beta_0 + \varepsilon,$$

where  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$  (note that  $\tilde{\mathbf{Y}}_h$  and  $\tilde{\mathbf{X}}_h$  are formed by partial nonparametric residuals adjusting for  $T$ ). Thus, in order to estimate  $\beta_0$ , we minimize the penalized least squares function

$$\mathcal{Q}(\beta) = \frac{1}{2} (\tilde{\mathbf{Y}}_h - \tilde{\mathbf{X}}_h \beta)' (\tilde{\mathbf{Y}}_h - \tilde{\mathbf{X}}_h \beta) + n \sum_{j=1}^{p_n} \mathcal{P}_{\lambda_{jn}}(|\beta_j|), \quad (4)$$

where  $\mathcal{P}_{\lambda_{jn}}(\bullet)$  is a penalty function with a tuning parameter  $\lambda_{jn}$ .

Once one has the Penalized Least Squares (PLS) estimator  $\hat{\beta}_0$ , a natural estimator for  $m(t)$  is

$$\hat{m}(t) = \sum_{i=1}^n w_{n,h}(t, T_i) (Y_i - \mathbf{X}_i' \hat{\beta}_0). \quad (5)$$

### 2.3 Variable selection

It is known that, in scalar settings (linear as well as partial linear), when suitable penalty function and tuning parameters are used, the penalized least squares approach gives sparse solutions (see Fan and Li (2001); Fan and Peng (2004); Liang and Li (2009), among others). Thus, if the estimate of the parameter  $\beta_{0j}$  ( $j = 1, \dots, p_n$ ) is not equal to zero, then the corresponding covariate  $X_j$  is selected in the final model.

Our current theoretical studies (still unpublished) show that the introduction of a functional covariate does not change the situation. Therefore, we have a methodology for variable selection and estimation of coefficients simultaneously in semi-functional PLR models.

## 3 Selection of the tuning parameters

In order to implement our methodology, it is necessary to have at hand a guideline for selecting the tuning parameters  $h$  and  $\lambda_j$ , as well as the functions  $K(\bullet)$ ,  $\mathcal{P}_{\lambda_j}(\bullet)$  and  $d(\bullet, \bullet)$ .

The kernel  $K(\bullet)$  has low impact on the estimates, and it is common to use the Epanechnikov kernel defined by  $K(u) = 0.75(1 - u^2)I_{(0,1)}$ . In contrast, the role of the penalty function  $\mathcal{P}_{\lambda_j}(\bullet)$  is fundamental for obtaining (from an asymptotic point of view) sparse estimates, stability of model selection and unbiased estimates. In this way, several authors recommended using the Smoothly Clipped Absolute Deviation (SCAD) penalty. The expression of this function is

$$\mathcal{P}_{\lambda}(z) = \begin{cases} \lambda z, & \text{if } 0 \leq z < \lambda, \\ \frac{(a^2-1)\lambda^2 - (z-a\lambda)^2}{2(a-1)}, & \text{if } \lambda \leq z < a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |z| \geq a\lambda, \end{cases} \quad (6)$$

where  $a > 2$ , being common to take  $a = 3.7$  (see Fan and Li (2001); Fan and Peng (2004), among others). A challenge of using the SCAD penalty is the minimization of (4), because (6) is nondifferentiable and nonconcave. Thus, several algorithms were proposed in the literature in order to optimize this kind of functions (see Fan and Li (2001) and Zou and Li (2008) for algorithms based on local quadratic and local linear approximation, respectively).

The semi-metric  $d(\bullet, \bullet)$  controls, from certain point of view, the concentration of the functional data. Therefore, it plays a major role to prevent the “curse of dimensionality” (remember that we are in an infinite dimensional context). Ferraty and Vieu (2006) suggested to use semi-metrics based on semi-norms. Then, they recommended, for smooth functional data, to take as semi-norm the  $L_2$  norm of some  $q$ -th derivative of the function. For the case of rough data curves, these authors suggested to construct a semi-norm based on the first  $q$  functional principal components of the data curves.

Finally, a data-driven procedure must be considered for the choice of the parameter  $q$  in the semi-norm, the tuning parameter  $\lambda_j$  in the penalty function, and the smoothing parameter  $h$  in the Nadaraya-Watson type weights (3). On the one hand, it should be noted that, when  $p_n$  is large, the task of selecting the tuning parameters  $\lambda_j$  ( $j = 1, \dots, p_n$ ) is hard. Thus, it is commonly considered that  $\lambda_j = \lambda sd(\hat{\beta}_{0,j,OLS})$ , where  $\hat{\beta}_{0,j,OLS}$  is the OLS estimate of  $\beta_{0,j}$  in the model (1), and  $sd(\hat{\beta}_{0,j,OLS})$  denotes its standard deviation. On the other hand, for finite sample sizes, empirical studies have shown a better behaviour of the Nadaraya-Watson type estimators when a bandwidth  $h_k$ , allowing to take into account  $k$  terms in (5), is used instead of  $h$ . Thus, we are interested in selection of  $q$ ,  $\lambda$  and  $k$ .

In statistical literature, there exist various procedures for selecting tuning parameters, including cross-validation and generalized cross-validation. Classical cross-validation works well, but is computationally expensive, especially if several parameters must be selected. The method of generalized cross-validation allows to quickly select the parameters, but its interpretation is not too intuitive. The data-driven procedure that we propose is the fivefold cross-validation method, which was used, among others, in Zou and Li (2008). It consists in selecting the value of  $\theta = (q, \lambda, k)$  that minimizes the criterion

$$CV(\theta) = \sum_{v=1}^5 \sum_{(Y_j, \mathbf{X}_j, T_j) \in D^v} \left( Y_j - \hat{r}_{\theta}^{(v)}(\mathbf{X}_j, T_j) \right)^2,$$

where  $D^v \subset D$  is a test sample ( $D$  denotes the full dataset), and  $\hat{r}_{\theta}^{(v)}(x, t)$  is the estimate of the regression function  $r(x, t) = x' \beta_0 + m(t)$  constructed using only the data in  $D - D^v$ , and the tuning parameter  $\theta$  (in order to obtain  $\hat{r}_{\theta}^{(v)}(\bullet, \bullet)$ ,  $r(\bullet, \bullet)$  is estimated via estimation of  $\beta_0$  and  $m(\bullet)$  by means of the techniques shown in the previous section). Assuming that the size of  $D^v$  relative to  $D$  is not too small, this procedure has shown to work well in practice, and is computationally less expensive than the classical cross-validation.

## 4 A simulation study

A modest simulation study was designed in order to illustrate the practical behaviour of the proposed procedure. The semi-functional PLR model

$$Y_i = X_{i1}\beta_{01} + X_{i2}\beta_{02} + \dots + X_{ip_n}\beta_{0p_n} + m(T_i) + \varepsilon_i, \forall i = 1, \dots, n,$$

was considered, and the size of the vector of parameters was  $p_n = \lceil 5n^{1/4} \rceil - 8$ .

The scalar data  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip_n})'$  were standard normal, and  $\varepsilon_i$  was  $N(0, \sigma_{\varepsilon})$  with  $\sigma_{\varepsilon} = 0.1(\max_T m(T) - \min_T m(T))$ . A dependence structure between the covariates was considered, and the results were compared with the independent case. More specifically, the correlation between  $X_{ij}$  and  $X_{ik}$

was  $\rho^{|j-k|}$  (values  $\rho = 0$  and  $\rho = 0.5$  were considered). The functional data were  $T_i(z) = a_i(z - 0.5)^2 + b_i$  ( $z \in [0, 1]$ ), with  $a_i$  and  $b_i$  being i.i.d. according to a  $U(0, 1)$  and a  $U(-0.5, 0.5)$ , respectively. These curves were discretized on the same grid of 100 equispaced points in  $[0, 1]$ .

The unknown vector of parameters was

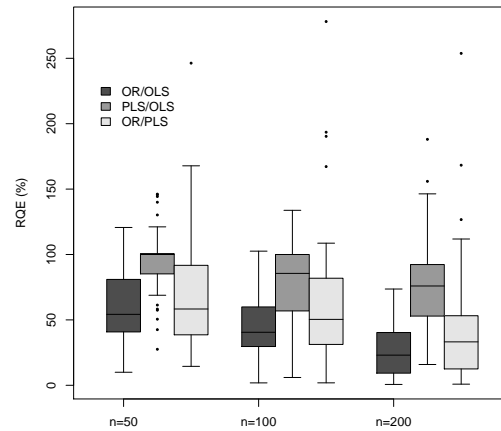
$$(\beta_{01}, \dots, \beta_{0p_n})' = (3, 1.5, 0, 0, 2, 0, \dots, 0),$$

while the unknown function  $m(\bullet)$  was

$$m(T_i) = \exp(-8f(T_i)) - \exp(-12f(T_i)),$$

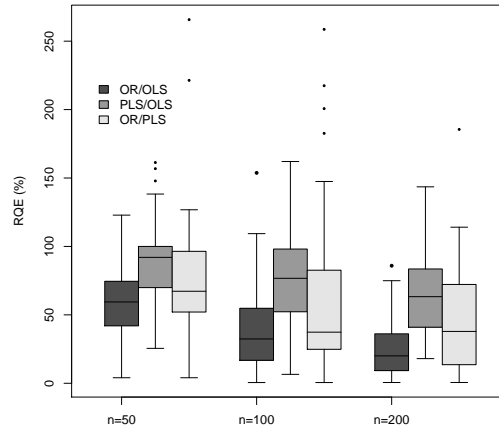
where

$$f(T_i) = \text{sign}(T_i'(1) - T_i'(0)) \sqrt{3 \int_0^1 (T_i'(z))^2 dz}.$$



**Fig. 1.** Relative quadratic errors when  $\rho = 0$ .

We simulated various sample sizes  $n = 50, 100, 200$ , and, for each of them, the same experiment was replicated  $M = 50$  times. For each of the  $M$  replicates, we compute both the PLS and the OLS estimates, as well as the Oracle (OR) estimate (that is, the OLS estimate based on the true submodel). Values for the tuning parameters  $\theta = (q, \lambda, k)$  were selected following the guideline given in the previous Section 3 (the smoothness of the curves  $T_i$  lead us to consider semi-metrics based on the  $L_2$  norm of the  $q$ -th derivative of the curves), and the size of the test samples  $D^v$  was  $0.25n$ . The Epanechnikov kernel was used, while the penalty function was the SCAD penalty ( $a = 3.7$ ).



**Fig. 2.** Relative quadratic errors when  $\rho = 0.5$ .

Figs. 1 and 2 display (for  $\rho = 0$  and  $\rho = 0.5$ , respectively) the  $M$  Relative Quadratic Errors (RQE) obtained for each pair of estimates (the quadratic error of  $\hat{\beta}$  for estimating  $\beta$  is defined as  $(\hat{\beta} - \beta)'(\hat{\beta} - \beta)$ ). In both figures, the value for the 38-th replicate corresponding to the combination  $n = 100$  and OR/PLS was not shown because it is a very large outlier (but was used to construct the box-plots).

In addition, Table 1 reports the averages (on the  $M$  replicates) of the number and the percentage of coefficients correctly set to zero (no coefficient was incorrectly set to zero).

$n$	$p_n$	Correct zero coefficients	
		$\rho = 0$	$\rho = 0.5$
50	5	0.42 [21%]	0.58 [29%]
100	7	1.42 [35%]	1.62 [40%]
200	10	2.92 [42%]	3.34 [48%]

**Table 1.** Averages of the number and the percentage of coefficients correctly set to zero.

The computation time (Intel Core 2 Duo T9400 processor, 2.53 GHz CPU) to obtain the value of our estimate was, approximately, 3.5, 9 and 19.25 minutes for the combinations  $(n, p_n) = (50, 5)$ ,  $(100, 7)$  and  $(200, 10)$ , respectively (including selection of the parameter  $\theta = (q, \lambda, k)$  in a grid of  $3 \times 50 \times n/2$  points).

## 5 Final comments

Naturally, the results of any simulation study are valid only for the models considered, and in that sense should be interpreted. Figs. 1 and 2 suggest that the PLS estimator performs better than the LS, especially as the sample size increases. However the OR estimator clearly surpasses. This may be indicative of the need for large sample sizes to obtain good results in the functional setting. In addition, Table 1 shows how, as the sample size increases, our estimator detects a greater percentage of nonsignificant variables. Finally, no significant effect of the dependence structure is observed.

In future works, we hope to provide theoretical advances and more illustrations on finite size Monte Carlo simulated samples.

**Acknowledgements.** The research of G. Aneiros was supported by Xunta de Galicia Grant PGIDIT07PXIB105259PR, and by the research group MODES. F. Ferraty and P. Vieu wish to thank all the participants of the working group STAPH on Functional Statistics in Toulouse for their numerous and interesting comments.

## References

- ANEIROS-PÉREZ, G. and VIEU, P. (2006): Semi-functional partial linear regression. *Statistics and Probability Letters* 76, 1102-1110.
- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004): Least angle regression. *The Annals of Statistics* 32, 407-499.
- LIANG, H. and LI, L. (2009): Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association* 104, 234-248.
- FAN, J. and LI, R. (2001): Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348-1360.
- FAN, J. and PENG, H. (2004): Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* 32, 928-961.
- FERRATY, F. and VIEU, P. (2006): *Nonparametric Functional Data analysis*. Springer, New York.
- HUANG, J., HOROWITZ, J. L. and MA, S. (2008a): Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics* 36, 587-613.
- HUANG, J., MA, S. and ZHANG, C-H (2008b): Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica* 18, 1606-1618.
- NI, X., ZHANG, H. H. and ZHANG, D. (2009): Automatic model selection for partially linear models. *Journal of Multivariate Analysis* 100, 2100-2111.
- TIBSHIRANI, R. (1996): Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 58, 267-288.
- XIE, H. and HUANG, J. (2009): SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics* 37, 673-696.
- ZOU, H. and LI, R. (2008): One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* 36, 1509-1533.



# Clustering Functional Data Using Wavelets

Anestis Antoniadis<sup>1</sup>, Xavier Brossat<sup>2</sup>, Jairo Cugliari<sup>2,3</sup>, and Jean-Michel Poggi<sup>3,4</sup>

<sup>1</sup> Université Joseph Fourier, Laboratoire LJK, Tour IRMA, BP53, 38041 Grenoble Cedex 9, France, *anestis.antoniadis@imag.fr*

<sup>2</sup> EDF R&D, 1 avenue du Général de Gaulle, 92141 Clamart Cedex, France, *xavier.brossat@edf.fr*

<sup>3</sup> Université Paris-Sud, Mathématique Bât. 425, 91405 Orsay, France  
*jairo.cugliari@math.u-psud.fr*, *jean-michel.poggi@math.u-psud.fr*

<sup>4</sup> Université Paris 5 Descartes, France

**Abstract.** This paper presents a method for effectively detecting patterns and clusters in high dimensional time-dependent functional data. It is based on wavelet-based similarity measures since wavelets are ideal for identifying highly discriminant local time and scale features. We consider the contribution of each scale to the global energy, in the orthogonal wavelet transform of each input function to generate a handy number of features that still makes the signals well distinguishable. Our new similarity measure combined with an efficient feature selection technique in the wavelet domain is then used within more or less classical clustering algorithms to effectively differentiate among high dimensional populations.

**Keywords:** clustering, functional data, wavelets

## 1 Introduction

In different fields of applications explanatory variables are not multivariate observations of classical statistics, but are functions observed either discretely or continuously. Typical examples of functional data can be found when studying electricity consumption, temporal gene expression analysis or ozone concentration in environmental studies to cite only a few.

Given a sample of curves, one important task is to search for homogeneous subgroups of individuals using clustering and classification. Clustering is one of the most frequently used data mining techniques, which is an unsupervised learning process for partitioning a dataset into sub-groups so that the instances within a group are similar to each other and are very dissimilar to the instances of other groups. In a functional context clustering helps to identify representative curve patterns and individuals who are very likely involved in the same or similar processes. Many functional clustering methods have been developed, ranging from heuristic approaches, such as variants of the  $k$ -means method (Tarpey and Kinader (2003)) and clustering after transformation and smoothing (Serban and Wasserman (2005)) to more formal

model-based procedures, such as clustering sparsely sampled functional data (James and Sugar (2003)) and, most recently, the  $k$ -centres functional clustering approach (Chiou and Li (2007)). In this paper we propose a wavelet-based methodology for clustering time series of smooth curves.

Our interest in time series of curves is motivated by an application in forecasting a functional time series when the most recent curve is observed. This situation arises frequently when a seasonal univariate time series is sliced into consecutive segments, for example days, and treated as a time series of functions. The idea of forming a functional time series from a seasonal univariate time series has been introduced by Bosq in 1990 and considered by several authors (see Antoniadis and Sapatinas (2003) and references within). The central issue in the analysis of such data consists in taking into account the temporal dependence of these functional observations. For most of the applications cited above the detrended functional times series are modeled as function-valued stationary processes allowing the development of efficient forecasting procedures. In practice, however, many observed functional time series cannot be modeled accurately as stationary. The important class of nonstationary time series includes, for example, those measured in a changing environment or describing an evolving phenomenon. Recognizing this, our aim is therefore to propose a clustering technique that clusters the functional times series into groups that may be considered as stationary so that in each group more or less standard functional prediction procedures can be applied.

The rest of the paper is organized as follows. The next section contains a reminder on multiresolution analysis and introduces the basis supporting our feature extraction algorithm by means of the energy operator. Following wavelet analysis we then cluster the data using the extracted features. Our clustering algorithm uses  $k$ -means as an unsupervised learning routine. Then, we present some other concurrent methods for clustering functional data that are available in the recent literature, and finally, an experimental evaluation of the proposed algorithm on simulated and on real data is provided.

## 2 Wavelets and energy distribution across scales

Let us introduce some basic ideas of the wavelet analysis. A compactly supported WT uses a orthonormal basis of waveforms derived from scaling and translations of a compactly supported scaling function  $\phi$  and a compactly supported mother wavelet  $\psi$ . Any function  $z \in \mathcal{H} = \mathcal{L}^2([0, \delta])$  can then be decomposed in terms of an orthogonal basis, given by the collection  $\{\phi_{j_0, k}, k = 0, 1, \dots, 2^{j_0} - 1; \psi_{j, k}, j \geq j_0, k = 0, 1, \dots, 2^j - 1\}$  for any  $j_0 \geq 0$ , of the following form:

$$z(t) = \sum_{k=0}^{2^{j_0}-1} c_{j_0, k} \phi_{j_0, k}(t) + \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} d_{j, k} \psi_{j, k}(t), \quad (1)$$

where  $c_{j,k}$  and  $d_{j,k}$  are called respectively the scale and the wavelet coefficients of  $z$  at the position  $k$  of the scale  $j$  defined as

$$c_{j,k} = \langle z, \phi_{j,k} \rangle_{\mathcal{H}} \quad d_{j,k} = \langle z, \psi_{j,k} \rangle_{\mathcal{H}}$$

For short, a wavelet is a smooth and quickly vanishing oscillating function with good localization properties in both frequency and time, this is more suitable for approximating time series curves that contain localized structures. The energy  $\mathcal{E}_Z = \|z\|_{\mathcal{H}}^2$  of the time series  $z$  via discrete wavelet decomposition is equal to the sum of the energy of its wavelet coefficients distributed across scales:

$$\mathcal{E}_z \approx \|\mathbf{Z}\|_2^2 = c_{0,0}^2 + \sum_{j=0}^{J-1} \|\mathbf{W}_j\|_2^2, \quad (2)$$

the approximation holding because of the truncation at scale  $J$  for the wavelet expansion of  $\mathbf{Z}$ , discarding finer scales. This relation justifies the use of the energy of wavelet coefficients for computing squared Euclidean distances between two series. However, since we are interested on how the energy of wavelet coefficients is distributed across scales other distance functions on DWT decompositions may be more appropriate for measuring the similarity between two series.

In what follows, define for  $j = 0, \dots, J-1$  the contribution and relative contribution of the scale  $j$  to the global energy of the centered function

$$\text{cont}_j = \|\mathbf{W}_j\|_2^2 \quad \text{rel}_j = \frac{\text{cont}_j}{\sum_{j=0}^{J-1} \text{cont}_j}.$$

A scale in the energy difference with a relative low relative contribution may be disregarded when comparing time series because it probably embeds lots of noise in the corresponding wavelet coefficients. We will therefore characterize each time series by the vector of its energy contributions or its relative contributions in order to define an appropriate measure of similarity that is going to be used for clustering.

### 3 A K-means like functional clustering procedure

The infinite-dimensional original objects are reduced to  $J$  features representing the dynamic of the curves across different scales. We handle two versions for the representation: the first one is the original absolute contribution (abbreviated (AC)) while the second is the relative contribution representation (RC). So, for (AC) we have a vector of positive components that sums up the global energy on the details  $\sum_j \|\mathbf{W}_j\|_2^2$  meanwhile for (RC) the vector has all its components positives summing to one. In other words we have a probability vector, which is then transformed in order to avoid this normalization, using the logit

$$p \rightarrow \log \left( \frac{p}{1-p} \right)$$

It is well known, as a consequence of the curse of dimensionality, that the  $k$ -means technique suffers from the increasing number of features. In our case, the number of features depends on the number of discretization points for which we acquire data. For  $N$  points, the number of features is  $J = \log_2(N)$  which can be relatively important. Moreover, since we are interested in the energy decomposition across scales, it is highly probable that several scales will not be informative for making the cluster. To decide which information we retain with we use a variable selection algorithm for no supervised learning proposed by Steinley and Brusco (2008).

Finally, to determine a convenient number  $K$  of clusters several data-driven strategies can be defined, at least in the classical case. The first one amounts in inspecting basically the within-cluster dissimilarity as a function of  $K$ . Many heuristics have been proposed trying to find a “kink” in the corresponding plot. A more formal argument has been proposed by Tibshirani et al. (2001) by comparing, using the gap statistic, the logarithm of the empirical within-cluster dissimilarity and the corresponding one for uniformly distributed data.

An information theoretic point of view provided by James and Sugar (2003), considering the transformed distortion curve  $d_K^{-p/2}$ , a kind of average Mahalanobis distance between data and the set of cluster centers as a function of  $K$ . Jumps in associated plot allow to select sensible values for  $K$  while the largest one may be the best choice for a mixture of  $p$ -dimensional distributions with common covariance. An asymptotic analysis (as  $p$  goes to infinity) states that, when the number of clusters used is smaller than the true number, then the transformed distortion remains close to zero, before jumping suddenly and increasing linearly.

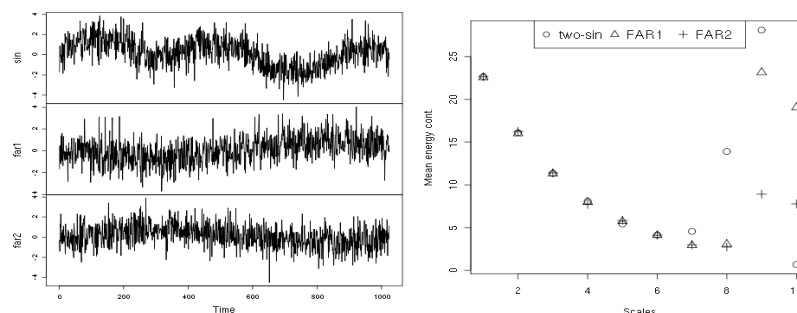
## 4 Simulations and comparisons on real data

We study the empirical performance of our clustering method by applying it to a simulated data set and to the electricity power consumption in France.

### 4.1 Simulated example

We start with a simple simulated example to show the adequacy of our procedures. We simulate  $K = 3$  clusters of 25 observations each: a 1024\*25 points trajectory is obtained for each cluster and each trajectory is divided in 25 non overlapping subintervals. The first cluster is a simple superposition of two sines and a white noise:  $f(x) = \sin(5\pi x/1024) + \sin(2\pi x/1024) + \epsilon$ . For the second and third cluster, we use two functional autoregressive processes: the first one has a diagonal covariance structure, the second one exhibits full

covariance matrix. These covariances are chosen to clearly distinguish the first model, dominated by a low frequency trend, from the two others whose differences are more intricate as we can see on the left side of Figure 1. However, if we calculate the (AR) for these models we can see (right side of Figure 1) how our representations shows a better discrimination for the models with only a few handy features.



**Fig. 1.** On the left, some typical simulated trajectories of the sinus model (top panel), the FAR1 model (middle), and the FAR2 model (bottom). On the right, the average energy contribution across scales over the simulated models.

To effectively detect which are the informative scales we use the Steinley-Brusco algorithm for variable selection in unsupervised learning. We retain scales 8 to 10 which are associated with the lowest frequencies. We use the  $k$ -means algorithm (that we initialize many times, retaining the minimum within cluster distance solution) where the input data are the selected scales of the (AR) and the number of cluster is the true one. Only 17 of the 75 observations were misclassified. The separation between FAR models and sine model is perfect (no error), thanks to the specific scale 8 on the scale domain that appears to clearly discriminate this model from the others. However the separation between the two FAR models is better than expected thanks to scale 10 which helps in this rather delicate task. While 12 observations from the FAR2 model were classed as FAR1, the specific error for FAR1 was only 5 observations. Our procedure gives largely better results than those obtained by clustering raw data using the  $L_2$  distance where 29 observations are wrongly classified and even the sinus model is not identified.

## 4.2 Power load supply

We now examine one year of national power load supply of the French producer EDF recorded each 30 minutes. We make segments of 48 points that represents daily profiles. We expect to find some well known facts by EDF engineers about french electricity consumption like the important thermo-sensitivity and highly dependence on social phenomena as the dichotomy be-

tween working-days, weekend-day and holidays. It is also usual to find intra weekly differences. But also we would like to exploit the descriptive power of a clustering analysis that could detect shapes of special days (affected by bank holidays for example).

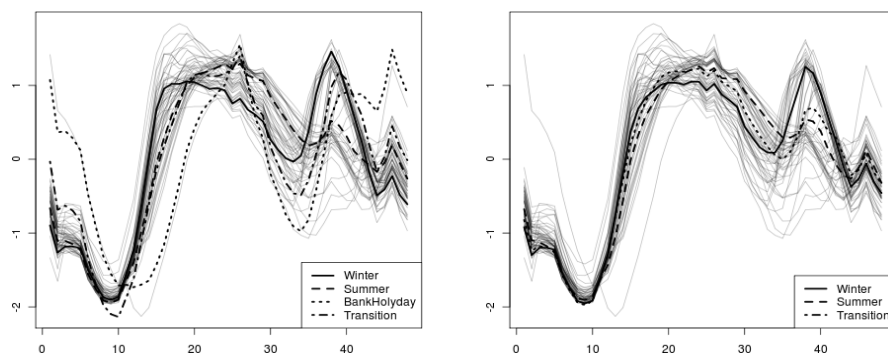
Raw data exhibit the effect of seasonal and week cycles making the profiles vary in mean level and in shape. Moreover, when profiles are centered and/or scaled there is still a great variation in the curves. This shows that higher order moments contribute also to the variability in the dynamic of the curves.

As before, we use the DWT over the 48-point discretized version of the daily profiles to compute the wavelet coefficients. Then we calculate both AC and RC representations reducing up to 6 handy features per day our originally functional daily data. For a reasonable wide range of number of clusters, we apply Steinley and Brusco's feature selection algorithm retains scales 3 to 5 that represent the dynamics of 1.5, 3 and 6 hours corresponding respectively for the AC representation. Meanwhile, for the RC representation the significative variables for detecting the cluster structure are 1 and 4 which corresponds to the 22.5 minutes and 3 hours cycles respectively. For both representations we choose the number of clusters by detecting the largest jump in the distortion curve (James and Sugar (2003)) obtaining 11 and 8 clusters for AC and RC respectively. Finally, we start many times the  $k$ -means algorithm with the selected variables and the chosen number of clusters. For each representation we retain the minimum within cluster distance solution.

Let us sketch some interesting facts from the empirical analysis. Two well defined periods are highlighted: the first one between May and October and the second covering from mid-December to March (electrical heating in these months is very important); with two clearly "half-season transitions" from November to mid-December and April. In each period we can see a clear dichotomy between working days and weekend days that evolves within each period showing a consistent behavior with respect to the two seasons structure.

Let us illustrate with the class of Fridays. Left panel of Figure 2 shows the 53 load curves for this type of day with four centroids for the AC. There is one clear different cluster that corresponds to the bank holidays cluster. Then, we also obtain a hot weather and a cold weather cluster with different shapes. For example, the maximum of daily demand is attained in the afternoon for the winter curves while for the summer curves we found it at midday. A centroid corresponding to the "transition" season is also included. The demand for these days shows less amplitude than the first part of the day.

We repeat the analysis for the RC representation. Right panel of Figure 2 shows three centroids and the same Friday's load curves. We omit bank holiday centroid because it is essentially the same. Centroids are more similar than in the AC representation: for example, the winter centroid has on the first half of the day more or less the same shape than the other two clusters. Remark that when clustering with RC, the dissimilarity measure does not



**Fig. 2.** On the left panel, four centroids for the AC representation and the centered and scaled load curves corresponding to Fridays. On the right panel, three centroids for the RC representation with centered and scaled Friday data.

take into account the differences in scales. Hence, using both variants of clustering drive us to detect weather the difference between two functions could be explained only by means of changes in scale. This is the case for the 'summer break period' (August) where the centered version detects a cluster of these days but the standardized version does not.

## 5 Discussion: towards using wavelet coherence

The success of any clustering algorithm depends on the adopted dissimilarity measure. Direct similarity measures such as  $L_p$  norms match two functional objects in their original representations without explicit feature extraction. When  $p = 2$ , this reduces to commonly used Euclidean distance.  $L_p$  norms are straightforward and easy to compute. However, in many cases such as in shifting and scaling, the distance of two sequences cannot reflect the "real" (dis)similarity between them. Furthermore,  $L_p$  distance has meaning only in the relative sense when used to measure (dis)similarity.

In the previous section we have proposed instead the usage of the discrete wavelet transform of two times series of equal length to define a weighted normalized Euclidian like distance between them as a measure of their similarity. Indeed this was supported by the fact that the similarity of time series data should be based on certain characteristics of the data rather than on the raw data itself by concentrating most of the energy in a small region of the scale-frequency domain.

We would like to comment here on another direct and intuitive similarity measure that could be used for matching sequential patterns. This adopted similarity measure is based on the wavelet coherence between two time series. This concept provides a way of analyzing local correlation of times series both in the time domain and in the frequency domain. In this, it fundamentally differs from Fourier coherence that relies upon the correlation of the two

series in the frequency domain only. In addition to locality, the continuous wavelet transform on which the wavelet coherence is based possesses the very desirable ability of filtering the polynomial behavior to some predefined degree and therefore is invariant to vertical or scale shifts. Therefore, correct characterization of time series is possible, in particular in the presence of nonstationarities like global or local trends or biases.

To conclude let us just say that wavelets offer an excellent framework when data are not stationary. For example, Gurley et al. (2003) develop a wavelet-coherence concept that is proved elsewhere to be convenient for clustering geophysical time series or to detect and cluster spikes on neural activity. With such motivations the wavelet transform property of time-frequency localization of the time series, we propose hereafter a time-series feature extraction algorithm using orthogonal wavelets for automatically choosing feature dimensionality for clustering.

## References

- ANTONIADIS, A. and SAPATINAS (2003): Wavelet methods for continuous-time prediction using Hilbert valued autoregressive processes. *Journal of Multivariate Analysis*, 87(1), 133–158.
- ANTONIADIS, A., PAPARODITIS, E. and SAPATINAS, T. (2006): A functional wavelet-kernel approach for time series prediction. *Journal Royal Statistical Society Series B Statistical Methodology*, 68(5), 834–857.
- BOSQ, D. (1990): Sur les processus autorégressifs dans les espaces de Hilbert. *Publ. Instit. Stat. Paris*, XXXV, 2, p. 3–17.
- CHIOU, J.-M. and LI, P.-L. (2007): Functional clustering and identifying substructures of longitudinal data. *J.R. Statist.Soc. B*, 69(4), 679–699.
- GURLEY, K., KIJEWski, T. and KAREEM, A. (2003): First-and higher-order correlation detection using wavelet transforms. *Journal of engineering mechanics*, 129(2), 188–201.
- JAMES, G.M. and SUGAR, C.A. (2003): Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical Association*, 98(462), 397–409.
- JAMES, G.M. and SUGAR, C.A. (2003): Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach. *Journal of the American Statistical Association*, 98(463), 750–764.
- MALLAT, S.G. (1999): A wavelet tour of signal processing. *Academic Press*.
- SERBAN, N. and WASSERMAN, L. (2005): Cats: clustering after transformation and smoothing, *J. Am. Statist. Ass.*, 100, 990–99.
- STEINLEY, D. and BRUSCO, M.J. (2008). A New Variable Weighting and Selection Procedure for K-Means Cluster Analysis. *Multivariate Behavioral Research*, 43(1), 77–108.
- STEINLEY, D. and BRUSCO, M.J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, 73(1), 125–144.
- TARPEY, T. and KINATEDER, K.K.J. (2003): Clustering functional data. *Journal of Classification*, 20(1), 93–114.
- TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(2), 411–423.



# Polynomial Methods in Time Series Analysis

Félix Aparicio-Pérez

Instituto Nacional de Estadística  
Castellana 183, 28046, Madrid, Spain *fapape@ine.es*

**Abstract.** Polynomial methods have been extensively used in system theory either as an alternative or in conjunction with state space methods. However, its use in time series analysis has been very limited. This paper highlights the main results in matrix polynomial algebra, like the solution of matrix polynomial equations or the transformation between a right and a left matrix fraction description and provides some mostly unknown applications in time series analysis, like the evaluation of the autocovariances of a VARMA process, the obtention of the model that follows a filter of a VARMA process or the computation of a multivariate Wiener-Kolmogorov filter.

**Keywords:** polynomial matrices, time series, Wiener-Kolmogorov filter, autocovariances

## 1 Polynomial matrices

To make the exposition simple, we can consider a polynomial matrix of dimension  $n \times m$  as an array consisting of  $n$  rows and  $m$  columns, where each element is a polynomial in a complex variable  $z$ . A more formal algebraic definition can be found, for example, in Callier and Desoer (1982), see also Kailath (1980), Chen (1984) and chapters 7 and 14 in Lancaster and Tismenetsky (1985) for a more detailed treatment, and Hannan and Deistler (1988), section 2 for some results that are of direct interest in statistical discrete-time stochastic systems.

Some definitions and properties of polynomial matrices are similar to those of scalar matrices, for example, the determinant of a square polynomial matrix is well defined (as a polynomial), but others are not, for example, the inverse of a polynomial matrix with non-zero determinant exists but may not be a polynomial matrix.

A square polynomial matrix is called unimodular if its determinant is a non-zero scalar. The inverse of a unimodular matrix is a polynomial matrix. As an example the polynomial matrix  $a(z) = \begin{pmatrix} 1-z & z \\ z & 1-0.5z \end{pmatrix}$  has determinant  $1 - 1.5z - 0.5z^2$ , and its inverse is the rational matrix  $a^{-1}(z) = (1 - 1.5z - 0.5z^2)^{-1} \cdot \begin{pmatrix} 1-0.5z & -z \\ -z & 1-z \end{pmatrix}$ , while the polynomial matrix  $b(z) =$

$\begin{pmatrix} 1 - z^2 & -2z \\ 2z & 4 \end{pmatrix}$  has determinant 4 and is thus unimodular, its inverse is the polynomial matrix  $b^{-1}(z) = \begin{pmatrix} 1 & 0.5z \\ -0.5z & 0.25 - 0.25z^2 \end{pmatrix}$ .

The degree of a  $n \times m$  polynomial matrix is defined as the maximum of the degrees of the  $nm$  polynomials that it has as elements. Sometimes we write  $a(z) = A_0 + z \cdot A_1 + \dots + z^p \cdot A_p$ , where  $p$  is the degree of  $a(z)$  (in the preceding example  $a(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + z \cdot \begin{pmatrix} -1 & 1 \\ 1 & -0.5 \end{pmatrix}$ ) and similarly for  $b(z)$  and other polynomial matrices. A basic result about  $n \times m$  polynomial matrices is that they can be reduced by means of pre(post)-multiplication by a unimodular polynomial matrix to row(column) Hermite form. We do not define a Hermite form here, but in the particular case that  $n = m$  this result implies that a square polynomial matrix can be transformed into an upper(lower) triangular form by means of pre(post)-multiplication by a unimodular polynomial matrix. Moreover, this can be done in a numerically reliable way, as explained in Henrion and Šebek (1998). The importance of this is that the determinant of a triangular polynomial matrix is equal to the product of its diagonal elements and the determinant of a unimodular matrix  $U(z)$  is simply the determinant of  $U_0$ .

For example,  $R(z) = U(z) \cdot a(z)$ , where  $U(z) = \begin{pmatrix} 1+z & z \\ -z & 1-z \end{pmatrix}$  is a unimodular matrix with determinant 1 and  $R(z) = \begin{pmatrix} 1 & 2z + 0.5z^2 \\ 0 & 1 - 1.5z - 0.5z^2 \end{pmatrix}$  is an upper triangular matrix, so  $\det(a(z)) = (\det(U_0))^{-1} \cdot \det(R(z)) = (1)^{-1} \cdot (1 - 1.5z - 0.5z^2)$ .

If  $P(z) = P_1(z) \cdot M(z)$  ( $P(z) = L(z) \cdot P_1(z)$ ) and  $Q(z) = Q_1(z) \cdot M(z)$  ( $Q(z) = L(z) \cdot Q_1(z)$ ) with  $M(z)$  ( $L(z)$ ) a polynomial matrix,  $M(z)$  ( $L(z)$ ) is called a common right divisor (common left divisor) of  $P(z)$  and  $Q(z)$  and  $P(z)$  and  $Q(z)$  are said to be multiples of  $P_1(z)$  and  $Q_1(z)$  respectively. If  $M(z)$  ( $L(z)$ ) is a multiple of any other common right (left) divisor of  $P(z)$  and  $Q(z)$ , it is called a greatest common right (left) divisor.

We are interested in VARMA processes of the form

$$a(B)y_t = b(B)\epsilon_t, \quad (1)$$

where  $y_t$  is a process of dimension  $n$ ,  $B$  is the backshift operator ( $By_t = y_{t-1}$ ),  $\epsilon_t$  is a gaussian white noise process of dimension  $n$  with covariance matrix  $\Sigma_\epsilon$  and  $a(z)$  and  $b(z)$  are  $n \times n$  polynomial matrices. We still need some more definitions and results.

A  $s \times m$  rational transfer function  $T(z)$  is a  $s \times m$  array that has as elements polynomial quotients.

A right coprime fraction (r.c.f) or right coprime matrix fraction description, of  $T(z)$  is a pair of polynomial matrices,  $(N_r(z), D_r(z))$ , of orders  $s \times m$  and  $m \times m$  respectively such that:

- (i)  $D_r(z)$  is non-singular (its determinant is not the zero polynomial).
- (ii)  $T(z) = N_r(z)D_r(z)^{-1}$ .
- (iii)  $(N_r(z), D_r(z))$  is right-coprime, that is, all its greatest common right divisors are unimodular matrices.

A left coprime fraction (l.c.f) or left coprime matrix fraction description, of  $T(z)$  is a pair of polynomial matrices,  $(N_l(z), D_l(z))$ , of orders  $s \times m$  and  $s \times s$  respectively such that:

- (i)  $D_l(z)$  is non-singular.
- (ii)  $T(z) = D_l(z)^{-1}N_l(z)$ .
- (iii)  $(N_l(z), D_l(z))$  is left-coprime, that is, all its greatest common left divisors are unimodular matrices.

In many system theory books and papers, the term matrix fraction description is reserved for systems written in the forward shift operator  $F = B^{-1}$ . In this paper we do not follow this convention and call for example the pair  $(a(B), b(B))$  of our VARMA process a left matrix fraction description, since if all the roots of  $(\det(a(z)))$  are outside the unit disk, then  $y_t$  has a transfer function given by  $a^{-1}(z)b(z)$ . We require that  $(a(z), b(z))$  be left coprime.

An important result states that given a  $n \times m$  rational transfer function  $T(z)$ , it can always be expressed as a r.c.f. or l.c.f.  $T(z) = D_l(z)^{-1}N_l(z) = N_r(z)D_r(z)^{-1}$ . An example of a  $2 \times 1$  transfer function expressed as a r.c.f. and a l.c.f. is:

$$T(z) = \begin{pmatrix} z(z-1)(z+2) \\ z+1 \end{pmatrix} ((z+1)(z-1))^{-1} = \begin{pmatrix} z+1 & z-1 \\ 0 & (z-1)^2 \end{pmatrix}^{-1} \begin{pmatrix} (z+1)^2 \\ z-1 \end{pmatrix}$$

In practice, if we have a, not necessarily coprime, left fraction or right fraction of  $T(z)$  we may be interested in transforming it into a l.c.f. and/or a r.c.f. Fortunately, this can also be done in a numerically reliable way, by solving a so-called minimal polynomial basis problem, see e.g. Chen(1984).

## 2 Some Applications

We now propose some polynomial methods, that can be used to solve certain problems that arise when working with VARMA processes. The method in section 2.1 is not new in system theory and signal processing, but it has some advantages over the methods that are used in time series analysis, while the methods in sections 2.2 and 2.3 are, to the author's best knowledge, new.

These methods are only illustrations of what can be done in the field of time series using polynomial techniques. Some recent work that is related to the work in this paper can be found in Anderson and Deistler (2009).

### 2.1 Matrix polynomial equations and autocovariances

Several kinds of polynomial equations arise in system theory and signal processing. Some of them are described in Kučera (1979). We will focus on the

so-called symmetric matrix polynomial equation, that has the form

$$A'(z^{-1})X(z) + X'(z^{-1})A(z) = B(z) \quad (2)$$

where  $A(z)$  and  $B(z)$  are given polynomial matrices with real coefficients and  $B(z)$  is para-Hermitian, that is  $B(z) = B_l(z^{-1}) + B_r(z)$ , with  $B_l(z) = B_r'(z)$ .

This equation has two main applications, one is spectral factorization, that we will use in sections 2.2 and 2.3, the other one is the computation of the autocovariances of a VARMA system, see Söderström, Ježek and Kučera (1998). For a stationary VARMA process of the form (1), the autocovariance generating function is  $G(z) = a^{-1}(z)b(z)\Sigma b'(z^{-1})a'^{-1}(z^{-1})$ , we are looking for a decomposition of the form  $G(z) = M(z) + M'(z^{-1})$ . Pre-multiplying by  $a(z)$ , post-multiplying by  $a'(z^{-1})$  and calling  $X'(z) = a(z)M(z)$  we get, after transposition,  $b(z^{-1})\Sigma b'(z) = a(z^{-1})X(z) + X'(z^{-1})a'(z)$ , this is equation (2) with  $B(z) = b(z^{-1})\Sigma b'(z)$  and  $A(z) = a'(z)$ . To find the autocovariances of the process we first solve this symmetric matrix polynomial equation for  $X$ , with the condition that  $X_0$  be symmetric, and then, since  $M(z) = (1/2)I_0 + zI_1 + z^2I_2 + \dots$ , ( $I_i$  is the lag- $i$  autocovariance of  $y_t$ ), we solve recursively (long division) the equation  $a(z)M(z) = X'(z)$  to get the first autocovariances, then the Yule-Walker equations can be used to obtain the next ones.

The solution of the symmetric matrix polynomial equation can be found in an efficient and numerically reliable way, as explained in Henrion and Šebek (1998). This method is more efficient than the methods that are usually employed in time series analysis (e.g. Reinsel(1997), pp. 59-60).

## 2.2 VARMA process filtering and matrix fraction descriptions

Given a VARMA process  $y_t$  of the form (1), sometimes it is necessary to compute the model that follows some linear combination(s) of its components. More in general, the linear combination(s) may include delayed components. This problem is usually addressed in time series using ad-hoc hand computations for each case, but these computations grow quickly in complexity with the dimension of  $y_t$  and with the number of linear combinations or if the white noise components are correlated.

The method in p. 436 of Lütkepohl (2005) has several drawbacks, first it obtains a model with a diagonal AR part (called a final form model), but such a model is not necessarily left coprime, second it involves the computation of the adjoint of a polynomial matrix. As we will see in section 3, this adjoint can be computed automatically and efficiently, but it is not explained in Lütkepohl (2005) how to do so, moreover the method that we will describe is more efficient.

Suppose that we want to compute the VARMA model that follows the process  $z_t = F(B)y_t$ , where  $F(z)$  is an  $s \times n$  polynomial matrix. After solving in (1) for  $y_t$  we pre-multiply by  $F(B)$  and obtain  $z_t = F(B)a^{-1}(B)b(B)\epsilon_t$ ,

but  $F(B)a^{-1}(B) = \tilde{a}^{-1}(B)\tilde{F}(B)$ , that is, we transform a right matrix fraction description into a left one. Finally we do the spectral factorization  $\tilde{F}(B)b(B) = c(B)u_t$ , where  $u_t$  is a new white noise with covariance matrix  $\Sigma_u$ . The final model is  $\tilde{a}(B)z_t = c(B)u_t$ . Both the transformation from a right to a left matrix fraction and the spectral factorization can be done using automatic and numerically reliable and efficient methods. The details of the implementation of this method and the extension to the case of a rational filter of the form  $G(B)z_t = F(B)y_t$  are in Aparicio-Pérez(2009a).

### 2.3 Exact multivariate Wiener-Kolmogorov Filtering

An exact method for the computation of a univariate Wiener-Kolmogorov filter based on a finite sample can be found in Burman(1980). The exact multivariate case based on a finite sample has been addressed in the literature before using state-space methods and, for some particular cases, like the signal plus noise model or the deconvolution problem, using polynomial methods (e.g. Ahlén and Sternad (1991)). We will provide a brief description of a new polynomial method that solves the general multivariate case, the details and the extension to the non-stationary case are in Aparicio-Pérez (2009b).

Assume that two multivariate processes,  $s_t$  and  $y_t$ , follow jointly a stationary, invertible and left coprime VARMA model with VAR part  $a(B)$  and MA part  $b(B)$  that we consider partitioned as in

$$\begin{pmatrix} a_{11}(B) & a_{12}(B) \\ a_{21}(B) & a_{22}(B) \end{pmatrix} \begin{pmatrix} s_t \\ y_t \end{pmatrix} = \begin{pmatrix} b_{11}(B) & b_{12}(B) \\ b_{21}(B) & b_{22}(B) \end{pmatrix} \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix} \quad (3)$$

Assume also that a finite sample of  $y_t$  is available, but no observations from  $s_t$  are available. We are interested in estimating the values of  $s_t$ . This is a fixed interval smoothing problem that can be solved exactly using the Kalman filter and smoother.

In a polynomial setting, we first transform the model into another one that has a diagonal AR part, this is accomplished by pre-multiplying (3) by  $Adj(a(B))$ , the adjoint of  $a(B)$ , the result is

$$det(a(B))I_n y_t = \begin{pmatrix} d_{11}(B) & d_{12}(B) \\ d_{21}(B) & d_{22}(B) \end{pmatrix} \begin{pmatrix} \hat{\epsilon}_{1t} \\ \hat{\epsilon}_{2t} \end{pmatrix} \quad (4)$$

where  $d(z) = Adj(a(z))b(z)L'$ , with  $L'L = \Sigma_\epsilon$  (Cholesky decomposition),  $\hat{\epsilon}_t = (L^{-1})'\epsilon_t$  is a standardized white noise process and  $I_n$  is the identity matrix of dimension  $n$ .

Now we will use the Wiener-Kolmogorov formula that assumes that we have a doubly infinite realization of  $y_t$ , see Caines(1988) p. 139. The key points are (i) the diagonal AR part cancels out and (2) since we actually have a finite sample, we use so many exact finite sample forecasts and backcasts of  $y_t$  as needed (we only need a few of them). Because of the properties of conditional expectations (or equivalently, because of the properties of Hilbert

space projections) this procedure will provide the exact Wiener-Kolmogorov filter based on the finite sample. The joint covariance generating function of  $s_t$  and  $y_t$  is  $G(z) = (\det(a(z)))^{-1} d(z) d'(z^{-1}) (\det(a(z^{-1})))^{-1}$  and the optimal filter is  $\hat{s}_t = G_{12}(B) \cdot G_{22}^{-1}(B) = [d_{11}(B) d_{12}(B)] \begin{bmatrix} d'_{21}(F) \\ d'_{22}(F) \end{bmatrix} \Theta'^{-1}(F) \Theta^{-1}(B) \hat{y}_t$ , where  $\Theta(z)$  is such that  $[d_{21}(z) d_{22}(z)] \begin{bmatrix} d'_{21}(z^{-1}) \\ d'_{22}(z^{-1}) \end{bmatrix} = \Theta(z) \Theta'(z^{-1})$ , (it is obtained using spectral factorization), and the notation  $\hat{y}_t$  stands for the exact prediction of  $y_t$  based on the finite sample, that is  $y_t$  itself for the time points in which there are observations and the exact backcasts or forecasts for another time points. So, we can compute the exact finite Wiener Kolmogorov filter running three cascaded filters, these are  $\hat{x}_t = \Theta^{-1}(B) \hat{y}_t$ , with time running forwards,  $\hat{v}_t = \tilde{\Theta}'^{-1}(F) \tilde{e}(F) \hat{x}_t$ , with time running backwards, where  $e'(z) = [d_{21}(z) d_{22}(z)]$  and  $e(z^{-1}) \Theta'^{-1}(z^{-1}) = \tilde{\Theta}'^{-1}(z^{-1}) \tilde{e}(z^{-1})$  (we transform a right fraction into a l.c.f) and  $\hat{s}_t = [d_{11}(B) d_{12}(B)] \hat{v}_t$  with time running forwards. We need the exact estimates  $\hat{y}_t$  for  $t \in \{0, -1, \dots, -r + 1\}$ , where  $r$  is the degree of  $\Theta(z)$ , in the first of the three cascaded filters. As in the univariate case, (a suggestion of Tunnicliffe-Wilson in Burman(1980)), they can be computed by solving a linear system of equations involving the coefficients of the VARMA model for  $y_t$  and those of the time-reversed model (the model that follows  $y_t$  when the time  $t$  runs from the future towards the past). In the univariate case, the time-reversed model is the same as the original model but with a different white noise, but in the multivariate case both models are different. The computation of the time-reversed model for the multivariate case can be done simply by finding a so-called echelon realization of the process that has as autocovariances the transposes of those of  $y_t$ , as explained in Aparicio-Pérez and Gómez(2008). See Hannan and Deistler (1988) for a definition and properties of echelon realizations.

We have solved the fixed interval smoothing problem, but the filtering, the fixed lag smoothing and the forecasting of  $s_t$  can be done in a similar way, provided that we compute the exact corresponding  $\hat{y}_t$  values based on the information available for each situation.

The exact forecasts and backcasts of  $y_t$  can be computed, for example, using the innovations algorithm (e.g. Brockwell and Davis (1991), p. 425) applied to the VARMA model of  $y_t$  and the corresponding time-reversed model respectively.

### 3 Numerical computations

When a result about polynomial matrices is proved, usually, the proof consists of applying elementary row and column operations to the polynomial matrices. These are (i) interchanging two rows or columns, (ii) multiplying a row or column by a non-zero polynomial and (iii) adding to a row (column) another row (column) multiplied by a polynomial. These operations

are of theoretical interest, but cannot be used in practice, except for simple problems, because numerical problems often arise.

All the polynomial techniques introduced in this paper can be implemented in a numerically reliable and efficient way, and some references to papers that explain how to do so have already been given.

A VARMA process can be expressed using several left matrix fractions  $(a(B), b(B))$ , all of them having the same transfer function  $T(z) = a^{-1}(z)b(z)$ . Thus, it is important to find a unique representative of the class of processes that have a given transfer function. One way to solve this problem is to find an echelon realization of a given process. Echelon realizations are coprime, so they can also be used to transform a left (or a right) matrix fraction into a l.c.f. and/or r.c.f. The method in Chen (1984), sections A-3 and G-6, computes an echelon realization of a right or left fraction, using a numerically stable method, it also reveals its algebraic structure, summarized by the so-called Kronecker indices, see section 2 in Hannan and Deistler (1988).

It is also possible to find an echelon realization of a process when we do not have a pair  $(a(B), b(B))$ , provided that we have enough  $T_i$  matrices in the expansion of its transfer function  $T(z) = T_0 + z \cdot T_1 + z^2 \cdot T_2 + \dots$  or enough autocovariances (enough here means a finite quantity that depends on the Kronecker indices of the process). To do so we can use the technique in section 2 of Hannan and Deistler (1988) for the transfer function case or the technique in Aparicio-Pérez and Gomez (2008) for the autocovariances case. This last technique can also be used to obtain an echelon realization of the time-reversed process, as was said before. Both techniques are numerically reliable.

The computation of the adjoint of a  $n \times n$  polynomial matrix  $a(z)$  is usually done by transforming first the polynomial matrix into an equivalent matrix pencil (a degree one polynomial matrix). We prefer to use another method that is simpler, but also reliable and efficient. It consists of first computing the determinant by transforming the matrix into a triangular form, as pointed out in section 2, and then computing the adjoint recursively (long division) from the condition that  $a(z)Adj(a(z)) = det(a(z))I_n$ .

Sometimes, the computation of the inverse of  $a(z)$  is necessary. Once the determinant and the adjoint have been obtained, one expression of the inverse as a left fraction is of the form  $(det(a(z))I_n, Adj(a(z)))$ , since  $a^{-1}(z) = (det(a(z))^{-1}I_n Adj(a(z)))$  as in the case of the inverse of a scalar matrix. This left fraction may not be coprime, but it can be transformed into a l.c.f or a r.c.f as we said before.

There are more polynomial techniques that may be needed, for example, the computation of a greatest common right or left divisor, row or column reduction of a polynomial matrix or transformation of a VARMA process into a regular form (one with  $A_p$  non-singular, where  $p$  is the degree of  $a(z)$ ).

Finally, transformations from VARMA form to State Space form and vice versa are also very important. The simple methods to do these transforma-

tions work well only for low dimensions, and more complex techniques have to be used for moderate or high dimensions.

## References

- AHLÉN, A., and STERNAD, M. (1991): Wiener Filter Design Using Polynomial Equations. *IEEE Transactions on Signal Processing*. 39 (1), 2387-99.
- ANDERSON, B.D.O. and DEISTLER, M. (2009): Properties of Zero-Free Spectral Matrices. *IEEE Transactions on Automatic Control*. 54 (10), 2365-75.
- APARICIO-PÉREZ, F. (2009a): *Matrix Polynomial Computations with Applications in Time Series Analysis*. Unpublished manuscript, INE, Madrid, Spain.
- APARICIO-PÉREZ, F. (2009b): *Multivariate Wiener-Kolmogorov Filtering by Polynomial Methods*. Unpublished manuscript, INE, Madrid, Spain.
- APARICIO-PÉREZ, F., and GÓMEZ, V. (2008): Inversión en el Tiempo de Procesos VARMA Mediante Técnicas Polinomiales. *Revista BEIO*. 24 (3), 23-26. (Downloadable from [http://www.seio.es/BEIO/files/BEIOv24n3\\_E0\\_FAparicio+VGomez.pdf](http://www.seio.es/BEIO/files/BEIOv24n3_E0_FAparicio+VGomez.pdf))
- BROCKWELL, P.J., and DAVIS, R.A. (1991): *Time Series: Theory and Methods, Second Edition*. Springer-Verlag, New York.
- BURMAN, J.P. (1980): Seasonal Adjustment by Signal Extraction. *Journal of the Royal Statistical Society, Ser. A*. 143, 321-337.
- CAINES, P.E. (1988): *Linear Stochastic Systems*. John Wiley.
- CALLIER, F.M. and DESOER, C.A. (1982): *Multivariable Feedback Systems*. Springer-Verlag, New York.
- CHEN, C.T. (1984): *Linear System Theory and Design*. Holt, Rinehart and Winston.
- HANNAN, E.J., and DEISTLER, M. (1988): *The Statistical Theory of Linear Systems*. John Wiley.
- HENRION, D., and ŠEBEK, M. (1998): Efficient numerical method for the discrete-time symmetric matrix polynomial equation. *IEEE Proceedings, Control Theory and applications*. 145 (5), 143-148.
- KAILATH, T., (1980): *Linear Systems*. Prentice-Hall, Englewood Cliffs, N.J.
- KUČERA, V. (1979): *Discrete Linear Control-The Polynomial Equation Approach*. Wiley, Chichester.
- LANCASTER, P., and TISMENETSKY, M. (1985): *The Theory of Matrices, Second Edition*. Academic Press, London.
- LÜTKEPOHL, H. (2005): *New Introduction to Multiple Time-Series Analysis*. Springer-Verlag, Berlin Heidelberg.
- REINSEL, G.C., (1997): *Elements of Multivariate Time Series Analysis, Second Edition*. Springer-Verlag, New York.
- SÖDERSTRÖM, T., JEŽEK, J. and KUČERA, V. (1998): An efficient and versatile algorithm for computing the covariance function of an ARMA process. *IEEE Transactions on Signal Processing* 46 (6), 1591-1600.



# Cointegrated Lee-Carter Mortality Forecasting Method<sup>★</sup>

Josef Arlt<sup>1</sup>, Markéta Arltová<sup>1</sup>, Milan Bašta<sup>1</sup>, and Jitka Langhamrová<sup>2</sup>

<sup>1</sup> Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic, *arlt@vse.cz, arltova@vse.cz, basta@vse.cz*

<sup>2</sup> Department of Demography, Faculty of Informatics and Statistics, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic, *langhamj@vse.cz*

**Abstract.** Since its introduction the Lee-Carter method has become an important statistical model for mortality forecasting. The classical Lee-Carter forecasting method is based on the assumption that the overall mortality index follows the random walk model with drift or the ARIMA model. If the two sexes in one country are forecasted separately, the forecasts of male and female mortalities can diverge increasingly over time despite the fact that this effect has not been observed in the past. This problem can be solved by the Cointegrated Lee-Carter method. On the example of some European countries it is shown that this method leads to more tied up forecasts of the overall mortality index with smaller standard errors in comparison with the classical Lee-Carter method.

**Keywords:** mortality, Lee-Carter method, cointegration

## 1 Lee-Carter Method

### 1.1 Model

Since its introduction (Lee and Carter, 1992) the Lee-Carter method has become the "leading statistical model of mortality in the demographic literature" (Deaton and Paxson, 2004). Lee and Carter developed their method for U.S. mortality data in years 1933-1987. Afterwards, this method has been than applied to analyze and mainly to forecast the mortality in many other countries.

The principle of the Lee-Carter method is relatively simple. If  $m_{xt}$  denotes the log of the mortality rate in the age group  $g = 1, 2, \dots, G$  and at time  $t = 1, 2, \dots, T$  for one country, the Lee-Carter model has form

$$m_{gt} = \alpha_g + \beta_g \gamma_t + \varepsilon_{gt}, \quad (1)$$

---

<sup>★</sup> This paper was written with the support of Grant Agency of the Czech Republic No. 402/09/0369 "Modelling of Demographic Time Series in Czech Republic".

where  $\alpha_g$ ,  $\beta_g$  and  $\gamma_t$  are parameters to be estimated and  $\varepsilon_{gt}$  are homoskedastic normally distributed random disturbances with mean 0 and variance  $\sigma_\varepsilon^2$ . The identification of the model is ensured by the constraints

$$\sum_{t=1}^T \gamma_t = 0 \quad \text{and} \quad \sum_{g=1}^G \beta_g = 1. \quad (2)$$

The parameters of model (1) have the following interpretation: Because of the constraint  $\sum_{t=1}^T \gamma_t = 0$  the parameter  $\exp \alpha_g$  represents the general shape of the mortality schedule. The parameter  $\beta_g$  indicates the dependence of the log mortality at age  $g$  on the time trend  $\gamma_t$ , which can be interpreted as the overall mortality index (the general time level of mortality). The shape of the function  $\beta_g$  shows how quickly the mortality rates decline over time in comparison with trend  $\gamma_t$ <sup>1</sup>.

## 1.2 Estimation

The parameters of the Lee-Carter model are usually estimated by the ordinary least squares method, specifically the estimators  $\hat{\alpha}_g$ ,  $\hat{\beta}_g$  and  $\hat{\gamma}_t$  should minimize the sum of squares

$$\sum_{g=1}^G \sum_{t=1}^T (m_{gt} - \alpha_g - \beta_g \gamma_t)^2. \quad (3)$$

As on the right-hand side of model (1) is not observed independent variable, which can be observed, it is not a simple regression model. The minimum of (3) is achieved by  $\hat{\alpha}_g = 1/T \sum_{t=1}^T m_{gt}$  (it follows from the constraint  $\sum_{t=1}^T \gamma_t = 0$ ) and by  $\hat{\beta}_g$  and  $\hat{\gamma}_t$  received from the singular value decomposition (SVD) of matrix  $(m_{gt} - \hat{\alpha}_g)$ . Lee and Carter also recommended the second stage of estimation, namely the adjustment of  $\hat{\gamma}_t$  (taking  $\hat{\alpha}_g$  and  $\hat{\beta}_g$  as given) to guarantee that the observed total number of death  $d_t = \sum_{g=1}^G d_{gt}$  equals to the total number of deaths predicted by the model, i. e.

$$d_t = \hat{d}_t = \sum_{g=1}^G p_{gt} e^{(\hat{\alpha}_g + \hat{\beta}_g \hat{\gamma}_t)}, \quad (4)$$

where  $p_{gt}$  is the mid-year population of the age group  $g$  at time  $t$ .

Wilmoth (1993) describes two other alternatives to the basic Lee-Carter estimation procedure: a) a weighted least square method, b) a maximum

<sup>1</sup> Girosi and King (2007) analyzed comprehensively the properties of the Lee-Carter method. They mentioned that the Lee-Carter method is a special case of the principal component method, "where the log-mortality data is summarized only using the first single principal component, with other variation ignored for the purpose of making forecasts". When the first principal component is insufficient to catch the majority of data variance than the Lee-Carter model is not be expected to forecast as well.

likelihood method. They are based on his insight that the reason why the observed number of deaths is generally differs from than the fitted number of deaths follows from the fact that the estimates of  $\gamma_t$  are computed by minimizing the sum of the least squares with the log-mortality instead of the mortality. The first method is not used in practice frequently, mainly as it leads to biased estimators and there are also some computation problems. The second one is more perspective. It follows from the fact that the number of deaths is a counting random variable, which can be modeled by a Poisson process. The model can be expressed in the form

$$d_{gt} \sim \text{Poisson}(p_{gt}e^{(\alpha_g + \beta_g \gamma_t)}). \quad (5)$$

The parameters are still subjected to the constraints (2). The parameter estimates are results of the maximization of the log-likelihood function based on the model (5). From the practical experiments made by authors of this article, the results of this estimation procedure and the classical two stage procedure are very similar.

### 1.3 Forecasting

Lee and Carter created the model mainly for forecasting purposes. They assumed  $\hat{\alpha}_g$  and  $\hat{\beta}_g$  remain constant over time and the overall mortality index  $\hat{\gamma}_t$ , which is in the fact the univariate time series, will be modeled by the principles of the Box-Jenkins methodology. The aim is to create its forecasts and consequently with the help of estimates  $\hat{\alpha}_g$  and  $\hat{\beta}_g$  the forecasts of the mortality. After testing of several ARIMA models, they find that the most appropriate model for forecasting of the US data is a random walk with drift, i. e.

$$\hat{\gamma}_t = c + \hat{\gamma}_{t-1} + a_t, \quad (6)$$

where  $a_t$  are normally distributed non-autocorrelated random disturbances with mean 0 and variance  $\sigma_a^2$ . This process is called as I(1) (process integrated of the first order).

In the application of the Lee-Carter method on the data of other countries the other models are used, for example ARIMA(0,1,1) with drift or ARIMA(2,1,1) with drift. It is essential that the overall mortality index  $\hat{\gamma}_t$  is a nonstationary time series of I(1) type with a declining trend.

### 1.4 Extending

Li and Lee (2005) pointed that the Lee-Carter method works well for a single population - for one sex or two sexes combined. But there is problem with the application of this method to forecast mortality for two sexes in the same country. In the case of the separate forecasting of two sexes, the male and female mortality can diverge increasingly over time despite of the fact that this effect has not been observed in history. A similar problem comes up in

the forecasting of the mortality in different provinces or races in a country, different countries in a region etc.

To avoid this problem Li and Lee suggested extending the Lee-Carter model in to the so-called common factor model. If  $m_{xti}$  denotes the log of mortality rate in the age group  $g = 1, 2, \dots, G$ , time  $t = 1, 2, \dots, T$  and for example sex  $i = 1, 2$  for one country, the common factor Lee-Carter model for sex  $i$  has form

$$m_{gti} = \alpha_{gi} + B_g \Gamma_t + \varepsilon_{gti}, \quad (7)$$

where  $\alpha_{gi}$ ,  $B_g$  and  $\Gamma_t$  are parameters to be estimated and  $\varepsilon_{gti}$  are homoskedastic normally distributed random disturbances with mean 0 and variance  $\sigma_\varepsilon^2$ . The parameter  $\exp \alpha_{gi}$  represents the general shape of the mortality schedule of sex  $i$ . In this model it is assumed that both sexes in the group have the same  $\beta_g$  and  $\gamma_t$ , which are denoted as  $B_g$  and  $\Gamma_t$ . The parameters are estimated in the similar way as in the classical case, SVD is used.

It is assumed that  $\hat{\Gamma}_t$  also follows the random walk model with drift or some of the nonstationary ARIMA models. The construction of the mortality forecasts is the same as in the classical Lee-Carter method.

Another form of model Li and Lee suggested is the combination of the classical Lee-Carter model and the common factor Lee-Carter model, it is called the augmented common factor Lee-Carter (ACFLC) model and has the form

$$m_{gti} = \alpha_{gi} + B_g \Gamma_t + \beta_{gi} \gamma_{ti} + \varepsilon_{gti}. \quad (8)$$

The factor  $\beta_{gi} \gamma_{ti}$  is specific for sex  $i$  and allows for a short-term or medium-term difference between the rate of change in sex  $i$  death rates and that rate of change implied by the common factor. The parameters of model are estimated by SVD.

The common time factor  $\hat{\Gamma}_t$  follows the random walk model with drift or some form of the nonstationary ARIMA model. If the sex specific time factors  $\hat{\gamma}_{ti}$  are nonstationary than the forecasts would be divergent. In the case of their stationarity the mortality forecasts would follow a non-divergent behavior of mortality of both sexes in the past.

## 2 Cointegrated Lee-Carter Method

### 2.1 Long-run relationship of the overall mortality indexes

The problem with the forecasting of both sexes in the same country, mentioned in the part 1.4, can be solved also by another way. It can be hypothesized that the overall mortality index of men  $\hat{\gamma}_{tM}$  and women  $\hat{\gamma}_{tF}$  are in some relationship. As noted in part 1.3, from the empirical analysis it follows that both indexes are the nonstationary time series, which can be generally represented by the ARIMA(., 1, .) models. Therefore, it is suitable to apply the cointegration analysis and to identify the presence of the long-run and the short-run relationships. The cointegration of the overall mortality

indexes  $\hat{\gamma}_{tM}$  and  $\hat{\gamma}_{tF}$  will lead to the error correction model which will be used for their forecasting. The presence of the long-run relationship will tie up both  $\hat{\gamma}_{tM}$  and  $\hat{\gamma}_{tF}$  forecasts and thus also the forecasts of male and female mortalities.

## 2.2 Cointegration

Consider a VAR model of order  $p$

$$\mathbf{X}_t = \phi_1 \mathbf{X}_{t-1} + \cdots + \phi_p \mathbf{X}_{t-p} + \delta \mathbf{D}_t + \mathbf{a}_t, \quad (9)$$

where  $\mathbf{X}_t$  is a  $k$ -dimensional vector of non-stationary I(1) variables,  $\mathbf{D}_t$  is  $d$ -dimensional vector of deterministic variables (drift, trend, etc.), and  $\mathbf{a}_t$  is a vector of white noise innovations. Following the Johansen (1991) cointegration methodology, the VAR model can be transformed into the vector error correction (VEC) model

$$\Delta \mathbf{X}_t = \Pi \mathbf{X}_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta \mathbf{X}_{t-i} + \delta \mathbf{D}_t + \mathbf{a}_t, \quad (10)$$

where  $\Pi = \sum_{i=1}^p \phi_i - \mathbf{I}$ ,  $\Gamma_i = -\sum_{j=i+1}^p \phi_j$ .

In the case of the I(1) time series there can be two situations:

- (A) The coefficient matrix  $\Pi$  has zero rank  $r = 0$ . It means that  $\Pi$  is the zero matrix and the model (10) does not contain non-differenced components. The  $k$ -dimensional time series is generated by the nonstationary vector process. The stationary time series can be obtained by differencing each time series individually.
- (B) The coefficient matrix  $\Pi$  has reduced rank  $r < k$ . Granger's representation theorem asserts that in this case  $k \times r$  matrices  $\alpha$  and  $\beta$  exist, each with rank  $r$  such that  $\Pi = \alpha \beta'$  and  $\beta' \mathbf{X}_t$  is I(0).  $r$  is the number of cointegrating relations and each column of  $\beta$  is the cointegrating vector. The elements of  $\alpha$  are known as the adjustment parameters in the VEC model. In the case of the cointegration of two time series,  $\alpha$  and  $\beta$  are two dimensional vectors and there is only one cointegration relation between time series.

## 3 Application

We apply the cointegrated Lee-Carter (CLC) method to forecast the two-sex overall mortality indexes of four countries - the Czech Republic, Slovakia, Austria and the Netherlands. We have got the mortality data for time period from 1950 to 2008. The source of data are the national statistical offices of the above mentioned countries. We estimate the Lee-Carter model parameters by the LCFIT software. First, we create the "ex post" forecasts, their estimated

standard errors and the "ex post" mean square errors for years 2001-2008. We use the VEC models and for comparison also the one dimensional ARIMA models.

**Table 1.**

	Czech Republic		Slovakia		Austria		Netherlands	
	M	F	M	F	M	F	M	F
ADF	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)
ARIMA	$(0, 1, 0)_c$	$(0, 1, 0)_c$	$(0, 1, 0)_c$	$(0, 1, 0)_c$	$(0, 1, 0)_c$	$(0, 1, 0)_c$	$(1, 1, 0)_c$	$(0, 1, 1)_c$
VAR <sup>2</sup>	3		3		2		3	
Johansen <sup>3</sup>	1		1		1		1	
VEC <sup>4</sup>	III		III		II		III	

Table 1 contains the results of the models creation process of the overall mortality indexes for the period 1950-2000. The ADF unit root test indicates that all time series are of I(1) type. The random walk with drift is the suitable model for the overall mortality indexes in all countries except the Netherlands. The Johansen cointegration test indicates that the overall mortality indexes of men and women are cointegrated in all analyzed countries.

Table 2 contains the estimated standard errors of the forecasts received by the VEC models and the ARIMA models. It can be seen that the VEC models lead to smaller standard errors and therefore to the narrower interval forecasts than the ARIMA models and from this point of view are better. In the last rows the first and second part of Table 2 the forecasts with the actual values are compared by the mean square errors. In the half cases the VEC models produce better "ex post" forecasts than the ARIMA models.

Table 3 shows the results of the models creation process of the overall mortality indexes for the period 1950-2008. The resulting VEC and ARIMA models are used for the creation of forecasts for years 2009-2050. In the majority of cases the estimated forecasts standard errors received from the VEC models are considerably smaller than the standard errors received from the ARIMA models.

The Figure 1 shows the point forecasts of the overall mortality indexes for the Czech Republic and Austria. In the both cases it is seen that the VEC models ties up the forecasts together more than the ARIMA models and from this point of view the VEC forecasts seems to be more realistic.

<sup>2</sup> The number of lags in VAR model.

<sup>3</sup> The number of cointegration relations, Johansen (1991) test:  $LR_{tr}(r/k) = -T \sum_{i=r+1}^k \ln(1 - \lambda_i)$ .

<sup>4</sup> The type of the VEC model, Johansen (1995):

- I.  $\Pi \mathbf{X}_{t-1} + \delta \mathbf{D}_t = \alpha \beta' \mathbf{X}_{t-1}, \delta \mathbf{D}_t = \mathbf{0}$
- II.  $\Pi \mathbf{X}_{t-1} + \delta \mathbf{D}_t = \alpha(\beta' \mathbf{X}_{t-1} + \beta_0), \delta \mathbf{D}_t = \alpha \beta_0$
- III.  $\Pi \mathbf{X}_{t-1} + \delta \mathbf{D}_t = \alpha(\beta' \mathbf{X}_{t-1} + \beta_0) + \alpha_{\perp} \gamma_0, \delta \mathbf{D}_t = \alpha \beta_0 + \alpha_{\perp} \gamma_0$
- IV.  $\Pi \mathbf{X}_{t-1} + \delta \mathbf{D}_t = \alpha(\beta' \mathbf{X}_{t-1} + \beta_0 + \beta_1 t) + \alpha_{\perp} \gamma_0, \delta \mathbf{D}_t = \alpha \beta_0 + \alpha_{\perp} \gamma_0 + \alpha \beta_1 t$
- V.  $\Pi \mathbf{X}_{t-1} + \delta \mathbf{D}_t = \alpha(\beta' \mathbf{X}_{t-1} + \beta_0 + \beta_1 t) + \alpha_{\perp}(\gamma_0 + \gamma_1 t), \delta \mathbf{D}_t = \alpha \beta_0 + \alpha_{\perp} \gamma_0 + (\alpha \beta_1 + \alpha_{\perp} \gamma_1) t$

**Table 2.**

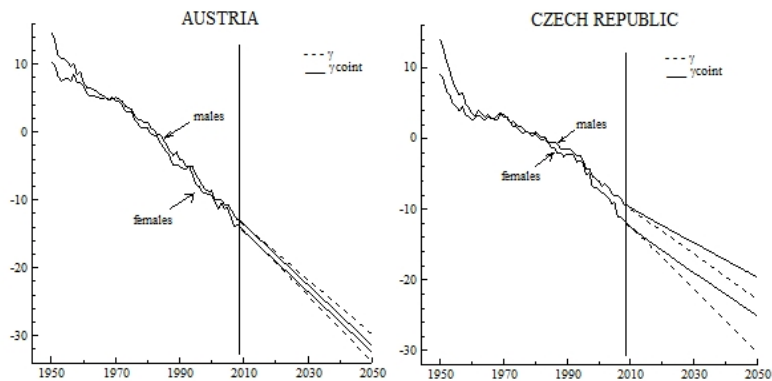
Year	Czech Republic				Slovakia			
	M		F		M		F	
	$SE_{coint}$	SE	$SE_{coint}$	SE	$SE_{coint}$	SE	$SE_{coint}$	SE
2001	0.408	0.495	0.445	0.624	0.276	0.359	0.620	0.832
2002	0.529	0.693	0.559	0.889	0.364	0.505	0.703	1.221
2003	0.683	0.862	0.702	1.132	0.457	0.636	0.818	1.491
2004	0.799	0.970	0.809	1.321	0.530	0.729	0.904	1.687
2005	0.900	1.099	0.929	1.464	0.584	0.821	0.989	1.876
2006	1.024	1.213	1.050	1.582	0.647	0.881	1.023	2.036
2007	1.124	1.319	1.169	1.721	0.704	0.952	1.089	2.224
2008	1.218	1.396	1.280	1.847	0.768	1.007	1.159	2.423
MSE	0.284	0.262	0.537	0.522	0.140	0.126	1.880	2.103

Year	Austria				Netherlands			
	M		F		M		F	
	$SE_{coint}$	SE	$SE_{coint}$	SE	$SE_{coint}$	SE	$SE_{coint}$	SE
2001	0.519	0.517	0.554	0.594	0.276	0.359	0.620	0.832
2002	0.731	0.730	0.711	0.820	0.364	0.505	0.703	1.221
2003	0.839	0.930	0.826	1.000	0.457	0.636	0.818	1.491
2004	0.994	1.078	0.935	1.147	0.530	0.729	0.904	1.687
2005	1.144	1.206	1.045	1.279	0.584	0.821	0.989	1.876
2006	1.234	1.318	1.111	1.415	0.647	0.881	1.023	2.036
2007	1.340	1.383	1.191	1.531	0.704	0.952	1.089	2.224
2008	1.473	1.490	1.309	1.650	0.768	1.007	1.159	2.423
MSE	0.221	0.466	0.379	0.473	2.088	2.475	0.932	0.505

**Table 3.**

	Czech Republic		Slovakia		Austria		Netherlands	
	M	F	M	F	M	F	M	F
ADF	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)
ARIMA	$(0, 1, 0)_c$	$(0, 1, 0)_c$	$(0, 1, 0)_c$	$(0, 1, 0)_c$	$(0, 1, 0)_c$	$(0, 1, 0)_c$	$(1, 1, 0)_c$	$(0, 1, 1)_c$
VAR	3		3		3		5	
Johansen	1		1		1		1	
VEC	III		IV		III		IV	

**Fig. 1.**

## 4 Conclusion

The classical Lee-Carter method assumes that the overall mortality index is represented by some ARIMA model (its special case is the random walk with drift). In the case of both sexes the overall mortality index there can be problem that the forecasts diverge increasingly over time which is unrealistic. Li and Lee (2005) suggested to solving this by the ACFLC method. In this paper we suggest the concurrent method, namely the CLC method. We applied this method to the forecasting of the overall mortality indexes of some European countries. In the comparison with the classical Lee-Carter method we received more tied up forecasts with smaller estimated standard errors. We have not statistically compared the CLC method with the ACFLC method, yet. But from the computational point of view the CLC method is simpler. These are the main arguments for deeper analysis of the CLC method properties and for their comparison with the ACFLC method properties we intend to do in our future research.

## References

- ARLT, J. and ARLTOVÁ, M. (2009): *Ekonomické časové řady*. Professional Publishing, Prague.
- BROUHNS, N., DENUIT, M. and VERMUNT, K.J. (2002): A Poisson Log-bilinear Regression Approach to the Construction of Projected Lifetables. *Insurance: Mathematics and Economics* 31, 373-393.
- CARTER, L.R., and LEE, R.D. (1992): Modelling and Forecasting U.S. Sex Differentials in Mortality. *International Journal of Forecasting* 8, 393-411.
- DEATON, A. and PAXSON, Ch. (2004): *Mortality, Income, and Income Inequality Over Time in the Britain and the United States*. Technical Report 8534 National Bureau of Economic Research Cambridge, MA. <http://www.nber.org/papers/w8534>.
- ENGLE, R.F. and GRANGER, C.W.J. (1987): Cointegration and Error Correction: Representation, Estimation and Testing. *Econometrica* 55, 251-276.
- GIROSI, F. and KING, G. (2007): *Understanding the Lee-Carter Mortality Forecasting Method*. Technical Report, Rand Corporation. <http://gking.harvard.edu/files/abs/lc-abs.shtml>
- JOHANSEN, S. (1991): Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica* 59, 1551-80.
- JOHANSEN, S. (1995): *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*, Oxford, Oxford University Press.
- LEE, R.D. (2000): The Lee-Carter Method for Forecasting Mortality, with Various Extensions and Applications. *North American Actuarial Journal* 4 (1), 80-93.
- LEE, R.D. and CARTER, L. (1992): Modeling and Forecasting the Time Series of U. S. Mortality. *Journal of the American Statistical Association* 87, 659-671.
- LI, N. and LEE, R. (2005): Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography* 42(3): 575-594.
- WILMOTH, J. (1993): *Computational Methods for Fitting and Extrapolating the Lee-Carter Model of Mortality Change*. Technical report Department of Demography, University of California, Berkeley.



# Empirical analysis of the climatic and social-economic factors influence on the suicide development in the Czech Republic<sup>\*</sup>

Markéta Arltová<sup>1</sup>, Jitka Langhamrová<sup>2</sup>, and Jana Langhamrová<sup>3</sup>

<sup>1</sup> Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic, *arltova@vse.cz*

<sup>2</sup> Department of Demography, Faculty of Informatics and Statistics, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic, *langhamj@vse.cz*

<sup>3</sup> Student of Statistics, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic, *xlanj18@vse.cz*

**Abstract.** The suicide rate is closely linked with social-economic and climatic factors. For the empirical analysis of this influence we made use of time series of the level of unemployment, average temperatures, the average duration of sunlight in hours and total precipitation. The analysis was carried out on monthly time series in the period from January 1999 to December 2007. The estimated model shows the influence of the duration of sunlight and the level of unemployment on the development of the number of suicides. From the viewpoint of the influence of weather on the suicide rate it was demonstrated that with the increase in hours of sunlight there is also an increase in the number of suicides. Weather clearly acts as a "trigger mechanism". Further mediating factors, such as, for example, psychic aspects and mental illness, must also be taken into account.

**Keywords:** suicide, climatic factors, social-economic factors, time series

## 1 Development of the number of suicides in the Czech Republic

The development of the number of suicides is closely connected with the social, economic and political situation and with current historical events on the territory of the Czech Republic (Fig. 1).

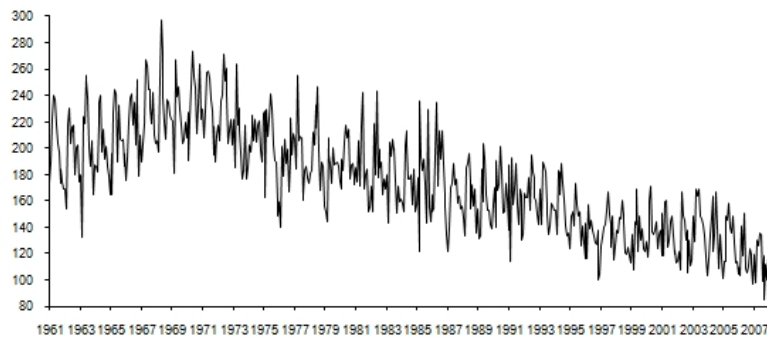
In our analysis we wish to concentrate on the factors that may influence suicidal behaviour in a certain manner. One of the factors that are expected to influence the number of suicides significantly is the time of year. From the monthly time series of the number of suicides in the individual months of the years 1961-2007 (Fig. 2) the seasonal nature of the time series is evident at first glance.

---

<sup>\*</sup> This paper was written with the support of Grant Agency of the Czech Republic No. 402/09/0369 "Modelling of Demographic Time Series in Czech Republic".



**Fig. 1.** Numbers of suicides in the Czech Republic in the years 1875-2008 (*Data source: Czech Statistical Office*)



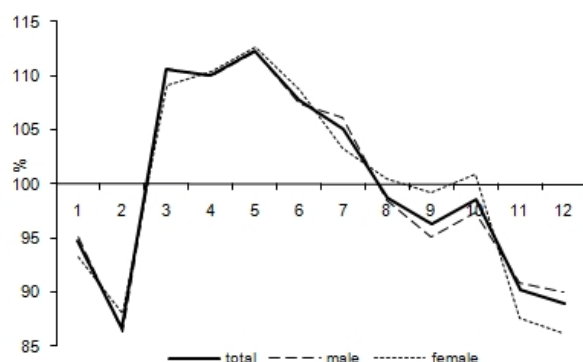
**Fig. 2.** Monthly development of the number of suicides in the Czech Republic in the years 1961-2007 (*Data source: Czech Statistical Office*)

In accordance with Yip, Chao, Chiu (2000) the highest values for the suicide rate in the Czech Republic are achieved in the spring months (Fig. 3). This can be explained in connection with climatic and social-economic factors (for instance the duration of sunlight, temperature, the phase of the moon or unemployment, social standing, marital status, etc.).

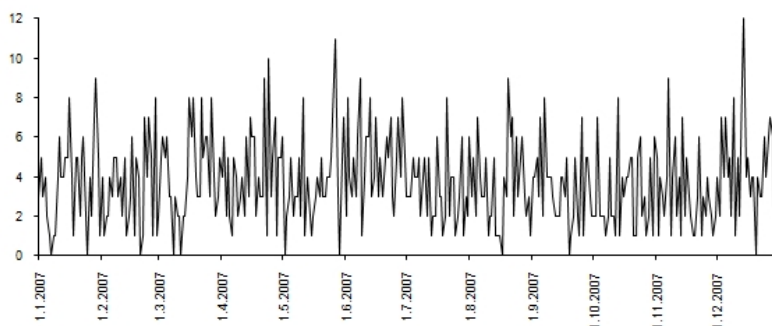
The period with above-average numbers of suicides in the years 1961-2007 was from March to July (in the case of women up to October), with a maximum in May, then below average from August to February, with the minimum in February (for women in December).

Also an interesting aspect is the distribution of suicides in the individual days of the week. For this analysis total daily data were used (without gender classification) from January 1, 1992 to December 31, 2007. In Fig. 4, due to the great amount of data, only the year 2007 is drawn in by way of illustration.

From Fig. 5 it can be seen that on average the least suicides fall on Tuesday, Wednesday and Sunday, with the minimum on Tuesday (4.42) and the



**Fig. 3.** Seasonal indexes of the number of suicides according to months in the years 1961-2007 (*Source: own calculations*)

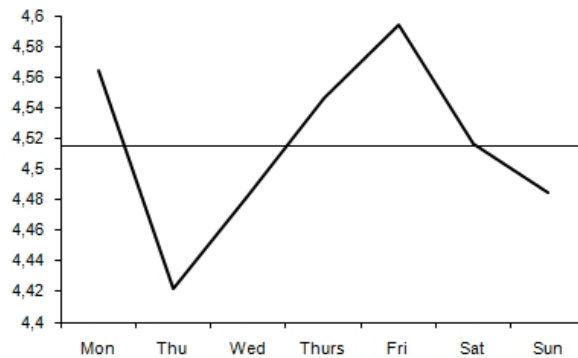


**Fig. 4.** Daily development of the number of suicides in the Czech Republic in 2007 (*Data source: Czech Statistical Office*)

most are on Monday, Thursday and Friday, when there is also the maximum (4.59). However the difference in the average daily numbers of suicides can be described as negligible.

## 2 The influence of climatic and social-economic factors on the development of the suicide rate in the Czech Republic

The idea of the influence of climatic factors on the suicide rate is based on the so-called bio-meteorological hypothesis (for example Zollner, Moller, Jensen, 2003). Temperature and changes in it should, in this theory, have a direct influence on suicidal tendencies. Warmth increases the sensitivity of the nervous system and causes an excess of energy in the organism, which is not consumed



**Fig. 5.** Development of the average number of suicides according to days in the week in the years 1992-2007 (*Source: own calculations*)

in the natural manner and is expended in other ways. The result is that in the summer period the overall activity of the organism increases, which may also appear as suicidal behaviour. Later in this theory further influences of climate were also taken into account, such as the length of sunlight, barometric pressure (this has an influence on the stability of the human psyche), solar activity, atmospheric humidity and the amount of rainfall (Krejčíková, 2009).

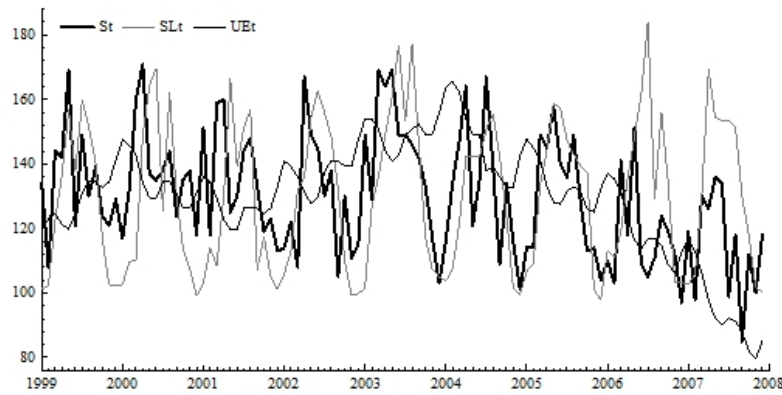
The suicide rate is, then, closely linked with social-economic and climatic factors. For the empirical analysis of this influence we used time series of the level of unemployment in % ( $UE$ ), average temperatures in  $^{\circ}C$  ( $T$ ), the average length of sunlight in hours ( $SL$ ) and the total precipitation in millimetres ( $P$ ). The analysis was carried out on monthly time series in the period 1/1999–12/2007. All the given time series contain a marked and regular seasonal component.

With regard to the selected time series we can expect that the influence of climatic factors and the level of unemployment will have an exogenous character and it is not therefore necessary to construct a VAR model; for the analysis it suffices to use a single-equation model (Arlt, Arltová, 2009).

For the estimate of the parameters of the model the econometric program EViews 6 was used. The resultant model was obtained after gradual elimination of time series with statistically insignificant estimates of parameters. Fig. 4 depicts time series of the average duration of sunlight and the level of unemployment in which influence on the number of suicides was demonstrated by the model in the given period (the time series are graphically converted to the same level). No influence of average temperature and total rainfall was demonstrated.

The estimated model in the form

$$S_t = 60.163 + 0.103SL_t + 6.263UE_t$$



**Fig. 6.** Development of the number of suicides, average values of sunlight and level of unemployment in the Czech Republic in the period 1/1999–12/2007 (*Data source: Czech Statistical Office, Czech Hydrometeorological Institute and Ministry of Labor and Social Affairs Czech Republic*)

**Table 1.**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	60.16322	13.68069	4.397674	0.0000
UE	6.262754	1.483807	4.220733	0.0001
SL	0.102982	0.018739	5.495522	0.0000
R-squared	0.290327	Mean dependent var		130.1574
Adjusted R-squared	0.276810	S.D. dependent var		19.13365
S.E. of regression	16.27137	Akaike info criterion		8.444076
Sum squared resid	27799.53	Schwarz criterion		8.518579
Log likelihood	-452.9801	Hannan-Quinn criter.		8.474284
F-statistic	21.47777	Durbin-Watson stat		1.746762
Prob(F-statistic)	0.000000			
Breusch-Godfrey Serial Corr. LM Test	1.600222	Prob. F(2,103)	0.2068	
Normality Test: Jarque-Bera	1.127098	Prob	0.5692	
Heterosced. Test: Breusch-Pagan-Godfrey	0.467497	Prob F(2,105)	0.6279	

demonstrates the influence of the length of sunlight and level of unemployment on the development of the number of suicides. It explains the 27,7% of variability (measured by the index of determination) of the number of suicides. Diagnostic tests showed that although the non-systematic component of the model has normal distribution (Jarque-Bera test=1.127 [0.569]) and is homoscedastic (Breusch-Pagan-Godfrey test=0.467 [0.628]), it is seasonally autocorrelated.

As we are not able to explain the seasonal aspects in the time series of the number of series by means of the model, a seasonal zero-one dummy variable ( $D_t$ ) was inserted in it.

After conversion the following model was obtained

$$S_t = 65.44 + 0.08SL_t + 6.12UE_t - 10.8D_2 + 14.6D_3 - 17.2D_4 - 10.2D_9 - 11.6D_{12}.$$

**Table 2.**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	65.44269	12.20377	5.362496	0.0000
UE	6.121026	1.327841	4.609759	0.0000
SL	0.076174	0.019059	3.996762	0.0001
D2	-10.85323	5.437236	-1.996094	0.0486
D3	14.60091	5.202987	2.806255	0.0060
D4	17.25236	5.139188	3.357021	0.0011
D9	-10.18461	5.107201	-1.994167	0.0489
D12	-11.61538	5.634958	-2.061308	0.0419
R-squared	0.475696	Mean dependent var		130.1574
Adjusted R-squared	0.438995	S.D. dependent var		19.13365
S.E. of regression	14.33115	Akaike info criterion		8.233936
Sum squared resid	20538.20	Schwarz criterion		8.432612
Log likelihood	-436.6325	Hannan-Quinn criter.		8.314492
F-statistic	12.96130	Durbin-Watson stat		1.890657
Prob(F-statistic)	0.000000			
Breusch-Godfrey Serial Corr. LM Test	1.458703	Prob. F(2,98)	0.2375	
Normality Test: Jarque-Bera	0.226108	Prob	0.8931	
Heterosced. Test: Breusch-Pagan-Godfrey	0.858967	Prob F(2,100)	0.5410	

It is evident that the parameters of the model did not change significantly with the inclusion of the dummy variables. Thanks to the seasonal dummy variables the model now explains better the variability of the number of suicides (43.899%), and in addition the diagnostic tests confirmed that the non-systematic component of the model is not autocorrelated (Breusch-Godfrey LM test=1.459 [0.237]), has normal distribution (Jarque-Bera test=0.226 [0.893]) and is homoscedastic (Breusch-Pagan-Godfrey test=0.859 [0.541]).

The explanation of roughly half of the variability the numbers of suicides represents an acceptable result for us. It is evident that it is not possible to capture all the factors, which influence the psyche of the individual, and especially it is impossible to give precise figures. We can therefore state that the influence of climatic and also of one of the social-economic factors on suicidal behaviour has been demonstrated. The remaining influence, not clarified by the model, may have its origin in biological, psychological or also other climatic factors.

### 3 Conclusion

On the basis of measurable factors we attempted to explain whether any of them can significantly influence the suicide rate; we concentrated on unemployment and the influence of weather. Unemployed persons account on average for 11% of all suicides; the influence of unemployment was also demonstrated in our model. From the viewpoint of the influence of weather on the suicide rate it was demonstrated that with the increase in hours of sunlight there is also an increase in suicides. Weather is clearly shown here as a "trigger mechanism", in addition to which other intermediating factors, such as psychic aspects and mental illness, must be taken into account.

## References

- ARLT, J. and ARLTOVÁ, M. (2009): *Ekonomické časové řady*. Professional Publishing, Prague.
- KREJČÍKOVÁ, J. (2009): *Analýza počtu sebevražd v České republice*. Thesis, University of Economics, Prague.
- POLÁŠEK, V. (2006): *Sebevraždy v České republice - 2001 až 2005*. Czech Statistical Office. <http://www.czso.cz/csu/2006edicniplan.nsf/p/4012-06>.
- YIP, P.S.F., CHAO, A. and CHIU, C.W.F. (2000): Seasonal variation in suicides: diminished or vanished. Experience from England and Wales, 1982–1996. *British Journal of Psychiatry*, 177, 366–369.
- ZOLLNER, L., MOLLER, S. and JENSEN B.F. (2003): Meteorological factors and seasonality in suicidal behaviour in Denmark 1970–2000. *Working papers*, Centre for Suicide Research, Odense, Denmark. <http://www.selvmordsforskning.dk/filecache/15948/1140704605/swingmeteorologicalfactorsandseasonality.pdf>
- Source of data:
- Czech Hydrometeorological Institute: <http://www.chmi.cz/meteo/ok/infklim.html>
- Czech Statistical Office: <http://www.czso.cz>
- Ministry of Labor and Social Affairs: <http://portal.mpsv.cz/sz/stat>





# Yield Curve Predictability, Regimes, and Macroeconomic Information: A Data-Driven Approach

Francesco Audrino<sup>1</sup> and Kameliya Filipova<sup>2</sup>

<sup>1</sup> Institute of Mathematics and Statistics, University of St. Gallen,  
Bodanstrasse 6, CH-9000 St. Gallen, Switzerland. *francesco.audrino@unisg.ch*

<sup>2</sup> Institute of Mathematics and Statistics, University of St. Gallen,  
Bodanstrasse 6, CH-9000 St. Gallen, Switzerland. *kameliya.filipova@unisg.ch*

**Abstract.** We propose an empirical approach to determine the various economic sources driving the U.S. yield curve. We allow the conditional dynamics of the yield at different maturities to change in reaction to past information coming from several relevant predictor variables. We consider both endogenous, yield curve factors and exogenous, macroeconomic factors as predictors in our model, letting the data themselves choose the most important variables. We find clear, different economic patterns in the local dynamics and regime specification of the yields depending on the maturity. Moreover, we present strong empirical evidence for the accuracy of the model in fitting in-sample and predicting out-of-sample the yield curve.

**Keywords:** yield curve modeling and forecasting, macroeconomic variables, tree-structured models, threshold regimes, bagging

## 1 Introduction

Beginning with Ang and Piazzesi (2003) the idea of incorporating macroeconomic variables on the top of yield curve factors for modeling bond yields plays a major role in today's term structure literature, giving raise to a new so called macro-finance modeling framework. Important contributions in that area include, for example, Dewachter and Lyrio (2006), Hoerdahl et al. (2008), Joslin et al. (2009), and Duffee (2006). Despite the various macro-finance modeling strategies proposed in the last years for the U.S. term structure of interest rates dynamics, several questions and controversies are still open. Most of the open issues in the recent macro-finance literature evolve around the central theme of how yields are associated with macro variables.

Various studies in the term structure literature (see, for example, Bansal et al. (2004), Dai et al. (2007), Audrino and De Giorgi (2007)) document that the regime-switching models better describe the nonlinearities in the yields' drift and the volatility found in the historical interest rate data.

In this paper we present a methodology to build and estimate a discrete-time regime-switching model for the term structure dynamics over time in

which for every maturity we are able to identify or infer, in a purely data-driven way, the most important macroeconomic and latent variables driving both the local dynamics and the regime shifts. As such, our modeling approach offers a clear interpretation and regime specification. Our basic framework for the yield curve is a macro-factor model, yet not the usual no-arbitrage factor representation typically used in the macro-finance literature. The methodology adopted in this paper is mainly motivated by Audrino's (2006) tree-structured model for the short rate. Our models belong to the class of threshold regimes models, where regimes are characterized by some threshold for the relevant predictor variables. In contrast to the existing term structure literature, the conditional mean dynamics is not exogenously given, but determined endogenously and is subjected to regime shifts depending on past values of certain relevant predictor variables.

Applying our model to U.S. interest rate data we draw a number of conclusions. We find clear, different economic patterns in the local dynamics and regime specification of the yields depending on the maturity. In addition, we conclude that our framework is consistent with the key stylized facts of the yield curve behavior. Finally, in order to improve the prediction accuracy of our model, we use bootstrap aggregating (bagging). We compare the out-of-sample forecasting ability of our model to that of several strong competitor models. Using the superior predictive ability (SPA) test of Hansen (2005), we find that such improvements are in most cases statistically significant.

## 2 The Model

To infer the yield curve behavior, we use a model with four distinctive features. First, to capture the cross-sectional dynamics of the yield curve, we employ two latent term structure factors often used in the finance literature, interpreted as level and slope. The two factors usually account for about 95% of the cross-sectional variation of yields. Second, we allow heteroscedasticity in the error term. Third, motivated by the interpretability and the improved forecasting performance of the macro-finance literature in comparison to the pure finance approach, we incorporate macroeconomic variables. Fourth, our model accommodates regime-switching behavior but still allows interpretation and clear endogenous regime specification.

### 2.1 The yield-macro model: specification

Let  $Y_t = (y(t, n_1), \dots, y(t, n_T))'$  be a  $T$ -dimensional vector of yields with maturities  $n_1, \dots, n_T$  observed at time  $t$  and let  $\Delta y(t, n_\tau) \equiv y(t, n_\tau) - y(t-1, n_\tau)$  denote the first difference of yields at time  $t$  with maturity  $n_\tau$ . Further, let us assume the following model for the term structure dynamics

$$\Delta y(t, n_\tau) = \mu_{t, n_\tau} + \varepsilon_{t, n_\tau}, \quad \tau = 1, \dots, T, \quad (1)$$

where  $\mu_{t,n_\tau} \equiv \mu(\Phi_{t-1,n_\tau}; \psi_{n_\tau})$  is a parametric function representing the conditional mean and  $\varepsilon_{t,n_\tau}$  is the error term of the yields' returns with maturity  $n_\tau$ . More formally,  $\varepsilon_{t,n_\tau}$  can be decomposed as  $\varepsilon_{t,n_\tau} = \sqrt{h(\Phi_{t-1,n_\tau}; \psi_{n_\tau})} z_t$ , where  $(z_t)_{t \in \mathbb{Z}}$  is a sequence of i.i.d. random variables with zero mean and unit variance, and where  $h(\Phi_{t-1,n_\tau}; \psi_{n_\tau})$  is the time-varying conditional variance. Above we denoted by  $\Phi_{t,n_\tau}$  all the relevant conditional information up to time  $t$  for maturity  $n_\tau$ . In our application (see Section 3),  $\Phi_{t,n_\tau}$  corresponds to a large number of term structure and macroeconomic variables.

In practice changes in monetary policy or macroeconomic shocks may cause interest rates to behave quite differently in different time periods, in terms of both level and volatility. An adequate characterization of this stylized fact requires building a term structure model with regime shifts. Rather than following the common Markovian regime-switching approach, here, similar to the standard classification and regression trees (CART) procedure, we impose regimes to be characterized by rectangular partition cells with edges determined by thresholds on the predictor variables. Such partition cells are practically constructed using a binary tree. The rectangular partition has several major advantages: First, it allows a clear interpretation of regimes in terms of relevant predictor variables. Second, it enables model estimation also in high-dimensional predictor spaces.

Third, it enables us to determine the current regime based solely on the realization of the state variables, macroeconomic variables, and the threshold structure. In particular, the conditional dynamics of the yields is given by:

$$\begin{aligned} \mu_{t,n_\tau} &= \sum_{j=1}^{K_{n_\tau}} (\alpha_{0,n_\tau}^j + \alpha_{1,n_\tau}^j \Delta y(t-1, n_\tau) + \beta_{n_\tau}^{j'} \mathbf{x}_{t-1} + \gamma_{n_\tau}^{j'} \mathbf{x}_{t-1}^{\text{ex}}) I_{[\Phi_{t-1,n_\tau} \in \mathcal{R}_{n_\tau}^j]}, \\ h_{t,n_\tau} &= \sum_{j=1}^{K_{n_\tau}} (\omega_{n_\tau}^j + a_{n_\tau}^j \epsilon_{t-1,n_\tau} + b_{n_\tau}^j h_{t-1,n_\tau}) I_{[\Phi_{t-1,n_\tau} \in \mathcal{R}_{n_\tau}^j]}, \end{aligned}$$

where  $\psi_{n_\tau} = (\alpha_{0,n_\tau}^j, \alpha_{1,n_\tau}^j, \beta_{n_\tau}^{j'}, \gamma_{n_\tau}^{j'}, \omega_{n_\tau}^j, a_{n_\tau}^j, b_{n_\tau}^j, j = 1, \dots, K_{n_\tau})$  is a parameter vector, which parameterizes the local dynamics in the different regimes,  $K_{n_\tau}$  is the number of regimes for maturity  $n_\tau$  (estimated from the data),  $I(\cdot)$  is the indicator function and  $\mathcal{R}_{n_\tau}^j$  represents a region of the partition  $\mathcal{P}_{n_\tau} = \{\mathcal{R}_{n_\tau}^1, \dots, \mathcal{R}_{n_\tau}^{K_{n_\tau}}\}$  of the predictor space  $G_{n_\tau}$  of the relevant conditional information  $\Phi_{t,n_\tau} = \{\Delta y(t, n_\tau), \mathbf{x}_t', \mathbf{x}_t^{\text{ex}'}\}$  such that

$$G_{n_\tau} = \bigcup_{j=1}^{K_{n_\tau}} \mathcal{R}_{n_\tau}^j, \quad \mathcal{R}_{n_\tau}^i \cap_{(i \neq j)} \mathcal{R}_{n_\tau}^j = \emptyset \quad \tau = 1, \dots, T.$$

Above we denoted by  $(\Delta y(t, n_\tau), \mathbf{x}_t')$  and by  $\mathbf{x}_t^{\text{ex}'}$  all the endogenous and all the exogenous information, respectively, available at time  $t$ .

## 2.2 Model estimation

In order to obtain an estimate for the parameters  $\psi$  we employ a two-step procedure. At Step 1, for each maturity, we select the optimal linear dynamics. Then, at Step 2, we allow for more complex, threshold structure. Note that our framework is flexible enough and does not rely on the assumption that the variables driving the optimal mean dynamics are necessary the ones determining the most relevant threshold structure.

**Step 1: Best subset selection (BSS)** One of the main questions in the term structure literature is how many and which yield curve factors and/or macro variables are needed to build a good model for the time series dynamics of yields? Do these factors always have the same impact on the yields with different maturities? A simple way to answer these questions is to use BSS.

To select the optimal linear dynamics, we create a pool of predictors and let the data themselves choose the most informative ones. This is achieved by finding for each possible number of variables the subset of the corresponding size that gives the smallest residual sum of squares. Here the optimal number of predictors is selected using Bayesian Information Criterion (BIC).

**Step 2: Regime specification** For any fixed sequence of partition cells our model can be estimated using quasi-maximum likelihood

$$-l(\psi_{n_\tau}; \Phi_{n_\tau} | \frac{n}{2}) = - \sum_{t=2}^n \log \left( \frac{1}{\sqrt{h_{t,n_\tau}(\psi_{n_\tau})}} p_Z \left( \frac{(\Delta y(t, n_\tau) - \mu_{t,n_\tau}(\psi_{n_\tau}))}{\sqrt{h_{t,n_\tau}(\psi_{n_\tau})}} \right) \right),$$

where  $p_Z(\cdot)$  is the density function of the distribution of the innovation  $z_t$ . The choice of the optimal partition cells (i.e., splitting variables and threshold values) involves a model choice procedure for non-nested hypotheses. As already discussed in Audrino and Bühlmann (2001), for general tree-structured models, the model selection of the optimal splitting variables and threshold values can be performed via a tree-structured partial search. Within any data-determined tree structure the optimal model is selected using BIC. For exact description and further applications of the algorithm we refer to Audrino and Bühlmann (2001), Audrino (2006), and Audrino and Trojani (2006).

## 2.3 Improving the forecasting ability: Bagging

To improve the prediction accuracy of our model, we use bagging. Bagging is a machine learning technique aimed at reducing the variance and thus improving the forecasting performance of unstable estimators such as trees. Bagging involves the following steps (see Inoue and Kilian (2004)): (i) Build a matrix, where the first column corresponds to our response variable and the next columns include all the potential predictors and construct bootstrap samples by randomly drawing with replacement blocks of rows from the matrix; (ii) for each bootstrap sample apply the two-step procedure described

in Section 2.2. Using the optimal parameters, compute the conditional mean of the yields; (iii) Average the forecasts of the conditional mean.

### 3 Empirical Results

The *term structure data* consist of one-month U.S. Treasury bills with eight different maturities: 3 and 6 months and 1, 2, 3, 5, 7, and 10 years taken from the Fama-Bliss files in the CRSP database. Since almost all the cross-sectional term structure information can be summarized in just a few variables, we build the endogenous predictors in the following way: we define the level as the 10-year yield and the slope as the difference between the longest (10-year) and the shortest (3-month) maturity in our sample.

*Macroeconomic data* including some of the leading U.S. indices of inflation (consumer price index of finished goods (CPI), producer price index of finished goods (PPI)), and real activity (the index of Help Wanted Advertising in Newspapers (HELP), unemployment (UE), the growth rate of industrial production (IP)) are available from the Datastream. To exploit additional macroeconomic information, we construct our measures of conditional volatility and variance of the macro indices by using 24-month rolling window technique and squaring the different indices, respectively.

We use the data between January 1961 and December 2001 as the in-sample period, whereas the remaining data from January 2002 to June 2005 are left to evaluate the out-of-sample forecasts of the different models.

#### 3.1 What is driving the Yield Curve Predictability?

Based on the observed patterns (see Table 1) the results can be summarized by three groups: short-, mid- and long-term maturities. Like the monetary policy rules documented in the macroeconomic literature (see for example Taylor (1993)), the short rate local dynamics is mainly driven by inflation, real activity, and an autoregressive component. It is not a surprise that the regimes here are linked to the level of inflation, since in times of high inflation, the Federal Reserve tends to raise the short end of the yield curve to provide economic stability. The mid-term maturities follow an autoregressive process (AR(1)-GARCH(1,1)), whose behavior is determined by the term structure slope and the level of real activity. The long rates capture strong macroeconomic effects. Here the volatility of inflation plays a major role in the threshold structure as well as in the local mean dynamics.

#### 3.2 Stylized facts

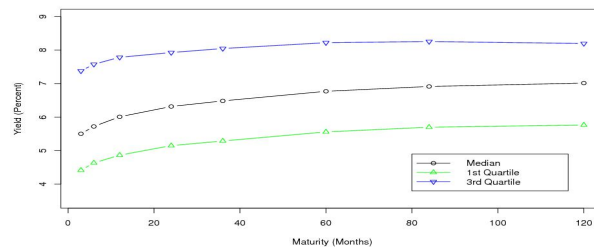
To show the goodness of our model, we illustrate our model's ability to replicate the most important stylized facts found in the term structure literature. The average upward-sloping form, the concavity, as well as the fact that short rates are more volatile than long rates are illustrated in Figure 1.

PANEL A: BEST SUBSET SELECTION								
Maturity ( $n_\tau$ )	$\Delta y_{n_\tau}$	slope	level	PPI	HELP	HELP.sq	vol.PPI	vol.CPI
3M	*	*	*	*		*	*	
6M	*	*	*	*		*	*	
1Y	*							
2Y	*							
3Y	*							
5Y			*		*		*	*
7Y			*		*		*	*
10Y			*		*		*	*

PANEL B: OPTIMAL REGIME STRUCTURE		
Maturity	Optimal Regime Structure	# Regimes
short-term maturities (3M and 6M)	$CPI_{t-1} \leq 3.5316$ $CPI_{t-1} > 3.5316$	2
mid-term maturities (1Y, 2Y and 3Y)	$HELP_{t-1} \leq 61.82$ $HELP_{t-1} > 61.82$ and $slope_{t-1} \leq -0.0662$ $HELP_{t-1} > 61.82$ and $slope_{t-1} > -0.0662$	3
long-term maturities (5Y and 10Y)	$volatilityPPI_{t-1} \leq 0.5935$ $volatilityPPI_{t-1} > 0.5935$	2

**Table 1.** BSS results (Panel A) and optimal regime structure (Panel B) found for every maturity. The variables we take into consideration are the following: the yield's first difference for maturity  $n_\tau$ ,  $\tau = 1, \dots, 8$  denoted by  $\Delta y_{n_\tau}$ , yield curve's level, yield curve's slope, the macroeconomic indices CPI, PPI, HELP, IP, UE, the square of the macroeconomic indices CPI.sq, PPI.sq, HELP.sq, IP.sq, UE.sq, and their conditional volatility vol.CPI, vol.PPI, vol.HELP, vol.IP, vol.UE.

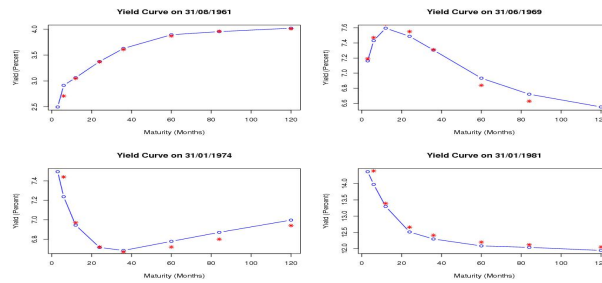


**Fig. 1.** Median fitted yield curve with interquartile range.

Figure 2 shows the ability of our model to replicate the broad variety of shapes the actual yield curve takes through time: upward-sloping, downward-sloping, humped, and inverted-humped.

### 3.3 Forecasting results

Apart from the economic linkage and the ability to replicate the most important stylized facts, a good term structure model should also be able to



**Fig. 2.** Fitted yield curves for selected dates (dotted lines), together with actual yields (stars).

provide a close out-of-sample fit. Here, we compare the out-of-sample performance of our model (**Macro Tree**) to those of several strong competitors (*i*) Random walk (RW); (*ii*) Nelson-Siegel proposed by Diebold and Li (2006) (NS AR(1)); (*iii*) Markovian regime switching model of Gray (1996) (Gray's RS); (*iv*) tree structured regime switching model of Audrino (2006) (Audrino Tree); and (*v*) the one regime version of our model (**Best Subset**).

PANEL A: OUT-OF-SAMPLE MEAN SQUARED ERROR (MSE)

	Macro Tree	Best Subset	NS AR(1)	RW	Audrino Tree	Gray's RS
3M	0.012 (0.101)	11.553 (0)	0.277 (0)	0.015 (0.247)	0.017 (0)	0.029 (0.006)
6M	0.021 (0.094)	12.995 (0)	0.184 (0)	0.017 (0.401)	0.036 (0)	0.050 (0)
1Y	0.035 (0.568)	0.032 (0.567)	0.085 (0.142)	0.035 (0.203)	0.039 (0.084)	0.549 (0)
2Y	0.094 (0.107)	0.090 (0.343)	0.128 (0.363)	0.090 (0.587)	0.091 (0.436)	0.143 (0.018)
3Y	0.117 (0.420)	0.118 (0.434)	0.164 (0.452)	0.124 (0.440)	0.117 (0.585)	0.148 (0.037)
5Y	0.134 (0.237)	0.229 (0.012)	0.215 (0.172)	0.135 (0.513)	0.127 (0.641)	0.122 (0.622)
7Y	0.128 (0.412)	0.129 (0.412)	0.280 (0)	0.122 (0.556)	0.126 (0.373)	0.144 (0.107)
10Y	0.101 (0.444)	0.115 (0.045)	0.099 (0.520)	0.103 (0.365)	0.097 (0.609)	0.121 (0)

PANEL B: OUT-OF-SAMPLE MSE FOR THE BAGGED MODELS

	Macro Tree	Best Subset	NS AR(1)	Audrino Tree	Gray's RS
3M	0.007 (0.595)	0.082 (0)	0.578 (0)	0.013 (0.126)	0.144 (0)
6M	0.009 (0.526)	0.036 (0.012)	0.432 (0)	0.019 (0.023)	0.080 (0.004)
1Y	0.028 (0.537)	0.065 (0)	0.242 (0)	0.036 (0.512)	0.375 (0)
2Y	0.082 (0.642)	0.091 (0.284)	0.085 (0.591)	0.088 (0.450)	0.311 (0)
3Y	0.115 (0.667)	0.145 (0.484)	0.155 (0.368)	0.114 (0.652)	0.294 (0)
5Y	0.124 (0.657)	0.143 (0.074)	0.168 (0.027)	0.123 (0.695)	0.261 (0)
7Y	0.116 (0.684)	0.116 (0.684)	0.411 (0)	0.112 (0.672)	0.271 (0.105)
10Y	0.092 (0.629)	0.010 (0.234)	0.123 (0.047)	0.095 (0.460)	0.209 (0)

**Table 2.** Results of out-of-sample 1-month-ahead forecasting using six models and their bagged versions, as described in detail in the text. The results are based on the out-of-sample period January 2002 - June 2006, for a total of 42 observations.

Without considering bagging (Table 2, Panel A), we find that our model has overall good performance at all eight maturities. Matters improve dramat-

ically, once we apply bagging (Table 2, Panel B). The SPA p-values [Hansen (2005)] based on all eleven model specifications presented in parenthesis in Table 2 reveal that the forecasts yield from the bagged versions of our model are significantly better than almost all of the alternative approaches.

## References

- ANG, A. and PIAZZESI, M. (2003): No-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary Economics*, 745-787.
- AUDRINO, F. (2006): Tree-structured multiple regime in interest rates. *Journal of Business and Economic Statistics*, 24 (3) 338-353.
- AUDRINO, F. and BÜHLMANN, P. (2001): Tree-structured GARCH models. *Journal of the Royal Statistical Society, Series B* 63, 727-744.
- AUDRINO, F. and DE GIORGI, E. (2007): Beta regimes for the yield curve. *Journal of Financial Econometrics* 5, 456-490.
- AUDRINO, F. and TROJANI, F. (2006): Estimating and predicting multivariate volatility regimes in global stock markets. *Journal of Applied Econometrics* 21 (3), 345-369.
- BANSAL, R. and TAUCHEN, G. and ZHOU, H. (2004): Risk premiums in the term structure and the business cycle. *Journal of Business and Economic Statistics* 22, 396-409.
- DAI, Q. and SINGLETON, K. and YANG, W. (2007): Regime shifts in a dynamic term structure model of U.S. Treasury bond yields. *Review of Financial Studies* 20, 1669-1706.
- DEWACHTER, H. and LYRIO, M. (2006): A structural macro model of the yield curve. *Computing in Economics and Finance* 236.
- DIEBOLD, F. and LI, C. (2006): Forecasting the term structure of government bond yields. *Journal of Econometrics* 130, 337-364.
- DUFFEE, G. (2006): Term structure estimation without using latent factors. *Journal of Financial Economics* 79, 507-536.
- GRAY, S. (1996): Modeling the conditional distribution of interest rates in a regime-switching process. *Journal of Financial Economics* 42, 27-62.
- HANSEN, P. R. (2005): A test for superior predictive ability. *Journal of Business & Economic Statistics* 23, 365-380.
- HÖRDAHL, P., TRISTANI, O., and VESTIN, D. (2008): The yield curve and macroeconomic dynamics. *Economic Journal, Royal Economic Society*, 118(533), 1937-1970.
- INOUE, A., and KILIAN, L. (2004): Bagging time series models. *CEPR Discussion Papers*, 4333.
- JOSLIN, S., PRIEBSCHE, M., and SINGLETON, K. (2009): Risk-premium accounting in macro- dynamic term structure models. *Working Paper*.
- LITTERMAN, R., and SCHEINKMAN J. (1991): Common factors affecting bond returns. *Journal of Fixed Income* 1, 51-61.
- TAYLOR, J. (1993): Discretion versus policy rules in practice. *Journal of Econometrics* 39, 195-214.



# Socioeconomic Factors in Circulatory System Mortality in Europe: A Multilevel Analysis of Twenty Countries

Sara Balduzzi, Lucio Balzani, Matteo Di Maso,  
Chiara Lambertini, and Elena Toschi

Faculty of Statistics

Via delle Belle Arti 41, Bologna, Italy, E-mail: *lucio.balzani@libero.it*

**Abstract.** This paper is the result of teamwork carried out during an advanced course of Health Statistics. Our aim was to apply the multilevel models learned on the course to some concrete research problems without using any fiscal resources. We used the WHO health databases to compare Standardized Death Ratios (SDR) due to circulatory system diseases in twenty representative European countries between 1992 and 2003 and to explain the role played by socio-economic and lifestyle indicators using a multilevel approach (years nested in countries). This paper underlines the power of research based on free online institutional databases, especially for health policy makers who often require accurate but expensive health information.

**Keywords:** multilevel models, circulatory system mortality, European Health database

## 1 Introduction

The following paper is the result of teamwork that emerged from an advanced course of Health Statistics at the University of Bologna. After a broad introduction to multilevel models, five students wished to apply this well-known statistical approach to a concrete problem in order to write a scientific paper for an international public health journal. Several possible topics were discussed and, in the end, mortality for circulatory system diseases in Europe emerged as the most interesting, both for statistical implications and practical importance. In fact health policy players are always in need of more information on the performance of health care systems (Nolte et al.(2005)). They are especially interested in the impact of socio-economic, lifestyle and healthcare indicators on the reduction of morbidity and mortality (Levi et al. (2002); Petrelli et al. (2006)). Although death due to circulatory system diseases is now declining in the majority of European countries, the continuing high incidence of this kind of disease still demands special attention be given to it. In our work, we used the European Health for All Database (HFA-DB) to compare Standardised Death Ratios (SDR) due to Circulatory System Diseases (CSD) between several European countries in different

years. We should highlight that these databases also contain indicators on demographic and socio-economic status, lifestyles, environment, health care resources, health care utilization and the expenditure of each country. Studies that utilize these available databases can be found in the literature, but they only evaluate trends of cause-specific mortality or avoidable mortality without considering factors such as those listed above. Furthermore, they are restricted to trends until 2000 (Treurniet et al. (2004), Evstifeeva et al. (1997)). Socio-economic development and health data show important interactions: health sector expenditure is an important driver of economic development, which in turn influences social circumstances, lifestyle and health. Health care resource statistics reflect numerous changes in health care systems, with the number of hospital beds declining in all considered countries (World Health Organization (2002)). We report the results of an analytical comparison of several macro indicators collected routinely from institutional sources in Europe. The main purpose of this study was to assess the association between circulatory system disease-related mortality trends and socio-economic and healthcare indicators among several countries between 1992 and 2003 using European databases. Multilevel modelling approach is appropriate to this data because it accommodates the multilevel structure of the data. It allows us to estimate, for the first time, the role of indicators collected from institutional sources as potential explanatory variables of mortality.

## 2 Material and methods

### 2.1 Data

SDR for circulatory system diseases and the healthcare, lifestyle and socio-economic indicators were derived from the Health for All (HFA) Statistical Database, WHO Regional Office for Europe, Copenhagen, Denmark, August 2009 version, integrated with health data published by EUROSTAT and OECD. To make comparisons between countries as valid as possible, data for each indicator has been taken from one common international source. Twenty countries in Europe, divided into five representative groups, have been chosen: Italy, Portugal and Macedonia (Southern Europe); Iceland, Sweden, Denmark and Finland (Northern Europe); France, the Netherlands, Switzerland and Austria (Central Europe); Slovakia, Hungary, the Czech Republic and Slovenia (Eastern Europe); Belarus, Ukraine, Azerbaijan, Kazakhstan and Uzbekistan (Former Soviet Republics). The choice of the countries was based on geographical position and the data availability for the outcome variable (SDR for circulatory system diseases) while the chosen time span (1992-2003) depended only on this last condition. Another twenty-four socioeconomic and lifestyle indicators taken from previous studies and articles were considered as possible explanatory factors of our outcome variable. Considering the particular nature of the database the determination of the variable level has been resulted full of difficulties. Each variable could have been considered as

a level-1 but a lot of them were evaluated as a level-2 for the presence of missing data and the low temporal variability.

So, to determine whether an indicator should be considered as a level-1 or a level-2 factor, we checked its internal structure. A high variance over time was the main characteristic of a level-1 variable. Otherwise we considered it a level-2 variable using its median or the value of the year with the widest data availability for all countries. A detailed list of indicators, each one associated to its level, is reported in Table 2.

## 2.2 Statistical Analysis

Before modeling our variables, we studied their structure in terms of mean, standard deviation, minimum and maximum. Then, line charts were built to study time trends for selected countries. A multilevel linear regression was used to model the relationship between circulatory system SDR and lifestyle, health expenditure, social and economic indicators. In fact, the repeated nature of panel data (across-country over a range of years) can be viewed in a hierarchical framework with annual observations (level-1) nested in higher level (level-2) groups (countries). A multilevel approach has particular appeal as it puts the emphasis on the estimation of the distribution of higher (country) level effects and their interpretation. We used this model as we thought there might be both a general Europe-wide trend and individual differences between countries in circulatory system mortality over a period of twelve years (1992-2003). Two-level linear regression models (years nested within countries) were built to investigate the impact of national socio-economic information on our outcome variable. Firstly, we calculated the null model to check the Variance Partition Coefficient (VPC) that showed us that 96% of the total variance was explained by the two-level structure of the dataset. Then we inserted the year, appropriately recoded for a better interpretation of the coefficient. After that, we put the level-1 variables in the model, one by one. The same forward procedure was used for level-2 variables. Then, to determine the model that best fitted our dataset, we used either AIC or BIC (Snijders and Bosker (1999)). Finally, we generated several possible interactions to check their eventual influence on our model. During all these steps we constantly monitored the VPC that decreased from 0.96 in the null model to 0.19 in the final model. The intercept was introduced as a random effect and each factor was treated as fixed effect. The final model, with a total number of 240 observations, can be written as follows (Hox (2002)):

$$Y_{ijk} = \gamma_{000} + \gamma_{p0}X_{ijp} + \gamma_{0q}Z_{jq} + \gamma_{pq}X_{ijp}Z_{jq} + u_{0j} + e_{ij} \quad (1)$$

where  $\gamma_{000}$  is the overall mean of our outcome variable in the final model,  $\gamma_{p0}$  is the increase in SDR for a unitary increase of the  $p$ -th first level variable,  $\gamma_{0q}$  is the increase in SDR for a unitary increase of the  $q$ -th second level

variable,  $\gamma_{pq}$  is the cross-level interaction coefficient,  $u_{0j}$  is the difference between the general mean and the  $j$ -th country specific mean,  $e_{ij}$  is the difference between the  $j$ -th country mean and the  $i$ -th year observation. All the predictor variables were centered on their grand-mean, thus the intercept term is the mean outcome in country  $j$ , adjusted for differences between years in the means of various characteristics. All estimations were done with the software **STATA 10.0**. For the identification of the best model and for a complete residuals analysis we used the **gllamm** command (Rabe-Hesketh and Skrondal (2001)).

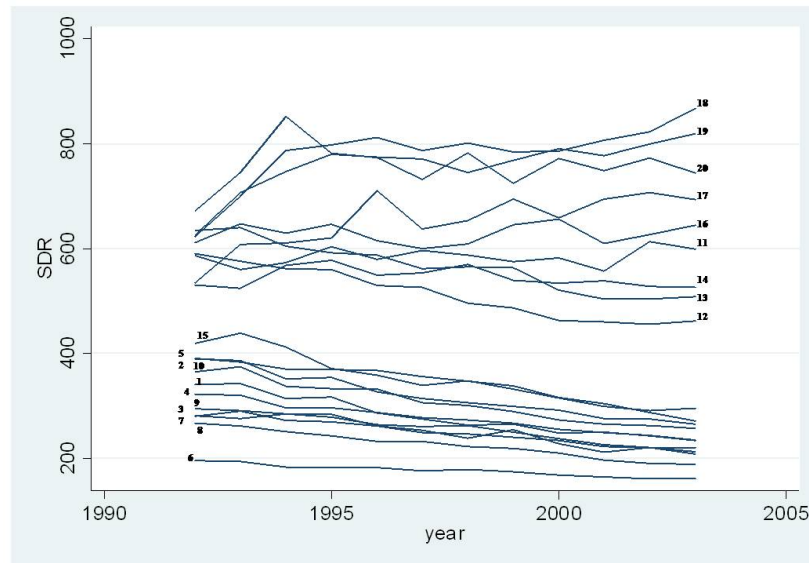
### 3 Results

Table 1 shows the values of SDR in different countries from end to end of the study (World Health Organisation (1992)).

COUNTRIES	MEAN(SD)	MIN-MAX	COUNTRIES	MEAN(SD)	MIN-MAX
<b>NORTHERN EUROPE</b>			<b>11.Macedonia</b>	584.49 (17.04)	557.26-613.57
<b>1.Denmark</b>	281.38 (38.29)	234.01-342.79	<b>EASTERN EUROPE</b>		
<b>2.Finland</b>	319.39 (42.52)	264.92-389.77	<b>12.Czech Rep.</b>	513.85 (49.66)	455.98-590.26
<b>3.Iceland</b>	249.91 (28.58)	208.06-284.59	<b>13.Hungary</b>	565.65 (48.68)	503.84-640.48
<b>4.Sweden</b>	276.51 (28.62)	234.38-321.47	<b>14.Slovakia</b>	544.97 (18.54)	524.11-577.89
<b>CENTRAL EUROPE</b>			<b>15.Slovenia</b>	351.42 (50.34)	290.52-439.06
<b>5.Austria</b>	341.45 (38.98)	270.66-390.33	<b>FORMER SOVIET REPUBLICS</b>		
<b>6.France</b>	176.31 (11.74)	160.53-195.14	<b>16.Azerbaijan</b>	628.46 (19.14)	599.81-656.26
<b>7.Netherlands</b>	249.55 (25.15)	212.07-289.70	<b>17.Belarus</b>	651.95 (52.47)	535.40-710.27
<b>8.Switzerland</b>	225.64 (26.56)	187.98-266.36	<b>18.Kazakhstan</b>	781.43 (62.8)	622.88-867.90
<b>SOUTHERN EUROPE</b>			<b>19.Ukraine</b>	758.94 (50.8)	625.68-819.67
<b>9.Italy</b>	257.15 (26.84)	219.91-293.92	<b>20.Uzbekistan</b>	758.59 (43.02)	672.22-852.55
<b>10.Portugal</b>	307.55 (40.36)	256.32-374.22			

**Table 1.** Mean (standard deviation) and range of SDR for diseases of circulatory system per 100,000 in twenty European countries.

In the selected countries as a whole, age-standardised mortality ratios fall from a mean value of 447.63 in 1992 to 420.51 in 2003, but diverging trends can be found between countries. Trends in separate countries are given in Figure 1. Three different time-trend groups of countries can be identified: in Western and Central European countries (Austria, Denmark, Finland, France, Iceland, Italy, the Netherlands, Portugal, Slovenia, Sweden and Switzerland) the age-standardized mortality ratio starts from lower values than those of the Eastern-European countries and continuously decreases between 1992 and 2003; the Eastern-European countries are divided into a decreasing mortality group (the Czech Republic, Hungary and Slovakia) and an increasing mortality group (Kazakhstan, Ukraine, Uzbekistan, TFYR Macedonia, Belarus, Azerbaijan) even if the last two countries, after an early period of increase, show a final period with no clear change in mortality.



**Fig. 1.** SDR for diseases of circulatory system per 100,000 between 1992 and 2003 in twenty European countries (the number near the line indicates the country code, as in Table 1).

A quite strong variability regarding all the indicators both in the time series and among the twenty countries can be observed. Within the first area of Table 2, identified as socio-economic, GDP per capita (US\$) is the most changeable indicator: it ranges from 389 (Uzbekistan) to 38,213 (Switzerland). The second area is related to lifestyle indicators: among them regular daily smokers (%) run from 12.50 to 36.30, while health expenditure per capita (US\$) which runs from 131.5 to 3,595 is of interest in the third area (healthcare). The selected indicators present relevant differences between Eastern and Western countries. In Eastern countries a lower investment in expenditure for health is evident. Other variables which have different mean values in the geographical areas are GDP per capita, the percentage of energy available from fat and the percentage of the male population over 65 years, with higher values in Western countries. The percentage of energy available from protein presents a much higher value in Iceland, whereas in the other countries it does not show a particular geographic pattern.

Table 3 summarizes the full random intercept multilevel model for SDR for circulatory system diseases which includes all the statistically significant covariates. The hierarchical model that only expresses variability among countries and years showed a strong geographic variability ( $VPC=0.97$ ) which is reduced after controlling for socio-economic indicators. In fact, the variance partition coefficient for the full model is 0.19. The model shows a general decrease in mortality: SDR declined at a rate of 5.09 for each year.

HFA CODE	VARIABLES	LEVEL	MEAN (SD)	MIN-MAX
<b>1320</b>	SDR, diseases of circulatory system (x100,000)	-	441.23(200.85)	160.53 - 867.90
<b>0031</b>	Male population aged 65+ years(%)	1	10.28(3.17)	2.87 - 16.27
<b>0260</b>	Gross Domestic Product per capita (\$)	2	15,118.8(12,626.78)	389 - 38,213
<b>2370</b>	Diabetes prevalence(%)	2	2.60(1.76)	0.08 - 6.37
<b>3210</b>	Total energy available from fat (%)	1	32.62(6.95)	13.88 - 42.16
<b>3210</b>	Total energy available from protein (%)	1	12.18(1.07)	10.09 - 15.79
<b>3010</b>	Regular daily smokers, age 15+(%)	2	27.31(6.16)	12.50 - 36.30
<b>5010</b>	Hospitals per 100,000	1	4.44(2.81)	0.87 - 10.73
<b>5010</b>	Hospital beds per 100,000	1	749.8(236.72)	363.69 - 1,346.19
<b>5290</b>	General Practitioners per 100,000	1	65.6(42.21)	9.47 - 167.12
<b>6720</b>	Total health expenditure (PPP\$ per capita)	2	1,595.13(1,108.97)	131.50 - 3,595
<b>6730</b>	Public sector health expenditure as % of total health expenditure	2	69.55(16.62)	18.55 - 90.05
<b>6770</b>	Total pharmaceutical expenditure as % of total health expenditure	2	16.08(6.96)	2.8 - 34
<b>6850</b>	Public sector expenditure on health as % of total government expenditure	2	12.35(3.49)	3.8 - 18.35

**Table 2.** List, level, mean (standard deviation) and range of variables used in the final model.

VARIABLES	COEF.(SE)	p-value
<i>First level:</i>		
<b>Year</b>	-5.10 (0.84)	<b>&lt;0.001</b>
<b>% of population aged 65+ years, male</b>	7.08 (2.41)	<b>0.003</b>
<b>% of total energy available from fat</b>	2.35 (1.28)	0.068
<b>% of total energy available from protein</b>	-22.20 (4.05)	<b>&lt;0.001</b>
<b>Hospitals</b>	-2.19 (9.12)	0.810
<b>Hospital beds</b>	0.20 (0.04)	<b>&lt;0.001</b>
<b>General practitioners</b>	-1.01 (0.15)	<b>&lt;0.001</b>
<i>Second level:</i>		
<b>Gross Domestic Product, (US\$)</b>	-0.01 (0.002)	<b>&lt;0.001</b>
<b>Diabetes prevalence, in %</b>	10.70 (4.91)	<b>0.029</b>
<b>% of regular daily smokers in the population, age 15+</b>	4.13 (0.77)	<b>&lt;0.001</b>
<b>Total health expenditure per capita</b>	-0.31 (0.05)	<b>&lt;0.001</b>
<b>Public sector health expenditure as % of total health expenditure</b>	3.09 (0.67)	<b>&lt;0.001</b>
<b>Total pharmaceutical expenditure as % of total health expenditure</b>	-13.89 (0.97)	<b>&lt;0.001</b>
<b>Public sector expenditure on health as % of total government expenditure</b>	-34.66 (5.41)	<b>&lt;0.001</b>
<i>Interactions:</i>		
<b>Public sector expenditure on health as % of total government expenditure x Total health expenditure per capita</b>	0.01 (0.003)	<b>&lt;0.001</b>
<b>Hospitals x Hospital beds</b>	-0.03 (0.01)	<b>&lt;0.001</b>
<b>Hospitals x Gross Domestic Product per capita</b>	0.001 (0.0002)	<b>&lt;0.001</b>

**Table 3.** Results of multilevel linear regression model between circulatory system disease mortality rates and socioeconomic, lifestyle and healthcare variables.

Our model confirms present knowledge of risk factors for circulatory diseases, showing a positive association between the SDR and the percentage of the male population over 65 years, the percentage of smokers in the popula-

tion, the prevalence of diabetes and the percentage of energy available from fat. In contrast, the percentage of energy available from protein is negatively associated with SDR. The model also gives us information about expenditure: pharmaceutical investment and a large number of general practitioners are important in reducing mortality for circulatory diseases. When other health expenditures are fixed, an increase in the percentage of public expenditure on the total health expenditure is associated with an increase in mortality: this could be explained by a lower quality use of public money than private, because of the restrictions which public money is often subjected to. We found a statistically significant interaction between the percentage of public health expenditure on the total of government expenditure and health expenditure per capita: the increase in the total health expenditure per capita is always a protective factor for mortality due to circulatory system diseases, whereas an increase of the part of government expenditure devoted to health is protective only for countries with total health expenditure per capita lower than 2,450 PPP \$ (Azerbaijan, Belarus, Czech Republic, Finland, Hungary, Italy, Kazakhstan, Portugal, Slovakia, Slovenia, TFYR Macedonia, Ukraine and Uzbekistan). The interaction between Gross Domestic Product per capita and the number of hospitals per 100,000 shows that countries with a higher value of GDP per capita are more protected from mortality from circulatory diseases only if the number of hospitals per 100,000 does not exceed 7: remembering that risk factors of cardiovascular diseases are strongly associated with welfare society lifestyles, we interpreted this result as meaning that those countries with higher GDP per capita are more protected from circulatory diseases only if they are poor in health services, because a higher GDP gives them the possibility of an improvement of services, but for those countries with an effective health service, higher GDP means more risk factors. Finally, the interaction between the number of hospitals per 100,000 and the number of hospital beds per 100,000 tells us that, when the other variables are fixed, an increase in the number of beds is protective only if the hospitals are in a satisfactory number. The residuals for our model were tested for consistency assumption of normal distribution and homoscedasticity.

## 4 Conclusion

In Europe there is a great variation in mortality time series for circulatory system diseases: the period between 1992 and 2003 shows a constant decline in Eastern, Central, Southern and Northern countries and an increase in Former Soviet Republics. According to our results, the change in the SDR for circulatory system diseases in Europe was due to various factors: an aging population, a high prevalence of diabetes, a high proportion of smokers, and a high fat diet are positively related to mortality while health and pharmaceutical expenditure and a high number of general practitioners are negatively related to it. The Former Soviet Republics have the highest values of SDR for

circulatory system diseases as a consequence of the previous factors and of the low GDP per capita. It is possible that a new health policy and economic development may improve the present situation in those countries.

We can conclude that these results (easily obtained from free online databases) confirm previous studies and papers. This model could be applied for health policy projects, often requiring accurate but expensive health information. Finally, it's important to underline the didactical importance of this work emerging from, as specified in the introduction, an advanced course of Health Statistics at the University of Bologna. For all five students, this experience has been far more interesting than an ordinary approach to the subject. So, mainly in advanced courses attended by students with a solid statistical background, it can be worth focusing more on the applications of well-known analysis techniques in order to expose students to concrete research problems and to teach them how to write a scientific paper. This approach, which deals with aspects of research usually ignored by traditional university courses, implies a greater effort from both students and professors but the results are certainly more satisfying.

## References

- EVSTIFEEVA, T.V., MACFARLANE, G.J., ROBERTSON, C. (1997): Trends in cancer mortality in Central European countries. *European Journal of Public Health* 7(2), 169-176.
- HOX, J. (2002): *Multilevel Analysis*. LEA.
- LEVI, F., LUCCHINI, F., NEGRI, E. and LA VECCHIA, C. (2002): Trends in mortality from cardiovascular and cerebrovascular disease in Europe and other areas of the world. *Heart* 88, 119-124.
- NOLTE, E., MCKEE, M. and GILMORE, A. (2005): Morbidity and Mortality in the transition countries of Europe. In: M. Macura, A. L. MacDonald and W. Haug (Eds.): *The New Demographic Regime Population Challenges and Policy Responses*. United Nations New York and Geneva, 153-176 (Chapter 9).
- PETRELLI, A., GNAVI, R., MARINACCI, C., COSTA, G. (2006): Socioeconomic inequalities in coronary heart disease in Italy: A multilevel population-based study. *Social Science & Medicine* 63, 446-456.
- RABE-HESKETH, S., PICKLES, A. and SKRONDAL, A. (2001): *GLLAMM Manual*. Department of Biostatistics and Computing, Institute of Psychiatry, King's College, University of London. Available at: <http://www.iop.kcl.ac.uk/iop/departments/biocomp/programs/gllamm.html>.
- SNIJDERS, T. and BOSKER, R. (1999): *Multilevel Analysis*. SAGE.
- TREURNIET, H.F., BOSHUIZEN, H.C., HARTELOH, P.P.M. (2004): Avoidable mortality in Europe (1980-1997): a comparison of trends. *J Epidemiol Community Health* 58, 290-295.
- WORLD HEALTH ORGANISATION (1992): *Health for All database*. Geneva: World Health Organisation.
- WORLD HEALTH ORGANISATION (1992): Health status overview for countries of Central and Eastern Europe that are candidates for accession to the European Union. *WHO*, 2002.



# Comparing ORF Length in DNA Code Observed in Sixteen Yeast Chromosomes

Anna Bartkowiak<sup>1 2</sup> and Adam Szustalewicz<sup>1</sup>

<sup>1</sup> Institute of Computer Science, University of Wrocław, Joliot Curie 15, 50-383, Wrocław, PL *aba@ii.uni.wroc.pl, asz@ii.uni.wroc.pl*

<sup>2</sup> Wrocław High School of Applied Informatics, Wejherowska 28, 54-239, Wrocław, PL

**Abstract.** We consider the distribution of ORF lengths (counted in amino-acids) observed in 16 yeast chromosomes. Using the asymptotic Likelihood Ratio (LR) test we investigate whether the ORF lengths may be modelled by one common Negative Binomial (NB) distribution. Apart from the formal asymptotic LR test, we construct confidence ellipses for the NB parameters and we perform also some simulation experiments supporting the thesis on the common NB distribution.

**Keywords:** DNA code, amino-acids, ORF length, negative binomial, parameter estimation, confidence ellipses

## 1 Introduction

The yeast genome has sixteen chromosomes containing together over 12 millions bases (basic nucleic acids A, C, G, T). The full sequencing in various layouts is available at <http://www.yeastgenome.org>. The genetic code providing templates for genes is inscribed into pieces of chromosomes called ORFs (Open Reading Frames). The code is composed from triplets of the four basic nucleotides; the triplets designate 20 amino-acids, see: Christianini and Hahn (2007), Bartkowiak (2008), Bartkowiak and Szustalewicz, (2009).

Our aim is to investigate the length of the ORFs, coded in amino-acids. There are altogether  $n = 6686$  ORFs inscribed into the sixteen yeast chromosomes. Each ORF contains the code (a kind of template) for one vital function of the organism, some of the functions are yet not exactly known. An excerpt of the amino-acid code in one ORF is shown in Bartkowiak and Szustalewicz (2009). Let  $X$  be the random variable (r.v.) denoting the length of an ORF when the length is counted in amino-acids. The distribution of  $X$  for ORFs found in four chromosomes was investigated in Bartkowiak (2008), also in Bartkowiak and Szustalewicz (2009). It was stated that  $X$  follows the negative binomial (NB) distribution. Now we consider all the sixteen chromosomes with together  $n = 6686$  ORFs. Firstly, we consider the NB distribution fitted to each chromosome separately. It appears that for all of them the hypothesis on the NB distribution can be sustained. Our next question is: may

the ORF length in all chromosomes be described by a common NB distribution? The answer is positive. This is interesting, because the parameters might constitute a kind of meta-characteristic of a species.

## 2 The negative binomial (NB) distribution and its parameters

The NB distribution may be derived from several probabilistic models and its parametrization may be different, see e.g. Fisz (1963), Durbin et al. (1999), Hilbe (2007), Bartkowiak and Szustalewicz (2009). In the following we use the parametrization from Matlab Stats Toolbox (2002). A discrete random variable  $X$  is said to have the negative binomial distribution, if its probability mass function (pmf) is given as

$$P(X = x) = f(x; r, p) = \binom{x+r-1}{x} p^r (1-p)^k \quad (1)$$

for  $x = 0, 1, 2, \dots$ , with real parameters  $r > 0$  and  $0 < p < 1$ . The parameter  $p$  characterizes deviation from the Poisson distribution. The variance  $\sigma^2$  of the NB is greater than its mean  $\mu$ , which is referred to as *overdispersion*.

For given empirical data, the parameters  $r$  and  $p$  of the NB distribution are most effectively estimated by the Maximum Likelihood (ML) method, which yields also asymptotic confidence intervals for the respective parameters. The values of the NB estimates for ORF length distributions found in yeast genome data are shown in Table 1.

The ML method finds the estimates by minimizing the negative likelihood, which is in the NB case a function of two unknown arguments:  $r$  and  $p$ . In the NB case there does not exist a direct solution and the minimum has to be evaluated iteratively. The iterations start usually with the moment estimates which are given by the formula ( $X$  below stands for a NB r.v.)

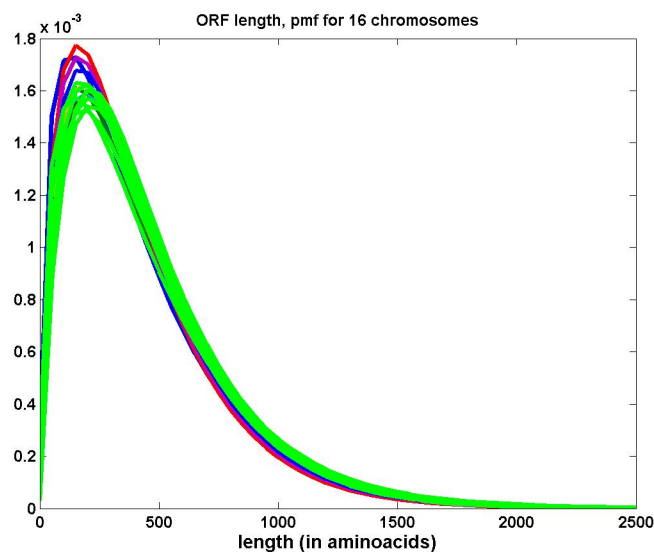
$$E\{X\} \equiv \mu = r \frac{1-p}{p}, \quad E\{(X - E(X))^2\} \equiv \sigma^2 = r \frac{1-p}{p^2}. \quad (2)$$

Using these estimates, the MATLAB function `nbinfit` obtains the ML estimates  $\hat{r}, \hat{p}$  in the following way: using the moment expression for  $\mu$  given in formula (2) and substituting there  $\mu := \bar{x}$ , the negative *Loglikelihood* is expressed as a function of one unknown argument only (of  $r$ ). Then a 1D search is applied for finding the minimum with respect to  $r$ . Let  $\hat{r}$  be the found minimum. With known  $\hat{r}$ , the ML estimate of  $p$  is obtained by the back substitution  $\hat{p} := \hat{r}/(\mu + \hat{r})$ .

The pmf's for the investigated NB distributions with parameters estimated from the 16 yeast chromosomes (see Table 1) are shown in Figure 1. Notice that there is a small difference in the pics of the curves, yet their shape is very similar.

chromosome	n	$r$	conf. int.	$p$	conf. interval
1	121	1.444	(0.726, 2.161)	0.00343	(0.00168, 0.00519)
2	462	1.551	(1.253, 1.849)	0.00345	(0.00278, 0.00411)
3	187	1.633	(1.069, 2.196)	0.00400	(0.00275, 0.00524)
4	855	1.738	(1.489, 1.986)	0.00381	(0.00326, 0.00437)
5	328	1.628	(1.203, 2.053)	0.00389	(0.00290, 0.00487)
6	143	1.695	(1.078, 2.312)	0.00380	(0.00248, 0.00511)
7	593	1.810	(1.483, 2.137)	0.00397	(0.00326, 0.00468)
8	325	1.664	(1.313, 2.016)	0.00387	(0.00311, 0.00463)
9	251	1.633	(1.094, 2.171)	0.00369	(0.00253, 0.00486)
10	404	1.626	(1.268, 1.983)	0.00343	(0.00265, 0.00420)
11	348	1.853	(1.475, 2.231)	0.00392	(0.00317, 0.00467)
12	588	1.462	(1.216, 1.709)	0.00316	(0.00267, 0.00366)
13	513	1.797	(1.458, 2.136)	0.00388	(0.00309, 0.00466)
14	439	1.901	(1.531, 2.272)	0.00417	(0.00334, 0.00500)
15	607	1.748	(1.447, 2.049)	0.00391	(0.00321, 0.00462)
16	522	1.900	(1.533, 2.266)	0.00416	(0.00337, 0.00496)
Grand Total	6686	1.699	(1.611, 1.786)	0.00376	(0.00357, 0.00395)

**Table 1.** Values of the NB parameters  $r$  and  $p$  estimated in 16 yeast chromosomes and in the Grand Total set, together with their confidence intervals constructed at the  $\alpha = 0.05/17$  confidence level.



**Fig. 1.** NB probability mass functions (pmf's) for parameters estimated from ORF lengths found in sixteen yeast chromosomes.

### 3 Constructing confidence intervals and confidence ellipses

To construct confidence intervals and confidence ellipses we need the covariance matrix of the estimates. It is obtained in the following way:

Let  $\boldsymbol{\theta}$  denote the vector of the parameters appearing in the likelihood function  $L(\boldsymbol{\theta})$ . In the NB case  $\boldsymbol{\theta} = (\theta_1, \theta_2)$ , with  $\theta_1 = r$ , and  $\theta_2 = p$ . Let  $\hat{\boldsymbol{\theta}}$  denote the ML estimate of  $\boldsymbol{\theta}$ . It is known that under some regularity conditions the sample distribution of the ML estimate  $\hat{\boldsymbol{\theta}}$  converges in probability (with sample size  $\rightarrow \infty$ ) to the normal distribution

$$\hat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}, \mathcal{I}^{-1}(\boldsymbol{\theta})), \quad (3)$$

where  $\mathcal{I}^{-1}(\boldsymbol{\theta})$  denotes the the inverse *Fisher information matrix* (see e.g. McLachlan and Peel (2000), pp. 41-42). It is common to approximate – for large samples – the Fisher information matrix  $\mathcal{I}(\boldsymbol{\theta})$  by the *observed information matrix*  $\mathbf{I}(\hat{\boldsymbol{\theta}})$  obtained as (see Stuart et al. (1999), p. 74, eq. 18.64)

$$\mathbf{I}(\hat{\boldsymbol{\theta}}) = -\partial^2 \log L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \quad (4)$$

In such a case, the covariance matrix of the estimates  $\hat{\boldsymbol{\theta}}$  appearing in formula (3) is approximated by  $\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})$ . Considering one single parameter, we may wish to construct for it a confidence interval; taking into account two parameters, we may wish to construct for them a confidence ellipse.

**Confidence intervals** For normally distributed  $\hat{\theta}_j$  ( $j = 1, 2$ ), using formulae (3) and (4), we obtain (approximately) the standard error (SE) of  $\hat{\theta}_j = (\hat{\boldsymbol{\theta}})_j$  as the square root of the  $(j, j)$ th element of the covariance matrix  $\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})$ :

$$s_j = SE(\hat{\theta}_j) \approx [(\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}))_{jj}]^{1/2}. \quad (5)$$

The confidence interval for  $\theta_j$  at the confidence level  $\beta = 1 - \alpha$  is then

$$\hat{\theta}_j - z_{\alpha/2} s_j < \theta_j < \hat{\theta}_j + z_{\alpha/2} s_j, \quad (6)$$

with  $z_{\alpha/2}$  denoting the upper  $1 - \alpha/2$  quantile in the normal  $N(0, 1)$  distribution. For a single statistical inference, usually  $\alpha = 0.05$ ,  $\alpha = 0.01$ , or  $\alpha = 0.001$  are assumed. It may be shown that the confidence interval defined in formula (6) covers the true value of  $\theta$  with probability  $\beta = 1 - \alpha$ .

**Confidence ellipses** Let  $\mathbf{x}$  be multivariate normal  $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Then, see e.g. Mardia et al. (1979), the probability density function (pdf) of  $\mathbf{x}$  has constant densities on ellipsoids or ellipses designated by the quadratic form

$$U = (\mathbf{x} - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})^T = c^2, \quad (7)$$

where  $c^2$  is a given positive constant. Moreover, it is known (see e.g. Mardia et al., 1979, Theorem 2.5.2) that the quadratic form  $U$  has the  $\chi_f^2$  distribution with  $f$  being the rank of  $\boldsymbol{\Sigma}$ . Thus  $U \sim \chi_f^2$ . This was proven for known  $\boldsymbol{\mu}$

and  $\Sigma$ . The formula holds asymptotically after substituting into (7) instead of the unknown parameters their ML estimates.

In our case, taking into account formula (3), and substituting into (7)  $\hat{\theta} - \mu = \mathbf{x} - \mu$ ,  $\Sigma^{-1} = (\mathbf{I}^{-1}(\hat{\theta}))^{-1}$ , we obtain the quadratic form  $\ddot{U} = (\theta - \hat{\theta}) \mathbf{I}(\hat{\theta})(\theta - \hat{\theta})^T$  which is also (asymptotically) distributed like the  $\chi_f^2$  variable.

Let  $\chi_f^2(\alpha)$  denote the upper  $\alpha$  quantile of the  $\chi_f^2$  distribution. Then, the equation

$$\ddot{U} = (\theta - \hat{\theta}) \mathbf{I}(\hat{\theta})(\theta - \hat{\theta})^T = \chi_f^2(\alpha) \quad (8)$$

designates a confidence ellipsoid for  $\theta$ . For fixed  $\alpha$ , the ellipsoid contains with probability  $\beta = (1 - \alpha)$  the 'true' value of the unknown parameters  $\theta$ .

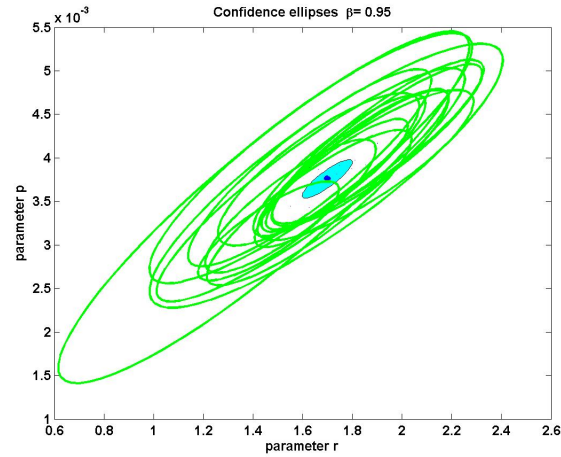
In the bi-variate case we obtain appropriate confidence ellipses located in the  $(\theta_1, \theta_2)$  plane. Examples of such confidence ellipses, constructed with parameters from individual chromosomes, are shown in Figure 2.

**Simultaneous confidence regions** When making several statements, e.g., constructing confidence intervals for the parameters of  $k$  groups, one should care that the overall significance level  $1 - \alpha$  is held simultaneously. This can happen when the individual intervals are wider than those constructed in the one-at-a-time manner – see Johnson and Wichern (2007), p. 226. One way of asserting the overall confidence level  $1 - \alpha$  is to use the Bonferroni inequality (see Johnson and Wichern (2007), p. 232). One variant of this inequality says that – to control the overall rate  $\alpha$  when making simultaneously  $k$  statements – the individual statements should be constructed at the significance level  $\alpha/k$ . Then the overall probability that all the  $k$  statements are true, will satisfy the inequality ( $C_i$  means the  $i$ th statement)

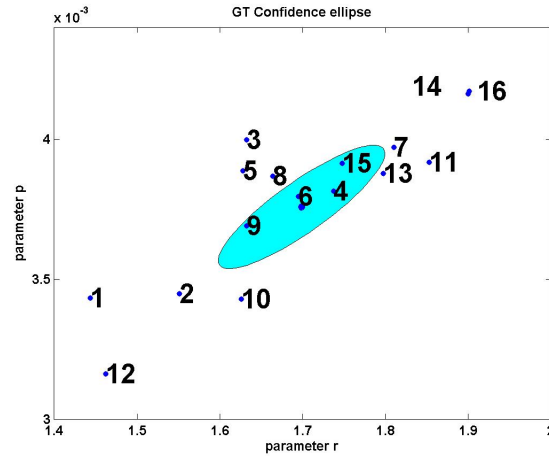
$$P[\text{all } C_i \text{ statements true}] \geq 1 - \alpha, \quad i = 1, \dots, k.$$

The number  $k$  denotes a batch of statements, for which we want to have the guarantee to have the overall confidence level  $1 - \alpha$ . In the following we will consider three batches of statements: two batches based on the univariate confidence intervals for  $\theta_1$  and  $\theta_2$ , and one batch concerned with confidence ellipses for  $(\theta_1, \theta_2)$  taken jointly. Each batch contains 17 statements: for the 16 chromosomes and for the Grand Total, which means that for all the batches we will assume  $k = 17$ . Moreover, we will assume  $\alpha = .05$ . Thus, to obtain for each batch the guarantee of an overall confidence level  $\beta = .95$ , we will construct each individual confidence interval or confidence ellipse with the individual significance level  $.05/17$ . In Table 1 we show the confidence intervals for the first two batches. The confidence ellipses obtained for individual chromosomes and for the entire data set are shown in Figures 2 and 3.

The individual ellipses are difficult to discern. Nonetheless, after careful examination, one may notice that each of the individual confidence ellipses has some (small) common part with the filled GT ellipse.



**Fig. 2.** Confidence ellipses constructed individually for all the 16 chromosomes and filled confidence ellipse for the Grand Total data. The individual ellipses were constructed at the  $\tilde{\alpha} = .05/17$  significance level.



**Fig. 3.** The GT confidence ellipse obtained from all  $n=6686$  ORFs, centered at  $\hat{r} = 1.6989$ ,  $\hat{p} = 0.0038$ . The individual chromosome estimates  $\hat{r}_i, \hat{p}_i$ ,  $i = 1, \dots, 16$ , are shown as small bullets.

Figure 3 shows the GT confidence ellipse (the same as in Figure 2) and values of the individual estimates  $(\hat{r}_j, \hat{p}_j)$ ,  $j = 1, \dots, 16$ , i.e. the centers of the confidence ellipses shown in Figure 2. One may see how the individual estimates deviate from the grand mean featured by the filled GT ellipse. Are the observed differences plausible when assuming one common model for all the 16 chromosomes?

**Test of equality of parameters for all chromosomes** We apply here a Likelihood Ratio test. Let  $\Omega = \{\theta_{i1}, \theta_{i2}, i = 1, \dots, 16\}$  denote the space of the parameters  $\theta_1, \theta_2$  which appear in the pmf of the considered random variable describing the length of ORFs.

The tested hypothesis is:  $H_0 : \{\theta_{i1} = \theta_{1*}, \theta_{i2} = \theta_{2*}\}, \forall i$ , against the alternative  $H_1$  being the negation of  $H_0$ . The hypothesis  $H_0$  imposes that the overall parameter space  $\Omega$  is restricted to a subspace  $\Omega_0 \subset \Omega$ . The Likelihood Ratio (LR) test calculates the likelihood of the observed events twice: firstly (i), under the hypothesis that the parameters  $\theta = \{\theta_1, \theta_2\}$  take values from the full parameter space  $\Omega$ , and secondly (ii), under the hypothesis that the parameters  $\theta = \{\theta_1, \theta_2\}$  are allowed to take values only from the restricted parameter space  $\Omega_0$ . The LR test formalized as the statistics  $\Lambda$  says then that we should reject  $H_0$  (in favor of  $H_1$ ), when the likelihood evaluated under (ii) is much smaller than that evaluated under (i), and we should reject  $H_0$  for small values of  $\Lambda$  (see, e.g. Johnson and Wichern, 2007, p. 211):

$$\text{Likelihood ratio} = \Lambda = \frac{\max_{\theta \in \Omega_0} L(\theta)}{\max_{\theta \in \Omega} L(\theta)} < c, \quad (9)$$

with the constant  $c$  controlling the significance level. When the sample is large and certain regularity conditions are satisfied, the sampling distribution of  $-2\ln\Lambda$  is well approximated by a  $\chi^2$  distribution with  $f$  degrees of freedom calculated as  $f = \nu - \nu_0 = (\text{dimension of } \Omega) - (\text{dimension of } \Omega_0)$ .

For our data  $\nu = 32, \nu_0 = 2, f = 30$ , the calculated value of  $\Lambda = 38.3211$  and the upper quantile  $\chi^2_{30}(0.05) = 43.7729$ . Thus the LR test does not reject the assumed hypothesis  $H_0$  on equality of the NB parameters in all the 16 chromosomes. Hence our thesis may be sustained.

## 4 Simulation Experiments

To support further the hypothesis on the common model, we have performed some computer simulations. One of the simulations was performed in the following way: Assuming the values  $r_0 = 1.6989, p_0 = 0.0038$  as the GT NB parameters, we have generated pseudo-random values from the  $NB(r_0, p_0)$  distribution using the Matlab function `nbirnd`. The pseudo-random values were generated in subgroups of the same lengths as the subgroups of the real ORFs appearing in the sixteen yeast chromosomes (the cardinalities of subsequent subgroups are shown in Table 1). All the generated values taken together gave a new GT data set. From the generated data new parameter estimates were obtained, both for the GT and the individually generated chromosomes. Then we have made displays similar to those shown in Figure 3. It is possible to present in real time an on-line slide-show depicting the results of subsequent simulations. The detailed analysis of the results of the simulations is beyond the scope of the paper. We may say only that the

simulation results confirm the conjecture that such scatter as observed in our data is quite plausible.

## 5 Concluding remarks

We have investigated the ORF length (counted in amino-acids) in the entire yeast genome and in individual chromosomes. We found together 6686 ORFs appearing in chromosomes of various length. By performing an asymptotic LR test and also by considering confidence intervals and confidence ellipses, we came to the conclusion that one negative binomial distribution with parameters estimated by the maximum likelihood method describes in a satisfactory manner the ORF length encountered in the yeast genome.

It is really amazing that such a simple probabilistic model like the NB distribution is able to describe such complicated phenomenon as the DNA code working in an environment subjected to cell divisions, cell repairs and mutational pressure.

One may ask also: is the NB characteristics specific for the yeast cell, or – may be it is valid also for other living organisms? If so, then the parameters  $r, p$  could constitute a kind of meta-characteristic of the species 'yeast'.

## References

- BARTKOWIAK, A. (2008): Orf length is negative binomial – why? In: P. Brito (Ed): *COMPSTAT 2008, Proceedings in Computational Statistics*, Contributed papers. Physica Verlag, a Springer Company, 291–298.
- BARTKOWIAK, A. and SZUSTALEWICZ, A. (2009): Are amino-acids counts in yeast ORFs negative binomial? *International J. of Biometrics* 1 (3), 268–287.
- CHRISTIANINI, N. and HAHN, M. W. (2007): *Introduction to Computational Genomics. A Case Studies Approach*. Cambridge University Press, UK.
- DURBIN, R., et al., (1999): *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge UK .
- FISZ, M. (1963): *Probability Theory and Mathematical Statistics*, Wiley, New York, also *Wahrscheinlichkeitsrechnung und Mathematische Statistik*, Veb Berlin (1976). Translation from the Polish 3rd Edition PWN, Warszawa 1967.
- HILBE, J. (2007): *Negative Binomial Regression*. Cambridge University Press, Cambridge, UK.
- JOHNSON R.A. and WICHERN D.W. (2007): *Applied Multivariate Statistical Analysis*. Sixth Edition. Pearson Education Inc/Prentice Hall, Upper Saddle River, NJ, USA.
- McLACHLAN, G. and PEEL, D. (2000): *Finite Mixture Models*. Wiley, New York, Chichester.
- MARDIA, K.V., et al. (1979): *Multivariate Analysis*. Academic Press, London, New York.
- MATLAB STATS TOOLBOX (2002): *Statistics Toolbox For Use with MATLAB*, Users Guide Version 4., ©1993 - 2002 by The MathWorks, Inc., 6th printing.
- STUART, A. et al. (1999): *Kendall's Advanced Theory of Statistics*, Volume 2A, Sixth Edition, Arnold, London.



# Influence of the Calibration Weights on Results Obtained from Czech SILC Data

Jitka Bartošová and Vladislav Bína

University of Economics in Prague, Faculty of Management, Jarošovská 1117/II,  
37701 Jindřichův Hradec, Czech Republic, {bartosov, bina}@fm.vse.cz

**Abstract.** The purpose of income sample survey is to obtain a representative data concerning level and structure of incomes and fundamental social-demographic characteristics of households and their members in the Czech Republic. The survey results are generalized to the whole population using the calibration weights created by Czech Statistical Office. An important question in the process of generalization of conclusions arising from the analysis of survey data is the influence of calibration weighting on the obtained results. The paper concerns the importance of calibration process in measuring the monetary poverty of Czech households (see also BARTOŠOVÁ, J. (2009) or BARTOŠOVÁ, J. and BÍNA, V. (2009a)). Various definitions of consuming unit are used in order to compare the influence of weighting on the relative measure of poverty in dependence on the chosen scale.

**Keywords:** calibration weights, EU-SILC, household incomes, poverty

## 1 Introduction

The results of Mikrocensus and EU - SILC surveys cannot be directly generalized on the whole population since they does not constitute a representative sample. This conclusion follows from the comparison with results of the Czech population census "Sčítání lidu, bytů a domů (SLBD)". Hence, the simple generalization of information obtained from sample data could lead to significant distortion of acquired information. There are few reasons leading to the corruption of sample representativeness.

The sample representativeness is primarily corrupted by various rates of success of interviewers in different regions (see BARTOŠOVÁ, J. and BÍNA, V. (2009b)). But there is also apparent non-uniformity of successfully returned forms in different social groups (the highest rate of success is among the households of retired and the lowest rate is among the self-employed). Also the mean size of households differs from the size ascertained in SLDB. Accordingly, it is impossible to accomplish the generalization of results using simple coefficients taking into account only the number of surveyed households in region in correspondence with its total count of inhabitants. And hence, for the construction of coefficients the iterative method of weight calibration was used. This method minimizes the difference of estimated and

recalculated sample characteristics chosen for each region separately. For the construction of weights the following characteristics were used<sup>1</sup>:

- *number of permanently occupied flats* – an estimate based on the results of SLDB and growth (resp. decrease) of flat counts
- *number of inhabitants per flat* – derived from the mean population according to demographical statistics; only inhabitants living in flats were surveyed the people populating the asylum institutions were subtracted from the demography data according to the statistics of social care
- *number of retirees (both working and not working)* – derived from the data of the Ministry of Labour and Social Affairs and Czech Social Security Administration; the number of individuals living in retirement homes and other facilities was subtracted
- *number of unemployed* – data of the Ministry of Labour and Social Affairs were increased by the estimate of unregistered unemployment provided by Labour force sample survey
- *number of self-employed* – estimate based on the Labour force sample survey data and results of SLBD.

To the above mentioned characteristics used for recomputation of Mikrocensus results in EU-SILC 2005 the following criterions was added:

- *age of the leading person of the household*
- *size groups of municipalities.*

In addition, there occurs some ten percent underestimate of incomes in the above mentioned survey. One of the reasons is the effort to suppress part of their incomes or the questioned person simply does not remember. This distortion is quite uneasy to quantify and hence the data are corrected by comparison with data concerning gross earnings. Similar approach is used in case of social benefits where the alleged values exceed reality<sup>2</sup>.

## 2 Calibration weights and their influence on the results

The above presented fact implies that the calibration weights of households is an important part of income sample data (see BÍNA, V. (2009)) and could considerably affect results of performed analyses. But there remains a question whether the influence of weighting is statistically significant or only marginal and thus nonessential.

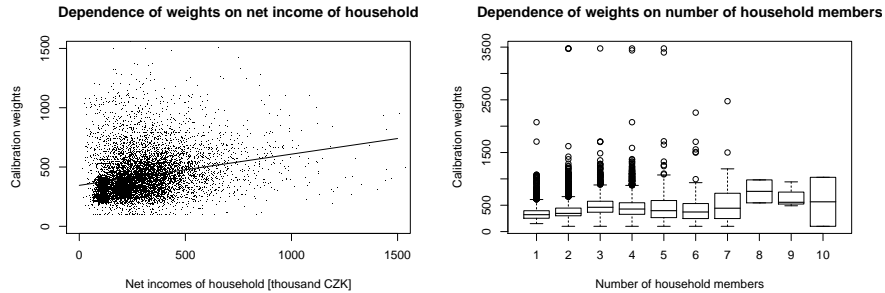
The calibration weights of households (variable *pkoe*) in sample data files of EU-SILC 2007 in most cases take the values from the interval  $\langle 100; 1000 \rangle$ . Their sum is cca 4 milion which corresponds with the number of households in the Czech Republic. Table 1 show some basic characteristics of this variable.

<sup>1</sup> Source: Methodology. Mikrocensus 2002.

<sup>2</sup> Some respondents incorrectly include minimum income benefits.

minimum	1st quart.	median	3rd quart.	maximum	mean	st. deviat.	weights sum
100.0	294.6	369.8	493.6	3475.0	417.9	205.5	4043341

**Table 1.** Basic statistical characteristics of calibration weights (EU-SILC 2007).



**Fig. 1.** Graphs of calibration weights in dependence on the household income (left part of figure) and on the number of household members (right part of figure).

According to the method of construction it can be assumed the the value of calibration weights is influenced by the number of household members (see right part of Figure 1), by age and social group of head of household (especially retired, unemployed or self-employed), but also by the size of municipality where the household is located. It is obvious that the size of calibration weights may depend on many other factors, e.g. the income of household (see left part of Figure 1), the region etc. We can observe that the conditional distributions of calibration weights contains considerable amount of outliers.

According to the sophisticated construction of calibration weights the question of their dependence on various factors is complicated and needs deeper analysis. From the practical point of view, the question of necessity of calibration coefficients in the problem of generalization to the complete population is more important. Therefore, we shall consider the influence of calibration weights on the outputs of analysis of the Mikrocensus 2002 and EU-SILC 2005, 2006, 2007 sample data. Concretely, we will focus on the detection of changes in the results of measuring the monetary poverty in the Czech Republic caused by use of calibration weights.

## 2.1 Measuring the poverty of households and individuals

In principle, we can distinguish two approaches to the definition of poverty – objective and subjective. The objective approach defines poverty using certain criterions concerning income or assets of individual. In contrary, subjective approach probes whether an individual perceive himself as poor whether

he feels the symptoms of poverty or whether he classifies himself in category of poor (see STANKOVIČOVÁ, I., BARTOŠOVÁ, J. (2009)).

In the frame of the objective approach we distinguish absolute and relative approaches. The relative approaches define poverty through the ratio to some important characteristic – mean income, median of income, distribution of income groups, etc. This second concept concerning the relative poverty is realistically described by the professor Peter Townsend from London School of Economics. He says that individuals and groups in population may be considered as poor if they lack resources to ensure the certain type of diet, living conditions and achievements common for community to which they belong. Using this approach we take into account even the level of development of given society and prevailing conditions. The importance of social context for determining the poverty is highlighted by the definition of European Commission (1984). As poor can be considered individuals, families or groups of persons whose resources (material, cultural and social) are so limited that they are excluded from the minimum acceptable way of life in states they live in.

Among the key indicators belongs the at-risk-of-poverty threshold and at-risk-of-poverty rate. At-risk-of-poverty threshold is set as 60 percent of median of equalized disposable income. In the other words as a monetary poor household is considered such a household in which the equalized disposable income is under the poverty threshold. At-risk-of-poverty rate is the proportion of individuals with equivalent disposable income under the threshold of 60 percent of median of national equivalent income.

## 2.2 Selecting the type of consuming unit

From the sample data we obtain information about total disposable income of household, about equivalent income per one representative and since 2005 also income per consumption unit which is in a way compromise of the preceding alternatives of measuring the financial power of household. Moreover, this approach considers the structure of household (i.e. the age and role of each member). Specifically defined consuming unit allows measuring of equivalent disposable income of household which is comparable with other states.

For the transformation of incomes to the equivalent scale the Czech Statistical Office uses dual methodology:

- a. OECD methodology where:
  - head of the household is considered with coefficient 1.0
  - children in age 0 – 13 with coefficient 0.5
  - other persons with coefficient 0.7
- b. EU methodology where:
  - head of the household is considered with coefficient 1.0
  - children in age 0 – 13 with coefficient 0.3
  - other persons with coefficient 0.5

Number of consuming units according to OECD definition ( $sj$ ) and EU definition ( $ej$ ) is given by

$$sj = 1 + 0.5 \cdot ych + 0.7 \cdot op \quad \text{and} \quad ej = 1 + 0.3 \cdot ych + 0.5 \cdot op$$

where  $ych$  means the number of *younger children* (children in age of 0 – 13 years in the household) and  $op$  is the number of *other persons* in household (besides the head of household).

For the comparison of financial power of households in the frame of EU the equivalent scale with number of consuming units  $ej$  is used. The value of  $ej$  is smaller than  $sj$  and brings different results in international comparison. This scale is especially suitable for comparison of Western Europe countries where households spend greater sums for common expenses (housing, water, energy, fuels etc.) than in the post-communistic countries of the Eastern Europe. The financial situation of eastern countries is more realistically represented by the methodology of OECD.

The change of weights in consuming unit definition gives us a possibility to change the point of view on the financial power of household and on the measuring of monetary poverty (see NICODEMO, C., LONGFORD, N., T. (2009)). Choosing both the weight of children and other individuals equal to zero we obtain the definition consuming unit *per household*

$$h = 1 + 0 \cdot ych + 0 \cdot op$$

and setting both coefficients equal to one we obtain definition *per representative*

$$r = 1 + 1 \cdot ych + 1 \cdot op.$$

It means that the total disposable income of household concerns the common expenditures and treats the household as a whole. In contrary, the income per one representative neglects the common expenditures and emphasizes the expenditures of individuals who are treated equivalent.

### 3 Results

As already mentioned in part 2.2 to the various types of consuming units we assign different values of incomes and different thresholds of poverty and it can be expected that also the influence of calibration weights differs.

Table 2 illustrates relatively significant influence of weighting on the estimate of poverty threshold in case of all types of consuming units. Weighted estimates are (according to expectation) higher; the only exception is the threshold of incomes per one representative in 2002. This unambiguous result can be interpreted by the slightly growing dependence of weights on the household incomes (see Figure 1).

The influence of weighting on the estimate proportion of Czech households which appeared in years 2002 – 2007 under the threshold of monetary

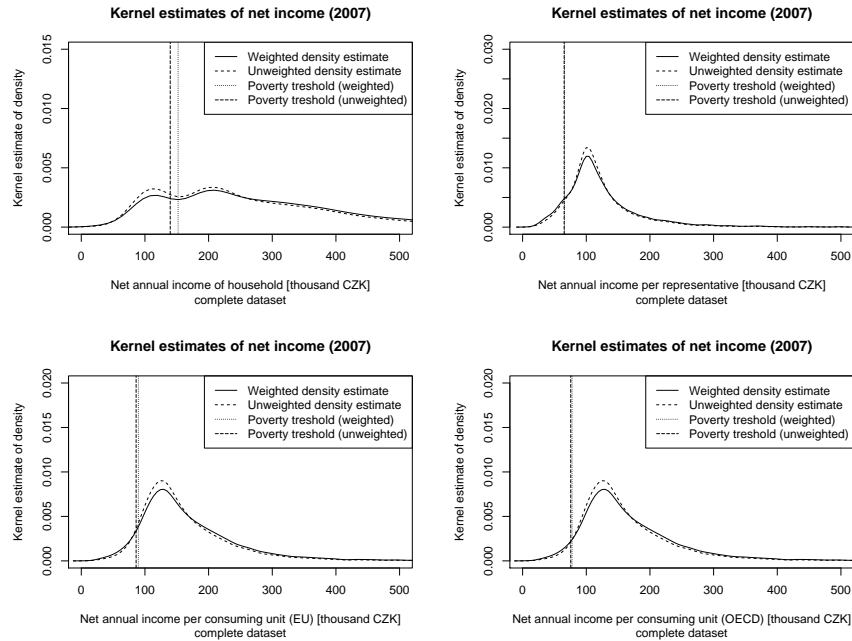
Year	Type of the consuming unit	The threshold of monetary poverty	
		weighted estimate	unweighted estimate
<b>2002</b>	household	116909	114554
	representative	52000	53522
<b>2005</b>	household	132549	123246
	representative	58200	58230
	def. EU	78786	76500
	def. OECD	68223	67199
<b>2006</b>	household	139743	128088
	representative	60912	60384
	def. EU	83052	79568
	def. OECD	72000	69926
<b>2007</b>	household	152069	139718
	representative	65850	65246
	def. EU	89611	86129
	def. OECD	89611	86129

**Table 2.** Influence of calibration weights on the threshold of monetary poverty for the different types of consuming units (Mikrocensus 2002, EU-SILC 2005 – 2007).

Year	Type of consuming unit	Freq. under risk-of-poverty threshold				Pearson in. test	
		weighted estimate		unweig. estimate		Statistics $\chi^2$	p-value
		absolute	relative	absolute	relative		
<b>2002</b>	household	1833	22.99 %	1782	22.35 %	0.894	0.3443
	representat.	672	8.43 %	757	9.49 %	5.424	<b>0.0199</b>
<b>2005</b>	household	1095	25.17 %	1012	23.26 %	4.211	<b>0.0402</b>
	representat.	439	10.09 %	439	10.09 %	0.001	0.9716
	def. EU	331	7.61 %	291	6.69 %	2.634	0.1046
	def. OECD	176	4.05 %	167	3.84 %	0.194	0.6594
<b>2006</b>	household	1878	25.10 %	1691	22.60 %	12.729	<b>0.0004</b>
	representat.	753	10.06 %	733	9.80 %	0.270	0.6035
	def. EU	570	7.62 %	469	6.27 %	10.343	<b>0.0013</b>
	def. OECD	297	3.97 %	253	3.38 %	3.490	0.0617
<b>2007</b>	household	2409	24.90 %	2193	22.67 %	13.179	<b>0.0003</b>
	representat.	858	8.87 %	832	8.60 %	0.405	0.5244
	def. EU	697	7.20 %	566	5.85 %	14.315	<b>0.0002</b>
	def. OECD	363	3.75 %	324	3.35 %	2.179	0.1400

**Table 3.** Influence of calibration weights on the rate of households under the threshold of monetary poverty for the different types of consuming units (Mikrocensus 2002, EU-SILC 2005 – 2007).

poverty is documented in Table 3. The results fully corresponds with the shift of poverty threshold shown in Table 2. We infer that the influence of weighting on the proportion of "poor" households is significant although it naturally differs according to the definition of consuming unit. This influence is apparent since the weights differ with the number of household members (as we



**Fig. 2.** Weighted and unweighted estimates of incomes (kernel density estimates with cosine kernel) and poverty thresholds for different types of consuming units (EU-SILC 2007). Upper-left – income per household, upper-right – income per representative, lower-left – income per consuming unit (EU), lower-right – income per consuming unit (OECD).

already shown in Figure 1). Setting of different coefficients to the members of household thus obviously changes (weakens or strengthens) the influence of calibration weights. The Pearson independence test shown that in nearly one half of cases (six of fourteen) we obtained statistically significant change (5 % significance level) in the number of households under poverty threshold (in years 2002 – 2007).

On the Figure 2 there are depicted the weighted and unweighted estimates of poverty threshold together with kernel estimates of income density of the Czech households in 2007. For depiction the four definitions of consuming unit were used. On the graphs we can observe both the change of density shape and the shift of poverty threshold together with the influence of the definition of consuming unit. It is obvious that the definition according to the OECD methodology is closer to the definition of income per representative than the recalculation of incomes according to the EU methodology. It appears that the equivalent scale given by EU definition emphasizes the common expenditures of the household whilst the OECD scale is closer to the concept of household as a group of "equivalent" individuals.

## 4 Conclusion

Sample survey Mikrocensus and the Czech modification of EU-SILC survey constitute the data base for acquiring information concerning the incomes and other social and demographic characteristics of Czech households. One of the variables in the sample files are the calibration weights that can significantly influence the results of realized analyses. It appears that the role of calibration (the size of weights) changes with the number of household members, grows with the growing incomes, etc.

The paper focuses on the strength of influence of the calibration weight on the risk of monetary poverty in the Czech Republic. We shown that the bias of results occurred in all cases (usually higher values) and in more than half cases this change was statistically significant (on the 5% level). It means that the unweighted results are slightly distorted but only in some half of cases the bias is statistically significant.

In order to create a complex insight on the problem of biasing the results of measuring the relative poverty by calibration weights, our analyses were based on the study of different definitions of consuming unit which handles the monetary poverty from different perspectives. We shown that the choice of scale can suppress or emphasize the influence of calibration weights.

An important outcome is the influence of consuming unit definition on the risk of poverty of Czech households. And therefore the suitable definition of consuming unit plays the key role in identifying of relative poverty in society.

## Acknowledgement

The paper was supported by Czech Science Foundation (project 402/09/0515: Analysis and Modelling of Financial Power of Czech and Slovak Households).

## References

- BARTOŠOVÁ, J. (2009): Analysis and Modelling of Financial Power of Czech Households. *Aplimat – J. of Appl. Math.* 2 (3), STU Bratislava, 31–36.
- BARTOŠOVÁ, J. and BÍNA, V. (2009a): Modelling of income distribution of Czech households in years 1996 - 2005. *Acta Oeconomica Pragensia* 17 (4), *Oeconomica, Prague*, 3–18.
- BARTOŠOVÁ, J. and BÍNA, V. (2009b): Financial Power of the Czech Households. In: *EURISBIS09: Book of Abstracts. TILAPIA, Cagliari*, 201–202.
- BÍNA, V. (2009): The role of callibration weights in SILC data. In: *Finanční potenciál domácností '09 (Proceedings of workshop of GAČR 402/09/0515). University of Economics in Prague, CDROM*.
- NICODEMO, C., LONGFORD, N.,T. (2009): A sensitivity analysis of poverty definitions used with EU-SILC. In: *Finanční potenciál domácností '09 (Proceedings of workshop of GAČR 402/09/0515). Univ. of Economics in Prague, CDROM*.
- STANKOVIČOVÁ, I., BARTOŠOVÁ, J. (2009): Príspevok k analýze subjektívnej chudoby v SR a ČR. *Forum Statisticum Slovaca* 5 (3), SŠDS, Bratislava, CDROM.



# Continuous Wavelet Transform and the Annual Cycle in Temperature and the Number of Deaths<sup>\*</sup>

Milan Bašta<sup>1</sup>, Josef Arlt<sup>1</sup>, Markéta Arltová<sup>1</sup>, and Karel Helman<sup>1,2</sup>

<sup>1</sup> Dept. of Statistics and Probability, Faculty of Informatics and Statistics,  
University of Economics, Prague, Czech Republic *milan.basta@vse.cz*

<sup>2</sup> Czech Hydrometeorological Institute

**Abstract.** The continuous wavelet transform applied to one time series allows the analysis of the temporal evolution and changes of the frequency content of this time series. The application of the cross-wavelet transform to two time series may reveal a complex relationship between the two time series - specifically, a relationship that differs from one frequency range to another and that is transient or evolves in time. As such, the wavelet transform is an intriguing tool for the analysis of demographic time series. In this paper we apply it to the analysis of the daily time series of the number of deaths due to cardiovascular diseases in Prague, Czech Republic and the daily time series of the average temperature in Prague, Czech Republic.

**Keywords:** wavelets, demography, time series, death rate, temperature

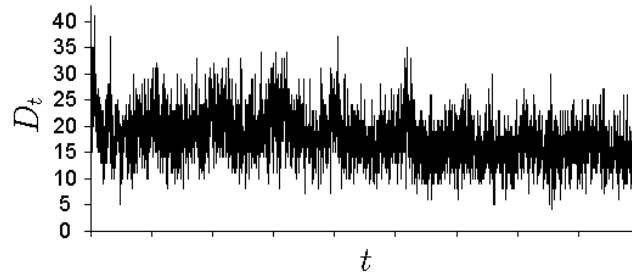
## 1 Data

In this paper we analyze the relationship between the daily time series  $D_t$  of the number of deaths due to diseases of the cardiovascular system in Prague, Czech Republic (see Figure 1) and the daily time series  $T_t$  of the average temperature in Prague, Czech Republic (see Figure 2). The time span of both time series is from Jan 1, 2000 till December 31, 2008 (altogether 3288 data points for each time series). The time series were provided by the Czech Statistical Office and the Czech Hydrometeorological Institute (see also <http://www.czso.cz/> and <http://www.chmi.cz/>).

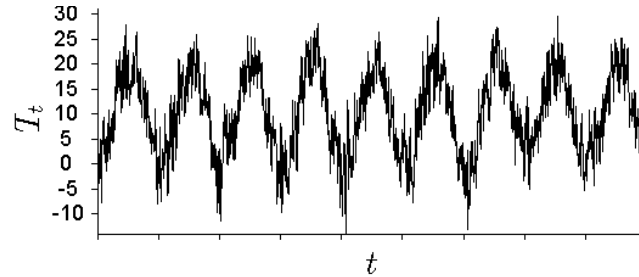
The behaviour of both time series may show transient events, that occur only from time to time. Moreover, it might be interesting to study the dynamics in various frequency regions. The continuous wavelet transform, which will be introduced in the next section, is a suitable tool for such an analysis, as it analyzes the temporal evolution of the frequency content of the time series. In this paper we will use the methodology introduced in physical and climatological sciences, specifically in papers Grinsted et al. (2004),

---

<sup>\*</sup> The paper was written with the support of the Grant Agency of the Czech Republic No. 402/09/0369, Modelling of Demographic Time Series in the Czech Republic



**Fig. 1.** The daily time series of the number of deaths due to cardiovascular diseases in Prague, Czech Republic from January 1, 2000 to December 31, 2008. The ticks on the  $x$ -axis denote the start of individual calendar years.



**Fig. 2.** The daily time series of the average temperature in Prague, Czech Republic from January 1, 2000 to December 31, 2008. The ticks on the  $x$ -axis denote the start of individual calendar years.

Torrence and Compo (1998) and Torrence and Webster (1999). The figures in this paper were made with the help of Aslak Grinsted's Cross Wavelet and Wavelet Coherence package.<sup>1</sup>

## 2 Wavelets

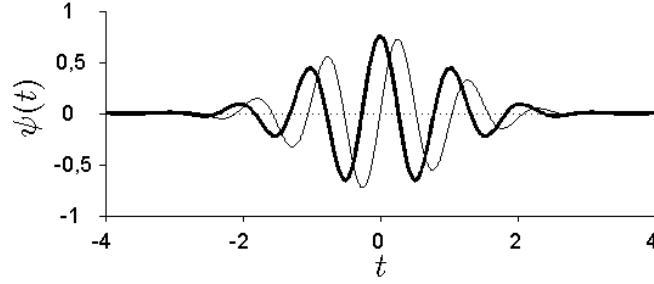
A wavelet  $\psi(t)$  is a function that fulfills the following two conditions (see also Percival and Walden (2000))

$$\int_{-\infty}^{\infty} \psi(t) dt = 0, \quad (1)$$

$$\int_{-\infty}^{\infty} \psi^2(t) dt = 1. \quad (2)$$

Moreover, additional conditions may also be required.

<sup>1</sup> <http://www.pol.ac.uk/home/research/waveletcoherence/>



**Fig. 3.** Morlet wavelet - the real part is depicted with a thick curve, the imaginary part with a thin curve.

Our choice of the  $\psi(t)$  function in this text is the Morlet wavelet, here defined as

$$\psi(t) \equiv \pi^{-1/4} \exp(i6t) \exp(-t^2/2). \quad (3)$$

The Morlet wavelet is depicted in Figure 3.<sup>2</sup> The reason for the choice of the Morlet wavelet in this paper is the fact that it is a complex wavelet and will thus enable to analyze the time delay between two time series (see also later in the text).

The daughter wavelets  $\psi_{s,\tau}(t)$  for the Morlet wavelet  $\psi(t)$  are defined as

$$\psi_{s,\tau}(t) \equiv \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right). \quad (4)$$

The scale parameter  $s$  serves for scaling of the function  $\psi(t)$ , the parameter  $\tau$  for changing its location in time. The factor  $\frac{1}{\sqrt{s}}$  ensures that

$$\int_{-\infty}^{\infty} \psi_{s,\tau}^2(t) dt = 1. \quad (5)$$

### 3 Continuous wavelet transform and wavelet power

The continuous wavelet transform  $W_x(\tau, s)$  of time series  $x_t : t = 0, \dots, N-1$  of length  $N$  is defined as

$$W_x(\tau, s) \equiv \sum_{t=0}^{N-1} x_t \psi_{s,\tau}^*(t) = \frac{1}{\sqrt{s}} \sum_{t=0}^{N-1} x_t \psi^*\left(\frac{t-\tau}{s}\right), \quad (6)$$

where  $\psi_{s,\tau}^*(t)$  and  $\psi^*(t)$  are the complex conjugates of  $\psi_{s,\tau}(t)$  and  $\psi(t)$ . The continuous wavelet transform  $W_x(\tau, s)$  is a function of the scale parameter  $s$ ,

<sup>2</sup> Morlet wavelet defined by the equation (3) does not fulfill the condition (1) exactly but only approximately. Despite of this fact the function defined by equation (3) is often referred to as the Morlet *wavelet*.

the location parameter  $\tau$  and is a complex function. The fact that the Fourier transform of the right-hand side of equation (6) turns the convolution of the input time series and the daughter wavelet into multiplication of their Fourier transforms offers a convenient way of the calculation of the continuous wavelet transform.

For a time series  $x_t : t = 0, \dots, N - 1$  the normalized wavelet power is defined as

$$\frac{|W_x(\tau, s)|^2}{\hat{\sigma}_x^2}, \quad (7)$$

where  $\hat{\sigma}_x^2$  is defined as

$$\hat{\sigma}_x^2 \equiv \frac{1}{N} \sum_{t=0}^{N-1} (x_t - \bar{x})^2, \quad (8)$$

where  $\bar{x}$  is the defined as

$$\bar{x} \equiv \frac{1}{N} \sum_{t=0}^{N-1} x_t. \quad (9)$$

For time series, which are realizations of a white noise process, the distribution of  $\frac{|W_x(\tau, s)|^2}{\hat{\sigma}_x^2}$  (as a random variable) is given by (see e.g. Torrence and Compo (1998))

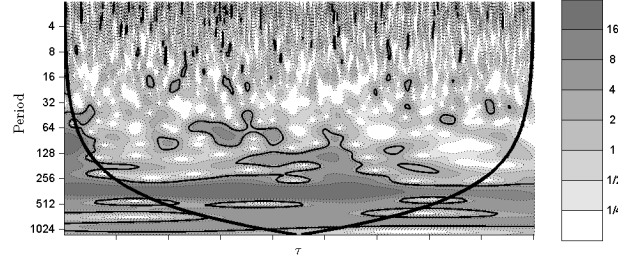
$$\frac{|W_x(\tau, s)|^2}{\hat{\sigma}_x^2} \sim \frac{1}{2} \chi^2[2]. \quad (10)$$

Significant deviations from the expected value of  $\frac{|W_x(\tau, s)|^2}{\hat{\sigma}_x^2}$  reveal regions in the time-scale plane where the hypothesis of a white noise process is rejected.

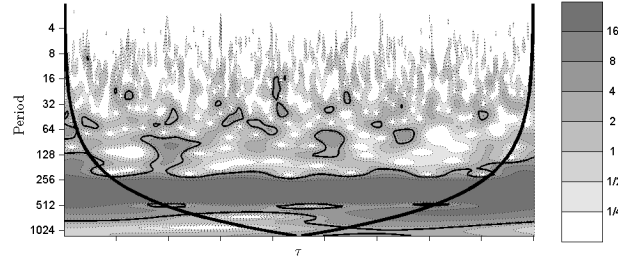
In Figures 4 and 5 the normalized wavelet power is shown for the time series  $D_t$  and  $T_t$ . On the  $x$ -axis the parameter  $\tau$  is drawn, on the  $y$ -axis the Fourier period  $\lambda$  is drawn. Through the study of the wavelet transform of cosine waves with known frequency, it can be shown that  $\lambda = 1.03 \times s$ , see e.g. Meyers et al. (1993) for details. The values of the normalized power are depicted in greyscale. The region underneath the thick U-shaped curve in both Figures 4 and 5 is called the cone of influence. In this region the calculation of  $W_x(\tau, s)$  is influenced by the assumptions that might not be valid (for details see Torrence and Compo (1998) or Grinsted et al. (2004)). Values  $\frac{|W_x(\tau, s)|^2}{\hat{\sigma}_x^2}$  which are higher than  $\frac{1}{2} \chi_{0.95}^2[2]$  point out the regions that are significantly different (at the 5 % significance level) from the dynamics of the white noise process.<sup>3</sup> These regions are marked by the thick contours in Figures 4 and 5.

Clearly, as could have been expected, there is a strong annual cycle in the time series  $T_t$ , with low temperatures in winter and high temperatures in summer. This annual cycle is clearly detected in the normalized wavelet power as a significant region within the range of periods 256 - 512 (days) in

<sup>3</sup>  $\chi_{0.95}^2[2]$  is the 95th percentile of the  $\chi^2[2]$  distribution.



**Fig. 4.** Normalized wavelet power for the number of deaths due to cardiovascular diseases.



**Fig. 5.** Normalized wavelet power for the temperature.

Figure 5. A strong annual component is also present in the time series  $D_t$ . This can be seen as the significant region in the range of periods 256 - 512 (days) in Figure 4. Dynamics of both time series,  $T_t$  as well as  $D_t$ , are clearly dominated by the annual component. Some transient minor contributions to the variability of the time series occur also at different frequencies.

#### 4 Cross-wavelet power

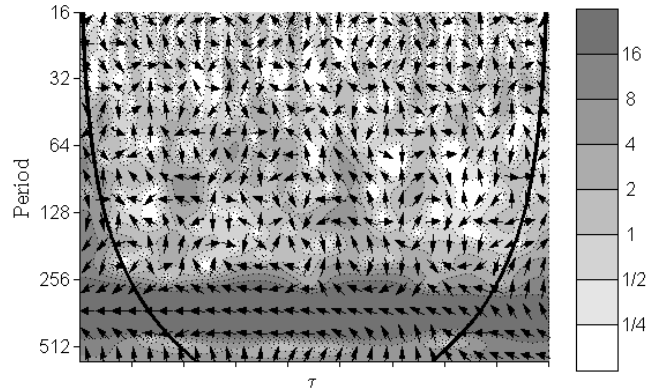
For two time series  $x_t$  and  $y_t$  the cross-wavelet transform  $W_{xy}(\tau, s)$  is defined as

$$W_{xy}(\tau, s) \equiv W_x(\tau, s)W_y^*(\tau, s), \quad (11)$$

where  $W_x(\tau, s)$  is the continuous wavelet transform of  $x_t$  and  $W_y(\tau, s)$  is the continuous wavelet transform of  $y_t$  and  $W_y^*(\tau, s)$  is the complex conjugate of  $W_y(\tau, s)$ . The cross-wavelet transform  $W_{xy}(\tau, s)$  is complex and may thus be rewritten as

$$W_{xy}(\tau, s) = |W_{xy}(\tau, s)| \arg(W_{xy}(\tau, s)). \quad (12)$$

The magnitude  $|W_{xy}(\tau, s)|$  is called the cross-wavelet power. The argument  $\arg(W_{xy}(\tau, s))$  can be interpreted as the local phase between the time series



**Fig. 6.** Normalized cross-wavelet power (in greyscale) and the local phase (in arrows) for the temperature and the number of deaths due to cardiovascular diseases.

$x_t$  and  $y_t$  in the time-scale plane. The normalized cross-wavelet power is defined as

$$\frac{|W_{x,y}(\tau, s)|}{\hat{\sigma}_x \hat{\sigma}_y}. \quad (13)$$

The normalized cross-wavelet power for the time series  $T_t$  and  $D_t$  is depicted in Figure 6. We can see that the highest value of the normalized cross-wavelet power occurs within the annual cycle region, which suggests a strong comovement of the two time series within the annual cycle. The tilt of the arrows shows the local value of  $\arg(W_{xy})$ . Let  $\alpha(\tau, s)$  be the angle the arrow makes with the horizontal line at the position  $[\tau, s]$  in the time-scale plane and let  $-\pi < \alpha(\tau, s) \leq \pi$ , with positive angles measured anticlockwise. Positive values of  $\alpha(\tau, s)$  suggest that the dynamics of the time series  $T_t$  is locally (at time  $\tau$ ) delayed behind the dynamics of  $D_t$  on scale  $s$ . We can clearly see that the time series  $T_t$  and  $D_t$  are in anti phase in the annual cycle region. This means that high temperature in the summer implies lower mortality due to diseases of the cardiovascular system and low temperature in winter implies higher mortality due to diseases of the cardiovascular system.

## 5 Wavelet coherence

Following Torrence and Webster (1999) we may define the wavelet squared coherence for two time series  $x_t$  and  $y_t$  as

$$R^2(\tau, s) = \frac{|S(s^{-1}W_{xy}(\tau, s))|^2}{S(s^{-1}|W_x(\tau, s)|^2) S(s^{-1}|W_y(\tau, s)|^2)}, \quad (14)$$

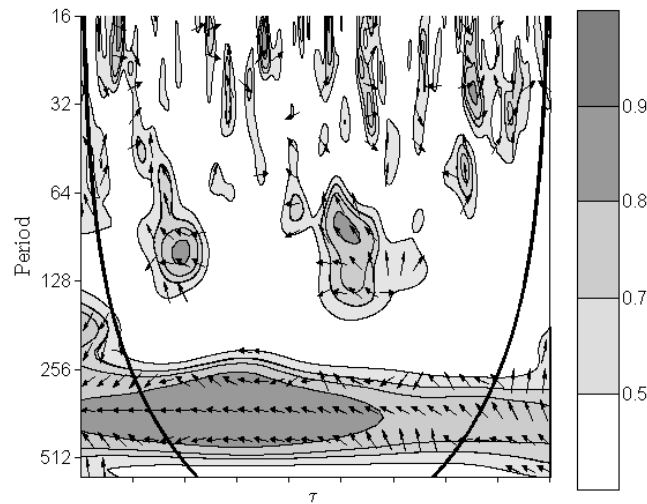
where  $S$  stands for a smoothing operator, which involves smoothing in time and scale. The wavelet squared coherence might be informally interpreted as the square of the local correlation coefficient.

The wavelet squared coherence for the time series  $T_t$  and  $D_t$  is shown in Figure 7. The regions in thick contour mark regions which are statistically significant (at 5 % significance level) with respect to the hypothesis of jointly stationary independent AR(1) processes. Specifically, for each time series ( $D_t$  as well as  $T_t$ ) the AR(1) parameter is estimated and several pairs of surrogate independent AR(1) time series with the same parameters (as are the estimated parameters) are generated. Consequently the value of  $R^2(\tau, s)$  is calculated for each pair and the distribution of  $R^2(\tau, s)$  under the hypothesis of independent AR(1) time series is obtained. Even though neither of the time series  $T_t$  and  $D_t$  is exactly an AR(1) time series, the significant regions are approximately valid (as Grinsted et al. (2004) have shown that AR(1) coefficients have little impact on the significance level; so presumably the parameterization has little impact in general, which might be intuitively expected from the definition of the wavelet squared coherence). The anti-phase behaviour of  $T_t$  and  $D_t$  in the annual cycle region suggests high temperatures are accompanied by low number of deaths and low temperatures by higher number of deaths.

Moreover, other regions also emerged as statistically significant, although this significance is only transient in nature. For example, a thorough calculation shows that statistically significant regions in the range of periods between 16 and 32 days occur much more often than what would be expected by pure chance. Moreover, the two time series are on average positively correlated in this region (see also Figure 7).

## 6 Conclusion

The relationship between the daily time series of the number of deaths due to cardiovascular diseases in Prague, Czech Republic and the daily time series of the average temperature in Prague, Czech Republic is very complex. The continuous wavelet transform is an innovative tool which enables to uncover this complex structure. Both the number of deaths and temperature exhibit pronounced annual components. These annual components are negatively correlated - high temperature in summer implies lower number of deaths, whereas low temperature in winter implies higher number of deaths. However, in other frequency regions correlation might be positive, though only transient in nature. For example, a positive correlation of components is present in the region of periods between 16 and 32 days - where an increase in temperature implies an increase in the number of deaths. We may thus conclude that the relationship of both time series is complex, being a function of time and frequency. Further research is necessary to provide detailed explanation for such a behaviour.



**Fig. 7.** Wavelet squared coherence (in greyscale) and the local phase (in arrows) for the temperature and the number of deaths due to cardiovascular diseases.

## 7 Acknowledgement

We acknowledge the support of the Grant Agency of the Czech Republic No. 402/09/0369, Modelling of Demographic Time Series in the Czech Republic. We acknowledge the use of data provided by the Czech Statistical Office and the Czech Hydrometeorological Institute. Crosswavelet and wavelet coherence software were provided by A. Grinsted.

## References

- GRINSTED, A., MOORE, J. and JEVREJEVA, S. (2004): Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics* 11, 561-566
- MEYERS, S., KELLY, B. and O'BRIEN, J. (1993): An introduction to wavelet analysis in oceanography and meteorology: With application to the dispersion of Yanai waves. *Mon. Weather Rev.* 121, 2858-2866
- PERCIVAL, D. and WALDEN, A. (2000): *Wavelet Methods for Time Series Analysis*. 1 edition. Cambridge University Press.
- TORRENCE, C. and COMPO, G. (1998): A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society* 79, 61-78
- TORRENCE, C. and WEBSTER, P. (1999): Interdecadal changes in the ENSO-Monsoon System. *Journal of Climate* 12 (8), 2679-2690



# EM-Like Algorithms for Nonparametric Estimation in Multivariate Mixtures

Tatiana Benaglia<sup>1</sup>, Didier Chauveau<sup>2</sup>, and David R. Hunter<sup>1</sup>

<sup>1</sup> Department of Statistics, Pennsylvania State University, USA

<sup>2</sup> Université d'Orléans, UMR CNRS 6628 - MAPMO - BP 6759  
45067 Orléans Cedex 2, France *didier.chauveau@univ-orleans.fr*

**Abstract.** We propose an iterative algorithm for nonparametric estimation for finite mixtures of multivariate random vectors which has connections with the EM algorithm. The vectors are assumed to have independent coordinates conditionally to their mixture component, but otherwise their density functions may be nonparametric, or may be partially specified (semiparametric). This algorithm is much more easily applicable than existing algorithms in the literature. Several versions of it can be defined, and in particular we discuss here adaptive bandwidth issues for the involved kernel density estimates. An illustration using our implementation in the *mixtools* package for the R statistical software is given.

**Keywords:** EM algorithm, kernel density estimation, multivariate mixture, nonparametric mixture

## 1 Introduction

Suppose the  $r$ -dimensional vectors  $\mathbf{X}_1, \dots, \mathbf{X}_n$  are a simple random sample from a finite mixture of  $m > 1$  distributions (see, e.g., McLachlan and Peel, 2000). For nonparametric mixtures, we do *not* assume that the component distributions come from a family of densities that may be indexed by a finite-dimensional parameter vector. However, it is necessary to restrict the family  $\mathcal{F}$  of multivariate density functions from which the component densities are drawn in order to avoid the problem of model non-identifiability, as discussed by Benaglia, Chauveau and Hunter (2009a) and several of the references therein. To this end, we assume that  $\mathcal{F}$  contains only densities equal to the product of their  $r$  univariate marginal densities. In other words, the coordinates of the  $\mathbf{X}_i$  vector are independent, conditional on the subpopulation or component (1 through  $m$ ) from which  $\mathbf{X}_i$  is drawn. This *conditional independence assumption* has appeared in a growing body of literature on non- and semi-parametric multivariate mixture models; see Benaglia et al. (2009a) for a discussion of the relevant literature. We avoid discussion of the identifiability question here, except to state that Allman et al. (2008) give mild sufficient conditions for identifiability whenever  $r \geq 3$ .

We let  $\boldsymbol{\theta}$  denote the vector of parameters, including the mixing proportions  $\lambda_1, \dots, \lambda_m$  and the univariate densities  $f_{jk}$  (here  $j$  indexes the component and  $k$  indexes the coordinate, so  $1 \leq j \leq m$  and  $1 \leq k \leq r$ ). Thus, under

the assumption of conditional independence, the mixture density evaluated at  $\mathbf{x}_i = (x_{i1}, \dots, x_{ir})^t$  is

$$g_{\boldsymbol{\theta}}(\mathbf{x}_i) = \sum_{j=1}^m \lambda_j \prod_{k=1}^r f_{jk}(x_{ik}). \quad (1)$$

In many models, there is reason to assume that *some* or *all* of the  $r$  coordinate densities  $f_{j1}(\cdot), \dots, f_{jr}(\cdot)$  are the same for all  $j$ . For instance, Elmore et al. (2004) and related articles assume that the coordinates are conditionally independent and identically distributed (iid), i.e.,  $f_{j1}(\cdot) = \dots = f_{jr}(\cdot)$  for all  $j$ . In order to encompass both the conditionally iid case and the more general case simultaneously in the model and our algorithm, we allow that the coordinates of  $\mathbf{X}_i$  are conditionally independent, and that there exist *blocks* of coordinates that are also identically distributed. These blocks may all be of size one so that the general case is still covered, or there may exist only a single block of size  $r$ , which is the conditional iid case. If we let  $b_k$  denote the block index to which the  $k$ th coordinate belongs, where  $1 \leq b_k \leq B$  and  $B$  is the total number of such blocks, the general model is

$$g_{\boldsymbol{\theta}}(\mathbf{x}_i) = \sum_{j=1}^m \lambda_j \prod_{k=1}^r f_{jb_k}(x_{ik}). \quad (2)$$

A motivating example of model (2) is an experiment involving  $n = 405$  children aged 11 to 16 years subjected to a water-level task as described by Thomas et al. (1993). Each child is presented with eight rectangular vessels on a sheet of paper, each tilted to one of  $r = 8$  clock-hour orientations: in order of presentation to the subjects, these orientations are 11, 4, 2, 7, 10, 5, 1, and 8 o'clock. Each vessel was on a separate sheet of paper. The children's task was to draw a line representing the surface of still liquid in the closed, tilted vessel in each picture. The acute angle formed between the horizontal and this line was measured for each response. In this "Water-level" data example a careful analysis reveals the differences among the coordinate distributions, and suggests that grouping the coordinates into four blocks of two i.i.d. coordinates each, by opposite orientations ((1,7), (2,8), ...) appears more appropriate.

## 2 The nonparametric EM algorithm

Benaglia et al. (2009a) recently propose an algorithm for estimating  $\boldsymbol{\theta}$ , that is based on the well-known family of EM algorithms for parametric mixture models. This "npEM" algorithm, which is implemented in the `mixtools` package (Young et al., 2009) for R (R Development Core Team, 2008), operates as follows: Given initial values  $\boldsymbol{\theta}^0 = (\boldsymbol{\lambda}^0, \mathbf{f}^0)$ , iterate the following steps for  $t = 0, 1, \dots$ :

- **E-step:** Letting  $Z_{ij}$  denote the (unobserved) indicator of the event that the  $i$ th observation is drawn from the  $j$ th component, calculate the “posterior” probabilities (conditional on the data and  $\theta^t$ ) of component inclusion,

$$p_{ij}^t \stackrel{\text{def}}{=} P_{\theta^t}(Z_{ij} = 1 | \mathbf{x}_i) \quad (3)$$

$$= \frac{\lambda_j^t \prod_{k=1}^r f_{jb_k}^t(x_{ik})}{\sum_{j'=1}^m \lambda_{j'}^t \prod_{k=1}^r f_{j'b_k}^t(x_{ik})}, \quad (4)$$

for all  $i = 1, \dots, n$  and  $j = 1, \dots, m$ .

- **M-step:** Set

$$\lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^n p_{ij}^t \quad (5)$$

for  $j = 1, \dots, m$ .

- **Nonparametric density estimation step:** For each component  $j \in \{1, \dots, m\}$  and each block  $\ell \in \{1, \dots, B\}$ , define the function

$$\begin{aligned} f_{j\ell}^{t+1}(u) &= \frac{\frac{1}{h} \sum_{k=1}^r \sum_{i=1}^n p_{ij}^t I\{b_k = \ell\} K\left(\frac{u - x_{ik}}{h}\right)}{\sum_{k=1}^r \sum_{i=1}^n p_{ij}^t I\{b_k = \ell\}} \\ &= \frac{1}{nh C_\ell \lambda_j^{t+1}} \sum_{k=1}^r \sum_{i=1}^n p_{ij}^t I\{b_k = \ell\} K\left(\frac{u - x_{ik}}{h}\right), \end{aligned} \quad (6)$$

where  $K(\cdot)$  is a kernel density function,  $h$  is a bandwidth, and

$$C_\ell = \sum_{k=1}^r I\{b_k = \ell\} \quad (7)$$

is the number of coordinates in the  $\ell$ th block.

For instance, we successfully analyzed the Water-level dataset (with  $m = 3$  or  $4$ ) with our algorithm in Benaglia et al. (2009a). The point is that our method appears to be the only one that is both fully general *and* easily extendible to any values  $m$  and  $r$  for which model (2) is identifiable, and we know of no other algorithm currently capable of producing similar results (e.g., the method of Hall et al. (2005) could potentially be extended to this case, but this method appears to be extremely complicated computationally for  $m > 2$  or  $r > 3$ ).

It is easy to derive modified versions of this npEM algorithm tailored for specific situations such as components differing only by a location or scale parameter (that is, semiparametric), univariate symmetric components as in Bordes, Chauveau and Vandekerckhove (2007), and others. Extensions of several of these versions are discussed in Benaglia et al. (2009a). For instance, a semiparametric restrictive case is the case in which the density function for

each component and block shares exactly the same shape, and the different components and blocks differ only by a possible location and scale parameter, so that for every component  $j$  and block  $\ell$ , the  $(j, \ell)$  density function becomes

$$f_{j\ell}(x) \equiv \frac{1}{\sigma_{j\ell}} f\left(\frac{x - \mu_{j\ell}}{\sigma_{j\ell}}\right). \quad (8)$$

In this case, because we wish to estimate only a single density  $f$ , the density estimation step in the algorithm would specify for any  $u \in \mathbb{R}$  that

$$f^{t+1}(u) = \frac{1}{nrh} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^r p_{ij}^t K\left(\frac{u - x_{ik} + \mu_{jb_k}}{h\sigma_{jb_k}}\right). \quad (9)$$

### 3 From fixed to adaptive bandwidth

The nonparametric density estimation step (6) of the npEM algorithm described above is a modified version of kernel density estimation, a well-studied topic in statistics (see, e.g., Silverman, 1986). The central decision in this step is the selection of an appropriate value for the bandwidth  $h$ , or smoothing parameter, since this choice affects density estimates dramatically (the choice of the kernel function  $K(\cdot)$  being not as influential). Silverman's rule (Silverman, 1986 p. 48) for univariate data gives  $h = 0.9 \min\{\text{SD}, \text{IQR}/1.34\} n^{-1/5}$ , where SD and IQR are respectively the standard deviation and interquartile range of the  $n$  data values.

Benaglia et al. (2009a) use a simplistic bandwidth-selection scheme, by applying Silverman's rule to the entire  $n \times r$  dataset treated as a vector of size  $nr$ , using thus a single bandwidth  $h$  for all components and blocks. We propose to extend that by allowing for (1) iteratively updating bandwidths; and (2) component- and block-specific bandwidths. Each of these two extensions improves the density estimations in certain cases for a different reason: The first solves the problem that it is difficult to estimate a bandwidth before knowing about the mixture structure, while the second takes care of cases in which different components or blocks have very different properties (e.g., support). In other words, we propose to modify the npEM algorithm by replacing  $h$  by  $h_{j\ell}^t$  in equation (6).

To adapt Silverman's rule of thumb to select the value of  $h_{j\ell}$  in an iterative procedure, for each iteration of the npEM algorithm, we need an estimate of the sample size, the sample standard deviation, and the interquartile range for each component and each block. Once these estimates are in place, the estimated bandwidths at the  $(t+1)$ th iteration, calculated just before the density estimation step of the algorithm, are given by:

$$h_{j\ell}^{t+1} = 0.9 \min\left\{\sigma_{j\ell}^{t+1}, \frac{\text{IQR}_{j\ell}^{t+1}}{1.34}\right\} (nC_\ell \lambda_j^{t+1})^{-1/5}, \quad (10)$$

where  $nC_\ell\lambda_j^{t+1}$  estimates the sample size for the  $\ell$ th block of coordinates in the  $j$ th component, and  $\sigma_{j\ell}^{t+1}$  and  $IQR_{j\ell}^{t+1}$  are the weighted standard deviation and empirical interquartile range for the  $j$ th component and  $\ell$ th block. Calculation of  $\sigma_{j\ell}^{t+1}$  is fairly straightforward if we augment each M-step to include

$$\mu_{j\ell}^{t+1} = \frac{\sum_{i=1}^n \sum_{k=1}^r p_{ij}^t I\{b_k = \ell\} x_{ik}}{\sum_{i=1}^n \sum_{k=1}^r p_{ij}^t I\{b_k = \ell\}} = \frac{\sum_{i=1}^n \sum_{k=1}^r p_{ij}^t I\{b_k = \ell\} x_{ik}}{n\lambda_j^{t+1} C_\ell} \quad (11)$$

and

$$\sigma_{j\ell}^{t+1} = \left[ \frac{1}{nC_\ell\lambda_j^{t+1}} \sum_{i=1}^n \sum_{k=1}^r p_{ij}^t I\{b_k = \ell\} (x_{ik} - \mu_{j\ell}^{t+1})^2 \right]^{1/2}. \quad (12)$$

We compute  $IQR_{j\ell}^{t+1}$  as the difference between the estimated 0.75 and 0.25 quantiles of the  $\ell$ th block of the  $j$ th component. To accomplish this, we first introduce the notion of a weighted quantile estimate:

Let  $a_1, \dots, a_\nu$  be real numbers and  $w_1, \dots, w_\nu$  be associated (nonnegative) weights, with  $W = w_1 + \dots + w_\nu$ . The first step in finding the weighted quantile estimate is to sort the  $a_i$  in non-decreasing order. To this end, let  $\tau(\cdot)$  be a permutation on the integers  $\{1, \dots, \nu\}$  such that

$$a_{\tau(1)} \leq a_{\tau(2)} \leq \dots \leq a_{\tau(\nu)}.$$

(The  $\tau$  permutation need not be unique if there are ties among the  $a_i$ .) Then for  $\alpha \in (0, 1)$ , we define the weighted  $\alpha$  quantile estimate to be  $a_{\tau(i_\alpha)}$ , where

$$i_\alpha = \min \left\{ s : \sum_{i=1}^s w_{\tau(i)} \geq \alpha W \right\}$$

is the smallest integer that gives at least a proportion  $\alpha$  of the total sum of weights  $W$ . Note that in the special case in which all  $w_i$  are the same, the weighted quantile estimate is simply a particular way to define the regular sample quantile.

To find  $IQR_{j\ell}^{t+1}$ , we first calculate the weighted 0.25 and 0.75 quantile estimate of the  $nC_\ell$  data values in block  $\ell$ , with corresponding weights given by the posterior probabilities  $p_{ij}^t$ . The weighted interquartile range is then the difference between these two quantiles. Note that when this calculation is performed as part of the npEM algorithm, the permutation  $\tau$  need only be calculated once due to the fact that the data and block structure do not change during the running of the algorithm. Furthermore, the sum  $W$  of the weights at iteration  $t$  is equal to  $\lambda_j^{t+1} C_\ell$  because of equation (5).

## 4 An example

Real-size examples showing the effect of allowing blocks in the model can be found in Benaglia, Chauveau and Hunter (2009a). For instance the Water-level data example described in the introduction is fitted there with the npEM algorithm, and compared with other methods from the literature. A large-scale simulation study is also presented, using three benchmark models from Hall et al. (2005) which shows that the npEM algorithm (with fixed bandwidth) dramatically outperforms Hall et al. (2005)'s inversion method.

We just show here a simple, toy example designed to illustrate the effect of allowing different bandwidths in the npEM algorithm. We generated 300 observations from a  $m = 2$  component mixture from the general model (2) with trivariate ( $r = 3$ ) observations,  $B = 1$  block (i.e.  $b_1 = b_2 = b_3 = 1$ , which means that we have three conditionally iid repeated measures), and parameter and densities:

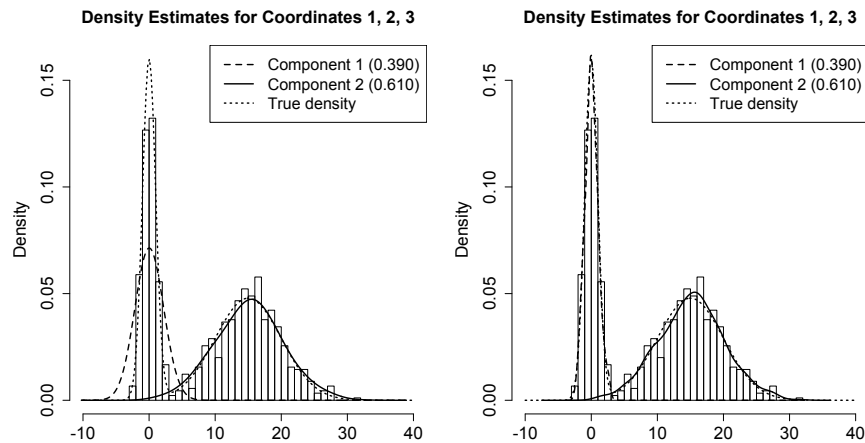
$$\lambda_1 = 0.4, \quad f_{11} \equiv \mathcal{N}(0, 1), \quad f_{21} \equiv \mathcal{N}(15, 25). \quad (13)$$

We estimate the model applying the npEM algorithm using both a single bandwidth given by Silverman's rule, and adaptive bandwidths computed by (10). The code used to generate the data and then apply the npEM algorithm in R using the `mixtools` package may be found in Benaglia et al. (2009b).

If we consider each component separately and apply Silverman's rule of thumb, the bandwidths would be  $h_1 = 0.9(0.4 \times 300)^{-1/5} \approx 0.35$ , and  $h_2 = 0.9 \times 5 \times (0.6 \times 300)^{-1/5} \approx 1.593$ , for the first and second component respectively. The npEM algorithm used  $\hat{h} = 1.932$  for a single bandwidth, and  $\hat{h}_{11} = 0.266$  and  $\hat{h}_{21} = 1.246$  when allowing different bandwidths. The left-hand plot of Figure 1 shows that the large value  $\hat{h}$  results in an over-smoothed density estimate for the first component. Allowing separate bandwidth estimates  $\hat{h}_{11}$  and  $\hat{h}_{21}$ , as in the right-hand plot of Figure 1, gives better results.

In Figure 1, we see from the legends that the final estimates of  $\lambda_1$  and  $\lambda_2$  are the same for the two algorithms, i.e.,  $\hat{\lambda}_1 = 0.390$  and  $\hat{\lambda}_2 = 0.610$ . This is not guaranteed to be the case: Since the  $\lambda$  estimates are ultimately functions of the density estimates in the npEM algorithm, the original (fixed, single  $h$ ) npEM algorithm can lead to different estimates than the modified algorithm we introduce here. However, the particular example we have chosen involves mixture components that are "well-separated" in the sense that it is fairly easy to assign each observation to one component or the other. Thus, the final estimate of  $\lambda_1$  is simply the proportion of observations that are classified as belonging to the first component, which is 117/300 in this example. We may observe that this is true by noting that all of the posterior probabilities  $p_{ij}$  at the final iteration are very close to zero or one.

Other examples, extensions (e.g. to semiparametric cases) and R code for running various versions of this algorithm with the `mixtools` package (Young



**Fig. 1.** Left: estimates from the npEM algorithm using the same bandwidth; Right: estimates using the modified npEM algorithm with adaptive bandwidths (10).

et al., 2009) can be found in Benaglia et al. (2009b). An empirical study of the rate of convergence of the original npEM algorithm is also provided in Benaglia et al. (2009a). The convergence properties of this EM-like family of algorithms is an ongoing work.

## References

- ALLMAN, E.S., MATIAS, C. and RHODES, J.A. (2008): *Identifiability of Latent Class Models with Many Observed Variables*, Tech. report arxiv 0809.5032v1.
- BENAGLIA, T., CHAUVEAU, D. and HUNTER, D. R. (2009a): An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures, *J. Comput. Graph. Statist.* 18 (2), 505-526.
- BENAGLIA, T., CHAUVEAU, D. and HUNTER, D. R. (2009b): Bandwidth Selection in an EM-like algorithm for nonparametric multivariate mixtures, Technical report.
- BORDES, L., CHAUVEAU, D., and VANDEKERKHOVE, P. (2007): A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics and Data Analysis* 51 (11), 5429-5443.
- ELMORE, R. T., HETTMANSPERGER, T. P., and THOMAS, H. (2004): Estimating component cumulative distribution functions in finite mixture models, *Communications in Statistics: Theory and Methods* (33), 2075-2086.
- HALL, P., NEEMAN, A., PAKYARI, R., and ELMORE, R. (2005): Nonparametric inference in multivariate mixtures, *Biometrika*, 92, 667-678.
- MCLACHLAN, G. and PEEL, D. A. (2000): *Finite Mixture Models*, New York: Wiley.

- R DEVELOPMENT CORE TEAM (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- SILVERMAN, B. W. (1986): *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability, Chapman & Hall.
- THOMAS, H., LOHAUS, A., and BRAINERD, C.J. (1993): Modeling Growth and Individual Differences in Spatial Tasks, *Monographs of the Society for Research in Child Development* 58 (9), 1–190.
- YOUNG, D. S., BENAGLIA, T., CHAUVEAU, D., and HUNTER, D. R. (2009), mixtools: An R Package for Analyzing Mixture Models, *Journal of Statistical Software* 32 (6), 1–29.



# On the use of Weighted Regression in Conjoint Analysis

Salwa Benammou<sup>1</sup>, Besma Souissi<sup>2</sup>, and Gilbert Saporta<sup>3</sup>

<sup>1</sup> Faculté de Droit et des Sciences Economiques et Politiques, Sousse, Tunisie,  
*saloua.benammou@fdseps.rnu.tn*

<sup>2</sup> Institut Supérieur de Gestion, Sousse, Tunisie, *besma.swissi@yahoo.fr*

<sup>3</sup> Chaire de statistique appliquée & CEDRIC, CNAM  
292 rue Saint Martin, Paris, France, *gilbert.saporta@cnam.fr*

**Abstract.** Conjoint analysis seeks to explain an ordered categorical ordinal variable according to several variables using a multiple regression scheme. A common problem encountered, there, is the presence of missing values in classification-ranks. In this paper, we are interested in the cases where consumers provide a ranking of some products instead of rating these products (i.e. explained variable presents missing values). In order to deal with this problem, we propose a weighted regression scheme. We empirically show (in several cases of weighting) that, if the number of missing values is not too large, the data remain useful, and our results are close to those of the complete order. A simulation study confirms these findings.

**Keywords:** conjoint analysis, missing values, weighted regression

## 1 Introduction

The conjoint analysis is a data analysis method. It links an explained categorical ordered variable to several explanatory variables either or unordered categories. It allows analyzing consumers' preferences for products defined by combinations of attributes, according to these last ones.

Initially developed by psychometrics, the conjoint analysis has been introduced in the marketing research field at the beginning of 1970s (Green and Rao 1971). Its use knew a considerable development at the end of 1970s and in the 1980s (Wittink and Cattin on 1989). The conjoint analysis is a complete methodology composed of three phases. The first one, based on experimental design, consists in collecting observations, generally, by direct interviews where each interviewee evaluates a set of real or hypothetical products.

The second phase corresponds to data processing and the parameters' estimation. The conjoint analysis decomposes then the preferences according to a model of additive utility which is specific for every interviewee. The last phase is dedicated to the simulation of market shares (Benammou et al. 2007).

We are interested here in the phase of treatment, and thus in the estimation of the parameters. When a consumer has to classify by order of preference a set of products, conjoint analysis is a particular case of the ordinary linear

model.

Generally, the estimation is done by ordinary least square method. It supposes that the consumer gives the same weight to all products and that the “distance” between two products of successive ranks is the same for all the ranks of classification. This hypothesis seems plausible in the case of the total order.

But, in actual fact, the cognitive capacity of the consumer decreases when the number of scenarios increases and the consumer tends to classify only the most favorite products. It is easier to imagine two or three products and to classify them rather than 10 or 12 products. At the end, we obtain missing ranks in the classification. Benammou et al (2003) show upon an example a good stability of the results when the missing values don’t exceed half of the classified scenarios.

In some cases the distance between ranks of classification can be different. The difference between the second classified product and the first one is much smaller than the one between the last product and the next to last before. Thus we propose the use of weighted least square method to model this behavior.

Furthermore, if we suppose that the consumer gives intuitively more (respectively less) importance to the most (respectively least) preferred products, it would then be logical to give a higher weight (respectively weak) to the first classified products (respectively the last ones).

We propose the use of decreasing weight functions; what give weak weights for the last classified ranks. These functions seem to better describe the behavior of the consumer. The use of fast decreasing weight functions can give an alternative solution to the problem of partial classification of products, especially when the number of non classified product is important.

## 2 Reminder on weighted regression

Let’s consider  $q$  products described by  $p$  qualitative variables  $X_1, X_2, \dots, X_p$  in  $m_1, m_2, \dots, m_p$  categories respectively. Generally, the associated linear model is given by (1)

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \quad (1)$$

Where  $\beta$  is the vector of parameters to estimate,  $\mathbf{y}$  the vector of classification ranks,  $\mathbf{X}$  the experimental matrix,  $(q, \sum_{i=1}^p m_i)$  the size of  $\mathbf{X}$ , and  $\mathbf{e}$  the random errors vector associated such as  $Var(\mathbf{e}) = \sigma^2 \Sigma$ .

Here  $\Sigma$  is the errors variance covariance matrix which equals the identity in the case of a classic linear model. The value of the Generalized Least Squares (GLS) estimator is then given by the relation (2)

$$\hat{\beta} = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} (\mathbf{X}' \Sigma \mathbf{y}) \quad (2)$$

For the weighted least squares (see for example Weisberg , 1985) every coordinate of errors vector  $e$  is correlated to all the others; but the variances can not be the same and the matrix  $\Sigma$  is then in the form (3).

$$\Sigma = \begin{bmatrix} W_1 & 0 & \cdots & 0 \\ 0 & W_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & W_q \end{bmatrix} \quad (3)$$

Where  $W_i > 0$  and  $\sum_{i=1}^q W_i = 1$ .

Let us suppose that a diagonal matrix  $\mathbf{C}$  such as  $\Sigma$  exists. The matrix  $\mathbf{C}$  is then the “square root” of the inverse of the matrix  $\Sigma$  and we have  $Var(\mathbf{C}e) = \sigma^2 \mathbf{I}$ . By an appropriate transformation of variables, we can recover the common linear model.

And so if we suppose  $\mathbf{y}_w = \mathbf{C}\mathbf{y}$ ,  $\mathbf{X}_w = \mathbf{C}\mathbf{X}$  and  $\mathbf{e}_w = \mathbf{C}e$ , the model becomes  $\mathbf{y}_w = \mathbf{X}_w\beta + \mathbf{e}_w$  and we recover the same shape as described by the relation (1). This model verifies all the hypotheses required by the classical linear model and we can use the ordinary least squares method to estimate its parameters (relation (4)).

$$\hat{\beta} = (\mathbf{X}_w' \mathbf{X}_w)^{-1} (\mathbf{X}_w' \mathbf{y}_w) \quad (4)$$

### 3 Exemple

#### 3.1 The data

To be able to compare our results with those of the literature, we take back the data of the example treated by Benammou et al (2003), relatives to 263 consumers classifying scenarios of mobile phone subscriptions.

TABLE 1. List and labels of variables

Label	Code	Categori	
Device price (in euro)	Device price	0	100
Subscription fee (in euro)	Sub fee	0	30
Peak hours definition	Peak hours def	$p_1$	$p_2$
Duration of subscription (in month)	Duration	6	24
Monthly subscription price (in euro)	Monthly sub	0	4, 5   9
Price/minute peak hours (in euro)	Peak hours	0, 5	0, 7   0, 9
Price/minute off – peak hours (in euro)	Off – peak hours	0, 008	0, 1   0, 15

The scenarios emanate of seven variables, where four variables are in two categories and three in three categories and the number of parameters linearly independent to be considered equals 10. The design used is a D-optimal one with 12 scenarios. The obtained products who presented to 263 consumers whom answered by a total ranking. We give variables description in table 1.

### 3.2 The results and their interpretations

#### 3.2.1 Studies of $R^2$

Goodness of fit is measured by multiple correlation coefficient  $R^2$ . This coefficient is an indicator of the quality of adjustment of the model. When this coefficient is low it denotes an inadequacy of the model, or an incoherence of interviewee's responses. Generally the second hypothesis is retained since the model seems realistic. We should eliminate interviewee for whom  $R^2$  is less than a critical value fixed by user. (Benammou et al 2003). In this example, Benammou et al (2003) showed that  $R^2$  is close to 1 for almost all individuals, in the total order case. The authors proposed three simple procedures for estimating missing values. The first one consists in attributing to all the non classified products the rank of the last classified product increased by 1. In the second one, all the non classified products receive the average of missing ranks. For the third one, all the non classified products receive the maximum rank. The results given by the three procedures being equivalent, we give here those of the first one only.

The  $R^2$  of the weighted linear model in the case of total order (see Fig. 1.) shows that they are better than those of the classic linear model. The adjustment quality improves with the decreasing speed of weight function. For example in the case of the functions  $f(x)=e^{-2x}$  and  $f(x)=(\frac{1}{2})^x$  the  $R^2$  is very close or equal to 1 for the majority of the individuals.

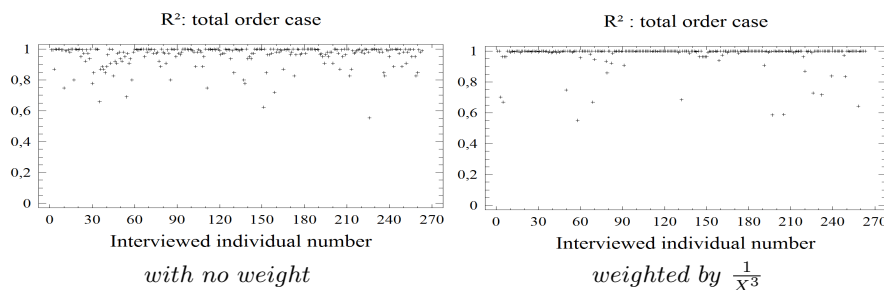


Fig. 1.  $R^2$  in the case of total order with and with no weight functions

### 3.2.2 Study of the individual utilities

Benammou et al ( 2003 ) showed that the individual utilities remain stable when the number of missing values does not exceed six -that is half of the classified scenarios. The use of the weight functions improves considerably this result (see Fig. 2.). We notice that in the case of the weight functions of the form  $f(x) = \frac{1}{(X)^n}$  and when the number of missing values equals 8 the correlation between individual utilities in the case of total order and the partial orders is  $\geq 0,8$ . It increases slightly with the decreasing speed of the weight function and stabilizes when  $n$  exceeds 4. For the same functions when  $n > 2$  the results remain useful with 9 missing values (the correlation between individual utilities reaches 0.9 for some factor levels).

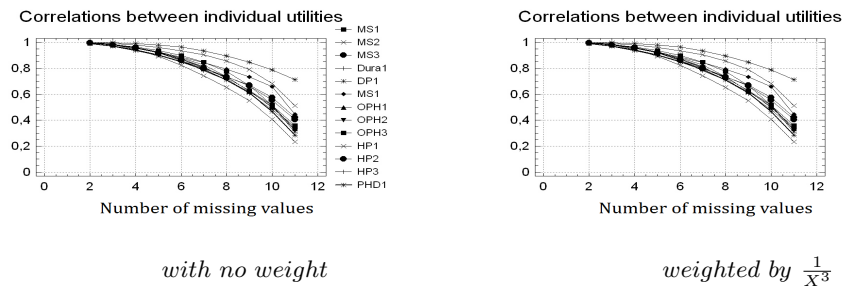


Fig. 2. Correlation between individual utilities in case of total order and various partial orders with and without weight functions

### 3.2.3 Study of the importance's of the utilities

Benammou et al (2003) showed that the importance of the factor utilities remain stable when the number of missing values does not exceed six. The use of the weight functions improves slightly this result (see fig. 3.). The importance of factors remains stable when the number of missing values is not very important (about 7). This stability degrades when this number increases. This is due to a strong correlation between these importances (and those of the total order case).

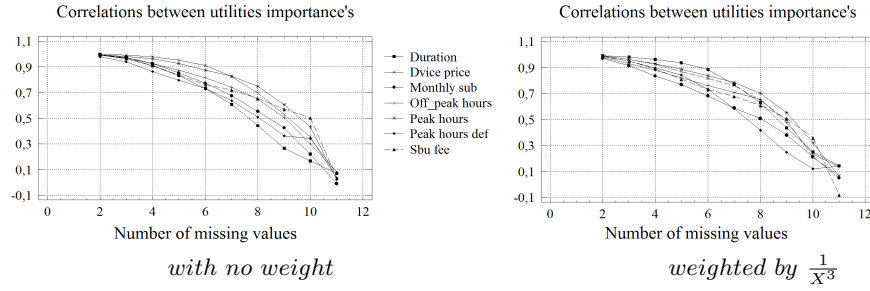


Fig. 3. Correlation between importance's utilities in case of total order and various partial orders with and without weight functions

## 4 Simulation

To generalize our results and to be able to compare them with the existing literature, we conduct a simulation study analogous to that used by Benammou et al (2003), where the authors have simulated ranks of classifications in a systematic way. To have a realistic model, we make choice of an utilities system and simulate ranks that are compatible and coherent adding noises  $\varepsilon$  with a given standard deviation to  $X$   $\beta$ .

### 4.1 $R^2$ study

The simulation of the classification rank is based on the coefficients of a real data model. This is done in order to guarantee the coherence of the simulated data with a multiple correlation coefficient close to 1. We give in Fig.4. the values of  $R^2$  for various values of  $\sigma^2$  for the weight function  $\frac{1}{X^3}$  which seems to give the best adjustment.

We point out that  $R^2$  values are close to 0,99 for the majority of the individuals. Other functions such as  $f(x) = \frac{1}{2^x}$ ,  $f(x) = \exp(-2x)$  or  $f(x) = \exp(\frac{1}{x})$  give comparable results.

### 4.2 Analysis of the individual utilities

The individual utilities remain stable even for an important number of missing values (about 8). This stability decreases when the number of missing values increases (see Fig. 5.). The results improve slightly when  $\sigma^2$  increases.

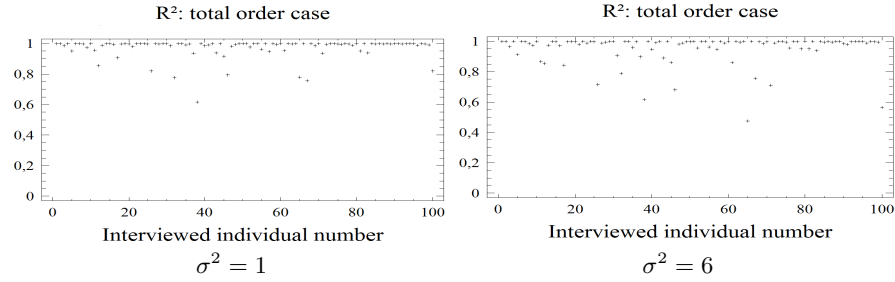


Fig. 4.  $R^2$  in the case of total order for different values of  $\sigma^2$  weighted by  $\frac{1}{X^3}$

The use of weight functions yields superior results to those obtained by Benammou et al (2003). As an example, we give, in Fig. 5. the correlations between individual utilities for various values of  $\sigma^2$  in the case of the weight functions  $\frac{1}{X^3}$ .

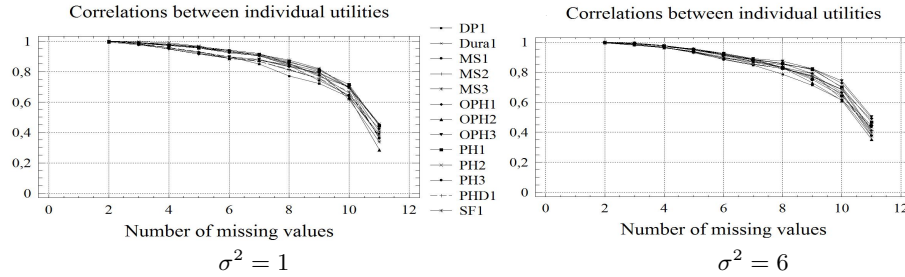


Fig. 5. Correlation between individual utilities in case of total order and various partial orders for different values of  $\sigma^2$  weighted by  $\frac{1}{X^3}$

### 4.3 Analysis of the importance's utilities

Importance's utilities remain stable for a large number of missing values (about 8). This stability decreases when the number of missing values increases (see Fig. 6.). The results improve slightly when  $\sigma^2$  increase. The use of the weight functions improves the results obtained by Benammou et al (2003). As an example, we give (Fig. 6.) the correlations between the importances of the utilities for various values of  $\sigma^2$  weighted by  $\frac{1}{X^3}$

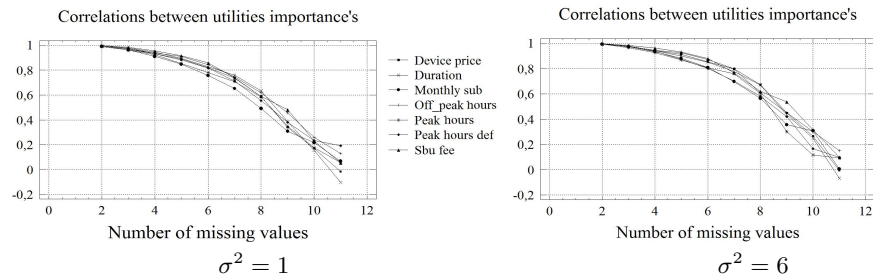


Fig. 6. Correlation between importance's utilities in case of total order and various partial orders for different values of  $\sigma^2$  weighted by  $\frac{1}{X^3}$ .

## 5 Conclusion

The case of partial ordering in conjoint analysis, is very frequent especially when the number of proposed product exceeds ten. In this paper, we propose the use of the weighted least squares to estimate the parameters. Our experimentations showed a good stability of the results under the three quarter ranked scenarios. We confirm our findings by simulation. It should be remarked that the results are better when the weight functions decreasing speed is faster.

## References

- Benammou, S., Harbi, S. and Saporta, G.(2003): Sur l'utilisation de l'analyse conjointe en cas de réponses incomplètes ou de non réponses. *Revue de Statistique Appliquée*, 51, 31-55.
- Benammou, S. , Saporta, G and Souissi, B.(2007): Une procédure de réduction du nombre de paires en analyse conjointe. *Journal de la Société Française de Statistique*, 148, (4).57- 76.
- Cattin, P. and Wittink, D.R. (1989): Commercial Use of Conjoint analysis: An Update. *Journal of Marketing*, 53, .91-96.
- Green, P.E. and Srinivasan, V. (1990): Conjoint analysis in marketing: new developments with implications for research and practice. *Journal of Marketing*, 3-19.
- Green, P.E. and Rao, V.R.(1971): Conjoint Measurement for Quantifying Judgment Data. *Journal of Consumers Research*, 5 September, 103-123.
- Weisberg, S. (1985): *Applied linear regression*. John Wiley and Sons, inc, second edition, New York.



# Wavelet-PLS Regression: Application to Oil Production Data

Salwa Benammou<sup>1</sup>, Kacem Zied<sup>1</sup>, Hedi Kortas<sup>1</sup>, and Dhifaoui Zouhaier<sup>1</sup>

<sup>1</sup> Computational Mathematical Laboratory, *saloua.benammou@yahoo.fr*

<sup>2</sup> *ZiedKacem2004@yahoo.fr*

<sup>3</sup> *kortashedi@yahoo.fr*

<sup>4</sup> *dh.zouhaier@yahoo.fr*

**Abstract.** This paper is devoted to the study of PLS regression in the presence of noise that could affect the quality of the results. To solve this problem, we suggest a hybrid approach which combines PLS regression and wavelet-based thresholding techniques. The proposed method is validated via a simulation study and subsequently applied to petroleum data. Empirical results show the relevance of the selected approach and contribute to a better modelling of the series of study.

**Keywords:** PLS regression, thresholding, minimax, wavelet-PLS

## 1 Introduction

In numerous data analysis applications, statisticians are confronted with several problems such as missing or incomplete data, the presence of a strong collinearity between the explanatory variables or the case where the number of variables exceeds the number of observations. To cope with these problems, several statistical approaches have been developed, among them, a data analysis method initially proposed by Wold and al. (1983). It is known as Partial Least Squares (PLS) regression.

Although PLS regression has proven to be of great performance in a wide range of applications, the model variables are usually corrupted by noise which may adversely affect the results drawn from the PLS regression in terms of modelling and prediction accuracy (Tenenhaus and al. (1995)).

To deal with this problem, we discuss, in this paper, a hybrid data analysis method based on the combination of wavelet thresholding techniques and PLS regression.

## 2 The Wavelet-PLS method

The Wavelet-PLS regression entails several steps. As a first step, we pre-process the variables in the following manner: if the explanatory data vectors are not of dyadic lengths (i.e. powers of 2), we extend the data samples by applying a so called "zero-padding" method. This method consists in adding

zeros to the beginning and/or end of each time domain sequence in order to attain the next dyadic length. This pre-processing step is needed for the implementation of the Discrete Wavelet Transform (DWT) algorithm (Mallat (2000)) which requires a dyadic length time series. It should be stressed here that the wavelet coefficients relating to the zero-padding operation are subsequently eliminated while performing the Inverse Discrete Wavelet Transform (IDWT) signal reconstitution.

In the second stage, the DWT is performed to the exogenous variables. This requires a precise choice of the wavelet system to be used. In this work, we rely on Daubechies wavelet bases possessing attractive properties such as vanishing moments, orthogonality and especially support compactness which results in significant computational gains (Daubechies (1992)).

In the third step, the obtained wavelet coefficients are thresholded by means of the wavelet-based denoising techniques (Donoho and Johnstone (1998)). In the fourth step, we carry out an IDWT to reconstruct the set of explanatory variables which are now practically noise-free. Finally, the conventional PLS regression is applied to the new set of regressors. The Wavelet-PLS approach is illustrated in Fig.1:

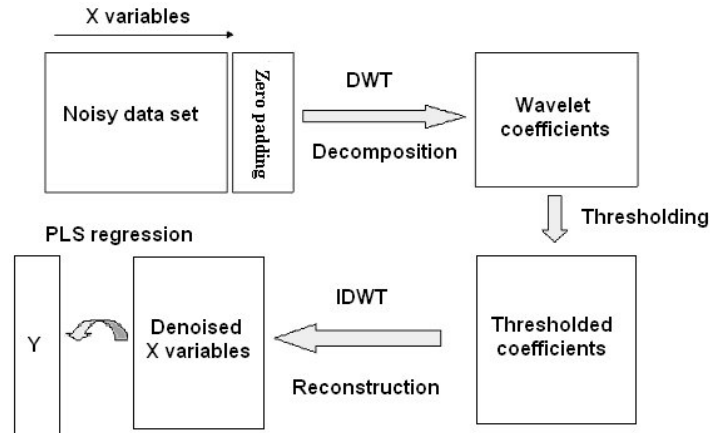


Fig. 1. Overview of the Wavelet-PLS regression.

### 3 Application

In order to assess the relevance of the Wavelet-PLS regression scheme, we consider a real world data set. The response (dependent) variable  $Y$  represents the crude oil (petroleum) production in barrels denoted "oil" in a given oil field composed of four wells. The data measurements are made on a daily basis during the period from May 1, 2003 to March 31, 2006 thus totalling 1024 observations. Here the response variable  $Y$  depends on 16 explanatory

variables corresponding to the features of the wells. The independent variables are:

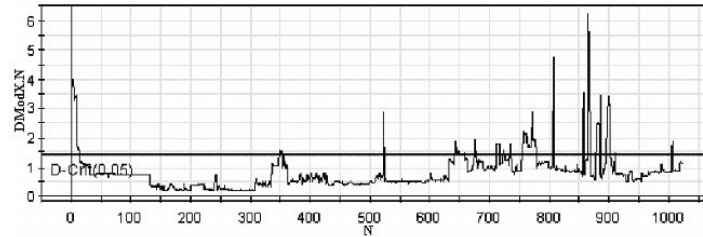
- (Choke  $i$ ),  $i = 1, \dots, 4$ : the choke valve position in the oil well  $i$ ,  $i = 1, \dots, 4$ . This variable takes integer values ranging from 1 to 64. In fact, the choke valve is variably positionable defining 64 incremental positions allowing to regulate the flow rate.
- (FTHP  $i$ ),  $i = 1, \dots, 4$ : Flowing Tubing Head Pressure of the well  $i$  (in bars). Actually, oil extraction is assured by the difference between the underground pressure in the oil reservoir and the pressure at the top of the well. The pressure at the top of the well, which is the extraction pressure, is a key parameter and is defined as the Flowing Top Head Pressure (FTHP).
- (Pressure at Choke  $i$ ),  $i = 1, \dots, 4$ : pressure on the level of the choke in the well  $i$  (in bars).
- (WC  $i$ ),  $i = 1, \dots, 4$ : (Water cut) Percentage of water. It is the ratio of water produced to the volume of total liquids extracted from the well  $i$ .

#### 4 PLS Regression on the raw data set

Using the cross validation technique, we retrain a PLS model with four components. The regression of  $y$  on  $t_1, t_2, t_3$  and  $t_4$  gives the following equation:

$$\hat{y} = (0.29716)t_1 + (0.199871)t_2 + (0.34454)t_3 + (0.167767)t_4$$

The normalized distances to the model in the  $X$  space are reported on Fig.2. Remember that the observations with  $ND_{modX}$  exceeding the critical limit



**Fig. 2.** Normalized distances to the model  $ND_{modX}$ .

at the 95% are regarded as outliers in the  $X$  space.

It is remarkable to note here that 98 observations i.e. 9.6 % of the total sample are regarded as outliers.

## 5 Wavelet-PLS regression results

### 5.1 Wavelet-based denoising

In order to eliminate the noise from the set of predictors, we first apply a DWT curtailed at the resolution level  $j = 5$  to each exogenous variable using the  $D(8)$  Daubechies compactly supported wavelet. The DWT results in five levels of wavelet detail coefficients and a single level approximation coefficients. Next, the obtained detail coefficients are subject to a wavelet thresholding operation. In our case, we opt for a soft thresholding function. The choice of the threshold value is done according to the Minimax procedure. This is due to the fact that the raw variables' series exhibit several discontinuities and abrupt changes. The Minimax procedure is well adapted for handling such features. It should be noted that associating the soft thresholding and the Minimax criterion defines the so called "Risk-Shrink" procedure (Donoho and Johnstone (1994)).

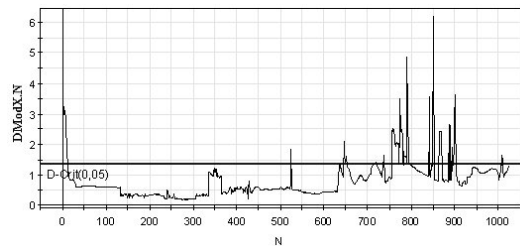
### 5.2 Wavelet-PLS regression on the denoised variables

In the following, we carry out a PLS regression using the obtained denoised regressors. According to the cross validation estimation results, we choose to retain four PLS components. The regression equation is then given by:

$$\hat{y} = (0.29816)t_1 + (0.165487)t_2 + (0.261839)t_3 + (0.319801)t_4$$

Estimation results for the Wavelet-PLS regression show a slight improvement for the determination coefficient with an  $R^2$  value of 0.919 compared to  $R^2 = 0.916$  for the PLS model performed on the raw data set.

Another interesting remark is that the Mean Square Error has decreased by 10.3% when carrying out a wavelet thresholding on the regressors' vectors. Fig.3 report the  $NDmodX$  values for the Wavelet PLS procedure. Observe



**Fig. 3.** Normalized distances to the model  $NDModX$  for the Wavelet-PLS regression.

the effect of noise removal on the regression results.

The obtained results show that 90 observations among 1024 are outliers representing 8.7 % of the total sample size. Thus we can state that the denoising procedure has reduced the number of outliers yielding better modelling results.

## 6 Simulation

In this simulation study, we apply the Wavelet-PLS regression procedure to multidimensional fractional Brownian motions with noisy components.

In order to present the  $n$ -dimensional fractional Brownian, we restrict ourselves to the simple case  $n = 2$ . The generalization is straightforward.

The process  $B_t = (B_t^1, B_t^2)^T$  is a correlated 2-dimensional fractional Brownian motion if:

- The increments  $B_1(t) - B_2(t)$  et  $B_2(t) - B_2(s), t > s$  are independent of  $B_1(y)$  and  $B_2(y), \forall 0 \leq y \leq s$ .
- $cov(B_1(t), B_2(t)) = E(B_1(t)B_2(t)) = \rho t$  where  $-1 \leq \rho \leq 1$ . This implies that:  $corr(B_1(t), B_2(t)) = \rho$ . Besides, we have:  $\forall t \neq s, cov(B_1(t), B_2(s)) = \rho \min(t, s)$ .

In this section, we synthesize 55 realisations of an  $n$ -dimensional fractional Brownian  $B_t = (B_t^1, B_t^2, \dots, B_t^n)^T$  with correlated components. Actually, this is a positive definite matrix whose elements are formally given by:  $corr(B_t^{(i)}, B_t^{(j)}) = \rho_{ij}$  where  $\rho_{ij}$  is the  $(i, j)^{th}$  entry of the matrix of  $\Sigma$ .

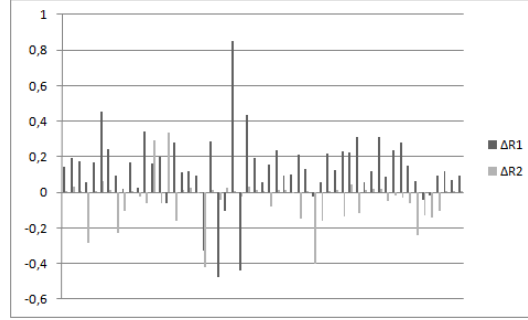
Issues related to construction and simulation of the multidimensional fractional Brownian motion are treated in details in the works of Haugh (2004) and Glasserman (2003).

The simulation study can be divided into three steps: In the first step, we apply the PLS regression scheme to the 55 realizations of the multidimensional fractional Brownian motion process. In the second phase, we add noise components to the simulated trajectories and we apply the PLS regression to the noisy dataset. It should be stressed here that the amount of the simulated noise is pre-specified by the "signal to noise ratio". This is the ratio of a signal power  $P_S$  to the noise power  $P_N$  present in the signal. Formally the SNR is defined as:

$$SNR = 10 \log_{10} \left( \frac{P_S}{P_N} \right)$$

We impose 17 different levels of additive Gaussian noise to the components of the simulated multidimensional fractional Brownian motions. It should be remarked here that choosing different SNR levels allows us to assess the robustness of the denoising technique to a change in the noise amplitude. The third step consists in applying the Wavelet-PLS method to the 55 noisy matrices so as to test the relevance of the proposed scheme.

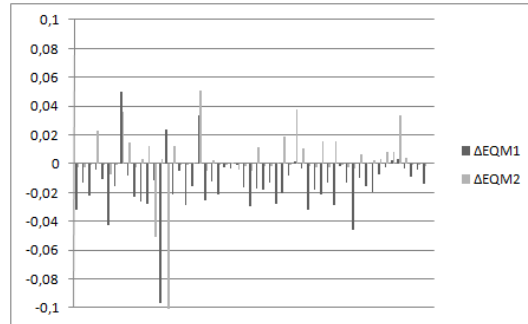
Fig.4 shows the values of  $\Delta R_1^2$  which are the differences between the goodness of fit values for the PLS regression performed on the initial dataset and those



**Fig. 4.** Normalized distances to the model NDMoX for the Wavelet-PLS regression.

of the PLS model performed on the noisy data. On the same figure, we have also plotted associated with simulated cases provided by the the values of  $\Delta R_2^2$  which are the differences between the  $R^2$  values for the PLS regression performed on the initial dataset and those of the PLS model performed on the denoised data. ( $\Delta R^2$ ) before and after introduction of noise

Remark that, overall, the  $\Delta R_1^2$  values are much closer to zero than the  $\Delta R_2^2$ . This shows the effectiveness of the wavelet techniques for noise removal.



**Fig. 5.** ( $\Delta MSE$ ) before and after introduction of noise

Fig.5 shows the values of  $\Delta MSE_1$  which are the differences between the Mean Squared Errors (MSE) for the PLS regression performed on the initial dataset and those of the PLS model performed on the noisy data. For the sake of comparison, we have also plotted the values of  $\Delta MSE_2$  which are the differences between the MSE values for the PLS regression performed on the initial dataset and those of the PLS model applied to the denoised data.

In view of these results, it is clear that the  $\Delta MSE_2$  are much smaller than those of  $\Delta MSE_1$ . This confirms the relevance of the Wavelets-PLS method.

## 7 Conclusion

In this work, a novel data analysis method has been proposed and discussed. It consists of utilizing wavelet based thresholding techniques in association with PLS regression.

By applying the Wavelet-PLS approach to oil production data sets, we were able to improve the modelling performance of the PLS regression model. Indeed, we succeeded in:

- diminishing the number of outliers
- reducing the Mean Square Error
- correcting the observations in the score plot
- ameliorating the model goodness of fit ( $R^2$ )

## References

- AMINGHAFARI M., CHEZE N. and POGGI J.M. (2006). Multivariate denoising using wavelets and principal component analysis. *Computational Statistics & Data Analysis* 50 (9), 2381-2398
- DAUBECHIES I. (1992). *Ten lectures on wavelets*, SIAM, Philadelphia
- DONOH O. D. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. theory*, 41 (3), 612-627.
- DONOH O. D. and JOHNSTONE I. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* 26 (3), 879-921.
- DONOH O. D. and JOHNSTONE I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425-455.
- MALLAT S. (2000). *Une exploration des signaux en ondelettes*. Les éditions de l'école Polytechnique, Ellipses edition.
- HAUGH M. (2004). "The Monte Carlo Framework, Examples from Finance and Generating Correlated Random Variables". Course Notes. [www.columbia.edu/mh2078/MCS04/MCS\\_framework\\_FEEgs.pdf](http://www.columbia.edu/mh2078/MCS04/MCS_framework_FEEgs.pdf).
- GLASSERMAN P. (2003). *Monte Carlo methods in financial engineering*. Springer-Verlag.
- TENENHAUS M. (1998). *La régression PLS : Théorie et Pratique*. Technip, Paris.
- TENENHAUS M. (1995). *Nouvelle Méthodes de Régression PLS*. Les cahiers de recherche, CR540.
- TENENHAUS M., GAUCHI J. P. and MENARDO C. (1995). Régression PLS et Applications. *Revue de Statistique Appliquée*, (3), 7-63.
- VIGNERON V., PARASCHIV-IONESCU A., AZANCOT A., JUTTEN C. and SIBONY O. (2003). Fetal electrocardiogram extraction based on non-stationary ICA and wavelet denoising. *Seventh IEEE International Symposium on Signal Processing and its applications*, (2), 69-72.
- WOLD S., MARTENS H. and WOLD H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In *Proc. Conf. Matrix Pencils, Ruhe A. and Kastrom B. (Eds), March 1982, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, 286-293*.





# Variable Selection and Parameter Tuning in High-Dimensional Prediction

Christoph Bernau<sup>1</sup> and Anne-Laure Boulesteix<sup>1,2</sup>

<sup>1</sup> Department of Medical Informatics, Biometry and Epidemiology  
University of Munich, Marchioninstr. 15, 81377 Munich, Germany  
*bernauc@ibe.med.uni-muenchen.de, boulesteix@ibe.med.uni-muenchen.de*

<sup>2</sup> Department of Statistics  
University of Munich, Ludwigstr. 33, 80539 Munich, Germany

**Abstract.** In the context of classification using high-dimensional data and nested cross-validation, it is still unclear how variable selection and parameter tuning should be combined if a classification method involves both variable selection and parameter tuning. It is well-known that variable selection should be repeated for each *external CV* iteration. However, should we also repeat variable selection for each *internal CV* iteration or rather perform tuning based on a fixed subset of variables? While the first strategy seems more adequate, it implies a huge computational expense and its benefit in terms of error rate remains unknown. In this paper, we assess both strategies quantitatively using real microarray data sets.

**Keywords:** class prediction, variable selection, parameter tuning, nested cross-validation, genomics

## 1 Background

In the context of classification using high-dimensional data such as microarray gene expression data, it is often useful to perform preliminary variable selection. For example, the  $k$ -nearest-neighbors classification procedure yields a much higher accuracy when applied on variables with high discriminatory power. Typical (univariate) variable selection methods for binary classification are, e.g., the two-sample t-statistic or the Mann-Whitney test.

In small sample settings, the classification error rate is often estimated using cross-validation (CV) or related approaches. From now on, we denote the whole data sample as  $S$ , the CV folds as  $T_1, \dots, T_J$  (with  $\cup_{j=1}^J T_j = S$  and  $T_{j_1} \cap T_{j_2} = \emptyset$  for  $j_1 \neq j_2$ ), and the corresponding CV learning sets as  $L_j = S \setminus T_j$ , for  $j = 1, \dots, J$ . If the chosen classification method involves a preliminary variable selection step, this step has to be applied for each CV iteration anew, i.e. for each considered learning set  $L_j$  successively. Performing variable selection based on the whole sample  $S$  before the CV procedure would yield a downwardly biased error rate estimate (Slawski et al., 2008).

CV may also be used to tune parameters involved in a classification method. For instance, the penalty in penalized regression is most often selected using CV. This type of CV is usually denoted as "internal CV" in

contrast to the "external CV" performed to estimate the error rate, while the term "nested CV" refers to the whole procedure embedding two CV loops. Similarly to the external CV, we denote the internal fold for the  $j$ th external CV iteration and  $i$ th internal CV iteration as  $T_{ij}$  and the corresponding learning sets as  $L_{ij} = L_j \setminus T_{ij}$ , for  $i = 1, \dots, I$  and  $j = 1, \dots, J$ . See Varma and Simon (2006) for a recent reference on cross-validation for tuning.

While variable selection and CV have been widely investigated in the context of high-dimensional classification, it is still unclear how variable selection and parameter tuning should be combined if a classification method involves both variable selection and parameter tuning. For example, the  $k$ -nearest-neighbors method usually requires variable selection and involves a tuning parameter: the number  $k$  of neighbors. Should we also repeat variable selection for each *internal CV* iteration based on  $L_{ij}$  or rather perform internal CV with the fixed subset  $\mathcal{V}_j$ , as implemented in the R package 'MCReestimate' (Ruschhaupt et al. (2004))?

In the latter variant, termed "V1" from now on, variables are selected based on the learning set  $L_j$  corresponding to the current external CV iteration. Hence, the well-known rule that variable selection should be performed without taking the test set into account is violated in the internal CV. This is because the subset of variables  $\mathcal{V}_j$  is derived from  $L_j = L_{ij} \cup T_{ij}$  that, from the point of view of internal CV, includes both the learning data set  $L_{ij}$  and test data set  $T_{ij}$ , for  $i = 1, \dots, I$ . As outlined above, variable selection can also be repeated for each internal CV iteration based on  $L_{ij}$  only, yielding variant "V2". In this case, the variable subset varies in each internal CV iteration. We denote the subset of variables selected in the  $i$ th internal CV iteration and  $j$ th external CV iteration as  $\mathcal{V}_{ij}$ .

In variant V1, the error rates computed in the internal CV are expected to be lower than 50% even if the class membership  $Y$  is random. That is because the variable subset  $\mathcal{V}_j$  is chosen to be associated with  $Y$  in  $L_j$ , for each  $j = 1, \dots, J$ . In other words, V1 performs tuning based on downwardly biased estimates of the error rate. The relative performance of parameter values in internal CV - and thus the result of the tuning procedure - may also be affected by the fact that the internal test data sets  $T_{ij}$  were not disregarded while selecting the subset  $\mathcal{V}_j$ .

In this sense, variant V2 seems more natural and adequate. However, it implies a higher computational expense and its benefit over V1 in terms of error rate remains unknown. An additional potential pitfall is that the variables that are used for tuning (i.e. that are selected at each internal CV iteration) are different from those that are eventually used to construct the classifier. For example, if we perform penalized regression based on variables of different scale, it would obviously be wrong to use a penalty parameter that was chosen based on other variables. This example is probably exaggerated and in this case the problem can be simply solved through appropriate scaling, but more subtle similar mechanisms may in general affect the accuracy of V2.

To our knowledge, V1 and V2 are both used in practice – most often rather implicitly and without much explanation. Their respective merits and pitfalls remain largely unexplored in the literature, although tuning issues are known to greatly affect accuracy in general. In this paper, we assess both variants V1 and V2 quantitatively using real microarray data sets. We focus on two representative examples:  $k$ -nearest-neighbors (with  $k$  as tuning parameter) and Partial Least Squares dimension reduction followed by linear discriminant analysis (with the number of components as tuning parameter). More precisely, we address the following questions: 1) Do V1 and V2 select the same parameter values? and, if yes, 2) Do the resulting classification accuracies differ substantially?

## 2 Methods and design of the study

### 2.1 Classification and variable selection methods

In this paper, V1 and V2 are compared for two completely different standard classification methods: the  $k$ -nearest-neighbors (kNN) algorithm with the number  $k$  of neighbors as tuning parameter, and Partial Least Squares (PLS) dimension reduction followed by linear discriminant analysis (PLS+LDA), with the number  $ncomp$  of PLS components as a tuning parameter. We refer to Boulesteix (2004) for details on PLS+LDA. For the purpose of reproducibility, we use the standardized implementations provided by the 'CMA' Bioconductor package (Slawski et al. (2008)). We consider the classical candidate parameter values  $k = 1, 3, 5, 7, 9$  for kNN and  $ncomp = 1, 2, \dots, 10$  for PLS+LDA.

Tens of variable selection criteria have been proposed in the context of microarray-based (binary) classification. In this study, we focus on two univariate methods. The first one selects the  $p^*$  genes with the highest absolute value of the two-sample t-statistic. The second one is the criterion provided by the Recursive Feature Elimination (RFE) approach by Guyon et al. (2002) based on support vector machines. For computational reasons, the number of "iterations" is set to 1. These two procedures are implemented in the 'CMA' package (Slawski et al. (2008)). In this study, the number of selected variables is fixed to  $p^* = 20, 50$  successively for the kNN method, and  $p^* = 100, 500$  for the PLS+LDA method, which are all common choices in the context of microarray data analysis (by experience, the number of relevant variables ranges between a few tens and a few hundreds, and PLS+LDA can better deal with weakly relevant variables).

### 2.2 Design of the comparison study

The study is based on two well-known real-life cancer data sets: the leukemia data set by Golub ( $n = 38$  observations,  $p = 3051$  variables) included in

the 'CMA' Bioconductor package (Slawski et al. (2008)) yielding very good accuracies with most standard classification methods, and the colon cancer data set by Alon included in the 'colonCA' Bioconductor package ( $n = 62$  observations,  $p = 2000$  variables) usually yielding error rates between 10% and 20% (see, e.g., Boulesteix (2004)). Both data sets include highly relevant variables as well as a large proportion of irrelevant variables.

The CV procedure is replicated several times using different random partitions, both in internal and external CV, which means that error rates are averaged over several random partitions instead of only one. This approach is commonly recommended to make the results more stable. In our study, external CV consists of 100 replications of  $J$ -fold CV with  $J = 6$ , whereas five replications of 3-fold internal CV are performed for tuning. In other words, external CV error rates are obtained by averaging over  $100 \times 6 = 600$  folds, while parameter values are selected based on the average error rate over three replications of 5-fold CV, i.e.  $3 \times 5 = 15$  internal folds.

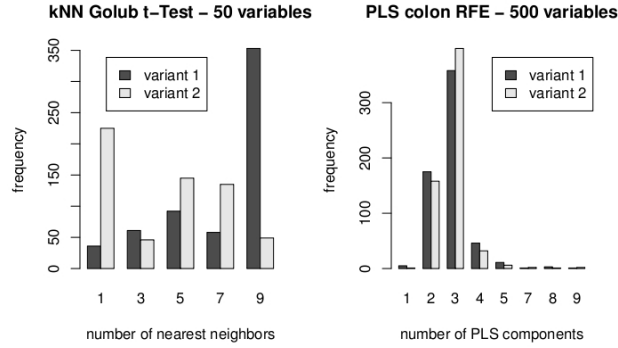
### 3 Results

#### 3.1 Do V1 and V2 select the same tuning parameter values?

At first we examine the tuning parameter values that are selected in the  $100 \times J$  tuning runs by internal CV using both variants V1 and V2. Do V1 and V2 select the same parameter values?

In about half of the setups, V1 and V2 yield similar results. For example, the barplot depicted in Figure 1 (right panel) shows the frequency of selection of each candidate number of PLS components with  $p^* = 500$  and RFE variable selection based on the colon data. The frequencies of selection do not differ substantially for V1 and V2. Clear differences between V1 and V2 are observed in the other half of the settings, with V2 consistently selecting more complex models. As illustrated in the left panel of Figure 1 for kNN with t-test variable selection and  $p^* = 50$  based on the Golub data, V2 noticeably selects smaller  $k$  values than V1 in kNN classification, i.e. more complex models. This general tendency of V2 to more complex models is also observed in several settings with PLS+LDA, where V2 selects higher numbers of PLS components.

The tendency of V1 to less complex models may be artificially enhanced by our convention that, if the lowest internal CV error rate is obtained with several parameter values, the least complex model is selected. Indeed, V1 performs tuning based on downwardly biased estimates of the error rate. With well-separated data sets such as Golub, it often occurs that all parameter values yield an error rate of 0%. In this case, the least complex model is selected, hence artificially increasing the frequency of selection of less complex parameter values (i.e. high  $k$  values or low  $ncomp$  values). However, this artificial mechanism resulting from our convention can only explain a moderate part of the tendency of V2 to more complex models.



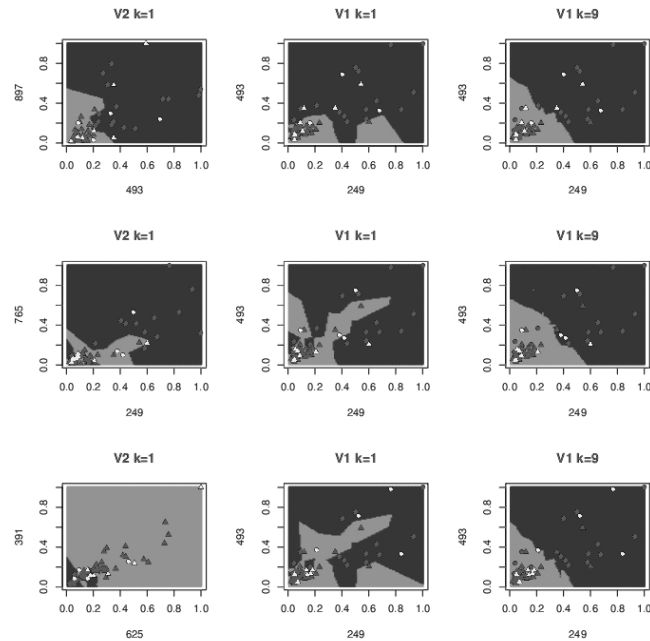
**Fig. 1.** Barplot of the frequency of selection of the candidate parameter values ( $k = 1, 3, 5, 7, 9$  for kNN and  $ncomp = 1, 2, \dots, 10$  for PLS+LDA) for both variants V1 and V2 in two different illustrative setups. Whereas the right panel (PLS+LDA, RFE,  $p^* = 500$ , colon data) shows similar frequencies of selection for each candidate parameter value, obvious differences can be observed in the left panel (kNN, t-test,  $p^* = 50$ , Golub data). The barplots sum to  $100 \times J = 600$ .

Roughly speaking, the higher complexity obtained with V2 can be partly explained as follows. With V2, the set of variables  $\mathcal{V}_{ij}$  used in the  $j$ th external iteration and  $i$ th internal iteration is selected based on  $L_{ij}$  only. Thus, in the learning set  $L_{ij}$ , they are more strongly associated to the response  $Y$  than the variables  $\mathcal{V}_j$  selected using the larger subsample  $L_j$ . As a consequence, a complex model fitted on  $L_{ij}$  based on variables  $\mathcal{V}_{ij}$  is likely to perform better than a complex model fitted with the "worse" variables  $\mathcal{V}_j$ . This mechanism can probably partly explain why V2 leads to the selection of more complex models than V1. It is illustrated in Figure 2 which shows the prediction regions of the kNN classifier (based on only  $p^* = 2$  for demonstration purposes) together with a scatterplot of the  $p^* = 2$  selected variables. In this example, complex models ( $k = 1$ ) in combination with V1 obviously yield overcomplex prediction regions leading to bad classification performance on the internal test data set  $T_{ij}$ .

### 3.2 Do the classification accuracies of V1 and V2 differ substantially?

In this section, we examine the differences in performance of V1 and V2 in external CV. On the whole, our conclusion is that both tuning variants yield approximately equal accuracies, with differences in accuracies smaller than 2.5% in all settings and almost no differences in standard deviations. As an example, the error rates obtained with the kNN method are summarized in Table 1. Similar differences are observed with PLS+LDA.

In the cases where V1 outperforms V2, the difference in performance is then largely due to the tendency of V1 to less complex models, as can be

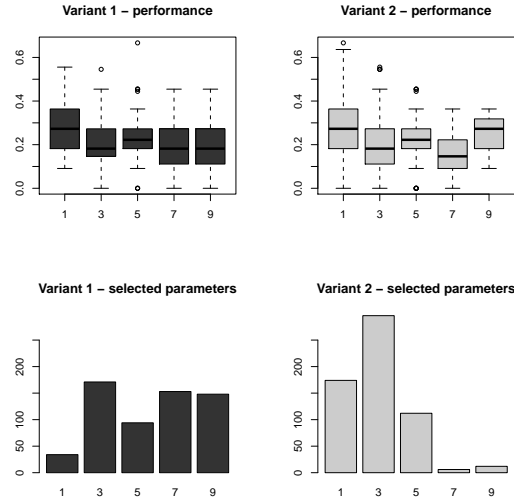


**Fig. 2.** Prediction regions of kNN classifiers with  $p^* = 2$  variables for three internal CV iterations based on variants V1 and V2. Each row corresponds to a particular internal CV iteration. **1st column:** V2 with  $k = 1$  neighbor. **2nd column:** V1 with  $k = 1$  neighbor. **3rd column:** V2 with  $k = 9$  neighbors. Circles stand for observations of class  $Y = 0$ , triangles stand for  $Y = 1$ . White symbols represent internal test observations from  $T_{ij}$ .

seen from the example depicted in Figure 3 (kNN with RFE, colon data). In this setup, the parameter value  $k = 1$  (complex model) is not often selected by V1, and yields higher error rates than the other  $k$  values, regardless of whether it is chosen by V1 or V2. V1 thus seems to benefit from its tendency to less complex models.

kNN		Golub data				colon cancer data			
		t-test		RFE		t-test		RFE	
		V1	V2	V1	V2	V1	V2	V1	V2
20 genes	mean MCR	7.8%	7.4%	5.8%	6.1%	16.8%	18.8%	21.6%	23.3%
	std. dev.	2.6%	2.8%	2.5%	2.9%	1.9%	2.4%	3.3%	4.1%
50 genes	mean MCR	5.9%	5.5%	1.9%	2.2%	16.4%	19.9%	16.9%	18.5%
	std. dev.	2.4%	2.7%	1.8%	1.7%	1.6%	1.9%	3.3%	3.0%

**Table 1.** Mean error rates (and standard deviations) with kNN using V1 and V2.



**Fig. 3.** kNN, RFE,  $p^* = 20$ , colon data. **Top:** Boxplots of the error rates in external CV for different values of  $k$  with V1 (left) and V2 (right). **Bottom:** Barplots of the frequencies of selection of the different  $k$  values with V1 (left) and V2 (right).

## 4 Discussion

In some setups, we see that overcomplex models, which are frequent with V2, are associated with an increased error rate. This may be partly explained as follows. Since the  $L_{ij}$  are smaller than the  $L_j$ , it is easier to find variables that separate the two classes well in  $L_{ij}$  than in  $L_j$ . For these variables  $\mathcal{V}_{ij}$ , complex models perform well when fitted on  $L_{ij}$  – probably even better than when they are fitted on variables  $\mathcal{V}_j$  in  $L_j$ . This may partly explain why the tendency of V2 to more complex models seems to be a disadvantage.

The variables used by V2 to construct the classifier in external cross-validation are not the same as those used for tuning in internal cross-validation. Beyond the examples considered in our paper, this may yield substantial problems in some cases, for instance when the tuning parameter controls the “amount of non-linearity” of a classifier. If some variables show linear relationships with  $Y$  while other substantially depart from linearity, performing parameter tuning and classifier fitting using different variables is obviously sub-optimal.

Finally, we point out that V1 may show worse performance in data sets with well-separated classes (like the Golub data) if all parameter values yield an error rate of 0% in internal CV. This may often occur in practice, since the internal CV error rates are strongly downwardly biased in V1. In this

case, no tuning is achieved by V1, while V2 is based on higher error rates that can be compared to perform parameter tuning.

Let us mention that, from a theoretical point of view, both variants V1 and V2 can be seen as imperfect workarounds for a computationally unfeasible task. The correct approach would be to select the tuning parameter and the variable subset jointly from a multi-dimensional grid in internal CV. Of course, an exhaustive search is unfeasible in high-dimensional data analysis. The development of simplified computationally efficient algorithms could be addressed in further research.

## 5 Conclusion

Our study shows that the two investigated tuning variants sometimes lead to clearly different tuning results. Variant V1 shows a general tendency to less complex models using both investigated data sets. Similar results are also obtained using further microarray data sets (data not shown). With regard to prediction accuracy, V1 and V2 yield similar accuracies and, in some settings, the seemingly inappropriate V1 approach even outperforms the more natural V2. Although V1 performs tuning based on severely biased internal CV error rates, the selected tuning parameter values yield acceptable accuracies in the settings considered in our study. Hence, the benefit of V2's higher computational expense in terms of prediction accuracy cannot be confirmed through our study.

## Acknowledgments

This project was supported by DFG project BO3139/2-1 and by the LMU-innovativ Project BioMed-S.

## References

- BOULESTEIX, A.-L. (2004): PLS dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology* 3, 33.
- GUYON, I., WESTON, J., BARNILL, S. et al. (2002): Gene selection for cancer classification using support vector machines. *Machine Learning* 46, 389–422.
- RUSCHHAUPT, M., HUBER, W., POUSTKA and MANSMANN, U. (2004): A Compendium to Ensure Computational Reproducibility in High-Dimensional Classification Tasks. *Statistical Applications in Genetics and Molecular Biology* 3, 37.
- SLAWSKI, M., DAUMER, M. and BOULESTEIX, A.-L. (2008): CMA - a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics* 9, 439.
- VARMA, S. and SIMON, R. (2006): Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7, 91.



# A Generative Model for Rank Data Based on Sorting Algorithm

Christophe Biernacki and Julien Jacques

Université Lille I & CNRS, Villeneuve d'Ascq, France,  
*christophe.biernacki@math.univ-lille1.fr, julien.jacques@polytech-lille.fr*

**Abstract.** Rank data arise from a sorting mechanism which is generally unobservable for the statistician. Assuming both that this mechanism relies on paired comparisons and that it aims to minimize their number, the insertion sorting algorithm is one of the best candidates. A Bernoulli event can be naturally introduced in the paired comparison step, leading to an original probabilistic generative model for rank data which depends on the initial presentation order. Its theoretical properties are studied among which unimodality, symmetry and identifiability. In addition, maximum likelihood principle can be easily performed through an EM algorithm thanks to an unobserved latent variables interpretation of the model. Finally, an illustration of adequacy between the proposed model and rank data resulting from a general knowledge quiz suggests the relevance of our proposal.

**Keywords:** EM algorithm, insertion algorithm, quiz data, rank data, sorting process.

## 1 Introduction

Ranking data are of great interest in human activities involving preferences, attitudes or choices like Web Page ranking, Sport, Politics, Economics, Educational Testing, Biology, Psychology, Sociology, Marketing, *etc.* Ranks are so meaningful that it is not unusual they result from a transformation of other kinds of data. Rank data are multivariate but highly structured data. From an inference point of view, parametric probabilistic models, if relevant and allowing easy parameter interpretation, are useful for summarizing and understanding such quite complex data and are a basis tool for density estimation, prediction or clustering. Major rank data models date from the mid 20th century and most of the current works on the topic uses these models. Pointing out that a rank data is the result of a sorting process, we suggest in this paper a generative model for rank data, based on a modeling of the sorting process which aims to be optimal in a sense explained in the next section.

## 2 Notation and usual rank data models

The rank datum, which is the statistical unit of interest in this paper, results from a ranking of  $m$  objects  $\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_m$  by a judge (human or not). Two

representations of these data are commonly used: Ranking or ordering. The *ranking* representation  $x^{-1} = (x_1^{-1}, x_2^{-1}, \dots, x_m^{-1})$  contains the ranks given to the objects, and means that  $\mathcal{O}_1$  is in the  $x_1^{-1}$ -th position,  $\mathcal{O}_2$  is in the  $x_2^{-1}$ -th position, and so forth. A ranking is then an element of  $\mathcal{P}_m$ , the set of permutations of the  $m$  first integers. The *ordering* representation  $x = (x_1, x_2, \dots, x_m)$  is also an element of  $\mathcal{P}_m$  and signifies that object  $\mathcal{O}_{x_1}$  is the first, object  $\mathcal{O}_{x_2}$  is the second, *etc.* In the following both ordering and ranking notations will be used for rank data.

The two most popular classes of models for rank data consist in modeling directly the hypothetical ranking process followed by the judge. For a complete review, refer to Marden (1995), Chap. 5 to 10. The first class is derived from a paired comparison process (Kendall and Smith, 1940; Mallows, 1957) and the second one consists in multistage models (Luce, 1959; Plackett, 1975; Fligner and Verducci, 1986; Fligner and Verducci, 1988). The ranking processes which have motivated these two classes of rank data models can be interpreted as two different sorting processes, in which stochastic errors are introduced to define a probability distribution on the whole rank data space. The natural question involved by this interpretation is whether the sorting algorithms used are the most appropriate. Effectively, in paired comparison models it seems not optimal to do so much comparisons since it leads to a sorting algorithm with excessively high computational complexity. In practice a human judge would probably not exhaustively proceed to all paired comparisons. For multistage models, the ranking process can be likened to a *selection* sorting algorithm. Even if this sorting algorithm is one of the most simple, it is well known for its lack of optimality (Knuth, 1973).

Here, we propose a generative model for rank data based on the (straight) *insertion* sorting algorithm, which is one of the most powerful among the usual sorts when  $m \leq 10$  (Knuth, 1973, Chap. 5). In addition, our proposal is potentially able to take into account the presentation order of the different objects to the judge, realistic situation which can have an impact on the resulting rank.

### 3 A generative model based on an insertion sorting

**Genesis** We assume there exists an ordering  $\mu = (\mu_1, \dots, \mu_m)$  on the  $m$  objects, so that a judge who perfectly sorts these objects returns this *reference* rank  $\mu$ . Moreover, we assume that the judge aims to minimize the number of paired comparison, and then adopts the *insertion* sorting algorithm which is optimal for reasonable number of objects ( $m \leq 10$ ). We also introduce the possibility for the judge of making mistakes regarding to  $\mu$  in his sorting, and such mistakes will be modeled by a random event in paired comparison. Merging both deterministic insertion algorithm and the random paired comparison leads to a meaningful generative model for rank data that is now presented at length.

Let the ordering  $\sigma = (\sigma_1, \dots, \sigma_m)$  be the presentation order of the objects to the judge, this latter using an insertion sorting algorithm to rank these objects. The current object to be sorted is placed on the left of the already sorted objects, and is compared to the first object on its right. If the relative position of both objects in this pair is correct (according to  $\mu$ ), this pair order is unchanged and the next object in  $\sigma$  is inserted far left. Otherwise, the pair order is reversed and a new pair comparison is performed with the next object on the right (if it exists). And so forth. The result of this deterministic sorting algorithm would be  $\mu$  if the judge was perfect. However, none judge is perfect and the mistakes he/she/it can do leads to a given rank  $x = (x_1, x_2, \dots, x_m)$ , which could be different from  $\mu$ . Since the sorting algorithm by insertion consists solely of a sequence of comparisons of pairs of objects, it is natural to model the reliability of the judge for the ranking by the risk of wrongly order a pair of objects. Each pair comparison can be interpreted as the result of a Bernoulli experiment whose outcome is a correct comparison (according to  $\mu$ ) with probability  $p$  and an incorrect comparison with probability  $1 - p$ . Moreover, it is reasonable to assume that each pair ranking operation is independent of the others. Based on this modeling of a stochastic insertion sorting, the first natural question is to calculate the probability  $pr(x|\sigma; \mu, p)$  to obtain a rank  $x$  from an initial presentation order  $\sigma$  and a reference rank  $\mu$ . To do so, let introduce the following notations, where  $j = 1, \dots, m$  denotes the step in the sorting algorithm consisting of ranking the object  $\mathcal{O}_{\sigma_j}$ .

- $\delta_{ii'}(\mu) = \mathbf{1}\{\mu_i^{-1} < \mu_{i'}^{-1}\}$ , which is equal to 1 if  $\mathcal{O}_i$  is correctly ranked before  $\mathcal{O}_{i'}$  (according to  $\mu$ ), 0 otherwise ( $i, i' = 1, \dots, m, i \neq i'$ ).
- $j^-(x, \sigma) = \{i : x_{\sigma_i}^{-1} < x_{\sigma_j}^{-1}, 1 \leq i < j\}$  is the set of the indices of the presentation order  $\sigma$  for which the already sorted objects  $\mathcal{O}_{\sigma_1}, \dots, \mathcal{O}_{\sigma_{j-1}}$  are ranked in  $x$  before the current object  $\mathcal{O}_{\sigma_j}$ , and consequently *on its left*. Its cardinal  $\#j^-(x, \sigma)$  is consequently the number of *all* comparisons of the current object with the objects already ranked (according to  $x$ ) on its left, if they exist.
- $j^+(x, \sigma) = \{i : i = \arg \min_{1 \leq i' < j} \{i' : x_{\sigma_{i'}}^{-1} > x_{\sigma_j}^{-1}\}\}$  is the index of the rank  $\sigma$  designating the object sorted in  $x$  just after (so *on the right* of)  $\mathcal{O}_{\sigma_j}$  among the already sorted objects  $\mathcal{O}_{\sigma_1}, \dots, \mathcal{O}_{\sigma_{j-1}}$ . This set has at most one element. Its cardinal  $\#j^+(x, \sigma)$  indicates that the current object  $\mathcal{O}_{\sigma_j}$  has been compared with the object ranked (according to  $x$ ) just *on its right*, if it exists.
- $\eta_j^-(x, \sigma, \mu) = \sum_{i \in j^-(x, \sigma)} \delta_{\sigma_i \sigma_j}(\mu)$  is the number of *good* comparisons (according to  $\mu$ ) of the current object  $\mathcal{O}_{\sigma_j}$  with the objects already ranked *on its left*, if they exist.
- $\eta_j^+(x, \sigma, \mu) = \sum_{i \in j^+(x, \sigma)} \delta_{\sigma_j \sigma_i}(\mu)$  is the indicator of *good* comparison (according to  $\mu$ ) of the current object  $\mathcal{O}_{\sigma_j}$  with the object already ranked just *on its right*, if it exists.

With these notations, the probability to obtain a rank  $x$  from a initial presentation order  $\sigma$  is:

$$pr(x|\sigma; \mu, p) = \prod_{j=1}^m p^{\eta_j^-(x, \sigma, \mu)} (1-p)^{\#j^-(x, \sigma) - \eta_j^-(x, \sigma, \mu)} p^{\eta_j^+(x, \sigma, \mu)} (1-p)^{\#j^+(x, \sigma) - \eta_j^+(x, \sigma, \mu)}. \quad (1)$$

The first term  $p^{\eta_j^-(x, \sigma, \mu)} (1-p)^{\#j^-(x, \sigma) - \eta_j^-(x, \sigma, \mu)}$  corresponds to the probability of shifting  $\#j^-(x, \sigma)$  times to the right the object  $\mathcal{O}_{\sigma_j}$  coming at the step  $j$ , and the second term  $p^{\eta_j^+(x, \sigma, \mu)} (1-p)^{\#j^+(x, \sigma) - \eta_j^+(x, \sigma, \mu)}$  is the probability for this object of being no longer shifted to the right. Finally, if the presentation order is unknown but of probability  $pr(\sigma)$ , the marginal distribution of  $x$  is

$$pr(x; \mu, p) = \sum_{\sigma \in \mathcal{P}_m} pr(x|\sigma; \mu, p) pr(\sigma). \quad (2)$$

In this paper, we assume the presentation orders are uniformly distributed, and then  $pr(\sigma) = 1/m!$  for all  $\sigma \in \mathcal{P}_m$ . In the following the rank data model defined by (2) will be named ISR for Insertion Sorting Rank data model. We will note shortly  $ISR(\mu, p)$  this model and its associated parameters.

**Properties** The main properties of the ISR model are the following: The possibility for the ISR distribution to be uniform for a special value of  $p$  (Proposition 6), the existence of modal and anti-modal ranks (Proposition 7 and 8), the symmetry of the ISR distribution (Proposition 9), and finally its identifiability (Proposition 10). The proofs are in Biernacki and Jacques (2010).

**Proposition 6.** *If  $p = 1/2$ ,  $\forall x, \mu \in \mathcal{P}_m$  then  $pr(x; \mu, 1/2) = 1/m!$ .*

**Proposition 7.** *If  $p > 1/2$ ,  $\forall x, \mu \in \mathcal{P}_m$ ,  $x \neq \mu$ , then  $pr(\mu; \mu, p) > pr(x; \mu, p)$ .*

**Proposition 8.** *If  $p > 1/2$ ,  $\forall x, \mu \in \mathcal{P}_m$ ,  $x \neq \bar{\mu}$ , then  $pr(\bar{\mu}; \mu, p) < pr(x; \mu, p)$ , where  $\bar{\mu}$  is the anti-modal rank (the rank the further from  $\mu$  for the Kendall distance).*

**Proposition 9.** *For all  $x, \mu \in \mathcal{P}_m$ , we have  $pr(x; \bar{\mu}, 1-p) = pr(x; \mu, p)$ .*

**Proposition 10.** *The ISR distribution is identifiable since  $p > 1/2$ .*

**Estimation of the model parameters** The ISR model for rank data has two parameters: The probability  $p$ , which is a real in  $[1/2, 1]$  and the reference rank, or modal rank,  $\mu$ , which can take its values in  $\mathcal{P}_m$ . The estimation is done by maximum likelihood, using the EM algorithm (Dempster et al., 1977) since the presentation order can be interpreted as missing data. For details refer to Biernacki and Jacques (2010).

## 4 Numerical illustration

In order to assess the adequacy of the ISR distribution to a real data set, we submitted the following quiz to our students, consisting of three items  $Q_1$ ,  $Q_2$  and  $Q_3$  of ascending difficulty:

- $Q_1$ . Rank the following numbers in ascending order:

$$\mathcal{O}_1 = \pi/3, \mathcal{O}_2 = \log 1, \mathcal{O}_3 = \exp 2, \mathcal{O}_4 = \frac{1+\sqrt{5}}{2}.$$

- $Q_2$ . Rank the following French writers in chronological order of birth:

$$\mathcal{O}_1 = \text{V. Hugo}, \mathcal{O}_2 = \text{Molière}, \mathcal{O}_3 = \text{A. Camus}, \mathcal{O}_4 = \text{J.-J. Rousseau}.$$

- $Q_3$ . Rank chronologically these Quentin Tarantino movies:

$$\mathcal{O}_1 = \text{Inglourious Basterds}, \mathcal{O}_2 = \text{Pulp Fiction}, \mathcal{O}_3 = \text{Reservoir Dogs}, \\ \mathcal{O}_4 = \text{Jackie Brown}.$$

The correct answers are  $\mu^* = (2, 1, 4, 3)$  for  $Q_1$ ,  $\mu^* = (2, 4, 1, 3)$  for  $Q_2$  and  $\mu^* = (3, 2, 4, 1)$  for  $Q_3$ . The answers of the 40 questioned students are in turn in Table 1.

Quiz $Q_1$		Quiz $Q_2$		Quiz $Q_3$	
ordering	frequency	ordering	frequency	ordering	frequency
(2, 1, 4, 3)	32	(2, 4, 1, 3)	15	(4, 3, 2, 1)	10
(2, 4, 1, 3)	6	(2, 4, 3, 1)	8	(4, 2, 3, 1)	9
(2, 1, 3, 4)	2	(2, 1, 4, 3)	4	(3, 2, 4, 1)	4
other	0	(4, 2, 1, 3)	4	(3, 4, 2, 1)	3
		(2, 3, 1, 4)	2	(1, 3, 2, 4)	2
		(1, 2, 3, 4)	1	(1, 3, 4, 2)	2
		(1, 3, 4, 2)	1	(2, 3, 1, 4)	2
		(1, 4, 2, 3)	1	(3, 1, 4, 2)	2
		(2, 1, 3, 4)	1	(1, 2, 3, 4)	1
		(2, 3, 4, 1)	1	(2, 3, 4, 1)	1
		(3, 1, 4, 2)	1	(2, 4, 3, 1)	1
		(3, 2, 1, 4)	1	(3, 2, 1, 4)	1
		other	0	(4, 1, 2, 3)	1
				(4, 3, 1, 2)	1
				other	0

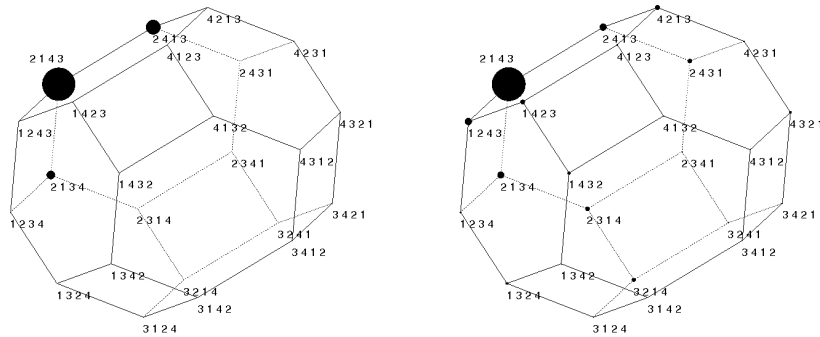
**Table 1.** Quiz answers of the 40 students.

For each item of the quiz, the ISR distribution is estimated and a  $\chi^2$  adequacy test, where the distribution under the null assumption is estimated by bootstrap (Efron and Tibshirani, 1993) based on 1000 replications, is performed:

- $Q_1$ .  $\hat{\mu} = (2, 1, 4, 3)$ ,  $\hat{p} = 0.962$  and p-value = 0.593,

- $Q_2$ .  $\hat{\mu} = (2, 4, 1, 3)$ ,  $\hat{p} = 0.815$  and p-value = 0.342,
- $Q_3$ .  $\hat{\mu} = (4, 3, 2, 1)$ ,  $\hat{p} = 0.754$  and p-value = 0.264.

We can first note that the adequacy of the ISR distribution is accepted for these three questions. This adequacy can be also found graphically on Figures 1, 2 and 3 displaying polytopes in the ranking space (Thompson, 1993a; Thompson, 1993b) (orderings are displayed on each node) of both the empirical distribution and the ISR estimated one. We remark also the decrease of the number of good answers when one moves away from the modal rank. The

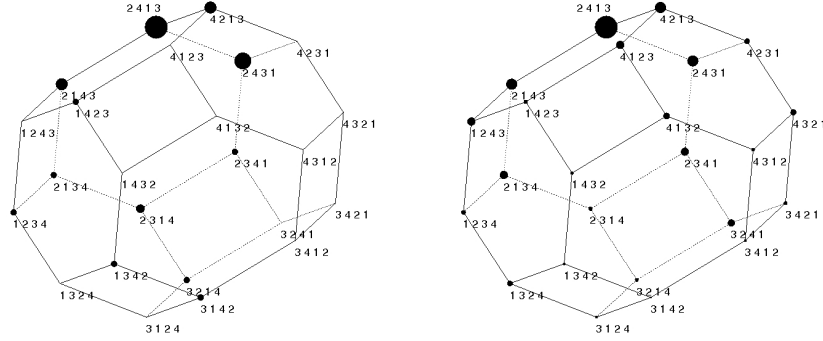


**Fig. 1.** Empirical (left) and estimated (right) distributions for quiz  $Q_1$ , related to numbers.

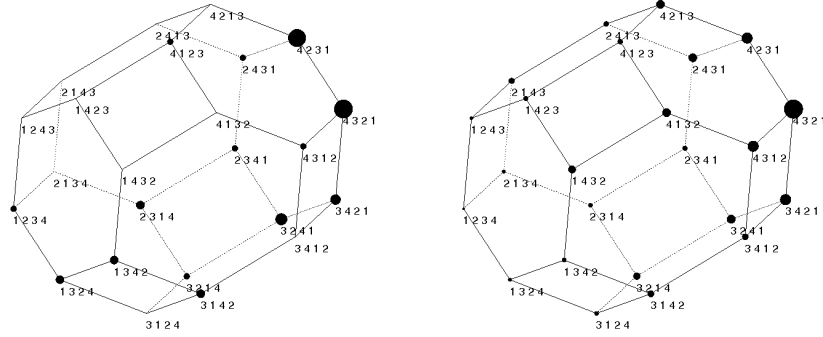
growth of the questions difficulty is reflected by a decrease in the probability  $p$ : For the easy first question, 80% of the students gives the right answer and 15% makes one mistake by reversing  $\mathcal{O}_1$  and  $\mathcal{O}_4$ . This leads to a high value of the probability  $p$ :  $\hat{p} = 0.962$ . For the second question, only 37.5% of the students gives the right answer, and the number of wrong answers decreases gradually with the number of bad comparisons made in the ranking process. Finally, the last question leads to more mixed answers with a smaller decrease in the number of responses gradually as the distance from the modal rank  $\hat{\mu} = (4, 3, 2, 1)$ , which moreover is different from the right rank  $\mu^* = (3, 2, 4, 1)$ .

## 5 Discussion

In this paper we suggest a probability distribution for rank data, modeling a stochastic version of the insertion sorting algorithm. The main force of the ISR distribution consists in the naturalness and the optimality of this sorting



**Fig. 2.** Empirical (left) and estimated (right) distributions for quiz  $Q_2$ , related to French writers.



**Fig. 3.** Empirical (left) and estimated (right) distributions for quiz  $Q_3$ , related to Quentin Tarantino movies.

algorithm for a moderate number  $m$  of objects to rank. In this sense, we can expect in many cases a higher modeling power of the ISR model than usual rank data models which can be interpreted as the modeling of poorly performing sorting algorithms. Obviously, this claim relies on the assumption that the judge is somewhat *optimal*. Moreover, the ISR model allows to take into account, if it is known, the presentation order of the objects to rank, what could be particularly interesting since this last could influence the ranking in such situations. Two other benefits to consider such a generative model is that it allows an interpretation of the ranking results *via* its parameters  $\mu$

(the modal ranking) and  $p$  (the probability of good paired comparison during the sorting), and leads to easy maximum likelihood estimation.

## References

- BIERNACKI, C. and JACQUES, J. (2010). A generative model for rank data based on sorting algorithm. Technical report, Preprint IRMA Lille Vol. 70-III.
- DEMPSTER, A. P., LAIRD, N. M., and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38. With discussion.
- EFRON, B. and TIBSHIRANI, R. J. (1993). *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York.
- FLIGNER, M. A. and VERDUCCI, J. S. (1986). Distance based ranking models. *J. Roy. Statist. Soc. Ser. B*, 48(3):359–369.
- FLIGNER, M. A. and VERDUCCI, J. S. (1988). Multistage ranking models. *J. Amer. Statist. Assoc.*, 83(403):892–901.
- KENDALL, M. G. and SMITH, B. B. (1940). On the method of paired comparisons. *Biometrika*, 31:324–345.
- KNUTH, D. (1973). *Sorting and Searching: Volume 3. The art of Computer Programming*. Addison-Wesley, Massachusetts.
- LUCE, R. D. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley & Sons Inc., New York.
- MALLOWS, C. L. (1957). Non-null ranking models. I. *Biometrika*, 44:114–130.
- MARDEN, J. I. (1995). *Analyzing and modeling rank data*, volume 64 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.
- PLACKETT, R. L. (1975). The analysis of permutations. *J. Roy. Statist. Soc. Ser. C Appl. Statist.*, 24(2):193–202.
- THOMPSON, G. L. (1993a). Generalized permutation polytopes and exploratory graphical methods for ranked data. *Ann. Statist.*, 21(3):1401–1430.
- THOMPSON, G. L. (1993b). *Probability models and statistical analyses for ranking data*, chapter Graphical techniques for ranked data, pages 294–298. Springer-Verlag, New-York.



# “Made in Italy” Firms Competitiveness: A Multilevel Longitudinal Model on Export Performance

Matilde Bini<sup>1</sup> and Margherita Velucchi<sup>2</sup>

<sup>1</sup> Università Europea di Roma, Via degli Aldobrandeschi 190 - 00163 Roma, Italy  
*mbini@unier.it, bini@ds.unifi.it*

<sup>2</sup> European University Institute, Badia Fiesolana, Via dei Roccettini 9, 50014, San Domenico di Fiesole, Firenze, Italy *margherita.velucchi@eui.eu*

**Abstract.** The competitiveness of the Italian industrial system during the last decade has shown a strong slowdown. To compete in international markets, Italian firms reduced their costs instead of fostering on innovation and investments, being largely influenced by small size. Only the so-called “Made in Italy” sectors succeeded in international markets. To analyze this phenomenon, we investigate, at firm and sector level, factors affecting export competitiveness in “Made in Italy” sectors using a multilevel longitudinal model in the period 1999-2005. We find that “Made in Italy” role in international markets strongly depends on firms’ geographical location and sector of activity and on their innovative capacity and productivity.

**Keywords:** Competitiveness, Made in Italy, multilevel models

## 1 Introduction

During the last decade, Italy has experienced a significant slowdown in its economic growth rate. Although other European countries have experienced similar problems in their economies, macro-economic indicators show that the Italian slowdown has been more marked (OECD, 2005). Several reasons lie behind the Italian poor performance: a sharp decline both in physical and human capital investments and in labor productivity. This phenomenon is well known and the Italian economy has been deeply hurt by it. However, in the same period at least part of the economic system has been very successful in international markets. These sectors, called “Made in Italy” sectors, have experienced similar difficulties but they have been capable to capture some opportunities that yielded them to be extremely competitive in international markets. “Made in Italy” sectors are the 3F of the Italian economy (Food, Fashion and Furniture) and are usually considered the most dynamic and creative sectors in Italy (ICE, 2005; Brandolini and Cipollone, 2003; Rabelotti et al., 2009). In a sort of “polarization” of the economy, “Made in Italy” sectors have become worldwide famous and successful while the whole Italian

economy was suffering from lost competitiveness. There can be several reasons behind this phenomenon; one of these relates the degree of internationalization of firms and firms heterogeneity competing in international markets (Meyer and Ottaviano, 2008; Castellani and Zanfei, 2007) and their labor productivity and size (Zeli and Mariani, 2009). Indeed, literature emphasizes that firms involved in international activities are “different” from purely domestic firms in several respect (labor productivity, wages, skill intensity, see for all Mayer and Ottaviano, 2007). The underlying idea is that there are relatively few firms ‘fit’ to cope with the more competitive international markets and these firms are more productive, pay higher wages, employ more skilled workers, invest more in R&D (Giovannetti et al. 2009). This paper investigates the factors affecting the export competitiveness of “Made in Italy” sectors in the period 1999-2005 at a firm level, distinguishing between firm-specific factors like size and labor productivity from context-specific factors like the geographical location and the presence of an industrial district in the region. We use a longitudinal multilevel approach to simultaneously model individual and context factors that affect the firms export competitiveness in presence of hierarchical data.

The paper is structured as follows: Section 2 briefly describes the “Made in Italy” characteristics, Section 3 introduces the model, Section 4 introduces the dataset, Section 5 discusses the results and Section 6 concludes.

## 2 “Made in Italy”: what is that?

Italy presents a specific technological specialization that has been the object of numerous economic analyses that have tried to explain over the years the diversity of the Italian economic model characterized by elements generally defined as a sign of backwardness, prevalence of small businesses (SMEs) and development and specialization in traditional sectors, which have been transformed into successful factors (the economic literature has often referred to the Italian case as *bumblebee Italy*; see Becattini, 2007). The traditional sectors also known as “Made in Italy” sectors include: food and wines, fashion furniture, marble, stone and ceramic tiles, metal products, machinery and domestic appliances, motorcycles, bicycles and yachts. They are sectors where the competition is strongest and where the Italian system has been successful over the years. Also, a relevant share of the production of Made in Italy is manufactured in industrial districts (IDs) (Conti and Menghinello, 1996; Oropallo, 2007). The industrial district is a structure of firms that involves and integrates economic and social environment into an economic one, creating a network in which firms produce and entrepreneurs and their families work and live.

In determining the competitiveness and success of “Made in Italy” specific elements play a role: immaterial factors like *tacit knowledge* and *learning by doing*, spillovers from the economic and social context, the capacity in cre-

ating and managing the demand for high-quality products. This is one of the reasons why, in order to protect “Made in Italy” specific innovation and quality, a wider use of trademarks and designs rather than patents are used. The performance of these firms is indeed more difficult to be measured and grasped because it is often related to tacit knowledge and skills instead of to new products. The aim of this paper is to analyze the export performance of “Made in Italy” firms including in the model the “spillover effect” from the environment in which they are working (both at a regional and sector level).

### 3 The Model

The Multilevel analysis combines information from more than one level of observation in studying the determinants of various forms of units’ behavior. Concerning firms, their behavior is not only influenced by individual goals and characteristics but it is also shaped by the social and economic environment. The multilevel approach, by combining elements from both levels allows greater concordance between the theoretical views and the models employed for studying firms’ behavior. Standard regression models (such as GLM), indeed, are not adequate when complex structure of data exist as they do not take into account a crucial feature of the problem, namely the data (hidden) hierarchical structure (Hox and Maas, 2005). To investigate the effect of some characteristics of “Made in Italy” firms on their export performance and competitiveness, we run regression analyses with a longitudinal multilevel approach due to the hierarchical structure of data set. In particular, in our dataset the measurements are repeated on the units (firms) over time, so a basic three hierarchical structure can be arranged with measurement occasions as first level, subjects as second level and the economic sector as third level. In this case, we also take into account the average pattern of changes over time as well as the variation of this pattern across subjects. The hierarchical model is a longitudinal three-level model with random intercept and random slope (Yang, Goldstein, 1996; Skrondal, Rabe-Hesketh, 2004) to allow more general results. Let the  $i$ -th occasion for the  $j$ -th firm of the  $k$ -th economic sector be expressed by a three level model as follows:

$$y_{ijk} = \underbrace{\beta_0 + \beta_1 a_{1ijk} + \beta_2 a_{2ijk} + \sum_{l=3}^L \beta_l x_{lijk} + \sum_{p=1}^P \zeta_p s_{pjk} + \sum_{h=1}^H \alpha_h d_{hjk} + \sum_{q=1}^Q \gamma_q t_{qijk} + \delta_1 e_{1jk}}_{\text{fixed part}} + \underbrace{u_{jk} + v_k + \varepsilon_{ijk}}_{\text{random part}}$$

where  $y_{ijk}$  is the response variable (propensity to export);  $i = 1, 2, \dots, I$  is the index of the number of occasions per unit;  $j = 1, 2, \dots, nk$  is the index of the number of units;  $k = 1, 2, \dots, K$  is the index of the number of economic sectors or districts.

## 4 The Data

To carry out performance and competitiveness analyses, The Italian National Statistical Institute (INSI) proposed a data base (1999-2005), for a population of 1,381,996 limited liability firms in industrial and service sectors. The information record for each firm is obtained combining data coming from the Balance sheet with data of Business Register (ASIA) and Export surveys data. Two additional sources provide information on legal structure and industry affiliation, year of the start of activity (age of firms), economic classification (ATECO, “ATTivit ECONomica”), localization, employment, events of reorganization like fusions, divisions, ceasing of activity, propensity to export (as a share of sales), presence of industrial district in the region and innovative capacity. In this paper, we select information on “Made in Italy” sectors (the 3F only) counting for 10.3% of the whole sample and 47.7% of the manufacturing sectors. For this sub-sample, we analyze the factors that may affect firms export performance: labor productivity (added value per employee) as a proxy of competition on the market, ROI, ROE, ROA for the stake-holders value, innovative capacity of the firm, presence of an industrial district in the region and size. A preliminary descriptive analysis (Table 1 versus Table 2) comparing manufacturing and “Made in Italy” sectors reveals that 98% (versus 77% of manufacturing) of firms in our sample has introduced an innovation (either process or product innovation) during the period and that “Made in Italy” sectors export on average 21% of their sales (versus 16% in the whole sample). The competitiveness proxy (added value/cost of labor) is very similar between the two groups and shows high heterogeneity in the data. Interestingly, “Made in Italy” firms tend to cluster more than the manufacturing firms (49% versus 43% of firms work in an industrial district area) and show slightly higher levels of productivity (added value is 10.92 versus 10.45). Finance indicators (roi, roa, roe) are very similar.

Variable	Description	Mean	Std.Dev.	Min	Max
f_exp	Export/Sales	0.16	0.25	0	1
lvaladd	Return on Investments	10.45	0.69	-0.45	12.83
roi	Return on Net Capital	0.17	0.30	-1	1
roe	Return on Earnings	0.07	0.36	-1	1
roa	Value Added (log)	0.05	0.11	-1	1
compet	Value Added/cost of labor	140.78	60.66	-19.20	567.23
dummy_dist	Belong to an ID	0.43	0.49	0	1
age	Age of Firm	17.55	12.21	0	140
employees	Number of Employees	34.37	167.44	0.25	18108
inno	Innovative Capacity	0.78	0.14	0	1

Note: 383677 observations.

**Table 1.** Descriptive Statistics on the manufacturing sectors

Variable	Description	Mean	Std.Dev.	Min	Max
f_exp	Export/Sales	0.21	0.28	0	1
lvaladd	Return on Investments	10.92	0.70	-0.45	12.83
roi	Return on Net Capital	0.16	0.30	-1	1
roe	Return on Earnings	0.06	0.36	-1	1
roa	Value Added (log)	0.05	0.11	-1	1
compet	Value Added/cost of labor	140.69	61.71	-19.14	567.23
dummy_dist	Belong to an ID	0.49	0.50	0	1
age	Age of Firm	17.95	12.55	0	140
employees	Number of Employees	34.31	131.35	1	8506
inno	Innovative Capacity	0.98	0.12	0	1

Note: 183421 observations

**Table 2.** Descriptive Statistics on the “Made in Italy” sectors

## 5 The Results

To run a multilevel model a two step procedure is required: 1) a null model is estimated to test second and third level variance significance to show the existence of a hierarchical structure in the data and 2) a general model is estimated including individual and context variables. The results of the likelihood ratio test on second and third level significance (region and sector) show that the multilevel approach is appropriate (LR  $\chi^2 = 224.23$ ,  $p\text{-value} < 0.001$ ). Test results show that a hierarchical structure in the data exists, confirming the use of a multilevel approach to describe and forecast Italian firms propensity to export. Then, we run a null and a general model, to select the best specification for our data. The best model specification has been detected inserting in the null model, firstly, the individual and, secondly, the context variables. Several models have been estimated to test the stability of the estimates both at individual and context level.

Table 3 reports the selected model showing that it fits well the data. The results show that all individual variables but Roe (returns on earnings) are positive and statistically significant. For example, one unit increase in the added value of a firm increases its average propensity to export by 1.3% while the role of the competitiveness proxy, age and size are less evident (although still significant and positive). The cost of labor has a positive and high coefficient, stressing that better skilled and better remunerated employees help the firm improving its role in international markets. As expected, export competitiveness strongly relies on the skills of labor and suffers by cuts in the remuneration of the labor factor. Firms investing in innovation and productivity have higher probability of succeeding in international markets, confirming Meyer and Ottaviano (2007). All individual factors positively affect the export competitiveness of “Made in Italy” firms and represent important elements in driving it. On the context variables side, geographical location and sector turn out to be important factors too. In particular, while working in some regions does not give a comparative advantage (see for example Val d’Aosta and Liguria) being and working in some other regions strongly affect the firms’ export performance. It is worth noticing that Veneto, Lombardia and Friuli Venezia Giulia represent the most stimulating places where establish and run a firm: a key

Dep. Var: f_exp	Coef.	Std.Err.	z	P> z
added value	0.0137	0.0009	15.86	<0.01
roi	-0.0112	0.0014	-7.96	<0.01
roe	0.0002	0.0011	0.21	0.83
roa	-0.0652	0.0056	-11.71	<0.01
ID	0.0033	0.0010	3.30	<0.01
Competitiveness Index	0.0002	0.0000	19.95	<0.01
age	0.0005	0.0000	14.98	<0.01
employees	0.0000	0.0000	17.62	<0.01
Innovation	0.0906	0.0045	20.22	<0.01
Region_2	-0.0209	0.0180	-1.16	0.25
Region_3	0.0022	0.0018	1.22	0.22
Region_4	-0.0020	0.0034	-0.61	0.55
Region_5	0.0139	0.0020	6.95	0.00
Region_6	0.0080	0.0029	2.77	0.01
Region_7	-0.0029	0.0029	-0.99	0.32
Region_8	-0.0053	0.0020	-2.67	<0.01
Region_9	-0.0074	0.0021	-3.59	<0.01
Region_10	-0.0198	0.0033	-6.03	<0.01
Region_11	-0.0010	0.0030	-0.33	0.75
Region_12	-0.0106	0.0019	-5.68	<0.01
Region_13	-0.0140	0.0034	-4.13	<0.01
Region_14	-0.0351	0.0070	-5.04	<0.01
Region_15	-0.0159	0.0022	-7.21	<0.01
Region_16	-0.0134	0.0027	-4.92	<0.01
Region_17	-0.0474	0.0082	-5.78	<0.01
Region_18	-0.0251	0.0037	-6.81	<0.01
Region_19	-0.0200	0.0025	-8.15	<0.01
Region_20	-0.0222	0.0030	-7.28	<0.01
Sector_17	0.0531	0.0035	15.21	<0.01
Sector_18	0.1048	0.0032	32.34	<0.01
Sector_19	0.1542	0.0036	43.22	<0.01
Sector_26	0.0368	0.0052	7.13	<0.01
Sector_29	0.0181	0.0065	2.78	<0.01
Sector_36	0.0446	0.0060	7.48	<0.01
Constant	-0.1825	0.0097	-18.78	<0.01

Note: In Appendix 1 we report the list of regions and sectors

**Table 3.** Longitudinal Multilevel Model Estimates (level 1: time, level 2: sector, level 3: region.)

factor that positively affects the “Made in Italy” firms performance. This result is extremely interesting because, as often discussed in the literature, it confirms that these regions (North-East regions) represent a fertile context for firms performance in the period 1999-2005. Moreover, disaggregating the role of sector, although all “Made in Italy” sectors perform quite well in the period considered, being in the

Fashion and Leather sectors represents an extra-positive factor for exporting firms. This is in line with several studies on “Made in Italy” boom that show the surprising performance of these sectors in international markets.

## 6 Conclusive Remarks

Italy’s competitiveness during the last decade has shown a strong slowdown: to compete in international markets Italian firms reduced their costs instead of fostering on innovation and investments. Although the whole economy has declined, a bunch of sectors, the so called “Made in Italy” sectors, has improved its role in international markets. In a sort of “polarization” of the economy, “Made in Italy” sectors have become worldwide famous and successful while the whole Italian economy was suffering from lost competitiveness. This paper investigates the factors affecting the export competitiveness of “Made in Italy” sectors in the period 1999-2005 at a firm level, using a longitudinal multilevel approach to simultaneously model individual and context factors that affect the firms export competitiveness in presence of hierarchical data. We show that there are a few factors driving the growth and success of these sectors over the period and they are strictly related to firms’ innovative capacity and to firms’ strategies on labor (both in terms of productivity and costs). Also, “Made in Italy” firms benefit from social capital that spill over the industrial districts improving their role in international markets. In particular, some sectors and regions turn out to be the most intriguing elements of this counter-cyclical phenomenon: firms in the North-East and in specific sectors, like fashion and leather, take advantage of a mix of experiences and skills that strongly help firms’ export competitiveness.

## References

- BECATTINI G., (2007), *Il Calabrone Italia. Ricerche e ragionamenti sulla peculiarit  economica italiana*, Bologna, Il Mulino.
- BRANDOLINI A. and CIPOLLONE P., (2003), Una nuova economia in Italia, in: S. Rossi (Ed.) *La Nuova Economia. I fatti dietro il mito*, Bologna, Il Mulino.
- CASTELLANI D. and ZANFEI A., (2007), Internationalisation, Innovation and Productivity: How Do Firms Differ in Italy?, *The World Economy*, 157-175.
- CONTI G. and MENGHINELLO S., (1996) , L’internazionalizzazione produttiva dei “sistemi locali”, *Rapporto ICE*, Rome.
- GIOVANNETTI G., RICCHIUTI G. and VELUCCHI M.,(2009), Size, Innovation and Internationalization: A Survival Analysis of Italian Firms, *Applied Economics*, forthcoming.
- HOX J.J. and MAAS C.J.M., (2005), Multilevel Analysis. In: K. Kempf-Leonard (Ed.), *The Encyclopedia of Social Measurement*. San Diego: Elsevier Academic Press, 785-793.
- ICE, Istituto Nazionale per il Commercio Estero, (2005), *La posizione competitiva dell’Italia nell’economia internazionale*, Italian International Trade Center, Rome.
- ISTAT, (2007), *Rapporto Annuale*, Roma.

- MAYER T. and OTTAVIANO G.M., (2007), The Happy Few: The Internationalization of European Firms, *Bruegel blueprint series*, n. 3.
- Organization for Economic Cooperation and Development (OECD), (2005), *Italy Country Report*, OECD Economic Surveys, Paris.
- OROPALLO F., (2007), Entrepreneurs' Behaviour and Performance: An Empirical Analysis on Italian Firms, *Rivista di Politica Economica*, 97, 3, 133-150.
- RABELLOTTI R., CARABELLI A. and HIRSCH G., (2009), Italian Industrial Districts on the Move: Where Are They Going?, *European Planning Studies*, 17:1, 19 -41.
- SKRONDAL A. and RABE-HESKET S., (2004), *Generalized latent variable modeling*, Chapman & Hall/CRC.
- YANG M. and GOLDSTEIN H. (1996), Multilevel Models for Longitudinal Data, in: Engel, U., Reinecke, J. (Eds.), *Analysis of Change. Advanced Techniques in Panel data Analysis*, Berlin, Walter de Gruyter.
- ZELI A. and MARIANI P., (2009), Productivity and profitability analysis of large Italian companies: 1998–2002, *International Review of Economics*, 56, 175–188.

## Appendix: Regions and Sectors

Istat Code	Region
1	Piemonte
2	Valle D'Aosta
3	Lombardia
4	Trentino Alto Adige
5	Veneto
6	Friuli Venezia Giulia
7	Liguria
8	Emilia-Romagna
9	Toscana
10	Umbria
11	Marche
12	Lazio
13	Abruzzi
14	Molise
15	Campania
16	Puglia
17	Basilicata
18	Calabria
19	Sicilia
20	Sardegna
ATECO Code	Sector
17	Food
18	Apparel
19	Fashion
26	Leather
29	Mechanics
36	Furniture



# Statistical Inference on Large Contingency Tables: Convergence, Testability, Stability

Marianna Bolla

Institute of Mathematics, Budapest University of Technology and Economics  
H-1111. Egrý József u. 1, Budapest, Hungary, *marib@math.bme.hu*

**Abstract.** Convergence of rectangular arrays with nonnegative, bounded entries is defined together with the limit object and cut distance. A statistic defined on a contingency table is testable if it can be consistently estimated based on a smaller, but still sufficiently large table which is selected randomly from the original one in an appropriate manner. By the above randomization, classical multivariate methods can be carried out on a smaller part of the array. This fact becomes important when our task is to discover the structure of large and evolving arrays, like genetic maps, social, and communication networks. Special block structures behind large tables are also discussed from the point of view of stability and spectra.

**Keywords:** convergence of contingency tables, testable contingency table parameters, block matrices, spectrum and stability

## 1 Introduction

In order to discover the structure of large rectangular arrays, e.g., microarrays, social, economic, or communication networks, classical methods of cluster and correspondence analysis may not be carried out on the whole table because of computational size limitations. In other situations, we want to compare contingency tables of different sizes. For basic notions see Section 2.

For the above causes, convergence and distance of general normalized arrays is introduced in Section 3. Roughly speaking, a sequence of contingency tables converges if their global structure becomes more and more similar which fact will be formulated in terms of the convergence of homomorphism densities of maps taking small 0-1 “probe” tables into the large one. The limit object is a measurable, bounded function on  $[0, 1]^2$  which can be regarded as generalization of graph limits, cf. Lovász and Szegedy (2006). As such a convergent sequence of contingency tables is also a Cauchy sequence in the so-called cut metric, we are able to define distance between contingency tables of different sizes. Relation to the Aldous–Hoover Representation Theorem (see Diaconis and Janson (2008)) is also discussed.

In Section 4, testable contingency table parameters are defined. In fact, they are statistics that can be consistently estimated based on a fairly large sample. Most parameters based on spectral and balanced classification properties of the table are testable. Hence, classical methods of variance, factor, or cluster analysis can be carried out on a smaller part of the table, obtained by an appropriate random selection of the rows and columns.

In Section 5, we generalize the famous Szemerédi's Regularity Lemma (see Frieze and Kannan (1999), Borgs et al. (2008)) to rectangular arrays. In this form, the theorem states that any  $m \times n$  rectangular array can be approximated by a matrix having a special block structure, where the number of blocks does not depend on  $m$  and  $n$ , it merely depends on the accuracy of the approximation. If the number of blocks is relatively small, both the original and the (correspondence) transformed table will have as many structural singular values as the rank of the block matrix, see Bolla et al. (2010). In the  $m = n$ , but not necessarily symmetric case, it is true for the number of structural eigenvalues too, the real parts of which determine the stability of the system, cf. May (1972), Erdi and Tóth (1990), Juhász (1996).

## 2 Preliminaries

Let  $C = C_{m \times n}$  be a contingency table of row set  $Row_C = \{1, \dots, m\}$  and column set  $Col_C = \{1, \dots, n\}$ . The nonnegative, real entries  $c_{ij}$ 's are interactions between the rows and columns, and they are normalized such that  $0 \leq c_{ij} \leq 1$ . Sometimes we have *binary* tables of entries 0 or 1. We may assign positive weights  $\alpha_1, \dots, \alpha_m$  to the rows and  $\beta_1, \dots, \beta_n$  to the columns expressing individual importance of the categories embodied by the rows and columns. (In correspondence analysis, these are the row- and column-sums.) A contingency table is called *simple* if all the row- and column-weights are equal to 1. Assume that  $C$  does not contain identically zero rows or columns, moreover  $C$  is dense in the sense that the number of nonzero entries is comparable with  $mn$ . Let  $\mathcal{C}$  denote the set of such tables (with any natural numbers  $m$  and  $n$ ).

Consider a simple binary table  $F_{a \times b}$  and maps  $\Phi : Row_F \rightarrow Row_C, \Psi : Col_F \rightarrow Col_C$ ; further

$$\alpha_\Phi := \prod_{i=1}^a \alpha_{\Phi(i)}, \quad \beta_\Psi := \prod_{j=1}^b \beta_{\Psi(j)}, \quad \alpha_C := \sum_{i=1}^m \alpha_i, \quad \beta_C := \sum_{j=1}^n \beta_j.$$

**Definition 1.** The  $F \rightarrow C$  homomorphism density is

$$t(F, C) = \frac{1}{(\alpha_C)^a (\beta_C)^b} \sum_{\Phi, \Psi} \alpha_\Phi \beta_\Psi \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)}.$$

If  $C$  is simple, then  $t(F, C) = \frac{1}{m^a n^b} \sum_{\Phi, \Psi} \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)}$ . In addition, if  $C$  is binary too, then  $t(F, C)$  is the probability that a random map  $F \rightarrow C$  is a homomorphism (preserves the 1's). The maps  $\Phi$  and  $\Psi$  correspond to sampling  $a$  rows and  $b$  columns out of  $Row_C$  and  $Col_C$  with replacement, respectively. In case of simple  $C$  it means uniform sampling, otherwise the rows and columns are selected with probabilities proportional to their weights.

To sampling without replacement, injective maps  $\Phi, \Psi$  correspond.

**Definition 2.** The injective and induced homomorphism densities of  $F \rightarrow C$  are

$$t_{inj}(F, C) = \frac{1}{(\alpha)_a a! (\beta)_b b!} \sum_{\Phi, \Psi \text{ inj.}} \alpha_\Phi \beta_\Psi \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)} \quad \text{and}$$

$$t_{ind}(F, C) = \frac{1}{(\alpha)_a a! (\beta)_b b!} \sum_{\Phi, \Psi \text{ inj.}} \alpha_\Phi \beta_\Psi \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)} \prod_{f_{ij}=0} (1 - c_{\Phi(i)\Psi(j)}),$$

where  $(\alpha)_a$  and  $(\beta)_b$  denote the  $a$ th and  $b$ th elementary symmetric polynomials of  $\alpha_1, \dots, \alpha_m$  and  $\beta_1, \dots, \beta_n$ , respectively.

For simple  $C$ ,  $(\alpha)_a = \binom{m}{a}$  and  $(\beta)_b = \binom{n}{b}$ . Clearly,  $t_{inj}(F, C)$  and  $t_{ind}(F, C)$  are zeroes if  $a \geq m$  or  $b \geq n$ . Typically,  $a$  and  $b$  are much smaller than  $m$  and  $n$ . As most maps into a large table are injective,  $t(F, C)$  and  $t_{inj}(F, C)$  are very close to each other. Namely, for simple  $C$ ,  $|t(F, C) - t_{inj}(F, C)| \leq \frac{ab}{m+n}$  that tends to zero for fixed  $a$  and  $b$  as  $m, n \rightarrow \infty$ . For not simple  $C$  the above difference also tends to zero if we assume that there are not dominant rows and columns in  $C$  in the sense that  $\max_i \frac{\alpha_i}{\alpha_C} \rightarrow 0$  and  $\max_j \frac{\beta_j}{\beta_C} \rightarrow 0$  as  $m \rightarrow \infty$  and  $n \rightarrow \infty$ , respectively.

The following simple binary random table  $\xi(a \times b, C)$  will play an important role in proving the theorems of Section 4. Select  $a$  rows and  $b$  columns of  $C$  with replacement, with probabilities  $\alpha_i/\alpha_C$  ( $i = 1, \dots, m$ ) and  $\beta_j/\beta_C$  ( $j = 1, \dots, n$ ), respectively. If the  $i$ th row and  $j$ th column of  $C$  are selected, they will be connected by 1 with probability  $c_{ij}$  and 0, otherwise, independently of the other selected row-column pairs, conditioned on the selection of the rows and columns. For large  $m$  and  $n$ ,  $\mathbb{P}(\xi(a \times b, C) = F)$  is very close to  $t_{ind}(F, C)$  that is reminiscent of a likelihood function.

### 3 Convergence of contingency tables

**Definition 3.** We say that the sequence  $(C_{m \times n})$  of contingency tables is convergent if the sequence  $t(F, C_{m \times n})$  converges for any simple binary table  $F$  as  $m, n \rightarrow \infty$ .

In view of Section 2, the convergence of  $t(F, C_{m \times n})$  is equivalent to the convergence of  $t_{inj}(F, C_{m \times n})$  and  $t_{ind}(F, C_{m \times n})$ , as well. The convergence means that the tables  $C_{m \times n}$  become more and more similar in small details as they are probed by smaller 0-1 tables ( $m, n \rightarrow \infty$ ).

The limit object is a measurable function  $U : [0, 1]^2 \rightarrow [0, 1]$  and we call it *contingon*. In the  $m = n$  and symmetric case,  $C$  can be regarded as the weight matrix of an edge- and node-weighted graph (the row-weights are equal to the column-weights, loops are possible) and the limit object was introduced as graphon, see Lovász and Szegedy (2006). The step-function contingon  $U_C$  is assigned to  $C$  in the following way: the sides of the unit square are divided into intervals  $I_1, \dots, I_m$  and  $J_1, \dots, J_n$  of lengths  $\alpha_1/\alpha_C, \dots, \alpha_m/\alpha_C$  and  $\beta_1/\beta_C, \dots, \beta_n/\beta_C$ , respectively; then over the rectangle  $I_i \times J_j$  the step-function takes on the value  $c_{ij}$ .

In fact, the above convergence of contingency tables can be formulated in terms of a special distance. First we define it for contingons.

**Definition 4.** The cut distance between the contingons  $U$  and  $V$  is

$$\delta_{\square}(U, V) = \inf_{\mu, \nu} \|U - V^{\mu, \nu}\|_{\square} \quad (1)$$

where the cut norm of the contingon  $U$  is defined by

$$\|U\|_{\square} = \sup_{S, T \subset [0, 1]} \left| \iint_{S \times T} U(x, y) dx dy \right|,$$

and the infimum in (1) is taken over all measure preserving bijections  $\mu, \nu : [0, 1] \rightarrow [0, 1]$ , while  $V^{\mu, \nu}$  denotes the transformed  $V$  after performing the measure preserving bijections  $\mu$  and  $\nu$  on the sides of the unit square, respectively.

An equivalence relation is defined over the set of contingons: two contingons belong to the same class if they can be transformed into each other by measure preserving map, i.e., their cut distance is zero. In the sequel, we consider contingons modulo measure preserving maps, and under contingency we understand the whole equivalence class. By a theorem of Borgs et al. (2008), the equivalence classes form a compact metric space with the  $\delta_{\square}$  metric.

**Definition 5.** The cut distance between the contingency tables  $C, C' \in \mathcal{C}$  is

$$\delta_{\square}(C, C') = \delta_{\square}(U_C, U_{C'}).$$

By the above remarks, the distance of  $C$  and  $C'$  is indifferent to permutations of the rows or columns of  $C$  and  $C'$ . In the special case when  $C$  and  $C'$  are of the same size,  $\delta_{\square}(C, C')$  is  $\frac{1}{mn}$  times the usual cut distance of matrices, cf. Frieze and Kannan (1999).

The following reversible relation between convergent contingency table sequences and contingons also holds, as a rectangular analogue of a theorem of Borgs et al. (2008).

**Theorem 1.** *For any convergent sequence  $(C_{m \times n}) \subset \mathcal{C}$  there exists a contingency such that  $\delta_{\square}(U_{C_{m \times n}}, U) \rightarrow 0$  as  $m, n \rightarrow \infty$ . Conversely, any contingency can be obtained as the limit of a sequence of contingency tables in  $\mathcal{C}$ . The limit of a convergent contingency table sequence is essentially unique: if  $C_{m \times n} \rightarrow U$ , then also  $C_{m \times n} \rightarrow U'$  for precisely those contingons  $U'$  for which  $\delta_{\square}(U, U') = 0$ .*

It also follows that a sequence of contingency tables in  $\mathcal{C}$  is convergent if, and only if it is a Cauchy sequence in the metric  $\delta_{\square}$ .

A simple binary random  $a \times b$  table  $\xi(a \times b, U)$  can also be randomized based on the contingency  $U$  in the following way. Let  $X_1, \dots, X_a$  and  $Y_1, \dots, Y_b$  be i.i.d., uniformly distributed random numbers on  $[0, 1]$ . The entries of  $\xi(a \times b, U)$  are independent Bernoulli random variables, namely the entry in the  $i$ th row and  $j$ th column is 1 with probability  $U(X_i, Y_j)$  and 0, otherwise. It is easy to see that the distribution of the previously defined  $\xi(a \times b, C)$  and that of  $\xi(a \times b, U_C)$  is the same. Further,  $\delta_{\square}(C_{m \times n}, \xi(a \times b, C_{m \times n}))$  tends to 0 in probability, for fixed  $a$  and  $b$  as  $m, n \rightarrow \infty$ . This fact also plays an important role in proving the theorems of Section 4.

Note, that in the above way, we can as well randomize an infinite simple binary table  $\xi(\infty \times \infty, U)$  out of the contingency  $U$  by generating countably infinitely many i.i.d. uniform random numbers on  $[0, 1]$ . The distribution of the infinite binary array  $\xi(\infty \times \infty, U)$  is denoted by  $\mathbb{P}_U$ . Because of the symmetry of the construction, this is an *exchangeable* array in the sense that the joint distribution of its entries is invariant under permutations of the rows and columns. Moreover, any exchangeable binary array is a mixture of such  $\mathbb{P}_U$ 's. More precisely, the Aldous–Hoover Representation Theorem (see Diaconis and Janson (2008)) states that for every infinite exchangeable binary array  $\xi$  there is a probability distribution  $\mu$  (over the contingons) such that  $\mathbb{P}(\xi \in A) = \int \mathbb{P}_U(A) \mu(dU)$ .

## 4 Testable contingency table parameters

A function  $f : C \rightarrow \mathbb{R}$  is called a *contingency table parameter* if it is invariant under isomorphism and scaling of the rows/columns. In fact, it is a statistic evaluated on the table, and hence, we are interested in contingency table parameters that are not sensitive to minor changes in the entries of the table.

**Definition 6.** A contingency table parameter  $f$  is testable if for every  $\varepsilon > 0$  there are positive integers  $a$  and  $b$  such that if the row- and column-weights of  $C$  satisfy

$$\max_i \frac{\alpha_i}{\alpha_C} \leq \frac{1}{a}, \quad \max_j \frac{\beta_j}{\beta_C} \leq \frac{1}{b}, \quad (2)$$

then

$$\mathbb{P}(|f(C) - f(\xi(a \times b, C))| > \varepsilon) \leq \varepsilon.$$

Consequently, such a contingency table parameter can be consistently estimated based on a fairly large sample. Now, we introduce some equivalent statements of the testability, indicating that a testable parameter depends continuously on the whole table. This is the generalization of a theorem of Borgs et al. (2008) applicable to simple graphs.

**Theorem 2.** For a testable contingency table parameter  $f$  the following are equivalent:

- For every  $\varepsilon > 0$  there are positive integers  $a$  and  $b$  such that for every contingency table  $C \in \mathcal{C}$  satisfying the condition (2),

$$|f(C) - \mathbb{E}(f(\xi(a \times b, C)))| \leq \varepsilon.$$

- For every convergent sequence  $(C_{m \times n})$  of contingency tables with no dominant row- or column-weights,  $f(C_{m \times n})$  is also convergent ( $m, n \rightarrow \infty$ ).
- $f$  is continuous in the cut distance.

For example, in case of simple binary tables the singular spectrum is testable, as  $C_{m \times n}$  can be regarded as part of the adjacency matrix of a bipartite graph on  $m+n$  vertices, where  $Row_C$  and  $Col_C$  are the two independent vertex sets; further, the  $i$ th vertex of  $Row_C$  and the  $j$ th vertex of  $Col_C$  are connected by an edge if and only if  $c_{ij} = 1$ . The non-zero real eigenvalues of the symmetric  $(m+n) \times (m+n)$  adjacency matrix of this bipartite graph are the numbers  $\pm s_1, \dots, \pm s_r$ , where  $s_1, \dots, s_r$  are the non-zero singular values of  $C$ , and  $r \leq \min\{m, n\}$  is the rank of  $C$ . Consequently, the convergence of adjacency spectra implies the convergence of the singular spectra. Therefore, by Theorem 2, any property of a large contingency table based on its singular value decomposition (e.g., correspondence decomposition) can be concluded from a smaller part of it. In Section 5, testability of some balanced classification properties is discussed.

## 5 Homogeneous partitions, spectra, and stability

Now, we shall prove that special blown up tables burdened with a general kind of noise are convergent.

**Definition 7.** The  $m \times n$  random matrix  $E$  is a noise matrix if its entries are independent, uniformly bounded random variables of zero expectation.

**Theorem 3.** *The cut norm of any sequence  $(E_{m \times n})$  of noise matrices tends to zero as  $m, n \rightarrow \infty$ , almost surely.*

**Definition 8.** The  $m \times n$  real matrix  $B$  is a blown up matrix, if there is an  $a \times b$  so-called *pattern matrix*  $P$  with entries  $0 \leq p_{ij} \leq 1$ , and there are positive integers  $m_1, \dots, m_a$  with  $\sum_{i=1}^a m_i = m$  and  $n_1, \dots, n_b$  with  $\sum_{j=1}^b n_j = n$ , such that the matrix  $B$ , after rearranging its rows and columns, can be divided into  $a \times b$  blocks, where block  $(i, j)$  is an  $m_i \times n_j$  matrix with entries all equal to  $p_{ij}$  ( $1 \leq i \leq a$ ,  $1 \leq j \leq b$ ).

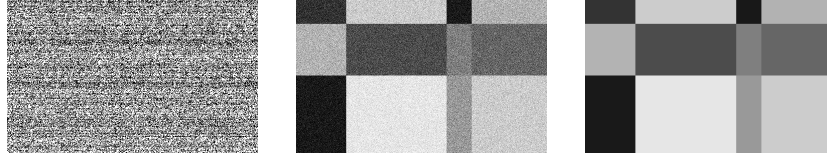
Let us fix the matrix  $P_{a \times b}$ , blow it up to obtain matrix  $B_{m \times n}$ , and let  $A_{m \times n} = B + E$ , where  $E_{m \times n}$  is a noise matrix. If the block sizes grow proportionally, the following almost sure statements are proved in Bolla et. al (2010): the noisy matrix  $A$  has as many structural (protruding) singular values of order  $\sqrt{mn}$  as the rank of the pattern matrix, all the other singular values are of order  $\sqrt{m+n}$ ; further, by representing the rows and columns by means of the singular vector pairs corresponding to the structural singular values, the  $a$ - and  $b$ -variances of the representatives tend to 0 as  $m, n \rightarrow \infty$ . Conversely, in the presence of structural singular values, with some additional conditions for the representatives, the block structure can be recovered.

**Theorem 4.** *Let the block sizes of the blown up matrix  $B_{m \times n}$  are  $m_1, \dots, m_a$  horizontally, and  $n_1, \dots, n_b$  vertically ( $\sum_{i=1}^a m_i = m$  and  $\sum_{j=1}^b n_j = n$ ). Let  $A_{m \times n} := B + E$  and  $m, n \rightarrow \infty$  is such a way that  $m_i/m \rightarrow r_i$  ( $i = 1, \dots, a$ ),  $n_j/n \rightarrow q_j$  ( $j = 1, \dots, b$ ), where  $r_i$ 's and  $q_j$ 's are fixed ratios. Under these conditions, the "noisy" sequence  $(A_{m \times n})$  converges almost surely.*

In many applications we are looking for clusters of the rows and columns of a rectangular array such that the densities within the cross-products of the clusters be homogeneous. E.g., in microarray analysis we are looking for clusters of genes and conditions such that genes of the same cluster equally influence conditions of the same cluster. The following theorem ensures the existence of such a structure with possibly many clusters. However, the number of clusters does not depend on the size of the array, it merely depends on the accuracy of the approximation.

**Theorem 5.** *For every  $\varepsilon > 0$  and  $C_{m \times n} \in \mathcal{C}$  there exists a blown up matrix  $B_{m \times n}$  of an  $a \times b$  pattern matrix with  $a + b \leq 4^{1/\varepsilon^2}$  (independently of  $m$  and  $n$ ) such that  $\delta_{\square}(C, B) \leq \varepsilon$ .*

The theorem is a consequence of the Szemerédi's Regularity Lemma (see Frieze and Kannan (1999), Borgs et al. (2008)) and can be proved by embedding  $C$  into the



**Fig. 1.** noisy table (a); table close to the limit (b); approximation by SVD (c).

adjacency matrix of an edge-weighted bipartite graph. The statement of the theorem is closely related to the testability of the following contingency table parameter:

$$S_{a,b}^2(C) = \min \sum_{i=1}^a \sum_{j=1}^b \sum_{k \in A_i} \sum_{l \in B_j} (c_{kl} - \bar{c}_{i,j})^2, \quad \bar{c}_{i,j} = \frac{1}{|A_i| \cdot |B_j|} \sum_{k \in A_i} \sum_{l \in B_j} c_{kl},$$

where the minimum is taken over balanced  $a$ - and  $b$ -partitions  $A_1, \dots, A_a$  and  $B_1, \dots, B_b$  of  $Row_C$  and  $Col_C$ , respectively; further, instead of  $c_{kl}$  we may take  $\alpha_k \beta_l c_{kl}$  in the row- and column-weighted case, provided there are no dominant rows/columns.

We applied our spectral partitioning algorithm for mixture of noisy data. Figure 1 shows the original  $300 \times 500$  contingency table (a); the  $1500 \times 2500$  blown up table close to the limit, with rows and columns sorted with respect to their cluster memberships obtained by k-means algorithm (b); eventually, the colour illustration of the average densities of the blocks formed by SVD (c).

The Gardner–Ashby’s connectance  $c_n$  of a not necessarily symmetric array  $A_{n \times n}$  is the percentage of nonzero entries in the matrix, that is the ratio of actual row-column interactions to all possible ones in the network. In social and ecological models, a random array  $A_{n \times n}$  of independent entries is considered. Suppose that the entries have symmetric distribution (consequently, zero expectation) and common variance  $\sigma_n^2$ , where  $\sigma_n$  is called average interaction strength. The stability of the system is characterized by the stability of the equilibrium solution 0 of the differential equation  $dx/dt = A_{n \times n}x$  (sometimes this is achieved by linearization techniques in the neighbourhood of the equilibrium solution). Based on Wigner’s famous semicircle law, May (1972) proves that the equilibrium solution is stable in the  $\sigma_n^2 n c_n < 1$ , and unstable in the  $\sigma_n^2 n c_n > 1$  case; further, the transition region between stability and instability becomes narrow as  $n \rightarrow \infty$ . Hence, it seems that high connectance and high interaction strength destroy stability, but only in this simple model. If  $A_{n \times n}$  is a block matrix, like a noisy matrix before, it has some structural, possibly complex eigenvalues, cf. Juhász (1996). If all their real parts are negative, the system is stable, see Érdi and Tóth (1990). In fact, in many natural ecosystems and other networks the interactions are arranged in blocks, at least an approximation of Theorem 5 works.

## References

- BOLLA, M., FRIEDL, K., and KRÁMLI, A. (2010): Singular value decomposition of large random matrices (for two-way classification of microarrays). *Journal of Multivariate Analysis* 101, 434–446.

- BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T., and VESZTERGOMBI, K. (2008): Convergent sequences of dense graphs I, subgraph frequencies, metric properties and testing. *Advances in Mathematics* 219, 1801-1851.
- DIACONIS, P. and JANSON, S. (2008): Graph limits and exchangeable random graphs. *Rendiconti di Matematica* 28 (Serie VII), 33-61.
- ÉRDI, P. and TÓTH, J. (1990): What is and what is not stated by the May-Wigner theorem? *J. Theor. Biol.* 145, 137-140.
- FRIEZE, A. and KANNAN, R. (1999): Quick approximation to matrices and applications. *Combinatorica* 19, 175-220.
- JUHÁSZ, F. (1996): On the structural eigenvalues of block random matrices. *Linear Algebra and Its Applications* 246, 225-231.
- LOVÁSZ, L. and SZEGEDY, B. (2006): Limits of dense graph sequences. *J. Comb. Theory B* 96, 933-957.
- MAY, R. M. (1972): Will a large complex system be stable? *Nature* 238, 413-414.



# A Class of Multivariate Type I Generalized Logistic Distributions

Salvatore Bologna

Department of Statistical and Mathematical Sciences, University of Palermo  
Faculty of Economics-Viale delle Scienze, Palermo, Italy, *bologna@unipa.it*

**Abstract.** The logistic distribution has found important applications in many different fields and several different forms of generalizations have been proposed in the literature. However it seems, with a few exceptions, that there are not in the literature forms of multivariate generalized logistic distributions. In this paper we focus on the type I generalized logistic distribution and, based on a procedure of multivariate transformation of multivariate exponential distributions, we introduce a class of multivariate type I generalized logistic distributions. We provide some examples of bivariate and multivariate distributions of this class.

**Keywords:** type I generalized logistic distribution, multivariate exponential distributions, multivariate type I generalized logistic distributions

## 1 Introduction

The logistic distribution arises as a natural model in numerous situations and has found important applications in many different fields, such as Biology, Psychology, Technology, Market. Several different forms of generalized logistic distributions have been proposed with various motivations in the literature and they are, usually, generically named (see, e.g., Johnson et al. (1995)): type I, type II, type III, type IV generalized logistic distribution. Most of these distributions are of theoretical interest, but type I has received additional attention in estimating its parameters for practical usage (Zelterman and Balakrishnan (1992)). In the literature we find applications of this distribution, for example, in factorial designs, as a distribution of error terms to evaluate two or more factors simultaneously, or in modelling the log odds ratio when the probability of an event is very small. Recently Gupta and Kundu (2010) have proposed an interpretation of type I generalized logistic distribution as a "proportional reversed hazard logistic distribution". However it seems, with a few exceptions (see Kotz et al. (2000)), that there are not in the literature forms of multivariate generalized logistic distributions to define dependence structures between random variables with some form of generalized logistic distribution. In this paper we focus on the type I generalized logistic distribution and, based on a procedure of multivariate transformation of multivariate exponential distributions, we introduce a class of multivariate distributions with marginals which are type I generalized logistic. We provide two examples of bivariate distributions belonging to this class, generated by two different forms of bivariate exponential distributions, that is Gumbel's model I and Gumbel's model II, and, also, an example of multivariate type I generalized logistic distribution generated by a multivariate exponential distribution.

## 2 A class of multivariate type I generalized logistic distributions

The probability density function of a type I generalized logistic random variable  $X$  is given by

$$f(x) = \frac{\alpha e^{-x}}{(1 + e^{-x})^{\alpha+1}}, \quad -\infty < x < +\infty, \quad \alpha > 0. \quad (1)$$

This density is negatively skewed for  $0 < \alpha < 1$  and positively skewed for  $\alpha > 1$ , and therefore  $\alpha$  can be termed as a skewness parameter. Expression (1) is the standard form of the type I generalized logistic distribution, but a location parameter and a scale parameter can be introduced. We propose a procedure of construction of multivariate versions of the distribution (1).

First we note that, as it can be easily shown by standard calculation, if the random variable  $X$  has a type I generalized logistic distribution with  $\alpha$  parameter, then the transformed random variable  $Y = \log(1 + e^{-X})^\alpha$  has a standard exponential distribution, that is  $g(y) = e^{-y}$ ,  $y > 0$ . Let  $X_1, \dots, X_k$  be random variables with probability density functions  $f(x_1), \dots, f(x_k)$  of the form (1) and skewness parameters  $\alpha_1, \dots, \alpha_k$  respectively. Let  $Y_1, \dots, Y_k$  be the random variables defined by the transformation  $Y_i = \alpha_i \log(1 + e^{-X_i})$  of  $X_i$ ,  $i = 1, \dots, k$ , and, therefore, each with a standard exponential distribution. Consider the random vector  $\mathbf{Y} = (Y_1, \dots, Y_k)^T$  with components  $Y_i$  defined above. Assume that  $\mathbf{Y}$  has some form of standard multivariate exponential distribution and denote with  $\psi_{\mathbf{Y}}(y_1, \dots, y_k)$ ,  $y_i > 0$ ,  $i = 1, \dots, k$ , the corresponding probability density function. Consider also the following one-to-one transformation from the  $k$ -dimensional space  $\mathfrak{R}_+^k$  to the  $k$ -dimensional space  $\mathfrak{R}^k$ :

$$x_i = g_i(y_i) = -\log(e^{\frac{y_i}{\alpha_i}} - 1), \quad i = 1, \dots, k, \quad (2)$$

so that the transformed random variables  $X_i = g_i(Y_i)$  define a new random vector  $\mathbf{X} = (X_1, \dots, X_k)^T$ . Denoting by  $y_i = g_i^{-1}(x_i)$ ,  $i = 1, \dots, k$ , the inverse  $k$ -dimensional transformation of (2) from  $\mathfrak{R}^k$  to  $\mathfrak{R}_+^k$ , then the probability density function of  $\mathbf{X} = (X_1, \dots, X_k)^T$  can be written

$$f_{\mathbf{X}}(x_1, \dots, x_k) = |J| \psi_{\mathbf{Y}} [g_1^{-1}(x_1), \dots, g_k^{-1}(x_k)], \quad (3)$$

where  $J$  is the Jacobian of the transformation, with

$$|J| = \left| \frac{\partial g_1^{-1}(x_1)}{\partial (x_1)} \times \dots \times \frac{\partial g_k^{-1}(x_k)}{\partial (x_k)} \right|.$$

The marginal probability density function  $f_{X_i}(x_i)$  of each component  $X_i$  of the vector  $\mathbf{X}$  is, as it can be easily shown, of the form (1). Thus, for example, the density of  $X_1$ ,

$$f_{X_1}(x_1) = \int_{\mathfrak{R}^{k-1}} |J| \psi_{\mathbf{Y}} [g_1^{-1}(x_1), \dots, g_k^{-1}(x_k)] dx_2 \dots dx_k,$$

can be easily determined, by standard calculation, making the following change of variables for multiple integrals:  $y_i = g_i^{-1}(x_i)$ ,  $i = 2, \dots, k$ . In this way we obtain that the marginal distribution of  $X_1$  is a type I generalized logistic distribution of the form (1) with skewness parameter  $\alpha_1$ . Analogous results are obtained for any marginal variable  $X_i$ ,  $i = 1, \dots, k$ , and we can define (3) as a "multivariate type

I generalized logistic distribution". Of course any possible different form of multivariate exponential distribution, with univariate marginal standard exponential distributions, will generate, by means of the multivariate transformation (2), a different form of multivariate type I generalized logistic distribution, thus defining a class of multivariate distributions.

### 3 Some examples of bivariate and multivariate type I generalized logistic distributions

We provide some examples of construction of bivariate and multivariate type I generalized logistic distributions according to the above procedure. There are indeed many classes of bivariate and multivariate exponential distributions having standard exponential marginals and we focus on bivariate "Gumbel's model I" and "Gumbel's model II", and on a special form of multivariate exponential distribution (Kotz et al. (2000)).

#### 3.1 Example I

Consider the bivariate exponential random vector  $\mathbf{Y} = (Y_1, Y_2)^T$  with probability density function expressed by the Gumbel's model I

$$\psi_{Y_1, Y_2}(y_1, y_2) = e^{-(y_1 + y_2 + \beta y_1 y_2)} [(1 + \beta y_1)(1 + \beta y_2) - \beta], y_1, y_2 > 0, 0 \leq \beta \leq 1,$$

with standard exponential marginal distributions. Consider, also, the bivariate random vector  $\mathbf{X} = (X_1, X_2)^T$ , with components

$$X_1 = g_1(Y_1) = -\log(e^{\frac{Y_1}{\alpha_1}} - 1), \quad X_2 = g_2(Y_2) = -\log(e^{\frac{Y_2}{\alpha_2}} - 1),$$

defined by the bivariate transformation

$$x_1 = g_1(y_1) = -\log(e^{\frac{y_1}{\alpha_1}} - 1), \quad x_2 = g_2(y_2) = -\log(e^{\frac{y_2}{\alpha_2}} - 1), \quad \alpha_1, \alpha_2 > 0. \quad (4)$$

We seek the bivariate distribution of  $\mathbf{X} = (X_1, X_2)^T$ . The probability density function of  $\mathbf{X} = (X_1, X_2)^T$  is expressed by

$$f_{X_1, X_2}(x_1, x_2) = |J| \psi_{Y_1, Y_2}(g_1^{-1}(x_1), g_2^{-1}(x_2)).$$

As  $y_1 = g_1^{-1}(x_1) = \alpha_1 \log(1 + e^{-x_1})$ ,  $y_2 = g_2^{-1}(x_2) = \alpha_2 \log(1 + e^{-x_2})$ , we have that

$$|J| = \left| \frac{\partial g_1^{-1}(x_1)}{\partial(x_1)} \right| \times \left| \frac{\partial g_2^{-1}(x_2)}{\partial(x_2)} \right| = \frac{\alpha_1 \alpha_2 e^{-x_1} e^{-x_2}}{(1 + e^{-x_1})(1 + e^{-x_2})},$$

and then the joint density of  $X_1$  and  $X_2$  will be

$$f_{X_1, X_2}(x_1, x_2) = \frac{\alpha_1 \alpha_2 e^{-x_1 - x_2}}{(1 + e^{-x_1})^{\alpha_1 + 1} (1 + e^{-x_2})^{\alpha_2 + 1}}$$

$$\times e^{-[\beta \alpha_1 \alpha_2 \log(1 + e^{-x_1}) \log(1 + e^{-x_2})]}$$

$$\times [(1 + \beta \alpha_1 \log(1 + e^{-x_1}))(1 + \beta \alpha_2 \log(1 + e^{-x_2})) - \beta], \quad (x_1, x_2) \in \mathbb{R}^2.$$

In this case it is immediate to verify that the marginal distributions of  $X_1$  and  $X_2$  are of the form (1). Thus, for example, the probability density function of  $X_1$ ,

$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_{X_1, X_2}(x_1, x_2) dx_2,$$

can be easily determined making the change of variable for integrals  $y_2 = \alpha_2 \log(1 + e^{-x_2})$  and obtaining an expression of the form (1) with skewness parameter  $\alpha_1$ .

### 3.2 Example II

Consider now the bivariate exponential random vector  $\mathbf{Y} = (Y_1, Y_2)^T$  with probability density function expressed by the following Gumbel's model II:

$$\psi_{Y_1, Y_2}(y_1, y_2) = e^{-y_1 - y_2} [1 + \beta(2e^{-y_1} - 1)(2e^{-y_2} - 1)], \quad y_1, y_2 > 0, \quad |\alpha| < 1,$$

which is a special form of Farlie-Gumbel-Morgenstern's bivariate distributions with standard exponential marginals. Consider, also, the bivariate random vector  $\mathbf{X} = (X_1, X_2)^T$ , with components defined by the same bivariate transformation (4) as in example I. Proceeding in exactly the same way as in example I, we derive the following joint probability density function of  $X_1$  and  $X_2$ :

$$f_{X_1, X_2}(x_1, x_2) = \frac{\alpha_1 \alpha_2 e^{-x_1 - x_2}}{(1 + e^{-x_1})^{\alpha_1 + 1} (1 + e^{-x_2})^{\alpha_2 + 1}} \times \left(1 + \beta \left[2(1 + e^{-x_1})^{-\alpha_1} - 1\right] \left[2(1 + e^{-x_2})^{-\alpha_2} - 1\right]\right), \quad (x_1, x_2) \in \mathbb{R}^2.$$

This form of bivariate distribution of the random vector  $\mathbf{X} = (X_1, X_2)^T$  is different from the one of the previous example because here we started from a different exponential bivariate distribution with respect to example I; however, the marginal distributions of  $X_1$  and  $X_2$  are, still, type I generalized logistic. So we can say that the above distributions of  $\mathbf{X} = (X_1, X_2)^T$  are "bivariate type I generalized logistic distributions".

### 3.3 Example III

Consider the multivariate exponential random vector  $\mathbf{Y} = (Y_1, \dots, Y_k)^T$  with probability density function expressed by

$$\psi_{Y_1, \dots, Y_k}(y_1, \dots, y_k) = (a + 1) \dots (a + k - 1) a^{-k+1} \left( \sum_{i=1}^k e^{\frac{Y_i}{a}} - k + 1 \right) e^{\sum_{i=1}^k \frac{Y_i}{a}},$$

$y_i > 0$ ,  $a > 0$ , and components  $Y_i$  with standard exponential marginal distributions (Kotz et al. (2000)). Let  $\mathbf{X} = (X_1, \dots, X_k)^T$  be the random vector with components

$$X_1 = g_1(Y_1) = -\log(e^{\frac{Y_1}{\alpha_1}} - 1), \dots, X_k = g_k(Y_k) = -\log(e^{\frac{Y_k}{\alpha_k}} - 1),$$

defined by the following multidimensional transformation:

$$x_1 = g_1(y_1) = -\log(e^{\frac{y_1}{\alpha_1}} - 1), \quad \dots, x_k = g_k(y_k) = -\log(e^{\frac{y_k}{\alpha_k}} - 1),$$

$\alpha_i > 0, i = 1, \dots, k$ . The probability density function of  $\mathbf{X} = (X_1, \dots, X_k)^T$  is expressed by the general form (3), with

$$|J| = \frac{\alpha_1 e^{-x_1}}{(1 + e^{-x_1})} \times \dots \times \frac{\alpha_k e^{-x_k}}{(1 + e^{-x_k})},$$

and

$$g_1^{-1}(x_1) = \alpha_1 \log(1 + e^{-x_1}), \dots, g_k^{-1}(x_k) = \alpha_k \log(1 + e^{-x_k}).$$

Thus we obtain the following expression for the probability density function of  $\mathbf{X} = (X_1, \dots, X_k)^T$ :

$$\begin{aligned} f_{X_1, \dots, X_k}(x_1, \dots, x_k) &= (a+1) \dots (a+k-1) a^{-k+1} \prod_{i=1}^k \alpha_i e^{-x_i} (1 + e^{-x_i})^{-1} \\ &\times \left( \sum_{i=1}^k (1 + e^{-x_i})^{\frac{\alpha_i}{a}} - k + 1 \right) \prod_{i=1}^k (1 + e^{-x_i})^{\frac{\alpha_i}{a}}, \quad (x_1, \dots, x_k) \in \mathbb{R}^k. \end{aligned}$$

In conclusion we can say that any possible different form of multivariate exponential distribution, with univariate marginal standard exponential distributions, can generate, correspondently, by means of the multivariate transformation (2), a different form of multivariate type I generalized logistic distribution (with type I generalized logistic marginals), thus defining a class of multivariate distributions. The properties of the distributions of this class and estimation problems are still to be studied.

#### 4 Appendix: Determination of the marginal distributions

The probability density function of the marginal variable  $X_1$  of the vector  $\mathbf{X} = (X_1, \dots, X_k)^T$  is expressed by

$$\begin{aligned} f_{X_1}(x_1) &= \int_{\mathbb{R}^{k-1}} |J| \psi_{\mathbf{Y}} [g_1^{-1}(x_1), \dots, g_k^{-1}(x_k)] dx_2 \dots dx_k \\ &= \int_{\mathbb{R}^{k-1}} \left| \frac{\partial g_1^{-1}(x_1)}{\partial(x_1)} \times \dots \times \frac{\partial g_k^{-1}(x_k)}{\partial(x_k)} \right| \psi_{\mathbf{Y}} [g_1^{-1}(x_1), \dots, g_k^{-1}(x_k)] dx_2 \dots dx_k. \end{aligned}$$

Since we can write

$$|J| = \left| \frac{\partial g_1^{-1}(x_1)}{\partial(x_1)} \right| \times \left| \frac{\partial g_2^{-1}(x_2)}{\partial(x_2)} \times \dots \times \frac{\partial g_k^{-1}(x_k)}{\partial(x_k)} \right| = |J_1| \times |J_{k-1}|,$$

(with obvious meaning of the notation), we have that

$$f_{X_1}(x_1) = |J_1| \int_{\Re^{k-1}} |J_{k-1}| \psi_{\mathbf{Y}} [g_1^{-1}(x_1), \dots, g_k^{-1}(x_k)] dx_2 \dots dx_k.$$

By considering the following change of variables for integrals:  $y_i = g_i^{-1}(x_i)$ , the Jacobian of the integral transformation will be

$$|J'_{k-1}| = \left| \frac{\partial g_2(y_2)}{\partial(y_2)} \times \dots \times \frac{\partial g_k(y_k)}{\partial(y_k)} \right|.$$

As we can write:  $|J'_{k-1}| = |J_{k-1}|^{-1}$ , making the suitable substitutions we find that the above integral provides the marginal probability density function of  $Y_1$ , which is written as  $f_{X_1}(x_1) = |J_1| \psi_{Y_1}(g_1^{-1}(x_1))$ . So the distribution of  $X_1$  is defined by the transformation  $x_1 = g_1(y_1) = -\log(e^{\frac{y_1}{\alpha_1}} - 1)$ , of the variable  $Y_1$  and, hence, it is expressed by (1). Analogously for any marginal variable  $X_i$ .

## References

- GUPTA, R.D. and KUNDU, D.: Generalized logistic distributions. *Journal of Applied Statistical Sciences*. (To appear).
- JOHNSON, N.L. KOTZ, S. and BALAKRISHNAN, N. (1995): *Continuous Univariate Distributions, vol.2*. Wiley and Sons, New York.
- KOTZ, S., BALAKRISHNAN, N., and JOHNSON, N.L. (2000): *Continuous Multivariate Distributions, vol.1*. Wiley and Sons, New York.
- ZELTERMAN, D. and BALAKRISHNAN, N. (1992): Univariate Generalized Distributions. In: N. BALAKRISHNAN (Ed.): *Handbook of the Logistic Distribution*. Marcel Dekker, New York, 209-221.

# Adaptive Mixture Discriminant Analysis for Supervised Learning with Unobserved Classes

Charles Bouveyron

Laboratoire SAMM, Université Paris 1 Panthéon–Sorbonne  
90 rue de Tolbiac, 75013 Paris, France, *charles.bouveyron@univ-paris1.fr*

**Abstract.** In supervised learning, an important issue usually not taken into account by classical methods is the possibility of having in the test set individuals belonging to a class which has not been observed during the learning phase. Classical supervised algorithms will automatically label such observations as belonging to one of the known classes in the training set and will not be able to detect new classes. This work introduces a model-based discriminant analysis method, called adaptive mixture discriminant analysis (AMDA), which is able to detect unobserved groups of points and to adapt the learned classifier to the new situation. Two EM-based procedures are proposed for the parameter estimation in an inductive and a transductive way respectively. Experimental studies will demonstrate the ability of the proposed method to deal with complex and real word problems.

**Keywords:** supervised classification, unobserved classes, adaptive learning, novelty detection, model selection

## 1 Introduction

The usual framework of supervised classification assumes that all existing classes in the data have been observed during the learning phase and does not take into account the possibility of having in the test set individuals belonging to a class which has not been observed. In particular, such a situation could occur in the case of rare classes or in the case of an evolving population. For instance, an important problem in Biology is the detection of novel species which could appear at any time resulting from structural or physiological modifications. Unfortunately, classical supervised algorithms, like support vector machines (SVM) or Linear Discriminant Analysis (LDA), will automatically label observations from a novel class as belonging to one of the known classes in the training set and will not be able to detect new classes. It is therefore important to find a way to allow the supervised classification methods to detect unobserved situations and to adapt themselves to the new configurations.

In statistical learning, the problem of classification with unobserved classes is a problem which has received very few attention. Indeed, both supervised and unsupervised classification contexts have been widely studied but intermediate situations have received less attention. We would like however to mention two related topics in statistical learning called semi-supervised classification and novelty detection. On the one hand, semi-supervised classification focuses on supervised classification with partially labeled data. Usually, unlabeled data are added to the learning data in order to improve the efficiency of the final classifier. Such an approach is

particularly useful when only few labeled observations are available for learning (applications with a high supervision cost). A good review on semi-supervised classification can be found in Krishnapuram et al. (2004). However, semi-supervised classification methods are not able to detect unobserved groups of points and, more importantly, will use them to re-estimate the model parameters of known classes and the estimates of known classes parameters will be deteriorated. On the other hand, novelty detection focuses on the identification of new or unknown data for which the learned classifier was not aware during the learning phase. This approach has become very popular in several application fields such as finance (fault and fraud detection), medical imaging (mass detection in mammograms) or web mining. In the last years, many methods have been proposed to deal with this problem which can be split into two main categories: statistical and neural network based approaches. An excellent review on both categories of novelty detection methods can be found in Markou and Singh (2003). However, even though all these methods are able to detect new or unobserved groups of points, no one of them is able to adapt the classifier to the new situation for classifying future observations.

The paper is organized as follows. Section 2 introduces an adaptive model-based discriminant analysis method which combines unsupervised and supervised learning for detecting unobserved classes. Section 3 presents experimental results highlighting the main features of the proposed method on simulated and real datasets. Finally, Section 4 proposes some concluding remarks and discusses further works.

## 2 Adaptive mixture discriminant analysis

### 2.1 The mixture model

Let us consider a classical parametric mixture model of  $K$  components: the observations  $\{x_1, \dots, x_n\} \in \mathbb{R}^p$  are assumed to be independent realizations of a random vector  $X$  with density:

$$f(x; \Theta) = \sum_{k=1}^K \pi_k f_k(x; \theta_k), \quad (1)$$

where  $\pi_k \geq 0$  for  $k = 1, \dots, K$  are the mixing proportions (with the constraint  $\sum_{k=1}^K \pi_k = 1$ ),  $f_k(x; \theta_k)$  is the density of the  $k$ th component of the mixture parametrized by  $\theta_k$  and finally  $\Theta = \{\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K\}$ . Among all parametric densities, the Gaussian model is probably the most used in classification. The Gaussian mixture model has been extensively studied in the last decades and used in many situations (see Banfield and Raftery (1993) for a review). Therefore, if the Gaussian model is chosen,  $f_k(x; \theta_k)$  will denote the density of a multivariate Gaussian density parametrized by  $\theta_k = \{\mu_k, \Sigma_k\}$  where  $\mu_k$  and  $\Sigma_k$  are respectively the mean and covariance matrix of  $k$ th component of the mixture. However, in some situations, modeling the data with a full covariance matrix can be too expensive in terms of number of parameters to estimate. In such a case, it is possible to use parsimonious Gaussian models (Celeux and Govaert (1995)) or Gaussian models designed for high-dimensional data (Bouveyron et al. (2007)). It is also possible to add a noise component (Banfield and Raftery (1993)) to improve the robustness of noisy dataset classification. For the mixture model (1), the log-likelihood has the



following form:

$$\mathcal{L}(x_1, \dots, x_n; \Theta) = \sum_{i=1}^n \sum_{k=1}^K s_{ik} \log(\pi_k f_k(x_i; \theta_k)),$$

where  $s_{ik} = 1$  if  $x_i$  belongs to the  $k$ th class and  $s_{ik} = 0$  otherwise. However, this work considers a specific learning situation in which only  $C$  classes are represented in the learning dataset  $\mathcal{X} = \{x_1, \dots, x_n\}$  with  $1 \leq C \leq K$ , *i.e.* one or several classes could be not represented in  $\mathcal{X}$ . Therefore, the mixture parameter estimation can not be done using the classical way and two alternative estimation procedures are proposed below.

## 2.2 Parameter estimation: transductive approach

The most intuitive way to identify unobserved classes in the test set is certainly the transductive approach which works on the gathering of learning and test sets. Indeed, since the learning sample  $\mathcal{X} = \{x_1, \dots, x_n\}$  and the test sample  $\mathcal{X}^* = \{x_1^*, \dots, x_{n^*}^*\}$  are assumed to come from the same population, both samples can be used to estimate model parameters. This would be the general framework of semi-supervised classification if  $C = K$  but semi-supervised classification methods can not be used in our context since  $K$  can be strictly larger than  $C$ . We therefore propose to adapt the classical EM algorithm (Dempster et al. (1977)) used in semi-supervised classification to the detection of unobserved classes. In the transductive learning case, the log-likelihood of model (1) has the following form:

$$\mathcal{L}(\mathcal{X}, \mathcal{X}^*; \Theta) = \sum_{i=1}^n \sum_{k=1}^C s_{ik} \log(\pi_k f_k(x_i; \theta_k)) + \sum_{i=1}^{n^*} \sum_{k=1}^K s_{ik}^* \log(\pi_k f_k(x_i^*; \theta_k)).$$

A constrained version of the EM algorithm is presented below to jointly estimate model parameters while searching for new classes. The joint estimation procedure alternates between the following E and M steps at each iteration  $q$ :

– **E step**: on the one hand, the conditional probabilities  $P(Z = k | X = x_i)$  remain fixed for the learning observations  $\{x_1, \dots, x_n\}$  and are equal to  $s_{ik}$ , for  $i = 1, \dots, n$  and  $k = 1, \dots, K$ , where  $s_{ik} = 1$  if  $x_i$  belongs to the  $k$ th class and  $s_{ik} = 0$  otherwise. On the other hand, the conditional probabilities  $t_{ik}^{*(q)} = P(Z = k | X = x_i^*)$  are updated for the test sample  $\{x_1^*, \dots, x_{n^*}^*\}$ , *i.e.* for  $i = 1, \dots, n^*$  and  $k = 1, \dots, K$ , according to the mixture parameters as follows:

$$t_{ik}^{*(q)} = \frac{\hat{\pi}_k^{(q-1)} f_k(x_i^*; \hat{\theta}_k^{(q-1)})}{f(x_i^*; \hat{\Theta}^{(q-1)})},$$

where  $\hat{\pi}_k^{(q-1)}$  and  $\hat{\theta}_k^{(q-1)}$  are the mixture parameters estimated in the M step at the step  $(q-1)$ .

– **M step**: the parameters of the  $C$  observed classes and of the  $K-C$  unobserved classes are estimated by maximizing the conditional expectation of the completed likelihood. Therefore, this step updates now the estimates of parameters  $\pi_k$  and  $\theta_k$  for  $k = 1, \dots, K$ . In the case of the Gaussian mixture, the update formulas for the

parameter estimates are, for  $k = 1, \dots, K$ :

$$\begin{aligned}\hat{\pi}_k^{(q)} &= \frac{n_k^{(q)} + n_k^{*(q)}}{n + n^*}, & \hat{\mu}_k^{(q)} &= \frac{1}{n_k^{(q)} + n_k^{*(q)}} \left( \sum_{i=1}^n s_{ik} x_i + \sum_{i=1}^{n^*} t_{ik}^* x_i^* \right), \\ \hat{\Sigma}_k^{(q)} &= \frac{1}{n_k^{(q)} + n_k^{*(q)}} \left( S_k^{(q)} + S_k^{*(q)} \right).\end{aligned}$$

where  $S_k^{(q)} = \sum_{i=1}^n s_{ik} (x_i - \hat{\mu}_k^{(q)})(x_i - \hat{\mu}_k^{(q)})^t$ ,  $S_k^{*(q)} = \sum_{i=1}^{n^*} t_{ik}^* (x_i^* - \hat{\mu}_k^{(q)})(x_i^* - \hat{\mu}_k^{(q)})^t$ ,  $n_k^{(q)} = \sum_{i=1}^n s_{ik}$  and  $n_k^{*(q)} = \sum_{i=1}^{n^*} t_{ik}^*$ .

### 2.3 Parameter estimation: inductive approach

The inductive learning context is, conversely to the previous situation, a more classical situation in supervised classification because it is more convenient to keep only model parameters to classify new observations than keeping all learning observations. In particular, the inductive approach is the only tenable approach for large dataset classification and real-time dynamic classification. However, the inductive approach poses a more complex problem since the mixture parameter estimation can not be done using the classical way. We therefore propose hereafter an inductive approach made of a learning phase and a discovery phase.

**The learning phase** In this first phase, only learning observation are considered and, since the data of the learning set are complete, *i.e.* a label  $z_i \in \{1, \dots, C\}$  is associated to each observation  $x_i$  of the learning set ( $i = 1, \dots, n$ ), we fall into the classical estimation framework of model-based discriminant analysis. In such a case, the maximization of the likelihood reduces to separately estimate the parameters of each class density by maximizing the associated conditional log-likelihood  $\mathcal{L}_k(\mathcal{X}; \Theta) = \sum_{i=1}^n s_{ik} \log(\pi_k f_k(x_i; \theta_k))$ , for  $k = 1, \dots, C$ , and this conduces to an estimation of  $\pi_k$  by  $\hat{\pi}_k = \frac{n_k}{n}$  where  $n_k = \sum_{i=1}^n s_{ik}$  is the number of observations of the  $k$ th class and to an estimation of  $\theta_k$  by  $\hat{\theta}_k$  which depends on the chosen component density. For instance, in the case of a Gaussian density, the maximization of  $\mathcal{L}_k(\mathcal{X}; \Theta)$  conduces to an estimation of  $\theta_k = \{\mu_k, \Sigma_k\}$  by  $\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^n s_{ik} x_k$  and  $\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i=1}^n s_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^t$ , for  $k = 1, \dots, C$ .

**The discovery phase** Usually, in discriminant analysis, the classification phase consists only in assigning new unlabeled observations to one of known classes. However, in this work, it is assumed that some classes could not be observed during the learning phase. It is therefore necessary to search for new classes before to classify the new observations for avoiding the misclassification of observations from an unobserved class (by assigning them to one of the observed classes). Using the model and the notations introduced above, it remains to find  $K - C$  new classes in the set of  $n^*$  new unlabeled observations  $\mathcal{X}^* = \{x_1^*, \dots, x_{n^*}^*\}$ . Since these new observations are unlabeled, we have to fit the mixture model in a partially unsupervised way. In this case, the log-likelihood has the following form:

$$\mathcal{L}(\mathcal{X}^*; \Theta) = \sum_{i=1}^{n^*} \left( \sum_{k=1}^C s_{ik}^* \log(\pi_k f_k(x_i^*; \theta_k)) + \sum_{k=C+1}^K s_{ik}^* \log(\pi_k f_k(x_i^*; \theta_k)) \right),$$

where the parameters  $\theta_k$  for  $k = 1, \dots, C$  have been estimated in the learning phase and the parameters  $\theta_k$  for  $k = C+1, \dots, K$  remain to estimate. Due to the constraint  $\sum_{k=1}^K \pi_k = 1$  on the parameters  $\pi_k$ , the mixture proportions of the  $C$  known classes have to be re-normalized according to the proportions of the  $K-C$  new classes which will be estimated on the new sample  $\{x_1^*, \dots, x_{n^*}^*\}$ . However, the test set  $\{x_1^*, \dots, x_{n^*}^*\}$  is an incomplete dataset since the labels  $z_i^*$  are missing and the  $s_{ik}^*$  are consequently unknown for all observations of this set. In such a situation, the direct maximization of the likelihood is an intractable problem and the EM algorithm can be used to estimate the mixture parameters by iteratively maximizing the likelihood. We propose below a constrained EM algorithm for estimating the parameters of the  $K-C$  unobserved classes which alternates between the following E and M steps at each iteration  $q$ :

– **E step**: the conditional probabilities  $t_{ik}^{*(q)} = P(Z = k | X = x_i^*)$ , for  $i = 1, \dots, n^*$  and  $k = 1, \dots, K$ , are updated according to the mixture parameters as follows:

$$t_{ik}^{*(q)} = \frac{\hat{\pi}_k^{(q-1)} f_k(x_i^*; \hat{\theta}_k^{(q-1)})}{f(x; \hat{\Theta}^{(q-1)})},$$

where  $\hat{\pi}_k^{(q-1)}$  and  $\hat{\theta}_k^{(q-1)}$  are the mixture parameters estimated in the M step at step  $(q-1)$ .

– **M step**: the parameters of the  $K-C$  unobserved classes are estimated by maximizing the conditional expectation of the completed likelihood whereas the estimated parameters of the observed classes remain fixed to the values obtained in the learning phase except for the proportions which are re-estimated. Therefore, this step only updates the estimates of parameters  $\pi_k$  for  $k = 1, \dots, K$  and  $\theta_k$  for  $k = C+1, \dots, K$ . In the case of the Gaussian mixture, the update formulas for the parameter estimates are:

$$\begin{cases} \text{for } k = 1, \dots, C & \hat{\pi}_k^{(q)} = \left(1 - \sum_{k=C+1}^K \frac{n_k^{*(q)}}{n^*}\right) \frac{n_k}{n}, \\ \text{for } k = C+1, \dots, K & \hat{\pi}_k^{(q)} = \frac{n_k^{*(q)}}{n^*} \end{cases}$$

where  $n_k^{*(q)} = \sum_{i=1}^{n^*} t_{ik}^{*(q)}$  and for  $k = C+1, \dots, K$ :

$$\hat{\mu}_k^{(q)} = \frac{1}{n_k^{*(q)}} \sum_{i=1}^{n^*} t_{ik}^{*(q)} x_i^*, \quad \hat{\Sigma}_k^{(q)} = \frac{1}{n_k^{*(q)}} \sum_{i=1}^{n^*} t_{ik}^{*(q)} (x_i^* - \hat{\mu}_k^{(q)})(x_i^* - \hat{\mu}_k^{(q)})^t.$$

## 2.4 Determining the number of components

Conversely to usual supervised classification, the total number  $K$  of classes is assumed to be unknown and has to be chosen. Therefore, this step is naturally a critical step in the search for unobserved classes. Classical tools for model selection in the mixture model framework are penalized likelihood criteria and include the AIC (Akaike (1974)) and BIC (Schwarz (1978)) criteria. The Bayesian Information Criterion (BIC) is certainly the most popular and consists in selecting the model which penalizes the likelihood by  $\frac{\gamma(\mathcal{M})}{2} \log(\nu)$  where  $\gamma(\mathcal{M})$  is the number of parameters in model  $\mathcal{M}$  and  $\nu$  is the number of observations. On the other hand, the AIC criterion penalizes the log-likelihood by  $\gamma(\mathcal{M})$ . The values of  $\gamma(\mathcal{M})$  and

$\nu$  are of course specific to the model proposed in this paper and depend on the chosen estimation procedure. For instance, if the classical Gaussian model is used within the transductive approach,  $\gamma(\mathcal{M})$  is equal to  $(K-1) + Kp + Kp(p+1)/2$  whereas it is equal to  $(K-1) + (K-C)p + (K-C)p(p+1)/2$  with the inductive approach. Finally, let us notice that, in practice, AIC turned out to be the most robust according to the size of  $\mathcal{X}^*$  and its use is therefore recommended.

## 2.5 Classification with the adapted classifier

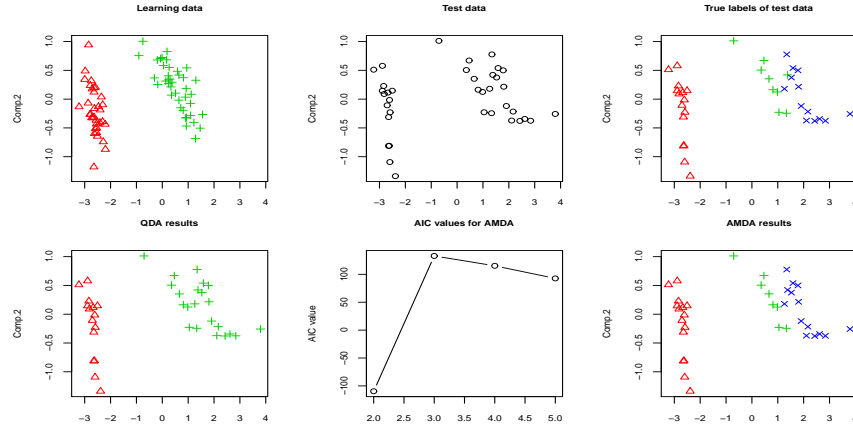
The previous paragraphs introduced a model-based discriminant analysis method which adapts its mixture model to a new situation. Therefore, the adapted model can be used to classify new observations in the future. In the classical discriminant analysis framework, new observations are usually assigned to a class using the *maximum a posteriori* (MAP) rule. The MAP rule assigns a new observation  $x \in \mathbb{R}^p$  to the class for which  $x$  has the highest posterior probability. In the case of the model described in this section, the posterior probability  $P(Z = k|X = x)$  can be expressed classically using the Bayes' rule. The posterior probabilities of the new observations depend therefore on both the classes observed in the learning phase and the classes discovered in the discovery phase.

## 3 Experimental results

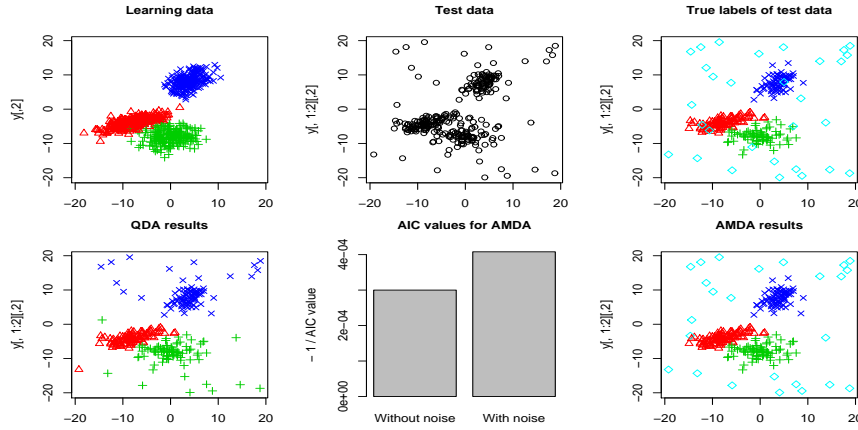
By lack of space, we present here only very short experiments but enough to illustrate the key ideas of the proposed approach. Figure 1 considers the well known Fisher's iris dataset. The scenario is the following: two classes (*setosa* and *versicolor*) have been observed during learning and one class (*virginica*) is new in the test set. Unsurprisingly, QDA classifies all observations from the unobserved class in the closest observed class. In contrast, AMDA (inductive version) successfully detects the unobserved class (3 classes are detected in the test set with the AIC criterion) and adapts the classifier to the new situation. Figure 2 presents a classification situation where the test data contains observations from a noise class which, as often in practical situations, has not been observed during the learning phase. Here as well, AMDA succeeds in detecting and classifying the observations of the noise class. The noise class has been in this experiment modeled by a uniform distribution.

## 4 Conclusion

This work has focused on the problem of learning a supervised classifier with unobserved classes. An adaptive model-based discriminant analysis method has been presented in this paper which is able to both detect unobserved groups of points in a new set of observations and to adapt the supervised classifier to the new situation. Two EM-based procedures have been proposed for parameter estimation. Experimental studies have shown that the proposed method is able to successfully detect different kinds of unobserved classes (Gaussian, uniform noise, ...). It remains however to deal with the problem of label switching when  $C - K > 1$ . A way to solve this problem could be to ask domain experts to classify some observations of the



**Fig. 1.** Classification with AMDA of the Iris dataset: the classes *setosa* (red triangles) and *versicolor* (green plus-es) have been observed during the learning phase whereas the class *virginica* (blue crosses) has not.



**Fig. 2.** Classification with AMDA of simulated data: 3 observed classes and 1 unobserved noise class (light blue diamonds) in  $\mathbb{R}^2$ .

new detected groups in order to associate a class name with the detected groups. Finally, it could be very interesting to study the evolution of the proposed strategy in the context of dynamic classification.

## References

- AKAIKE, H. (1974): A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.
- BANFIELD, J. and RAFTERY, A. (1993): Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821.

- BOUYEYRON, C., GIRARD, S., and SCHMID, C. (2007): High Dimensional Discriminant Analysis. *Comm. in Statistics: Theory and Methods*, 36(14):2607-2623.
- CELEUX, G. and GOVAERT, G. (1995): Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781-793.
- DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society*, 39(1):1-38.
- KRISHNAPURAM, B., WILLIAMS, D., XUE, Y., HARTEMINK, A., CARIN, L. and FIGUEIREDO, M.(2004): On semi-supervised classification. *In NIPS*.
- MARKOU, M. and SINGH, S. (2003): Novelty detection: A review - part 1: Statistical approaches. *Signal Processing*, 83(12):2481-2497.
- MARKOU, M. and SINGH, S. (2003): Novelty detection: A review - part 2: Neural network based approaches. *Signal Processing*, 83(12):2499-2521.
- SCHWARZ, G. (1978): Estimating the dimension of a model. *The Annals of Statistics*,6:461-464.

# Forecasting a Compound Cox Process by means of PCP

Paula R. Bouzas<sup>1</sup>, Nuria Ruiz-Fuentes<sup>2</sup>, and Juan Eloy Ruiz-Castro<sup>3</sup>

<sup>1</sup> Dept. Statistics and Operations Research, University of Granada  
Faculty of Pharmacy, Granada, Spain, *paula@ugr.es*

<sup>2</sup> Dept. Statistics and Operations Research, University of Jaén  
Campus Las Lagunillas, Jaén, Spain, *nfuentes@ujaen.es*

<sup>3</sup> Dept. Statistics and Operations Research, University of Granada  
Faculty of Sciences, Granada, Spain, *jeloy@ugr.es*

**Abstract.** A compound Cox process (CCP) is a generalization of a Cox process in which the events have an associated mark. The counting statistics of a CCP with marks in a specific subset are presented and the expression of the mode is derived. The representation theorems extended to the CCP, an ad hoc FPCA estimation of the mean process and principal components prediction are the basis to forecast the mean and mode of the CCP in the future. Several simulations illustrate the forecasting process.

**Keywords:** Cox process, functional principal components analysis, prediction with principal components

## 1 Introduction

The compound doubly stochastic Poisson process or compound Cox process (CCP) is the natural extension of the compound Poisson process where the intensity is a stochastic process influenced by the so-called information process. Therefore, a CCP is a marked point process in which the point process is a Cox process and the marks of the events are independent and identically distributed and also independent of the Cox process. Even this is quite a general and flexible model, there are not many real processes modelled by it. The main reason can be the difficulty of estimating the parameters of the CCP (mean or intensity) that characterize it.

Bouzas et al. (2007) proposed a methodology for estimating the mean process and some counting statistics of a CCP based on the representation theorems extended to CCP. In this paper, we try to go further in the study of a CCP, not only providing the expressions for counting statistics but also firstly deriving the expression of the mode and secondly estimating several statistics in a future instant of time by means of prediction with principal components .

It is remarkable that a CCP generalizes some particular type of processes. A CCP is a Cox process if the mark space is denumerable with equal marks or a Cox process with random deletions if the mark space is  $\{0, 1\}$  where the deleted points are marked with 0. Relaxing the property of orderliness, a CCP can represent a Cox process with simultaneous occurrences when the mark space is  $\mathbb{N}$  so that each mark indicates the number of occurrences. A multichannel Cox process can be considered

a CCP in which the mark indicates the region where the point occurs. A time-space Cox process can also be modeled by a CCP in time where the mark on a point is continuously distributed and indicates the spatial position of the point.

The paper is structured in the following way. Section 2 shows the counting statistics of a CCP having specific marks (see Bouzas et al. (2007)) and presents as a first novelty the study of the mode of the process. Section 3 provides a summary of functional principal components analysis (FPCA) and prediction with principal components. Then, Section 4 presents the second novelty of the paper, the application of prediction with principal components to the mean process of a CCP having specific marks in order to forecast its counting statistics. Finally, Section 5 illustrates the forecasting method using several simulations of a CCP.

## 2 Statistics of a CCP having specific marks

A CCP is defined as a Cox process  $\{N(t); t \geq t_0\}$  with intensity  $\{\lambda(t, x(t)); t \geq t_0\}$  where  $\{x(t); t \geq t_0\}$  is the information process and with marks associated to the arrival times. The marks are i.i.d. variables and independent to the Cox process. The  $n$ -th arrival time will be denoted by  $w_n$  and its mark by  $u_n$  which is a realization of the random variable  $U$  in  $\mathcal{U}$ .

Let us define  $\{N(t, B); t \geq t_0\}$  as the process that counts the points having mark in the subset  $B \subset \mathcal{U}$ . In this section, using the representation theorems of a CCP from Bouzas et al. (2007), the expression for the basic counting statistics of  $N(t, B)$  taking into account that the mark space is nondenumerable (their expressions are analogous in the denumerable case) are provided in terms of the mean process.

### 2.1 Counting statistics

Having a CCP with intensity  $\{\lambda(t, x(t)); t \geq t_0\}$ , it is well known that if the mean of the process,  $\{\Lambda(t, x(t)); t \geq t_0\}$ , is absolutely continuous then,  $\Lambda(t, x(t)) = \int_{t_0}^t \lambda(\sigma, x(\sigma)) d\sigma$ . Due to that, the explicit expressions of the counting statistics of  $\{N(t, B); t \geq t_0\}$  can be expressed in terms of the mean process in the following way.

The **probability mass function** is

$$P[N(t, B) = n] = E_x \left[ \frac{1}{n!} \left( \Lambda(t, x(t)) \int_B P_u(dU) \right)^n \exp \left[ -\Lambda(t, x(t)) \int_B P_u(dU) \right] \right] \quad (1)$$

where  $P_u$  is the probability distribution of  $U$ .

As well, we can write the **mean** of the CCP as

$$E[N(t, B)] = E_x \left[ \Lambda(t, x(t)) \int_B P_u(dU) \right] = E_x [\Lambda(t, x(t))] \int_B P_u(dU)$$

We introduce in this paper the study of the **mode** of the CCP. The mode,  $n_{\max} \in \mathbb{N}$ , is the number of occurrences which makes the probability mass function to get its maximum value. It is the most probable number of occurrences until a certain time point. The mode is given in the following proposition.



**Proposition 11.** *The number of occurrences of the Cox process with marks in  $B$  with maximum probability ( $\max P[N(t, B) = n]$ ) will be denoted by  $n_{\max} \in \mathbb{N}$  and it is  $\lceil \Lambda(t, x(t)) \int_B P_u(dU) \rceil - 1$  if this expression is in  $\mathbb{N}$ ; it is  $\text{int} \left[ \left( \Lambda(t, x(t)) \int_B P_u(dU) \right) - 1 \right]$  or  $\text{int} \left[ \left( \Lambda(t, x(t)) \int_B P_u(dU) \right) - 1 \right] + 1$  if it is not in  $\mathbb{N}$  and  $n_{\max}$  is 0 if the expression is negative. Note:  $\text{int} = \text{integer part}$ .*

*Proof.* The probability  $P[N(t, B) = n]$  is given in equation (1). Having known the value of  $\Lambda(t, x(t))$  given the information process, let us calculate

$$\max P[N(t, B) = n] = \max \left[ \frac{1}{n!} \left( \Lambda(t, x(t)) \int_B P_u(dU) \right)^n \exp \left[ -\Lambda(t, x(t)) \int_B P_u(dU) \right] \right]$$

and as  $\exp \left[ -\Lambda(t, x(t)) \int_B P_u(dU) \right] = \text{constant}$ , we just have to maximize  $\frac{1}{n!} \left( \Lambda(t, x(t)) \int_B P_u(dU) \right)^n$ . Let us call  $\Lambda(t, x(t)) \int_B P_u(dU) = A$  in order not to complicate the following reasoning.

- i) It is known that for a fixed  $A \in \mathbb{R}$ ,  $\frac{A^n}{n!} \xrightarrow{n \rightarrow \infty} 0$  so there exists  $n_0 \in \mathbb{N}$  from which the series decreases in  $n$ . Then,

$$\frac{A^n}{n!} \geq \frac{A^{n+1}}{(n+1)!} \text{ if and only if } n \geq A - 1$$

- ii) Also,

$$\frac{A^n}{n!} \leq \frac{A^{n+1}}{(n+1)!} \text{ if and only if } n \leq A - 1$$

Therefore, if  $A \in \mathbb{N}$ ,  $n_{\max} = A - 1$  and if  $A \notin \mathbb{N}$ ,  $n_{\max} = \text{int}(A - 1)$  or  $n_{\max} = \text{int}(A - 1) + 1$ .

We have to take into account the possibility of that the mentioned  $n_{\max}$  becomes negative; it could happen if  $A < 1$ . Then,  $n_{\max}$  has to be taken as 0.

Taking into account that  $A = \Lambda(t, x(t)) \int_B P_u(dU)$ , the proof is complete.

### 3 Functional analysis for CCP

In this section, we will briefly summarize the technique of functional principal components analysis (FPCA) as well as the prediction with principal components developed by Aguilera et al. (1997).

#### 3.1 Functional principal components analysis for CCP

Let us remember that by analogy with the multivariate case, the functional principal components of a second order and quadratic mean stochastic process with sample paths in the space  $L^2[t_0, t_p]$  of square integrable functions are defined as uncorrelated generalized linear combinations of the process variables whose weight functions (principal factors) are obtained as the eigenfunctions of the sample covariance kernel, see Ramsay and Silverman (1997). In order to obtain the principal components of a stochastic process with sample paths in a finite dimension space

generated by a set of linearly independent functions, Ocaña et al. (1999) proved a theorem that shows the equivalence between FPCA with respect to the usual inner product in  $L^2[t_0, t_p]$  and standard multivariate principal components analysis in  $\mathbb{R}^{2(p+1)}$ .

An ad hoc FPCA for the mean process of a Cox process has been developed in Bouzas et al. (2006), therefore in this section we work in terms of the mean. This FPCA method can be applied to CCP in order to estimate some of its statistics. It is interesting to estimate the number of points with the mark in a given subset of the mark space.

Applying FPCA to the mean process of the CCP regardless of its marks, it is obtained the estimated truncated orthogonal expansion in terms of the principal components in the observed time interval  $[t_0 = T_0, T_1]$

$$\Lambda^q(t) = \mu_\Lambda(t) + \sum_{j=1}^q \xi_j f_j(t); \quad t \in [t_0 = T_0, T_1]$$

In case that the observed data are the data of just one long sample path, we can split it up in several trajectories. Rescaling the time, we can presume they begin in the same time point with  $N(t_0, B) = 0$  due to the independence of increments of a Poisson process, see Bouzas et al. (2006). Therefore, the sample paths can be for example observed data on several years (or another time interval) of the process. Then, if we want to give the result in the real time point we will have to add the number of points before the initial time point of the new sample path.

### 3.2 Prediction with principal components for CCP

Prediction by means of principal components of a stochastic process gives a continuous prediction of the process in a future time interval from discrete observations of the process in the past which was introduced by Aguilera et al. (1997).

Having known the evolution of an stochastic process  $\{X(t); t \in [T_0, T_1]\}$ , the prediction with principal components models estimate it in a future interval  $\{X(t); t \in [T_1, T_2]\}$  using FPCA. The process must be of second order, continuous in quadratic mean and squared integrable sample paths in their corresponding intervals. In case that the stochastic evolution of the process is not known but several sample paths of it are available, the prediction with principal components model has to be estimated (see also Aguilera et al. (1999) and Valderrama et al. (2000) for a deeper study).

Firstly, FPCA of the process in both intervals, past and future, is estimated and selected the number of principal components to be considered (those that accumulate at least a fixed percentage of explained variance) in order to truncate the process Karhunen-Loève expansion. Secondly, the principal components of the past that predict the principal components of the future are selected by means of having significantly high correlation. Finally, any other sample path of the process observed in the past predicts its evolution in the future interval using the FPCA in the future with the principal components predicted by the past ones.

In this paper, we will consider the prediction of the mean process of a CCP in a future interval  $(T_1, T_2]$  from its sample paths observed or estimated in a past interval  $[T_0, T_1]$ .

Applying FPCA to the mean process in  $[t_0 = T_0, T_1]$  and in  $(T_1, T_2]$ , it is obtained the estimated truncated orthogonal principal components decomposition in both time intervals. Let us denote them in the following way

$$\begin{aligned}\widehat{A^1(s)} &= A^{q_1}(t) = \mu_A^1(t) + \sum_{j=1}^{q_1} \xi_j f_j(t); & t \in [t_0 = T_0, T_1] \\ \widehat{A^2(s)} &= A^{q_2}(s) = \mu_A^2(s) + \sum_{j=1}^{q_2} \eta_j g_j(s); & s \in (T_1, T_2)\end{aligned}\quad (2)$$

Let us denote by  $\tilde{\eta}_j^{p_j} = \sum_{i=1}^{p_j} b_i^j \xi_i$  the estimator of  $\eta_j$ ,  $j = 1, \dots, q_2$  in terms of the  $p_j$  principal components  $\xi_j$ . Therefore, we can rewrite (2) so that

$$\tilde{A}^{q_2}(s) = \mu_A^2(s) + \sum_{j=1}^{q_2} \left( \sum_{i=1}^{p_j} b_i^j \xi_i \right) g_j(s); \quad s \in (T_1, T_2) \quad (3)$$

This is the estimated stochastic structure of the mean process in a future time interval. The prediction with principal components model will be denoted as PCP( $q_2; p_1, \dots, p_j$ ). The selected PCP model contains those pairs of future-past principal components with significant linear correlation, which are included in order of magnitude of the proportion of future variance explained by a PCP model only including the pair, until the relative proportion of future variance explained by the model is as close to one as possible. For each observed sample path of the mean process in  $[T_0, T_1]$ , its evolution in  $(T_1, T_2]$  can be predicted estimating the principal components of the mean sample path and using equation (3). An analogous procedure can be developed for the intensity process using the estimation methodology by the adapted FPCA proposed in Bouzas et al. (2010).

## 4 Forecasting the statistics

It can be interesting to introduce the prediction derived in equation (3) as well as to use the independence between the marks and the occurrences of a CCP in order to predict several of its statistics in a future time point  $s \in (T_1, T_2)$  having observed the process in  $[T_0, T_1]$ . Even it is not possible to take advantage of the mean process estimation in all of them, it can be applied in the most practical ones.

As mentioned before, if a sample path has been observed till time  $T_1$  and a forecasting in  $s$  is needed, due to the independence of increments of the CCP, it is possible to estimate  $N(s, B)$  in  $(T_1, T_2)$  and add the estimated number of occurrences to the observed ones.

**Mean of  $N(s, B)$ ,  $s$  in  $(T_1, T_2)$**

$$E[N(s, B)] = E_x \left[ A^2(s, x(s)) \int_B P_u(dU) \right] = \mu_A^2(s) \int_B P_u(dU)$$

**Mode of  $N(s, B)$ ,  $s$  in  $(T_1, T_2)$**  (number of occurrences with maximum probability until time  $s$ ). Then, having  $s \in (T_1, T_2)$ , taking into account eq. (1) and Proposition 11, we have that

$$P[N(s, B) = n] = \frac{1}{n!} \left( \tilde{A}^{q_2}(s) \int_B P_u(dU) \right)^n \exp \left[ -\tilde{A}^{q_2}(s) \int_B P_u(dU) \right]$$

gets its maximum in

$$\begin{cases} n_{\max} = \tilde{\Lambda}^{q_2}(s) \int_B P_u(dU) - 1, & \tilde{\Lambda}^{q_2}(s) \int_B P_u(dU) \in \mathbb{N} \\ n_{\max} = \begin{cases} \text{int} \left( \tilde{\Lambda}^{q_2}(s) \int_B P_u(dU) - 1 \right) \\ \text{or} \\ \text{int} \left( \tilde{\Lambda}^{q_2}(s) \int_B P_u(dU) - 1 \right) + 1 \end{cases}, & \tilde{\Lambda}^{q_2}(s) \int_B P_u(dU) \notin \mathbb{N} \\ n_{\max} = 0, & \tilde{\Lambda}^{q_2}(s) \int_B P_u(dU) < 1 \end{cases}$$

## 5 Simulation

This section illustrates the theoretical calculations presented in the former sections by means of simulations. Several have been made with different types of intensity processes and mark distributions and the method of forecasting always gives proper results. Four examples are presented below.

In each example, 100 sample paths have been simulated in  $[T_0, T_2) = [0, 10]$  plus a new sample path also simulated in the whole interval  $[T_0, T_2) = [0, 10]$  in order to preserve its future part to be able to validate the results. Therefore, we have chosen the future instant of time  $s$  as one of the knots of the future time interval so it could be possible to calculate the sample values of the mean and the mode in that instant of time and then, compare them with the estimated ones. It is remarkable that the continuous reconstruction of the mean process allows us to estimate the mean and the mode in any  $s \in (T_1, T_2)$ .

In the first simulation, the intensity is distributed as a Gamma (5,0.4), the marks are distributed as Binomial(10,0.4) and  $B = \{u; 4 \leq u \leq 6\}$  so  $\int_B P_u(dU) = p = 0.5630$ . Table 1 shows the technical data of two examples with different future time instants and that the estimated mean ( $E[N(s, B)] = E_s$ ) and mode ( $n_{\max}$ ) are very similar to the sample ones ( $\bar{E}[N(s, B)] = \bar{E}_s$  and  $\bar{n}_{\max}$ , respectively).

Marks $B(10, 0.4)$	and	$B = \{u; 4 \leq u \leq 6\}$
$T_1 = 7, \quad s = 8$		$T_1 = 5, \quad s = 7$
PCP (2; 3, 1)		PCP (4; 2, 1, 2, 1)
$E_s = 9.02 \quad \bar{E}_s = 9.8$	$E_s = 7.88$	$\bar{E}_s = 8.87$
$n_{\max} = 9 \quad \bar{n}_{\max} = 8$	$n_{\max} = 7$	$\bar{n}_{\max} = 8$

**Table 1.** Mean and mode of two examples of the first simulation type with different future time points.

In the second simulation, the CCP is formed by four simultaneous Poisson processes with uniformly distributed intensity randomly occurring being controlled by a boolean vector, the marks are distributed as lognormal(1,0.5) and  $B = \{u; 2 \leq u \leq 5\}$  so  $\int_B P_u(dU) = p = 0.6188$ . Regardless of the marks, this type of CCP is also known as a mixed Cox process. Table 2, analogous to Table 1, gives data for this second type of simulation.

Marks $\lg n(1, 0.5)$ and		$B = \{u; 2 \leq u \leq 5\}$	
$T_1 = 4, \quad s = 5$		$T_1 = 5, \quad s = 9$	
PCP (3; 2, 2, 1)		PCP (2; 3, 2)	
$E_s = 3.11$	$\bar{E}_s = 3.93$	$E_s = 5.64$	$\bar{E}_s = 6.33$
$n_{\max} = 3$	$\bar{n}_{\max} = 1, 3, 4$	$n_{\max} = 5$	$\bar{n}_{\max} = 4, 7$

**Table 2.** Mean and mode of two examples of the second simulation type with different future time points.

## 6 Acknowledgments

This work was partially supported by projects MTM2007-63793 and MTM 2007-66791 of Plan Nacional I+D+I, Ministerio de Ciencia e Innovación, P06-FQM-01470 from Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía and grant FQM-307 of Conserjería de Innovación de la Junta de Andalucía, all of them in Spain.

## References

- AGUILERA, A.M., OCAÑA, F.A. and VALDERRAMA, M.J. (1997): An approximated principal component prediction model for continuous-time stochastic processes. *Applied Stochastic Models and Data Analysis* 13, 61-72.
- AGUILERA, A.M., OCAÑA, F.A. and VALDERRAMA, M.J. (1999): Forecasting Time Series by Functional PCA. Discussion of Several Weighted Approaches. *Computational Statistics*, 14, 443-467.
- BOUZAS, P.R., VALDERRAMA, M.J., AGUILERA, A.M. and RUIZ-FUENTES, N. (2006): Modelling the mean of a doubly stochastic Poisson process by functional data analysis. *Computational Statistics and Data Analysis* 50 (10), 2655-2667.
- BOUZAS, P.R., RUIZ-FUENTES, N. and OCAÑA, F.M. (2007): Functional approach to the random mean of a compound Cox process. *Computational Statistics* 22, 467-479.
- BOUZAS, P.R., AGUILERA, M.J. and RUIZ-FUENTES, N. (2010): Functional estimation of the intensity of a Cox process. *Methodology and Computing in Applied Probability*. Accepted for publication.
- BREMAUD, P. (1981): *Point processes and queues: Martingale dynamics*. Springer-Verlag, N.Y.
- OCAÑA, F.A., AGUILERA, A.M. and VALDERRAMA, M.J. (1999): Functional principal components analysis by choice of norm, *J.Multiv.Analysis*, 71, 262-276.
- RAMSAY, J.O. and SILVERMAN, B.M. (1997): *Functional Data Analysis*. Springer-Verlag, N.J.
- VALDERRAMA, J.M., AGUILERA, A.M. and OCAÑA, F.A. (2000): *Predicción dinámica mediante análisis de datos funcionales*. La Muralla, Madrid.



# Cutting the Dendrogram through Permutation Tests

Dario Bruzzese<sup>1</sup> and Domenico Vistocco<sup>2</sup>

<sup>1</sup> Dipartimento di Medicina Preventiva, Università di Napoli - Federico II  
Via S. Pansini 5, Napoli, Italy, [dario.bruzzese@unina.it](mailto:dario.bruzzese@unina.it)

<sup>2</sup> Dipartimento di Scienze Economiche, Università di Cassino  
Via S. Angelo S.N., Cassino, Italy, [vistocco@unicas.it](mailto:vistocco@unicas.it)

**Abstract.** This paper introduces an innovative approach for detecting a sub-optimal partition starting from the dendrogram produced by a hierarchical clustering technique. The approach exploits permutation tests and it can be used regardless of the agglomeration method and distance measure used in the classification process because it relies on the same criteria used for producing it. Moreover, the proposed approach can detect partitions not necessarily identifiable using a traditional cut approach, as the resulting clusters could correspond to different heights of the tree.

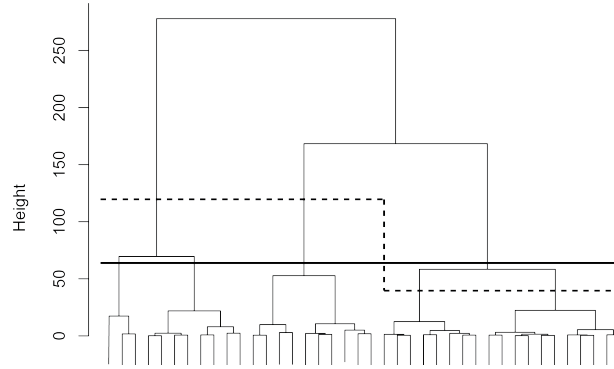
**Keywords:** hierarchical clustering, permutation tests, cluster detection

## 1 Introduction and motivation

Hierarchical clustering represents one of the most widespread analytical approach to face with classification problems mainly due to the visual power of the associated graphical representation, the dendrogram, and because of the directness of the cluster generation process. All these aside, the requirement of choosing properly the “optimum” number of clusters still represents the main difficulty for the final user. Actually, different (semi)automatic criteria can be devised to reach the final classification; very often, the very informal solution adopted is that of finding the height in the dendrogram where large changes in fusion level occur.

Broadly speaking, all these criteria determine a threshold value of the ultrametric used to grow the dendrogram such that all the units with a dissimilarity index below this threshold will belong to the same cluster; however such approach allows to search for the solution only within a small set of the whole family of partitions housed in the dendrogram: those which stem from *horizontal* cuts of the dendrogram. This induces a biunivocal relation between the number  $k$  of clusters and the partition set such that by fixing one element of the relation (e.g. the number of clusters) the other is univocally determined.

It could happen, however, that clusters differ in terms of their own internal coherence in a way that the same threshold value wouldn’t be suitable for all of them. Figure 1 shows the dendrogram obtained on a simulated dataset. The dataset contains 4 different clusters of the same cardinality generated from multivariate normal distributions with different mean vectors and variance-covariance matrices. The Ward criterion with the Euclidean distance was used to grow the tree.



**Fig. 1.** Two different partitions in 4 clusters of a simulated dataset. Solid line refers to a traditional horizontal criterion while the dashed line refers to a possible solution offered by the proposed algorithm.

The solid line in Figure 1 highlights the 4-clusters solution obtained by cutting the dendrogram with a traditional horizontal criterion; this partition, which actually is the only one that can produce a 4-clusters solution, isolates a very small cluster on the left side of the dendrogram while leaving ungrouped the two clusters on the right that, on the contrary, contain units belonging to different populations. A different solution, inside those that still comply with the hierarchical classification process, could thus be the one described by the dashed line and characterized by two local thresholds located at different heights; actually it turns out that this non-conventional cut can better recover the original cluster structure (according to the misclassification index the first partition produces an error rate of 0.58 while the second is characterized by an error rate of 0.40).

The possibility of merging clusters at different heights (thus conflicting with the *classificability* principle previously described) makes mandatory the implementation of a procedure able to automatically explore the complete set of partitions by tracing the partial thresholds whenever two clusters plainly reflect specific characteristics.

The proposed algorithm exploits the theoretical framework of permutation tests in order to reach this goal. The most important by-product of such approach is the automatic identification of the number of clusters.

The paper is organized as follows: the idea used for detecting the partition is introduced in Section 2, the notation and the proposed procedure is detailed in Section 3; Section 4 shows some results on a genetic dataset and a simulation study in order to explore the influence of tuning parameters on the algorithm output. Finally, some concluding remarks and future work directions follow in the final section.

## 2 The basic intuition

The proposed algorithm exploits a permutation test approach to automatically detect a partition starting from a dendrogram resulting from a hierarchical cluster. The algorithm retraces down-ward the tree, starting from the root of the dendrogram where all objects are classified in a unique cluster and moving down a *partial*



*threshold* until a link joining two clusters is encountered. A permutation test is thus performed in order to verify whether the two clusters must be accounted as a unique group (the null hypothesis) or not (the alternative one). If the null cannot be rejected, the corresponding branch will become a cluster of the final partition and none of its sub-branches will be longer processed. Otherwise each of them will be further visited in the course of the procedure. In fact, in both cases, the partial threshold will continue its path and the next branch of the dendrogram will be processed. The algorithm stops when there are no more branches that stand the test (i.e. the null cannot be rejected any more).

The permutation test on which the whole procedure is based can be summarized in this way. Under the *Null*, if all the units belonging to each of the two clusters are mixed up together and then randomly split up, with the only constraint of the group cardinality, the distance among the shuffled clusters should not be very different from the original one. Repeating the shuffling  $m$  times, a Montecarlo  $p$ -value can be computed as the number of permuted distances at least as extreme as the original one.

The whole algorithm is detailed in the next section.

### 3 The algorithm

Let denote with  $n$  the number of objects to classify, with  $C_L^k$  and  $C_R^k$  the two classes merged at level  $k$  ( $k = 1, \dots, n$ ), with  $h(C_L^k \cup C_R^k)$  the height necessary to merge  $C_L^k$  and  $C_R^k$ . Finally we denote with  $h(C_j^k)$  the height at which  $C_j^k$  has been obtained ( $j \in \{L, R\}$ ). In Figure 2 the adopted formalism is superimposed, for  $k \in \{1, 2\}$ , on the dendrogram shown in the previous section.

For each  $k$ , the difference between  $\max_{j \in \{L, R\}} h(C_j^k)$  and  $\min_{j \in \{L, R\}} h(C_j^k)$  can be considered as the *minimum cost* necessary to merge the two classes. Minimum because, at least, the dissimilarity measure used in the agglomeration process, must raise from  $\min_{j \in \{L, R\}} h(C_j^k)$  to  $\max_{j \in \{L, R\}} h(C_j^k)$  in order to merge the two clusters.

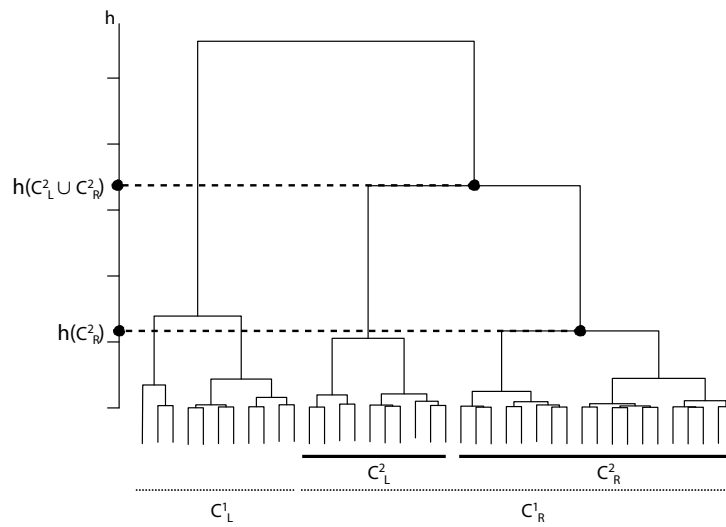
The difference between  $h(C_L^k \cup C_R^k)$  and  $\max_{j \in \{L, R\}} h(C_j^k)$  can be, instead, considered as the *cost* actually incurred for merging  $C_L^k$  and  $C_R^k$ . The ratio between these two costs:

$$\text{cost}(C_L^k \cup C_R^k) = \frac{\max_{j \in \{L, R\}} h(C_j^k) - \min_{j \in \{L, R\}} h(C_j^k)}{h(C_L^k \cup C_R^k) - \max_{j \in \{L, R\}} h(C_j^k)}$$

is thus a measure that characterizes the aggregation process resulting in the new class  $C_L^k \cup C_R^k$  and is indeed used in the permutation test approach for automatically detecting the clusters.

The proposed procedure is detailed in Algorithm 1. In particular we denote with *aggregationLevelsToVisit* a vector containing the heights of the dendrogram to be explored and with *permClusters* an object storing the clusters detected by the procedure.

The permutation test step is embedded in the row 6 of the Algorithm 1. In particular, for each  $k$  a permutation test is designed to test the Null Hypothesis



**Fig. 2.** Exemplification of the main notation adopted with respect to the dendrogram reported in Figure 1.

that the two groups  $C_L^k$  and  $C_R^k$  really belong to the same cluster, i.e.

$$H_0 : C_L^k \equiv C_R^k.$$

Under  $H_0$ , mixing up (i.e. permuting) the statistical units of  $C_L^k$  and  $C_R^k$  should not alter the aggregation process resulting in their merging in.

**Input:** A dataset and its related dendrogram

**Output:** A partition of the dataset

- a. **inicialization:**
- b.  $\text{aggregationLevelsToVisit} \leftarrow h(C_L^1 \cup C_R^1)$
- c.  $\text{permClusters} \leftarrow []$
- d.  $i \leftarrow 1$
- e. **repeat**
- f.     **if**  $C_L^i \equiv C_R^i$
- g.         add  $C_L^i \cup C_R^i$  to  $\text{permClusters}$
- h.     **else**
- i.         add  $h(C_L^i)$  and  $h(C_R^i)$  to  $\text{aggregationLevelsToVisit}$
- j.         sort  $\text{aggregationLevelsToVisit}$  in descending order
- k.     **end**
- l.     remove the first element from  $\text{aggregationLevelsToVisit}$
- m.      $i \leftarrow i+1$
- n. **until**  $\text{aggregationLevelsToVisit}$  is empty

**Algorithm 1:** The proposed *PermClust* algorithm

Let  ${}_mC_L^k$  and  ${}_mC_R^k$  be the two new classes obtained by permuting the elements in  $C_L^k$  and  $C_R^k$ . As a matter of fact, the hierarchical clustering process is invariant with respect to the permutation of the original observations and thus growing a single dendrogram on the permuted set would simply re-establish the same structure. For this reason, after  ${}_mC_L^k$  and  ${}_mC_R^k$  have been obtained, a new dendrogram is generated for each of them. The heights at which each of the two classes are built up again, clearly correspond to the heights of the root nodes of the corresponding dendrograms. The ratio:

$$\text{cost}({}_mC_L^k \cup {}_mC_R^k) = \frac{\max_{j \in \{L,R\}} h({}_mC_j^k) - \min_{j \in \{L,R\}} h({}_mC_j^k)}{h(C_L^k \cup C_R^k) - \max_{j \in \{L,R\}} h({}_mC_j^k)}$$

is thus a measure that characterizes the aggregation process resulting in the new (*potential*) class  ${}_mC_L^k \cup {}_mC_R^k$ . Under  $H_0$ , the aggregation process resulting in the new cluster  $C_L^k \cup C_R^k$  should be very similar to the one that *potentially* would have produced  ${}_mC_L^k \cup {}_mC_R^k$ ; thus the two values  $\text{cost}(C_L^k \cup C_R^k)$  and  $\text{cost}({}_mC_L^k \cup {}_mC_R^k)$  should be close enough.

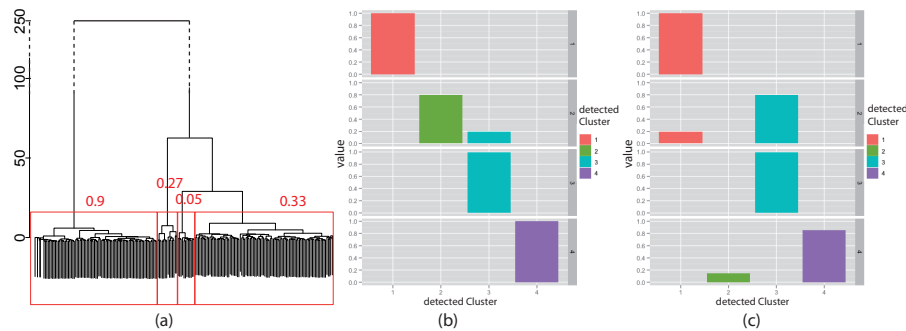
The permutation procedure is repeated  $M$  times and each time a new couple  ${}_mC_L^k, {}_mC_R^k$  is obtained. The pvalue Montecarlo (Good, 1994) is thus computed as:

$$p = \frac{\#\{\text{cost}({}_mC_L^k \cup {}_mC_R^k) \leq \text{cost}(C_L^k \cup C_R^k)\} + 1}{M + 1}$$

## 4 Some results

The PermClust algorithm has been applied both on real and synthetic datasets; in the following the main results will be presented. In all the computations, the dendrograms have been generated with the Euclidean distance and the Ward agglomeration criterion (Maechler et al., 2005). Unless differently specified, p-values

less than 0.01 were considered statistically significant in the permutation test step. Figure 3(a) shows (a zoom of) the dendrogram obtained on the Yeast galactose dataset which describes a subset of 205 genes reflecting four functional categories of the Gene Ontology (Ideker et al., 2001)<sup>1</sup>. The obtained partition is highlighted



**Fig. 3.** (a) The dendrogram obtained on the Yeast Galactose dataset with the partition selected by PermClust algorithm. Numbers refer to the p-values of the associated permutation test. (b) Visual representation of the confusion matrix resulting from PermClust algorithm and (c) from a *k-means* with  $k=4$ .

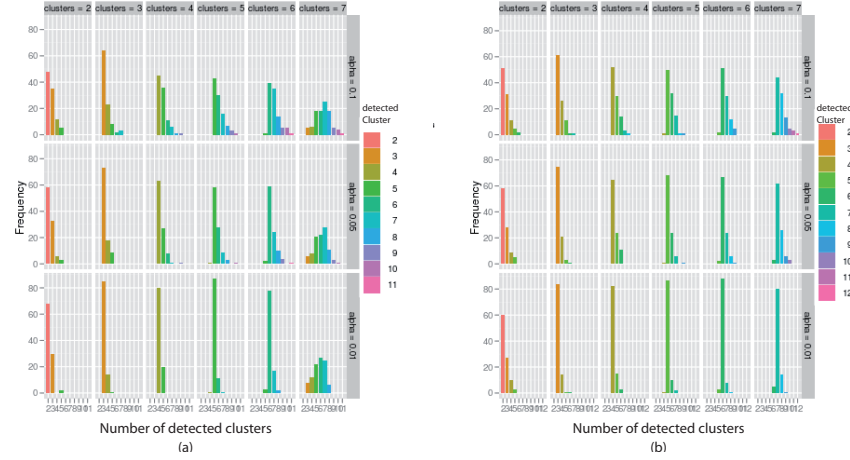
using red rectangles and clearly reveals the 4-clusters structure originally contained in the dataset. Panels (b) and (c) show the confusion matrices related to the proposed algorithm (b) and to a *k-means* procedure (c) with  $k$  equal to 4. The different sub panels depict the original clusters in the dataset while the different bars refer to the clusters detected by the classification procedures. It can be noticed that the proposed algorithm correctly assigns the units to the first and the fourth class while a small fraction of units belonging to the second cluster is misclassified into the third cluster. K-means, on the contrary, is unable to grasp second cluster (whose units are misclassified in the third cluster). Small misclassification rates characterize also the first and the fourth cluster. The misclassification rate was 1.5% for PermClust and 8.3% for the k-means procedure. It is worth of notice that the partition selected by the proposed algorithm agrees with the *hortodox* 4-clusters solution but it has been automatically detected by the algorithm.

The PermClust algorithm has been also tested on artificial datasets. In particular, Figure 4 shows the results of the algorithm on artificial datasets generated according to the random cluster generation method proposed in Qiu and Joe (2006a, 2009). Generated data differ in terms of the number of clusters ( $k = 2, 3, 4, 5, 6, 7$ ) and of the number of variables ( $p = 5, 10$ )<sup>2</sup>. The artificial data have been generated using a value of 0.01 for the separation index (Qiu and Joe, 2006b) between

<sup>1</sup> For this application, the algorithm, written in the R language, uses almost 50 secs. on a Intel Core 2 Duo 2.26 GHz machine with 4 GB of RAM. More efficiency could be achieved optimizing the code and implementing it using a compiled language.

<sup>2</sup> With  $p=15$  the performance of the algorithm is almost equal to  $p=10$ . The corresponding figure is not reported for sake of brevity.

any cluster and its nearest neighbor cluster which reflects a close cluster structure. For each combination of  $k$  and  $p$ ,  $s = 100$  different datasets have been generated.



**Fig. 4.** Distribution of the number of clusters detected by the PermClust algorithm for artificial datasets in case of 5 variables (a) and 10 variables (b).

Figure 4(a) shows the number of clusters composing the partition detected by the PermClust algorithm using  $p = 5$  variables. Different columns of the Figure depict the different value of  $k$ , while the rows refer to the significance level used in the permutation test step of the algorithm (see row 6 of Algorithm 1). The barplots in each panel show the distribution of the numbers of clusters detected by the algorithm in the  $s$  simulations. The same structure is used for the case of a dataset with 10 variables (Figure 4(b)).

As can be noticed, the stability of the algorithm strictly depends on the combination among the significance level, the cardinality of the cluster structure and the number of variables. In particular, while a significance level of 0.01 (last row of Figure 4(a) and (b)) always allows to achieve the best results, the accuracy of the solution is inversely proportional to the ratio between  $k$  and  $p$ . In case of a simple cluster structure ( $k=2,3$ ), the algorithm seems to fail even with a large number of available variables.

## 5 Concluding remarks and further developments

The output of hierarchical clustering methods is typically displayed as a dendrogram describing a family of partitions indexed by an ultrametric distance. Actually, after the tree structure of the dendrogram has been set up, the most tricky problem is that of cutting the tree with a suitable threshold in order to take out a sub-optimal classification. Several (more or less) objective criteria may be used to achieve this goal, e.g. the deepest step, but most often the partition relies on a subjective choice leaded by interpretation issues. Additionally, whatever the chosen criterion is, only

one solution can be obtained for each desired granularity, i.e. the one where clusters are joined at consecutive heights starting from the adopted threshold.

In this paper we propose an algorithm, exploiting the methodological framework of permutation test, allowing to find out automatically a sub-optimal partition where clusters do not necessarily obey to the afore-mentioned principle. The algorithm allows us to explore partitions which are not directly achievable using a standard cut-level approach.

Further works should concern a comparison of the obtained partition with respect to partitions of the same dataset deriving from common partitioning methods. A comparison in terms of quite common quality indexes (Rand, 1971) should strength the proposal. Furthermore the study of the stability of the obtained partitions with respect to tuning parameters used in the permutation test procedure and the study of the computational complexity are topics of interest for further research.

## References

- GOOD P. (1994). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer, New York.
- IDEKER T., THORSSON V., RANISH J.A., CHRISTMAS R., BUHLER J., ENG J.K., BUMGARNER R.E., GOODLETT D.R., AEBERSOLD R., HOOD L. (2001) Integrated genomic and proteomic analyses of a systemically perturbed metabolic network. *Science*, 292: 1929-1934.
- MAECHLER M., ROUSSEEUW P., STRUYF A., HUBERT M. (2005). Cluster Analysis Basics and Extensions. *unpublished*.
- QIU W.L., JOE, H. (2006a) Generation of Random Clusters with Specified Degree of Separation. *Journal of Classification*, 23(2), 315-334.
- QIU W.L., JOE, H. (2006b) Separation Index and Partial Membership for Clustering. *Computational Statistics and Data Analysis*, 50, 585-603.
- QIU W. L., JOE H. (2009). clusterGeneration: random cluster generation (with specified degree of separation). *R package version 1.2.7*.
- R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, url = <http://www.R-project.org>.
- RAND W.M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, December 1971, 66, 336, 846-850.

# Design of Least-Squares Quadratic Estimators Based on Covariances from Interrupted Observations Transmitted by Different Sensors

R. Caballero-Águila<sup>1</sup>, A. Hermoso-Carazo<sup>2</sup> and J. Linares-Pérez<sup>2</sup>

<sup>1</sup> Dpto. de Estadística e I.O., Universidad de Jaén, 23071 Jaén, Spain,  
*raguila@ujaen.es*

<sup>2</sup> Dpto. de Estadística e I.O., Universidad de Granada, 18071 Granada, Spain,  
*ahermoso@ugr.es, jlinares@ugr.es*

**Abstract.** The least-squares quadratic estimation problem of discrete-time signals from uncertain noisy observations coming from multiple sensors is addressed. The uncertainty about the signal being present or missing in the observation of each sensor is modelled by a set of Bernoulli random variables whose probabilities are not necessarily the same for all the sensors. It is assumed that only information on the moments (up to the fourth-order ones) of the signal and observation noise is available. The recursive quadratic estimation algorithm is derived from a linear estimation algorithm for a suitably defined augmented system.

**Keywords:** least-squares quadratic estimation, uncertain observations, multiple sensors

## 1 Introduction

In many real systems the signal to be estimated can be randomly missing in the observations due, for example, to intermittent failures or random interruptions in the observation mechanism. These situations are characterized by including in the observation equation a multiplicative noise modelling the possibility that the signal may not be present in the observations (*uncertain observations*).

On the other hand, in some practical situations the state-space model of the signal is not available and another type of information must be processed for the estimation. In the last years, the estimation problem from uncertain observations has been investigated using covariance information and algorithms with a simpler structure than those obtained when the state-space model is known have been derived (see e.g. Nakamori et al. (2003)).

Recently, the least-squares linear estimation problem using uncertain observations transmitted by multiple sensors, whose statistical properties are assumed not to be the same, has been studied by several authors under different approaches and hypotheses on the processes (see e.g. Hounkpevi and Yaz (2007) for a state-space approach, and Jiménez-López et al. (2008) for a covariance approach). In this paper a least-squares quadratic filtering algorithm is proposed for this multi-sensor observation model. For the quadratic estimation approach the augmented signal and observation vectors are introduced by assembling the original vectors with their

second-order powers and, by using an innovation approach, the linear estimator of the augmented signal based on the augmented observations is obtained, providing the required quadratic estimator.

## 2 Problem formulation

The problem at hand is to determine the least-squares (LS) quadratic estimator of an  $n$ -dimensional discrete signal,  $z_k$ , from noisy measurements coming from multiple sensors which may not contain the signal with different probabilities. In this section, we present the observation model and the hypotheses about the signal and noise processes underlying it.

Consider  $m$  scalar sensors whose measurements at each sampling time,  $k$ , denoted by  $y_k^i$ , may either contain the signal to be estimate,  $z_k$ , or be only additive noise,  $v_k^i$ ; the uncertainty about the signal being present or missing in the observation is modelled by Bernoulli variables,  $\gamma_k^i$ . Thus, the observation model is described as follows:

$$y_k^i = \gamma_k^i H_k^i z_k + v_k^i, \quad k \geq 1, \quad i = 1, \dots, m \quad (1)$$

If  $\gamma_k^i = 1$ , then  $y_k^i = H_k^i z_k + v_k^i$  and the measurement of the  $i$ th sensor contains the signal; otherwise, if  $\gamma_k^i = 0$ , then  $y_k^i = v_k^i$ , which means that such measurement is only noise. Therefore, the variables  $\{\gamma_k^i; k \geq 1\}$  model the uncertainty of the observations coming from the  $i$ th sensor.

To simplify the notation, the observation equation (1) is rewritten in a compact form as follows:

$$y_k = \Upsilon_k H_k z_k + v_k, \quad k \geq 1, \quad (2)$$

where  $y_k = (y_k^1, \dots, y_k^m)^T$ ,  $H_k = (H_k^{1T}, \dots, H_k^{mT})^T$ ,  $\Upsilon_k = \text{Diag}(\gamma_k^1, \dots, \gamma_k^m)$  and  $v_k = (v_k^1, \dots, v_k^m)^T$ .

It is known that if the signal  $z_k$  and the observations  $y_1, \dots, y_k$  have finite second-order moments, the LS linear filter of  $z_k$  is the orthogonal projection of  $z_k$  on the space of  $n$ -dimensional random variables obtained as linear transformations of  $y_1, \dots, y_k$ . So, by defining the random vectors  $y_i^{[2]} = y_i \otimes y_i$  ( $\otimes$  denotes the Kronecker product, Magnus and Neudecker (1999)) and, if  $E[y_i^{[2]T} y_i^{[2]}] < \infty$ , the LS quadratic estimator of  $z_k$  based on the observations up to the sampling time  $k$ , is the orthogonal projection of  $z_k$  on the space of  $n$ -dimensional linear transformations of  $y_1, \dots, y_k$  and their second-order powers  $y_1^{[2]}, \dots, y_k^{[2]}$ . To guarantee the existence of the second-order moments of the vectors  $y_i^{[2]}$ , the pertinent assumptions about the processes in (1) are now stated:

### **Hypotheses about the model:**

**(H1)** The  $n \times 1$  signal process  $\{z_k; k \geq 1\}$  has zero mean and its autocovariance function,  $K_{k,s}^z$ , as well as the autocovariance function of the second-order powers,  $K_{k,s}^{z[2]}$ , are expressed in a semi-degenerate kernel form,

$$K_{k,s}^z = A_k B_s^T, \quad s \leq k, \quad K_{k,s}^{z[2]} = a_k b_s^T, \quad s \leq k,$$

where the  $n \times M$  matrix functions  $A$ ,  $B$ , and the  $n^2 \times L$  matrix functions  $a$ ,  $b$ , are known. Moreover, it is assumed that the covariance function of the signal and its second-order powers,  $K_{k,s}^{zz[2]}$ , can also be expressed as



$$K_{k,s}^{zz[2]} = \begin{cases} \alpha_k \beta_s^T, & s \leq k, \\ \varepsilon_k \delta_s^T, & k \leq s, \end{cases}$$

where  $\alpha, \beta, \varepsilon$  and  $\delta$  are  $n \times N$ ,  $n^2 \times N$ ,  $n \times P$  and  $n^2 \times P$  known matrix functions, respectively.

(H2) For  $i = 1, \dots, m$ , the sensor additive noises,  $\{v_k^i; k \geq 1\}$ , are zero-mean white processes and their moments, up to the fourth one, are known; we will denote  $R_k = Cov[v_k]$ ,  $R_k^{(3)} = Cov[v_k, v_k^{[2]}]$  and  $R_k^{(4)} = Cov[v_k^{[2]}]$ .

(H3) For  $i = 1, \dots, m$ , the noise  $\{\gamma_k^i; k \geq 1\}$  is a sequence of independent Bernoulli variables with known probabilities,  $P[\gamma_k^i = 1] = p_k^i, \forall k \geq 1$ .

(H4) The signal process,  $\{z_k; k \geq 1\}$ , and the noise processes,  $\{\gamma_k^i; k > 1\}$  and  $\{v_k^i; k \geq 1\}$ , for  $i = 1, \dots, m$ , are mutually independent.

### 3 Quadratic estimation problem

Given the observation model (1) with assumptions (H1)-(H4), the problem is to find the LS quadratic filter,  $z_{k/k}^q$ , of the signal,  $z_k$ . The technique used to obtain this estimator consists of augmenting the signal and observation by assembling the original vectors and their second-order powers,  $\mathcal{Z}_k = \begin{pmatrix} z_k^T, z_k^{[2]T} \end{pmatrix}^T$ ,  $\mathcal{Y}_k = \begin{pmatrix} y_k^T, y_k^{[2]T} \end{pmatrix}^T$  and deriving the estimator  $z_{k/k}^q$  as the vector constituted by the first  $n$  entries of the LS linear filter of  $\mathcal{Z}_k$ .

To obtain this linear estimator, the first and second-order statistical properties of the augmented vectors  $\mathcal{Z}_k$  and  $\mathcal{Y}_k$  are now analyzed.

**Properties of the augmented vectors.** By using the Kronecker product properties and denoting  $\mathcal{D}_k^\gamma = Diag(\Upsilon_k, \Upsilon_k^{[2]})$ ,  $\mathcal{H}_k = Diag(H_k, H_k^{[2]})$  and

$$\mathcal{V}_k = \begin{pmatrix} v_k \\ (I_{m^2} + K_{m^2})((\Upsilon_k H_k z_k) \otimes v_k) + v_k^{[2]} \end{pmatrix},$$

( $I_{m^2}$  is the  $m^2 \times m^2$  identity matrix and  $K_{m^2}$  is the  $m^2 \times m^2$  commutation matrix, Magnus and Neudecker (1999)) we obtain another model with uncertain observations,

$$\mathcal{Y}_k = D_k^\gamma \mathcal{H}_k \mathcal{Z}_k + \mathcal{V}_k, \quad k \geq 1.$$

It should be noted that the signal,  $\mathcal{Z}_k$ , and the noise,  $\mathcal{V}_k$ , in this new model have non-zero mean. Nevertheless, this handicap can be overcome by defining the centered augmented vectors  $Z_k = \mathcal{Z}_k - E[\mathcal{Z}_k]$  and  $Y_k = \mathcal{Y}_k - E[\mathcal{Y}_k]$  which, taking into account that  $E[D_k^\gamma \mathcal{H}_k \mathcal{Z}_k] = E[D_k^\gamma] \mathcal{H}_k E[\mathcal{Z}_k]$ , satisfy:

$$Y_k = D_k^\gamma \mathcal{H}_k Z_k + V_k, \quad k \geq 1 \quad (3)$$

where

$$V_k = \begin{pmatrix} v_k \\ (I_{m^2} + K_{m^2})((\Upsilon_k H_k z_k) \otimes v_k) + v_k^{[2]} - vec(R_k) \end{pmatrix} + (D_k^\gamma - E[D_k^\gamma]) \mathcal{H}_k E[\mathcal{Z}_k]$$

being  $vec$  the operator that vectorizes a matrix (Magnus and Neudecker (1999)).

Note that the LS linear estimator of  $\mathcal{Z}_k$  based on  $\mathcal{Y}_1, \dots, \mathcal{Y}_k$  is obtained from the LS linear estimator of  $Z_k$  based on  $Y_1, \dots, Y_k$ , just adding the mean vector

$E[Z_k]$ . Hence, since the first  $n$  components of  $E[Z_k]$  are zero, the required quadratic estimator  $z_{k/k}^q$  is just the vector constituted by the first  $n$  entries of the LS linear filter of  $Z_k$ . Henceforth, these centered vectors will be referred to as the augmented signal and observation vectors, respectively.

The signal and noise processes  $\{Z_k; k \geq 1\}$  and  $\{V_k; k \geq 1\}$  involved in model (3) are zero-mean. In the following propositions the second-order statistical properties of these processes are established.

**Proposition 1.** If the signal process  $\{z_k; k \geq 1\}$  satisfies (H1), the autocovariance function of the augmented signal process  $\{Z_k; k \geq 1\}$  can be expressed in a semi-degenerate kernel form, namely

$$K_{k,s}^Z = E[Z_k Z_s^T] = \mathcal{A}_k \mathcal{B}_s^T, \quad s \leq k,$$

where

$$\mathcal{A}_k = \begin{pmatrix} A_k & \alpha_k & 0_{n \times P} & 0_{n \times L} \\ 0_{n^2 \times M} & 0_{n^2 \times N} & \delta_k & a_k \end{pmatrix}, \quad \mathcal{B}_k = \begin{pmatrix} B_k & 0_{n \times N} & \varepsilon_k & 0_{n \times L} \\ 0_{n^2 \times M} & \beta_k & 0_{n^2 \times P} & b_k \end{pmatrix}.$$

**Proposition 2.** Under (H1)-(H4), the noise  $\{V_k; k \geq 1\}$  is a sequence of mutually uncorrelated random vectors with covariance matrices given by

$$E[V_k V_k^T] = \bar{R}_k = \begin{pmatrix} R_k & R_k^{(3)} \\ R_k^{(3)T} & R_k^{22} \end{pmatrix} + Cov[C_k^\gamma] \circ \left( \mathcal{H}_k E[Z_k] E[Z_k]^T \mathcal{H}_k^T \right)$$

where

$$R_k^{22} = (I_{m^2} + K_{m^2}) \left( (E[C_{T_k} C_{T_k}^T] \circ (H_k A_k B_k^T H_k^T)) \otimes R_k \right) (I_{m^2} + K_{m^2}) + R_k^{(4)},$$

$$C_k^\gamma = \left( C_{T_k}^T, C_{T_k}^{[2]T} \right)^T \text{ with } C_{T_k} = (\gamma_k^1, \dots, \gamma_k^m)^T, \circ \text{ denotes the Hadamard product}$$

and  $E[Z_k] = (0_{n \times 1}^T, (vec(A_k B_k^T))^T)^T$ . Moreover,  $\{V_k; k \geq 1\}$  is uncorrelated with the processes  $\{Z_k; k \geq 1\}$  and  $\{D_k^\gamma \mathcal{H}_k Z_k; k > 1\}$ .

**Linear estimation of the augmented signal.** As indicated above, to obtain the LS quadratic estimators of the signal,  $z_k$ , based on the observations (1), we consider the LS linear estimation problem of the augmented signal,  $Z_k$ , based on the augmented observations (3). To address this problem, we use the innovation approach, which simplifies considerably the derivation of the filtering algorithm, since the innovations constitute a white process.

Let  $\nu_i$  be defined as  $\nu_i = Y_i - \hat{Y}_{i/i-1}$ , where  $\hat{Y}_{i/i-1}$  denotes the LS linear estimator of  $Y_i$  based on the previous observations,  $Y_1, \dots, Y_{i-1}$ . For each  $i$ ,  $\nu_i$  may be regarded as a measure of the new information or the *innovation* provided by the observation  $Y_i$ . It is known that the innovations  $\{\nu_i, i \leq k\}$  can be determined from the observations  $\{Y_i, i \leq k\}$  by means of a causal and causally invertible linear transformation. Therefore, each set can be replaced by the other with no loss of information and, consequently, the LS linear filter of the signal  $Z_k$  based on the observations  $Y_1, \dots, Y_k$ , which is denoted by  $\hat{Z}_{k/k}$ , is equal to the LS linear estimator given the innovations  $\nu_1, \dots, \nu_k$ . Since the innovations constitute a white process, from the Orthogonal Projection Lemma it is easily proven that the filter is given by

$$\hat{Z}_{k/k} = \sum_{i=1}^k S_{k,i} \Pi_i^{-1} \nu_i \quad (4)$$

where  $S_{k,i} = E[Z_k \nu_i^T]$  and  $\Pi_i = E[\nu_i \nu_i^T]$ .

## 4 Quadratic filtering algorithm

Using the properties of the processes involved in equation (3), as established in propositions 1 and 2, and expression (4) for the filter, we derive a recursive algorithm for the linear filtering estimators,  $\hat{Z}_{k/k}$ , of the augmented signal  $Z_k$ . The first  $n$  entries of these estimators provide the required quadratic filter of the original signal  $z_k$ .

**Theorem 1.** The quadratic filter,  $z_{k/k}^q$ , of the original signal  $z_k$  is given by

$$z_{k/k}^q = \Xi \hat{Z}_{k/k}, \quad k \geq 1,$$

where  $\Xi$  is the operator which extracts the first  $n$  entries of  $\hat{Z}_{k/k}$ , the linear filter of the augmented signal  $Z_k$ , which is obtained by

$$\hat{Z}_{k/k} = \mathcal{A}_k O_k, \quad k \geq 1,$$

where the vectors  $O_k$  are recursively calculated from

$$O_k = O_{k-1} + J_k \Pi_k^{-1} \nu_k, \quad k \geq 1; \quad O_0 = 0.$$

The matrix function  $J$  is given by

$$J_k = [\mathcal{B}_k^T - r_{k-1} \mathcal{A}_k^T] \mathcal{H}_k^T D_k^p, \quad k \geq 1,$$

with  $r_k$  being recursively obtained from

$$r_k = r_{k-1} + J_k \Pi_k^{-1} J_k^T, \quad k \geq 1; \quad r_0 = 0.$$

The innovation,  $\nu_k$ , satisfies

$$\nu_k = Y_k - D_k^p \mathcal{H}_k \mathcal{A}_k O_{k-1}, \quad k \geq 1,$$

and  $\Pi_k$ , the covariance matrix of the innovation, verifies

$$\Pi_k = E [C_k^\gamma C_k^{\gamma T}] \circ (\mathcal{H}_k \mathcal{A}_k \mathcal{B}_k^T \mathcal{H}_k^T) - D_k^p \mathcal{H}_k \mathcal{A}_k r_{k-1} \mathcal{A}_k^T \mathcal{H}_k^T D_k^p + \bar{R}_k, \quad k \geq 1.$$

Finally, as a measure of the estimation accuracy, we have calculated the filtering error covariance matrices,  $\Sigma_{k/k} = E [Z_k Z_k^T] - E [\hat{Z}_{k/k} \hat{Z}_{k/k}^T]$ , which clearly are obtained by  $\Sigma_{k/k} = \mathcal{A}_k [\mathcal{B}_k^T - r_k \mathcal{A}_k^T]$ ,  $k \geq 1$ .

## 5 Numerical simulation example

To illustrate the application of the proposed filtering algorithm a numerical simulation example is shown now. To check the effectiveness of the proposed quadratic filter, we ran a program in MATLAB, simulating at each iteration the signal and the observed values and providing the linear and quadratic filtering estimates, as well as the corresponding error covariance matrices.

This program has been applied to a scalar signal  $\{z_k; k \geq 1\}$ , generated by the following first-order autoregressive model,

$$z_k = 0.95 z_{k-1} + w_{k-1}, \quad k \geq 1$$

where the initial state is a zero-mean Gaussian variable with  $Var[z_0] = 1$  and  $\{w_k; k \geq 0\}$  is a zero-mean white Gaussian noise with  $Var[w_k] = 0.1$ .

The autocovariance functions of the signal and their second-order powers are given in a semidegenerate kernel form, specifically,

$$K_{k,s}^z = 1.025641 \times 0.95^{k-s}, \quad K_{k,s}^{zz^2} = 2.1038795 \times 0.95^{2(k-s)}, \quad s \leq k$$

and the covariance function of the signal and their second-order powers is given by  $K_{k,s}^{zz^2} = 0$ ,  $\forall s, k$ . According to hypothesis (H.1), the functions constituting these covariance functions can be defined as follows:

$$A_k = 1.025641 \times 0.95^k, \quad B_k = 0.95^{-k}, \quad a_k = 2.1038795 \times 0.95^{2k}, \quad b_k = 0.95^{-2k}, \\ \alpha_k = \beta_k = \varepsilon_k = \delta_k = 0.$$

Consider two sensors which whose measurements, according to our theoretical study, are perturbed by sequences of independent Bernoulli random variables  $\{\gamma_k^i; k \geq 1\}$ ,  $i = 1, 2$ , and by independent additive white noises,  $\{v_k^i; k \geq 1\}$ ,  $i = 1, 2$ ; that is:

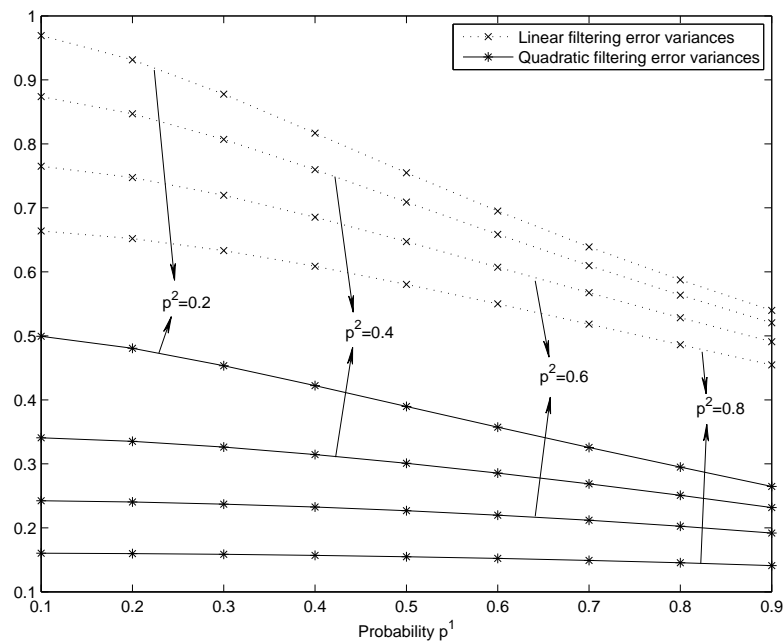
$$y_k^i = \gamma_k^i z_k + v_k^i, \quad k \geq 1, \quad i = 1, 2.$$

Assume that,  $P[\gamma_k^i = 1] = p^i$ , for all  $k \geq 1$ , and that the additive white noises have the following probability distributions:

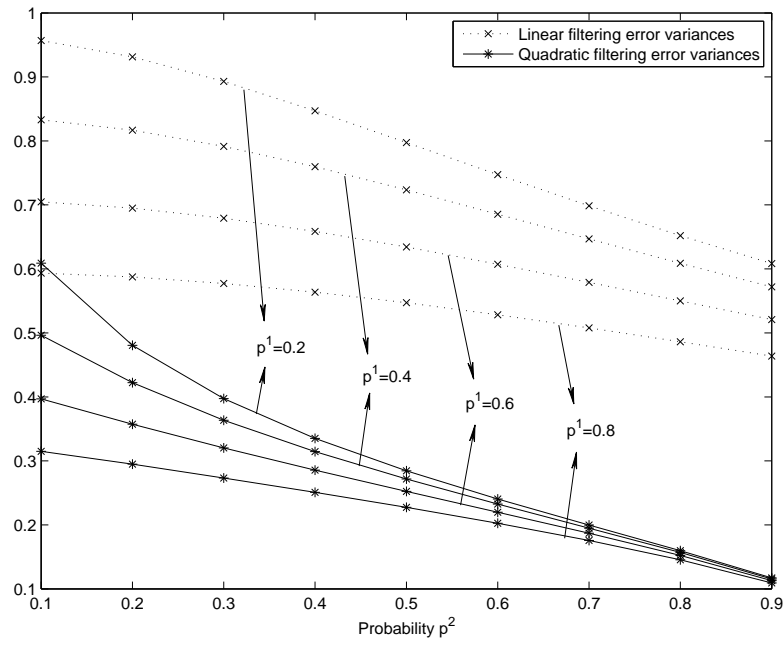
$$P[v_k^1 = -8] = \frac{1}{8}, \quad P\left[v_k^1 = \frac{8}{7}\right] = \frac{7}{8}, \quad \forall k \geq 1, \\ P[v_k^2 = 1] = \frac{15}{18}, \quad P[v_k^2 = -3] = \frac{2}{18}, \quad P[v_k^2 = -9] = \frac{1}{18}, \quad \forall k \geq 1.$$

To compare the performance of the linear and quadratic filtering estimators, the filtering error variances are calculated for different values of  $p^1$  and  $p^2$ . Such variances show insignificant variation from the 5th iteration onwards and, consequently, only the error variances at a specific iteration are considered. In Figure 1 the linear and quadratic filtering error variances at  $k = 50$  are displayed versus  $p^1$  (for constant values of  $p^2$ ) and versus  $p^2$  (for constant values of  $p^1$ ). From these figures it is gathered that, as  $p^1$  or  $p^2$  increase (and, consequently, the probability that the signal is equal to one in the observations increases), the error variances become smaller and, hence, better estimations are obtained. Note that this improvement is more significant for small values of  $p^1$  or  $p^2$ ; that is, when the probability that the signal is present in the observations coming from one of the sensors is small. On the other hand, both figures show that, for all the values of  $p^1$  and  $p^2$ , the error variances for the quadratic filter are smaller than those of the linear filter, which confirms the superiority of the quadratic filter over the linear one in estimation accuracy.

- JIMÉNEZ-LÓPEZ, J.D., LINARES-PÉREZ, J., NAKAMORI, S., CABALLERO-ÁGUILA, R. and HERMOSO-CARAZO, A. (2008): Signal estimation based on covariance information from observations featuring correlated uncertainty and coming from multiple sensors. *Signal Processing*, 88, 2998–3006.
- MAGNUS J.R. and NEUDECKER H. (1999): *Matrix differential calculus with applications in statistics and econometrics (revised edn.)*. Wiley, New York.
- NAKAMORI, S., CABALLERO-ÁGUILA, R., HERMOSO-CARAZO, A. and LINARES-PÉREZ, J. (2003): Linear estimation from uncertain observations with white plus coloured noises using covariance information. *Digital Signal Processing*, 138, 552–568.



**Fig. 1.** Linear and quadratic filtering error variances versus probability  $p^1$ , with  $p^2 = 0.2, 0.4, 0.6, 0.8$ .



**Fig. 2.** Linear and quadratic filtering error variances versus probability  $p^2$ , with  $p^1 = 0.2, 0.4, 0.6, 0.8$ .

# Pseudo-Bayes Factors

Stefano Cabras<sup>1</sup>, Walter Racugno<sup>1</sup>, and Laura Ventura<sup>2</sup>

<sup>1</sup> Department of Mathematics, University of Cagliari  
Via Ospedale 72, Cagliari, Italy, *s.cabras@unica.it*, *racugno@unica.it*

<sup>2</sup> Department of Statistics, University of Padova  
Via C. Battisti 241, Padova, Italy, *ventura@stat.unipd.it*

**Abstract.** The use of Bayes factors (BF) in hypothesis testing may encounter difficulties in the presence of unknown nuisance parameters. Indeed, their elimination requires in general both the computation of multidimensional integrals and the elicitation of prior distributions. In modern frequentist and Bayesian literature, the elimination of nuisance parameters can be carried out by resorting to pseudo-likelihood functions. Here, we propose to substitute in the BF the integrated likelihood with a suitable pseudo-likelihood of the parameter of interest only. A new formulation of the BF is derived, called pseudo-Bayes factor. The properties of the proposed pseudo-Bayes factors are investigated through two examples.

**Keywords:** frequentist risk, hypothesis testing, marginal posterior distribution, pseudo-likelihoods

## 1 Introduction

Consider a sampling model  $p(y; \theta)$  with parameter  $\theta \in \Theta \subseteq \mathbb{R}^p$ ,  $p > 1$ . Suppose that  $\theta = (\psi, \lambda)$ , where  $\psi$  is a scalar parameter of interest and  $\lambda$  a  $(p-1)$ -dimensional nuisance parameter. For the elimination of the nuisance parameter, modern statistical literature indicates that non-Bayesian methods based on pseudo-likelihoods can be usefully incorporated into classical Bayesian analyses (see, among others, Reid (1995), Severini (1999), Chang and Mukerjee (2006), Racugno et al. (2009), Ventura et al. (2009)). This approach has the remarkable advantages of avoiding elicitation on the nuisance parameters and computation of multidimensional integrals.

In Bayesian model selection, when hypotheses involve unknown nuisance parameters, also Bayes factors (see e.g. Kass and Raftery (1995)) have the drawback of calling for priors on these parameters and present computational difficulties associated with the evaluation of marginal density functions. In this setting, this paper introduces in the formulation of the Bayes factor a suitable pseudo-likelihood for  $\psi$  only, which substitutes the integrated likelihood. In this way, we obtain a new index of evidence, named pseudo-Bayes factor, which mimics the genuine Bayes factor. The aim of this paper is to discuss inferential properties of pseudo-Bayes factors through examples. Moreover, we give a criterion, based on the frequentist risk function, to compare pseudo-Bayes factors with genuine Bayes factors.

The remainder of the paper is organized as follows. In Section 2 posterior distributions derived from pseudo-likelihoods are briefly reviewed. In Section 3 we introduce pseudo-Bayes factors and we propose a criterion to compare them with

genuine Bayes factors. In Section 4 we discuss two examples on the use of the proposed pseudo-Bayes factor, also in comparison with the genuine Bayes factor. Some final remarks are given in Section 5.

## 2 Posterior distributions from pseudo-likelihoods

Let  $y = (y_1, \dots, y_n)$  be a random sample from  $p(y; \psi, \lambda)$ . In the Bayesian framework, elimination of the nuisance parameter  $\lambda$  may be carried out using an appropriate pseudo-likelihood  $L^*(\psi)$ , i.e. a function of the parameter of interest only, with properties similar to those of a genuine likelihood function. Examples of pseudo-likelihoods for a parameter of interest are the conditional, the marginal, the profile, the modified profile, and the integrated likelihoods (see e.g. Pace and Salvan (1997), Severini (2000)).

If  $L^*(\psi)$  is treated as a true likelihood, then the following posterior distribution for  $\psi$  can be formally considered

$$\pi^*(\psi|y) \propto \pi(\psi)L^*(\psi), \quad (1)$$

where  $\pi(\psi)$  is a prior on  $\psi$  only. Although (1) cannot always be considered as orthodox in a Bayesian setting, the use of alternative likelihoods is nowadays widely shared, and several papers are devoted to Bayesian interpretation and applications of some well-known pseudo-likelihoods. See, among others, Monahan and Boos (1992), Severini (1999), Lazar (2003), Chang and Mukerjee (2006), Greco et al. (2008), Racugno et al. (2009), Ventura et al. (2009), and references therein.

There are two main advantages in using  $\pi^*(\psi|y)$  instead of the marginal posterior distribution

$$\pi_m(\psi|y) \propto \int_{\Lambda} L(\psi, \lambda) \pi(\psi, \lambda) d\lambda, \quad (2)$$

where  $L(\psi, \lambda)$  denotes the complete likelihood function and  $\pi(\psi, \lambda)$  is a prior on  $(\psi, \lambda)$ . First, elicitation over  $\lambda$ , for hypothesis testing on  $\psi$ , is not necessary and, second, computation of the integral in (2) is avoided by using  $L^*(\psi)$ . See Reid (1995), Racugno et al. (2009) and Ventura et al. (2009) for several examples and applications.

## 3 Pseudo-Bayes factors

Suppose it is of interest to test  $H_0 : \psi \in \Psi_0$  versus  $H_1 : \psi \in \Psi_1$  in the presence of the additional nuisance parameter  $\lambda$  in  $\Lambda$ . Let  $\pi_0(\psi, \lambda)$  and  $\pi_1(\psi, \lambda)$  be the prior probabilities conditionally on  $H_0$  and  $H_1$ , respectively. The Bayes factor of  $H_0$  against  $H_1$  is (see e.g. Kass and Raftery (1995))

$$BF = \frac{\int_{\Psi_0} \int_{\Lambda} L(\psi, \lambda) \pi_0(\psi, \lambda) d\lambda d\psi}{\int_{\Psi_1} \int_{\Lambda} L(\psi, \lambda) \pi_1(\psi, \lambda) d\lambda d\psi}. \quad (3)$$

The Bayes factor (3) presents computational difficulties associated with the evaluation of marginal density functions of the form (2), which can be heavy when the dimension of  $\lambda$  is high. Moreover, it requires the elicitation of a prior on  $\theta \in \Theta$ .



Consider a pseudo-likelihood  $L^*(\psi)$  and let  $\pi_0(\psi)$  and  $\pi_1(\psi)$  two proper priors on the parameter of interest only, conditional on  $H_0$  and  $H_1$ , respectively. The pseudo-Bayes factor of  $H_0$  against  $H_1$  is thus defined as

$$BF^* = \frac{\int_{\Psi_0} L^*(\psi) \pi_0(\psi) d\psi}{\int_{\Psi_1} L^*(\psi) \pi_1(\psi) d\psi} . \quad (4)$$

Associated to (4), also the weight of evidence  $w^* = \log BF^*$  can be considered, with the usual interpretation: the greater the value of  $w^*$ , the stronger the evidence against  $H_0$  (see e.g. Good (1985)).

An advantage in using the pseudo-Bayes factor instead of (3) is that  $BF^*$  involves only scalar integrals and then there is no need of Monte Carlo methods, which are often needed for the genuine Bayes factors (see e.g. Chen et al. (2000)).

To compare  $BF^*$  with  $BF$ , we consider the frequentist risk, in terms of the usual 0-1 loss function for hypothesis testing, where  $H_0$  is selected if  $BF^* > 1$  (or  $BF > 1$ ). Formally, the decision,  $d(BF^*)$  (or  $d(BF)$ ) based on the observed data  $y$  is

$$d(BF^*) = d(BF^*(y)) = \begin{cases} H_0 & \text{if } BF^*(y) > 1 \\ H_1 & \text{if } BF^*(y) < 1 \end{cases} , \quad (5)$$

and the loss function for the true hypothesis  $H_i$  is

$$l(d(BF^*), H_i) = \begin{cases} 0 & \text{if } d(BF^*) = H_i \\ 1 & \text{if } d(BF^*) \neq H_i \end{cases} , \quad i = 0, 1 , \quad (6)$$

and, similarly, for  $l(d(BF), H_i)$ ,  $i = 0, 1$ . The frequentist risk  $R^*$  (or  $R$ ) under the true hypothesis is then given by

$$R^* = \Pr(BF^*(y) < 1 | H_0) + \Pr(BF^*(y) > 1 | H_1) .$$

If  $BF^*$  outperforms  $BF$ , we expect a lower risk.

## 4 Examples

Two examples in the context of nested models are here discussed. We note, however, that  $BF^*$  can be applied also for non-nested models, such as for model discrimination between separate scale and regression models (see e.g. Pace et al. (2006, 2009)).

Risks  $R$  for the genuine Bayes factor and risks  $R^*$  for the pseudo-Bayes factor are compared. In both the examples, an orthogonal parametrization between the parameter of interest  $\psi$  and the nuisance parameter  $\lambda$  is used. In view of this, in the evaluation of (3), independence between  $\psi$  and  $\lambda$  can be assumed. In particular, the full priors can be written in the following form

$$\pi_0(\psi, \lambda) = \pi_0(\psi)\pi(\lambda) , \quad \pi_1(\psi, \lambda) = \pi_1(\psi)\pi(\lambda) ,$$

where  $\pi_0(\psi)$  and  $\pi_1(\psi)$  are the same priors used in  $BF^*$ , while  $\pi(\lambda)$  is a suitable prior on  $\lambda$  chosen under a favorable scenario for  $BF$ , i.e. such that  $\pi(\lambda)$  gives high probability to the true values of  $\lambda$ .

The frequentist risk has been approximated at a certain point  $(\psi, \lambda)$  using a Monte Carlo sum with 1000 terms. All integrals have been computed using adaptive quadrature.

#### 4.1 Example 1: Stress-strength model

A stress-strength model is concerned with the statistical problem of evaluating the reliability parameter  $\psi = \Pr(X < Y)$ , where  $X$  and  $Y$  are independent random variables. For example, in a clinical study,  $X$  is the response of a control group,  $Y$  the response of a treatment group and  $\psi$  measures the effectiveness of the treatment. Moreover, in a reliability study,  $X$  is the stress applied to the system,  $Y$  the strength of a system and  $\psi$  measures the chance that the system does not fail (see Kotz et al. (2003)).

Let us assume that  $X$  and  $Y$  are independent exponentially distributed, with rates  $\alpha$  and  $\beta$ , respectively. In this case,  $\psi = \Pr(X < Y) = \alpha/(\alpha + \beta)$ . Let  $(x_1, \dots, x_n)$  be a random sample of size  $n$  from  $X$  and let  $(y_1, \dots, y_m)$  be a random sample of size  $m$  from  $Y$ . Moreover, let  $\theta = (\psi, \lambda)$ , with  $\lambda = \alpha$ , and assume that it is of interest to test  $H_0 : \psi > 1/2$  against  $H_1 : \psi < 1/2, \forall \lambda > 0$ .

The genuine  $BF$  is based on the full likelihood  $L(\psi, \lambda) = \lambda^{n+m} \left(\frac{1-\psi}{\psi}\right)^m \exp\left\{-\lambda\left(n\bar{x} + m\bar{y}\frac{1-\psi}{\psi}\right)\right\}$ , with  $\bar{x} = \sum x_i/n$  and  $\bar{y} = \sum y_i/m$ , and on the priors  $\pi_0(\psi, \lambda) = \pi_0(\psi)\pi(\lambda)$  and  $\pi_1(\psi, \lambda) = \pi_1(\psi)\pi(\lambda)$ , with uniform priors for  $\psi$ , i.e.  $\pi_0(\psi) = U(0, 1/2)$  and  $\pi_1(\psi) = U(1/2, 1)$ , and  $\pi(\lambda) = \text{Gamma}(1, 1)$ . The Bayes factor

$$BF = \frac{\int_0^{1/2} \int_{\lambda>0} L(\psi, \lambda) \pi_0(\psi, \lambda) d\psi d\lambda}{\int_{1/2}^1 \int_{\lambda>0} L(\psi, \lambda) \pi_1(\psi, \lambda) d\psi d\lambda} = \frac{\int_0^{1/2} \frac{1-\psi}{\psi} \left(b + s_x + \frac{1-\psi}{\psi} s_y\right)^{-(m+n+a)} d\psi}{\int_{1/2}^1 \frac{1-\psi}{\psi} \left(b + s_x + \frac{1-\psi}{\psi} s_y\right)^{-(m+n+a)} d\psi}$$

has not an analytical form and a double integral is required.

The pseudo-Bayes factor can be based on the modified profile likelihood (see e.g. Severini (2000)), given by

$$L_{mp}(\psi) = -(n+m-2) \log\left(n\bar{x} + m\bar{y}\frac{1-\psi}{\psi}\right) + m \log \frac{1-\psi}{\psi},$$

and the uniform priors  $\pi_0(\psi)$  and  $\pi_1(\psi)$ . The pseudo-Bayes factor,

$$BF^* = \frac{\int_0^{1/2} L_{mp}(\psi) \pi_0(\psi) d\psi}{\int_{1/2}^1 L_{mp}(\psi) \pi_1(\psi) d\psi},$$

has not an analytical form.

Some values for the risks  $R$  and  $R^*$  corresponding to  $BF$  and  $BF^*$ , respectively, are given in Table 1. A picture of the differences between  $R$  and  $R^*$  is given in Figure 1. From Table 1 and Figure 1, it may be noted that  $R^*$  is lower than  $R$  and does not depend on  $\lambda$ .

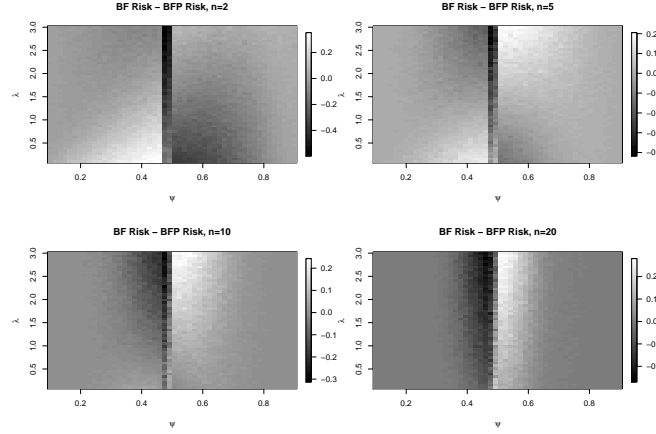
#### 4.2 Example 2: Logistic Regression

The logistic regression model has likelihood function

$$L(\beta) = \exp\left\{\sum_{i=1}^n y_i \sum_{j=1}^p \beta_j x_{ij} - \sum_{i=1}^n \log\left(1 + e^{\sum_{j=1}^p \beta_j x_{ij}}\right)\right\},$$

	$R$		$R^*$	
	$\lambda = 0.5$	$\lambda = 2.5$	$\lambda = 0.5$	$\lambda = 2.5$
$\psi = 0.4$	10%	33 %	21%	23 %
$\psi = 0.6$	29%	22 %	21%	26 %

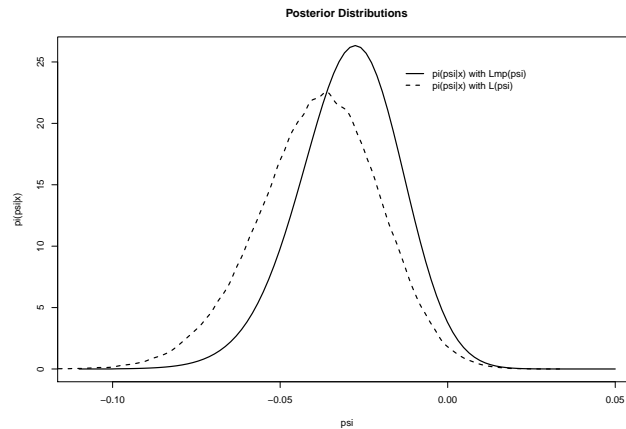
**Table 1.** Values of risks (with  $n = m = 5$ ).



**Fig. 1.** Difference between  $R$  and  $R^*$  (positive values for  $R^*$  lower than  $R$ ).

with  $\beta = (\beta_1, \dots, \beta_p)$  unknown regression coefficient and  $x_{ij}$  fixed constants,  $i = 1, \dots, n$  and  $j = 1, \dots, p$ . Let us assume that the parameter of interest is  $\psi = \beta_p$  and take  $\lambda = (\beta_1, \dots, \beta_{p-1})$  to be the nuisance parameter. Suppose it is of interest to test  $H_0 : \psi < 0$  against  $H_1 : \psi > 0$ ,  $\forall \lambda$ . Accurate inference about  $\psi$  may be based on the modified profile likelihood  $L_{mp}(\psi)$  (see e.g. Severini (2000)), which is a third-order frequentist approximation to the conditional distribution of  $\sum_{i=1}^n x_{ip} y_i$  given the maximum likelihood estimator of  $\lambda$ . For comparison, also the simpler profile likelihood can be considered, given by  $L_p(\psi) = L(\psi, \hat{\lambda}_\psi)$ , with  $\hat{\lambda}_\psi$  restricted maximum likelihood estimate of  $\lambda$  for fixed  $\psi$ . To compute the pseudo-Bayes factors, we assume the Truncated Normal priors  $\pi_0(\psi) = TN(0, 1, \underline{\psi} = -\infty, \bar{\psi} = 0)$  and  $\pi_1(\psi) = TN(0, 1, \underline{\psi} = 0, \bar{\psi} = \infty)$ .

For illustration, we analyze the urine data of Davison and Hinkley (1997) on the presence or absence of calcium oxalate crystals in urine samples together with the values of six quantitative covariates measured for 77 individuals. Assume that  $\psi$  is the coefficient of the effect of the variable urea concentration. These data have been considered also in Brazzale et al. (2008), using the profile likelihood  $L_p(\psi)$  and the modified profile likelihood  $L_{mp}(\psi)$  for frequentist inference. These two pseudo-likelihoods induce the posterior distributions shown in Figure 2. The weights of evidence  $w^*$  based on  $L_p(\psi)$  and  $L_{mp}(\psi)$  are similar, being 4.2 and 3.8, respectively. These values express strong evidence in favour of  $H_0$ , according to the Jeffreys' scale, which confirm a negative effect of the urea concentration on the



**Fig. 2.** Posterior distributions based on  $L_p(\psi)$  and  $L_{mp}(\psi)$ , and  $\pi_m(\psi|y)$ .

presence of calcium oxalate crystals. For comparison, the genuine  $BF$  is computed assuming a flat prior on all the coefficients in  $\lambda$ , i.e. the normal density  $N(0, (10^6)^2)$ . The corresponding marginal posterior distribution is given in Figure 2. In this case the weight of evidence is 4.5, quite similar to  $w^*$ .

The two  $BF^*$  have been compared in a simulation study with  $n = 300$  and twenty covariates positively correlated. Five coefficients are assumed positive (assumed equal to 1), five are negative (assumed equal to -1), and the remaining coefficients are zero. The empirical weights of evidence, in 1000 replications of the data set, for each coefficient  $\psi$  are illustrated in Figure 3. Positive and negative coefficients are clearly recognizable from the  $w^*$  and inference based on  $L_p(\psi)$  and  $L_{mp}(\psi)$  is very similar.

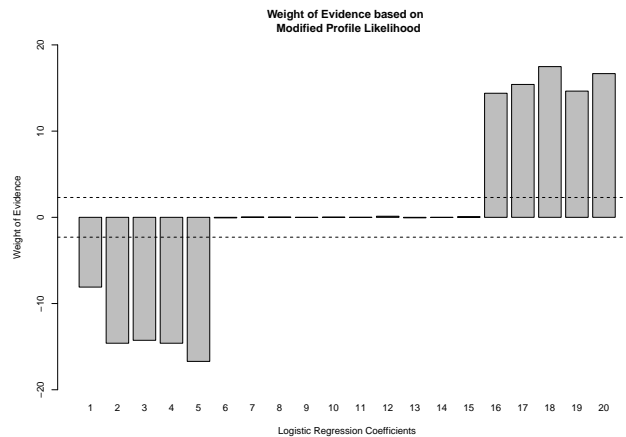
## 5 Conclusions

This paper explores the potentiality of hypothesis testing based on pseudo-Bayes factors. In the considered examples, the risk function for  $BF^*$  is generally lower than that based on the genuine Bayes factor. Moreover,  $BF^*$  does not depend on the values of nuisance parameters.

As a final remark, note that the proposed pseudo-Bayes factors can be computed also using pseudo-profile likelihood functions derived from estimating equations, such as quasi- and empirical profile likelihoods. These pseudo-likelihoods allows to deal those problems in which Bayesian integrated likelihoods may be not available (see Greco et al. (2008) and Ventura et al. (2009b)).

## References

BRAZZALE, A.R., DAVISON, A.C. and REID, N. (2007): *Applied Asymptotics*. Cambridge University Press, Cambridge.



**Fig. 3.** Empirical weights of evidence for positive and negative coefficients. Horizontal dash lines are the levels of strong evidence.

- CHANG, H. and MUKERJEE, R. (2006): Probability matching property of adjusted likelihoods. *Statistics and Probability Letters* 76 (8), 838–842.
- CHEN, M., SHAO, Q.M. and IBRAHIM, J.G. (2000): *Monte Carlo Methods in Bayesian Computation*. Springer, New York.
- GOOD, I.J. (1985): Weight of evidence: A brief survey. *Bayesian Statistics 2*, 249–269.
- GRECO, L., RACUGNO, W. and VENTURA, L. (2008): Robust likelihood functions in Bayesian inference. *Journal of Statistical Planning and Inference* 138 (5), 1258–1270.
- KASS, R.E. and RAFTERY, A.E. (1995): Bayes factors. *Journal of the American Statistical Association* 90 (430), 773–795.
- KOTZ, S., LUMELSKII, Y. and PENSKY, M. (2003): *The Stress-Strength Model and its Generalizations*. World Scientific, Singapore.
- LAZAR, N.A. (2003): Bayesian empirical likelihood. *Biometrika* 90 (2), 319–326.
- MONAHAN, J.F. and BOOS, D.D. (1992): Proper likelihoods for Bayesian analysis. *Biometrika* 79 (2), 271–278.
- PACE, L. and SALVAN, A. (1997): *Principles of Statistical Inference from a Neo-Fisherian Perspective*. World Scientific, Singapore.
- PACE, L., SALVAN, A. and VENTURA, L. (2006): Likelihood based discrimination between separate scale and regression models. *Journal of Statistical Planning and Inference* 136 (10), 3539–3553.
- PACE, L., SALVAN, A. and VENTURA, L. (2009): Remedying the Neyman-Scott phenomenon in model discrimination. *Journal of Statistical Computation and Simulation*, in press.
- RACUGNO, W., SALVAN, A. and VENTURA, L. (2009): Bayesian analysis in regression models using pseudo-likelihoods. *Communications in Statistics - Theory and Methods*, in press.
- REID, N. (1995): Likelihood and Bayesian approximation methods. *Bayesian Statistics 5*, 351–368.

- SEVERINI, T.A. (1999): On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters. *Statistica Sinica* 9 (3), 713–724.
- SEVERINI, T.A. (2000): *Likelihood Methods in Statistics*. Oxford University Press.
- VENTURA, L., CABRAS, S. and RACUGNO, W. (2009): Prior distributions from pseudo-likelihoods in the presence of nuisance parameters, *Journal of the American Statistical Association* 104 (486), 768–777.
- VENTURA, L., CABRAS, S. and RACUGNO, W. (2010): Default prior distributions from quasi- and quasi-profile likelihoods. *Journal of Statistical Planning and Inference* to appear.

# Diagnostic Checking of Multivariate Normality Under Contamination

Andrea Cerioli

Dipartimento di Economia, Università di Parma  
via Kennedy 6, 43100 Parma, Italy, *andrea.cerioli@unipr.it*

**Abstract.** The normal distribution has a central place in the analysis of multivariate data. It is also important in the development of high-breakdown methodologies, since it is often assumed to describe the genesis of the “good” part of the data. In this paper we describe a simple and effective way to assess multivariate normality which can be used when the data contain outliers. Our proposal is based on accurate distributional results and can be seen as a robust version of the classical diagnostic methods applied to detect departures from multivariate normality.

**Keywords:** outliers, quantile plot, reweighted MCD, robust distances

## 1 Introduction

The normal distribution has a central place in the analysis of multivariate data, since it underlies most of the classical multivariate statistical methodologies. It is thus not surprising that much effort has been devoted to solve the problem of assessing multivariate normality. This task is more complex than in the univariate case, requiring consideration of the interrelationships among the variables. Given a sample  $y = (y_1, \dots, y_n)'$  of  $n$   $v$ -variate observations, many of the available methods are based on inspection of the squared Mahalanobis distances

$$d_i^2 = (y_i - \hat{\mu})' \hat{\Sigma}^{-1} (y_i - \hat{\mu}) \quad i = 1, \dots, n, \quad (1)$$

where  $\hat{\mu}$  and  $\hat{\Sigma}$  are the sample mean and the sample (unbiased) covariance matrix. See, e.g., Gnanadesikan (1997, pp. 187–220) or Mecklin and Mundfrom (2004) for a review.

The multivariate normal distribution is also important in the development of robust high-breakdown methodologies, where it is often assumed to describe the genesis of the “good” part of the data. For instance, under the multivariate normal model it is possible to obtain workable estimates of the correction factors which ensure consistency and approximate unbiasedness of the robust estimators of scatter (Hubert et al. (2008)), or to define precise outlier identification rules (Hardin and Roche (2005); Riani et al. (2009); Cerioli (2010)). However, the assessment of multivariate normality carries additional difficulties when multiple outliers are present, because the classical estimates  $\hat{\mu}$  and  $\hat{\Sigma}$  may break down and the squared Mahalanobis distances (1) become unreliable. An even more fundamental issue is that, under contamination, the hypothesis of multivariate normality should be checked

only for a portion of the available data. Strictly speaking, an effective test of multivariate normality based on the whole sample should be able to reject the null hypothesis when outliers are present.

In this paper we suggest a simple and effective graphical way to assess multivariate normality which can be used when the data contain outliers. Our proposal is based on accurate distributional results for a robust version of the squared Mahalanobis distances (1). To motivate the method, in Section 2 we show examples of departure from multivariate normality which are hard to discriminate through the available outlier identification rules. Our proposal is described in Section 3 and then applied in Section 4 to solve the problem.

## 2 Outliers and non-normal models

We consider the squared robust distances which are computed from the reweighted Minimum Covariance Determinant (RMCD) estimator of  $\mu$  and  $\Sigma$ . This estimator is given by

$$\hat{\mu}_{(\text{RMCD})} = \frac{1}{m} \sum_{i=1}^n w_i y_i \quad (2)$$

and

$$\hat{\Sigma}_{(\text{RMCD})} = \frac{k_{(\text{RMCD})}}{m-1} \sum_{i=1}^n w_i (y_i - \hat{\mu}_{(\text{RMCD})})(y_i - \hat{\mu}_{(\text{RMCD})})', \quad (3)$$

where  $k_{(\text{RMCD})}$  is a consistency factor and  $m = \sum_{i=1}^n w_i$ . In the version of (2) and (3) considered here, the weights  $w_i$  are defined as follows:

$$\begin{aligned} w_i &= 0 \quad \text{if the squared robust distance for } y_i \text{ computed from the raw MCD} \\ &\quad \text{estimator of Rousseeuw and Van Driessen (1999) exceeds the 0.975} \\ &\quad \text{quantile of Hardin and Rocke (2005) scaled } F \text{ distribution;} \\ w_i &= 1 \quad \text{otherwise.} \end{aligned} \quad (4) \quad (5)$$

Being based on a high-breakdown estimator, the squared robust distances

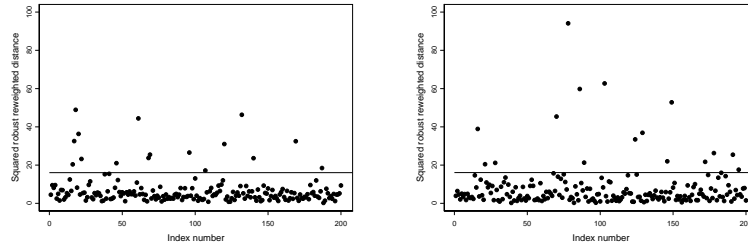
$$d_{i(\text{RMCD})}^2 = (y_i - \hat{\mu}_{(\text{RMCD})})' \hat{\Sigma}_{(\text{RMCD})}^{-1} (y_i - \hat{\mu}_{(\text{RMCD})}) \quad i = 1, \dots, n, \quad (6)$$

can be used in the place of (1) to identify multivariate outliers without suffering from masking and swamping. Cerioli (2010) shows how to construct accurate yet powerful detection rules starting from these distances when the uncontaminated part of the data is normally distributed.

However, the information provided by the robust distances may become less straightforward when the whole sample  $y$ , or a large majority of it, comes from a multivariate non-normal distribution. An example of the ambiguity that can arise in such a situation is provided in Figure 1. The left-hand panel shows the values of the robust reweighted distances (6) for a sample of  $n = 200$  observations on  $v = 5$  variables. Of these observations, 184 come from the postulated  $N(0, I)$  distribution. The remaining 16 observations are simulated from the location-shift contamination model  $N(0 + \lambda e, I)$ , where  $\lambda$  is a positive scalar and  $e$  is a column vector of ones. In this example  $\lambda = 2.0$ , a moderate amount of contamination. The right-hand panel gives the same information for a sample of  $n = 200$  observations simulated from



the 5-variate  $t$  distribution on 6 degrees of freedom, again a moderate deviation from normality. The threshold displayed in each picture is the 0.99 quantile of the scaled  $F_{v,m-v}$  distribution suggested by Cerioli (2010) for the squared distances of the units trimmed in the reweighting step.



**Fig. 1.** Robust squared distances and scaled  $F$  cut-off values for multiple outlier detection in two samples with  $n = 200$  and  $v = 5$ . Left: 184 observations from  $N(0, I)$  and 16 observations from a location-shift model. Right: 200 observations from the 5-variate  $t$  distribution on 6 degrees of freedom.

The message conveyed by the two pictures is surprisingly similar, with about the same number of observations labelled as outliers and a few borderline units. The main perceivable difference between the plots is the extreme distance originated by the multivariate  $t$  distribution, but in practice this single effect might be attributed to an additional source of contamination. Therefore, we argue that it is hard to distinguish between these two alternative situations using an outlier identification rule based on the extreme distances. A less perceivable difference between the two panels of Figure 1 is a relative shortage of small distances in the right-hand plot. Correspondingly, the number of moderately large distances (but still below the threshold) is higher. We elaborate this idea in the next section.

### 3 Robust distances for checking normality

The following result is useful for the purpose of this paper. It is derived by Cerioli (2010). Atkinson et al. (2008) also considered a related version of it in a different context.

Assume that  $y_i \sim N(\mu, \Sigma)$ , with  $\mu$  and  $\Sigma$  known. The squared distance (1) then becomes

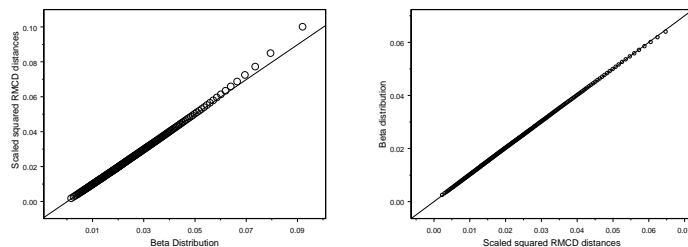
$$D_i^2 = (y_i - \mu)' \Sigma^{-1} (y_i - \mu) \quad i = 1, \dots, n.$$

Correspondingly, replace condition (4) by

$$w_i = 0 \text{ if } D_i^2 \text{ exceeds the 0.975 quantile of the } \chi_v^2 \text{ distribution, say } \chi_{v,0.975}^2, \quad (7)$$

since  $D_i^2 \sim \chi_v^2$ . In the scatter formula (3), take

$$k_{(\text{RMCD})} = \frac{0.975}{P(\chi_{v+2}^2 < \chi_{v,0.975}^2)}. \quad (8)$$



**Fig. 2.** Q-Q plot of the scaled squared distances  $m/(m-1)^2 d_{i(\text{RMCD})}^2$ , with weights (4) and (5), against the Beta distribution. Left: all  $n$  ordered distances. Right: only the first  $m$  ones.

If we ignore the variability in  $m$ , the distribution of  $d_{i(\text{RMCD})}^2$  is then approximately scaled Beta for the units for which  $w_i = 1$ :

$$d_{i(\text{RMCD})}^2 \approx \frac{(m-1)^2}{m} \text{Beta} \left( \frac{v}{2}, \frac{m-v-1}{2} \right) \quad \text{if} \quad w_i = 1. \quad (9)$$

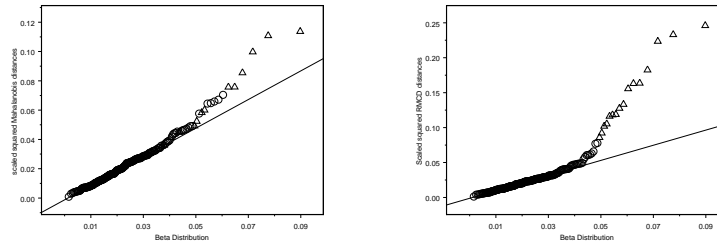
The above distributional result suggests that also the squared reweighted distances (6), based on the sample weights (4) and (5), could be compared to the given scaled Beta distribution, thus mimicking the behaviour of the classical squared Mahalanobis distances. The accuracy of the suggested approximation is evaluated in Figure 2. This picture, for  $n = 200$  and  $v = 5$ , compares the empirical quantiles of the scaled squared reweighted distances

$$\frac{m}{(m-1)^2} d_{i(\text{RMCD})}^2 \quad i = 1, \dots, n, \quad (10)$$

with weights (4) and (5), to the theoretical ones derived from (9). The empirical quantiles of the squared distances are estimated by running 1000 simulations from  $N(0, I)$ . We set  $m = 0.975n = 195$  to compute the scaling in (10) and the quantiles of (9). The left-hand panel plots the whole set of  $n$  ordered scaled squared distances, while the right-hand panel provides a zoom into the first  $m = 195$  ones. It is clearly seen that the approximation for the first 195 ordered distances is very good, even if the weights are estimated.

Result (9) holds under the multivariate normal model and concerns the observations for which  $w_i = 1$ , which contribute to the computation of (2) and (3). Therefore, it can be used to check the adequacy of the normal model without being affected even by a relatively large number of outliers, such as those displayed in Figure 1. If  $n$  is not too small, the correlations between the squared distances  $d_{i(\text{RMCD})}^2$  may be ignored. In the next section we thus rely on Q-Q plots that compare the robust squared distances  $d_{i(\text{RMCD})}^2$ , computed on the units for which  $w_i = 1$ , to the corresponding quantiles of distribution (9). These plots may be considered as a robust extension of the classical graphical techniques based on the squared Mahalanobis distances  $d_i^2$  (Healy (1968); Small (1978)). Formal statistical testing procedures that exploit approximation (9) are not considered here, but will be studied elsewhere.

Finally, it is important to note that our approach is not equivalent to the simplistic procedure which computes the squared Mahalanobis distances (1) on the



**Fig. 3.** Simulated data set with 184 observations from  $N(0, I)$  (circles) and 16 observations from a location-shift model (triangles). Left-hand panel:  $Q$ - $Q$  plot of the  $n$  scaled squared Mahalanobis distances against the  $\text{Beta}(2.5, 97)$  distribution. Right-hand panel: same plot, but for scaled robust squared distances.

subset of observations that remain after removal of the outliers, as suggested, e. g., by Singh (1993). The factor  $k_{(\text{RMCD})}$  given in Equation (8) is crucial to allow for trimming of the outliers (at the specified trimming proportion) and to reach the accurate approximation shown in Figure 2.

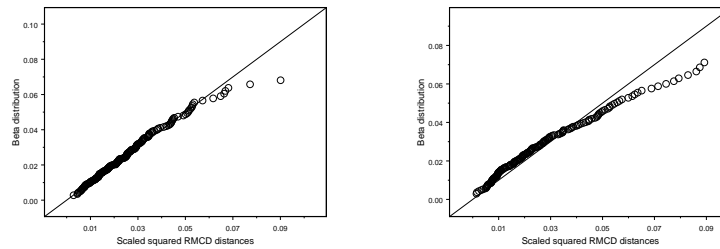
## 4 Data Analysis

### 4.1 Simulated data

We start our analysis of the simulated data sets introduced in Section 2 by first appreciating the importance of relying on robust distances. The left-hand panel of Figure 3 is a  $Q$ - $Q$  plot of the scaled squared Mahalanobis distances  $\{n/(n-1)^2\}d_i^2$ ,  $i = 1, \dots, n$ , for the contaminated normal data set with 184 “good” observations and 16 moderate outliers. The right-hand panel repeats the plot for the robust squared distances  $d_{i(\text{RMCD})}^2$  with the same scaling. In both panels the reference distribution is  $\text{Beta}(v/2, (n-v-1)/2)$  and the 16 contaminated observations are displayed as triangles. Even in presence of a mild amount of contamination, masking and swamping heavily affect the classical Mahalanobis distances (1). On the other hand, the robust reweighted distances (6) show the structure of contamination. Singh (1993) suggested to use the latter representation also to detect departures from normality. However, the distribution of  $d_{i(\text{RMCD})}^2$  is not scaled Beta if  $w_i = 0$ .

The right-hand panel of Figure 3 looks similar to the analogous plot (not shown) drawn from the multivariate  $t$  data set described in Section 2. This reinforces the idea that the information provided by the extreme robust distances is less useful when the goal is to detect departures from normality for the central part of the data. We need to inspect more closely the distances that do not lie in the tail of the distribution in order to see these departures.

We now draw  $Q$ - $Q$  plots of the scaled robust squared distances  $\{m/(m-1)^2\}d_{i(\text{RMCD})}^2$  of the units for which  $w_i = 1$ , against the quantiles of the Beta distribution suggested by (9). These theoretical quantiles need to allow for trimming of the 2.5% largest distances when reshaping the plot from the whole set of  $n$  distances to the subset of the  $m$  largest ones, as shown in Figure 2. The left-hand panel of Figure 4 refers to the simulated data set with 184 observations from



**Fig. 4.**  $Q$ - $Q$  plot of the scaled robust squared distances  $\{m/(m-1)^2\}d_{i(\text{RMCD})}^2$  of the units for which  $w_i = 1$ , against the Beta distribution. Left-hand panel: simulated data set with 184 observations from  $N(0, I)$  and 16 observations from a location-shift model ( $m = 183$ ). Right-hand panel: Simulated data set with 200 observations from the 5-variate  $t$  distribution on 6 degrees of freedom ( $m = 175$ ).

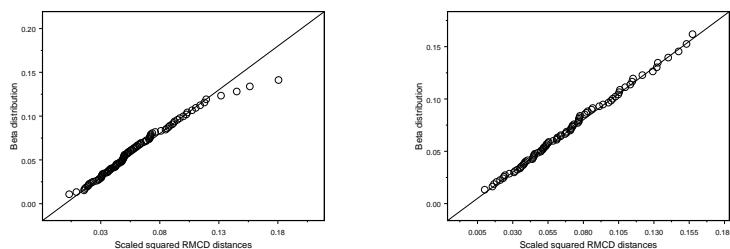
$N(0, I)$ , for which we obtain  $m = 183$ . The effect of the outliers is removed from the plot, which shows appreciable closeness to the linear fit, except for two borderline units. We can thus correctly conclude that the hypothesis of multivariate normality is supported by the subset of “good” observations. The right-hand panel repeats the analysis for the simulated data set from the 5-variate  $t$  distribution on 6 degrees of freedom, for which we obtain  $m = 175$ . The excess of relatively large distances is now revealed by the systematic pattern in the upper part of the plot. Mild non-linearity is also present in the central part. Therefore, we are now able to assess that the normality assumption is not tenable for these data even when the extreme observations are trimmed from the sample.

The lesson provided by this simulated example is that our diagnostic technique for multivariate normality check adds further insight to the outlier detection rule shown in Figure 1. By making use of the central part of the data, it allows us to discriminate between the two reported contamination schemes, even if the amount of contamination is moderate and the sample sizes are relatively small.

## 4.2 Swiss banknotes

The data introduced by Flury and Riedwyl (1988) on Swiss banknotes are often used to describe the results of classical multivariate methods. These data contain information on six variables measuring the size and other features of 200 notes, 100 of which are classified as genuine and 100 as forged. The hypothesis of multivariate normality might be reasonable for the first sample, which is obtained under supposedly tighter quality control rules. On the other hand, the group of forged notes is known to be heterogeneous, with 15 well separated outliers, perhaps due to the action of different forgers (see, e.g., Atkinson et al. (2004); Garcia-Escudero and Gordaliza (2005); Cerioli (2010)). We now analyze the two groups of banknotes separately to check the normality assumption without being affected by outliers.

The left-hand panel of Figure 5 compares the scaled squared robust distances of the  $m = 97$  genuine notes for which  $w_i = 1$  to the hypothesized Beta distribution. The plot looks similar to the representation given in the left-hand panel of Figure 4. Except for few units, the data are very close to the linear fit and the appropriateness of the normality assumption for the bulk of the data is confirmed. The right-hand



**Fig. 5.** Swiss banknotes:  $Q$ - $Q$  plot of the scaled squared robust distances  $\{m/(m-1)^2\}d_{i(\text{RMCD})}^2$  of the units for which  $w_i = 1$ , against the Beta distribution. Left-hand panel: genuine notes ( $m = 97$ ). Right-hand panel: forged notes ( $m = 84$ ).

panel is for the group of forged notes, for which we obtain  $m = 84$ . When the outliers are trimmed, it is seen that also the production of forged notes is very well approximated by a multivariate normal model.

## Acknowledgments

This research was partially supported by the grant “Nuovi metodi multivariati robusti per la valutazione delle politiche sull’e-government e la società dell’informazione” of Ministero dell’Università e della Ricerca – PRIN 2008.

## References

- ATKINSON A. C., RIANI M. AND CERIOLO, A. (2004): *Exploring Multivariate Data with the Forward Search*. Springer, Berlin.
- ATKINSON A. C., RIANI M. AND CERIOLO, A. (2008): Monitoring random start forward searches for multivariate data. In: Brito, P. (Ed.): *COMPSTAT 2008*. Physica-Verlag, Heidelberg, 447–458.
- CERIOLO, A. (2010): Multivariate outlier detection with high-breakdown estimators *Journal of the American Statistical Association*, 105 (489), 147–156.
- FLURY, B. and RIEDWYL, H. (1988): *Multivariate Statistics: A Practical Approach*. Chapman and Hall, London.
- GARCIA-ESCUDERO, L. A. and GORDALIZA, A. (2005): Generalized radius processes for elliptically contoured distributions *Journal of the American Statistical Association* 100 (471), 1036–1045.
- GNANADESIKAN, R. (1997): *Methods for Statistical Data Analysis of Multivariate Observations*. Second Edition. Wiley, New York.
- HARDIN, J. and ROCKE, D. M. (2005): The distribution of robust distances *Journal of Computational and Graphical Statistics* 14 (4), 928–946.
- HEALY, M. J. R. (1968): Multivariate normal plotting *Applied Statistics* 17, 157–161.
- HUBERT, M., ROUSSEEUW, P. J. and VAN AELST, S. (2008): High-breakdown robust multivariate methods *Statistical Science* 23 (1), 92–119.

- MECKLIN, C. J. and MUNDFROM, D. J. (2004): An appraisal and bibliography of tests for multivariate normality. *International Statistical Review* 72 (1), 123–138.
- RIANI, M., ATKINSON, A. C. and CERIOLI, A. (2009): Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society B* 71 (2), 447–466.
- ROUSSEEUW, P. J. and VAN DRIESSEN, K. (1999): A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41 (3), 212–223.
- SINGH, A. (1993): Omnibus robust procedures for assessment of multivariate normality and detection of multivariate outliers. In: Patil, G.P. and Rao, C.R. (Eds.): *Multivariate Environmental Statistics*. Elsevier, Amsterdam, 445–488.
- SMALL, N. J. H. (1978): Plotting squared radii. *Biometrika* 65 (3), 657–658.

# On Computationally Complex Instances of the $c$ -optimal Experimental Design Problem: Breaking *RSA*-based Cryptography via $c$ -optimal Designs

Michal Černý,<sup>1</sup> Milan Hladík<sup>2</sup> and Veronika Skočdoplová<sup>1</sup>

<sup>1</sup> University of Economics Prague, Department of Econometrics  
Winston Churchill Square 4, 130 67 Prague, Czech Republic, [cernym@vse.cz](mailto:cernym@vse.cz),  
[veronika.skocdoplova@vse.cz](mailto:veronika.skocdoplova@vse.cz)

<sup>2</sup> Charles University, Department of Applied Mathematics  
Malostranské náměstí 25, 118 00 Prague, Czech Republic, [hladik@mff.cuni.cz](mailto:hladik@mff.cuni.cz)

**Abstract.** We study the computational complexity of the problem to find a  $c$ -optimal experimental design over a finite experimental domain. We construct instances of the problem which are computationally very difficult: we show how any algorithm for  $c$ -optimality can be used for integer factoring and hence for breaking the *RSA* cryptographic protocol. These ‘hard’ instances can also be used as a benchmark for testing algorithms for finding  $c$ -optimal designs.

**Keywords:**  $c$ -optimal experimental design, cryptography, *RSA*, integer factoring

## 1 Introduction

Consider the regression model  $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$  where  $\boldsymbol{\beta} \in \mathbb{R}^M$  is the unknown vector of regression parameters,  $\mathbf{X} \in \mathbb{R}^{N \times M}$  and  $\mathbf{y}$  is a vector of  $N$  uncorrelated (say, normal) variables with variance  $\sigma^2 \in (0, \infty)$ . Let a nonzero vector  $\mathbf{c} \in \mathbb{R}^M$  be given; the number

$$\text{var}_c(\mathbf{X}) = N \cdot \mathbf{c}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{c}$$

is called *the  $c$ -variance of  $\mathbf{X}$*  (here  $^{-1}$  denotes a matrix pseudoinverse). The  $c$ -variance of  $\mathbf{X}$  is proportional to the variance of the OLS-estimator  $\mathbf{c}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  of  $\mathbf{c}^T \boldsymbol{\beta}$  and is often used as a measure of the contribution of the matrix  $\mathbf{X}$  to the total variance of the OLS-estimator.

Let  $M$  and  $N$  be fixed. Given a set  $\mathcal{X} \subseteq \mathbb{R}^M$ , called *the experimental domain*, a matrix  $\mathbf{X}$  is called  $\mathcal{X}$ -correct if for any row  $\mathbf{x}^T$  of  $\mathbf{X}$  it holds  $\mathbf{x} \in \mathcal{X}$ . An  $\mathcal{X}$ -correct matrix  $\mathbf{X} \in \mathbb{R}^{N \times M}$  is called  $c$ -optimal if for any  $\mathcal{X}$ -correct matrix  $\mathbf{X}' \in \mathbb{R}^{N \times M}$  it holds  $\text{var}_c(\mathbf{X}) \leq \text{var}_c(\mathbf{X}')$ .

The  *$c$ -optimal experimental design problem* is the following problem: *given  $N$ ,  $\emptyset \neq \mathcal{X} \subseteq \mathbb{R}^M$  and  $\mathbf{0} \neq \mathbf{c} \in \mathbb{R}^M$ , find (any)  $c$ -optimal  $\mathcal{X}$ -correct matrix  $\mathbf{X}$* . The complexity of this optimization problem clearly depends on  $\mathcal{X}$ . We study a (seemingly) simple case of  $\mathcal{X}$  finite. This case occurs if we can control the experiment conditions only in discrete steps, if the domain  $\mathcal{X}$  is so intricate that only a description of  $\mathcal{X}$

in terms of a grid is available or if the domain points have been generated by some kind of a nondeterministic or random process.

*Remark.* In practice,  $\mathcal{X}$  is often a finite union of intervals or a finite union of compact sets. Observe that our construction implies that the case of union of compact sets cannot be computationally easier than the finite-domain case.

Let  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_K\}$  be the finite domain. Now the optimal matrix  $\mathbf{X}$  is (up to a permutation of rows) uniquely determined by a vector  $\boldsymbol{\xi} \in \mathbb{R}^K$  such that  $\boldsymbol{\xi} \geq \mathbf{0}$ ,  $\mathbf{1}^T \boldsymbol{\xi} = 1$  and for any  $i$ ,  $N \cdot (\boldsymbol{\xi})_i$  is integral. (The symbol  $(\cdot)_i$  denotes the  $i$ -th component of a vector and  $\mathbf{1}$  is an all-one vector.) The vector  $\boldsymbol{\xi}$ , called *design*, has the meaning that  $N \cdot (\boldsymbol{\xi})_i$  observations of the dependent variable are to be made in the  $i$ -th point of the domain  $\mathcal{X}$ . I. e., the design matrix  $\mathbf{X}$  has  $N$  rows  $\mathbf{X}_1, \dots, \mathbf{X}_N$ , such that

$$\begin{aligned} \mathbf{X}_1 = \dots = \mathbf{X}_{N \cdot (\boldsymbol{\xi})_1} = \mathbf{x}_1^T; \quad \mathbf{X}_{1+N \cdot (\boldsymbol{\xi})_1} = \dots = \mathbf{X}_{N \cdot ((\boldsymbol{\xi})_1 + (\boldsymbol{\xi})_2)} = \mathbf{x}_2^T; \dots; \\ \mathbf{X}_{1+N \cdot \sum_{i=1}^{k-1} (\boldsymbol{\xi})_i} = \dots = \mathbf{X}_N = \mathbf{x}_k^T. \end{aligned}$$

The main result of this text is that this problem is computationally so complex that any algorithm for this problem can be used for breaking a wide class of cryptographic protocols used in practice. On one hand, this shows an unexpected connection between a purely statistical problem, the experimental design, and cryptography (in particular, number theory). On the other hand, this shows that the general formulation of the design problem, which is usual in statistical literature, admits ‘exotic’ instances and it would be worth considering whether these instances are relevant for practical statistics or not. We shall discuss this problem in Sect. 4. By the way, we construct a set of ‘complex’ instances of the design problem that may serve as a benchmark for testing new algorithms for the design problem.

The best known cryptographic protocol is RSA, by Ron Rivest, Adi Shamir and Leonard Adleman. Let us briefly recall that it is a public-key protocol based on the following idea. Let  $p_1, p_2$  be two distinct primes and  $z := p_1 p_2$ . Let  $\mathbb{Z}_K$  be the group  $\{1 \leq \kappa < K : \kappa \text{ is coprime with } K\}$  where the group operation is multiplication mod  $K$ . Denote  $\Phi(K) = |\mathbb{Z}_K|$ . Let  $\kappa \in \mathbb{Z}_{\Phi(z)}$  and  $\kappa^{-1}$  be the inverse to  $\kappa$  in  $\mathbb{Z}_{\Phi(z)}$ . The *public key* is  $(z, \kappa)$  and the *private key* is  $(z, \kappa^{-1})$ . A message  $x$  is regarded as a number  $x \in \mathbb{Z}_z$ ; the *encryption function* is  $E(x) = \text{mod}(x^\kappa, z)$  and the *decryption function* is  $D(y) = \text{mod}(y^{\kappa^{-1}}, z)$ . *Breaking the RSA protocol* means *being able to compute the decryption function given the public key  $(z, \kappa)$* , or, in other words, *being able to compute  $\kappa^{-1}$  given  $(z, \kappa)$* . Observe that *if, given  $z$ , we are able to find its prime factors  $p_1$  and  $p_2$ , then computing  $\kappa^{-1}$  is trivial*. Said otherwise: the security of RSA relies on the assumption that *it is computationally extremely hard to find  $p_1$  and  $p_2$  given  $z$* . We show that any algorithm for the experimental design problem is necessarily capable of doing this. In other words, given  $z$ , we construct an instance of the experimental design problem such that from the optimal design  $\boldsymbol{\xi}$  it is possible to reconstruct  $p_1$  and  $p_2$



and hence to break the protocol. Or, more in general, *any algorithm for the experimental design problem is capable of breaking any cryptographic protocol relying on hardness of prime factoring.*

The proof has two steps: first we construct a boolean formula formalising multiplication of natural numbers and then we convert the formula into an instance of the experimental design problem.

## 2 The formula

We shall write down boolean formulae in the basis  $\&, \vee, \neg$ . The symbols  $x + y$  and  $x = y$  stand for  $(x \vee y) \& \neg(x \& y)$  and  $(x \& y) \vee (\neg x \& \neg y)$ , respectively. Bitvectors are typeset in bold. Natural numbers are written down in binary as bitvectors. First, we consider the formula

$$S(\alpha_{2n-1}, \alpha_{2n-2}, \dots, \alpha_1; \beta_{2n-1}, \beta_{2n-2}, \dots, \beta_1; d_{2n-1}, d_{2n-2}, \dots, d_1; z_{2n}, z_{2n-1}, \dots, z_1)$$

with the following meaning: if  $\alpha = \alpha_{2n-1}\alpha_{2n-2} \dots \alpha_1$ ,  $\beta = \beta_{2n-1}\beta_{2n-2} \dots \beta_1$  and  $z = z_{2n}z_{2n-1} \dots z_1$  are natural numbers, then  $z$  is the sum of  $\alpha$  and  $\beta$ . The  $d_i$ 's are uniquely determined by  $\alpha_i$ 's and  $\beta_i$ 's; they stand for carry bits. The formula  $S$ , formalising the 'school' addition algorithm, can be written as

$$\begin{aligned} z_1 &= [\alpha_1 + \beta_1] \& \\ d_1 &= [\alpha_1 \& \beta_1] \& \\ z_2 &= [\alpha_2 + \beta_2 + d_1] \& \\ d_2 &= [(\alpha_2 \& \beta_2) \vee (\alpha_2 \& d_1) \vee (\beta_2 \& d_1)] \& \\ z_3 &= [\alpha_3 + \beta_3 + d_2] \& \\ d_3 &= [(\alpha_3 \& \beta_3) \vee (\alpha_3 \& d_2) \vee (\beta_3 \& d_2)] \& \\ &\vdots \\ z_{2n-1} &= [\alpha_{n-1} + \beta_{n-1} + d_n] \& \\ d_{2n-1} &= [(\alpha_{n-1} \& \beta_{n-1}) \vee (\alpha_{n-1} \& d_{n-2}) \vee (\beta_{n-1} \& d_{n-2})] \& \\ z_{2n} &= d_{2n-1}. \end{aligned} \tag{1}$$

Now consider the following formula  $g$  ( $uv$  is an abbreviation for  $u \& v$ ):

$$\begin{aligned} &S(\mathbf{0}_n, x_n y_1, x_{n-1} y_1, \dots, x_1 y_1; \\ &\quad \mathbf{0}_{n-1}, x_n y_2, x_{n-1} y_2, \dots, x_1 y_2, \mathbf{0}_1; \mathbf{d}^1; \mathbf{z}^1) \& \\ &S(\mathbf{0}_{n-2}, x_n y_3, x_{n-1} y_3, \dots, x_1 y_3, \mathbf{0}_2; \mathbf{z}^1; \mathbf{d}^2; \mathbf{z}^2) \& \\ &S(\mathbf{0}_{n-3}, x_n y_3, x_{n-1} y_3, \dots, x_1 y_3, \mathbf{0}_3; \mathbf{z}^2; \mathbf{d}^3; \mathbf{z}^3) \& \\ &S(\mathbf{0}_{n-4}, x_n y_4, x_{n-1} y_4, \dots, x_1 y_4, \mathbf{0}_4; \mathbf{z}^3; \mathbf{d}^4; \mathbf{z}^4) \& \\ &\vdots \\ &S(\mathbf{0}_1, x_n y_n, x_{n-1} y_n, \dots, x_1 y_n, \mathbf{0}_{n-1}; \mathbf{z}^{n-2}; \mathbf{d}^{n-1}; \mathbf{z}^{n-1}) \& \\ &[\neg x_1 \vee x_2 \vee \dots \vee x_n] \& [\neg y_1 \vee y_2 \vee \dots \vee y_n]. \end{aligned} \tag{2}$$

The formula  $g$  says:  $z^{n-1}$  is the product of  $x$  and  $y$  &  $x \neq 1$  &  $y \neq 1$ . The symbol  $\mathbf{0}_i$  is a shorthand for  $\underbrace{0, 0, \dots, 0}_{i \text{ times}}$ .

By substituting (1) into (2) we get an expression for the formula  $g$  which is in an almost conjunctive normal form. It can be easily converted into the conjunctive normal form; so assume that  $g$  has this property.

Let the  $2n$ -bit number  $z$  be given and  $\mathbf{z}$  be the vector of bits of its binary form. Substitute the bits  $\mathbf{z}$  for  $z^{n-1}$  into  $g$ . Call this formula  $g_{\mathbf{z}}$ . The formula contains constants 0 and 1. It may be simplified into a form without constants. If a clause contains “0”, the constant “0” can be deleted from the clause; if a clause contains “1”, the entire clause may be deleted.

Observe that

- the formula  $g_{\mathbf{z}}$  has  $(2n - 1)^2$  variables;
- the formula (1) can be written down in the conjunctive normal form with at most  $16n$  clauses, and hence the formula  $g_{\mathbf{z}}$  has at most  $16n^2$  clauses.

This will be useful in Sect 3. By construction of the formula, we get the following lemma.

**Lemma 2.** *Let  $p_1$  and  $p_2$  be two distinct primes of at most  $n$  bits each. Let  $z = p_1 p_2$  and  $\mathbf{z}$  be the binary form of  $z$ , with zeros added so that  $\mathbf{z}$  has exactly  $2n$  bits. Then the formula  $g_{\mathbf{z}}$  has exactly two satisfying assignments. Denote them  $\mathbf{v}$  and  $\mathbf{v}'$ . The following holds:*

- the  $(x_n, \dots, x_1)$ -components of  $\mathbf{v}$  are bits of  $p_1$  and the  $(y_n, \dots, y_1)$ -components of  $\mathbf{v}$  are bits of  $p_2$ ;
- the  $(x_n, \dots, x_1)$ -components of  $\mathbf{v}'$  are bits of  $p_2$  and the  $(y_n, \dots, y_1)$ -components of  $\mathbf{v}'$  are bits of  $p_1$ .  $\square$

Thus, any algorithm being able to find any satisfying assignment for  $g_{\mathbf{z}}$  will find the prime decomposition of  $z$ . In the next section, we shall prove the following theorem:

**Theorem 1.** *Any algorithm for finding a  $\mathbf{c}$ -optimal design over a finite experimental domain is able to find a satisfying argument for  $g_{\mathbf{z}}$ .  $\square$*

Hence, any algorithm for the design problem is able to break the RSA cryptographic protocol. This result may be understood either positively or negatively. The positive statement is that *finding a good algorithm for the design problem yields a good algorithm for the factoring problem*, and hence investigation in algorithms for the design problems may show results interesting for computational number theory and cryptography. The negative statement is that *the design problem is no easier than the factoring problem*; or, in other words, *there exist extremely computationally hard instances of the design problem*.

### 3 Construction for Theorem 1

The following theorem was proved by Harman and Jurík (2008).

**Theorem 2.** *Let  $\mathbf{x}_1, \dots, \mathbf{x}_K$  be the experimental domain and let  $\mathbf{c} \neq \mathbf{0}$  and  $N$  be given. Let  $\Xi$  be a matrix with columns  $\mathbf{x}_1, \dots, \mathbf{x}_K$ . Then, for any optimal solution  $(\mathbf{y}, \mathbf{z}, w)$  of the optimization problem*

$$\begin{aligned} \max w \quad \text{subject to} \quad & \Xi(\mathbf{y} - \mathbf{u}) = w\mathbf{c}, \quad \mathbf{1}^\top(\mathbf{y} + \mathbf{u}) = 1, \\ & N \cdot (\mathbf{y} + \mathbf{u}) \text{ integral}, \quad \mathbf{y} \geq \mathbf{0}, \quad \mathbf{u} \geq \mathbf{0}, \quad w \geq 0, \end{aligned} \quad (3)$$

it holds that  $\xi := \mathbf{y} + \mathbf{u}$  is the  $\mathbf{c}$ -optimal design.  $\square$

Let the formula  $g_{\mathbf{z}}$  be given and let  $C_1, \dots, C_k$  be its clauses. Let us give new names  $v_1, \dots, v_{(2n-1)^2}$  to all of the variables occurring in  $g_{\mathbf{z}}$ . Define a  $k \times (2n-1)^2$ -matrix  $\mathbf{Q}$ :

$$(\mathbf{Q})_{ij} = \begin{cases} -1, & \text{if } v_j \text{ occurs in } C_i \text{ positively,} \\ 1, & \text{if } v_j \text{ occurs in } C_i \text{ negated,} \\ 0, & \text{if } v_j \text{ does not occur in } C_i. \end{cases}$$

Now let  $\nu(C_i)$  be the number of negative literals in  $C_i$ . Define a  $k$ -component vector  $\mathbf{q}$ :

$$(\mathbf{q})_i = \frac{1 - \nu(C_i)}{32n^3}.$$

Let  $\mathbf{I}$  be a unit matrix,  $\mathbf{1}$  an all-one matrix and  $\mathbf{0}$  a zero matrix (with sizes in the subscript). Let

$$\Xi = \begin{pmatrix} \mathbf{Q} & \mathbf{0}_{k \times (2n-1)^2} & \mathbf{I}_{k \times k} & \mathbf{0}_{k \times 1} \\ \mathbf{I}_{(2n-1)^2 \times (2n-1)^2} & \mathbf{I}_{(2n-1)^2 \times (2n-1)^2} & \mathbf{0}_{(2n-1)^2 \times k} & \mathbf{0}_{(2n-1)^2 \times 1} \\ \mathbf{1}_{1 \times (2n-1)^2} & \mathbf{1}_{1 \times (2n-1)^2} & \mathbf{1}_{1 \times k} & 1 \end{pmatrix} \quad (4)$$

and

$$\mathbf{c} = \begin{pmatrix} \mathbf{q} \\ \frac{1}{32n^3} \cdot \mathbf{1}_{k \times 1} \\ 1 \end{pmatrix}. \quad (5)$$

This choice of  $\Xi$  and  $\mathbf{c}$ , by the construction by Černý and Hladík (2010), leads to the following lemma. The lemma states that  $\Xi$  and  $\mathbf{c}$  are suitable ‘encodings’ of the integer factoring problem into an instance of the design problem.

**Lemma 3.** *Let  $(\mathbf{y}, \mathbf{u}, w)$  be any optimal solution to the optimization problem*

$$\begin{aligned} \max w \quad \text{subject to} \quad & \Xi(\mathbf{y} - \mathbf{u}) = w\mathbf{c}, \quad \mathbf{1}^\top(\mathbf{y} + \mathbf{u}) = 1, \\ & 32n^3 \cdot (\mathbf{y} + \mathbf{u}) \text{ integral}, \quad \mathbf{y} \geq \mathbf{0}, \quad \mathbf{u} \geq \mathbf{0}, \quad w \geq 0. \end{aligned}$$

Then, the first  $(2n-1)^2$  components  $\xi_1, \dots, \xi_{(2n-1)^2}$  of the vector  $\mathbf{y} + \mathbf{u}$  satisfy: for  $i = 1, \dots, (2n-1)^2$ ,

- $\xi_i \in \{0, \frac{1}{32n^3}\}$ ,
- set  $v_i = \text{TRUE}$  if  $\xi_i = \frac{1}{32n^3}$ , and  $v_i = \text{FALSE}$  if  $\xi_i = 0$ . Then,  $(v_1, \dots, v_{(2n-1)^2})$  is a satisfying assignment to  $g_z$ .  $\square$

If  $A$  is any algorithm for the design problem, we run  $A$  with the input  $(\Xi, \mathbf{c}, N := 32n^3)$  where  $\Xi$  and  $\mathbf{c}$  are given by (4) and (5). By construction of  $g_z$ , there are exactly two optimal designs in this setting. Either of them, found by the algorithm  $A$ , determines a satisfying assignment for  $g_z$ , and any such assignment shows the prime decomposition  $p_1, p_2$  of  $z$ . Hence, the algorithm  $A$  can be used for breaking the RSA protocol. The proof is concluded by the observation that the construction of  $\Xi$  and  $\mathbf{c}$  given  $z$  can be done in computation time which is polynomial in  $n$ .

## 4 Conclusion

It is possible to argue that the construction produces instances of the design problem that are artificial from the statistical point of view. In complexity theory, such a situation often occurs: in the large space of all admissible instances of a problem we are searching for a complexity core, but the core instances are ‘untypical’ for the theory (e.g. statistics) that motivated the formulation of that problem. So, the major question to be further studied is: *is it possible to define, in an exact sense, what is a ‘natural’ instance of the design problem?* Having seen that there exist ‘artificial’ instances, it would be valuable to define a property capturing the substance of the problem that is relevant for statistics and ruling out unnaturalness.

Of course, further questions come to mind: is it computationally feasible to distinguish between the ‘natural’ and ‘artificial’ instances? Is this property syntactic, in the sense that it can be observed from the *form* of the vector  $\mathbf{c}$  and the experimental domain? Finally, is the restriction of the general problem onto the ‘natural’ instances polynomial-time solvable, or is it again as hard as integer factoring? Such considerations should clarify what constitutes the intrinsic hardness of the problem; whether its complexity is induced by instances that ‘a statistician would never think of’ or whether there are ‘statistically-relevant’ instances that are as hard as the problem in general.

\*\*\*

**Acknowledgement.** The research has been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project MSM6138439910.

## References

ATKINSON, A., DONEV, A. and TOBIAS, R. (2007): *Optimum Experimental Designs with SAS*. Oxford University Press, Oxford.

- ČERNÝ, M. and HLADÍK, M. (2010): Complexity of designing a  $c$ -optimal experiment over a finite experimental domain. Submitted in *Computational Optimization and Applications*. Preprint available at: <http://nb.vse.cz/~cernym/design.pdf>.
- GAREY, M. R. and JOHNSON, D. S. (1979): *Computers and Intractability. A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York.
- HARMAN, R. and JURÍK, T. (2008): Computing  $c$ -optimal experimental designs using the simplex method of linear programming. *Computational Statistics and Data Analysis* 53, 247–254.
- PAPADIMIRIOU, C. H. (1995): *Computational Complexity*. Addison-Wesley Longman.
- PÁZMAN, A. (1986): *Foundations of Optimum Experimental Design*. Reidel Publishing Company, Dordrecht.
- PUKELSHEIM, F. and RIEDER, S. (1992): Efficient rounding in approximate designs. *Biometrika* 79, 763–770.



# Estimation and Detection of Outliers and Patches in Nonlinear Time Series Models

Ping Chen

Department of Mathematics, Southeast University, Nanjing, 210096, China,  
*cp18@263.net.cn*

**Abstract.** In this paper, we propose a Gibbs sampling algorithm to detect additive isolated outliers and patches of outliers in ARMAX and bilinear time series models with Bayesian view. First, we use some methods to delete the influence of input process in ARMAX model, and then mining outliers and patches in ARMAX series based on the former work. Second, we also detect the outliers and patches in bilinear models by analogous method. It is shown that our procedure could reduce possible masking and swamping effects, which is an improvement and extension on ARMA models over the existing detection methods. At last, simulated examples show that we acquire better results.

**Keywords:** nonlinear time series, ARMAX model, bilinear models, outlier patches, Gibbs sampler

## 1 Introduction

Time series observations are often perturbed by some unusual events, such as sudden political factor, economic crises, and even typing or recording errors. Such values are usually referred to as outliers. There may be isolated outliers or patches of outliers in a time series. Outliers may have a significant impact on model identification and parameter estimation for time series. A special case of multiple outliers is a patch of additive outliers, Justel et al.(2001) proposed a procedure to detect outlier patches in an autoregressive process. Chen(1997) did a lot in the detection of additive outliers in bilinear time series. We know that the ARMAX model is more complex than ARMA model. It is widely used in engineering, finance and signal management.

In this paper, based on some different prior distributions, an adaptive Gibbs sampling algorithm is proposed for identifying additive isolated outliers and patches of outliers in nonlinear time series. First, we introduce Outliers models and identification of ARMAX series. Second, we propose Gibbs samp-

---

The research is supported by National Natural Science Foundation of China(No. 10671032).

Correspondence: CHEN Ping, Department of Mathematics, Southeast University, Nanjing 210096, China

ling methods to mine outliers and patches in the view of Bayesian. At last, some case studies show that the algorithm is effective in detecting the locations of outliers and patches and in estimating their size for the ARMAX models and bilinear models.

## 2 Outliers models and identification of ARMAX series

An ARMA model with input process is called ARMAX model, which is defined as

$$Z_t = \sum_{i=1}^d v_i(B) X_{i,t} + n_t, \quad (2.1)$$

where  $v_i(B) = (\omega_i(B)/\delta_i(B))B^{k_i}$  is the transfer function of  $i$ th input process,  $n_t = (\theta(B)/\phi(B))\varepsilon_t$  is noise process,  $\{Z_t\}$  is called response process. And  $X_{i,t}$  denotes the  $i$ th input process or the difference of  $i$ th input process at time  $t$ ,  $k_i$  presents the influence's time delay of  $i$ th input process,  $\omega_i(B)$  is the numerator factors and  $\delta_i(B)$  is the denominator factors of transfer function of  $i$ th input process,  $\{\varepsilon_t\}$  is normal white noise process,  $\phi(B)$  and  $\theta(B)$  is defined as:  $\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ ,  $\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$ , and the  $B$  is the backshift operator. When  $v_i(B) = 0$ ,  $i = 1, \dots, d$ , (2.1) is ARMA model, when some  $v_i(B)$  is nonzero constant,  $i = 1, \dots, d$ , then (2.1) is regression model with ARMA error.

The additive outliers(AO) model is as follows:

Suppose that only the  $j$ th point  $z_j$  is an AO, with influence magnitude  $\beta_{tj}$ , then we have

$$Z_t = \sum_{i=1}^d v_i(B) X_{i,t} + \beta_{tj} \delta_{t,tj} + n_t, \quad (2.2)$$

where  $\delta_{t,tj}$  is Kronecker symbol: If  $t = tj$ , then  $\delta_{t,tj} = 1$ , else  $\delta_{t,tj} = 0$ .

Considerable simplification in the identification process would occur if the input to the system were white noise. Similar to (2.1), suppose that the ARMAX model of only one input process is as follows:

$$Z_t = \delta^{-1}(B)\omega(B)X_{t-m} + n_t = v(B)X_t + n_t, \quad (2.3)$$

where  $\delta(B) = 1 - \delta_1 B - \dots - \delta_{r1} B^{r1}$ ,  $\omega(B) = \omega_0 - \omega_1 B - \dots - \omega_{r2} B^{r2}$  and  $v(B) = \delta^{-1}(B)\omega(B)B^m$ . Suppose the input process  $X_t$  is stationary and is able to be represented by some member of the general linear class of autoregressive-moving average models. Then, given a set of data, we can carry out our usual identification and estimation methods to obtain a model for the  $X_t$  process  $\phi(B)\theta^{-1}(B)X_t = \xi_t$ , which, to a close approximation, transforms the correlated input series  $X_t$  to the uncorrelated white noise series  $\xi_t$ . At the same time, we can obtain an estimate  $s_\xi^2$  of  $\sigma_\xi^2$  from the



sum of squares of the  $\hat{\xi}'$ s. If we now apply this same transformation to  $Z_t$  to obtain  $\eta_t = \phi(B)\theta^{-1}(B)Z_t$ , and let  $\varepsilon_t = \phi(B)\theta^{-1}(B)n_t$ , then the model (2.3) may be written  $\eta_t = v(B)\xi_t + \varepsilon_t$ . Multiplying  $\xi_{t-k}$  on both sides and taking expectations, we obtain  $\gamma_{\xi\eta}(k) = v_k\sigma_\xi^2$ , where  $\gamma_{\xi\eta}(k) = E[\xi_{t-k}\eta_t]$  is the cross covariance at lag  $k$  between  $\xi$  and  $\eta$ . Thus  $v_k = [\rho_{\xi\eta}(k)\sigma_\eta]/[\sigma_\xi]$ ,  $k = 0, 1, 2, \dots$

Hence, after 'prewhitening' the input, the cross correlation function between the prewhitened input and correspondingly transformed output is directly proportional to the response function. In practice, we do not know the theoretical function  $\rho_{\xi\eta}(k)$ , so we must substitute estimates in  $v_k$  to give  $\hat{v}_k = [r_{\xi\eta}(k)s_\eta]/[s_\xi]$ ,  $k = 0, 1, 2, \dots$ , where  $r_{\xi\eta}(k) = c_{\xi\eta}(k)/[s_\xi s_\eta]$ ,  $c_{\xi\eta}(k) = \frac{1}{n} \sum_{t=1}^{n-k} (\xi_t - \bar{\xi})(\eta_{t+k} - \bar{\eta})$  and  $s_\xi = \sqrt{c_{\xi\xi}(0)}$ ,  $s_\eta = \sqrt{c_{\eta\eta}(0)}$ ,  $k = 0, 1, 2, \dots$

The preliminary estimates  $\hat{v}_k$  can provide a rough basis for selecting suitable transfer function model. First, we may use the estimates  $\hat{v}_k$  so obtained to make guesses of the order  $r1$  and  $r2$  of  $\delta(B)$  and  $\omega(B)$ , and of the delay parameter  $m$ . Second, we do not consider the noise  $n_t$  now, substituting  $Z_t = \hat{v}(B)X_t$  in the equation  $\delta(B)Z_t = \omega(B)B^m X_t$ , based on equating coefficients of  $B$ , to obtain initial estimates of the parameters  $\delta(B)$  and  $\omega(B)$  in (2.3).

### 3 Outliers detection for ARMAX model via standard Gibbs sampling

We detect AO type outliers in ARMAX model by Gibbs sampling based on Bayesian method. The idea is as follows: Suppose the probability that observation is outlier has some prior information. Based on the method of conjugate priors, we proved some theorems which gives the expressions of some posterior distributions, then we compute the posterior probabilities for each data point to be an AO type outlier using techniques of Gibbs sampling. If the posterior probability is larger than some prescribed value, then we consider it as an AO type outlier.

Since output  $z_t$  may be an outlier at each time point, we let  $\delta_t = 1$  if the observation at this time point is an additive outlier. Let  $\delta_t = 0$  if it is an outlier-free time point, and denote  $P(\delta_t = 1) = \alpha$ . Then the general ARMA model with additive outlier is as follows

$$\begin{cases} y_t = \sum_{i=1}^p \phi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j} \\ z_t = y_t + \delta_t \beta_t, \end{cases} \quad \varepsilon_t \sim N(0, \sigma^2). \quad (3.1)$$

It means that the observation  $z_t$  may be AO with probability  $\alpha$ , its magnitude is  $\beta_t$  at time  $t$ . For simplicity, assume that  $y_1, \dots, y_p$  are fixed and  $z_t = y_t$  for  $t = 1, \dots, p$ , i.e. there exist no outliers in the first  $p$  observations. The indicator vector of outliers then becomes  $\delta = (\delta_{p+1}, \delta_{p+2}, \dots, \delta_n)'$  and the size vector is  $\beta = (\beta_{p+1}, \beta_{p+2}, \dots, \beta_n)'$ . Let  $\hat{\varepsilon}_t$  denote the residual estimation

of model (3.1) without AO. And let  $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{n-1})'$ ,  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_{n-1})'$ ,  $\Theta = (\phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots, \theta_q)'$ ,  $z = (z_1, z_2, \dots, z_n)'$  and  $\Phi_{t-1} = (y_{t-1}, y_{t-2}, \dots, y_{t-p}, \hat{\varepsilon}_{t-1}, \dots, \hat{\varepsilon}_{t-q})'$ . By  $y_t = \Theta' \Phi_{t-1} + \varepsilon_t$ , we can obtain the likelihood function

$$L(\Theta, \sigma^2, \delta, \beta, \alpha \mid z, \hat{\varepsilon}) \propto \sigma^{n-p} \exp\left\{-\frac{1}{2\sigma^2} \sum_{t=p+1}^n (y_t - \Theta' \Phi_{t-1})^2\right\}, \quad (3.2)$$

where  $y_t = z_t - \delta_t \beta_t$ .

For computational reason, we use conjugate prior distribution for parameters  $\Theta$  and  $\sigma^2$ , which distributed as multidimensional uniform distribution on  $[0, 1]$  region and inverted-Gamma distribution  $IG(\frac{\nu}{2}, \frac{\nu\lambda}{2})$  respectively. Assume that the outlier indicator  $\delta_t$  and the outlier magnitude  $\beta_t$  are independent and distributed as *Bernoulli*( $\alpha$ ) and  $N(0, \tau^2)$  respectively for all  $t$ . Then, the prior probability of being contaminated by an outlier is the same for all observations, namely  $P(\delta_t = 1) = \alpha$ , for  $t = p+1, \dots, n$ . The prior distribution of the contamination parameter  $\alpha$  is *Beta*( $\gamma_1, \gamma_2$ ), and  $(\beta_t)'$ s are *i.i.d.* for all  $t$ . The hyperparameters in our model are  $\lambda, \nu, \gamma_1, \gamma_2$  and  $\tau^2$ , all of which are assumed known.

It is obvious that we must obtain the conditional posterior distributions of parameters  $(\Theta, \beta, \delta, \alpha, \sigma^2)$  for detecting outliers in the model. The important thing is to conduct the conditional posterior distributions of  $\delta_j = 1$  and  $\beta_j$ . Let  $z = (z_1, z_2, \dots, z_n)$  denote the observed vector of model (3.1). Using the standard Bayesian method, under the above conditions, we have the following results:

**Theorem** (1) For the conditional posterior distribution of  $\delta_j = 1$ , we have

$$p(\delta_j = 1 \mid z, \delta_{(-j)}, \beta, \Theta, \sigma^2, \alpha) = \left[1 + \frac{1 - \alpha}{\alpha} B_{10}(j)\right]^{-1}$$

where  $\delta_{(-j)}$  is obtained from  $\delta$  by eliminating the element  $\delta_j$ ,  $T_j = \min\{n, j + p\}$ , and

$$B_{10}(j) = \exp\left\{\frac{1}{2\sigma^2} \left[\sum_{t=j}^{T_j} a_t^2(1) - \sum_{t=j}^{T_j} a_t^2(0)\right]\right\}$$

where  $a_t(1) = (y_t - \Theta' \Phi_{t-1})_{\delta_j=1}$ ,  $a_t(0) = (y_t - \Theta' \Phi_{t-1})_{\delta_j=0}$  and  $a_t(0) = a_t(1) + \varphi_{t-j} \beta_j$ . When  $p = 0$ , then  $\varphi_i = 0$  for all  $i$ ; when  $p > 0$ , we have that  $\varphi_i = -1$  if  $i = 0$ ,  $\varphi_i = \phi_i$  if  $i = 1, \dots, p$ , and  $\varphi_i = 0$  if  $i \geq p+1$ .

(2) When  $\delta_j = 0$ , there is not any new information about the posterior distribution of  $\beta_j$ , namely,  $\beta_j$  distributed as  $N(0, \tau^2)$ . When  $\delta_j = 1$ , since  $Z_t$  contains information of  $\beta_j$ , so we have that

$$p(\beta_j \mid z, \delta, \beta_{(-j)}, \Theta, \sigma^2, \alpha) \sim N(\beta_j^*, \tau_j^*),$$

where  $\beta_{(-j)}$  is obtained from  $\beta$  by eliminating the element  $\beta_j$ , and  $\beta_j^* =$

$$\frac{\mathbb{A} \tau^2}{\mathbb{B} \tau^2 + \sigma^2}, \quad \tau_j^* = \frac{\sigma^2 \tau^2}{\mathbb{B} \tau^2 + \sigma^2}, \quad \text{where } \mathbb{A} = -\sum_{t=j}^{T_j} a_t(1) \varphi_{t-j}, \text{ and } \mathbb{B} = \sum_{t=j}^{T_j} \varphi_{t-j}^2. \quad \square$$

We give the Gibbs method for detecting AO in ARMAX model as follows:  
**a.** Given the starting point  $(\lambda, \nu, \gamma_1, \gamma_2, \tau^2, \alpha)$ , whereafter this algorithm iterates the following loop: **b.** sample  $\Theta^{(t)}$  from  $p(\Theta|z, \delta^{(t-1)}, \beta^{(t-1)}, \alpha^{(t-1)}, \sigma^{2(t-1)})$ ,  
**c.** sample  $\sigma^{2(t)}$  from  $p(\sigma^2|z, \Theta^{(t)}, \delta^{(t-1)}, \beta^{(t-1)}, \alpha^{(t-1)})$ , **d.** sample  $\delta_j^{(t)}$  from  $p(\delta_j|z, \Theta^{(t)}, \sigma^{2(t)}, \beta^{(t-1)}, \alpha^{(t-1)})$ , **e.** sample  $\beta_j^{(t)}$  from  $p(\beta_j|z, \Theta^{(t)}, \sigma^{2(t)}, \delta_j^{(t)}, \alpha^{(t-1)})$ ,  
**f.** sample  $\alpha^{(t)}$  from  $Beta(\gamma_1 + k, \gamma_2 + n - p - k)$ , repeat the above steps till it is convergence. From the Bayesian principle, we suppose that if the outlying posterior probability is larger then  $c_1$ , then believe it is an outlier.

#### 4 Detection of outlier patches via adaptive Gibbs sampling

Similar to Justel et al.(2001), our procedure also consists of two Gibbs runs. In the first run, the standard Gibbs sampling based on the results of Section 3 is carried out. The results of this Gibbs run are then used to implement a second Gibbs sampling that is adaptive in treating identified outliers and in using block interpolation to reduce possible masking and swamping effects. Let  $\hat{\Theta}^{(s)}$ ,  $\hat{\sigma}^{(s)}$ ,  $\hat{\beta}^{(s)}$  and  $\hat{\alpha}^{(s)}$  be the posterior means of  $\Theta$ ,  $\sigma^2$ ,  $\beta$  and  $\alpha$  respectively based on the  $s$  iterations of the first Gibbs run. First, we select a appropriate critical value  $c_1$  to identify potential outliers. An observation  $z_j$  is identified as an outlier if the posterior probability  $\hat{p}_j^{(s)} > c_1$ . Let  $\{t_1, \dots, t_m\}$  be the collection of time indexes of outliers identified by the first Gibbs run. Second, let  $c_2, c_2 \leq c_1$  be another appropriate critical value to specify the beginning and end points of a potential outlier patch. We select a window of length  $2p$  around the identified outlier to search for the boundary points of a possible outlier patch by a forward-backward method. For example, consider an identified outlier  $z_{t_i}$ . For the  $p$  observations before  $z_{t_i}$ , if their posterior probabilities  $\hat{p}_j^{(s)} > c_2$ , then these points are regarded as possible outlier patch associated with  $z_{t_i}$ . We then select the farthest point from  $z_{t_i}$  as the begining point of the outlier patch. Denote the point by  $z_{t_i-k_i}$ . Then we do the same for the  $p$  observations after  $z_{t_i}$  and select the farthest point from  $z_{t_i}$  with  $\hat{p}_j^{(s)} > c_2$  as the end point of the outlier patch. Denote the end point by  $z_{t_i+v_i}$ . Combine the two blocks to form a possible outlier patch associated with  $z_{t_i}$ , which denoted by  $(z_{t_i-k_i}, \dots, z_{t_i+v_i})$ . Consecutive or overlapping patches should be merged to form a larger patch. Lastly, draw Gibbs samples jointly within a patch. Suppose that a patch of  $k$  outliers starting at time index  $j$  is specified. Denote the vectors of outlier indicators and magnitudes by  $\delta_{j,k} = (\delta_j, \dots, \delta_{j+k-1})'$  and  $\beta_{j,k} = (\beta_j, \dots, \beta_{j+k-1})'$ , respectively, associated with the patch. Similar to the Theorem 1 of Justel et al.(2001), we may obtain the conditional posterior distributions of  $\delta_{j,k}$  and  $\beta_{j,k}$ .

For the second adaptive Gibbs sampling, we use the results of the first Gibbs run to start the second Gibbs sampling and to specify prior distri-

butions of the parameters. For each outlier patch, we use the conditional posterior distributions to draw  $\delta_{j,k}$  and  $\beta_{j,k}$  in the second Gibbs sampling, which is also run for  $s$  iterations. The starting values of  $\delta_t$  are as follows:  $\delta_t^{(0)} = 1$  if  $\hat{p}_t^{(s)} > 0.3$ , otherwise,  $\delta_t^{(0)} = 0$ . Then the prior distributions of  $\beta_t$  are as follows.

(a) If  $z_t$  is identified as an isolated outlier, then the prior distribution of  $\beta_t$  is  $N(\hat{\beta}_t^{(s)}, \tau^2)$ , where  $\hat{\beta}_t^{(s)}$  is the Gibbs estimate of  $\beta_t$  from the first Gibbs run.

(b) If  $z_t$  belongs to an outlier patch, then the prior distribution of  $\beta_t$  is  $N(\tilde{\beta}_t^{(s)}, \tau^2)$ , where  $\tilde{\beta}_t^{(s)}$  is the conditional posterior mean as follows:

$$\tilde{\beta}_{j,k} = (D_{j,k} \sum_{t=j}^{T_{j,k}} \Pi_{t-j} \Pi'_{t-j} D_{j,k})^{-1} (- \sum_{t=j}^{T_{j,k}} e_t(0) D_{j,k} \Pi_{t-j})$$

(c) If  $z_t$  does not belong to any outlier patch, and is not an isolated outlier, then the prior distribution of  $\beta_t$  is  $N(0, \tau^2)$ .

## 5 Simulated examples and conclusions

**Example A.** In the simulations, we consider the ARMAX(1,1,2) model:

$$\begin{cases} (1 - 0.78B + 0.3B^2)x_t = e_t \\ (1 - 0.7B)y_t = (1 - 0.7B)(1 - 0.48B)x_t + (1 - 0.27B + 0.96B^2)\varepsilon_t \\ z_t = y_t - 11\delta_{t,31} + 10\delta_{t,32} - 9\delta_{t,33} + 10\delta_{t,34} - 9\delta_{t,35} + 10\delta_{t,50}, \end{cases} \quad (5.1)$$

where  $\{e_t\}$  and  $\{\varepsilon_t\}$  are all normal white noise, their means are zero and variance  $\sigma^2 = 1$ .

We create 101 observations  $x_0, x_1, \dots, x_{100}$  of  $x_t$  and 100 observations  $z_1, z_2, \dots, z_{100}$  of  $z_t$  by simulation. It is obvious that the input process is an AR(2) series, the transfer function of the ARMAX model is  $(1 - 0.48B)$ , a patch of five consecutive additive outliers have been introduced from  $t = 31$  to  $t = 35$ , a single AO has been add at  $t = 50$ , and the outlier magnitudes are  $\beta_{31} = -11$ ,  $\beta_{32} = 10$ ,  $\beta_{33} = -9$ ,  $\beta_{34} = 10$ ,  $\beta_{35} = -9$  and  $\beta_{50} = 10$  respectively. Applying our method to the above simulate series  $\{z_t\}$  and prewhitening the input series. Making  $\{x_t\}$  follows an ARMA model:  $(1 - 0.41117B)x_t = \varepsilon_t$ . Then we take the same manipulation of prewhitening the  $\{z_t\}$ . By analyzing filtered cross correlation coefficient of  $\{z_t\}$  and  $\{x_t\}$ , we obtain the transfer function  $1.00597 - 0.1568B$  for  $\{x_t\}$ . Note that the transfer function was influenced by outliers. In order to delete the influence of input process  $\{x_t\}$  in response process  $\{z_t\}$ , we let  $z_t^* = z_t - (1.00597 - 0.1568B)x_t$ . Then we can apply the method described above to detect the outliers in  $\{z_t^*\}$ , which are just the outliers of  $\{z_t\}$ .

Let  $\gamma_1 = 5$ ,  $\gamma_2 = 95$ ,  $\nu = 3$ ,  $\lambda = \tilde{\sigma}^2/3$ ,  $\alpha = 0.5$ ,  $c_1 = 0.5$ ,  $c_2 = 0.3$  and  $\tau = 3\tilde{\sigma}$ , where  $\tilde{\sigma}^2$  is mean square error of  $\{z_t\}$ . Here  $\gamma_1 = 5$  means that we

believe the prior probability of each point is an outlier approximate to 0.05. First, we detect the outliers in  $\{z_t\}$  using the standard Gibbs sampling. Limit to the computational ability, we take 100 iterations by usual Gibbs algorithm. We obtain the posterior probabilities that each observation is outlier, the Figure shows that the posterior probabilities of being an outlier for data points at  $t = 31, 33, 34, 35, 36, 50$  are large. Meanwhile, the outlying posterior probabilities of other observations are low. We see that the algorithm fails to detect the inner point at  $t = 32$  of the patch, resulting in the masking effect. On the other hand, the algorithm misspecifies the 'good' data point at  $t = 36$  as outlier because the outlying posterior probability of this point is larger, so the 'good' data point at  $t = 36$  is swamped by the patch of outliers. Second, in order to avoid the presence of masking and swamping problem, we use the method given by section 4, and take 900 iterations by the adaptive Gibbs algorithm. The Figure shows the posterior probability of outlier for each data point, which clearly shows that the patch of outliers and the isolated outlier in  $\{z_t\}$  process are detected triumphantly, and there is not any misjudgement. The posterior means of the sizes of these outliers are  $\hat{\beta}_{31} = -5.9849$ ,  $\hat{\beta}_{32} = 14.4163$ ,  $\hat{\beta}_{33} = -16.5910$ ,  $\hat{\beta}_{34} = 14.4383$ ,  $\hat{\beta}_{35} = -14.6142$  and  $\hat{\beta}_{50} = 3.8853$ , respectively.

Similar to the method above, we also could detect the outliers and patches in bilinear time series models by adaptive Gibbs sampling algorithm. We omit the theoretics and give an example as follows

**Example B.** In the following example, we consider the BL(1,1,1,1) model:

$$\begin{cases} y_t = 0.7y_{t-1} + \varepsilon_t - 0.3\varepsilon_{t-1} + 0.31y_{t-1}\varepsilon_{t-1} \\ z_t = y_t + 4\delta_{t,40} - 3\delta_{t,41} + 3\delta_{t,42}, \end{cases} \quad (5.2)$$

where  $\{\varepsilon_t\}$  is standard normal white noise.

We create a set of observations of the bilinear model (5.2) by simulation, where a patch of three consecutive additive outliers have been introduced from  $t = 40$  to  $t = 42$ , and the outlier magnitudes are  $\beta_{40} = 4$ ,  $\beta_{41} = -3$  and  $\beta_{42} = 3$  respectively. The figure shows that the curve of observations has large volatility, it would be very difficult to distinguish between 'outliers' and normal points of nonlinear model.

First, we detect the outliers in  $\{z_t\}$  using the standard Gibbs sampling. We obtain the posterior probabilities that each observation is outlier, the Figure obviously shows that the posterior probabilities of being outlier only at  $t = 40$  is large than 0.5. Meanwhile, the outlying posterior probabilities of other observations are low, the posterior probability of being an outlier at  $t = 41$  is even small than 0.2. However, the posterior probability of being an outlier at  $t = 57$  is larger than at  $t = 41, 42$ . We see that the standard Gibbs sampling fails to detect the inner and border points at  $t = 41, 42$ . resulting in the masking effect. On the other hand, the algorithm is apt to misspecify the 'good' data point at  $t = 57$  as outlier because the outlying posterior probability of this point is larger than every points but the point at  $t = 40$ ,

so that the 'good' data point at  $t = 57$  may be swamped by the patch of outliers.

Second, in order to avoid the presence of masking and swamping problem, we apply the similar method given by section 4, and take 900 iterations by the adaptive Gibbs algorithm. In the process of running, we let the initial distribution of  $\Theta$  be  $N(\mathbf{0}, 0.3\mathbf{I})$ . The Gibbs sampler was repeated several times with different hyper-parameters and different numbers of iterations to reanalyze the data. The results show that the locations of possible outliers and patch are stable, even though the estimated outlying probabilities may vary slightly between the Gibbs samples. The figure gives the time plot of the estimated posterior probability that each point is an outlier via adaptive Gibbs sampling, the window width of search was 4.

It obviously shows that the posterior probability of being an outlier obtained by adaptive Gibbs algorithm for data points at  $t = 40, 41, 42$  are large. Meanwhile, the outlying posterior probabilities of other observations are low. Actually, if select the critical value  $c_2 = 0.3$ , then we could identify the patch. On the other hand, many normal points which were similar to outliers do not be misspecified as outliers because the outlying posterior probabilities of these points are smaller than 0.3, which show that the adaptive Gibbs sampling is effective in mining the additive outlier patch of bilinear time series.

Furthermore, the posterior means of the sizes of these outliers are  $\hat{\beta}_{40} = 4.1652$ ,  $\hat{\beta}_{41} = -5.7693$  and  $\hat{\beta}_{42} = 3.4172$ , respectively. By a number of simulations which detect the patches in bilinear model by adaptive Gibbs sampling, we discovered that the critical value  $c_2$  should be selected smaller than ARMA model. Investigate its reason, it may be the volatility of bilinear series is larger than ARMA series, and itself could often produce some normal points which appear to be outliers.

## References

- BOX, G.E.P., JENKINS, G.M. and REINSEL, G.C.(1994): *Time Series Analysis: Forecasting and Control, Third edition*. Prentice-Hall & Englewood Cliffs, NJ.
- CHEN, C. W. S.(1997):Detection of additive outliers in bilinear time series. *Comput. Statist. Data Anal.* 24, 283-294.
- CHEN, P., YAN, F.R., WU, Y.Y. and CHEN, Y. (2009): Detection of outliers in ARMAX time series models. *Advances in Systems Science and Applications* 9, 97-103.
- JUSTEL, A., PEÑA, D. and TSAY, R.S. (2001): Detection of outlier patches in autoregressive time series. *Statistica Sinica* 11, 651-673.
- MCCULLOCH, R.E., TSAY, R.S.(1994): Bayesian analysis of autoregressive time series via the Gibbs sampler. *Journal of Time Series Analysis* 15, 235-250.

# Two-way Classification of a Table with non-negative entries: Validation of an Approach based on Correspondence Analysis and Information Criteria

Antonio Ciampi<sup>1</sup>, Alina Dyachenko<sup>1</sup>, and Yves Lechevallier<sup>2</sup>

<sup>1</sup> Department of Epidemiology, Biostatistics and Occupational Health  
McGill University, Montreal, Qc., Canada

<sup>2</sup> INRIA, Paris-Rocquencourt  
78153 Le Chesnay cedex, France *Yves.Lechevallier@inria.fr*

**Abstract.** We present a validation of a rule to choose the number of dimension in Correspondence Analysis and of an AIC/BIC based selection approach to block clustering. An example of micro-array analysis is also shown.

**Keywords:** 2-way clustering, dimension reduction, number of clusters, microarrays

## 1 Introduction

Exploratory techniques based on 2-way clustering and the heat map, are now used in a broad variety of fields. One major example is the analysis of gene expression data. In this case, data are in the form of a rectangular table, in which entry  $n_{ij}$  is a number that measures the expression intensity of a *gene*  $i$  in a particular *tissue*  $j$ . In a now historical paper, Eisen et al. (1998) detected, using 2-way clustering and the heat map, clearly distinct blocks of genes and tissues with characteristics expression patterns. These blocks could be given a clinical interpretation, which helped develop an effective prognostic classification of tumor subtypes. Another example is mining of web data: see Charrad et al.(2009), who develop an approach based on block clustering (Govaert (1984)), and Ciampi et al (2009) who use hierarchical clustering.

In this paper we develop an earlier proposal (Ciampi et al. (2005), Ciampi et al. (2009)) by focusing on the choice of dimensionality of the representation space, and on the choice of the number of rows and columns defining a partition of the data table. Our general framework for 2-way clustering of tables with non-negative entries is reviewed in Section 2. In Section 3 we highlight the basic choices mentioned above and propose statistically motivated approaches for guiding these choices. In Section 4 the proposals are evaluated by simulations. Section 5 is devoted to an example of analysis of

a microarray gene expression data set. Section 6 concludes the paper with a discussion and an outline of current and future research.

## 2 A general approach to 2-way clustering

We assume that the reader is familiar with Correspondence Analysis (CA), the  $\chi^2$  distance and its properties, the notion of inertia, the scree plot, and the basic techniques for clustering rows and columns of a contingency tables based on the  $\chi^2$  distance (Greenacre (2007)). Our approach consists of three steps:

1. Given a data table  $T$ , perform a CA and select  $k \ll m$ , so that the inertia contained in the subspace  $E^k$  spanned by the first  $k$  eigenvectors is a large proportion of the total inertia. This is our representation space for the rows and the columns of  $T$ .
2. Calculate the distances between the points that represent rows (columns) in  $E^k$  and apply a hierarchical clustering algorithm to the points of the two clouds. Then use a non-unique order induced by the dendrograms to rearrange rows and columns of the table.
3. Cut the dendrograms, obtaining a partition of the row cloud into  $p_r$  classes and a partition of the column cloud into  $p_c$  classes, so as to keep essential information.

The selection of  $k$ ,  $p_r$  and  $p_c$  is described in the next section. The choice of the hierarchical clustering algorithm is open; in this paper we have used the Ward algorithm. Our choice, easily replaceable, is based on pragmatic considerations:

- i) it is frequently used in approaches that, like ours, apply clustering to a Euclidean representation of the data obtained by preliminary scaling;
- ii) it tends to produce regularly shaped clusters (in Euclidean geometry);
- iii) it works well in simulations with data from mixtures of multivariate normal distributions.

## 3 Choosing dimensionality and cutting the row- and column-dendrograms

In order to choose the dimensionality  $k$ , we develop further the approach described in Ciampi et al. (2005). A summary description of our new proposal is as follows:

1. From the data table  $T$ , randomly generate an artificial data table  $T^*$  with the same marginal totals as  $T$  but with independent rows and columns. Repeat this construction  $M$  times, with  $M$  large, thus obtaining a family  $\{T^{*(m)}, m = 1, \dots, M\}$ .



2. Draw on the same graph  $S$ , the scree plot of  $T$ , and the envelope of the scree plots of  $\{T^{*(m)}, m = 1, \dots, M\}$ . This defines a band, to be compared with the single scree plot of  $T$ .
3. Determine the first point which falls below the simulation band. Choose as the dimension of the representation space as the abscissa (number of axes) of the point immediately preceding this point.

To choose  $p_r$  and  $p_c$ , i.e. where to cut the dendrograms of rows and columns, we associate a statistical model to any specific clustering of rows and columns of a 2-way table. Then, information criteria such as the Aikake Information Criterion (AIC), or the Bayesian Information criterion (BIC), become a natural tool for comparing classifications and for choosing a particular one, out of a family of candidates. Several variants are possible, including but not limited to the following possibilities:

- a) Develop a 2-way hierarchical clustering algorithm, by separately clustering rows and columns; then choosing to cut each dendrograms at the level corresponding to the minimum AIC (BIC);
- b) Obtain a hierarchical classification of rows and columns as in a); but now, for each pair of levels of the two dendrograms, calculate the AIC (BIC) of the corresponding cross-classification, and then choose the cross-classification with the minimum AIC (BIC);

## 4 Evaluation of the approach

The approach presented in this work rests on two key ideas: a) a preliminary reduction of dimension is obtained, if necessary, by choosing the number of axes in a Correspondence Analysis (CA) of the original data; b) columns and rows are clustered using a hierarchical clustering algorithm; then, if one wishes to choose an optimal clustering among the resulting cross-classifications, the minimum AIC or BIC criterion is used for the selection. We have carried out two simulation experiments to study these choices.

### 4.1 Choice of number of axes $k$

Our goal is to show that by the simple graphical approach described above we can actually retrieve the correct number of dimension.

To simulate matrices with a chosen number of underlying dimensions, we have started by generating one  $100 \times 10$  contingency table of total sum 10,000, under the hypothesis of independence of rows and columns. This initial matrix was used to simulate other matrices with lower underlying dimension as follows: by the Singular Value Decomposition (SVD) applied to the matrix of standardized residuals we extracted two matrices, we extracted Singular Values (SV) and two matrices of dimension  $10 \times 10$  and  $100 \times 10$  respectively: the matrices of left and right singular vectors. These results were

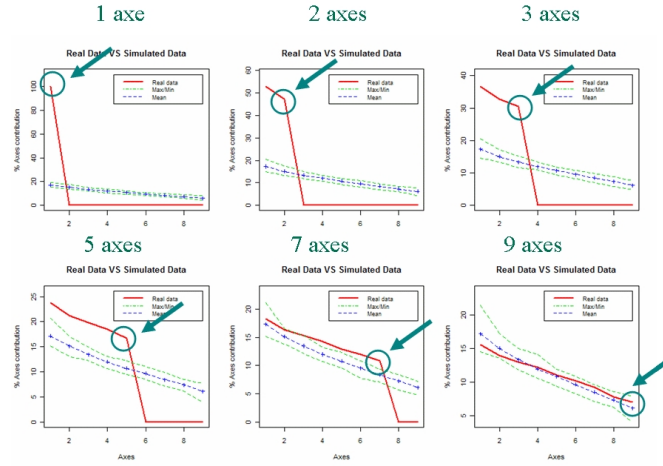


Fig. 1. Simulations: choice of number of axes

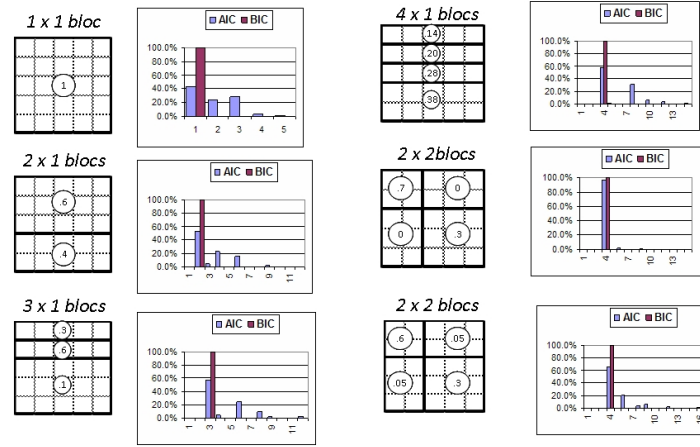
used to build  $10 - 1 = 9$  contingency tables with underlying dimension ranging from 1 to 9. The  $i^{th}$  matrix,  $i = 1, \dots, 9$ , was constructed by reconstituting the SVD but setting all eigenvalues from  $i+1$  to 9 equal to zero. A selected subset of the simulation results of is represented in Figure 1.

In all cases shown above (and others not shown) the simple rule of thumb is confirmed: the number of axes corresponds to the point in the curve preceding the first point which falls below the simulation band. Notice that in the last graph ( $k = 9$ ) the curve is entirely comprised within the simulation band, suggesting that no data reduction is possible and we have to use all the dimensions.

#### 4.2 Choice of number of clusters by AIC/BIC

Our goal is to study the behaviour of the AIC and BIC in choosing the correct block structure in a simulation experiment. We assume that the contingency table is generated by a multinomial distribution corresponding to a block structure; for each clustering of rows and columns we calculate the likelihood ratio statistics of each model with respect to the saturated model, and correct this likelihood ratio by adding the difference of the number of free parameters multiplied by 2 for the AIC and by the logarithm of the total size for the BIC.

We have simulated contingency tables having 5 rows and five columns, of total sum 10,000, according to several multinomial models. Each contingency table is defined by a number of blocks (1 to 4) with a pre-defined probability for a unit to belong to each block; within each block, all cells have the same probability. Each table has been generated 1000 times. A summary of our result is given in Figure 2 above. Each panel of the figure shows the model



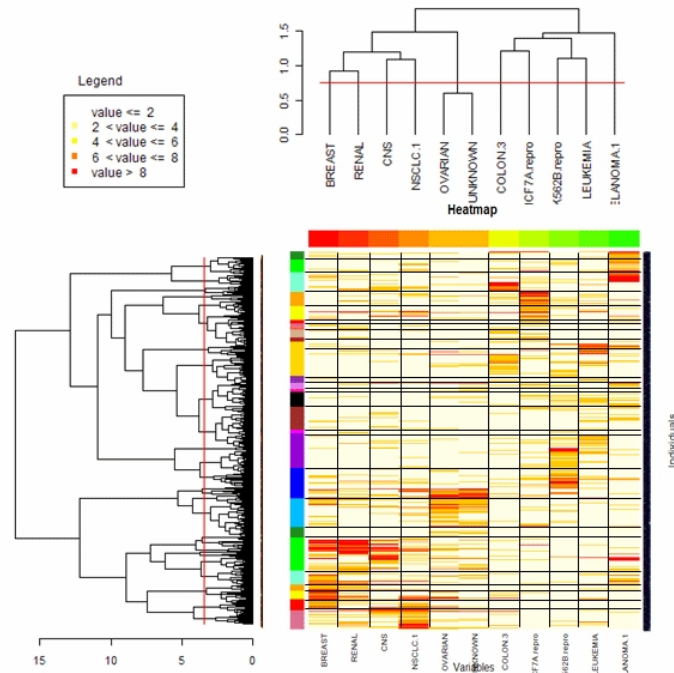
**Fig. 2.** Simulation: Block models and barplots of number of clusters by minimum AIC/BIC selection

and the bar plot of the number of clusters identified by the minimum AIC and BIC rules. For example, the top leftmost panel represents a trivial model with only one homogeneous block, i.e. each cell has a probability of  $1/25$ ; as it can be seen, 100% of the 1000 simulated samples were recognized by the BIC as being constituted of 1 block; in contrast, the AIC results tend to vary, though in the relative majority of the cases (42.3%) the correct structure is retrieved. As another example, the bottom rightmost panel represents a situation with 4 homogeneous blocks of varying probabilities (0.05 for the off-diagonal block and 0.6 and 0.3 for the diagonal blocks). Again, the BIC retrieves the correct model in all cases and the AIC in the majority of cases (65.0%), with some worrisome cases in which the number of blocks is very high (up to 16).

## 5 Example: gene expression NCI data

The publicly available NCI data presented here are a classic example of micro-array analysis. See Hastie et al. (2009) for details. The original file contains expression measures of 6830 genes in 64 cancerous tissues. To simplify the problem and reduce the amount of calculations, we have selected 2000 genes at random and have chosen 11 distinct malignancies out of the original data set; thus we work with a  $2000 \times 11$  data matrix. The malignancies are: CNS, RENAL, BREAST, NSCLC, UNKNOWN, OVARIAN, LEUKEMIA, K562B.repro, COLON, MELANOMA, MCF7A.repro. Our task is to identify distinct blocks of genes with distinct profiles for the tumor tissues.

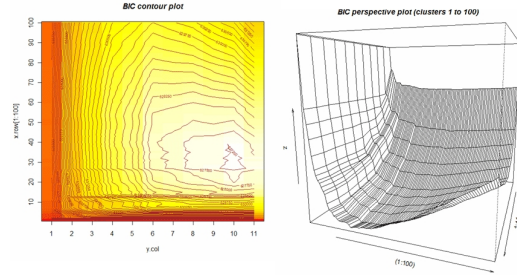
As first step, we pre-processed the data to obtain non-negative entries (the original data are in the logarithmic scale). Then we applied CA and ap-



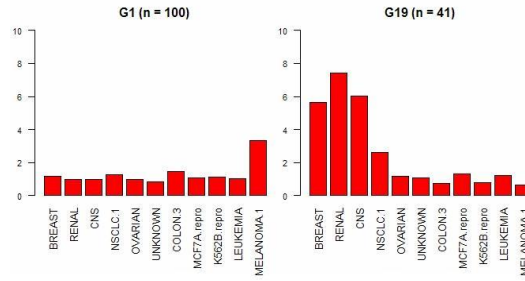
**Fig. 3.** Dendrogram and Heatmap, 2000 rows  $\times$  11columns

plied the Ward hierarchical clustering algorithm to both rows and columns of the data matrix, using the Euclidean representation in 11-dimensional space defined by CA. Figure 3 shows the heat map of the data and the row- and column-dendrograms obtained from the clustering. In the figure, both row- and column-dendrograms are cut according the red lines. These lines were obtained by inspecting the  $2000 \times 11$  table containing the BIC of every model corresponding to a possible cut of the row-dendrogram and a possible cut of the column-dendrogram. We chose the cuts corresponding to the minimum value in this table. The block structure is indicated by horizontal and vertical lines as well as by distinct colors on the margins. Figure 4 shows graphically the contour and perspective plots of the BIC table. We also obtained (data not shown) a similar table and plots for the AIC; as one might expect from general experience, as well as from the simulations presented above, the AIC tends to find more clusters than the BIC, and perhaps more clusters than the actual number.

The minimum BIC rule identifies 27 groups of genes and 10 groups of tissues (OVARIAN and UNKNOWN tissues merge according to this clustering). We remark here that if we cluster separately rows and columns, we obtain also 27 groups of genes but only a unique group of tissues, a result which is



**Fig. 4.** BIC table: contour plot and perspective plot



**Fig. 5.** Cluster descriptions: selected gene classes profiles

not helpful and unintuitive: the overall picture of the BIC given in Figure 4 helps exclude this solution, since it is represented on the steep portion of the surface. Figure 5 contains two of the 27 cluster profiles (selected for reason of space) corresponding to gene groups in the cancer tissues. Substantive validation of these results was not attempted here, as we were working with public data and had no access to experts.

## 6 Discussion and future work

We have presented a new step in our long-term methodological development aiming to produce interpretable two-way classifications of data tables with non-negative entries. The progress achieved here is twofold: a) a validation of the data reduction proposed in Ciampi et al. (2005), namely the choice of the dimension of the factor space in which rows and columns are represented; b) an AIC/BIC based approach to choosing how to cut the dendrograms of rows and columns. We stress here the utility of the latter and in particular of the contour and perspective plots, which provide powerful graphical tools for exploring various possible choices of block clustering. The example highlights a few important points of the methodology. First, if our purpose is a true two-way classification, i.e. if we seek to find rectangular blocks of the data matrix, then it is potentially misleading to only look at the row- and

column-dendrograms separately. Second, while the AIC and the BIC yield a reasonable choice of an optimal classification from the hierarchical classification, this choice should not be seen as unique: by means of the contour and perspective plots, we can actually identify a region within which the resulting models can be seen as statistically very close, or even equivalent. This is in keeping with our aim to provide exploratory tools rather than unique classifications. On the other hand, the identification of a region of useful classifications, can serve as a guide and a preliminary restriction of the search space, should we wish to model the data as a mixture of distributions. Finally, although the simulation indicates the superiority of the BIC with respect to the AIC, this work does not allow us to conclude that this apparent superiority holds in practice. Indeed, in real data analysis we do not know the underlying structure of the data, and therefore actual model choice should also be based on pragmatic and expert considerations.

There are several important points that this work does not address. We are currently working on more extensive simulations and on a comparison with block clustering (e.g. CROKI2). Preliminary results are encouraging and indicate that the two approaches could be used in a complementary way, e.g. in choosing the number of clusters. Other areas for further research are: a) comparison of the AIC and the BIC performance with that of other criteria which have been proposed for model selection, such as the ICL, the CS and the NEC; b) comparison of CA with other techniques of scaling; c) effect of replacing Ward with other hierarchical classification algorithms; d) variable and object selection; e) introduction of expert input in cluster validation.

## References

- EISEN, M., SPELMAN, P., BROWN, P. and BOLTSEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science USA*, 95, 14863-14868.
- CHARRAD, M., LECHEVALLIER, Y., BEN AHMED, M. and SAPORTA, G. (2009). IDEAL, Vol. 5788 of LNCS, 260-267. Springer,
- GOVAERT, G. (1984). Algorithmes de classification d'un tableau de contingence. In : Diday E et al. (Eds), *Data Analysis and Informatics*, 3, 223-236. North-Holland.
- CIAMPI, A., DYACHENKO, A., GONZALEZ-MARCOS, A. and LECHEVALLIER, Y. (2009) Two-way classification of a data table with non-negative entries: the role of the 2 distance and Correspondence Analysis. *6<sup>th</sup> St. Petersburg Workshop on Simulation*, 523-527.
- CIAMPI, A., GONZALEZ-MARCOS, A. and CASTEJON-LIMAS, M. (2005). Correspondence analysis and two-way clustering. *SORT*, 29, 27-42.
- GREENACRE, M. (2007). *Correspondence Analysis in practice*. Chapman & Hall/CRC, Boca Raton.
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

# Computational Statistics: the Symbolic Approach

Rose Colin

Theoretical Research Institute, Sydney  
66 Drumalbyn Road, Bellevue Hill, NSW 2023, Australia *colin@tri.org.au*

**Abstract.** There is a somewhat old-fashioned tendency to think of computational statistical software as a numerical tool for working with data or models. By contrast, in this paper, we illustrate the current state of the symbolic approach to computational statistics, providing examples using *mathStatistica* (2010) which is based on top of the *Mathematica* computer algebra system. Symbolic toolkits enable one to derive exact arbitrary new theoretical results, essentially in real-time. Of course, new tools bring new problems . . . and we comment briefly on the changing nature of proof and epistemology in such a world.

**Keywords:** computer algebra systems, symbolic methods, *mathStatistica*

## 1 Introduction

The 21<sup>st</sup> century has brought with it a conceptually new methodology for conducting computational statistics: symbolic / exact methods. Recent texts applying the symbolic framework include Andrews and Stafford (2000), Rose and Smith (2002), and Drew et al (2008). This should not be confused with the area of 'symbolic data analysis' which is a completely different topic.

Traditional 20<sup>th</sup> century computer packages usually have numerical engines (as distinct from algebraic ones). Because they either cannot perform algebra, or have limited algebraic capabilities, they tend to be designed around the concept of a cookbook of specific known solutions. Each recipe has its own name which is used to grab the known solution from a look-up table, very much like the way we might look up a known answer in a textbook appendix. For example, a standard statistics package might provide a set of functions/names such as: *NormalDistMean*, *NormalDistVar*, *NormalDistCDF*, *NormalDistMGF*, *NormalDistCF*, *NormalDistSkewness*, *NormalDistKurtosis*, *NormalDistRNG*, *NormalDistQuantile* and so on. If this package supports 30 standard distributions, then it would need some 300 names/functions just to provide this particular functionality.

Unfortunately, the cookbook approach has 4 main disadvantages: First, the approach usually becomes unstuck when one wants to work with a non-standard distribution, of which there are, of course, an infinitely large number. Second, even if one starts off with a standard distribution, we again become unstuck as soon as we want to find the distribution of  $X^2$ , or the

expectation of  $\ln(x)$ , or truncate, fold or censor the random variable. Third, as a matter of design, the cookbook approach leads to a plethora of package functions and names that can make any package appear cumbersome to use, and which inevitably makes such packages appear complicated rather than simple. Fourth, and perhaps most importantly, ultimately, such a cookbook provides a set of recipes, whereas what we might truly desire from a computational statistical package is a small set of *statistical* operators that provide us with the ability to be our own *cordon bleu* chef, namely: an expectations operator, a probability function, a transformations operator, a Fisher Information function, and so on.

By contrast, symbolic / exact methods are built on top of computer algebra systems . . . programs such as *Mathematica* and Maple that understand algebra and mathematics, in addition to having numerical and graphical engines. Accordingly, symbolic algorithms can provide exact general solutions . . . not just for specific distributions / models. Symbolic computational statistical packages include mathStatICA (based on top of *Mathematica*) and APPL (based on top of Maple).

Symbolic methods include: automated expectations for arbitrary distributions, probability, combinatorial probability, transformations of random variables, products of random variables, sums and differences of random variables, generating functions, inversion theorems, maxima/minima of random variables, symbolic and numerical maximum likelihood estimation (using exact methods), curve fitting (using exact methods), non-parametric kernel density estimation (for arbitrary kernels), moment conversion formulae, component-mix and parameter-mix distributions, copulae, pseudo-random number generation for arbitrary distributions, decision theory, asymptotic expansions, order statistics (for identical and non-identical parents), unbiased estimators (h-statistics, k-statistics, polykays), moments of moments, etc

We illustrate some of the latest algorithms in the mathStatICA symbolic suite. In particular, we illustrate code for finding products of piecewise random variables, solving many-to-one transformations, finding the pdf of  $\min(X, Y, Z, \dots)$ , calculating order statistics with non-identical parent distributions, multivariate moments of moments, . . . so as to give a brief flavour of interesting and difficult problems that are now easy to solve using the symbolic approach.

Finally, new tools bring new problems . . . and we comment briefly on the changing nature of proof and epistemology in such a world.



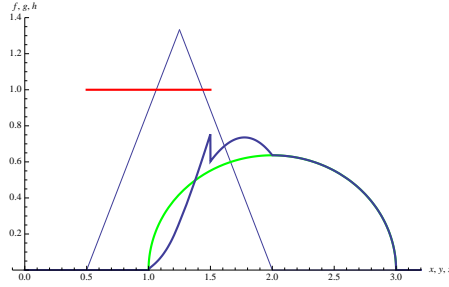


Fig. 1. plots the pdf of the maximum  $\phi(w)$ , together with the 3 underlying pdf's

## 2 Symbolic examples

### 2.1 Example: Deriving the pdf of $\max(X, Y, Z, \dots)$

Consider, say, three different distributions defined over three different domains of support. Let

$$\begin{aligned} X &\sim \text{Triangular}(1/2, 2) && \text{with pdf } f(x) \\ Y &\sim \text{Uniform}(1/2, 3/2) && \text{with pdf } g(x) \\ Z &\sim \text{half-Halo} && \text{with pdf } h(x) \end{aligned}$$

which we enter as:

$$f = \text{PDF} \left[ \text{TriangularDistribution} \left[ \left\{ \frac{1}{2}, 2 \right\} \right], x \right]; \text{domain}[f] = \{x, -\infty, \infty\};$$

$$g = 1; \quad \text{domain}[g] = \left\{ y, \frac{1}{2}, \frac{3}{2} \right\};$$

$$h = \frac{2}{\pi} \sqrt{1 - (z - 2)^2}; \quad \text{domain}[h] = \{z, 1, 3\};$$

We seek the pdf of  $W = \max(X, Y, Z)$ . The solution pdf, say  $\phi(w)$ , is simply:

$$\phi = \text{Maximum}[\{f, g, h\}, w]$$

$$\begin{cases} \frac{2\sqrt{-3+4w-w^2}}{\pi} & 2 < w < 3 \\ -\frac{2(\sqrt{-3+4w-w^2}(55-64w+16w^2)+8(-2+w)\text{ArcCos}[2-w])}{9\pi} & \frac{3}{2} \leq w \leq 2 \\ \frac{\sqrt{-3+4w-w^2}(101-261w+192w^2-40w^3)-3(13-24w+8w^2)\text{ArcCos}[2-w]}{9\pi} & \frac{5}{4} < w < \frac{3}{2} \\ \frac{2(1-2w)^2((-7+5w)\sqrt{-3+4w-w^2}+3\text{ArcCos}[2-w])}{9\pi} & 1 < w \leq \frac{5}{4} \end{cases}$$

## 2.2 Example: Products of Random Variables (piecewise functions)

Let random variable  $X \sim \text{Pareto}(a, b)$  with pdf  $f(x)$ :

$$f = ab^a x^{-(a+1)}; \quad \text{domain}[f] = \{x, b, \infty\} \&\& \{a > 0, b > 0\};$$

and let random variable  $Y$  have a standard Triangular distribution with pdf  $g(y)$  defined in piecewise form:

$$g = \begin{cases} \frac{2y}{c} & 0 \leq y \leq c \\ \frac{2(1-y)}{1-c} & c < y \leq 1 \end{cases} \quad \text{domain}[g] = \{y, -\infty, \infty\} \&\& \{0 < c < 1\};$$

We seek the pdf of the product of the above random variables  $X$  and  $Y$ , *i.e.* the pdf of  $V = XY$ . The solution is a piecewise pdf, and it can be simply obtained with `mathStatica 2` as:

**TransformProduct** $[v, \{f, g\}]$

$$\begin{cases} \frac{2av}{(2+a)b^2c} & 0 < v < bc \\ \frac{2a(b^2c(\frac{bc}{v})^a - (2+a)bv + (1+a)v^2)}{(2+3a+a^2)b^2(-1+c)v} & bc < v < b \\ \frac{2ab^a(-1+c^{1+a})v^{-1-a}}{(2+3a+a^2)(-1+c)} & v > b \end{cases}$$

## 2.3 Example: Order statistics with non-identical parent distributions

Standard order statistic calculations assume that we are dealing with samples of independent and identically distributed (iid) variables. By contrast, `mathStatica`'s new `OrderStatNonIdentical` function generalises to independent non-identical distributions. This is an enormously flexible and powerful capability.

Suppose we have three completely different distributions defined over three different domains of support:

- $f(x)$  is the pdf of an Exponential ( $\lambda$ ),
- $g(x)$  is the pdf of a standard Normal, and
- $h(x)$  is the pdf of a Uniform  $(-1, 1)$  random variable:

$$f = \frac{1}{\lambda} e^{-x/\lambda}; \quad \text{domain}[f] = \{x, 0, \infty\} \&\& \{\lambda > 0\};$$

$$g = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}; \quad \text{domain}[g] = \{x, -\infty, \infty\};$$

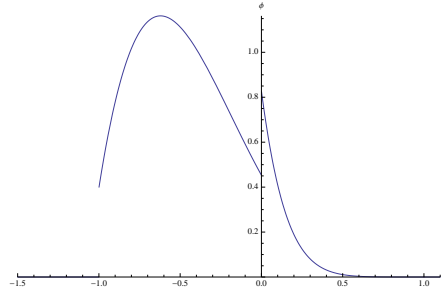


Fig. 2. pdf of the solution

$$h = \frac{1}{2}; \quad \text{domain}[h] = \{x, -1, 1\};$$

**Problem:** Consider a random sample of size  $n = 10$ . Of this sample, 4 values are drawn from the Exponential, 1 value is drawn from the Normal, and 5 from the Uniform. Find the pdf of the 2<sup>nd</sup> smallest value from the sample, namely the second order statistic.

**Solution:** The solution pdf, say  $\phi(x)$  is simply:

$$\phi = \text{OrderStatNonIdentical}[2, \{f, g, h\}, \{4, 1, 5\}]$$

$$\begin{cases} \frac{5}{64} e^{-\frac{x^2}{2}} (-1+x)^3 \left( -e^{\frac{x^2}{2}} (5+3x) + \sqrt{\frac{2}{\pi}} (-1+x^2) + e^{\frac{x^2}{2}} (3+5x) \text{Erf}\left[\frac{x}{\sqrt{2}}\right] \right) & -1 < x \leq 0 \\ -\frac{1}{64\lambda} e^{-\frac{x(8+x\lambda)}{2\lambda}} (-1+x)^3 \left( 4e^{\frac{1}{2}x(x+\frac{1}{\lambda})} (-1+x)(-3+3x-5\lambda) + 4e^{x/\lambda} \sqrt{\frac{2}{\pi}} (-1+x)^2 \lambda + \right. \\ \quad \left. + \sqrt{\frac{2}{\pi}} (1+8x-9x^2) \lambda + e^{\frac{x^2}{2}} (12-28x^2+5\lambda+x(16+35\lambda)) + \right. \\ \quad \left. - e^{\frac{x^2}{2}} (4-36x^2+4e^{x/\lambda}(-1+x)(-3+3x-5\lambda)-5\lambda+x(32+45\lambda)) \text{Erf}\left[\frac{x}{\sqrt{2}}\right] \right) & 0 < x < 1 \\ 0 & \text{True} \end{cases}$$

Here is quick plot of the solution pdf.

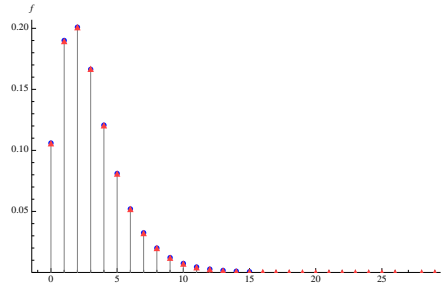
#### 2.4 Example: Find the covariance between arbitrary sample moments

Let  $(X_1, \dots, X_n)$  denote a random sample of size  $n$  drawn from population random variable  $X$ . Find the covariance between the sample mean  $\frac{1}{n} \sum X_i$  and  $(\frac{1}{n} \sum X_i^3)(\sum X_i^2)$ .

**Solution:** The *lingua franca* for such problems is the power sum  $s_r = \sum_{i=1}^n X_i^r$ . Using this notation, we are seeking  $\text{Cov}\left(\frac{s_1}{n}, \frac{s_3 s_2}{n}\right) = \mu_{1,1}\left(\frac{s_1}{n}, \frac{s_3 s_2}{n}\right)$ , i.e. the covariance is just the  $\{1, 1\}$ <sup>th</sup> product central moment. The solution with mathStatca is then simply:

$$\text{CentralMomentToCentral}\left[\{1, 1\}, \left\{\frac{s_1}{n}, \frac{s_3 s_2}{n}\right\}\right]$$

$$\frac{(-1+n)\mu_3^2}{n} + \frac{(-1+n)\mu_2\mu_4}{n} + \frac{\mu_6}{n}$$



**Fig. 3.** Comparison of the empirical pmf to the exact theoretical pmf

## 2.5 Example: Using general symbolic methods to solve numerical problems

Here is the ugliest discrete distribution we could find: Holla's distribution. The pmf contains a BesselK function, it has no closed form cdf, it is non-invertible *etc*:

$$f = \frac{1}{x!} \left( e^{\lambda/\mu} \sqrt{\frac{2}{\pi}} \sqrt{\lambda} \left( \frac{2}{\lambda} + \frac{1}{\mu^2} \right)^{\frac{1}{4}(1-2x)} \text{BesselK} \left[ \frac{1}{2} - x, \sqrt{\lambda \left( 2 + \frac{\lambda}{\mu^2} \right)} \right] \right)$$

mathStatistica's `DiscreteRNG` function takes less than  $\frac{1}{2}$  a second to generate 1 million pseudo-random drawings from this beast:

```
data = DiscreteRNG[1000000, f /. {μ → 3, λ → 10}]; //Timing
{0.432788, Null}
```

EMPIRICAL and TRUE distributions are compared in Fig. 3.

### Example: Multivariate moments of moments

Cook (1951, pp.187-195) derived explicit results for product cumulants of various bivariate k-statistics; some of the simpler results are listed in Stuart and Ord (1994, Section 13.3). In this example, we illustrate not only how to obtain these known and published results, but more generally how one can obtain *any* such desired product cumulant ... not just the simpler cases that are already known / published.

To illustrate, we will work here with the bivariate k-statistics  $k_{2,1}$  and  $k_{3,0}$ :

$$\mathbf{k21} = \mathbf{KStatistic}[\{\mathbf{2}, \mathbf{1}\}][[2]]$$

$$\frac{2s_{0,1}s_{1,0}^2 - 2ns_{1,0}s_{1,1} - ns_{0,1}s_{2,0} + n^2s_{2,1}}{(-2+n)(-1+n)n}$$

$$\mathbf{k30} = \mathbf{KStatistic}[\{\mathbf{3}, \mathbf{0}\}][[2]]$$

$$\frac{2s_{1,0}^3 - 3ns_{1,0}s_{2,0} + n^2s_{3,0}}{(-2+n)(-1+n)n}$$

We will now find the product cumulant  $\kappa_{1,1}(k_{3,0}, k_{2,1})$ . In the notation of Cook (1951, p.190) and Stuart and Ord (1994, equation (13.9)), this is the expression  $\kappa \begin{pmatrix} 3 & 2 \\ 0 & 1 \end{pmatrix}$ :

**CumulantMomentToCumulant[{1, 1}, {k30, k21}]**

$$\frac{6n\kappa_{1,1}\kappa_{2,0}^2}{(-2+n)(-1+n)} + \frac{9\kappa_{2,1}\kappa_{3,0}}{-1+n} + \frac{6\kappa_{2,0}\kappa_{3,1}}{-1+n} + \frac{3\kappa_{1,1}\kappa_{4,0}}{-1+n} + \frac{\kappa_{5,1}}{n}$$

Here is  $\kappa_{1,1,1}(k_{3,0}, k_{2,1}, k_{2,1})$ ; in the notation of Cook (1951, p.191), this is the expression  $\kappa \begin{pmatrix} 3 & 2 & 2 \\ 0 & 1 & 1 \end{pmatrix}$ :

**CumulantMomentToCumulant[{1, 1, 1}, {k30, k21, k21}]**

$$\begin{aligned} & \frac{24n(-5+2n)\kappa_{1,2}\kappa_{2,0}^3}{(-2+n)^2(-1+n)^2} + \frac{24n(-26+11n)\kappa_{1,1}\kappa_{2,0}^2\kappa_{2,1}}{(-2+n)^2(-1+n)^2} + \frac{24n(-17+7n)\kappa_{1,1}^2\kappa_{2,0}\kappa_{3,0}}{(-2+n)^2(-1+n)^2} + \\ & + \frac{12n(-12+5n)\kappa_{0,2}\kappa_{2,0}^2 + \kappa_{3,0}}{(-2+n)^2(-1+n)^2} + \frac{6(148-132n+31n^2)\kappa_{2,1}^2\kappa_{3,0}}{(-2+n)^2(-1+n)^2} + \frac{12(-29+12n)\kappa_{2,0}\kappa_{2,2}\kappa_{3,0}}{(-2+n)(-1+n)^2} + \\ & + \frac{6(56-48n+11n^2)\kappa_{1,2}\kappa_{3,0}^2}{(-2+n)^2(-1+n)^2} + \frac{24(-26+11n)\kappa_{2,0}\kappa_{2,1}\kappa_{3,1}}{(-2+n)(-1+n)^2} + \frac{24(-17+7n)\kappa_{1,1}\kappa_{3,0}\kappa_{3,1}}{(-2+n)(-1+n)^2} + \\ & + \frac{2(-55+31n)\kappa_{2,0}^2\kappa_{3,2}}{(-2+n)(-1+n)^2} + \frac{36(-5+2n)\kappa_{1,2}\kappa_{2,0}\kappa_{4,0}}{(-2+n)(-1+n)^2} + \frac{12(-26+11n)\kappa_{1,1}\kappa_{2,1}\kappa_{4,0}}{(-2+n)(-1+n)^2} + \\ & + \frac{6(-12+5n)\kappa_{0,2}\kappa_{3,0}\kappa_{4,0}}{(-2+n)(-1+n)^2} + \frac{(-55+31n)\kappa_{3,2}\kappa_{4,0}}{(-1+n)^2n} + \frac{4(-50+29n)\kappa_{1,1}\kappa_{2,0}\kappa_{4,1}}{(-2+n)(-1+n)^2} + \frac{2(-50+29n)\kappa_{3,1}\kappa_{4,1}}{(-1+n)^2n} + \\ & + \frac{2(-23+17n)\kappa_{3,0}\kappa_{4,2}}{(-1+n)^2n} + \frac{2(-20+11n)\kappa_{1,1}^2\kappa_{5,0}}{(-2+n)(-1+n)^2} + \frac{4(-7+4n)\kappa_{0,2}\kappa_{2,0}\kappa_{5,0}}{(-2+n)(-1+n)^2} + \frac{(-34+19n)\kappa_{2,2}\kappa_{5,0}}{(-1+n)^2n} + \\ & + \frac{4(-13+10n)\kappa_{2,1}\kappa_{5,1}}{(-1+n)^2n} + \frac{16\kappa_{2,0}\kappa_{5,2}}{(-1+n)n} + \frac{(-10+7n)\kappa_{1,2}\kappa_{6,0}}{(-1+n)^2n} + \frac{10\kappa_{1,1}\kappa_{6,1}}{(-1+n)n} + \frac{\kappa_{0,2}\kappa_{7,0}}{(-1+n)n} + \frac{\kappa_{7,2}}{n^2} + \end{aligned}$$

### 3 On knowledge and proof

Symbolic algorithms can derive solutions to problems that have never been posed before — — — they place enormous technological power into the hands of end-users. Of course, it is possible (though rare) that an error may occur (say in integration, or by mis-entering a model). In a sense, this is no different to traditional reference texts and journal papers which are also not infallible, and which are often surprisingly peppered with typographical or other errors. For instance, Rose and Smith (2002, p.269) show that results obtained by Fisher (1928) and published for over 80 years in texts such as Stuart and Ord (1994, eqn 12.170) are substantially in error.

In this regard, the symbolic approach offers both greater exposure to danger, as well as the tools to avoid it. The ‘danger’ is that it has become extremely easy to generate output in real-time. The sheer scale and volume of calculation has magnified, so that the average user is more likely to encounter an error, just as someone who drives a lot is more likely to encounter an accident. *Proving* that the computer’s output is actually correct can be

very tricky, or impractical, or indeed impossible for the average practitioner to do, just as the very same practitioner will tend to accept a journal result at face value, without properly checking it, even if they could do so. The philosopher, Karl Popper, argued that the aim of science should not be to prove things, but to seek to refute them. Indeed, the advantage of the computational statistical approach is that it is often possible to check one's work using two completely different methods: both numerical and symbolic. Here, numerical methods take on a new role of checking symbolic results. One can throw in some numbers in place of symbolic parameters, and one can then check if the solution obtained using symbolic methods (the exact theoretical solution) matches the solution obtained using numerical methods (typically, numerical integration or Monte Carlo methods). If the numerical and symbolic solutions do *not* match, there is an obvious problem and we can generally immediately reject the theoretical solution (*a la* Popper). On the other hand, if the two approaches match up, we still do not have a proof of correctness ... all we have is just one point of agreement in parameter space. We can repeat and repeat and repeat the exercise with different parameter values ... and as we do so, we effectively build up, not an absolute proof in the traditional sense, but, appropriately for the statistics profession, an ever increasing degree of confidence, ... effectively a proof by probabilistic induction ... that the theoretical solution is indeed correct. This is an extremely valuable (though time-consuming) skill to develop, not only when working with computers, but equally with textbooks and journal papers.

## References

- ANDREWS, D. F. and STAFFORD, J. E. H. (2000): *Symbolic Computation for Statistical Inference*. Oxford University Press, New York.
- COOK, M. B. (1951): Bivariate k-statistics and cumulants of their joint sampling distribution, *Biometrika*, **38**, 179-195.
- DREW, J. H., EVANS, D. L., GLEN, A. G. and LEEMIS, L. M. (2008): *Computational Probability*. Springer.
- FISHER, R. A. (1928): Moments and product moments of sampling distributions, *Proceedings of the London Mathematical Society*, series 2, volume 30, 199—238 (reprinted in Fisher, R. A. (1950), *Contributions to Mathematical Statistics*. Wiley, New York).
- mathStatica (2010), [www.mathStatica.com](http://www.mathStatica.com), Sydney.
- ROSE, C. (2010): Computational Statistics. In: M. Lovric (Ed.): *International Encyclopedia of Statistical Science*. Springer-Verlag.
- ROSE, C. and SMITH, M. D. (2002): *Mathematical Statistics with Mathematica*. Springer, New York.
- STUART, A. and ORD, J. K. (1994): *Kendall's Advanced Theory of Statistics*. Volume 1, 6th edition, Edward Arnold, London.

# A Mann-Whitney Spatial Scan Statistic for Continuous Data

Lionel Cucala

Université des Sciences et Techniques du Languedoc  
Place Eugène Bataillon, Montpellier, France, *lcucala@math.univ-montp2.fr*

**Abstract.** A new scan statistic is proposed for identifying clusters of high or low values in georeferenced continuous data. On the one hand, it relies on a concentration index which is based on the Mann-Whitney statistic and thus is completely distribution-free. On the other hand, the possible spatial clusters are given by an original graph-based method. This spatial scan test seems to be very powerful against any arbitrarily-distributed cluster alternative. These results have applications in various fields, such as the epidemiological study of rare diseases or the analysis of astrophysical data.

**Keywords:** cluster detection, epidemiology, scan statistics, spatial marked point processes

## 1 Introduction

Let  $X_1, \dots, X_n$  be continuous random variables associated to their respective random locations  $S_1, \dots, S_n$  in the observation domain  $D \in \mathbb{R}^d$ . In other words,  $\{(S_1, X_1), \dots, (S_n, X_n)\}$  is a point process with continuous marks. See Møller and Waagepetersen (2003) for a review of the models, the inference methods and the simulation techniques associated to these mathematical objects. However, the aim of this paper is quite different since we would like to identify the areas where the continuous marks are significantly higher (or lower) than elsewhere: these areas will be called clusters.

Identifying level sets of a spatial field when knowing its value only in a few locations is a quite standard problem which has been extensively studied for many years (Cressie (1993)). However, in our case, these techniques are not appropriate since we cannot assume any dependence structure between the  $X_i$ 's: they are not issued from a spatial field having smoothness properties.

On another hand, the question of looking for clusters in unmarked spatial point processes has been thoroughly investigated since Kulldorff (1997) introduced the first spatial scan statistic. The principle is the following: the first step consists in creating a collection of potential clusters, usually circular or elliptic; a concentration index is chosen and the scan statistic is the maximum (or minimum) concentration over the potential clusters; the significance value of the scan statistic is estimated through a Monte-Carlo procedure. Actually,

looking for clusters in marked point processes is not that different since only an appropriate concentration index is needed.

Section 2 describes a new method to obtain a family of arbitrarily-shaped potential clusters. Then a new concentration index adapted to continuous marks is introduced and the scan statistic is defined. Finally, this spatial scan statistic is applied to an astrophysical data set. The paper is concluded with a discussion.

## 2 A set of data-based possible clusters

As mentioned earlier, the classical spatial scan statistic looks for the maximal concentration in a set of circular or elliptic windows. These windows depend on the spatial distribution of the events, but also on various parameters: the coordinates of the centroids, the shape and angle of the ellipses. Remark that their shape is predetermined and may not be adapted to the shape of the real cluster, specially if we consider only circular windows. Moreover, the number of possible clusters to test becomes huge when the number of events increases or when the data are three-dimensional, and so does the computation time. Recently, a couple of arbitrarily-shaped spatial scan statistics have been proposed (Patil and Taillie (2004), Duczmal and Assunção (2004), Tango and Takahashi (2005)) but they only deal with grouped data, that is when the individual locations are unknown (we only know to which subdomain each event belongs). Following this idea, we introduce a set of possible clusters that are completely data-based, do not depend on any parameter and which number is quite limited.

On what follows, we describe a technique recently introduced by Bar-Hen et al. (2007): it consists in associating a family of graphs to the original point process. Let  $S_1, \dots, S_n$  be defined as in the Introduction. For any  $\delta \in \mathbb{R}^+$ , a graph, denoted by  $\mathcal{G}(\delta)$ , is defined: its set of vertices is  $\{1, \dots, n\}$  and its set of edges is  $\{(i, j) : d(S_i, S_j) \leq \delta, 1 \leq i < n, i < j \leq n\}$ , where  $d(.,.)$  stands for the Euclidian distance. Since each vertex  $i$  is associated to the event with location  $S_i$ , this consists in linking only the events whose distance is less than  $\delta$ . The connected component of the vertex  $i$  in this graph is denoted by  $\mathcal{N}_i(\delta)$ . Let  $A_i(\delta) = \{s \in D : \exists j \in \mathcal{N}_i(\delta), d(s, S_j) \leq \delta\}$  be the  $\delta$ -neighbouring of the set of vertices  $\mathcal{N}_i(\delta)$ .

Since a cluster usually contains events which are a small distance away from at least another event of the cluster, it seems quite logical to restrict the possible clusters to all the  $\delta$ -neighbourings associated to the events locations. Thus, we decide to restrict the set of possible clusters to the areas

$$\{A_i(\delta) : 1 \leq i \leq n, \delta \in \mathbb{R}^+\}.$$

At first sight, the number of possible clusters to test may appear quite large. However, it can be drastically reduced. First, the set of distances  $\delta$  to analyse is just the set of distances  $d_{i,j} = d(S_i, S_j)$ , since the graph  $\mathcal{G}(\delta)$  remains the



same when  $\delta$  is between two consecutive  $d_{i,j}$ 's. Moreover, a new edge is always added to the graph  $\mathcal{G}(\delta)$  when  $\delta$  reaches  $d_{i,j}$  but the connected components of the graph may remain the same: in that case, the number of events in  $A_i(\delta)$  remains the same and the concentration cannot be maximised for  $\delta = d_{i,j}$ . Finally, only the potential cluster  $A_i(d_{i,j}) = A_j(d_{i,j})$  should be analysed when  $\delta$  reaches  $d_{i,j}$  since this is the only neighbouring in which the number of events increases. Let  $\mathcal{G}^-(\delta)$  denote the graph whose set of vertices is  $\{1, \dots, n\}$  and whose set of edges is  $\{(i, j) : d(S_i, S_j) < \delta, 1 \leq i < n, i < j \leq n\}$ . The connected component of the vertex  $i$  in this graph is denoted by  $\mathcal{N}_i^-(\delta)$ . The final set of possible clusters is thus

$$\mathcal{C} = \{A_i(d_{i,j}) : 1 \leq i < n, i < j \leq n, \mathcal{N}_i^-(d_{i,j}) \neq \mathcal{N}_j^-(d_{i,j})\}.$$

The process we just described is similar to the creation of an Euclidian spanning tree by linking the closest vertices as long as no loop is created. Hence the number of possible clusters in  $\mathcal{C}$  is exactly the number of edges linking the  $n$  vertices of a graph without any loop, the number of edges in a spanning tree, that is  $n - 1$ , which is very small compared to the number of possible clusters explored by the classical scan statistic.

### 3 A rank-based concentration index

Now that the potential clusters are defined, we need a concentration index  $I(Z)$  for any  $Z \in D$ . The one introduced by Kulldorff (1997) to look for clusters of events was based on the likelihood ratio between two hypotheses: the null hypothesis that the events are issued from an uniform distribution over  $D$  and the alternate distribution that they are issued from the mixture of an uniform distribution over  $Z$  and another uniform distribution over  $\bar{Z} = \{s \in D : s \notin Z\}$ . Following this idea, Kulldorff et al. (2009) recently introduced a concentration index for continuous data: under the null hypothesis, all the  $X_i$ 's are issued from the same Gaussian distribution; under the alternate hypothesis, the  $X_i$ 's associated to locations inside  $Z$  are issued from a certain Gaussian distribution and the ones associated to locations outside  $Z$  are issued from another Gaussian distribution. Denote  $X_Z = \sum_{i:S_i \in Z} X_i$ ,  $XX_Z = \sum_{i:S_i \in Z} X_i^2$ ,  $n_Z = \text{Card}\{i : S_i \in Z\}$  and  $\mu_Z = \frac{X_Z}{n_Z}$ . This concentration index thus reduces to

$$I_{LR}(Z) = \frac{1}{\sigma_Z^2} = \frac{n}{XX_Z - 2X_Z\mu_Z + n_Z\mu_Z^2 + XX_{\bar{Z}} - 2X_{\bar{Z}}\mu_{\bar{Z}} + n_{\bar{Z}}\mu_{\bar{Z}}^2}.$$

When looking for clusters of events, Cucala (2008) and Cucala et al. (2009) showed that the concentration indices issued from a likelihood ratio may be overpowered by concentration indices not relying on any alternate hypothesis. Based on this, we introduce a new concentration index for continuous data. The Mann-Whitney statistic (Mann and Whitney (1947)) is very useful to

test whether two samples are issued from the same distribution, without any assumption about this distribution. More precisely it relies on the null hypothesis that the two populations have the same median. For any window  $Z \subset D$ , we propose to compute the Mann-Whitney statistic associated to the sample of observations in  $Z$ , and to consider the significance value of this statistic as a concentration index of high marks in  $Z$ . Let  $X_{(1)}, \dots, X_{(n)}$  be the order statistics associated to the  $X_i$ 's, and  $R_j$  be the rank of  $X_j$  among the  $X_i$ 's, such that  $X_{(R_j)} = X_j, \forall j = 1, \dots, n$ . The Mann-Whitney statistic associated to the window  $Z$  is the sum of ranks  $SR(Z) = \sum_{i:S_i \in Z} R_i$ , with mean  $M(Z) = \frac{n_Z(n_Z+1)}{2}$  and variance  $V(Z) = \frac{n_Z n_Z (n_Z+1)}{12}$  under the null hypothesis. Since the distribution of the ratio  $\frac{SR(Z) - M(Z)}{\sqrt{V(Z)}}$  is very close to the standard Gaussian distribution, even for moderate  $n$  and  $n_Z$  (Hollander and Wolfe (1999)), the significance value of  $SR(Z)$  is close to  $F\left(\frac{SR(Z) - M(Z)}{\sqrt{V(Z)}}\right)$ , where  $F(\cdot)$  is the standard Gaussian distribution function. The rank-based concentration index thus reduces to

$$I_{MW}(Z) = \frac{SR(Z) - M(Z)}{\sqrt{V(Z)}}.$$

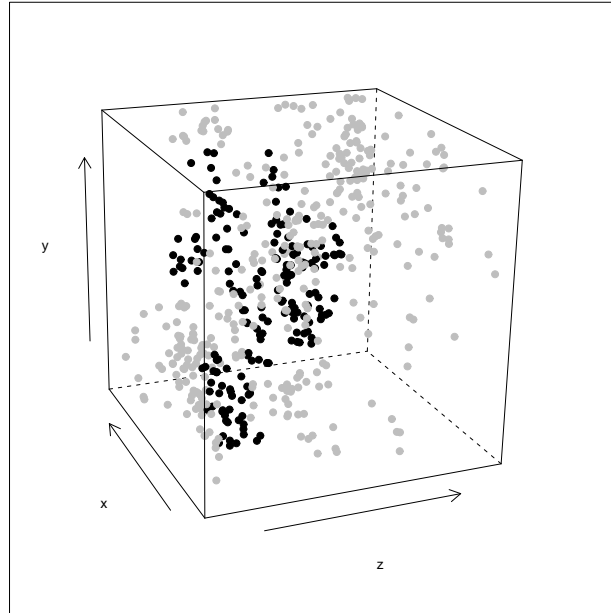
Using this concentration index and the set of possible clusters defined in the previous section, we introduce the Mann-Whitney spatial scan statistic

$$A_{MW} = \sup_{Z \in \mathcal{C}} I_{MW}(Z).$$

This statistic is just the maximal concentration observed in the defined neighbourhoods. When the graph-based scan statistic has been computed, its significance value is assessed by a Monte Carlo permutation procedure, just as a classical scan statistic.

## 4 An application to astrophysical data

Based on SDSS (Sloan Digital Sky Survey) data, we have compiled a sample of galaxies 10 Mpc around the Ursa Major galaxy supercluster center ( $11^h 30^m + 55^\circ$ ). This system is a close ( $z \simeq 0.06$ ), compact and fairly isolated supercluster in the local Universe, and thus it can be used to study how galaxy clusters develop in space in the absence of external effects (Kopylov and Kopylova (2001), Kopylova and Kopylov (2007)). Equatorial coordinates were converted to supergalactic (cartesian) ones to probe the 3D galaxy distribution. Color indices were obtained from the  $g - r$  band magnitudes subtraction. This quantity is a very important astrophysics variable. Colors are related to type and distribution of galaxies in space. Indeed, cluster elliptical galaxies are among the reddest objects in the Universe. Most of them are located at the central regions of galaxy systems. Thus, colors can be



**Fig. 1.** Detection of red galaxies clusters.

properly used to detect significant overdensities in optical surveys of galaxies (Gal (2007)). In this context, the reddest galaxies in our sample were taken as marks of connected points to find galaxy clusters using the Mann-Whitney spatial scan statistic. The coordinates of the galaxies correspond to  $S_1, \dots, S_n$  and the associated color measurements to  $X_1, \dots, X_n$ . Figure 1 gives the results.

The black dots represent the 187 galaxies contained in the most significant cluster, whose p-value is 0.001. It includes two clusters catalogued by Kopylova and Kopylov (2007) in Ursa Major out to 10 Mpc, with locations at (0.4686, 0.0832, -0.6737) and (6.4663, 1.3902, -2.8560). These clusters were obtained using a kernel density estimation method. Hence, this single cluster should be encompassing the two galaxy systems previously found in this

field. The result is very interesting because we can link them through a red galaxies connection.

We have used the SaTScan software to compute the scan statistic based on likelihood ratio. However, as we are dealing with three-dimensional data, this program only explores circular windows and the most concentrated circular window is clearly non-significant since its p-value is 0.338.

## 5 Discussion

The Mann-Whitney spatial scan tests allows one to detect clusters in a spatial marked point process without assuming anything about the clustering structure, and without setting up any parameter. On the other hand, the widely-used circular scan statistic is very restrictive concerning the shape of the possible clusters. The recently-introduced elliptic scan statistic is more flexible but is computationally very consuming when the control data are not aggregated and of course it is not adapted to three-dimensional data.

A simulation study may confirm that the likelihood-based concentration indicators are not always the most efficient. As when the point process is unmarked, an hypothesis-free concentration index may be more adapted to cluster detection. We have to mention that this Mann-Whitney index may also be used with grouped data.

Finally, the results obtained from the astrophysical data are very encouraging and a further analysis, based on a larger dataset, will be the subject of a future work.

## References

- BAR-HEN, A., KOSKAS, M. and PICARD, N. (2007): Spatial cluster detection using the number of connected components of a graph. *Technical report*.
- CRESSIE, N. A. (1993): *Statistics for Spatial Data*. Wiley, New York.
- CUCALA, L. (2008): A hypothesis-free multiple scan statistic with variable window. *Biometrical Journal* 50, 299-310.
- CUCALA, L., DEMATTEI, C., LOPES, P. and RIBEIRO, A. (2009): Spatial scan statistics for case event data based on connected components. *Submitted*.
- DUCZMAL, L. and ASSUNÇÃO, R. (2004): A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis* 45, 269-286.
- GAL, R. (2007): *A pan-chromatic view of clusters of galaxies and the large scale structure*. Springer, Berlin.
- HOLLANDER, M. and WOLFE, D.A. (1999): *Nonparametric Statistical Methods*. Wiley, New York.
- KOPYLOV, A.I. and KOPYLOVA, F.G. (2001): The Ursa Major supercluster of galaxies - I. The luminosity function. *Astronomical Letters* 27, 140.
- KOPYLOVA, F.G. and KOPYLOV, A.I. (2007): Structure and dynamics of the Ursa Major Supercluster of Galaxies. *Astronomical Letters* 33, 211.

- KULLDORFF, M. (1997): A spatial scan statistic. *Communications in Statistics. Theory and Methods* 26, 1481-1496.
- KULLDORFF, M., HUANG, L. and KONTY, K. (2009): A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics* 8: 58.
- MANN, H. and WHITNEY D. (1947): On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18, 50-60.
- MØLLER, J. and WAAGEPETERSEN, R. P. (2004): *Statistical inference and simulation for spatial point processes*. Chapman & Hall, Boca Raton.
- PATIL, G.P. and TAILLIE, C. (2004): Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* 11, 183-197.
- TANGO, T. and TAKAHASHI, K. (2005): A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 4: 11.



# Quantile Regression for Group Effect Analysis

Cristina Davino<sup>1</sup> and Domenico Vistocco<sup>2</sup>

<sup>1</sup> Dipartimento di Studi sullo Sviluppo Economico, Università di Macerata  
Piazza Oberdan 3, Macerata, Italy, *cdavino@unimc.it*

<sup>2</sup> Dipartimento di Scienze Economiche, Università di Cassino  
Via S. Angelo S.N., Cassino, Italy, *vistocco@unicas.it*

**Abstract.** This paper aims to propose an innovative approach to identify group effects in a quantile regression model. Quantile regression is a quite recent regression technique that allows to focus on the effects that a set of explanatory variables have on the entire conditional distribution of the dependent variable. The proposal concerns the use of a stratification variable in order to detect effects attributable to different group membership. An empirical analysis is also provided to measure the changes on job satisfaction owing to modification of the evaluation of different job features and taking into account the type of job (self-employed, private employee or public employee). This latter variable is used to estimate the group effects.

**Keywords:** quantile regression, group effects

## 1 Introduction

The aim of the paper is to propose an innovative approach based on quantile regression to identify a typology. The detection of a typology could derive either from the clustering of units into groups or from the analysis of the differences among a priori defined groups. In spite of this double potential meaning, the present paper focuses only on the analysis of the differences among groups using an available stratification variable.

The methodological framework is represented by quantile regression, as introduced by Koenker and Basset (1978). This method may be considered as an extension of classical least squares estimation of conditional mean models to the estimation of a set of conditional quantile functions. The use of quantile regression offers a more complete view of the relationships among variables, providing a method for modelling the rates of changes in the response variable at multiple points of its conditional distribution. As the independent variables could affect the response variable in different ways at different locations of its conditional distribution, useful insights derive from extracting information at other places other than the expected value.

It is a matter of fact that if two units have similar features/behaviours or belong to the same group of a stratification variable, the dependence structure of a regression model is more alike. The approach proposed in this paper aims to estimate group effects in a regression model taking into account the impact of the regressors on the entire conditional distribution of the dependent

variable. This goal can be achieved using different approaches, such as the estimation of different models for each group or the introduction of dummy variables among the regressors denoting group membership. Notwithstanding, the first solution does not allow to identify the impact of each group on the dependent variable and it requires tools for model comparisons, while the second solution is able to catch the effect of each group but it does not provide the impact of the levels on each regressor. Multilevel modelling (Gelman and Hill, 2007), also known as hierarchical linear models, mixed models and random effect models, represents a widespread approach to take into account and explore dependencies in hierarchical population structures. In particular, multilevel modelling restricts to the analysis of group differences in the mean of the dependent variable and it is a parametric model requiring distributional hypothesis.

The paper is organized as follows: basic notation is introduced in Section 2 together with quantile regression model. The proposed approach for typology identification is detailed in Section 3 from a methodological point of view while an empirical analysis for the evaluation of job satisfaction is provided in Section 4. Some concluding remarks and future work directions are described in the final section.

## 2 Basic notation

The proposed approach aims to explain the conditional distribution of a dependent variable starting from a set of predictors taking into account that each unit can belong to a different group of a stratification variable.

Let us consider a data matrix composed of a dependent variable vector  $\mathbf{y}_{[n]}$  and a matrix  $\mathbf{X}_{[n \times p]}$  of regressors. Let the data matrix be row-partitioned in  $G$  strata. The generic elements of the response vector and of the regressors matrix are respectively  ${}_g y_i$  and  ${}_g x_{ij}$  ( $i=1, \dots, n$ ;  $j=1, \dots, p$ ;  $g=1, \dots, G$ ) where  $n$  denotes the number of units,  $p$  the number of regressors and  $G$  the number of groups or levels. It follows that  $n_g$  is the number of units in group  $g$  and the total sample size can be expressed as  $n = \sum_{g=1}^G n_g$ .

In a classical linear regression model, for a given group  $g$  it is possible to define the sample regression function as:

$${}_g \mathbf{y} = {}_g \mathbf{X}_g \beta + {}_g \mathbf{e} \quad (1)$$

Quantile regression (QR) can be viewed as an extension of classical LS estimation for conditional quantile functions. The QR model for a given quantile  $\theta$  and for a given group  $g$  follows:

$$Q^\theta({}_g \mathbf{y} | {}_g \mathbf{X}) = {}_g \mathbf{X}_g \beta(\theta) \quad (2)$$

where  $0 < \theta < 1$  and  $Q^\theta(\cdot | \cdot)$  denotes the conditional quantile function for the  $\theta^{th}$  quantile.



The parameter estimates in QR linear models have the same interpretation as those of any other linear model: they measure the change in the conditional quantile of  $\mathbf{y}$  per unit change in the corresponding regressor, holding the values of the others regressors constant.

QR provides, for each group  $g$ , a coefficients matrix  ${}_g\hat{\mathbf{B}}(\theta)_{[p \times \Theta]}$  whose generic element  ${}_g\hat{\beta}(\theta)_j$  can be interpreted as the rate of change of the  $\theta^{th}$  quantile of the dependent variable distribution per unit change in the value of the  $j^{th}$  regressor. The value of  $\Theta$  is determined by the number of conditional quantiles that have been estimated.

Starting from the estimated regression quantiles, the density estimation of the response variable can be an useful tool to go into more depth on the effect of a given regressor. Exploiting the quantile regression estimates, indeed, it is straightforward to estimate the response variable conditional distribution for a given group  $g$  as follows:

$${}_g\hat{\mathbf{y}} = {}_g\mathbf{X}_g\hat{\beta}(\theta), \text{ for } 0 < \theta < 1. \quad (3)$$

The estimated conditional distribution is strictly dependent on the values used for the covariates. It is then possible to use different potential scenarios in order to evaluate the effect on the conditional response variable, carrying out a what-if study.

### 3 An innovative approach for typology identification

If a single analysis is performed for each group, a methodological problem arises because it is necessary to compare  $G$  sets of coefficients:  ${}_g\hat{\mathbf{B}}(\theta)_{[p \times \Theta]}$ , for  $g = 1, G$ .

The proposed approach is mainly based on the undoubted potentiality of QR to explore the entire conditional distribution of the response variable and it aims to discover the best model for each group. The approach is structured in the following steps:

- a. Global estimation;
- b. Identification of the best model for each unit;
- c. Identification of the best model for each group;
- d. Partial estimation.

In the first step, a QR model is estimated without taking into account the group variable:

$$Q^\theta(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}}(\theta) \quad (4)$$

In the second step, the coefficients matrix  $\hat{\mathbf{B}}(\theta)$  and the regressors data matrix  $\mathbf{X}$  are used to estimate the response variable conditional distribution matrix:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}(\theta) \quad (5)$$

The generic element  $\hat{\mathbf{y}}_i(\theta)$  of the  $\hat{\mathbf{Y}}$  ( $n \times \Theta$ ) matrix represents the value of the density estimation for the  $i^{th}$  units according to the  $\theta^{th}$  quantile.

The best model for each unit  $i$  is identified through the quantile model able to better estimate the response variable, namely to minimize the difference between the observed and the estimated value:

$$\theta_i : \underset{\theta=1,\Theta}{\operatorname{argmin}} y_i - \hat{y}_i(\theta) \quad (6)$$

The identification of the best quantile model for each unit allows to select from the  $\hat{\mathbf{Y}}$  matrix the best density estimation vector:  $\hat{\mathbf{y}}_{\theta}^{best}$ .

In the third step, units are partitioned according to the group variable and the best model for each group is identified synthesizing the quantiles assigned in the previous step to each units belonging to the same group:  ${}_g\theta^{best}$ , for  $g = 1, G$ . The synthesis can be performed through mean or median but the choice of the proper central tendency index depends on the particular distribution of  $\hat{\mathbf{y}}_{\theta}^{best}$  in each group. Group effects can be identified analyzing differences among the  ${}_g\theta^{best}$ .

In the last step, QR is again executed on the total sample but only retaining the quantiles assigned to each group in the previous step. Differences in the explaining capability of the regressors according to the group membership can be easily identified through the inspection of a single coefficient matrix called  $\hat{\mathbf{B}}(\theta)^{best}$  of dimension  $[p \times G]$ .

## 4 An empirical analysis: the evaluation of job satisfaction

### 4.1 The dataset

The proposed approach is exploited on a real dataset, aiming to measure how the evaluation of several job features affects the overall job satisfaction taking into account that this effect can be different for unsatisfied and satisfied workers.

This evaluation is based on a random sample of 400 students graduated at University of Macerata (Davino, Vistocco 2007) and in a working condition at the time of the interview. The submitted questionnaire concerns the evaluation of the different aspects related to the working experience: syllabus, University background, consistent training, career chance, skill, personal interest, free time, salary, office location, job stability, human relationships, amusing job, independence. Finally an overall opinion on the job is recorded. All the variables have been measured on a ten-levels scale.

The main descriptive statistics of the overall job satisfaction evaluation (Minimum=1, Q1=7.00, Mean=7.68, Median=8.00, Q3=8.85, Maximum=10) show the presence of skewness in the distribution.

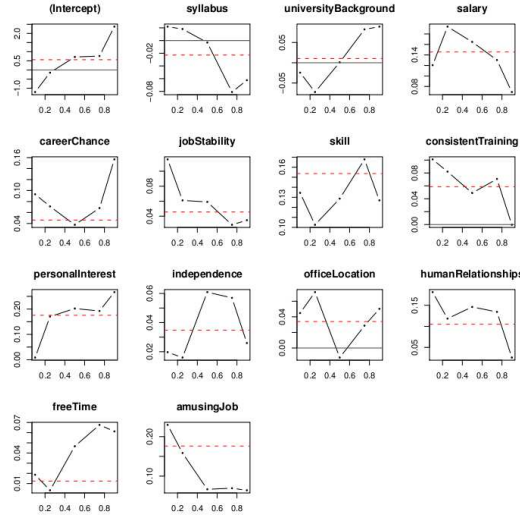


Fig. 1. LS and QR coefficients.

## 4.2 Main results

In Figure 1, LS and QR coefficients are graphically represented for the different evaluations of job features. LS coefficients measure a change in the conditional mean while QR coefficients measure a change on a given conditional quantile. The horizontal axes display the different quantiles while the effect of each feature holding the others is represented on the vertical axes. The dashed lines parallel to the horizontal axis correspond to LS coefficients. The graphical representation allows to visually catch the different effect of the evaluations of the several job features on the overall job satisfaction: all the coefficients related to fulfillment aspects (career chance, skill, personal interest, free time), increase moving from lower to upper quantiles while the coefficients related to tangible aspects (salary, office location, job stability) move on the contrary with the exception of the office location. Figure 1 can be useful to explore from a descriptive point of view the coefficient trends along the different quantiles and in comparison with LS ones. In order to take into account the generalization capability of the model, the coefficients estimated by LS and QR (the following quantiles are considered: 0.1; 0.25; 0.5; 0.75; 0.9) are shown in Table 1 highlighting in bold significant coefficients at  $\alpha=0.1$ . Classical LS estimates are able to extract a reduced information about the job features contributing to the overall job satisfaction while QR results widen the set of significant coefficients giving a detailed description of the factors influencing the whole conditional distribution of the job satisfaction.

Let us consider that the effects on the job satisfaction could be conditioned by the type of job: 1.self-employed, 2.private employee or 3.public employee. The QR based approach proposed in Section 3 is applied for ty-

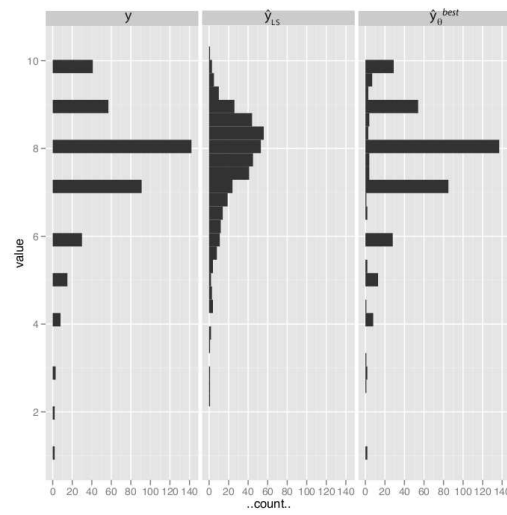
Variable	LS	$\theta=0.1$	$\theta=0.25$	$\theta=0.5$	$\theta=0.75$	$\theta=0.9$
Intercept	0.403	<b>-1.211</b>	-0.149	0.711	0.761	<b>2.370</b>
syllabus	-0.009	0.022	0.018	-0.003	-0.081	-0.062
University background	0.004	-0.024	-0.072	0.001	0.082	0.089
salary	<b>0.146</b>	<b>0.120</b>	<b>0.194</b>	<b>0.165</b>	<b>0.130</b>	0.069
career chance	<b>0.078</b>	0.093	0.071	0.037	0.068	<b>0.157</b>
job stability	<b>0.061</b>	<b>0.116</b>	0.061	0.059	0.028	0.035
skill	<b>0.117</b>	0.134	0.102	<b>0.129</b>	<b>0.168</b>	0.127
consistent training	0.043	0.101	0.082	0.049	0.070	-0.000
personal interest	<b>0.187</b>	0.008	<b>0.170</b>	<b>0.202</b>	<b>0.192</b>	<b>0.267</b>
independence	0.051	0.019	0.016	0.061	0.056	0.026
office location	0.031	0.044	0.072	-0.012	0.029	0.050
human relationships	0.126	<b>0.181</b>	0.118	<b>0.146</b>	<b>0.134</b>	0.026
free time	0.017	0.189	0.003	0.047	<b>0.067</b>	0.061
amusing job	<b>0.147</b>	<b>0.230</b>	<b>0.158</b>	0.066	0.069	0.064

Table 1. LS and QR coefficients

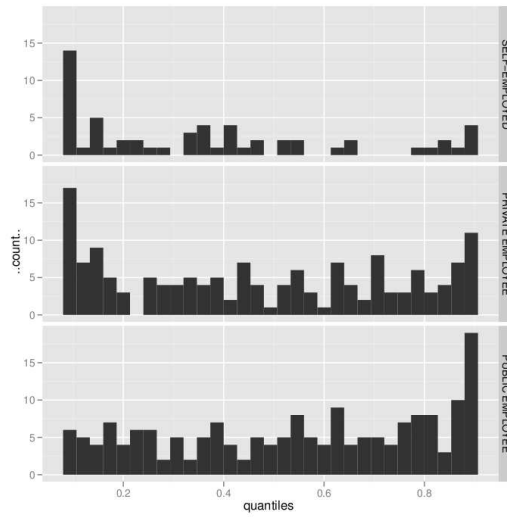
polity identification, namely for the estimation of the group effects. Results in Table 1 and Figure 1 provide indeed global estimations without considering the type of job. Therefore, the second step requires the identification of the best model for each unit through the quantile able to better estimate the response variable. The added value in considering the reconstructed estimated response variable  $\hat{\mathbf{y}}_{\theta}^{best}$  instead of a LS estimated response variable is evident from Figure 2 reproducing the histograms of the dependent variable (left panel) and of the estimated dependent variable using LS (middle panel) and the proposed QR approach (right panel). In the third step, units are partitioned according to the type of job. Figure 3 shows the distribution of the “best” quantiles assigned to units grouped according to the type of job. The best model for each category of the grouping variable is identified through the mean value of the “best” quantiles assigned to units belonging to the  $g^{th}$  group:  $\theta_1^{best}=0.371$ ;  $\theta_2^{best}=0.474$ ;  $\theta_3^{best}=0.548$ .

Finally, QR is again executed on the total sample but only retaining the quantiles assigned to each category of the grouping variable. In Table 2 for each regressor (Table rows) and for each type of job (Table columns) the values of the final coefficient matrix  $\hat{\mathbf{B}}(\theta)^{best}$  are shown (in bold significant coefficients at  $\alpha=0.1$ ).

Each category of the group variable is “best” represented by a different quantile. QR coefficients with group effects reveal slight differences among the groups if regressors are ranked according to their coefficients. Notwithstanding, differences among the groups can be identified looking at the different magnitude of the coefficients. For example, evaluation on salary and amusing job has a major impact on the overall satisfaction in the group of self-employed. Private employees mainly differentiate from the other groups for the effect played by the evaluation on human relationships. Finally, the



**Fig. 2.** Distribution of the dependent variable (left panel) and of the estimated dependent variable using LS (middle panel) and the proposed QR approach (right panel).



**Fig. 3.** Distribution of the “best” quantiles assigned to each unit grouped according to the type of job.

highest coefficients in the group of public employees are on skill and personal interest.

Variable	self-employed	private employee	public employee
intercept	0.646	0.683	0.694
syllabus	-0.007	-0.012	-0.035
University background	-0.030	0.006	0.026
salary	<b>0.201</b>	<b>0.152</b>	<b>0.160</b>
career chance	0.012	0.037	-0.008
job stability	0.049	0.034	0.054
skill	<b>0.118</b>	<b>0.156</b>	<b>0.184</b>
consistent training	0.065	0.066	0.064
personal interest	<b>0.200</b>	<b>0.175</b>	<b>0.202</b>
independence	0.022	0.035	0.035
office location	0.011	-0.006	0.007
human relationships	<b>0.114</b>	<b>0.152</b>	<b>0.107</b>
free time	0.018	0.032	0.026
amusing job	<b>0.148</b>	<b>0.124</b>	<b>0.141</b>

Table 2. QR coefficients with group effects

## 5 Concluding remarks and further issues

The QR based approach introduced in the paper allows to take into account group effects when the impact of the regressors on the entire conditional distribution of a dependent variable is of interest. This paper focuses on the availability of a priori defined groups exploiting a stratification variable.

The approach can represent a valid tool to cluster units according to the dependence structure without a priori information but only using the observed similarities among them in terms of conditional quantile estimates.

A further development could include an evaluation of the statistical significance of the differences among the “best” quantiles assigned to the different groups and the inclusion of time as grouping variable.

## References

- DAVINO, C., VISTOCCO, D. (2007): The evaluation of university educational processes: a quantile regression approach. *STATISTICA*, n.3, pp. 267-278.
- EIDE, E. SHOWALTER, M.H. (1998): The effect of school quality on student performance: a quantile regression approach. *Economics Letters* 58, 345-350.
- FURNO, M. (2010): Quantile regression analysis of the Italian school system. *Statistical Modelling*, vol. 4, 2010, in press.
- GELMAN, A. HILL, J. (2006): *Data analysis using regression and multi-level/hierarchical models*. Cambridge University Press.
- LOCKHEED, M.E. HANUSHECK, E.R. (1994): Concepts of Educational Efficiency and Effectiveness, in Torsten Husn and T. Neville Postlethwaite (ed.), *International Encyclopedia of Education*, second edition, Volume 3 (Oxford: Pergamon, 1994), pp. 1779-1784.
- KOENKER, R., BASSET, G.W. (1978): Regression Quantiles, *Econometrica* 46, 33-50.
- KOENKER, R. (2005): *Quantile Regression*. Econometric Society Monographs.

# Regularized Directions of Maximal Outlyingness

Michiel Debruyne<sup>1</sup>

Department of mathematics and computer science, Universiteit Antwerpen,  
Middelheimlaan 1G, 2020 Antwerpen, Belgium, *michiel.debruyne@ua.ac.be*

**Abstract.** The following problem is posed: once an outlier is detected in a multivariate data set, how to find the subset of variables that contribute most to this outlyingness? It turns out that a direction of maximal outlyingness can be rewritten as the normed solution of a classical least squares regression problem. We propose to add a  $L_1$  penalty term in this expression, thus replacing the classical least squares regression by the LASSO. This yields a path of regularized directions of maximal outlyingness. Based on such a path, an algorithm is proposed to select a subset of variables that are most relevant to the outlyingness of the outlier under consideration.

**Keywords:** robust statistics, outlyingness, variable selection, LASSO

## 1 Introduction

Let  $X = (x_1, \dots, x_n)^t$  be a sample of multivariate observations  $x_i \in \mathbb{R}^p$  generated from a  $p$ -dimensional elliptical distribution. It is well known that the classical mean and the classical covariance matrix are very sensitive to outliers in the data. Many robust alternatives are available, e.g. the MCD estimator (Rousseeuw, 1984), S-estimators (Rousseeuw and Leroy, 1987),  $\tau$ -estimators (Lopuhaä, 1991) and many more, all returning robust estimates of location and scatter  $\bar{x}_r$  and  $\Sigma_r$ . The squared robust Mahalanobis distance of an observation  $x_j$  is defined by

$$m(x_j; \bar{x}_r, \Sigma_r)^2 = (x_j - \bar{x}_r)^t \Sigma_r^{-1} (x_j - \bar{x}_r).$$

The Mahalanobis distance measures the distance between  $x_j$  and the robust center taking into account the elliptical distribution of the data. Weights can be assigned to every observation as  $w_i = w(m(x_i; \bar{x}_r, \Sigma_r))$  with  $w(\cdot)$  some weight function. In the case of the MCD estimator for example one typically computes the robust Mahalanobis distances using the raw MCD estimators of location and scatter. Afterward weights are defined by  $w_i = I(m(x_i; \bar{x}_r, \Sigma_r)^2 \leq \chi_{p,0.975}^2)$ , relying on the fact that the squared Mahalanobis distances are  $\chi_p^2$  distributed under the assumption of multivariate normal data. These weights can be used to compute a weighted mean, a weighted covariance matrix and corresponding Mahalanobis distances  $m(x_j; \bar{x}_w, \Sigma_w)$ .

For brevity we will refer to this specific distance using weighted estimators as the outlyingness of  $x_j$ , denoted  $o(x_j) = m(x_j; \bar{x}_w, \Sigma_w)$ .

Nowadays the MCD and many other robust algorithms can easily handle relatively large data sets with high dimension  $p$ . If an outlier is detected for such data, thus having large outlyingness, it is often of practical interest to find out more about this observation. For instance in genetics it is perfectly reasonable that an observation deviates from the majority of data points only for a few genes, not for all of them. Obviously, finding this subset of genes would be of high practical interest. This motivates the following problem: given multivariate data  $X$ , given weights as the result of a robust estimator, given that observation  $x_i$  is an outlier with large outlyingness  $o(x_i)$  and given  $1 \leq k < p$ , we want to find the subset of  $k$  variables contributing most to the outlyingness of  $x_i$ . A simple idea to find relevant variables is to check the univariate direction in which the observation is most outlying. If a coefficient of this direction is very large, it seems that the corresponding variable contributes a lot to the outlyingness. However two problems arise with this strategy. First this direction highly depends on the covariance structure of the regular data. Secondly the estimated coefficients turn out to be very unreliable in high dimensional situations. We propose a solution to both problems by computing an entire path of regularized directions of maximal outlyingness. To this end it is shown in Section 2 that the problem of estimating a direction of maximal outlyingness can be rewritten as a standard least squares regression problem. Replacing the latter by a LASSO (Tibshirani, 1995) type of penalized regression, we obtain a path of sparse directions of maximal outlyingness with non-zero coefficients only at those variables that are likely to contribute most to the outlyingness of an outlier. In Section 3 a specific algorithm is proposed to select a good  $k$ -subset of relevant variables using the LASSO path. Section 4 contains some examples.

## 2 Outlyingness as a regression problem

Let  $X = (x_1^t, \dots, x_n^t)^t$  be a sample of multivariate observations  $x_i \in \mathbb{R}^p$ . Denote  $w_i$  the weights with  $\sum_{i=1}^n w_i = n$ . Denote  $\bar{x}_w = \frac{1}{n} \sum_{i=1}^n w_i x_i$  the weighted multivariate mean. Denote  $X_w = (w_1(x_1^t - \bar{x}_w^t), \dots, w_n(x_n^t - \bar{x}_w^t))^t$  the weighted sample, centered around the weighted multivariate mean. Denote  $\Sigma_w = \frac{1}{n-1} X_w^t X_w$  the weighted covariance matrix. We define the outlyingness of  $x_j$  as  $o(x_j)^2 = (x_j - \bar{x}_w)^t \Sigma_w^{-1} (x_j - \bar{x}_w)$ . The following proposition is well known in the unweighted case and is easily shown to hold in the weighted case as well.

**Proposition 12.** *The outlyingness can be expressed as the solution of a maximization problem as follows:*

$$o(x_j) = \max_{a \in \mathbb{R}^p, \|a\|=1} \frac{|x_j^t a - \bar{x}_w^t a|}{\sqrt{a^t \Sigma_w a}}.$$



This can be interpreted as searching for the direction  $a$  such that the distance between the projected observation  $x_j^t a$  and the projected center  $\bar{x}_w^t a$ , standardized by a measure of spread of the projected observations  $\sqrt{a^t \Sigma_w a}$ , is maximal. We will call the direction for which the maximum in Proposition 1 is attained the direction of maximal outlyingness and denote it by  $a(x_j)$ . This direction of maximal outlyingness is potentially interesting, because its coefficients can learn us about the variables that are relevant in determining the outlyingness. It is an easy consequence of the Cauchy-Schwarz inequality that  $a(x_j) = \Sigma_w^{-1}(x_j - \bar{x}_w)$ , but it turns out that this direction of maximal outlyingness can also be expressed as a normalized least squares problem.

**Theorem 1.** Denote  $Y_w = e_j$  with  $e_j$  the  $j$ th basis vector in  $\mathbb{R}^n$  containing 1 at component  $j$  and 0 elsewhere. If all weights  $w_i > 0$ , then

$$a(x_j) = \frac{\theta}{\|\theta\|} \text{ with } \theta = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y_w - X_w \beta\|^2.$$

To perform variable selection we propose to not only consider the latter, but to consider an entire path of regularized directions of maximal outlyingness as follows.

**Definition 9.** A path of regularized directions of maximal outlyingness  $a(\lambda, x_j)$  is defined by

$$a(\lambda, x_j) = \frac{\theta(\lambda)}{\|\theta(\lambda)\|} \text{ with } \theta(\lambda) = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\{ \|Y_w - X_w \beta\|^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}.$$

By definition  $\theta(\lambda)$  is the solution of a least squares regression problem with an  $L_1$  penalty. Consequently the entire solution path for all  $\lambda$  can be computed in a very fast way using the LARS algorithm (Efron et al., 2004) with response vector  $Y_w$ , data matrix  $X_w$  and no intercept.

### 3 Selecting a $k$ -subset

Once the path  $a(\lambda, x_j)$  is obtained, we want to use those coefficients to select a subset of  $k$  variables contributing most to the outlyingness. There are two obvious approaches which we will call forward and backward selection. However, the best results are obtained when using both in a combined algorithm.

#### 3.1 Backward selection

Consider the original, unpenalized direction of maximal outlyingness  $a(0, x_j)$ . A very simple idea is to select the subset of  $k$  variables for which the components  $a_i(0, x_j)$  are largest in absolute value. Since we use the unconstrained solution  $a(0, x_j)$  and thus start from the full model, we will call this backward

selection in analogy with regression. As in regression however, this comes with some obvious drawbacks. If the dimension  $p$  is large in comparison to the sample size  $n$ , using all variables will lead to estimates with a very high variance. Therefore, the values of the components can be quite unreliable indications of the relevance of individual variables.

### 3.2 Forward selection

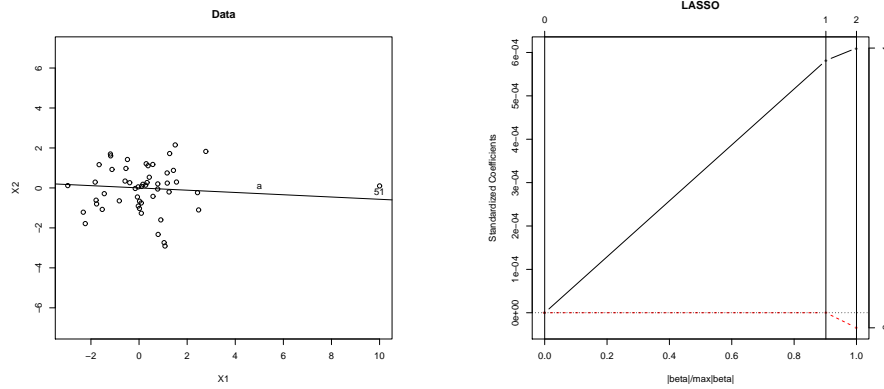
Since  $a(\lambda, x_j)$  is the normed solution of a  $L_1$  penalized regression problem, these vectors are generally sparser as  $\lambda$  increases. For large  $\lambda$  only a few coefficients of  $a(\lambda, x_j)$  are non-zero. Along the regularization path there is always at least one  $\lambda$  for which exactly  $k$  coefficients of  $a(x_j, \lambda)$  are non-zero. A possible strategy is to select the  $k$  variables corresponding to these non-zero coefficients. If there are multiple  $\lambda$ 's for which exactly  $k$  coefficients are non-zero, one can simply compute the outlyingness for every corresponding  $k$ -subset and select the one for which the outlyingness is largest. We will refer to this strategy as forward selection.

### 3.3 Combined algorithm

Both forward and backward selection can be successful in some situations, but fail in others. However, their behavior is rather complementary: if backward selection fails, forward selection often succeeds and vice versa. Therefore it turns out a good idea to compare both results if they return different  $k$ -subsets. Recall that our goal is to find a  $k$ -subset for which the outlyingness is largest. Thus if we have several different  $k$ -subsets, we can simply select the one corresponding to the largest outlyingness. We propose to create  $k+1$  potentially different  $k$ -subsets  $S_i$ , with  $S_0$  the  $k$ -subset obtained by forward selection,  $S_k$  by backward selection and  $S_i$ ,  $i = 1, \dots, k-1$  a mixture of variables from both.

- a. For  $i = 1, \dots, k$ , let  $B_i$  be the  $i$ -subset of  $i$  variables selected by backward selection and  $F_i$  the  $i$ -subset obtained by forward selection.
- b. Let  $S_0 = F_k$ .
- c. For  $i = 1, \dots, k$ , let  $S_i = \{F_{k-i} \cup B_i\}$ .
- d. For every of the  $k+1$  subsets of  $k$  variables, compute the classical unregularized outlyingness in the corresponding  $k$ -dimensional space. Retain the subset  $S_i$  for which the outlyingness of  $x_j$  is largest.

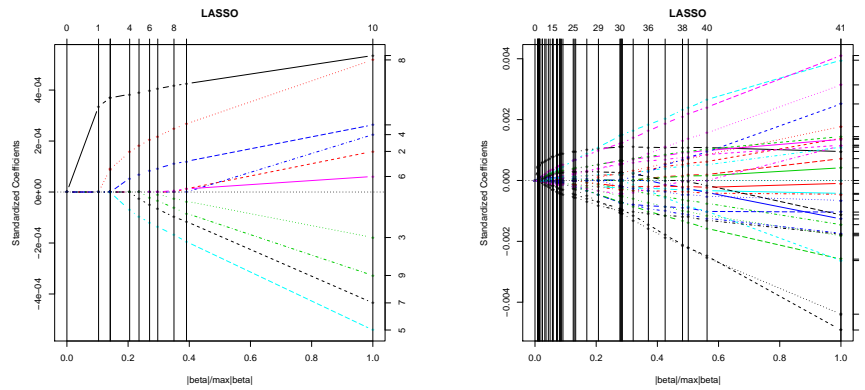
Note that the computational cost of this procedure is almost the same as the cost of running LARS once, since the sets  $B_i$ ,  $F_i$  and  $S_i$  can be determined immediately from the regularization path. Step 4 requires the computation of  $k+1$  outlyingnesses, which basically comes down to computing and inverting a  $k \times k$  weighted covariance matrix  $k+1$  times. Consequently the algorithm is very fast



**Fig. 1.** Example 1: data set on the left with unregularized direction of maximal outlyingness  $a$ . Right: corresponding LASSO path.

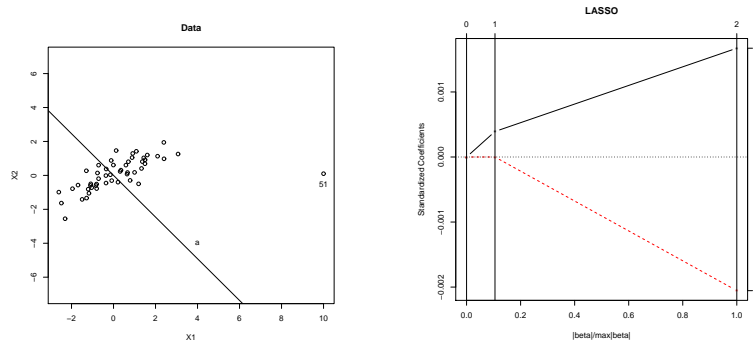
## 4 Examples

We consider some simple examples to illustrate the regularization paths of directions of maximal outlyingness. All examples use MCD estimators to obtain robust estimates of location and scatter and the corresponding weights. Figure 1 considers 50 points generated from a bivariate normal distribution with independent standard normal components. One outlier (51) is put at position (10, 0). By construction the first variable contributes most to the large outlyingness of 51. The classical direction of maximal outlyingness  $a$  equals (0.99, 0.14). Thus using backward selection yields the first variable. On the right hand side of Figure 1 the regularization path of the unnormalized solution  $\theta(\lambda)$  is plotted as defined in Definition 1. The plot is scaled in the default LASSO way, with the most regularized solution on the left and the unregularized solution on the right, thus  $\lambda$  decreases along the  $x$ -axis. For large  $\lambda$  a sparse solution is obtained with only the coefficient of the first variable non-zero. Consequently forward selection also selects the first variable as 1-subset. Next independent standard normal noise variables are added to this data set such that the dimension of the data is  $p > 2$ . By construction the first variable is still the only variable for which 51 is outlying. Figure 2 shows the regularization paths for  $p = 10$  on the left side and  $p = 30$  on



**Fig. 2.** Example 1: LASSO path when (left) 8, (right) 28 noise variables are added.

the right side. Clearly, the curse of dimensionality plays a nasty role now. When  $p = 10$ , the first variable is still the largest and would be selected by backward selection. However, the difference with variable 8 is rather small, although this is a pure noise variable. In the case  $p = 30$  backward selection completely fails. Plenty of noise variables have coefficients larger than the first variable. However the forward strategy still works perfectly. The sparsest solution has only one non-zero coefficient: the coefficient corresponding to the first variable. In Figure 3 a similar example is considered. On the left the data is shown which is generated from a bivariate normal distribution with highly correlated components. Again one outlier (51) is added at position (10, 0). On the right side of Figure 3 the corresponding regularization path is shown. Backward selection now (incorrectly) selects the second variable as 1-subset. This is due to the correlation between both variables, causing a rotation of the unregularized direction in comparison to the first example with independent components. However, forward selection still correctly returns the first variable, since the sparse solutions always have the first coefficient non zero. When noise variables are added (not shown here for brevity), forward selection remains a solid strategy returning the first variable as 1-subset and the first two variables as 2-subset. Backward selection on the other hand only returns noise variables not identifying the correct variables. Figure 4 shows an example where backward selection works, but forward se-

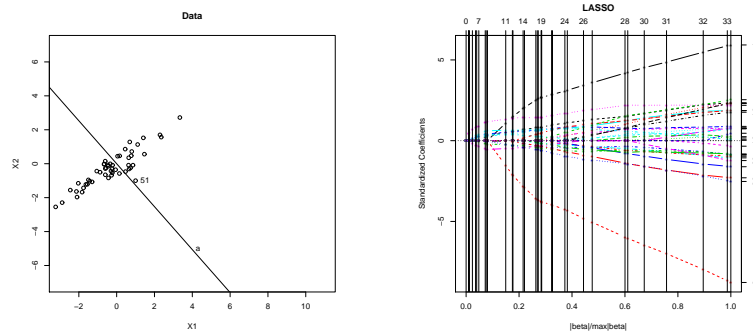


**Fig. 3.** Example 2: data set on the left with unregularized direction of maximal outlyingness  $a$ . Right: corresponding LASSO path.

lection fails. Again data are generated from a bivariate normal distribution with large correlation between both variables. One outlier (51) is added at position  $(1, -1)$ . Next 28 independent normally distributed noise variables are added such that the dimension equals  $p = 30$ . By construction the outlier is only outlying when the first two variables are both considered. When only the first or the second variable is considered separately, point 51 is not outlying. This explains the failure of forward selection. These regularization paths are essentially forward selection methods in the sense that as  $\lambda$  decreases, variables are generally added to the model one at a time. If two variables are only relevant when considered both at the same time, those variables will not necessarily be added first. This can be seen on the right side of Figure 4 where 6 noise variables are added to the model before the truly important ones 1 and 2 are added. However, once taken in the model, their coefficients rapidly increase. In the final unregularized solution backward selection does select the first two variables as 2-subset.

## 5 Conclusion

The problem of identifying relevant variables for outliers is posed. Regularization paths of directions of maximal outlyingness are defined. Three approaches are considered: forward selection, backward selection and a combination of both. Examples illustrate the differences between forward and



**Fig. 4.** Example 3: data set on the left with unregularized direction of maximal outlyingnes  $a$ . Right: corresponding LASSO path when 28 noise variables are added.

backward selection. More extensive simulation results (not reported here due to space constraints) confirm that the combined algorithm works well. Since the proposed methodology starts with weights obtained through a robust estimator such as MCD, it is of course important that those weights are reliable. This will be the case in classical contamination models where a majority of observations is fully uncontaminated, but might not hold in more complex contamination models affecting individual variables (Alqallaf et al., 2009). Extending the current framework to those models might provide an interesting future research topic.

## References

- ALQALLAF, F., VAN AELST, S., YOHAI, V.J. and ZAMAR, R.H. (2009): Propagation of outliers in multivariate data. *Annals of Statistics*, 37, 311-331.
- EFRON, B., HASTIE, T., JOHNSTONE, T. and TIBSHIRANI, R. (2004): Least Angle Regression. *Annals of Statistics*, 32, 407-499.
- LOPUHAÄ, H. P. (1991): Multivariate  $\tau$ -estimators for location and scatter. *The Canadian Journal of Statistics* 19, 307-321.
- ROUSSEEUW, P. J. (1984): Least median of squares regression. *Journal of the American Statistical Association* 79, 871-880.
- ROUSSEEUW, P. J. and LEROY, A. M. (1987): *Robust Regression and Outlier Detection*. Wiley, New York.
- TIBSHIRANI, R. (1996): Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58, 267-288.

# A New Approach to Robust Clustering in $\mathbb{R}^p$

Catherine Dehon<sup>1</sup> and Kaveh Vakili<sup>1</sup>

European Center for Advanced Research in Economics and Statistics.  
50, Avenue Roosevelt CP 114 Brussels, Belgium *kvakili@ulb.ac.be*

**Abstract.** In this note we present a fresh look at an old problem: that of identifying groups of cohesive observations lying in moderately large spaces when the dataset is potentially contaminated by an unknown number of outliers. The solution we introduce here is invariant to affine transformations, does not place assumptions on the number of clusters, the function governing their distribution or the share of contamination by outliers. Finally, our procedure is supported by a scalable, stable and efficient.

## 1 Introduction

Cluster analysis aims to partition a dataset into groups of cohesive observations disjoint from one another. Defining a cluster  $\zeta$  in  $p$ -dimensional space as an area of high object density surrounded by areas of lower object density, we present a procedure to identify the hyperplanes separating  $\zeta$  from the rest of the set. Our method belongs to the class of affine equivariant clustering algorithms designed to withstand the presence of outliers and noise in the dataset (for a related contribution, see García-Escudero et al. (2008).)

In our approach, clusters are identified sequentially and in a non arbitrary order: starting from an observations  $x_\phi$  located in an area of high local density, we find the coordinates of  $\kappa$  points  $W_i^{\tau*}$  such that the local density along a line joining any of the  $W_i^{\tau*}$  to  $x_\phi$  never falls below a multiple of that of the rest of the dataset. Finally, the convex hull of the  $W_i^{\tau*}$ , denoted  $CH(W_\kappa^{\tau*})$ , is our estimator of the boundary of  $\zeta$ .

The method we describe below is general by three aspects. First, it is robust to the presence of noise and outliers because observations separated from  $x_\phi$  by an area of low object density are not awarded any weight in the determination of  $CH(W_\kappa^{\tau*})$ . Second, it is model-free, meaning that we do not impose any assumption on the number of clusters or the functions governing their distributions, as long as they do not overlap. Third, it does not require the sample size to grow exponentially with  $p$  or the share of contamination by noisy and/or outlying points to be smaller than some preset value. Finally, by formulating the boundary of  $\zeta$  as the solution of a series of related linear optimization problems we ensures that repeated runs on a given dataset will essentially yield the same result and that our procedure is supported by a stable, scalable, efficient (i.e. with polynomial complexity) and parallel algorithm.

## 2 The focal point

Given a collection of independent  $p$ -variate random vectors  $X_n = (x_1, \dots, x_n) \in \mathbb{R}^{n \times p}$  lying in the so-called general configuration and a given value of  $d \in \mathbb{N}$ , the focal point is the observation with index  $\Phi^d$  defined by:

$$\Phi^d = \underset{i \in 1, \dots, n}{\operatorname{Argmin}} \max_{u_j \in \mathbb{R}^p: \|u_j\|=1} (\phi_{(i)j} - \phi_{(i-d)j}) \mathbf{I}(i > d) \quad (1)$$

where  $z_{(i)}$  denotes the vector of ordered values of  $z_i$ ,  $\Phi^d \in \{1, \dots, n\}$  and  $\phi_{ij} = x'_i u_j$ . In words,  $\phi_{(i)j} - \phi_{(i-d)j}$  is the distance between observation  $x_i$  and its  $d^{\text{th}}$  nearest neighbor in the direction  $u_j$  and  $\Phi^d$  is the index of the observation with minimal value of  $\phi_{ij}^d$  among all directions  $u_j$ 's. In practice, a finite but arbitrary large number of directions is found by re-sampling  $p$ -subsets of observations from  $X_n$ .  $d$  in equation (1) is (understood to be) a small number increasing with  $p$  (i.e.  $2p$  or  $3p$ ), as  $d$  large would compromise the required local character of the estimator whereas  $d = 1$  would imply too little a variation in the  $\Phi^d$  for any given  $i$ . To simplify notations a bit, we drop the  $d$  index to denote  $\Phi^d$  as  $\Phi$ , define  $X_{-\Phi} \in \mathbb{R}^{(n-1) \times p}$  the observations with indexes  $\{i : i \neq \Phi\}$  and the set  $S_0 = \{\Phi\}$ .

## 3 Path of expansion along $v_1$

The basic idea is to find an initial direction along which to expand the boundary of  $S_0$ . We define  $v_1$  as the direction given by  $x_\Phi$  and an observation drawn from  $X_n$  with weight vector  $\omega_i^0 = \mathbf{I}(i \notin S_0)$  and  $w(m_0, v_1) = \min_{i \in 1, \dots, n: v_1(x_i - x_\Phi) > 0} \{v_1 \|v_1\|^{-1} (x_i - x_\Phi)\} + x_\Phi$ . Then, we define:

$$\begin{aligned} & \min_{\gamma \in \mathbb{R}^{p+2}} (w(m_j, v_1), 1, 1)' \gamma \\ & \text{u.c. } (x_i, 1, 2)' \gamma \geq -\|x_i\|^2 \quad \forall i \neq \Phi \\ & (x_\Phi, 1, 1)' \gamma = -\|x_\Phi\|^2 \\ & (w(m_j, v_1), 1, 1)' \gamma \leq -\|w(m_j, v_1)\|^2. \end{aligned} \quad (2)$$

Equation (2) defines a  $p+2$  dimensional linear objective function subject to  $n+1$  linear inequalities and can be solved efficiently, for given values of  $(m_j, v_1)$ , by the simplex method, even for  $n, p$  large.

Set, initially,  $m_j = m_0$  and let  $\gamma(m_j, v_1)^*$  be the corresponding vector of solutions to equation (2). While a formal proof is beyond the scope of this note, it can be shown (see also Raković et al. (2004) for the derivation of a solution to a related problem) that provided that  $X_n$  lies in the general position, then,  $\|\gamma(m_j, v_1)^*\|$  is bounded and unique iff  $w(m_j, v_1) \in CH(X_n)$ .

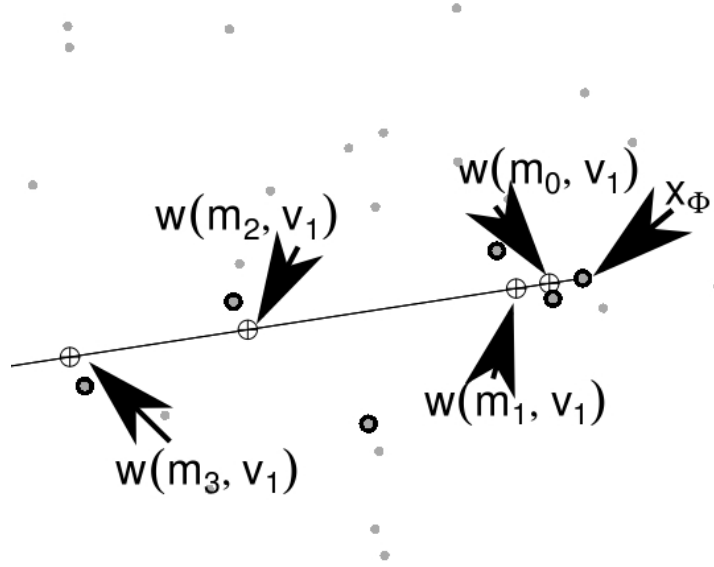
This in turn implies that any finite values  $\|\gamma(m_j, v_1)^*\|$  lies at the intersection of  $p+2$  of the  $n+1$  constraints of equation (2). Define  $t(m_j, v_1) = \{i :$



$(x_i, 1, 2)' \gamma(m_j, v_1)^* = 0\} \cup \Phi$ , then, by the mapping between the constraints in equation (2) and the observations,  $T(m_j, v_1) = \{x_i : i \in t(m_j, v_1)\}$  forms a triogram in the  $p$ -dimensional observation space. Then, it can be shown that  $T(m_j, v_1)$  encloses the point with coordinates  $w(m_j, v_1)$  and further that this triogram has smallest determinant among all triograms enclosing  $w(m_j, v_1)$  and having  $x_\Phi$  and  $p$  points from  $X_{-\Phi}$  as edges. Given  $t(m_j, v_1)$ , we define:

$$w(m_{j+1}, v_1) = \max_{k \in t(m_j, v_1)} \{v_1 \|v_1\|^{-1} (x_k - x_\Phi)\} + x_\Phi \quad (3)$$

That is, the largest scalar projection of an edge of  $T(m_j, v_1)$  unto  $v_1$ . It follows from equation (3) that  $w(m_{j+1}, v_1) \notin T(m_j, v_1)$  and also that  $\|w(m_{j+1}, v_1) - x_\Phi\| > \|w(m_j, v_1) - x_\Phi\|$ .



**Fig. 1.**  $\{w(m_j, v_1)\}_{j=0}^3$ ,  $x_\Phi$  and  $\{t(m_j, v_1)\}_{j=0}^3$  for a dataset of randomly generated values in  $\mathbb{R}^2$ .

A key feature of equation (2) is the relationship between  $\gamma(m_{j+1}, v_1)^*$  and  $\gamma(m_j, v_1)^*$ : given a finite solution  $\gamma(m_j, v_1)^*$  to equation (2) we either have that  $\gamma(m_{j+1}, v_1)^*$  is unbounded (iff  $w(m_{j+1}, v_1) \notin CH(X_n)$ ) or that  $\gamma(m_{j+1}, v_1)^*$  is adjacent to  $\gamma(m_j, v_1)^*$  (two simplex solutions are said to be adjacent if the line segment connecting them is an edge of the feasible region) because  $w(m_{j+1}, v_1)$  is immediately adjacent to  $T(m_j, v_1)$ . This feature of equation (2) in turn implies that  $\gamma(m_j, v_1)^*$  given  $\gamma(m_{j+1}, v_1)^*$  can be found at very little cost by parametric programming (that is by using  $\gamma(m_j, v_1)^*$  as a so-called warm start).

#### 4 Path of expansion along $v_i, i > 1$

By the arguments of the previous section, we have that  $\|\gamma(m_j, v_1)^*\|$  unbounded implies that  $\|\gamma(m_{j+1}, v_1)^*\| \forall m_{j+1} > m_j > 0$  will be unbounded as well and that  $\|\gamma(m_{j+1}, v_1)^*\|$  bounded implies that  $\|\gamma(m_j, v_1)^*\|$  will be bounded as well.

We define  $J(v_1)$  as the last value of  $j$  for which  $\|\gamma(m_j, v_1)^*\|$  is bounded along the direction  $v_1$ . For our initial direction  $v_1$ , we find the chain of indexes  $\{t(m_j, v_1)\}_{j=0}^{J(v_1)}$  by solving equation (2) recursively by parametric linear programming along the chain of values  $\{w(m_j, v_1)\}_{j=0}^{J(v_1)}$ . Then,  $S_1 = \cup_{j=0}^{J(v_1)} t(m_j, v_1)$  are the indexes of all observations used as edges of triograms enclosing the points  $\{w(m_j, v_1)\}_{j=0}^{J(v_1)}$  along the expansion path of  $x_\Phi$  in the direction  $v_1$ .

By way of illustration, the gray dots in Figure 1 depicts some dataset  $X_n$  lying in  $\mathbb{R}^2$ . The large black dot labeled  $x_\Phi$  is the focal point of this dataset and the 4 aligned crossed dots each depicts one of the chain of points with coordinates  $\{w(m_j, v_1)\}_{j=0}^3$ . These, in turn, yield (in that order) a chain of minimization problems each described by equation (2) with corresponding objective functions  $\{(w(m_j, v_1), 1, 1)\}_{j=0}^3$  for some direction  $v_1$ . Note that the chain of points with coordinates  $\{w(m_j, v_1)\}_{j=0}^3$  (and therefore the corresponding objective functions) are ordered: they are located monotonically further away from  $x_\Phi$  in the direction  $v_1$ . The remaining (unlabeled for clarity) black dots are points from the set with indexes  $\cup_{j=0}^{J(v_1)} t(m_j, v_1)$  obtained by recursively solving the linear program described by equation (2) for the chain of objective functions  $\{(w(m_j, v_1), 1, 1)\}_{j=0}^3$ . Finally, notice that  $J(v_1)$  for this combination of dataset  $X_n$ , direction  $v_1$  and focal point  $x_\Phi$  equals 2 since the point with coordinates  $w(m_3, v_1)$  lies outside the convex hull of  $X_n$  (implying that  $\|\gamma(m_3, v_1)^*\|$  is unbounded).

Given a chain of previous directions  $\{v_i\}_{i=1}^{l-1}$  and vector of weights  $\omega^{l-1} = \mathbf{I}(i \notin \cup_{i=0}^{l-1} S_i)$ , a new direction  $v_l$  is given by  $x_\Phi$  and an observation drawn from  $X_n$  with probability vector  $\omega^{l-1}$ . Just as before, this new direction is used to construct the chain of points  $\{w(m_j, v_l)\}_{j=1}^{J(v_l)}$  obtained by substituting  $v_1$  by  $v_l$  in the algorithm described in section 3 and yields a  $l^{th}$  chain of indexes  $\{t(m_j, v_l)\}_{j=0}^{J(v_l)}$  and a corresponding exclusion set  $S_l$ .

Each run of the algorithm described in section 3 increases the share of covered observations (defined as the members of the set  $\cup_{i=0}^{l-1} S_i$ .) A natural rule is to stop the exploration of  $X_n$  around  $x_\Phi$  along new directions  $v_i$  when the proportion of covered observations is larger than some preset value  $\alpha \in (0, 1]$ , formally:

$$\kappa = \min\{k : n^{-1} \# \cup_{i=1}^k \cup_{j=0}^{J(v_i)} t(m_j, v_i) = \alpha\}. \quad (4)$$

Note that for a given value of  $\alpha$ ,  $\kappa$  is an increasing function of both  $n$  and  $p$  (because the  $J(v_l)$  will tend to decrease as  $p$  increases.)

## 5 Pruning

$\kappa$  runs of the algorithm as described in section 3 yields  $\kappa$  chains of irregularly spaced points with coordinates  $\cup_{i=1}^{\kappa} \cup_{j=0}^{J(v_i)} w(m_j, v_i)$  forming a fan around  $x_{\Phi}$  each joining  $x_{\Phi}$  to a point  $w(m_{J(v_i)}, v_i)$  near  $CH(X_n)$ . Each of these fans corresponds to the expansion path of  $x_{\Phi}$  along a particular direction  $v_i, i \in \{1, \dots, \kappa\}$ . The next step consists in pruning each  $\{w(m_j, v_i)\}_{j=0}^{J(v_i)}$  at some index  $j(v_i)^*$ . For a given  $v_i, i \in \{1, \dots, \kappa\}$ , we define:

$$\begin{aligned} D(m_{j+1}, v_i) &= ||w(m_{j+1}, v_i) - x_{\Phi}|| - ||w(m_j, v_i) - x_{\Phi}|| \\ P(m_j, v_i) &= \max_{j \in \{1, \dots, j\}} D(m_j, v_i) \end{aligned} \quad (5)$$

Here,  $D(m_{j+1}, v_i)$  is a non parametric and affine equivariant proxy for the local density of  $X_n$  about  $w(m_{j+1}, v_i)$  while  $P(m_{j+1}, v_i)$  proxies for the size of the largest (local) density gap one has to cross to join  $x_{\Phi}$  and  $w(m_{j+1}, v_i)$ . Note that for a given direction  $v_i$ ,  $P(m_j, v_i)$  is a monotonously increasing, irregularly spaced staircase function of  $m_j$ . Figure 2 depicts values of  $\{w(m_j, v_i)\}_{j=0}^{J(v_i)}$  for a given dataset (left pan), the upper right pan depicts the corresponding values of  $||w(m_j, v_i) - x_{\Phi}||$  and the lower right pan depicts the corresponding values of  $\{D(m_j, v_i)\}_{j=1}^{J(v_i)}$  (dots) and  $\{P(m_j, v_i)\}_{j=1}^{J(v_i)}$  (line) all as a function of  $j$ . For each value of  $\tau_t$  along a grid of  $T$  regularly spaced values of  $\tau_t : \min P(m_j, v_i) < \{\tau_t\}_{t=1}^T < \max P(m_j, v_i)$ , we compute:

$$P_G(\tau_t) = \frac{\sum_{i=1}^{\kappa} \#\{j : P(m_j, v_i) \leq \tau_t\}}{\sum_{i=1}^{\kappa} J(v_i)} - \frac{\tau_t}{\max P(m_j, v_i)}. \quad (6)$$

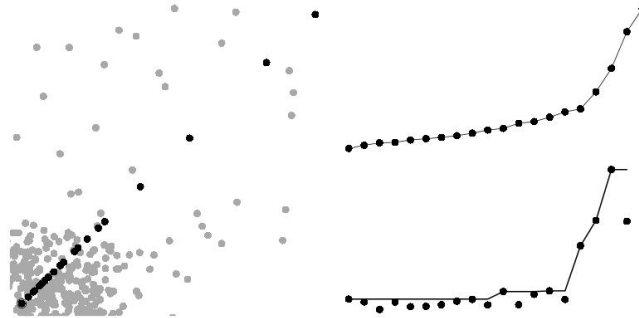
$P_G(\tau_t)$  is the Gini transform of the  $P(m_j, v_i)$  and proxies for the density of the set of observations located at most  $\tau_t$  away from  $x_{\Phi}$ . Defining  $\tau^* = \text{ArgMax } P_G(\tau_t)$  we set :

$$W_i^{\tau^*} = w(m_j, v_i) : \max_{1 \leq j \leq J(v_i)} \{j : P(m_j, v_i) \leq \tau^*\} \quad (7)$$

in words,  $W_i^{\tau^*}$  marks the maximal extension (in the direction  $v_i$ ) of an area enclosing  $x_{\Phi}$  where our local measure of density never falls below  $\tau^*$ . We denote  $W_{\kappa}^{\tau}$  the collection of all  $\kappa$  points  $W_i^{\tau^*}$ . Finally, note that the sharpness of  $P_G(\tau_t)$  around  $\tau^*$  will depend on the degree of separation between the eventual clusters.

## 6 Constructing the envelope of $\zeta$

At this stage, the result of the algorithm described in Sections 2-5 is a matrix  $W_\kappa^{\tau*} \in \mathbb{R}^{\kappa \times p}$  of coordinates where  $W_i^{\tau*}$  is the maximal extension meeting our density restriction along the expansion path of  $x_\Phi$  in the direction  $v_i$ . The last step of the algorithm naturally consists in using these points to build a convex envelope around  $x_\Phi$  separating a zone of high density around  $x_\Phi$  from the rest of the set. When it can be assumed that the eventual clusters are not overlapping (i.e. the clusters are disjoint from one another), then, the boundary of the cluster of observations to which  $x_\Phi$  belongs, (which we denote  $\zeta$ ), is well approximated by the largest convex polytope with edges in  $W_\kappa^{\tau*}$ . The convex hull of  $W_\kappa^{\tau*}$  is therefore a natural and affine equivariant (Zuo and Serfling (2000)) estimator of the boundary of  $\zeta$ . When  $p \leq 8$ , we can use the classical algorithm of Barber et al. (1996) to compute the convex hull of  $W_\kappa^{\tau*}$  with general time and memory complexity of  $O(\kappa^{\frac{p}{2}})$ . When  $p > 8$ , computing the exact convex hull of  $W_\kappa^{\tau*}$  becomes computationally intractable. Vakili (2009) proposes an algorithm to approximate the convex hull of a set of point in  $\mathbb{R}^p$  in polynomial time. A full description of this procedure would, however, far exceed the scope of this note. In essence, his algorithm is based on the resolution of a series of linear programs whose structure is very similar to that of equation (2).



**Fig. 2.** Values of  $\|w(m_j, v_i) - x_\Phi\|$  (top right pan),  $P(m_j, v_i)$ ,  $D(m_j, v_i)$  (bottom right pan) for the for the points  $\{w(m_j, v_1)\}_{j=0}^{20}$  shown in the left pan.

## 7 Application

To give an idea of the stability of the new algorithm, we used the fruit dataset (Hubert and Van Driessen (2004).) This dataset has  $n = 2818$  observations and 256 dimensions. For our purpose we use the coordinates obtained from the projection on there first  $p = 2$  principal components as our  $X_n$ . The upper

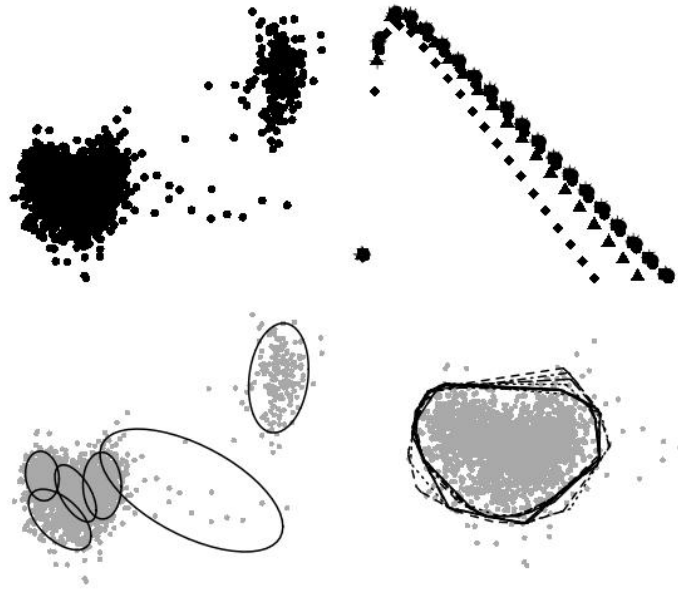
left corner of Figure 3 shows  $X_n$ , where two separated clusters and a trail of outliers are clearly visible. The dark gray scatter-plot in the lower right corner in Figure 3 magnifies the cloud of points around the focal point of this dataset (found using 500  $p$ -subsets and  $d = 4$ .) The series of dashed lines in this image mark our estimations of  $CH(\zeta)$  obtained from repeated runs of the algorithm which we ran 10 times to outline the stability of the solutions. Each run took about 10 seconds on a recent quad-core desktop computer running Linux OS (in our implementation, steps 2-4 are carried in parallel over all cores.) Because our procedure only uses the focal point as an initial starting point, the end result does not require  $x_\phi$  to be particularly close to the mode of any of the clusters. The upper right plot shows the values of  $P_G(\tau_t)$  for a grid of  $T = 20$  obtained from the repeated runs of the algorithm. Note that although the focal points were generally different across runs, the location of  $\tau^*$  remains relatively stable around 0.2. Finally for comparison purposes, the sub-plot on the lower right corner of Figure 3 shows the clustering found by the **R** implementation of the MClust algorithm of Fraley and Raftery (2002), using the best solution (lowest BIC value of  $-13143$ ) out of 10 runs. Each run of MClust took over 5 minutes (the algorithm is not parallelizable and uses only one of the 4 cores). Notice that MClust partitions  $\zeta$  into four clusters, due to the fact that observations inside  $\zeta$  follow an asymmetric distribution and are therefore difficult to parsimoniously cluster with ellipses.

## 8 Conclusions

The major limitation of our procedure is the assumption that the clusters are disjoint: the last step of our algorithm will bundle overlapping clusters together, which we define as at least two area of high object density not separated by areas of lower object density. This limitation stems from the absence of restrictions on the shape of the clusters and the subsequent difficulty in finding an alternative definition of what a cluster is. Still, even with this limitation, our approach offers the promise of an alternative to existing algorithm from the robust literature such as the Fast-MCD algorithm (Rousseeuw and Van Driessen (1999)) for cases involving datasets of moderately large dimensions where the user is not willing to assume that a known density function governs the distribution of a majority of observations or from the clustering literature for cases where the clusters are more likely to be separated than elliptical.

## 9 Acknowledgments

The authors are grateful to N. Gothelf, M. Gassner and V. Verardi (ULB) M. Hubert and P.J. Rousseeuw (KUL) for there numerous suggestions, the participants to the doctoral seminars at the ULB and KUL as well as two anonymous referees for there helpful comments.



**Fig. 3.** First two principal components (top left), area around the focal point (bottom right) best MClust partition (bottom left) and values of  $P_G(\tau_l)$  for the fruit dataset.

## References

- BARBER, C.B., DOBKIN, D.P., and HUHDANPAA, H.T. (1996) The Quickhull algorithm for convex hulls. *ACM Trans. on Mathematical Software*.
- FRALEY, C. and RAFTERY, A. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, Vol. 97, No. 458.
- GARCÍA-ESCUADERO, L. A., GORDALIZA, A., MATRÁN, C. and MAYO-ISCAR, A. (2008) A General Trimming Approach to Robust Cluster Analysis. *The Annals of Statistics*, Vol. 36, No. 3.
- HILLIER, F. S. and LIBERMAN, G. J. (2001) Introduction to Operations Research, 7<sup>th</sup> edition. *McGraw-Hill*.
- HUBERT, M. and VAN DRIESSEN, K. (2004) Fast and Robust Discriminant Analysis. *Comput. Statist. Data Anal.*, Vol. 45.
- RAKOVIĆ, S.V., GRIEDER, P and JONES, C. (2004). Computation of Voronoi diagrams and Delaunay triangulation via parametric linear programming. *ETH Technical Report AUT04-03*.
- ROUSSEEUW, P.J. and VAN DRIESSEN, K. (1999) A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, Vol. 41, No. 3.
- VAKILI, K. (2009) *Fast mixture partitioning in  $\mathbb{R}^p$* . unpublished Master thesis, Université Libre de Bruxelles.
- ZUO, Y. and SERFLING, R. (2000) General Notions of Statistical Depth Function *The Annals of Statistics*, Vol. 28, No. 2.

# An Exploratory Segmentation Method for Time Series

Christian Derquenne

Electricité de France R&D

1, avenue du Général de Gaulle, Clamart, France, *christian.derquenne@edf.fr*

**Abstract.** The method proposes to segment a time series. It offers an original process with a first step of preparing data which is crucial to build the most adequate structure to initialize the second step of modelling an heteroskedastic linear model including the different trends, levels and variances. This method can be used in a lot of domains and to set up several objectives. Building of sub-models on each detected segment, achieving stationarity of time series with a segmentation model, building symbolic curves to cluster series, modelling multivariate time series are so many examples in this context.

**Keywords:** segmentation, change-point, time series, variance components

## 1 Issues and motivations

Generally, time series are decomposed in several behaviours: trend, seasonality, volatility and noise. Due to the regularity of the series, it could be more or less easy to decompose them as this scheme. For instance, observed hourly temperature series in France on one year seems to a sinusoidal curve which allows to estimate a reference temperature on 100 years. In this case, residual will correspond to an approximation of noise. However, irregular phenomena exist, in the sense that they are difficult to forecast, as financial series: energy market prices, CAC40, etc. These have volatility which can be extracted with difference log ratio and trend and seasonality occur less frequently and less regularly. Behaviours breaks could characterize series. For example, these changes can be peaks (high energy price on a very short period), level breaks, trend changes (impact of a rules change) or in volatility (FTSE 100 ratio). The data modelling is very delicate, as it asks many experience in the application domain and in Statistics. In a lot of cases, forecast these series can be close to an utopian view. Then it can be interesting to detect behaviour breakpoints when pre or post-treating data. Building of sub-models on each detected segment, achieving stationarity of time series with a segmentation model, building symbolic curves to cluster series, modelling multivariate time series are so many examples in this context. Many methods have been and are developed to answer different issues in economics, finance, human sequence, meteorology, energy management, etc. Several methods exist: from exploring the segmentation space for the assessment of multiple change-point models

[Guédon, 2008], to inference on the models with multiple breakpoints in multivariate time series, notably to select optimal number of breakpoints [Lavielle et al., 2006], testing for multiple structural changes in cointegrated regression models [Perron et al., 2008], sequential change-point detection when the pre- and post-change parameters are unknown [Lai et al. 2009] and on-line detection of breakdown to build meta-model coming from weighting of predictors. Most algorithms use dynamic programming to decrease computation complexity of segmentations, because it would be illusory to calculate all segmentations. Indeed, the number of segmentations for a series with length  $T$  and a number  $S$  of segments is  $\binom{T-1}{S-1}$  whereas for overall segments  $s = 1, T$ , the total number of segmentations is  $2^{T-1}$ . For instance, in case of exploring the segmentation space, complexity is generally in  $O(ST^2)$  for the time and in  $O(ST)$  for the linear clustered space whose complexity increases with the length of series. But, this complexity can be decrease in  $O(T^2)$  [Lavielle et al., 2006], even in the frame of multiple breakpoints for  $M$  multivariate time series, whereas it could be in  $O(MT^2)$ . These methods of breakpoints detection aims at answering three detection problems [Lavielle et al., 2006] : change mean with a constant variance, change of variance with a constant mean and change for overall distribution of time series without change of level, in dispersion and on the distribution of errors. Furthermore this method proposes to resolve a fourth problem : detection of trend [Perron et al., 2008], but also to reduce the computation complexity in  $O(KT)$ , where  $K$  is the smoothing degree, which is generally less than to  $\sqrt{T}$ . Then our method offers solutions of segmentation containing segments with increasing or decreasing trend, constant level and different standard-deviations. Lastly, it is very important to keep in mind, this method is not a replacement of classical time series modelling as ARMA.

## 2 The proposed method

### 2.1 The model and its inference

Let's  $(Y_t)_{t=1,T}$  be a time series, we suppose that it is decomposed in accordance with an heteroskedastic linear model (or variance components) [Rao et al., 1988, Searle et al, 1992] as follows :

$$Y_t = \sum_{s=1}^S (\beta_0^{(s)} + \beta_1^{(s)}t + \sigma_s \epsilon_t) 1_{[t \in \tau_s]} \quad (1)$$

where  $\beta_0^{(s)}$ ,  $\beta_1^{(s)}$  and  $\sigma_s > 0$  are respectively the level, trend and standard-deviation parameters for each segment  $\tau_s$ , and  $\epsilon_t$  from a standard normal distribution. For each segment, there are  $T_s$  observations with  $\sum_{s=1}^S T_s = T$ . Each segment contains values  $y_t$  for  $t = U_{s-1} + 1$  to  $U_s$ , where  $U_s = U_{s-1} + T_s$  and  $U_S = T$ .  $3S$  parameters need to be estimated, but the number  $S$  of segments is unknown. In case of heteroskedastic linear model, several estimators



are available : Ordinary Least Squares (OLS), Maximum Likelihood (ML) and Restricted or Residual Maximum Likelihood (REML) [Bartlett, 1937, Harville, 1977].

These three estimators give the same solutions for the  $\beta_0^{(s)}$  and  $\beta_1^{(s)}$ . But only ML and REML allow to estimate  $\sigma_s$ . However, these parameters can be obtained a posteriori with OLS, such as :

$$\hat{\sigma}_s = \sqrt{\sum_{t \in \tau_s} (y_t - \hat{\beta}_0^{(s)} - \hat{\beta}_1^{(s)}t)^2 / (T - 2)} \quad (2)$$

which is an unbiased estimator of  $\sigma_s$  for  $s = 1, S$ .

The advantage of REML on ML is that it provides unbiased estimators of variance and covariance components which correspond to (2).

Concerning statistical test associated to parameters, ML and REML allow to make an inference taking into account the variance components. But, standard-deviation errors of  $(\beta_0^{(s)}, \beta_1^{(s)})$  of the  $t$ -statistics coming from REML and ML are different. For the first one, we have :

$$\hat{\sigma}^{(REML)}(\hat{\beta}_0^{(s)}) = \hat{\sigma}_s^{(REML)} \sqrt{1/T_s + \bar{t}_s^2 / \sum_{t \in \tau_s} (t - \bar{t}_s)^2} \quad (3)$$

$$\hat{\sigma}^{(REML)}(\hat{\beta}_1^{(s)}) = \hat{\sigma}_s^{(REML)} / \sqrt{\sum_{t \in \tau_s} (t - \bar{t}_s)^2} \quad (4)$$

where  $\bar{t}_s$  is mean of  $t \in \tau_s$  and  $\hat{\sigma}_s^{(REML)}$  corresponds to (2) unbiased estimator of  $\sigma_s$ , whereas for ML  $\hat{\sigma}_s^{(REML)}$  is replaced by  $\hat{\sigma}_s^{(ML)}$ , biased estimator of  $\sigma_s$ . Of course the model (1) is statistically correct only if null hypothesis of homoskedasticity is rejected. The statistic used corresponds to two times of log-likelihood with and without variance components in the linear model. Under null hypothesis, this statistic is distributed as a  $\chi_{S-1}^2$ . If the model is homoskedastic, then for  $s = 1, S$ ,  $\sigma_s = \sigma$ , estimated by  $\hat{\sigma}$ .

## 2.2 The overall process of segmentation

The proposed method is mainly original in its process step by step to provide a decision aid for data segmentation. Indeed, statistical tools used for modelling are classical, but necessary for the method. The overall process of segmentation includes two steps. The first one consists in preparing data and the second one in a successive and adaptive modelling iterations.

The **data preparing** is as follows. First, it consists in smoothing the series to eliminate data impurity. Then it is necessary to differentiate smoothed data and given smoothing degrees to count the number of positive, negative or null values. These series will constitute initial segments. The smaller smoothing

degrees is, the higher the number of segments is. These initial segments will be used to estimate by REML an initial model (1), final result of the data preparing. Given this first model, it will be possible to begin the **data modelling** step of an iterative process. As a conclusion of homoskedasticity test,  $t$ -tests on trend and intercept parameters are different. Their results will allow to build a simplified model in which some  $\beta_1^{(s)}$  trend parameters could be put to zero. But this simplified model will have probably too much segments. We can repeat this simplification until the number of segments is satisfactory. In state of the work, the precise convergence criteria is empirical (four times). However, this one is the result for one smoothing degree. So, the global process in two steps is repeated for several smoothing degrees. The maximum smoothing degree is  $T$ , but in practice it is less than  $\sqrt{T}$ . The empirical and theoretical complexities are respectively in  $O(T\sqrt{T})$  and in  $O(T^2)$ . Lastly an assessment step allows to evaluate the proposed segmentation.

### 2.3 The detailed process of segmentation

Both steps are repeated for a set of smoothing degrees.

#### 2.3.1 Preparing data

The **step of smoothing** aims at simplifying time series to keep only significant and robust trends, to prepare the data for step of differentiation that follows. To smooth data, we have chosen moving median because it is more robust than moving average. Let us note  $j$  a smoothing degree corresponding to the number of observations included in moving median  $m_j(t)$  for  $t = 1, T - j$ . Theoretically,  $j = 1, T$ , but in practice  $j \leq \sqrt{T}$ . We have:

$$m_j(t) = \underset{t \in [a_j(t), b_j(t)]}{med} (y_t) \quad (5)$$

where for  $j$ :  $a_j(t) = t$  and  $b_j(t) = t + j - 1$ , with  $t = 1, T - j + 1$ .

The more  $j$  increases, the less irregularity of data is taken into account. The **step of differentiation** allows to detect the trends of smoothed data. This differentiation must be sufficiently high to reveal trend deviations, but not too much otherwise it could be skipped. Then, we have taken into account the property of moving median in calculating a difference at time  $t$  and at time  $k = t - j/2$ , if  $j$  is even and  $k = t - (j + 1)/2$ , if  $j$  is odd. We have:

$$d_j(t) = (m_j(t) - m_j(t - k)) / m_j(t - k) \quad (6)$$

The denominator allows to obtain a relative deviation. It is very useful to have comparable quantities. But, it is only a visual choice and not a theoretical choice. The step of differentiation has allowed to build a serie of positive, negative or null differences. The **step of counting** the number of values with the same sign is reasonably linked to the smoothing degree. Indeed, the smaller smoothing degrees is, the smaller size of series of differences with

same sign is. Each serie will correspond to an initial segment. The first segment  $\tau_{j,1}^{(0)}$  for a smoothing degree  $j$  will contain the  $T_{j,1}^{(0)}$  observations having the same sign, the second segment  $\tau_{j,2}^{(0)}$  will include the  $T_{j,2}^{(0)}$  observations having the same sign, but different to this  $\tau_{j,1}^{(0)}$ , etc. At the end of this process, we will obtain a vector of segments  $(\tau_{j,1}^{(0)}, \dots, \tau_{j,s}^{(0)}, \dots, \tau_{j,S}^{(0)})$ , with the size  $(T_{j,1}^{(0)}, \dots, T_{j,s}^{(0)}, \dots, T_{j,S}^{(0)})$  and  $\sum_{s=1}^S T_{j,s}^{(0)} = T$ .

### 2.3.2 Modelling

#### *Initial step of modelling*

There are several successive steps of modelling. The initial model found at the end of the first step has probably too much segments. The goal is so to simplify the proposed initial model. It contains several sub-steps of modelling. Let us indicate that the estimated coefficients  $(\beta_0, \beta_1, \sigma)$  are different at each sub-step. The full model for a smoothing degree  $j$  has the following form (1):

$$Y_t = \sum_{s=1}^S (\beta_0^{(j,s)} + \beta_1^{(j,s)} t + \sigma_{j,s} \epsilon_t) 1_{[t \in \tau_{j,s}^{(0)}]} \quad (7)$$

This model is estimated by REML, test of homoskedasticity is applied, and if the null hypothesis of constant variance is not rejected, then the following model is estimated:

$$Y_t = \sum_{s=1}^S (\beta_0^{(j,s)} + \beta_1^{(j,s)} t) 1_{[t \in \tau_{j,s}^{(0)}]} + \sigma_j \epsilon_t \quad (8)$$

The structure of a simplified model is built with  $S$  tests on the coefficients  $\beta_1^{(j,s)} = 0$  on the full model. The  $t$ -statistic is used:  $\hat{\beta}_1^{(j,s)} / \hat{\sigma}^{REML}(\hat{\beta}_1^{(j,s)})$ , for the heteroskedastic model, otherwise standard-deviation of errors is replaced by  $\hat{\sigma}^{OLS}(\hat{\beta}_1^{(j,s)})$ . Then the simplified model is the following:

$$Y_t = \sum_{s=1}^S (\beta_0^{(j,s)} + \beta_1^{(j,s)} t 1_{[\beta_1^{(j,s)} \neq 0]} + \sigma_{j,s} \epsilon_t) 1_{[t \in \tau_{j,s}^{(0)}]} \quad (9)$$

This model is simplified again to obtain a model with aggregated segments. But, only consecutive segments are regrouped:  $\tau_{j,s}^{(0)}$  and  $\tau_{j,s+1}^{(0)}$  are compared in order to merge them, if they seem statistically valid. Each segment is structured by three parameters:  $(\beta_0^{(j,s)}, \beta_1^{(j,s)}, \sigma_{j,s})$ , if the model is heteroskedastic or by  $(\beta_0^{(j,s)}, \beta_1^{(j,s)}, \sigma_j)$ , if the model is homoskedastic.

So we must first test homoskedasticity on following model:

$$Y_t = (\beta_0^{(j,s)} + \beta_1^{(j,s)} t + \sigma_{j,s} \epsilon_t) 1_{[t \in \tau_{j,s}^{(0)}]} + (\beta_0^{(j,s+1)} + \beta_1^{(j,s+1)} t + \sigma_{j,s+1} \epsilon_t) 1_{[t \in \tau_{j,s+1}^{(0)}]} \quad (10)$$

If the variances are equal and if the two parameters  $\beta_1^{(j,s)}$  and  $\beta_1^{(j,s+1)}$  are different from zero, then we use a  $t$ -test to compare them, such as:

$$|\hat{\beta}_1^{(j,s)} - \hat{\beta}_1^{(j,s+1)}| / \hat{\sigma}^{OLS}(\hat{\beta}_1^{(j,s)}, \hat{\beta}_1^{(j,s+1)}) \quad (11)$$

where

$$\hat{\sigma}^{OLS}(\hat{\beta}_1^{(j,s)}, \hat{\beta}_1^{(j,s+1)}) = \hat{\sigma}_{j,s+1} \sqrt{1 / \sum_{t \in \tau_{j,s}^{(0)}} (t - \bar{t}_{j,s})^2 + 1 / \sum_{t \in \tau_{j,s+1}^{(0)}} (t - \bar{t}_{j,s+1})^2}$$

with  $\hat{\sigma}_{j,s+1}^2 = (T_{j,s}^{(0)} \hat{\sigma}_{j,s}^2 + T_{j,s+1}^{(0)} \hat{\sigma}_{j,s+1}^2) / (T_{j,s}^{(0)} + T_{j,s+1}^{(0)})$ . If these two coefficients are equal, then we use a  $t$ -test to compare  $\beta_0^{(j,s)}$  and  $\beta_0^{(j,s+1)}$  with the same principle. If these last ones are equal, then  $\tau_{j,s}^{(0)}$  and  $\tau_{j,s+1}^{(0)}$  are merged. At the end of this process, the number of groups obtained  $S_1 \leq S$  will correspond to the new segments  $(\tau_{j,1}^{(1)}, \dots, \tau_{j,s}^{(1)}, \dots, \tau_{j,S_1}^{(1)})$ , to include in model with regrouped segments, such as:

$$Y_t = \sum_{s=1}^{S_1} (\beta_0^{(j,s)} + \beta_1^{(j,s)} t + \sigma_{j,s} \epsilon_t) 1_{[t \in \tau_{j,s}^{(1)}]} \quad (12)$$

### ***Further steps of modelling***

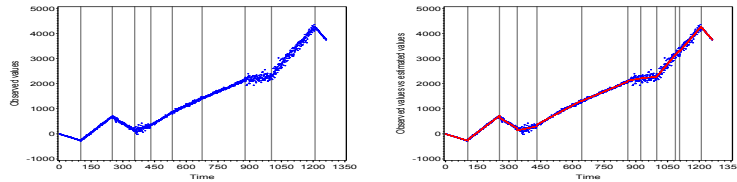
In the second step, (12) is submitted to the same process of successive tests as presented previously. We can repeat this simplification until the number of segments is satisfactory. In state of the work, the precise convergence criteria is empirical (four times).

### **2.3.3 Models assessment**

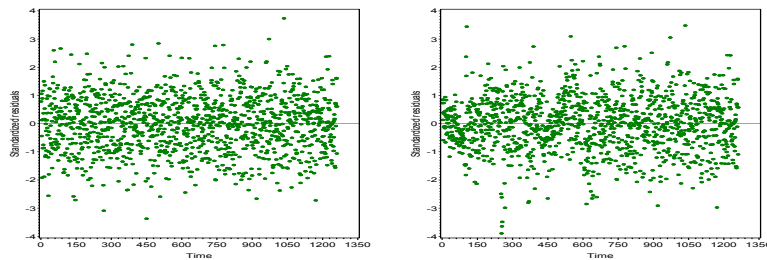
The both stages described previously are repeated several times in accordance with the smoothing degree chosen by the user. It is natural to think that for some smoothing degrees, a final model will provide a good reconstitution of data which would correspond to obtain a good segmentation. Even if the  $T$  smoothing degrees are tried, the optimal segmentation is not guaranteed with a probability equal to one, but the goal of this method is not this one. Indeed, as the model is complex because it allows to reveal the trends, levels and variances in time series, the aim is to propose some interesting segmentations, in terms of decision aid. To evaluate each final model and to offer some possible segmentations, we have chosen the value of REML which has allowed to estimate the model and the MAPE (average of absolute relative errors). Then the smaller values of these last ones are preferred to decide the quality level of the segmentation. These measures are heuristic choices because they can have an impact in the process of segmentation, notably to select one or several uninteresting segmentations.

### 3 Application

We have applied this method on simulated data. The time series is simulated on 10 segments, in accordance with model (1). For each segment, number observations, coefficients  $\beta_0$  and  $\beta_1$ , and  $\sigma$  are generated. To evaluate the quality of each estimated segmentation with respect to simulated segmentation, we compare the distributions of simulated and estimated segments with aid of correlation measures: Cramer's  $V$ , Kendall's  $\tau_b$ , Stuart's  $\tau_c$  and the percent of missclassified. For the 10 best segmentations in accordance with different criteria introduced previously, 11 smoothing degrees appear the most appropriate, because it is the only one to obtain good scores in REML, MAPE=9.61% and 87% of errors are less than 10%. In addition,  $V = 0.88$ ,  $\tau_b = 0.97$ ,  $\tau_c = 0.95$  and the missclassified rate is equal to 15%. This estimated segmentation contains 12 segments (fig. 1(b)) *vs* 10 segments for the real segmentation (fig. 1(a)). The figure 1(b) shows a very good fitness between estimated segmentation and simulated data. Indeed, the real and estimated breakpoints are closed. On the other hand, the standardized residuals coming from real and estimated segmentation (fig. 2(a) & 2(b)) have very near behaviours. This last result shows the interest of this method to achieve stationarity a time series.



**Fig. 1. (a) Observed segmentation – Fig. 1. (b) Estimated segmentation**



**Fig. 2. (a) Observed errors — Fig. 2. (b) Estimated residuals**

## 4 Contributions, applications and further researches

The proposed method allows to segment a time series. It offers an original process containing a stage of preparing data which is essential to build the most adequate structure to initialize stage of modelling, in accordance with an heteroskedastic linear model including the different trends, levels and variances. The goal of this method is not to provide the optimal segmentation as the majority of the methods discussed in introduction, but to provide a decision aid. Indeed, even if the minimum complexity of the other methods is in  $O(T^2)$ , it stays high, however. The method introduced in this paper uses only assessment criteria, such as values of REML, MAPE and percentage of relative errors less than 10%. But its complexity is in  $O(T)$  for each smoothing degree and the number of this last one is rarely greater than  $\sqrt{T}$ . Indeed, for high smoothing degree, the quality of segmentations decreases rapidly, because they move away optimality, even if this one is empirical. As said previously, this method can be used in a lot of domains of application and for a lot of objectives: searching of segments, achieving stationarity, building of different models on a same time series having different behaviours, simplifying (symbolic approach) of several time series to make clustering of curves, etc. Lastly, this method is rather preliminary and we work to improve some steps of this method, particularly on the detection of volatility in data and on the evaluation and the validation tools of segmentations to obtain a better means to have a hierarchy of these last ones.

## References

- BARTLETT, M.S. (1937): Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, Series A* 160, 268-282.
- GUEDON, Y. (2008): Exploring the segmentation space for the assessment of multiple change-point models. Institut National de Recherche en Informatique et en Automatique, *Cahier de recherche* 6619.
- HARVILLE, DA. (1977): Maximum likelihood approaches to variance component estimation and to related problems. *J Amer Stat Assoc* 72, 320-340.
- LAI, TL. and XING, H. (2009): Sequential Change-point Detection when the pre- and post-change parameters are unknown. *Technical report 2009-5*, Stanford University, Department of Statistics.
- LAVIELLE, M. and TEYSSIERRE, G. (2006): Détection de ruptures multiples dans des séries temporelles multivariées. *Lietuvos Matematikos Rinkinyys*, Vol 46.
- PERRON, P. and KEJRIWAL, M. (2006): Testing for Multiple Structural Changes in Cointegrated Regression Models. Boston University, C22.
- RAO, CR. and KLEFFE, J. (1988): *Estimation of variance components and applications*. North Holland series in statistics and probability, Elsevier.
- SEARLE, SR., CASELLA, G. and Mc CULLOCH, CE. (1992): *Variance components*. Wiley & sons, New-York.

# Using Auxiliary Information Under a Generic Sampling Design

Giancarlo Diana<sup>1</sup> and Pier Francesco Perri<sup>2</sup>

<sup>1</sup> Department of Statistical Sciences, University of Padova  
Via Cesare Battisti, 241, 35121 Padova, Italy, *giancarlo.diana@unipd.it*

<sup>2</sup> Department of Economics and Statistics, University of Calabria  
Via P. Bucci, 87036 Arcavacata di Rende, Italy, *pierfrancesco.perri@unical.it*

**Abstract.** Under a generic sampling design we consider a class of asymptotically unbiased estimators for the mean of a finite population when auxiliary information is available. After obtaining the minimum variance bound of the class we find that the regression estimator is the best one to use and that recent estimators proposed by Bacanli and Kadilar (2008) are unsatisfactory solutions. Finally, we carry out a simulation study under unequal probability sampling - with estimated first and second order inclusion probabilities - to shed light on the performance of different estimators.

**Keywords:** Horvitz-Thompson estimator, estimated inclusion probabilities, pps sampling, simulation

## 1 Introduction

The estimation of population parameters is a persistent issue in sampling from finite population when auxiliary information is available. Many efforts have been made to estimate the mean (or total) through the *ratio*, *product* and *regression* estimation methods and a great deal of literature has been produced according to simple random sampling without replacement (srswor). Among others, here we mention Kadilar and Cingi (2006) who introduced a class of estimators for the population mean starting from some ratio-type estimators discussed in Sisodia and Dwivedi (1981), Singh and Kakran (1993) and Upadhyaya and Singh (1999). Recently, Bacanli and Kadilar (2008) (BK hereafter) extended the aforementioned ratio-type estimators to the probability proportional to size (pps) sampling by means of the Horvitz-Thompson estimator. The new estimators have also been analytically and numerically compared with the ratio estimator, neglecting the regression estimator which is the best estimator in situations such as those examined.

Motivated by this recent work we aim to provide some guidelines in the efficient use of the auxiliary information. In Section 2, we consider - under a generic sampling design - a simple class of estimators for the population mean which includes, among others, BK proposals. We prove that the best

estimator in the class is the regression estimator. In Section 3, some theoretical efficiency considerations are drawn and it is pointed out that BK estimators are not optimum in the class. Section 4 is devoted to a simulation study in order to compare different estimators when their mean square error is computed with estimated inclusion probabilities. Some final remarks conclude the work.

## 2 The suggested class of estimators

Consider a finite population  $U = \{1, \dots, N\}$  from which a sample  $s$  of size  $n$  is selected according to a generic sampling design  $p(s)$ . Let  $(x_i, y_i)$  be the values of the auxiliary characteristic  $x$  and the study variable  $y$  for the  $i$ -th population unit,  $i = 1, \dots, N$ . Let us denote by  $\bar{X}$  and  $\bar{Y}$  the population means of  $x$  and  $y$ . The unknown quantity of interest is  $\bar{Y}$ , while  $\bar{X}$  is assumed to be known in advance. According to  $p(s)$ , let  $\hat{\bar{X}}$  and  $\hat{\bar{Y}}$  be two unbiased estimators of  $\bar{X}$  and  $\bar{Y}$ , respectively. Then, we consider the following class of estimators for  $\bar{Y}$

$$\hat{Y}_{pr} = \hat{\bar{Y}} \frac{\bar{X} + \tau}{\hat{\bar{X}} + \tau} \quad (1)$$

where  $\tau$  is a constant that may be related to population parameters. In order to investigate the efficiency of the class, we expand it in a Taylor's series (Wolter (1985), p. 225). Let  $\delta_y = (\hat{\bar{Y}} - \bar{Y})/\bar{Y}$  and  $C(\hat{\bar{Y}}) = \sqrt{Var(\hat{\bar{Y}})}/\bar{Y}$ . Analogously for the auxiliary variable  $x$ . Let  $C(\hat{\bar{X}}, \hat{\bar{Y}}) = Cov(\hat{\bar{X}}, \hat{\bar{Y}})/\bar{X}\bar{Y}$  and let  $\rho_{\hat{\bar{X}}, \hat{\bar{Y}}}$  denote the coefficient of correlation between estimators  $\hat{\bar{X}}$  and  $\hat{\bar{Y}}$ . Then, the class can be expressed as

$$\hat{Y}_{pr} = \frac{\bar{Y}(1 + \delta_y)(\bar{X} + \tau)}{\bar{X}(1 + \delta_x) + \tau}. \quad (2)$$

If  $n$  is "sufficiently large", expanding  $[\bar{X}(1 + \delta_x) + \tau]^{-1}$  at the point  $\delta_x = 0$  in a second order Taylor's series and taking in the expanded expression (2) the expectation term-by-term up to include  $\delta$  terms of power two, we obtain the bias (B) and the mean square error (MSE) of the class to the first order of approximation

$$\begin{aligned} B(\hat{Y}_{pr}) &= \frac{1}{\bar{X} + \tau} \left[ \frac{\bar{Y}Var(\hat{\bar{X}})}{\bar{X} + \tau} - Cov(\hat{\bar{X}}, \hat{\bar{Y}}) \right] \\ MSE(\hat{Y}_{pr}) &= Var(\hat{\bar{Y}}) + \frac{\bar{Y}^2 Var(\hat{\bar{X}})}{(\bar{X} + \tau)^2} - \frac{2\bar{Y}Cov(\hat{\bar{X}}, \hat{\bar{Y}})}{\bar{X} + \tau}. \end{aligned} \quad (3)$$

Minimization of (3) with respect to  $\tau$  is achieved when

$$\tau = \bar{X} \frac{[C(\hat{\bar{X}})]^2 - C(\hat{\bar{X}}, \hat{\bar{Y}})}{C(\hat{\bar{X}}, \hat{\bar{Y}})}. \quad (4)$$



For this optimum choice, the class is unbiased and the minimum MSE (variance) is given by

$$\min MSE(\hat{Y}_{pr}) = Var(\hat{Y})(1 - \rho_{\hat{X}, \hat{Y}}^2). \quad (5)$$

We observe that (5) is the variance of the regression estimator  $\hat{Y}_{lr} = \hat{Y} + \beta_{\hat{Y}, \hat{X}}(\bar{X} - \hat{X})$ ,  $\beta_{\hat{Y}, \hat{X}}$  being the regression coefficient of  $\hat{Y}$  on  $\hat{X}$ .

**Remark.** Class (1) can be further generalized by considering the form  $\hat{Y}_{pr, \alpha} = \hat{Y} \left[ \frac{\bar{X} + \tau}{\hat{X} + \tau} \right]^\alpha$ ,  $\alpha$  being a suitable constant to be chosen. Despite its more complex structure, the class does not make any improvement to  $\hat{Y}_{pr}$  in terms of efficiency, at least to the first order of approximation. In fact, it is easy to verify that the optimum choice of  $\tau$  and  $\alpha$  yields (5).

### 3 Efficiency considerations

The result stated in (5) emphasizes that all the estimators belonging to the class can be only, at best, as efficient as the regression estimator. They are equivalent to it only when the constant  $\tau$  satisfies relation (4). Since this condition is very easy to verify, one can immediately ascertain whether the estimator to be used may be improved by the simple and well known regression estimator. To better understand this point, consider, for instance, the estimators in Table 1 first proposed in srswor. These estimators are not opti-

Authors	Estimators	$\tau$
Sisodia and Dwivedi (1981)	$\hat{Y}_{SD} = \hat{Y} \frac{\bar{X} + C_x}{\hat{X} + C_x}$	$C_x$
Singh and Kakran (1993)	$\hat{Y}_{SK} = \hat{Y} \frac{\bar{X} + \beta_2(x)}{\hat{X} + \beta_2(x)}$	$\beta_2(x)$
Upadhyaya and Singh (1999)	$\hat{Y}_{US1} = \hat{Y} \frac{\bar{X}\beta_2(x) + C_x}{\hat{X}\beta_2(x) + C_x}$	$C_x/\beta_2(x)$
	$\hat{Y}_{US2} = \hat{Y} \frac{\bar{X}C_x + \beta_2(x)}{\hat{X}C_x + \beta_2(x)}$	$\beta_2(x)/C_x$

**Table 1.** Some estimators in the class. The parameters  $C_x$  and  $\beta_2(x)$  are the coefficients of variation and kurtosis of the auxiliary variable  $x$ .

mum in the class since relation (4) does not hold. Therefore, we can conclude that they are less efficient than the regression estimator.

As already mentioned in Section 1, BK have recently analyzed the efficiency of the estimators in Table 1 according to pps sampling without replacement

by replacing  $\hat{Y}$  and  $\hat{X}$  with the Horvitz-Thompson (HT hereafter) estimator

$$\hat{T}_{HT} = \frac{1}{N} \sum_{i \in s} \frac{t_i}{\pi_i}, \quad t = x, y \quad (6)$$

with

$$Var(\hat{T}_{HT}) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) t_i t_j, \quad t = x, y \quad (7)$$

where  $\pi_i = \sum_{s \ni i} p(s)$  and  $\pi_{ij} = \sum_{s \ni (i,j)} p(s)$  are, respectively, the first and second order inclusion probabilities ( $\pi_i = \pi_{ii}$ ). Although the idea seems interesting, it contains some basic inaccuracies. The modified estimators are compared in terms of MSE with the ratio estimator  $\hat{Y}_r = (\hat{Y}_{HT}/\hat{X}_{HT})\bar{X}$  (Thompson (1992), p. 67) and the conditions under which the proposed estimators can outperform  $\hat{Y}_r$  are given. Some numerical comparisons are carried out by using expressions for  $\pi_i$  and  $\pi_{ij}$  inherited from the *adaptive cluster sampling* (Thompson and Seber (1996), pp. 95-96) that appear rather inappropriate for the pps sampling. We are not able to find any reasonable match between pps sampling and adaptive cluster sampling: formulas for the inclusion probabilities do not work correctly for BK problem. In addition, we observe that BK estimators can be more efficient than  $\hat{Y}_r$  but, for our earlier discussion, it is clear that they cannot improve the regression estimator  $\hat{Y}_{lr} = \hat{Y}_{HT} + \beta_{\hat{Y}_{HT}, \hat{X}_{HT}}(\bar{X} - \hat{X}_{HT})$  (Thompson (1992), p. 82). Motivated by these considerations, therefore, we perform a simulation study in order to quantify the efficiency gain of the best estimator in class (1), say the regression estimator, upon BK estimators. In so doing, we will implement a heuristic algorithm for estimating the inclusion probabilities when their exact determination becomes prohibitive (Fattorini (2006)).

## 4 A simulation study

The HT estimator is unbiased for the population mean with a sampling variance dependent on first and second order inclusion probabilities. Under a generic pps sampling scheme, the explicit derivation of these probabilities becomes prohibitive when the population size and/or the sample size increase due to computational time and/or memory allocation problems. In this case, a heuristic solution can be adopted. If the inclusion probabilities do not depend on the unknown population values  $y_i$ , they can be reliably simulated. The simulation algorithm is briefly described as follows.

From the population  $U$ , we consider samples without replacement of size  $n$ . The number of possible distinct samples of size  $n$  is  $\binom{N}{n}$ . According to pps sampling, we assume that each population unit has an unequal selection probability  $p_i$  which is proportional to its *size*. Let  $z_i$  denote the size of the  $i$ -th population unit. Then, in our simulation,  $p_i = z_i / \sum_{j=1}^N z_j$  in such a way

the greater  $z_i$ , the higher the selection probability of the unit. If the exact computation of  $\pi_i$  and  $\pi_{ij}$  on all the  $\binom{N}{n}$  possible samples is unfeasible, we independently select  $M < \binom{N}{n}$  samples and obtain an estimate of the inclusion probabilities as  $\hat{\pi}_i = M_i/M$  and  $\hat{\pi}_{ij} = M_{ij}/M$ , where  $M_i$  and  $M_{ij}$  represent the number of samples that contain unit  $i$  and units  $(i, j)$ , respectively. The computational algorithm has been implemented in the R environment and the command `sample(U,n,replace=FALSE,prob=p)` has been used to select the pps samples. As a consequence, the HT estimators in (6) and their variances are modified by replacing  $\pi_i$  and  $\pi_{ij}$  with  $\hat{\pi}_i$  and  $\hat{\pi}_{ij}$ .

To have an idea of the accuracy of the estimated probabilities and, at the same time, to compare the estimators shown in Section 3, we consider a numerical study performed on some real data used by BK and presented in Cochran (1977, p. 34). These data concern the weekly expenditure on food ( $y$ ), the weekly family income ( $x$ ) and the number of persons per family ( $z$ ). In order to have comparable results with those in BK, the family labeled 5 is excluded from the analysis since it is considered as an outlier in BK. Three sample sizes are adopted:  $n = 10, 15, 20$ . The total number of distinct samples to be investigated for the exact determination of  $\pi_i$  and  $\pi_{ij}$  is  $\binom{32}{n} = 64\,512\,224, 565\,722\,720, 225\,792\,840$  according to  $n = 10, 15, 20$ , respectively. Despite the modest sizes of the population under study and samples, allocation memory problems occur when carrying out the analysis with a standard personal computer. Therefore, in this study we consider a limited number of samples,  $M = 100\,000$ . To evaluate the performance of the estimated probabilities, we compare the efficiency of the best estimator in our class ( $\hat{Y}_{lr}$ ) with that of the estimators discussed in BK. First, we assume that units are selected according to srswor for which  $\pi_i = n/N$  and  $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ . The choice of assuming the srswor as a reference sampling is motivated by the fact that the inclusion probabilities are known in advance for each population unit.

In Table 2 we report the first order approximate MSE computed by using both the exact and estimated inclusion probabilities. The maximum absolute difference ( $D$ ) between exact and estimated MSEs is also shown. We observe

$n$	$\hat{Y}_r$	$\hat{Y}_{lr}$	$\hat{Y}_{SD}$	$\hat{Y}_{SK}$	$\hat{Y}_{US1}$	$\hat{Y}_{US2}$	$D$
10	6.119	6.098	6.118	6.122	6.119	6.154	0.035
	<b>6.108</b>	<b>6.063</b>	<b>6.107</b>	<b>6.112</b>	<b>6.109</b>	<b>6.148</b>	
15	3.152	3.141	3.152	3.154	3.152	3.170	0.008
	<b>3.148</b>	<b>3.14</b>	<b>3.145</b>	<b>3.147</b>	<b>3.145</b>	<b>3.162</b>	
20	1.669	1.663	1.669	1.670	1.669	1.678	0.004
	<b>1.669</b>	<b>1.667</b>	<b>1.668</b>	<b>1.669</b>	<b>1.669</b>	<b>1.678</b>	

**Table 2.** MSE computed with exact and estimated (in bold) first and second order inclusion probabilities in srswor.

that, despite of the severe reduction of the cardinality of the sample space, no striking differences appear in the precision of the estimators. The magnitude of the “exact” MSE is achieved when probabilities are estimated and small variations tend to disappear as the sample size increases: modified estimators are asymptotically ( $M \rightarrow \binom{N}{n}$ ) equivalent to the exact ones. This simulation provides evidence that the cumbersome problem of determining the exact inclusion probabilities can be overcome by means of simulated probabilities over a limited number of samples. Therefore, we do not expect substantial changes when an unequal sampling design is considered for which the inclusion probabilities cannot be easily computed except for small sample sizes, or for particular designs such as the Midzuno scheme. For this reason, we are motivated in performing a second simulation study assuming a pps sampling for which the exact probabilities may be not known in advance and, hence, they need to be estimated. Using the same settings of the previous investigation, we compare the estimators according to two pps samplings: (i) the unequal sampling with estimated inclusion probabilities previously discussed; (ii) the Midzuno selection scheme in which the first unit is selected according to a pps mechanism and the  $n - 1$  more units from the remaining  $N - 1$  population units according to srswor. The Midzuno method is not cumbersome to implement and allows us to determine the exact inclusion probabilities in a nice fashion as  $\pi_i = p_i + (1 - p_i) \frac{n-1}{N-1}$  and  $\pi_{ij} = \frac{n-1}{N-1} \left[ \frac{N-n}{N-2} (p_i + p_j) + \frac{n-2}{N-2} \right]$ . The results, given in terms of first order approximate MSE, are reported in Table 3. Cross checks with the results shown in Table 2 emphasize the fact

$n$	Method	$\hat{Y}_r$	$\hat{Y}_{lr}$	$\hat{Y}_{SD}$	$\hat{Y}_{SK}$	$\hat{Y}_{US1}$	$\hat{Y}_{US2}$
10	Midzuno	5.978	<b>5.498</b>	5.976	5.988	5.980	6.064
	Est. prob.	5.813	<b>4.945</b>	5.796	5.917	5.832	6.665
15	Midzuno	3.102	<b>2.963</b>	3.102	3.106	3.103	3.134
	Est. prob.	2.877	<b>2.065</b>	2.870	2.925	2.886	3.267
20	Midzuno	1.649	<b>1.606</b>	1.649	1.650	1.649	1.663
	Est. prob.	1.405	<b>0.886</b>	1.401	1.425	1.408	1.571

**Table 3.** MSE computed with exact Midzuno inclusion probabilities and estimated inclusion probabilities in pps sampling.

that the best estimator in the class outperforms all the others whatever the sample size and the sampling design. Sampling with unequal estimated probabilities offers the best solution in terms of efficiency if compared with the Midzuno scheme and the srswor. It is worth noting that the gain in efficiency rises as the sample size increases.

The above considerations are drawn on the basis of the approximate MSE. The validity of this approximation relies on the realistic assumption that the sample size is “sufficiently large”. But, when can a sample size be considered as “sufficiently large”? Do the sample sizes assumed in this paper offer

guarantee for valid first order MSE-based conclusions? To shed light on the matter, we have performed another Monte Carlo experiment consisting in drawing from the real population under study  $M = 100\,000$  independent samples of size  $n = 10, 15, 20$ . For each sample, the estimate of  $\bar{Y} = 27.063$  is obtained and an estimation of the MSE for estimator  $\hat{Y}_{(\cdot)}$  is computed as  $M^{-1} \sum_{k=1}^M \left( \hat{Y}_{(\cdot)}^{(k)} - \bar{Y} \right)^2$ . Simulation outcomes are given in Table 4. No

$n$	Method	$\hat{Y}_r$	$\hat{Y}_{lr}$	$\hat{Y}_{SD}$	$\hat{Y}_{SK}$	$\hat{Y}_{US1}$	$\hat{Y}_{US2}$
10	srswor	6.103	<b>6.089</b>	6.102	6.107	6.104	6.143
	Midzuno	5.980	<b>5.887</b>	5.978	5.990	5.982	6.068
	Est. prob.	5.923	<b>4.783</b>	5.905	6.035	5.944	6.846
15	srswor	3.159	<b>3.145</b>	3.159	3.161	3.159	3.178
	Midzuno	3.107	<b>3.072</b>	3.107	3.111	3.108	3.141
	Est. prob.	2.903	<b>2.404</b>	2.895	2.953	2.912	3.309
20	srswor	1.675	<b>1.664</b>	1.674	1.676	1.675	1.685
	Midzuno	1.661	<b>1.643</b>	1.661	1.663	1.661	1.676
	Est. prob.	1.420	<b>1.203</b>	1.416	1.441	1.423	1.592

**Table 4.** Estimated MSE from srswor and pps sampling with exact Midzuno inclusion probabilities and estimated inclusion probabilities.

significant discrepancy is evident with respect to the results in Tables 2-3. We observe that a further refinement (not reported here) between first order approximate MSE and estimated MSE has been achieved over one million of simulation experiments where the MSEs are nearly identical up the second decimal digit. Therefore, we can be fairly confident of the validity of the comparisons carried out since they are not affected by the sample size.

## 5 Conclusion

Researchers and practitioners interested in the estimation of population mean (or total) can find a plethora of proposals in the literature. Does this wide choice offer the best solution for the problem at hand? New estimators are usually proposed by modifying the structure of existing ones without providing any reasonable and convincing reason. Very often, new proposals are compared with estimators that are less efficient and, therefore, they seem to be innovative and offer a means to improve the estimation of the population mean. In many cases, however, these estimators are equivalent to existing proposals and no practical gain is produced. As a matter of fact, this practice has inundated the literature on survey sampling with papers whose theoretical and practical relevance is rather questionable.

From these considerations showing that it may well be fruitless to consider

single estimators, we have focused on a class of estimators and found unsurprisingly that the best estimator in the class turns out to be the well known regression estimator. The class has a very simple expression, includes recent proposals such as Bacanli and Kadilar (2008), and has the merit of showing clearly that, *ceteris paribus*, no improvement can be achieved upon the regression estimator. It seems that many researchers, though are aware of this result, often choose to ignore it.

In order to evaluate the performance of Bacanli-Kadilar estimators with respect to the regression estimator when sampling is performed according to probability proportional to size, we have tackled the problem of estimating the first and second order inclusion probabilities for cases where their exact determination is unfeasible. A simple estimation algorithm has been considered and its stability ascertained through a number of simulation experiments. The main finding is that the estimated inclusion probabilities allow us to achieve satisfactory results both in terms of accuracy and in the reduction of time and memory required. Finally, the results demonstrate that Bacanli-Kadilar estimators do not work well when compared with the best estimator of the class considered in this paper.

**Acknowledgments.** Work supported by the Italian Ministry of University and Research, MIUR-PRIN 2007: “Efficient use of auxiliary information at the design and at the estimation stage of complex surveys: methodological aspects and applications for producing official statistics”

## References

- BACANLI, S., KADILAR, C. (2008): Ratio estimators with unequal probability designs. *Pakistan Journal of Statistics* 24 (3), 167-172.
- COCHRAN, W.G. (1977): *Sampling Techniques*. John Wiley & Sons, New York.
- FATTORINI, L. (2006): Applying the Horvitz-Thompson criterion in complex designs: a computer-intensive perspective for estimating inclusion probabilities. *Biometrika* 93 (10), 269-278.
- KADILAR, C., CINGI, H. (2006): Improvement in estimating the population mean in simple random sampling. *Applied Mathematics Letters* 19 (1), 75-79.
- SINGH, H.P., KAKRAN, M.S. (1993): A modified ratio estimator using known coefficient of kurtosis of an auxiliary character. *Unpublished manuscript*.
- SISODIA, B.V.S., DWIVEDI, V.K. (1981): A modified ratio estimator using coefficient of variation of an auxiliary character. *Journal of Indian Society of Agricultural Statistics* 33 (2), 13-18.
- THOMPSON, S.K. (1992): *Sampling*. John Wiley & Sons, New York.
- THOMPSON, S.K., SEBER, G.A.F. (1996): *Adaptive Sampling*. John Wiley & Sons, New York.
- UPADHYAYA, L.N., SINGH, H.P. (1999): Use a transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal* 45 (5), 627-636.
- WOLTER, K.M. (1985): *Introduction to Variance Estimation*. Springer-Verlag, New York.

# Improving Overlapping Clusters obtained by a Pyramidal Clustering

Edwin Diday<sup>1</sup>, Francisco de A. T. de Carvalho<sup>2</sup>, and Luciano D.S. Pacifico<sup>2</sup>

<sup>1</sup> LISE-CEREMADE, Université Paris-IX Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris 16 ième, France, *diday@ceremade.dauphine.fr*

<sup>2</sup> Centro de Informatica -CIn/UFPE, Av. Prof. Luiz Freire, s/n -Cidade Universitaria, CEP 50740-540, Recife-PE, Brésil, *{fatc,ldsp}@cin.ufpe.br*

**Abstract.** Indexed standard or spatial hierarchical clustering produce partitions if they are cut at a given level. Such partition can be improved by using a K-means like clustering. In case of a standard or spatial pyramid a cut at a given level produces an overlapping clustering (where some observations can belong to several clusters). In order to improve such overlapping clustering we need an extension of K-means like algorithm to a new kind of algorithm giving at output a better overlapping clustering for a given criterion. The aim of this paper is to provide such algorithm and to show that it improves at each step a given criterion.

**Keywords:** overlapping clustering, pyramidal clustering, dynamic clustering

## 1 Introduction

Standard classification structures are for instance, indexed hierarchies (Johnson (1967)) which produce nested partitions or indexed pyramids (Diday (2008)) which produce nested overlapping clusters covering the population. Spatial pyramids concern any finite set, of units described by a finite set of standard or symbolic variables and for which a dissimilarity, denoted  $d$ , can be given. We recall that a tessellation is a tiling pattern that covers space without leaving any gap. The aim of a spatial pyramid is to associate each unit to a vertex of a tessellation and to produce simultaneously an overlapping classification structure "compatible" with the tessellation and which induces a dissimilarity  $d'$  fitting  $d$  as best as possible. Its advantage in comparison with standard approaches is that it provides not only overlapping clusters but also their mapping on a given tessellation (for example, on a grid).

The results already obtained in standard hierarchies and pyramids to spatial pyramids compatible with a grid, can be extended to other kinds of classes (for instance, of maximal or connected classes instead of convex), to other kinds of grids (as triangular or hexagonal) and to multidimensional grids. For instance, in the case of a cubic grid we can obtain a  $3 - D$  Yadidean dissimilarity defined by blocks which are  $2 - D$  Yadidean dissimilarities increasing from the main diagonal in rows and columns. In that way, we can

go more generally, from a  $n - D$  Yadidean dissimilarity to a  $(n + 1) - D$  one. In the  $3 - D$  Yadidean dissimilarity case, the classes of the associated classification structure are volumes as they merge cells of the  $3 - D$  grid. They form a partitioning or an overlapping of the  $3 - D$  grid depending on the fact that the  $3 - D$  associated Yadidean dissimilarity is “ultrametric” or not, etc. Many other directions remain open, such as how to get the closest Yadidean dissimilarity of a given dissimilarity and what is the statistical distribution of a quality criterion between a given dissimilarity and different kinds of Yadidean dissimilarity (weakly large, large, weakly strict, strict).

*Example 1.* Each class of the hierarchy or the pyramid has a height given by a mapping called “index”. For example, in the indexed hierarchy given in Figure 2, like in the indexed pyramid given in Figure 3, the height of the class  $\{A, B\}$  is 1. In standard indexed hierarchies or pyramids, the tessellation is reduced to a chain on a straight line (defined by the positions of A, B, C, D on the straight line given in Figures 2 and 3). It is always possible to induce a dissimilarity from such classification structures by associating to any couple  $(x_1, x_2)$  of elements of  $\Omega$  their dissimilarity  $d'(x_1, x_2)$ , defined by the height of the class of lowest height containing them (for instance, the indexed pyramid, shown in Figure 3, induces  $d'(B, C) = 1$  and  $d'(A, C) = \sqrt{2}$ ). Johnson (1967) has shown that a hierarchy induces by this way an ultrametric and Diday (1986) has shown that a pyramid induces by the same way a Robinsonian dissimilarity which contains as a special case ultrametries. Diday (2008) has shown that a spatial pyramid induces (also, by the same way), a new kind of dissimilarity called “Yadidean” (“Yadid” means “friend” for the people of the Bible who has contributed in building the Egyptian Pyramids). A dissimilarity, denoted  $d_T$ , can also be induced from a tessellation  $T$ . For instance, the length of the shortest path connecting the vertices associated to two units measures their dissimilarity. This will be our choice in the following. The “compatibility” between a classification structure and a tessellation (where each vertex is associated to an unit) can then be measured by the compatibility between the dissimilarity induced by the classification structure  $d'$  and  $d_T$ . The “compatibility” between two dissimilarities  $d_1$  and  $d_2$  denoted  $Comp(d_1, d_2)$ , can be measured for instance, by the number of times where the largest dissimilarity among a pair in a triple of units is the same one for  $d_1$  and  $d_2$ . The fit between  $d_1$  and  $d_2$  can be measured by  $|d_1 - d_2|$  or the number of different dissimilarity values. Hence, the “quality” of a classification structure can be measured by the compatibility and fit between  $d$ ,  $d'$  and  $d_T$ .

It can also be shown that not only the fit between  $d$  and  $d'$  is the best for the spatial pyramid for this example, but also the compatibility between  $d$  and  $d_T$ . Let the tessellation  $T$  be defined by the sequence of segments  $(A, B)$ ,  $(B, C)$ ,  $(C, D)$  on the straight line induced by the hierarchy and the pyramid. We get  $d_T(A, D) = 3$  as there are three segments between  $A$  and  $D$ :  $AB$ ,  $BC$  and  $CD$  (see Figures 2 and 3). Therefore,  $d_T(A, D)$  is maximal



among the couples of the triples  $(A, C, D)$  and  $(D, A, B)$ . This is not the case, for the tessellation denoted  $M$  induced by the spatial pyramid as  $d_M(A, D) = 1$ , like for the initial dissimilarity  $d$  as  $d(A, D) = 1$ . The maximal value for  $d$  and  $d_M$  among the couples of the triples  $(A, C, D)$  (resp.  $(D, A, B)$ ) is  $AC$  (resp.  $BD$ ). For the three dissimilarities the maximal value among the couples of the triples  $(A, B, C)$  (resp.  $(B, C, D)$ ) is  $AC$  (resp.  $BD$ ). Therefore, we obtain finally  $Comp(d, d_T) = 2$  and  $Comp(d, d_M) = 4$ . This means that the compatibility induced by the tessellation associated to the spatial pyramid is better than the one induced by the standard hierarchy and pyramid.

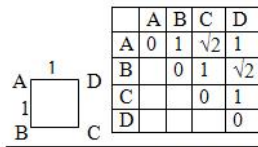


Fig. 1 A square cell of a grid and the distances  $d$  between its vertices

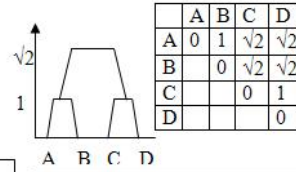


Fig. 2 A single Hierarchy with  $d$  as input and its induced ultrametric  $d'$ : 2 values differ

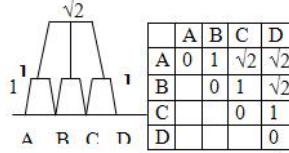


Fig. 3 A single Pyramid with  $d$  as input and its induced Robinsonian dissimilarity  $d'$ : 1

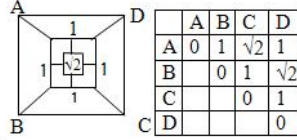


Fig. 4 A single Spatial Pyramid with  $d$  as input and its induced Yadidean dissimilarity  $d'$ : no

*Example 2.* In order to gain insight by an example, we use as input the dissimilarities induced by a square of side of length equal to 1. This square and the dissimilarities (which is the Euclidean distance) between its four vertices  $A, B, C, D$  are given in Figure 1. In Figure 2 a “complete linkage” hierarchy is shown. As usual, “complete” means that the classes are aggregated bottom up by the largest dissimilarity  $d$  between the units of the two classes, until all units become aggregated. This indexed hierarchy induces an ultrametric  $d'$ , also given in Figure 2. It results that two values of  $d'$  differ from  $d$  ( $d'(B, C)$  and  $d'(A, C)$ ). A “complete linkage” Pyramid with  $d$  as input and its induced Robinsonian dissimilarity is given in Figure 3. It results that only 1 value differs from  $d$  as  $1 = d(A, D) \neq d'(A, D) = \sqrt{2}$ . A complete linkage Spatial Pyramid from a top view with  $d$  as input and its induced Yadidean dissimilarity  $d'$  are given in Figure 4. This spatial pyramid has two levels, the height of the first level is 1, the height of the second is  $\sqrt{2}$ . No value of its induced dissimilarity  $d'$  differs from  $d$ . Hence, this result shows, on this example, that with a smaller number of levels (2, instead of 3 for the standard hierarchy and 4 for the standard pyramid), the spatial pyramid gives the best result in comparison with the standard approaches.

The aim of this paper is to provide an algorithm which starts from an overlapping clustering produced by a pyramid and to show that it improves it at each step for a given criterion.

Several authors addressed the problem, for example by extending hierarchies to weak hierarchies (Bertrand and Janowitz, (2003)), and more recently Cleuziou (2008) with the OKM algorithm.

In this paper, we first present the algorithm, then we show its convergence and finally we give some examples with results and comparisons.

## 2 Principle of the algorithm and proof of its convergence

Let  $\Omega = \{e_1, \dots, e_n\}$  be a set of objects described by  $p$  variables  $\{y_1, \dots, y_p\}$ . We denote  $R = (R_1, \dots, R_K)$  a covering of the set of observations  $\Omega$  which means that  $\Omega = \bigcup_{k=1}^K R_k$ . Our aim is to cluster the set of objects into  $K$  overlapping clusters, each cluster  $R_k$  ( $k = 1, \dots, K$ ) having a representative  $g_k$ . The overlapping clustering algorithm aims to give a list of overlap clusters  $R = (R_1, \dots, R_K)$  and the corresponding list of prototypes  $G = (g_1, \dots, g_K)$  by optimizing the following adequacy criterion:

$$W(R, G) = \sum_{k=1}^K \frac{1}{n_k} \sum_{e_i \in R_k} d(e_i, g_k) \quad (1)$$

where  $d$  is a dissimilarity function and  $n_k = \text{cardinal}(R_k)$ .

As each observation is weighted by the number of observations of each class to which it belongs, the criterion takes more care to the observations which belong in small classes.

The algorithm is based on the following insertion-deletion process:

- each individual  $e'$  is added to a cluster  $C$  belonging to  $R$ , of cardinality  $n_C$ , if  $d(e', g_C) < \frac{1}{n_C} \sum_{e \in C} d(e, g_C)$ ;
- if  $e'$  is added to a cluster  $C$ , it can be deleted from its own cluster  $C'$  if the following condition is satisfied:  $d(e', g_{C'}) > \frac{1}{n_{C'}} \sum_{e \in C'} d(e, g_{C'})$ .

This insertion-deletion process is justified by the following lemmas. The first one gives a necessary and sufficient condition for adding an observation to a class, the second one gives a necessary and sufficient condition for deleting an observation from its class (if it has been added to another class).

**Lemma 4.** *A necessary and sufficient condition to obtain  $I_C = \frac{\sum_{e \in C} d(e, g_C)}{n_C} > I_{C \cup \{e'\}} = \frac{(\sum_{e \in C} d(e, g_C) + d(e', g_C))}{n_C + 1}$  is that  $d(e', g_C) < \frac{1}{n_C} \sum_{e \in C} d(e, g_C)$ .*

*Proof.*  $I_C = \frac{\sum_{e \in C} d(e, g_C)}{n_C} > \frac{(\sum_{e \in C} d(e, g_C) + d(e', g_C))}{n_C + 1} = I_{C \cup \{e'\}}$  is equivalent to  $\sum_{e \in C} d(e, g_C) \left( \frac{1}{n_C} - \frac{1}{n_C + 1} \right) = \frac{\sum_{e \in C} d(e, g_C)}{n_C(n_C + 1)} > \frac{d(e', g_C)}{n_C + 1}$ , which is equivalent to  $I_C = \frac{\sum_{e \in C} d(e, g_C)}{n_C} > d(e', g_C)$ . Which proves the lemma 1.

**Lemma 5.** *A necessary and sufficient condition for having  $I_C = \frac{\sum_{e \in C} d(e, g_C)}{n_C} > I_{C - \{e'\}} = \frac{(\sum_{e \in C} d(e, g_C) - d(e', g_C))}{n_C - 1}$  is that  $d(e', g_C) > \frac{1}{n_C} \sum_{e \in C} d(e, g_C)$ .*

*Proof.* The proof of this lemma 2 is analogous to the proof of lemma 1.

Given a dissimilarity  $d$  defined on  $\Omega \times \Omega$ , the algorithm can start from a covering  $R^{(1)} = (R_1^{(1)}, \dots, R_K^{(1)})$  given for example by a standard or spatial pyramid. Then it follows three steps:

- 1) Having  $R^{(t)} = (R_1^{(t)}, \dots, R_K^{(t)})$  obtained at step  $t-1$ ,  $G^{(t)} = (g_1^{(t)}, \dots, g_K^{(t)})$  is such that each  $g_k^{(t)}$  minimizes the sum of the dissimilarities to all the observations of  $R_k^{(t)}$ ;
- 2) The insertion-deletion process is applied to  $R^{(t)}$  in order to obtain  $R^{(t+1)}$ ;
- 3) We repeat 1) with  $R^{(t+1)}$  instead of  $R^{(t)}$  in order to obtain  $(R^{(t+1)}, G^{(t+1)})$  until the criterion  $W$  converge.

This convergence can be proved in the following way.

**Proposition 13.** *The criterion  $W(R, G) = \sum_{k=1}^K \frac{1}{n_k} \sum_{e_i \in R_k} d(e_i, g_k)$  decreases at each iteration.*

*Proof.* Our aim is to prove that:  $v_t = W(R^{(t)}, G^{(t)}) = \sum_{k=1}^K \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t)})}{n_k^{(t)}} \geq v_{t+1}$ .

First step: getting  $g_k^{(t+1)}$  from  $R_k^{(t)}$ . When we have numerical variables which describe each observation, and we use an Euclidean dissimilarity we can write,  $u_t(x) = \sum_{k=1}^K \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, x)}{n_k^{(t)}}$  and we have,  $u_t^* = \min_{x \in \mathbb{R}^p} u_t(x) = \sum_{k=1}^K \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t+1)})}{n_k^{(t)}}$  which means that  $g_k^{(t+1)}$  ( $k = 1, \dots, K$ ) is the mean of the cluster  $R_k^{(t)}$ . If we cannot calculate an Euclidean dissimilarity,  $g_k^{(t+1)}$  ( $k = 1, \dots, K$ ) is the observation such that the mean of its dissimilarities to the other observations of the cluster  $R_k^{(t)}$  is minimum. In other words  $u_t^* = \min_{x \in \mathbb{R}^p} u_t(x) = \sum_{k=1}^K \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t+1)})}{n_k^{(t)}}$ . So by construction in both cases we have  $u_t^* \leq v_t$ .

Second step : In order to get the next covering,  $R_k^{(t+1)}$ , we use the following steps: each observation  $e'$  is inserted to a cluster  $C$  of the covering  $R^{(t)}$  if  $d(e', g_C) < \frac{1}{n_C} \sum_{e \in C} d(e, g_C)$ . It results from lemma 1 that :  $\frac{\sum_{e \in C} d(e, g_C)}{n_C} >$

$$I_{C \cup \{e'\}} = \frac{(\sum_{e \in C} d(e, g_C) + d(e', g_C))}{n_C + 1}. \text{ If } C = R_j^{(t)} \text{ and we denote } R_j^{(t)+} \text{ the new class obtained by adding } e' \text{ to } R_j^{(t)} \text{ we obtain the inequality } u_t^* \geq u_t^{*+} \text{ defined as follows: } u_t^* = \sum_{k=1}^K \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t+1)})}{n_k^{(t)}} = \sum_{k=1}^{(j-1)} \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t+1)})}{n_k^{(t)}} + \frac{\sum_{e_i \in R_j^{(t)}} d(e_i, g_j^{(t+1)})}{n_j^{(t)}} + \sum_{k=(j+1)}^K \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t+1)})}{n_k^{(t)}} \geq \sum_{k=1}^{(j-1)} \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t+1)})}{n_k^{(t)}} + \frac{\sum_{e_i \in R_j^{(t)} \cup \{e'\}} d(e_i, g_j^{(t+1)})}{n_j^{(t)} + 1} + \sum_{k=(j+1)}^K \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t+1)})}{n_k^{(t)}} = u_t^{*+}.$$

If  $e'$  belongs to  $C = R_l^{(t)}$  and moreover:  $d(e', g_C) > \frac{1}{n_C} \sum_{e \in C} d(e, g_C)$ . If we make the assumption for example that  $l > j$  (the proof is identical in the other case) it results from the lemma 2 the inequality  $u_t^* \geq u_t^{*-}$  defined as follows:  $u_t^{*+} \geq \sum_{k=1}^{(j-1)} \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t+1)})}{n_k^{(t)}} + \frac{\sum_{e_i \in R_j^{(t)}} d(e_i, g_j^{(t+1)})}{n_j^{(t)}} + \sum_{k=(j+1)}^{(l-1)} \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t+1)})}{n_k^{(t)}} + \frac{\sum_{e_i \in R_l^{(t)}} d(e_i, g_l^{(t+1)})}{n_l^{(t)} - 1} + \sum_{k=(l+1)}^K \frac{\sum_{e_i \in R_k^{(t)}} d(e_i, g_k^{(t+1)})}{n_k^{(t)}} = u_t^{*-}$ . By applying this process of insertion on all the individuals several times until the process converge as the criterion decreases at each step, we obtain a new covering  $R^{(t+1)}$  and so  $v_{t+1} = W(R^{(t+1)}, G^{(t+1)}) \leq v_t = W(R^{(t)}, G^{(t)})$ , which proves that the sequence  $v_t$  converges as it is positive and decreasing.

### 3 Applications

To illustrate the interest of the overlapping clustering algorithm introduced in this paper, given a data set we will apply a pyramidal clustering algorithm (see the sodas software: <http://www.info.fundp.ac.be/asso/>) in order to obtain a covering for each data sets. The overlapping clustering algorithm will starts from this covering and will alternate its steps until the convergence of the clustering criterion in order to obtain improved overlapping clusters.

Fats and oils interval-valued data set (Ichino and Yaguchi, 1994) consists of a set of 8 objects described by 4 interval-valued variables. In this application, the 4 interval variables - *Specific Gravity*, *Freezing Point*, *Iodine Value* and *Saponification Value* - were considered for clustering purposes.

Car interval-valued data set consists of a set of 33 car models described by 8 interval-valued variables. In this application, the 8 interval variables - *Price*, *Engine Capacity*, *Top Speed*, *Acceleration*, *Step*, *Length*, *Width* and *Height* - were considered for clustering purposes.

The module HYPYR of the sodas software was applied on these data sets, in order to obtain pyramidal classifications. From these pyramidal structures it was obtained a covering of the fats and oils data set into 3 overlapping clusters as well as a covering of the car data set in 4 overlapping clusters, which have been used as initial covering (*a priori* classification) in order to start the overlapping clustering algorithm.

### 3.1 Results

The fats and oils data set *a priori* classification was as follows:

Cluster 1:  
3-cotton seed oil 4-sesame oil 5-camellia oil 6-olive oil 7-beef tallow 8-hog fat  
Cluster 2:  
1-linseed oil 2-perilla oil 3-cotton seed oil 4-sesame oil 5-camellia oil 6-olive oil  
Cluster 3:  
2-perilla oil 3-cotton seed oil 4-sesame oil 5-camellia oil 6-olive oil 8-hog fat

The overlapping clustering algorithm starts from this *a priori* classification until convergence to a stationary value of the adequacy criterion. The starting and final values of the adequacy criterion  $W$  were, respectively, 11437.80 and 8130.63. The fats and oils data set final classification is as follows:

Cluster 1:  
7-beef tallow 8-hog fat  
Cluster 2:  
1-linseed oil 3-cotton seed oil 4-sesame oil 5-camellia oil 6-olive oil  
Cluster 3:  
2-perilla oil 3-cotton seed oil 4-sesame oil 5-camellia oil 6-olive oil

The car data set *a priori* classification was as follows:

Cluster 1:  
21-Alfa 156/B 22-Skoda Octavia/B 23-Audi A3/U 24-Alfa 145/U 25-Rover 25/U  
26-Focus/B 27-Lancia Y/U 28-Twingo/U 29-Nissan Micra/U 30-Skoda Fabia/U  
31-Fiesta/U 32-Punto/U 33-Corsa/U  
Cluster 2:  
7-Mercedes SL/S 8-Mercedes Classe S/L 9-Audi A8/L 10-Bmw serie 7/L  
Cluster 3:  
8-Mercedes Classe S/L 9-Audi A8/L 10-Bmw serie 7/L 11-Mercedes Classe E/L  
12-Audi A6/B 13-Bmw serie 5/L 14-Lancia K/L 15-Alfa 166/L  
16-Rover 75/B 17-Passat/L 18-Mercedes Classe C/B 19-Bmw serie 3/B  
20-Vectra/B 21-Alfa 156/B 22-Skoda Octavia/B 23-Audi A3/U  
24-Alfa 145/U  
Cluster 4:  
1-Lamborghini/S 2-Aston Martin/S 3-Ferrari/S 4-Honda NSK/S  
5-Maserati GT/S 6-Porsche/S 7-Mercedes SL/S

The overlapping clustering algorithm starts from this *a priori* classifications until convergence to a stationary value of the adequacy criterion. The starting and final values of the adequacy criterion  $W$  were, respectively,  $3.94383e + 010$  and  $2.40101e + 010$ . The car data set final classification is as follows:

Cluster 1:  
32-Punto/U 33-Corsa/U  
Cluster 2:  
7-Mercedes SL/S  
Cluster 3:  
5-Maserati GT/S 9-Audi A8/L 11-Mercedes Classe E/L 12-Audi A6/B  
13-Bmw serie 5/L 14-Lancia K/L 15-Alfa 166/L 16-Rover 75/B  
17-Passat/L 18-Mercedes Classe C/B 19-Bmw serie 3/B 20-Vectra/B  
21-Alfa 156/B 22-Skoda Octavia/B 23-Audi A3/U 24-Alfa 145/U  
25-Rover 25/U 26-Focus/B 27-Lancia Y/U 28-Twingo/U  
29-Nissan Micra/U 30-Skoda Fabia/U 31-Fiesta/U 32-Punto/U  
33-Corsa/U  
Cluster 4:  
1-Lamborghini/S 2-Aston Martin/S 3-Ferrari/S 4-Honda NSK/S  
6-Porsche/S 7-Mercedes SL/S 8-Mercedes Classe S/L 10-Bmw serie 7/L

These final classifications were improved in comparison with the *a priori* classifications concerning the clustering homogeneity evaluated by the adequacy criterion.

## 4 Conclusion

This paper introduced an extension of the K-means algorithm in order to obtain overlapping clusters given by a pyramidal classification. The paper gives the principles of the algorithm and the proof of its convergence by improving a criterion from any *a priori* overlapping clusters. Applications having as entry overlapping clusters of interval-valued data sets given by pyramidal classifications showed the usefulness of this overlapping clustering algorithm.

## References

- AUDE, J.C. (1999): Analyse de génomes microbiens, apports de la classification pyramidal. *Thèse de Doctorat. Université Paris-IX Dauphine*.
- BENZECRI, J.P. (1973): *L'Analyse des données: la Taxinomie*. Dunod, Paris.
- BERTRAND, P. and JANOWITZ, M.F. (2002): Pyramids and weak hierarchies in the ordinal model for clustering. *Discrete Applied Mathematics*, 122(1–3), 55–81.
- BERTRAND, P. (1995): Structural properties of pyramidal clustering. *Dimacs Ser. Theor Comput. Sci.*, 19, 35–53.
- BRITO, P. (1994): Order structure of symbolic assertion objects. *IEEE Transactions on Knowledge and Data Engineering*, 6 (5), 830–85.
- CLEUZIOU, G. (2008): An extended version of the k-means method for overlapping clustering. In: *Proceedings of the Nineteenth International Conference on Pattern Recognition (ICPR 2008)*: 1–4.
- DIDAY, E. (1986): Orders and Overlapping clusters in pyramids. In: J. De Leeuw, et al., (Eds.): *Multidimensional Data Analysis*. DSWO Press, 201–234.
- DIDAY, E. (2004): Spatial Pyramidal Clustering Based on a Tessellation. In: D. Banks, L. House, F.R. McMorris, P. Arabie, W. Gaul (Eds.): *Classification, Clustering, and Data Mining Applications*. Springer, Berlin, 105–120.
- DIDAY, E. (2008): Spatial classification. *Discrete Applied Mathematics*, 156 (8), 1271–1294.
- JOHNSON, S.C. (1967): Hierarchical clustering schemes. *Psychometrika*, 32, 241–254.
- ICHINO, M. and YAGUCHI, H. (1994): Generalized Minkowsky metrics for mixed feature-type data analysis. *IEEE Transactions on System, Man and Cybernetics*, 24 (4), 698–708.
- RODRIGUEZ ROJAS, O. (2000): Classification et Modèles linéaires en Analyse des Données Symboliques. *Thèse de Doctorat. Université Paris-IX Dauphine*.

# Forecasting by Beanplot Time Series

Carlo Drago<sup>1</sup> and Germana Scepi<sup>1</sup>

University of Naples “Federico II”

Complesso Universitario Monte Sant’Angelo via Cinthia, Naples, Italy,

*carlo.drago@unina.it, germana.scepi@unina.it*

**Abstract.** In this paper, we propose a new approach for the aggregation, the parameterization and the forecasting of complex time series. This approach is based on a peculiar density plot, called beanplot (Kampstra (2002)). These types of new aggregated time series can be fruitfully used when there is an overwhelming number of observations, for example in High Frequency financial data. At the same time, they can be useful for analyzing the complex behaviour of the markets where we can discover important patterns in the long time (complex patterns of dependency over the time).

**Keywords:** forecasting, symbolic data analysis, beanplot

## 1 Introduction

Scalar Time Series Forecasting is the use of a statistical model to estimate and forecast the future values of a time series. The research developments in that field were very relevant and important results have been obtained from seventies (De Gooijer and Hyndman (2006)). Anyway there are real cases in which scalar time series do not permit to correctly deal a phenomena, in particular when the dataset contains an huge quantity of observations and the visualization cannot be possible. Another important case could happen when we are interested not in a single value but in a specific distribution of a variable in a given temporal interval (for example Arroyo and Matè (2006) refer on the variable as outcomes the daily time-varying demand of energy). In this cases we are trying to forecast distributions where we force them to be a single value. The case is typical in high frequency financial datasets in which data are collected at a given high frequency (for examples minutes), but sometimes they need to be analyzed at a lower frequency (daily): in this case the need of a statistical aggregation arises naturally. Anyway the aggregation does not measure the intra-day dynamics where data are observed only at some equilibrium levels and is neglected how this equilibrium value is reached (Engle and Russel (2004)). In this case the aggregation does not faithfully represent the underlying phenomenon and a time series of distributions can be more useful than the other forms of aggregated time series (Arroyo and Matè (2009)). In particular already Schweitzer says that: “Distributions are the number of the future!” (Schweizer (1984)) where a possibility, followed

in literature, is to cope directly with the distributions and not on with the original data. Various approaches in that sense are followed to obtain suitable data representations. The estimation of the distribution can be obtained by using parametric or nonparametric methods. Arroyo and Matè (2009) purpose the Histograms as nonparametric method. This approach can be related to the Symbolic Data Analysis (Diday and Noirhomme (2008) Billard and Diday (2000,2006)). In this sense the Symbolic Data Analysis propose an alternative way to manage huge datasets: by transforming the original data in symbolic one as Intervals, Histograms, Lists and so on by retaining the key knowledge. In these symbolic datasets items are described by symbolic variables (Arroyo and Matè (2009)) and the cells can contain distributions (Diday (2002)). In our paper we propose to transform original data in new aggregate data: the beanplots (see also Drago Scepi (2009)). They synthetize the location, represented by their beanline, the size, represented by the difference between minima and maxima, and the shape, that can be considered an estimation of the distribution as well. We purpose in this paper to represent beanplot time series by a peculiar parametrization and to use it for a forecasting aim. This work is organized as follows: in the second paragraph we define the Beanplots time series and how it is possible to handle these types of series and to use these types of data in exploratory data analysis. In the third paragraph we purpose a coherent way to obtain a beanplot parametrization, by using Polynomials. Finally, in the fourth paragraph we introduce the Forecasting procedures for beanplot time series by using VAR (Vector Autoregressive) models.

## 2 Definition of Time Series Beanplot

A time series beanplot is an ordinated sequence of beanplots over the time. Each temporal interval can be considered as a domain of values that is related to the chosen interval temporal (daily, week, and month). The choice of the temporal interval is an a priori choice and depends on the specific data features the analyst wants to study (Drago and Scepi 2009). The beanplot can be considered as a particular case of an interval-valued modal variable at the same time like boxplots and histograms (see Arroyo and Maté (2006)). In a beanplot variable we are taking into account at the same time the intervals of minimum and maximum and the density in form of a kernel nonparametric estimator (the density trace see Kampstra (2008)). The density trace is combined with an 1-d scatterplot where every single dot can be represented for each observation. The beanline can be considered as a measure of the centre and could be represented by the mean, or the median. So the same beanplot can be considered an interval composed by considering the two consecutive intervals through the beanline (the radii of the beanplot).

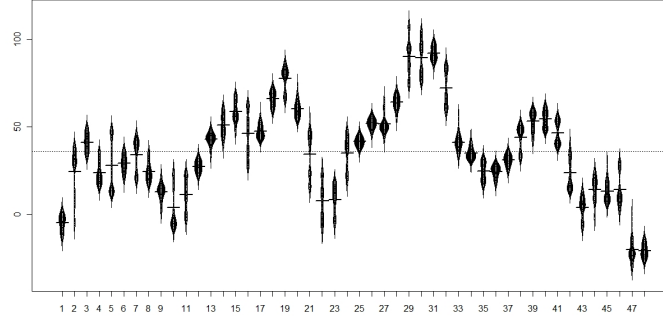
The beanplot is a kernel density plot based on the following kernel density estimate:



$$\hat{f}_{x,h} = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \quad (1)$$

where  $K$  is a Kernel,  $h$  is a smoothing parameter defined as a bandwidth, and  $x_i$  each observation centred and scaled by the factor  $h$ .  $K$  can be a gaussian function with mean zero and variance 1 (See also Hintze and Nelson 1998).

We can consider as fundamental the  $h$  parameter. In fact higher the  $h$  parameter (the bandwidth of the density) and more irregular is the curve. So we need to choose carefully the parameter for the bandwidth. In particular this parameter is obtained by the Sheather-Jones method (see Kampstra (2008)). The beanplot time series show the complex structure of the underlying phenomena by representing jointly the data location (the beanline) the size (the interval minimum and maximum) and the shape (the density trace) over the time. See figure 1, for an example of beanplot time series. In particular the bumps are representing the value of maximum density, and they can show important equilibrium values reached in a single temporal interval (for example to trading purposes). Bumps can also show the intra-period patterns over the time and more in general the beanplot shape show the intra-period dynamics. When the beanplot increases (so increases the difference between minimum and maximum) that can be interpreted with the presence of a structural change on the underlying time series (fig.1). The beanlines permit us to compute the trend for the Beanplot time series. By choosing a suitable temporal interval it is possible to visualize, as well, also intra period seasonality patterns. In general, the beanplots seem to preserve the structure of the time series, but showing additional relevant patterns in data, for example by showing bumps (or equilibrium levels over the time). Another important reason in using the beanplot is that these types of data can show long-run structures where they can summarize an high quantity of data over the time. With respect to other complex objects used in literature beanplots data we are leaving data free to show the empirical structure for each temporal interval, and we obtain a smooth visualization of the underlying phenomena. Histograms and beanplots seem complementary: where histograms can be usefully compared, beanplots tend to show the data structure, and they can show for example observations that could be considered as outliers in a time series. Boxplot can be useful to detect and to identify outliers. In applications: histograms can be useful in setting trading systems where beanplots seem to be very useful in risk management to analyze the occurrences of financial crashes. In every case it is simple to provide a transformation from a data as the beanplot to other symbolic data. For example it is very simple to transform a beanplot time series in an interval-valued time series.



**Fig. 1.** A beanplot time series on a simulated dataset.

### 3 Beanplot parametrization

Our idea is to transform and parametrize each beanplot by using a specific approximation function. We assume that the original beanplot is constituted by two parts: a structured one (defined as a "model") plus a residual (in this sense we follow the approach to histogram approximation in Signoriello (2008)):

$$B_T = M_T + E_T \quad (2)$$

where  $B$  is the beanplot at temporal interval  $T$ ,  $M$  represent the model used, and  $E$  is the residual part. We need to parametrize the structured part and minimize the residual part. In that way we can deal with the densities data by considering directly the model function parameters that could be usefully interpreted.

Each beanplot can be parametrized by considering the mean, the minimum and the maximum of the observed values  $x$  for each  $T$ , as a measure of the location and the size. Furthermore we can consider a parametrization of the density  $y$  as a measure of the shape.

First of all we start to suppose a polynomial regression model as representing the density at interval temporal  $T$  of the beanplot. We estimate a polynomial of order  $n$  or degree  $n - 1$ :

$$y = a_0 + a_1x + a_1x^2 + a_2x^3 + \dots + a_nx^{n-1} + \varepsilon = \sum_{j=1}^n a_jx^{j-1} + \varepsilon \quad (3)$$

where  $y$  represent the density at interval temporal  $T$  of the beanplot and  $x$  is representing the observed values. The obtained curve can be considered as indicator of the beanplot shape.

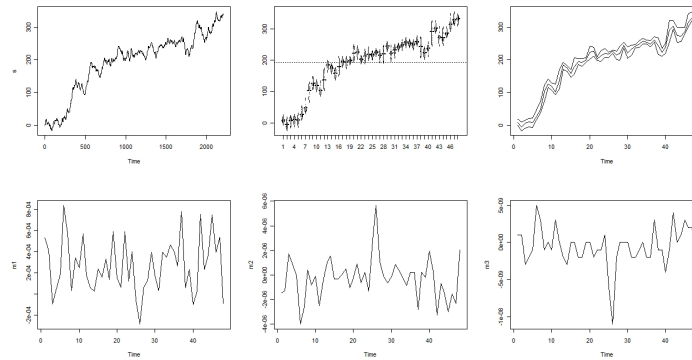
We need to obtain a good model fit for each temporal interval at time  $T$ , so we stop the procedure of parameters estimation by optimizing a specific

stopping criterion in the model selection, as the adjusted  $R^2$ . It is important to note that it is necessary to obtain a minimization of the residuals  $\varepsilon$  but not in every case we are interested in characterizing and replicating the exact structure of the beanplot. Sometime we are interested only in obtaining a specific parametrization in the curve estimation, sometimes we want to characterize the entire density and considering the double or the triple structures due to a lower  $h$  (and representing the structural changes in original data). It is possible in any case to interpretate the coefficient we are obtaining both considering the structure of the single parametrization of the shape for each temporal interval  $T$  and considering the dynamics of the coefficients over the time.

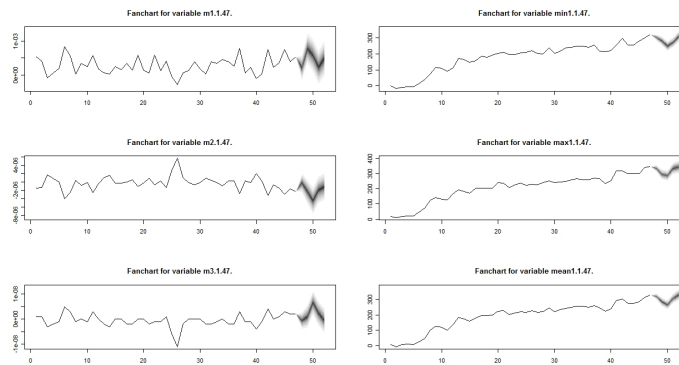
#### 4 Parameters Interpretation: Some Experiments on Simulated Datasets

We simulate various types of time series with different structural characteristics in terms of trend, seasonality, volatility and so on and we compare the different parametrization results for obtaining some general interpretation rules.

For the single beanplot considered, we obtain a minimum, a maximum, a mean (the value of the beanline) and the coefficients representing the curve. We can try to interpretate the parametrization obtained for every  $T$  for example we report the results of a simulation study on  $X$  observations aggregated in  $X$  beanplots with underlined AR(1) model as illustrative example in figure 2. We can observe the original time series (the first graph by starting at top left), the beanplot time series (the second graph) and the dynamical parametrization obtained both the minimum, the maximum (the third graph at the right of the beanplot time series) and (the graphs from the fourth to the six) each term of the polynomial estimated at each temporal interval  $T$  (we have a graph for each of the three coefficients estimated over the time). The single coefficients obtained at time  $T$  are useful to interpretate the different beanplot shapes. In fact, by observing the single beanplot parametrization, we obtain for each time a specific static evaluation of the size, location and shape. In particular the higher order coefficients are very important because they can show some complex structures in data related. The different coefficient over the time show if there is a change in the shapes due to the temporal shocks. In particular we expect a very stable parametrization over the time, and some oscillations from the initial values and where we find a trend in the coefficient dynamics that could be interpreted as a change over the time of the beanplot shapes. For example in figure 2 we can observe a example of parametrization results on simulated data somewhat stable situation in the first beanplots, where at  $T = 7$  we detect a structural change, that could be observed at the same time for all the different parametrizations: that means there is a change of the shape, where the situation tends to stabilize after a



**Fig. 2.** A simulated beanplot time series: time parametrizations.



**Fig. 3.** A simulated beanplot time series: forecasting

specific shock. By a financial level, the increase of the size, can be seen as an increase of the risk, due to the volatility of time series. At the same time, the size can be considered an indicator of risk and volatility where the shape, represents an indicator of market stability. When the shape of the beanplot (and the parametrization) tends to change, that could be a valid indicator of financial instability. The worst scenario is related to a situation where there are more than one bump and an increasing difference between minimum and maximum, that means an increasing instability over the time, probably due to asymmetrical shocks into the financial systems. In some cases and in certain markets we can describe these effects as "domino effects" as an indicator of financial crisis or structural instability.

## 5 Beanplot forecasting

The experiments in this section are related to forecasting, to measure the accuracy of the forecasts, and the capability of the methods and the algorithms, to capture some complex features of financial time series by considering some simulations. Anyway, here, it is very relevant to understand the difference between density forecasting (see Tay and Wallis (2000)) where we are forecasting time series of values with associated interval of confidence (and in this case on the uncertainty of the forecast) and the beanplot forecasting where we want to forecast explicitly the distribution (the beanplot density) in the considered temporal interval (Arroyo and Matè (2009)). At the same time, it is possible to obtain a forecast for the next beanplot (in a similar way as for the prediction in scalar time series) by obtaining forecasts for each different parameter of the beanplot and in particular for the minimum, the maximum, the beanline (the mean) and either for the shape (represented by the curve estimated) of the models of beanplot data. A similar forecasting procedure can be applied to other data in general as boxplots, intervals, and other type of data. In all these cases we extract for each data a suitable possible parametrization (in particular minima and maxima for the intervals and the values representing the boxplots) and we'll use this one to obtain suitable forecasts (see Maia et al. (2008) Arroyo et al. (2009)). So, in the beanplot case we start from a vector with the values of the coefficients each time for the shape, and at the same time the minimum, the maximum and the mean for the location and the shape. We divide the procedure in two distinct parts. In the first part we forecast the mean, the minimum and maximum, where in the second part we forecast the shape by considering the parametrizations of the density. We consider we consider for simplicity's sake a Vector Autoregressive Model (see Lutkepohl 2005) representing the relationships among the location, the size and the shape at different times. Each phase of the forecasting process is validated by using the appropriate statistical tests and indices to verify the forecasting model adequacy of the beanplot time series. In the figure 3 we show the results of the forecasting procedure for the three coefficients (at left) and for the minimum, the maximum and the beanline (at the right). In the charts presented we show the baseline predictions and the interval confidence related. In the first graph at top (figure 3) we can observe at left the forecast for the coefficients of the shape, in the right the forecasts for the minimum, the maximum, and in the third the forecast for the beanline. In general we can observe as forecast a predicted decreasing value of the minimum and the maximum and the mean parameters, where we can expect a changing beanplot shape, visualized by the forecasts on the future dynamics of the model coefficients. These model coefficients can be interpreted by considering the entire fitted regression function so they represents jointly indicators of shape transformations. An higher value mean a big change in the future shapes of the beanplots.

## References

- ARROYO, J. and MATÈ, C. (2006): Introducing Interval Time Series: Accuracy Measures. In *Proceedings in Computational Statistics*. Physica-Verlag, Heidelberg, 1139–1146.
- ARROYO, J. and MATÈ, C. (2009): Forecasting Histogram Time Series with the K-nearest neighbour methods. *International Journal of Forecasting* 25, 192–207.
- ARROYO, J., GONZÁLES-RIVERA, G. and MATÈ, C. (2009): Forecasting with interval and histogram data: Some financial applications. *Working Paper*.
- BILLARD, L. and DIDAY, E. (2006): *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, Chichester.
- BOCK, H.H and DIDAY, E. (eds.) (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin.
- DE GOOIJER, J.G. and HYNDMAN, R.J. (2006): 25 Years of Time Series Forecasting. *International Journal of Forecasting* 22(3), 443–473.
- DIDAY E. (2002): An Introduction to Symbolic Data Analysis and the Sodas Software. *Journal of Symbolic Data Analysis*, 0 (0) ISSN 1723-5081.
- DIDAY E. and NOIRHOMME, M. (2008): *Symbolic data and the SODAS software*. Chichester: Wiley and Sons.
- DRAGO, C. and SCEPI, G. (2009): Univariate and Multivariate Tools for Visualizing Financial Time Series. In Ingrassia S. and Rocci R. (eds.) *Proceedings of Seventh Meeting of the Classification and Data Analysis Group of the Italian Statistical Society* Cleup editore, Catania 481–485.
- ENGLE, R.F. and RUSSEL, J.R. (2004): Analysis of High Frequency Financial Data. *Working Paper*.
- HINTZE, J.L and NELSON (1998): R.D Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, 52, (2), 181–184.
- KAMPSTRA, P. (2008): Beanplot: A Boxplot Alternative for Visual Comparison of Distributions *Journal of Statistical Software*, 28.
- LUTKEPOHL, H. (2005): *New Introduction to Multiple Time Series Analysis*. Springer.
- MAIA A.L.S., DE CARVALHO, F.A.T. and LUDERMIR, T.B.(2008): Hybrid Approach for interval-valued time Series Forecasting. *Neurocomputing*, 71, (16–18), 3344–3352.
- SCHWEIZER B. (1984): Distributions are the numbers of the future. In *Proceedings of the Mathematics of Fuzzy Systems Meeting*. University of Naples, Naples, 137–149.
- SIGNORIELLO S. (2008): Contributions to Symbolic Data Analysis: A Model Data Approach. Ph.D. thesis, Dep. of Mathematics and Statistics, University of Naples Federico II.
- TAY A. and WALLIS K.F. (2000): Density Forecasting: A Survey. *Working Paper*.

# M-estimation in INARCH Models with a Special Focus on Small Means

Hanan El-Saied<sup>1</sup> and Roland Fried<sup>1</sup>

Department of Statistics, TU Dortmund University  
Vogelpothsweg 87, 44221 Dortmund, Germany, {saied,  
fried}@statistik.tu-dortmund.de

**Abstract.** We treat robust M-estimation of INARCH-models for count time series. These models assume the observation at each point in time to follow a Poisson distribution conditionally on the past, with the mean being a linear function of previous observations. This simple linear structure allows to transfer M-estimators for autoregressive models to this situation, with some simplifications being possible because the conditional variance given the past equals the conditional mean. The situation of a small mean deserves special attention because of the strong asymmetry of such Poisson distributions.

**Keywords:** time series, outliers, robustness, Huber M-estimator, Tukey M-estimator

## 1 Introduction

Integer-valued GARCH (briefly: INGARCH) models have been studied by Ferland et al. (2006) and Fokianos et al. (2009). An INGARCH( $p, q$ ) process ( $Y_t : t \geq 1$ ) of orders  $p$  and  $q$  is defined through the relationships

$$Y_t | \mathcal{F}_{t-1} \sim \text{Poisson}(\mu_t),$$
$$\mu_t = \beta_0 + \sum_{j=1}^q \beta_j Y_{t-j} + \sum_{i=1}^p \alpha_i \mu_{t-i}, \quad (1)$$

for  $t \geq 1$ . The dynamics of the process are modeled via the conditional mean  $\mu_t = E(Y_t | \mathcal{F}_{t-1})$  of  $Y_t$ , where  $\mathcal{F}_{t-1}$  stands for the  $\sigma$ -field generated by  $\{Y_{1-q}, \dots, Y_{t-1}, \mu_{1-p}, \dots, \mu_0\}$  representing the whole information up to time  $t-1$ . Here,  $\beta_0 > 0$  is an intercept and  $\beta_j, j = 1, \dots, q$ , and  $\alpha_i, i = 1, \dots, p$ , are non-negative regression parameters. A stationary process fulfilling (1) with mean  $\beta_0 / (1 - \sum_{i=1}^p \alpha_i - \sum_{j=1}^q \beta_j)$  exists if  $\sum_{i=1}^p \alpha_i + \sum_{j=1}^q \beta_j < 1$ . Model (1) closely resembles the GARCH( $p, q$ )-model (Bollerslev (1986)) since the mean of the Poisson distribution equals its variance.

The detection of different types of outliers and intervention effects in INGARCH(1,1) models has been treated by Fokianos and Fried (2010). Our interest here is robust estimation of the parameters of INGARCH models in

the presence of additive outliers, since these can severely influence the maximum likelihood (ML) estimators. We focus on the INGARCH(1,0) model, more briefly called INARCH(1) model, which is the simplest interesting variant of this type of models.

Section 2 reviews M-estimation of location in case of i.i.d. Poisson data and introduces modified Tukey M-estimators with bias correction. Several estimators are compared by simulation in case of clean and outlier-contaminated data. Section 3 extends the idea of M-estimation to the INARCH(1) model and investigates the performance of the resulting generalized M-estimators using simulation. Section 4 concludes.

## 2 M-estimation for independent Poisson variables

Given observations  $y_1, \dots, y_n$ , an M-estimator of a location parameter  $\mu$  is defined as a solution of a minimization problem,

$$\hat{\mu} = \arg \min_m \sum_{t=1}^n \rho(y_t, m), \quad (2)$$

where the function  $\rho$  measures the agreement between an observation  $y_t$  and any possible value  $m$  of  $\mu$ . Using  $\rho(y_t, m) = (y_t - m)^2$  leads to the least squares estimator, while  $\rho(y_t, m) = -\log f_m(y_t)$ , i.e. the negative logarithm of the model density, gives the maximum likelihood estimator (in case of a time series, the conditional density  $f_m(y_t | \mathcal{F}_{t-1})$  of the data given the past is used). If  $\rho$  has a derivative with respect to its second argument, denoted by  $\psi$ , we need to solve the following equation for  $\hat{\mu}$ :

$$\sum_{t=1}^n \psi(y_t, \hat{\mu}) = 0. \quad (3)$$

Popular choices are the Huber and the Tukey  $\rho$  and  $\psi$  functions, which both depend on a tuning constant  $k$ . The Huber  $\psi$  function is given by

$$\psi_k(y_t, m) = \begin{cases} y_t - m, & |y_t - m| \leq k \\ k \cdot \text{sign}(y_t - m), & |y_t - m| > k \end{cases}, \quad (4)$$

whereas Tukey's biweight  $\psi$  function can be written as

$$\psi_k(y_t, m) = \begin{cases} (y_t - m) \left( 1 - \left( \frac{y_t - m}{k} \right)^2 \right)^2, & |y_t - m| \leq k \\ 0, & |y_t - m| > k \end{cases}. \quad (5)$$

The Huber  $\psi$  function is monotone and thus generally leads to a unique solution of (3). The Tukey  $\psi$  function is redescending to zero so that there are possibly several solutions of (3) and we need good initial values to find the



global minimizer of (2). The calculations are commonly done by iteratively reweighted least squares, derived from writing the solution of equation (3) as a weighted mean with weights depending on the distances between the data points and the current solution. For both functions, choosing a larger value of  $k$  increases the efficiency but reduces the robustness to outliers.

If the model distribution  $F$  is a normal distribution with a unit scale, it is reasonable to choose the tuning constant  $k$  of the Huber function within the interval  $[1, 3]$ , since such a distribution rarely generates values with distances from the mean larger than 3 (standard deviations), whereas all values within the range  $[-1, 1]$  are typical. Larger values of  $k$  between 3 and 5 are used for the Tukey function because then  $k$  does not limit the range of typical but the range of plausible observations generated from  $F$ . If  $F$  depends on an unknown scale parameter  $\sigma$ ,  $k$  can be chosen as a corresponding multiple of an estimate  $\hat{\sigma}$ , which can be calculated a-priori or simultaneously.

In case of the Poisson distribution, the variance  $\sigma^2$  equals the mean, so that some modifications are reasonable since we can use the (current) estimate of  $\mu$  for both centering and scaling. A consistent estimator of the Poisson parameter is derived using a modified Huber  $\psi$  function

$$\psi_{k,a}(y_t, m) = \left( \frac{y_t - m}{\sqrt{m}} - a \right) \min \left( 1, \frac{k\sqrt{m}}{|y_t - m - \sqrt{m}a|} \right),$$

see Simpson et al. (1987). The additional constant  $a$  is a bias correction for balancing the symmetric truncation of the asymmetric Poisson distribution by the Huber function. Asymptotically unbiased estimators are obtained by choosing  $a = a(\mu)$  such that

$$E_{\mu} \psi_{k,a}(Y, \mu) = 0.$$

In the iterative calculation of the modified Huber estimator,  $a(\hat{\mu})$  is used, where  $\hat{\mu}$  is the current estimate of  $\mu$ , see Cadigan and Chen (2001). Using the formulas given by these authors we can calculate the asymptotic efficiencies of the modified Huber M-estimators relatively to the sample mean as measured by the ratio of the variances. Figure 1 illustrates that these asymptotic relative efficiencies are quite stable if  $\mu$  is large, say  $\mu > 4$ , while very large efficiencies are difficult to achieve if  $\mu$  is very small. A possibility is to choose  $k$  adaptively, similarly to the choice of  $a$ . In practice this means that even the lower bound 0 of the range of the Poisson distribution lies in the interval where  $\psi_{k,a}$  increases linearly if the current estimate of  $\mu$  is small, say  $\hat{\mu} < 1.5$ , since we should choose  $k$  large then.

Like the Huber function, Tukey's biweight function also treats positive and negative deviations from the mean symmetrically. Therefore we modify Tukey's  $\psi$  function by introducing a bias correction  $a$ , similar to the above

modification of Huber's  $\psi$  function. The modified Tukey  $\psi$  function reads

$$\psi_{k,a}(y_t, m) = \left( \frac{y_t - m}{\sqrt{m}} - a \right) \left( k^2 - \left( \frac{y_t - m}{\sqrt{m}} - a \right)^2 \right)^2 I_{[-k,k]} \left( \frac{y_t - m}{\sqrt{m}} - a \right), \quad (6)$$

where  $I_A$  is the indicator function for a real set  $A$ , and  $a = a(\mu)$  again needs to fulfill  $E_\mu \psi_{k,a}(Y, \mu) = 0$ . In our implementation we search for of a suitable root of (3) which is close to the median of the data.

The right hand side of Figure 1 illustrates the efficiencies of the Huber and Tukey M-estimators for  $n = 50$  and several values of  $k$ , measured by the percentage mean square error relatively to the maximum likelihood estimator, which is the sample mean. The relative efficiencies for several values of  $\mu \in \{0.1, 0.2, 0.3, 0.5, 0.8, 1.3, 2.1, 3.4, 5.5, 8.9, 14.4, 23.3\}$  have been derived from 5000 simulation runs each. If  $\mu$  is very small, the ordinary Huber M-estimator `huberM` in the R-package `robustbase` is not efficient at all because it is strongly biased. The corrected Huber M-estimator with the same  $k$  achieves much better efficiencies there. The corrected Tukey estimator with an adaptive choice of  $k \in \{5, 6\}$  achieves better efficiencies than the versions with one of these values of  $k$  being fixed in case of a Poisson with a small mean. The Huber M-estimator implemented in the function `glmrob` (R-package `robustbase`), which is based on the work of Cantoni and Ronchetti (2001), achieves very stable and good efficiencies for fitting a constant mean.

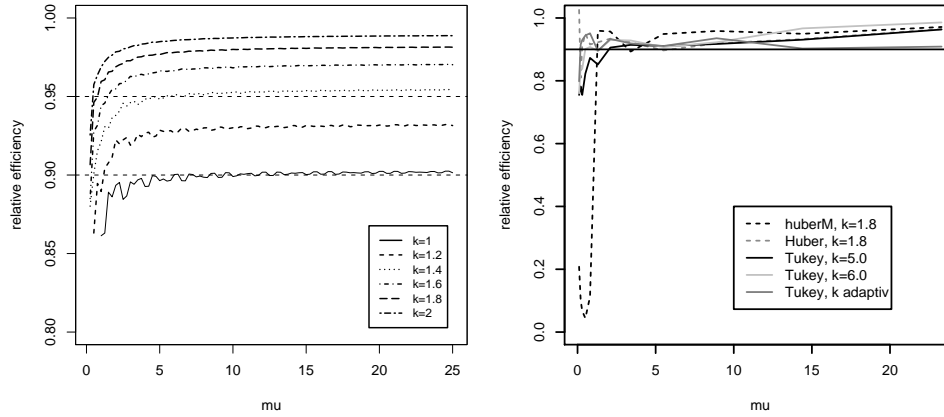
To get some information on the robustness of the estimators we add increasingly large values  $j\sqrt{\mu}$  (rounded to the next integer) to an increasing number  $j$  of observations in a Poisson sample of size  $n = 50$  with mean  $\mu \in \{0.5, 5\}$ . In this way we create increasing numbers of additive outliers of increasing size. Figure 2 illustrates the efficiencies (again relatively to the sample mean) resulting from 2000 simulation runs each on a logarithmic scale. Obviously, in this exercise the corrected Tukey M-estimators perform more robustly than the Huber M-estimators with similar efficiencies, because the former become less biased by many large outliers.

### 3 M-estimation in the INARCH model

Application of conditional likelihood in the INARCH( $p$ ) model, conditioning on the first  $p$  observations, gives the following set of estimation equations:

$$\sum_{t=p+1}^n \left( \frac{y_t}{\mu_t} - 1 \right) \frac{\partial \mu_t}{\partial \theta} = \sum_{t=p+1}^n \left( \frac{y_t - \mu_t}{\sqrt{\mu_t}} \right) \frac{1}{\sqrt{\mu_t}} \frac{\partial \mu_t}{\partial \theta} = 0 \quad (7)$$

where  $\theta = (\beta_0, \beta_1, \dots, \beta_p)$  is the vector of all parameters and  $\partial \mu_t / \partial \theta = (1, y_{t-1}, \dots, y_{t-p})'$ . Downweighting the influence of unusual observations in these equations leads to a straightforward robustification of the conditional

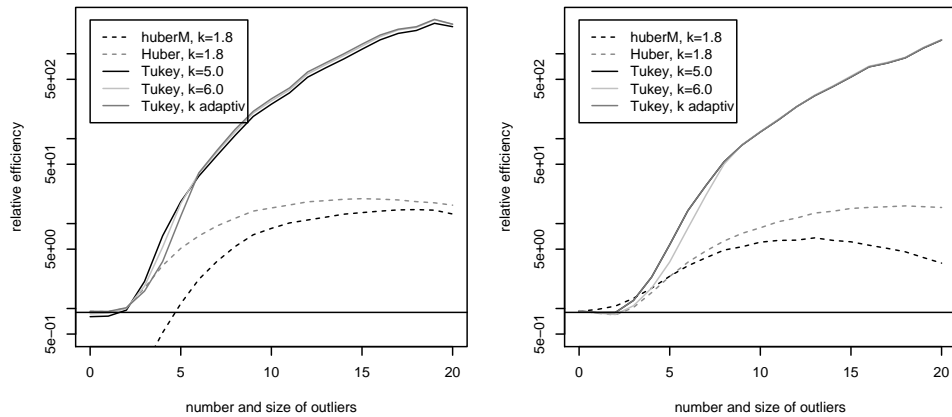


**Fig. 1.** Relative asymptotic efficiencies of the Huber M-estimator relative to the sample mean for several tuning constants (left) and relative efficiencies of the Huber and the Tukey M-estimator for sample size  $n = 50$  (right), as a function of the true mean for several tuning constants  $k$ .

likelihood estimators. For this, we truncate observations with large standardized residuals  $(y_t - \mu_t)/\sqrt{\mu_t}$  using Huber's or Tukey's  $\psi$  function, and do the same with regressors  $y_{t-1}, \dots, y_{t-p}$  which are outlying w.r.t. the marginal distribution. This leads us to the following set of estimating equations:

$$\sum_{t=p+1}^n \psi\left(\frac{y_t - \mu_t}{\sqrt{\mu_t}}\right) \frac{1}{\sqrt{\mu_t}} \begin{pmatrix} 1 \\ \sigma\psi\left(\frac{y_{t-1} - \mu}{\sigma}\right) + \mu \\ \vdots \\ \sigma\psi\left(\frac{y_{t-p} - \mu}{\sigma}\right) + \mu \end{pmatrix} = 0, \quad (8)$$

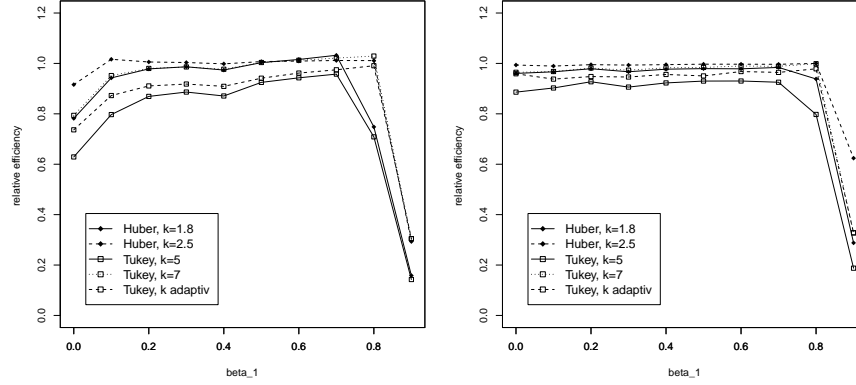
where  $\mu$  and  $\sigma^2$  are the marginal mean and variance for the given set of parameters, respectively. We focus our attention on the basic case of the INARCH(1) model in the following. The marginal mean and variance in this model are  $\mu = \beta_0/(1 - \beta_1)$  and  $\sigma^2 = \mu(1 + \beta_1^2/(1 - \beta_1^2))$ , see Fokianos et al. (2009).



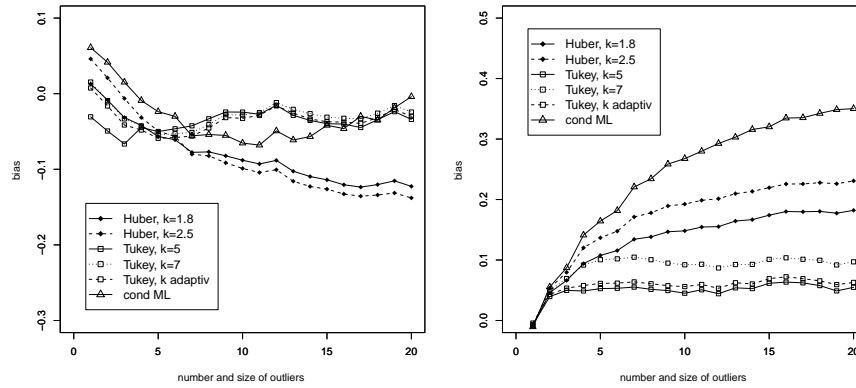
**Fig. 2.** Efficiencies relatively to the sample mean of different implementations of Huber and Tukey M-estimators for  $\mu = 0.5$  (left) and for  $\mu = 5$  (right) in case of an increasing number  $j$  of outliers of size  $[j\sqrt{\mu}]$  for different values of the tuning constant  $k$ .

Figure 3 illustrates the efficiencies of the arising (generalized) Huber and Tukey M-estimators for different values of  $k$  in case of a sample size  $n = 100$ ,  $\beta_0 = 1$  and  $\beta_1 \in \{0, 0.1, \dots, 0.9\}$ , derived from 2000 data sets for each value of  $\beta_1$ . The estimators generally achieve good efficiencies, but have some problems as  $\beta_1$  approaches 1, which is the non-stationary case.

To inspect the robustness of the estimators we consider an increasing number  $j$  of additive outliers of increasing size  $[j\sigma]$  (rounded multiples of the marginal standard deviation) in a time series of length  $n = 100$  from the INARCH(1) model with  $\beta_0 = 1$  and  $\beta_1 = 0.4$ . The bias curves generated from 1000 time series in Figure 4 confirm the good robustness properties of the Tukey M-estimator. The conditional ML estimator shows little bias for the intercept  $\beta_0$  in this situation because  $\beta_1$  absorbs almost all of the outlier effect. The function `glmrob` often produced errors because of invalid starting values when using it for fitting an INARCH(1) model. For our own functions we use  $(\hat{\mu}, 0.001)$  to initialize the estimator  $(\hat{\beta}_0, \hat{\beta}_1)$ , where  $\hat{\mu}$  is the corresponding corrected Huber or Tukey M-estimator of the marginal



**Fig. 3.** Relative efficiencies for  $\beta_0$  (left) and  $\beta_1$  (right) relatively to the conditional maximum likelihood estimator as a function of  $\beta_1 = 0, 0.1, \dots, 0.9$ , with  $\beta_0 = 1$  and sample size  $n = 100$ .



**Fig. 4.** Bias in case of an increasing number  $j$  of outliers of increasing size  $[j\sigma]$  for  $\beta_0$  (left) and for  $\beta_1$  (right) in case of  $\beta_0 = 1$  and  $\beta_1 = 0.4$ , sample size  $n = 100$ .

mean with the same tuning constant, which have been discussed in Section 2. Application of this idea to provide feasible starting values to the function `glmrob` gave results similar to those for our implementation of the Huber M-estimator.

## 4 Conclusions

We have constructed Tukey M-estimators with bias correction for the Poisson parameter in case of independent data, as well as generalized Huber and

Tukey M-estimators for the parameters of INARCH(1) models. The Tukey M-estimators show good performance in our simulations, except if the regression coefficient describing the influence of the previous observation is close to unity. The implementation of suitable bias corrections in the time series case and the extension of the estimators to general INGARCH models is the scope of ongoing work.

## References

- BOLLERSLEV, T. (1986): Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307-327.
- CADIGAN, N.G. and CHEN, J. (2001): Properties of robust M-estimators for Poisson and negative binomial data. *Journal of Statistical Computation and Simulation* 70, 273-288.
- CANTONI, E. and RONCHETTI, E. (2001): Robust inference for generalized linear models. *Journal of the American Statistical Association* 96, 1022-1030.
- FERLAND, R., LATOUR, A. and ORAICHI, D. (2006): Integer-valued GARCH processes. *Journal of Time Series Analysis* 27, 923-942.
- FOKIANOS, K., RAHBK, A. and TJØSTHEIM, D. (2009): Poisson autoregression. *Journal of the American Statistical Association* 104, 1430-1439.
- FOKIANOS, K. and FRIED, R. (2010): Interventions in INGARCH processes. *Journal of Time Series Analysis*, forthcoming.
- SIMPSON, D.G., CARROLL, R.J. and RUPPERT, D. (1987): M-estimation for discrete data: asymptotic distribution theory and implications. *Annals of Statistics* 15, 657-669.

# Score Moment Estimators

Zdeněk Fabián<sup>1</sup>

Institute of Computer Science, Academy of Sciences of the Czech republic  
Pod vodárenskou věží 2, 182 00 Prague [zdenek@cs.cas.cz](mailto:zdenek@cs.cas.cz)

**Abstract.** Thanks to the application of newly introduced concept of the scalar score, the score moments are introduced and used for parametric estimation. In cases of heavy-tailed distributions, the variances of score moment estimates are slightly higher than variances of the maximum likelihood estimates, but the estimates of all parameters are robust.

**Keywords:** generalized moments, score moments, robust estimators

## 1 Introduction

Let  $X$  be random variable with continuous distribution  $F$  with support  $\mathcal{X} \subseteq \mathbb{R}$  and probability density  $f$ . Its usually used numerical characteristics are the mean  $\mu_1 = EX = \int_{\mathcal{X}} xf(x) dx$  and the central moments  $\mu_k = E(X - EX)^k, k = 2, 3, \dots$  Unfortunately, moments of heavy-tailed distributions may not exist. The method of moments is the oldest method of finding points estimators, having the virtue of being quite simple to use. Unfortunately, moment estimators often have low efficiencies.

If  $\varphi$  is a continuous monotone function,  $E\varphi^k(X)$  represents generalized moments. We show that the generalized moment estimators, function  $\varphi$  of which is the scalar score, can be used even in cases of distributions with infinite moments, and that could be a good alternative to the maximum likelihood estimators.

## 2 Scalar Score

Let  $\eta : \mathcal{X} \rightarrow \mathbb{R}$  be an increasing continuous mapping. Based on old idea of Johnson (1949), Fabián (2001) suggested to view any distribution with support  $\mathcal{X} \neq \mathbb{R}$  as a transformed “prototype” distribution  $G$  with support  $\mathbb{R}$ . Using the lesson drawn from Hampel et al. (19860), as an important function of  $G$  was identified the *score function*

$$S_G(y) = -\frac{1}{g(y)} \frac{dg(y)}{dy}, \quad (1)$$

describing the relative change of its density with respect to the probability.

As an important function describing the transformed distribution

$$F(x) = G(\eta(x)), \quad x \in \mathcal{X} \quad (2)$$

was identified by Fabián (2001) the transformed score function of the prototype,

$$T_F(x) = S_G(\eta(x)). \quad (3)$$

From (1) and (2),

$$T_F(x) = -\frac{1}{f(x)} \frac{d}{dx} \left( \frac{1}{\eta'(x)} f(x) \right), \quad (4)$$

where  $\eta'(x) = d\eta(x)/dx$  is the Jacobian of the transformation.

For comparison of properties of different distributions, it turned out to be necessary to use one concrete  $\eta$  for all distributions with the given support. According to the principle of parsimony, that one providing the simplest mathematical forms of (3) for a large amount of distributions used in statistical practice should be used. According to Johnson (1949),  $\eta$  was defined as

$$\eta(x) = \begin{cases} x & \text{if } \mathcal{X} = \mathbb{R} \\ \log(x-a) & \text{if } \mathcal{X} = (a, \infty) \\ \log \frac{(x-a)}{(b-x)} & \text{if } \mathcal{X} = (a, b) \end{cases} \quad (5)$$

so that, for instance, for distributions supported by  $\mathcal{X} = (a, \infty)$

$$T_F(x) = -1 - (x-a)f'(x)/f(x).$$

Function (3) with  $\eta$  given by (5) is called the *transformation-based score* or shortly the *t-score*. It expresses the relative change of “basic component”  $g(\eta(x))$  of the density of distribution  $F$  (remind that  $f(x) = g(\eta(x))\eta'(x)$ ).

The t-scores of two particular classes of distributions are well known in statistics. The t-score (score function) of distribution  $G(y-\mu)$  on  $\mathbb{R}$  with location parameter  $\mu$  equals the maximum likelihood score function for  $\mu$ . Indeed,  $S_G(y-\mu) = \frac{\partial}{\partial \mu} \log g(y-\mu)$ . The transformed distribution  $F(x; \mu) = G(\log x - \mu)$  with support  $\mathcal{X} = (0, \infty)$  is, by setting

$$\tau = \eta^{-1}(\mu) = \exp(\mu), \quad (6)$$

a log-location distribution  $F(x; \tau) = G(\log x - \log \tau) = G(\log \frac{x}{\tau})$  (c.f. Lawless, 2003). According to Fabián (2001), Theorem 1, the t-score of a log-location distribution is proportional to its maximum likelihood score for  $\tau$ ,

$$\eta'(\tau)T_F(x; \tau) = \frac{\partial}{\partial \tau} \log f(x; \tau). \quad (7)$$

On the other hand, the t-scores of other distributions are new functions.



As a characteristic of central tendency of distributions was introduced by Fabián (2001, 2007) the solution  $x^*$  of equation

$$T_F(x) = 0$$

(which exists and is unique if the density of the prototype is unimodal), called the *transformation-based mean* or shortly the *t-mean*. The t-mean of location distributions is  $\mu$ , the t-mean of log-location distributions is  $\tau$  (6), for other distributions t-mean is a new characteristic which can be used instead of the mean.

Parameter  $\tau$  given by (6) is the transformed location of the prototype. This fact was used by generalizing function (7) for arbitrary distributions by using the t-mean instead of  $\tau$ . Function

$$S(x) = \eta'(x^*)T_F(x) \quad (8)$$

which we call a *scalar score*, is the score for t-mean either  $x^*$  is a real parameter of the distribution or not. It is a new function which has a similar sense as the vector of likelihood scores for location and log-location distributions: it is an inference function which can be used for adapting the data to the assumed model.

### 3 Score Moments

**Definition.** Let  $X$  be a random variable with distribution  $F$  supported by  $\mathcal{X}$ , with density  $f$  and scalar score  $S$  given by (8). For every  $k \in \mathcal{N}$ , define the  $k$ -th order score moment by

$$ES^k = \int_{\mathcal{X}} S^k(x) f(x) dx.$$

Score moments of distributions with linear scores have affinity to central moments. The score function of the normal distribution  $\mathcal{N}(\mu, \sigma)$  is  $S(x) = S_G(x; \mu, \sigma) = \frac{x - \mu}{\sigma^2}$  so that  $ES^k = \frac{1}{\sigma^{2k}} \mu_k$ . Particularly,  $ES^2 = 1/\sigma^2$  is Fisher information for  $\mu$ . The score of the exponential distribution with density  $f(x; \tau) = \frac{1}{\tau} e^{-x/\tau}$  is  $S(x; \tau) = \frac{1}{\tau} T_F(x; \tau) = \frac{1}{\tau} (\frac{x}{\tau} - 1)$  so that  $ES^k = \frac{1}{\tau^{2k}} \mu_k$ . Particularly,  $ES^2 = 1/\tau^2$  is Fisher information for  $\tau$ .

The score moments of other distributions are new characteristics. Analogically to the location and log-location distributions,  $ES^2$  can be considered as Fisher information for t-mean of an arbitrary continuous distribution. Assuming that  $F$  is regular in the sense that  $0 < ES^2 < \infty$  (it corresponds to the usual regularity conditions), its reciprocal value

$$\omega^2 = \frac{1}{ES^2} \quad (9)$$

can be considered to be a characteristic of variability of distribution  $F$ . It was introduced by Fabián (2007) under the name t-variance, but it seems better to call it the *score variance*.

The t-score, t-mean and the second t-score moment of some distributions are given in Table 1. Scalar scores can be simply determined by (8) and  $\omega^2$  by (9).

Model	$\mathcal{X}$	$f(x; \theta)$	$T_F(x; \theta)$	$x^*$	$ET_F^2(\theta)$
Weibull	$(0, \infty)$	$\frac{c}{x} \left(\frac{x}{\tau}\right)^c e^{-\left(\frac{x}{\tau}\right)^c}$	$c\left(\left(\frac{x}{\tau}\right)^c - 1\right)$	$\tau$	$c^2$
Fréchet	$(0, \infty)$	$\frac{c}{x} \left(\frac{x}{\tau}\right)^{-c} e^{-\left(\frac{x}{\tau}\right)^{-c}}$	$c\left(1 - \left(\frac{x}{\tau}\right)^c\right)$	$\tau$	$c^2$
gamma	$(0, \infty)$	$\frac{\gamma^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\gamma x}$	$\gamma x - \alpha$	$\alpha/\gamma$	$\alpha$
inv. gamma	$(0, \infty)$	$\frac{\gamma^\alpha}{x\Gamma(\alpha)} x^{-\alpha} e^{-\gamma/x}$	$\alpha - \gamma/x$	$\gamma/\alpha$	$\alpha$
loggamma	$(1, \infty)$	$\frac{\gamma^\alpha}{\Gamma(\alpha)} \frac{(\log x)^{\alpha-1}}{x^{\gamma+1}}$	$\gamma \log x - \alpha$	$e^{\alpha/\gamma}$	$\alpha$
log-logistic	$(0, \infty)$	$\frac{c}{x} \frac{(x/\tau)^c}{((x/\tau)^c + 1)^2}$	$c \frac{(x/\tau)^c - 1}{(x/\tau)^c + 1}$	$\tau$	$\frac{c^2}{3}$
beta prime	$(0, \infty)$	$\frac{1}{B(p, q)} \frac{x^{p-1}}{(x+1)^{p+q}}$	$\frac{qx-p}{x+1}$	$p/q$	$\frac{pq}{p+q+1}$
Burr	$(0, \infty)$	$\frac{kc}{x} \frac{x^c}{(x^c+1)^{k+1}}$	$\frac{c(kx^c-1)}{x^c+1}$	$1/k^{1/c}$	$c^2 \frac{k}{k+2}$
Cauchy	$\mathbb{R}$	$\frac{1}{\sigma\pi} \frac{1}{1+\left(\frac{x-\mu}{\sigma}\right)^2}$	$\frac{1}{\sigma} \frac{2\frac{x-\mu}{\sigma}}{1+\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mu$	$\frac{1}{2\sigma^2}$

Table 1. The t-score, t-mean and  $ET_F^2$  of some distributions

## 4 Estimation

Let  $\{\mathcal{F}_\theta : \theta \in \Theta \subseteq \mathbb{R}^m\}$  be a parametric family of distributions and  $X_1, \dots, X_n$  random variables iid according  $F_\theta \in \mathcal{F}_\theta$  with  $\theta$  unknown. The scalar function  $S$  can be used for finding estimators of true  $\theta$  even if  $m > 1$ , by the use of the generalized moment method. According to Fabián (2001), the score moment estimator is a solution of the system of equations

$$\hat{\theta}_M : \quad \frac{1}{n} \sum_{i=1}^n S^k(X_i; \theta) = ES^k(\theta), \quad 1 \leq k \leq m. \quad (10)$$

The score moment estimators have promising properties:

i) According to (8), for the purposes of estimating parameters it suffices to consider the t-score moments. For two-parameter distributions with  $\theta = (\theta_1, \theta_2)$ , for instance, equations (10) turn into

$$\sum_{i=1}^n T(x_i; \theta_1, \theta_2) = 0 \quad (11)$$

$$\frac{1}{n} \sum_{i=1}^n T^2(x_i; \theta_1, \theta_2) = ET^2(\theta_1, \theta_2). \quad (12)$$

ii) t-score moments of related distributions are identical, i.e., for  $F$  and  $G$  related by (2),  $ET_F^k = ET_G^k$  (Fabián (2001), Proposition 1).

iii)  $\hat{\theta}_M$  exists, is strongly consistent and asymptotically normal with the asymptotic variance-covariance matrix given by Fabián (2001).

iv) It follows from (1) that if  $g(y) = O(e^{-y})$  then  $T_G(y) = O(1)$ . According to ii), t-scores of heavy-tailed distributions are bounded. The score moment estimator is sensitive to observations far from the bulk of the data if the distribution is light-tailed and it is robust if the distribution is heavy-tailed. Since the variable occurs in equations (10) only in powers of the t-score, it holds true for all the parameters.

v) The scores are functions matching distributions so that the score moments are often expressed by elementary functions of parameters.

vi) For simple heavy-tailed distributions, the asymptotic relative efficiencies of the score moment estimators are reasonably near to one (Fabián (2001)). Variances of the score moment estimates in our simulation experiments described below were only little larger than variances of maximum likelihood estimates.

Equations (10) are to be solved by an iterative way. For some distributions, equations (11) and (12) separate into individual ones, yielding in some cases solutions in closed formulas. Let us present some examples.

In the case of Weibull distribution we obtain  $\hat{c}$  from (12)

$$n \sum_{i=1}^n x_i^{2c} = 2 \left( \sum_{i=1}^n x_i^c \right)^2$$

and then  $\hat{\tau} = \left( \frac{1}{n} \sum_{i=1}^n x_i^c \right)^{1/\hat{c}}$  from (11). Analogically, in the case of Fréchet distribution one obtains  $\hat{c}$  from

$$n \sum_{i=1}^n 1/x_i^{2c} = 2 \left( \sum_{i=1}^n 1/x_i^c \right)^2$$

and  $\hat{\tau} = \left( \frac{1}{n} \sum_{i=1}^n 1/x_i^{\hat{c}} \right)^{1/\hat{c}}$ .

In the case of beta-prime distribution we obtain from (11)

$$\hat{x}^* = \widehat{p/q} = \frac{\sum_{i=1}^n x_i/(1+x_i)}{\sum_{i=1}^n 1/(1+x_i)}$$

and then from (12)  $\hat{q} = (\hat{x}^*/\rho - 1)/(\hat{x}^* + 1)$ , where  $\rho = n / \sum_{i=1}^n \left( \frac{x_i - \hat{x}^*}{x_i + 1} \right)^2$ .

In the case of inverted gamma distribution one obtains from (11)  $1/\hat{x}^* = \frac{1}{n} \sum_{i=1}^n 1/x_i = 1/\bar{x}_h$ , where  $\bar{x}_h$  is the harmonic mean. From (12)  $1/\hat{\alpha} = \bar{x}_h^2 / \bar{x}_{2h} - 1$ , where  $1/\bar{x}_{2h} = \frac{1}{n} \sum_{i=1}^n 1/x_i^2$  and  $\hat{\gamma} = \hat{\alpha} \bar{x}_h$ .

In the case of log-gamma distribution, an alternative mapping  $\eta : (1, \infty) \rightarrow \mathbb{R}$  in the form  $\eta(x) = \log \log x$  can be used. As  $\eta'(x) = 1/(x \log x)$ , the “loglog” t-score is, by (3),

$$T(x) = \frac{1}{f(x)} \frac{d}{dx} (-(\log x)^\gamma x^{-\alpha}) = \gamma \log x - \alpha.$$

From (11) we obtain  $\hat{x}^* = \hat{\alpha}/\gamma = \bar{x}_l$ , where  $\bar{x}_l = \frac{1}{n} \sum_{i=1}^n \log x_i$  and from (12)  $1/\hat{\alpha} = \bar{x}_{2l}/\bar{x}_l^2 - 1$ , where  $\bar{x}_{2l} = \frac{1}{n} \sum_{i=1}^n \log^2 x_i$ .

## 5 Estimation of the Threshold Parameter

Consider the uniform distribution on  $[0, \gamma]$  with  $\gamma$  unspecified. The maximum likelihood estimator of  $\gamma$  is  $\hat{\gamma}_{ML} = x_{(n)} = \max(x_1, \dots, x_n)$ . According to (5),  $\eta(x) = \log \frac{x}{\gamma-x}$  so that the t-score of the uniform distribution is

$$T(x) = \frac{d}{dx} \left( -\frac{x(\gamma-x)}{\gamma} \right) = \frac{2x}{\gamma} - 1.$$

By (11),  $\frac{1}{n} \sum_{i=1}^n \frac{2x_i}{\gamma} = 1$ . The score moment solution in form

$$\hat{\gamma}_M = \max(x_{(n)}, 2\bar{x})$$

takes into account that the cut-off can be higher than the largest observed value. For  $n = 5, 10, 20$  and  $50$  we obtained after 10 000 experiments  $\hat{\gamma}_{ML} \approx 0.87, 0.91, 0.95$  and  $0.98$ , respectively, whereas  $\hat{\gamma}_M = 1$  with accuracy to three decimal points.

In Table 2 are listed two distributions with the threshold parameter.

Distribution	$\mathcal{X}$	$f(x)$	$T_F(x)$	$x^*$	$ET_F^2(\theta)$
exponential	$(\gamma, \infty)$	$\frac{1}{\tau} e^{-\frac{x-\gamma}{\tau}}$	$\frac{(x-\gamma)}{\tau} - 1$	$\gamma + \tau$	1
Pareto	$(\gamma, \infty)$	$\frac{c\gamma^c}{x^{c+1}}$	$c(1 - x^*/x)$	$\frac{c+1}{c}$	$\frac{c}{c+2}$

Table 2. t-score, t-mean and  $ET_F^2(\theta)$  of some distributions.

In the exponential case,  $\hat{\tau} = (\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2)^{1/2}$  from (12) and  $\hat{\gamma} = \bar{x} - \hat{\tau}$  from (11). In the Pareto case  $\hat{x}^* = \bar{x}_h$  from (11). Denoting  $1/\rho = \bar{x}_h^2/\bar{x}_{2h} - 1$ , we obtain  $\hat{c} = \sqrt{1 + \rho} - 1$  from (12) and  $\hat{\gamma} = \bar{x}_h \hat{c}/(\hat{c} + 1)$ . In both cases, the solution

$$\hat{\gamma}_M = \min(\hat{\gamma}, x_{(1)})$$

takes into account that the cut-off can be less than the lowest observed value.

## 6 Simulation Experiments

In simulation experiments, score moment estimators were compared with the maximum likelihood ones. Samples of length  $n = 100$  were generated from contaminated distribution  $\Phi(x) = 0.9F(\omega_0) + 0.1F(\omega)$ , where  $\omega^2$ , given by (9), was taken as the measure of variability of  $F$ . Average values of the estimates were computed after 1 000 replications of every experiment.

Due to a linear t-score, the score moment estimator of the parameters of beta distribution equals to the ordinary moment estimator. However, since the t-score is bounded on  $\mathcal{X} = (0, 1)$ , the moment estimates should be less influenced by “outlying” values near zero and one, generated from U-shaped  $F(\omega)$  for large  $\omega$  than the maximum likelihood estimates. It is, surprisingly, the case, as it is apparent from Figure 1.

Average values of the maximum likelihood and score moment estimates of the t-mean of beta-prime distribution with  $x^* = 1$  and Pareto distribution with  $x^* = 1.66$  together with standard deviations of estimates in the Pareto case are compared in Figure 1. Score moment estimates are less distorted by large values generated from  $F(\omega)$ , but have larger variances for small  $\omega$ .

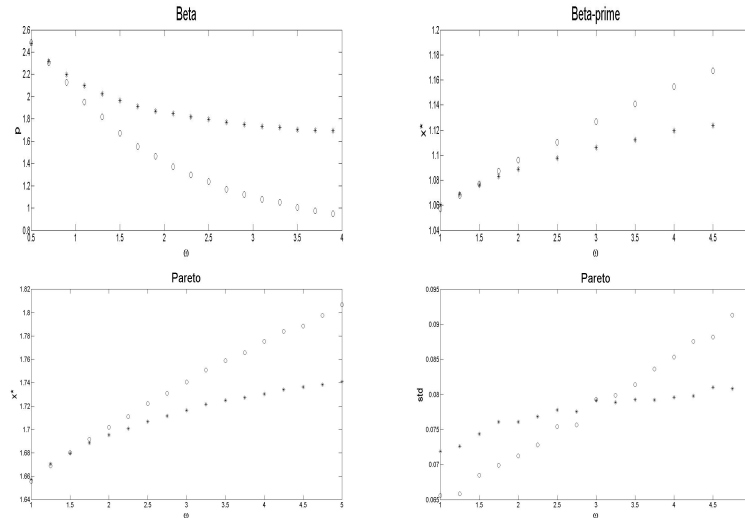


Fig. 1. Plots of average maximum likelihood (o) and score moment (\*) estimates as functions of increasing variability of distributions

In Figure 2 are plotted average maximum likelihood and score moment estimates of the threshold parameter of distributions from Table 2 as functions of  $\omega$ . Whereas the score moment estimator is undoubtedly better for Pareto, in the exponential case one could use, as a rule of thumb,  $\hat{\gamma} = \frac{1}{3}\hat{\gamma}_{ML} + \frac{2}{3}\hat{\gamma}_M$ .

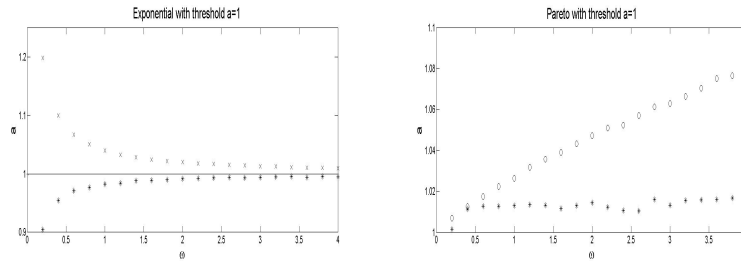


Fig. 2. Plots of average maximum likelihood (o) and score moment (\*) estimates of the threshold parameter of exponential and Pareto distributions

## 7 Conclusions

The score moment estimators seem to be a good alternative to the maximum likelihood estimators. They are often simple, for heavy-tailed distributions reasonably efficient and robust for all the parameters.

**Acknowledgement.** This work was supported by projects AV0Z10300504 and GACR 205/09/1079.

## References

- 1.FABIÁN, Z. (2001): Induced cores and their use in robust parametric estimation. *Comm. Statist. Theory Methods* 30, 537-556.
- 2.FABIÁN, Z. (2007): Estimation of simple characteristics of samples from skewed and heavy-tailed distribution. In C. Skiadas (Ed.): *Recent Advances in Stochastic Modeling and Data Analysis*. World Scientific, Singapore, 43-50.
- 3.FABIÁN, Z. (2008): New measures of central tendency and variability of continuous distributions. *Comm. Statist. Theory Methods* 37, 159-174.
- 4.FABIÁN, Z. and STEHLÍK, M. (2008): A note on favorable estimation when data is contaminated. *Comm. Dependability and Quality Management* 11, 36-43.
- 5.FABIÁN, Z. (2009): Confidence intervals for a new characteristic of central tendency of distributions. *Comm. Statist. Theory Methods* 38, 1804-1814.
- 6.HAMPEL, F. R., ROUSSEEUW, P. J., RONCHETTI, E. M. and STAHEL, W. A. (1986): *Robust Statistic. The Approach Based on Influence Functions*, Wiley, New York.
- 7.JOHNSON, N.L. (1949): Systems of frequency curves generated by methods of translations. *Biometrika* 36, 149-176.
- 8.JOHNSON, N. L., KOTZ, S., BALAKRISHNAN, N. (1994, 1995): *Continuous univariate distributions 1, 2*. Wiley, New York.
- 9.LAWLESS J.F., (2003): *Statistical models and methods for lifetime data*, 2nd ed. Wiley, Hoboken.

# Testing the Number of Components in Poisson Mixture Regression Models

Susana Faria<sup>1</sup> and Fátima Gonçalves<sup>2</sup>

<sup>1</sup> Department of Mathematics and Applications, Mathematical Research Centre ,  
University of Minho, 4800-058 Guimarães, Portugal *sfaria@math.uminho.pt*

<sup>2</sup> University of Minho, 4800-058 Guimarães, Portugal *fat.rod.goncalves@sapo.pt*

**Abstract.** Estimating the number of mixture components is one of the major difficulties in the application of finite mixture models. The likelihood ratio test is a general statistical procedure to use. Unfortunately, a number of specific problems arise and the classical theory fails to hold. In this paper we investigate the testing of hypotheses concerning the number of components in Poisson regression models (PMR) via parametric and nonparametric bootstrap. We also compare the performance of these procedures with criteria AIC and BIC in testing the number of components in these models.

**Keywords:** EM algorithm, mixture Poisson regression models, likelihood ratio test, resampling

## 1 Introduction

Finite mixture models are a well-known method for modelling unobserved heterogeneity (see e.g. McLachlan and Peel (2000) and Fruhwirth-Schnatter (2006) for a review). In particular, Poisson mixture regression models (PMR) are commonly used to analyze heterogeneous count data.

Let the random variable  $Y_i$  denote the  $i$ th response variable, and let  $(\mathbf{x}_i, y_i), i = 1, \dots, n$  denote observations where  $y_i$  is the observed value of  $Y_i$  and  $\mathbf{x}_i$  a  $(p+1)$ -dimensional covariate vector. It is assumed that the marginal distribution of  $Y_i$  follows a mixture of Poisson distributions,

$$Y_i \sim \sum_{j=1}^J \pi_j f_j(y_i | \mathbf{x}_i, \lambda_{i|j}) \quad (1)$$

where

$$f_j(y_i | \lambda_{i|j}) = \frac{\exp(-\lambda_{i|j})(\lambda_{i|j})^{y_i}}{y_i!}, \quad i = 1, \dots, n, j = 1, \dots, J \quad (2)$$

and  $\lambda_{i|j} = \exp(\boldsymbol{\beta}_j^T \mathbf{x}_i)$ , with  $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jp})^T$  denoting the  $(p+1)$ -dimensional vector of regression coefficients for  $j$ th component. The proportions  $\pi_j$  are the mixing probabilities ( $0 < \pi_j < 1$ , for all  $j = 1, \dots, J$  and

$\sum_j \pi_j = 1$ ) and can be interpreted as the unconditional probability that an observation arises from component  $j$  of the mixture.

An important but difficult problem in practice is to determine the number of components that fits data the best. A common method of testing for the number of components in a model is the likelihood ratio test (LRT). Unfortunately, the limit distribution of the likelihood ratio statistic does not follow the usual  $\chi^2$  distribution (see McLachlan and Peel (2000)).

One way to overcome this problem is to use parametric bootstrap techniques (see e.g. Karlis and Xekalaki (1999), Schlattmann (2005)). Using these techniques implies simulating from a certain mixture model with  $k$  components under  $H_0$ . For these simulated data a model with  $k$  and  $k + 1$  components is fitted allowing to compute the likelihood ratio statistic (LRS) for this sample. Replicating this procedure  $B$  times provides the distribution of LRS for testing  $k$  against  $k + 1$  components.

Another approach to determine the appropriate number of components is based in a penalized form of the log likelihood, yielding what are called information criteria. Two widely used information criteria are the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Wang et al. (1996) discuss the use of AIC and BIC to determine the number of components in a PMR. Even though they show that both AIC and BIC are reliable methods, their study suggests BIC is more reliable and thus its usage is more recommendable. When AIC misidentified the correct number of components it always yielded a model with too many components suggesting that AIC may underpenalize the number of parameters in the mixtures.

In this work we investigate the testing of hypotheses concerning the number of components in a PMR via a nonparametric bootstrap by resampling from the residuals of the null model. A similar procedure was suggested by Turner (2000) to determine the number of components in mixtures of Gaussian regression models. The performance of this approach is compared with a parametric bootstrap where the bootstrap samples are generated by random selection with replacement. A comparison between bootstrap techniques and criteria AIC and BIC is also performed.

The paper is organized as follows: in Section 2, a procedure based on a nonparametric bootstrap for determining the number of components in a PMR is described. Section 3 provides a simulation study investigating the performance of the bootstrap techniques and criteria AIC and BIC in testing the number of components in a PMR. The performance of these techniques in real data sets are studied in section 4. In Section 5 the conclusions of the study are drawn and additional comments are made.

## 2 Hypotheses testing

Consider the hypothesis  $H_0$ : the number of components in a PMR is  $k$  versus the hypothesis  $H_1$ : the number of components in the mixture is  $k + 1$ . The



test statistic is  $Q = -2 \log \lambda$  where  $\lambda$  represents the likelihood ratio. Due to the non-applicability of the standard asymptotic theory to mixture models, we adopt a resampling approach to compute the  $p$ -value associated with the LRS. It consists of the following steps:

- a. Fit Poisson regression models with  $k$  and  $k + 1$  components to original data.
- b. Calculate the log-likelihood ratio statistic  $Q$ .
- c. For  $b = 1, \dots, B$ :
  - create a bootstrap data  $(\mathbf{x}_i, y_i^*)$  where

$$y_i^* = \text{round}(\hat{\lambda}_{i|j} + r_i^*) \quad (3)$$

where  $\text{round}(\cdot)$  is the function that rounds a value to the nearest integer,  $\hat{\lambda}_{i|j} = \exp(\hat{\beta}_j^T \mathbf{x}_i)$  and  $r_i^*$  are the residuals given by

$$r_i^* = y_i - \hat{\mu}_i \quad (4)$$

with

$$\hat{\mu}_i = \sum_{j=1}^J \hat{\pi}_j \hat{\lambda}_{i|j}; \quad (5)$$

- fit Poisson regression models with  $k$  and  $k + 1$  components to bootstrap data;
  - calculate the corresponding 'bootstrap' log-likelihood ratio statistic. Denote this value by  $Q_b^*$ .
- d. Compute the 'bootstrap'  $p$ -value as

$$p_B = \frac{1}{B} \sum_{b=1}^B I\{Q_b^* \geq Q\} \quad (6)$$

where  $I\{\cdot\}$  denotes the indicator function of an event.

This algorithm is implemented by first testing 1 versus 2 components. If the obtained  $p_B$  for this test is lower than a significance level  $\alpha$ , one claims statistical significance and proceed to test 2 versus 3 components. If not, algorithm is stopped and it is claimed that there is not statistically significant evidence for a 2-component fit. The procedure continues in this manner until the null hypothesis cannot be rejected for the first time, i.e. when there is no sufficient evidence that adding one more component will significantly improve the likelihood.

This procedure for testing the number of components is based on a forward search technique aiming mainly at reducing the computational effort.

Notwithstanding, it is not clear yet how to form a bootstrap observation because for each observation there are  $J$  possible estimated mean values,  $\hat{\lambda}_{i|j}$ , one for each component of the mixture. This situation can be circumvented by

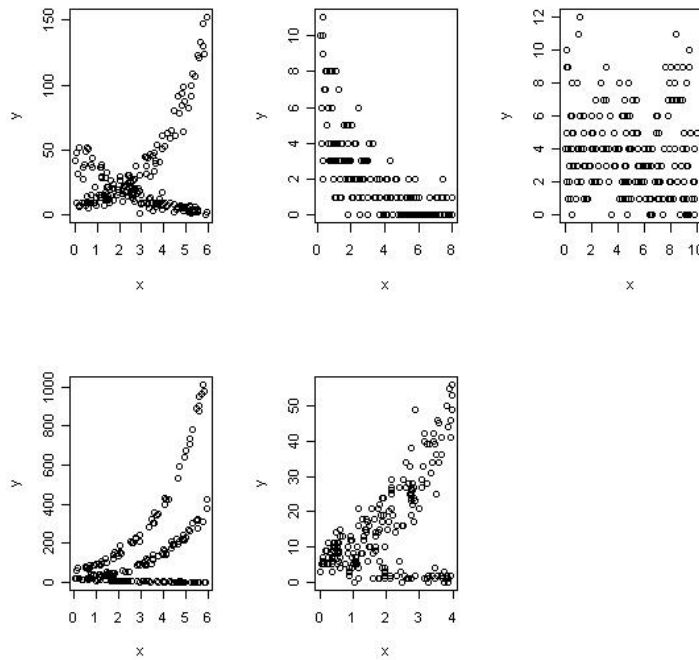
independently selecting a component  $j$  from  $1, \dots, J$  according to probabilities equal to  $\mathbf{w}_i$ , where  $\mathbf{w}_i$  is the vector of probabilities  $(w_{i1}, \dots, w_{iJ})$  where  $w_{ij}$  is the conditional probability that the  $i$ th observation was generated by the  $j$ th component of the mixture, given that observation.

### 3 Simulation study

This section concerns the simulation study which was performed. The main interest lies in the ability of the procedures to determine the correct number of components.

The scope was limited to the study of models with two and three components. We used the freeware R to develop the simulation program.

#### 3.1 Design of the study



**Fig. 1.** Scatter plots of simulated data from models with 2 and 3 components and sample size  $n = 500$ . The sets of true parameter values are A2, B1, C1, D1 and E2, respectively.

*Data set.* Each datum  $(x_i, y_i)$  was generated according to the following scheme. First, a random number  $c_i$  was generated from a Uniform distribution

over the interval  $(0,1)$  in order to select a component  $J$  from the mixture of regressions model. Next,  $x_i$  was randomly generated from a Uniform  $[x_L, x_U]$  distribution. Finally, the value  $y_i$  was simulated as a random number from a Poisson distribution with parameter  $\lambda_{i|j} = \exp(\beta_{j0} + \beta_{j1} x_i)$ .

*Number of Samples.* For each type of simulated data set, we generated 500 samples of sizes  $n = 50, 100, 500, 1000$ .

*Parameter Estimates.* Estimates of the unknown parameters were obtained via the EM algorithm. (Dempster et al. (1977)).

*Initial Conditions.* We ran the algorithm 10 times from random initial position. Parameter estimates corresponding to the best result with respect to the log-likelihood were subsequently used as initial values for the EM algorithm.

*Stopping Rules.* Iterations were stopped when the relative change in log-likelihood between two successive iterations was less than  $10^{-10}$ .

*Number of Bootstraps.* The bootstrap techniques were applied using 100 bootstrap samples ( $B = 100$ ) for assessing the null distribution of each test statistic.

*Significance Level.* For all the tests we used  $\alpha = 5\%$ .

*AIC and BIC criteria.* Models with the number of components ranging from 1 to 5 were fitted to the data. The one yielding the smallest AIC, BIC respectively, was considered the optimal number of components.

Samples of four different sizes were generated for each set of true parameter values (Table 1). Typical scatter plots of simulated data with a sample size of 500 are shown in Figure 1.

Cases	$\beta_{10}$	$\beta_{11}$	$\beta_{20}$	$\beta_{21}$	$\pi_1$	Cases	$\beta_{10}$	$\beta_{11}$	$\beta_{20}$	$\beta_{21}$	$\beta_{30}$	$\beta_{31}$	$\pi_1$	$\pi_2$
A1	4	-0.5	2	0.5	0.2	D1	3	-0.5	3	0.5	4	0.5	0.4	0.2
A2	4	-0.5	2	0.5	0.5	D2	3	-0.5	3	0.5	4	0.5	1/3	1/3
A3	4	-0.5	2	0.5	0.9	E1	2	-0.5	2	0.5	1.9	0.5	0.4	0.2
B1	2	-0.5	2.1	-0.5	0.5	E2	2	-0.5	2	0.5	1.9	0.5	1/3	1/3
B2	2	-0.5	2.1	-0.5	0.9									
C1	2	-0.2	1	0.1	0.5									

**Table 1.** True parameter values for both two and three component essays.

### 3.2 Correct number of components

Tables 2 and 3 present the relative frequencies of the number of components detected by the methods.

It is clear that all methods are very successful in determining the number of components in a PMR. Although among all of them BIC criterion is the one performing better, the nonparametric bootstrap procedure arises as a very good competitor. Nonparametric bootstrap procedure performs slightly better than parametric bootstrap method and AIC criterion. We also observed that as the sample size increases the performance of the methods increases as well. However, when some components are highly overlapped none of the methods was able to distinguish them. This explains the performance of the

Case	n	Parametric Bootstrap				Non Parametric Bootstrap			AIC				BIC		
		1	2	3	4	1	2	3	1	2	3	4	1	2	3
A1	50		0.96	0.04			0.97	0.03		0.95	0.05			0.97	0.03
	100		0.95	0.05			0.99	0.01		0.95	0.05			0.99	0.01
	500		0.96	0.04			1.00			0.96	0.04			1.00	
	1000		0.96	0.04			1.00			0.96	0.04			1.00	
A2	50		0.96	0.04			1.00			0.95	0.04	0.01		1.00	
	100		0.95	0.04	0.01		1.00			0.96	0.04			1.00	
	500		0.98	0.02			1.00			0.96	0.04			1.00	
	1000		0.98	0.02			1.00			0.97	0.03			1.00	
A3	50	0.04	0.96			0.04	0.96		0.04	0.96			0.04	0.96	
	100	0.04	0.96			0.02	0.98		0.02	0.97	0.01		0.02	0.98	
	500		1.00				1.00			1.00				1.00	
	1000		1.00				1.00			1.00				1.00	
B1	50	0.96	0.04			0.96	0.04		0.96	0.04			1.00		
	100	0.97	0.03			0.98	0.02		0.96	0.04			1.00		
	500	0.99	0.01			1.00			0.98	0.02			1.00		
	1000	1.00				1.00			0.99	0.01			1.00		
B2	50	0.97	0.03			0.98	0.02		0.96	0.04			0.99	0.01	
	100	0.96	0.04			0.98	0.02		0.96	0.04			0.99	0.01	
	500	0.98	0.02			0.99	0.01		0.98	0.02			1.00		
	1000	1.00				1.00			1.00				1.00		
C1	50	0.09	0.91			0.09	0.91		0.07	0.91	0.02		0.07	0.93	
	100	0.04	0.96			0.04	0.96			0.98	0.02		0.03	0.97	
	500		1.00				0.99	0.01		0.98	0.02			1.00	
	1000		1.00				1.00			1.00				1.00	

**Table 2.** Relative frequencies of the estimated number of components among 500 simulated samples from 2 component models.

Case	n	Parametric Bootstrap			Non Parametric Bootstrap			AIC			BIC		
		2	3	4	2	3	4	2	3	4	2	3	4
D1	50		0.97	0.03		0.97	0.03		0.97	0.03		1.00	
	100		0.98	0.02		0.99	0.01		0.99	0.01		1.00	
	500		0.99	0.01		1.00			0.99	0.01		1.00	
	1000		1.00			1.00			1.00			1.00	
D2	50		0.97	0.03		0.98	0.02		0.97	0.03		1.00	
	100		0.97	0.03		0.97	0.03		0.97	0.03		1.00	
	500		0.98	0.02		1.00			0.98	0.02		1.00	
	1000		0.99	0.01		1.00			1.00			1.00	
E1	50	0.96	0.04		0.99	0.01		0.96	0.04		1.00		
	100	0.97	0.03		1.00			0.97	0.03		1.00		
	500	0.98	0.02		1.00			0.98	0.02		1.00		
	1000	0.99	0.01		1.00			0.99	0.01		1.00		
E2	50	0.97	0.03		0.98	0.02		0.97	0.03		0.98	0.02	
	100	0.97	0.03		0.99	0.01		0.97	0.03		1.00		
	500	0.98	0.02		1.00			0.98	0.02		1.00		
	1000	0.99	0.01		1.00			0.99	0.01		1.00		

**Table 3.** Relative frequencies of the estimated number of components among 500 simulated samples from 3 component models.

methods for cases B1, B2, E1 and E2, and the fact that when a specific method misidentified the correct number of components it always yielded a model with less components than the true one.

## 4 Real Data Sets

In the present section the aforementioned procedures for determining the number of components are applied to two real data set examples.

First, we consider the real data *Michigan* as given by Hurn et al.(2003). They relate the monthly unemployment rate with the monthly number of accidents (in thousands) in the state of Michigan, from 1978 to 1987. The

response variable is the number of accidents and the logarithm of the corresponding unemployment rate is used as independent variable. From the results reported in Table 4 there is no evidence for a two-component fit in *Michigan* data. All the procedures support this conclusion, which ultimately means that a mixture model is unnecessary.

The *Fabric Faults* data set consists of 32 observations of number of faults in rolls of textile fabric with varying length. The dataset is analyzed using finite mixtures of Poisson regression models in Aitkin (1996). The response variable is the number of faults and the covariate is the logarithm of length of role. As can be seen in Table 4 for assessing the number of components both parametric and nonparametric bootstrap procedures provide statistically significant evidence that two components are more appropriate than one, and the p-values for testing  $k = 2$  versus  $k = 3$  indicate there is not statistically significant evidence for selecting more than two components. Also, both AIC and BIC criteria agree with the decision of selecting two components.

Data	Number of components					Bootstrap	P-value		
	1	2	3	4	5		1 vs 2	2 vs 3	
<i>Michigan</i>	AIC	667.94	673.65	679.65	685.65	691.65	Parametric	0.84	-
	BIC	673.31	687.06	701.10	715.15	729.20	Nonparametric	0.86	-
<i>Fabric</i>	AIC	191.83	179.77	183.42	189.42	189.42	Parametric	0.00	0.20
	BIC	194.77	187.10	195.14	205.54	205.54	Nonparametric	0.00	0.69

**Table 4.** Values for AIC and BIC criteria and bootstrap p-values for both parametric and nonparametric techniques for *Michigan* and *Fabric Faults* data.

## 5 Conclusions

In this paper we have investigated bootstrap techniques for testing hypotheses concerning the number of components in a PMR: a parametric bootstrap where the bootstrap samples are generated by random selection with replacement and a nonparametric bootstrap where the bootstrap samples are obtained by resampling from the residuals of the null model. We have also compared their performance to determine the number of components with AIC and BIC criteria.

Our simulation study shows that all methods work successfully in determining the number of components in these models and the proposed nonparametric bootstrap is an alternative to select the correct number of components.

### Acknowledgements

S. Faria wants to acknowledge the financial support provided by the Research Centre of Mathematics of the University of Minho through the FCT Pluriannual Funding Program.

## References

- AITKIN, M. (1996): A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6: 251-262.
- DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc., Ser. B* 39: 1-38.
- FRUHWIRTH-SCHNATTER (2006): *Finite Mixture and Markov Switching Models*, Springer, Heidelberg.
- HURN, M., JUSTEL, A. and ROBERT, C.P.(2003): Estimating Mixtures of Regressions. *Journal of Computational and Graphical Statistics*, 12: 55-79.
- KARLIS, D. and XEKALAKI, E. (1999): On testing for the number of components in finite Poisson mixtures. *Ann. Inst. of Stat. Math.*, 51: 149-161.
- MCLACHLAN, G.J. and PEEL, D. (2000): *Finite Mixture Models*, Wiley, New York.
- SCHLATTMANN, P. (2005): On bootstrapping the number of components in finite mixtures of Poisson distributions. *Statistics and Computing* 15(3): 179-188 .
- TURNER, T. (2000): Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Applied Statistics* 49 (3): 371-384 .
- WANG, P., PUTERMAN, M.L., COCKBURN, I.M. and LE, N. (1996): Mixed Poisson Regression Models with Covariate Dependent Rates. *Biometrics* 52 (2): 381-400.

# Support Vector Machines for Large Scale Text Mining in R

Ingo Feinerer<sup>1</sup> and Alexandros Karatzoglou<sup>2</sup>

<sup>1</sup> Database and Artificial Intelligence Group  
Institute of Information Systems  
Vienna University of Technology  
Austria *Ingo.Feinerer@tuwien.ac.at*

<sup>2</sup> LITIS, INSA de Rouen  
Avenue de Universite  
76801 Saint-Etienne du Rouvray  
France *alexis@ci.tuwien.ac.at*

**Abstract.** SVM are an established tool in machine learning and data analysis. Though many implementations of SVM exist often specific applications require tailor made algorithms. In text mining in particular the data often comes in large sparse data matrices. Typical SVM algorithms like SMO do not take advantage of the sparsity, and do not scale well to data sets with millions of entries. In this paper we present an implementation of linear SVM's for R that address both of these issues.

**Keywords:** SVM, text mining, large scale

## 1 Introduction

Many applications in machine learning and data mining require the classification of massive amounts of data like e.g. spam filtering or text topic identification. As R is becoming the tool of choice for data intensive operations in both research and industry it is vital that basic learning tasks can be performed on large scale data sets. In particular in the area of text mining which has seen some significant research activity due to the importance of the web, one has to usually deal with data that comes in the form of large sparse matrices as produced by the term-document extraction process.

Term-document matrices which essentially represent each document by the frequency of the occurring words can be easily produced in R due to the availability of modern text processing and mining functionality provided by the `tm` package. There are though not many methods that can utilize large scale sparse matrices and in particular there are no support vector machines (SVM) implementations ([?]) that can easily deal with large scale data sets. SVMs are central in many text mining tasks like topic assignment and filtering and thus an SVM implementation that is tailored for large scale sparse data matrices is particularly important.

Current support vector machine implementations in R support sparse matrix formats as inputs but the algorithms that are used in the optimization task do not take advantage of the sparsity in that data. Moreover all SVM implementations in R solve the dual optimization problem utilizing algorithms like SMO ([?]) which are known to scale super-linear with the size of the data.

In this paper we introduce a new implementation of SVMs in R which is tailored for large scale sparse data. The implementation is based on the algorithms introduced by [?] and by [?], and in addition adds multi-class functionality and handling of all sparse matrix formats in R. Linear learning techniques are often the tool of choice when massive amounts of data are available.

## 2 Large scale linear support vector machines

Many SVM variants have been developed over the years. One that is particularly well suited for large scale sparse data is the  $l_2$ -SVM ([?]). Given  $m$  binary labeled examples  $\{x_i, y_i\}$  with  $i \in 1, \dots, m$  and  $y_i \in \{+1, -1\}$  the  $l_2$ -SVM optimization problem can be written in its Lagrangian form as

$$w^* = \underset{w \in \mathbb{R}}{\operatorname{argmin}} f(w) = \frac{1}{2} \sum_{i=1}^m c_i l_2(y_i w^T x_i) + \frac{\lambda}{2} \|w\|^2 \quad (1)$$

where  $l_2$  is the  $l_2$ -SVM loss given by  $l_2(z) = \max(0, 1 - z)^2$ ,  $\lambda$  is the regularization parameter, and  $c_i$  is the relative importance or weight for each data point. The decision function is then given by  $f = \operatorname{sign}(w^* x)$ . The difference between a standard SVM and this formulation is in the loss function. A standard SVM utilizes the well known hinge loss  $l(z) = \max(0, 1 - z)$  while the  $l_2$ -SVM uses a squared hinge loss. This seemingly trivial difference has a big impact on the optimization procedure one can use in order to find the optimal  $w^*$  for Equation 1. The main difference between the hinge loss and its square form is that the squared form is differentiable in any point  $z$ . Note also that the  $l_2$ -SVM does not include an explicit bias term as in the standard SVM formulation but the bias term is included in  $w^*$  and computed with the addition of an intercept term in  $x$ . This in effect forces a regularization on the bias. Moreover the  $l_2$ -SVM preserves the large margin characteristics that come with the theoretical guarantees for good generalization properties of the standard SVM.

Note here that Equation 1 is a strictly convex differentiable and a piecewise quadratic function of  $w$ . A very effective optimization procedure in order to find the single optimal  $w^*$  was used by [?] and is based on the Newton method. In the Newton method a second order approximation of the objective function is used to iteratively update the  $w$  parameter vector, that is

$$w^{t+1} = w^t - \eta^t [\nabla^2 f(w^t)]^{-1} \nabla f(w^t) \quad (2)$$



where  $\eta^t$  is the step size parameter,  $\nabla f(w^t)$  is the gradient vector and  $\nabla^2 f(w^t)$  is the Hessian matrix. The Hessian does not exist for every  $w$  since  $f(w)$  is not twice differentiable for all  $w$  and thus a generalized approximation of the Hessian is used. The step direction is then given by first solving a least squares problem over a subset  $s(w^t)$ :

$$\left[ \lambda I + X_{s(w^t)}^T C_{s(w^t)} X_{s(w^t)} \right] \bar{w}^t = X_{s(w^t)}^T C_{s(w^t)} Y_{s(w^t)} \quad (3)$$

where  $I$  is the identity matrix. Given  $\bar{w}^t$  the direction of the update is then given by  $w^{t+1} = w^t + \eta^k (\bar{w}^t - w^t)$  and the optimal step size  $\eta^t$  is found by a simple line-search on the optimization problem  $\eta^t = \underset{\eta}{\operatorname{argmin}} f(w^t + \eta(\bar{w}^t - w^t))$ .

The least squared optimization problem 3 is solved using a conjugate gradient method that performs particularly well when  $X$  is sparse ([?]).

The whole optimization procedure can be performed very fast and the algorithm for solving Equation 1 scales linear to the number of non-zero entries in the data matrix. Note also here that in contrast to many conventional SVM algorithms all optimization computations are performed in primal space.

### 3 svmlin R package

The  $l_2$ -SVM algorithm is implemented in the **svmlin** R package. It is licensed under the GPL and is available via the CRAN archive. The implementation extends the original C++ version of **svmlin** provided by [?] and introduces multi-class classification, cross validation and handling of a range of sparse matrix formats. The **svmlin** R package supports all available standard sparse matrix formats provided by the **SparseM**, **Matrix**, and **slam** packages. The function implementing the algorithm boost a formula interface along with a default interface and the data matrix has to be in sparse matrix format or in term-document matrix format as supported by the **tm** package. A call to the **svmlin** function is done simply by

```
svmodel <- svmlin(matrix, labels, lambda = 0.1, cross = 3)
```

where the resulting object contains the weights parameter vector  $w$ , the offset term along with the training and the 3-fold cross-validation error. In this call we set the regularization parameter value of  $\lambda = 0.1$ . The resulting object can be used with the **predict** function along with test data. The **svmlin** R package provides multi-class classification functionality by implementing two popular voting schemes: the one-against-one and the one-against-all.

The one-against-one or pairwise classification method ([?,?]) constructs  $\binom{k}{2}$  classifiers ( $k$  as the number of classes) where each one is trained on data from two classes. Prediction is done by voting where each classifier gives a prediction and the class which is predicted more often wins ("Max Wins"). This method has been shown to produce robust results when used with SVMs ([?]).

The one-against-all method is a somewhat simpler approach where  $k$  classifiers are constructed that always separate one class from the rest. The  $i$ -th SVM classifier is trained with all the examples of the  $i$ -th class with positive labels and all other with negative ones. At the classification phase a sample  $x$  is assigned to class  $i$  when the decision value of the classifier for class  $i$ ,  $f_i$  produces the largest value.

## 4 tm R package

The `tm` ([?,?]) package provides a framework for text mining applications in R. It offers functionality for managing corpora and text documents, abstracts the process of document manipulation and eases the usage of heterogeneous text formats in R. An advanced meta data management is implemented for collections of text documents to alleviate the usage of large and meta data enriched document sets. `tm` provides easy access to preprocessing mechanisms such as stemming, stopword exclusion, or removal of punctuation marks. Out of the box `tm` also includes functionality for processing the Reuters-21578 data sets in native XML, and for processing the e-mail messages (including meta data in headers) for the SpamAssassin and 20 newsgroups data sets. In addition `tm` can construct and export term-document matrices in a sparse representation from corpora.

## 5 Experiments

### 5.1 Data

Our primary research data set is the Reuters-21578 data set ([?]) containing stories collected by the Reuters news agency in 1987. The data set is publicly available and has been widely used in text mining research within the last decade. It contains 21578 short to medium length documents in XML format covering a wide range of topics, like mergers and acquisitions, finance, or politics.

Our second data set is the SpamAssassin public mail corpus (<http://spamassassin.apache.org/publiccorpus/>). It is freely available and offers authentic e-mail communication with classifications into normal (ham) and unsolicited (spam) mail of various difficulty levels (easy ham, hard ham, and spam). In total we have 4150 ham documents and 1896 spam documents.

Our final data set is the 20 newsgroups text collection (<http://kdd.ics.uci.edu/databases/20newsgroups/>). It consists of 19997 e-mail messages taken from 20 different newsgroups (however cross posting was allowed) and is publicly available due to a donation by Tom Mitchell. The newsgroups cover a wide field of unique topics dealing e.g. with atheism, computer graphics, motorcycles, or politics in the middle east.

## 5.2 Protocol

The main aim of the experiments is to illustrate the significant speedups obtained by the use of the `svmlin` package in particular compared to a standard SVM implementation like the function `svm` in package `e1071`. To this end we use the data sets to train with both SVM implementations. For the `e1071` implementation we make sure to use the linear kernel and to set the cost parameter to  $\frac{1}{\lambda}$  to get equivalent models.

In order to compare the scaling behavior we sample from our data sets first  $\frac{1}{10}$  of the data for training and increase the training data amount by  $\frac{1}{10}$  before training again up to the whole data set. We repeat this procedure for both implementations. We also compare the classification performance of the implementations on the data sets using 10-fold cross-validation. We tune both models for the regularization and cost parameters.

The creation of the sparse term-document matrices took about 42 seconds for the Reuters-21578 XML data set, about 31 seconds for the SpamAssassin data set, and about 75 seconds for the 20 newsgroups data set. The Reuters-21578 term-document matrix has 65973 terms, 21578 documents, and a size of about 24 MB in memory, the SpamAssassin matrix has 151029 terms, 6046 documents, and is about 24 MB big, whereas the 20 newsgroups matrix has 175685 terms, 19997 documents, and a memory footprint of about 46 MB.

## 5.3 Results

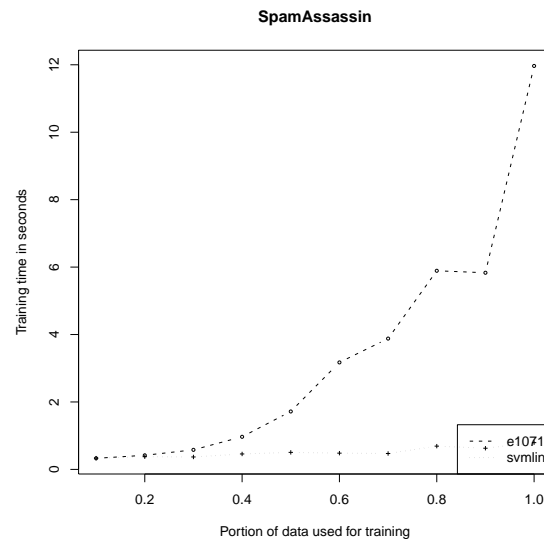
The results obtained are presented in the following plots. In Figures 1 and 2 we observe that although for very small portions of the 20 newsgroup data the `e1071` implementation is faster as the data size increases `svmlin` performs better. This highlights the better scaling behavior of the `svmlin` implementation. Note also that most of the data handling and splitting is done in R in the `svmlin` implementation and thus represents an overhead compared to the `e1071` method where almost all data splittings necessary for the one-against-one voting scheme are done in C++.

In Figure 3 the difference in training time is much clearer since this is also the largest data set in our experiment setup. `svmlin` is significantly faster than the `e1071` implementation. Similarly in Figure 1 which is a relatively small binary classification data set the `svmlin` implementation has a faster training time than `e1071`.

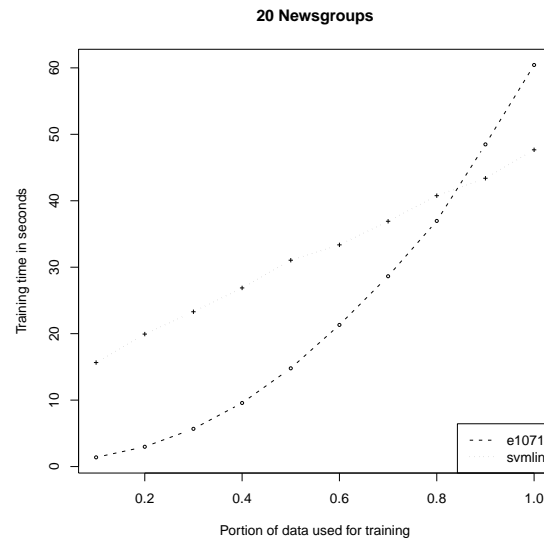
We also compared the results of the two implementations in terms of classification performance with 10-fold cross-validation and found no significant performance advantages for either implementation.

## 6 Conclusion

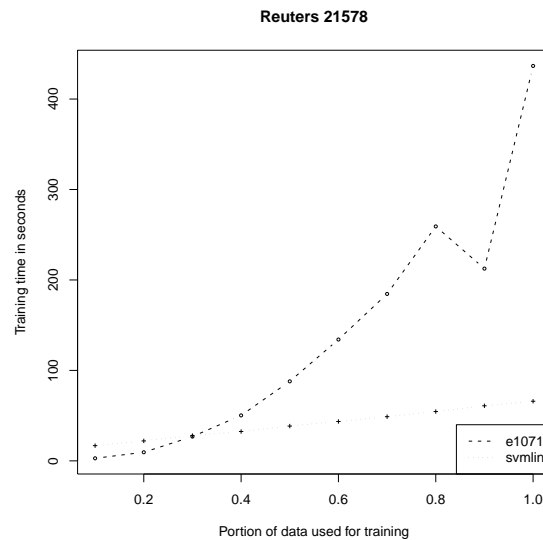
We presented the `svmlin` SVM implementation for large scale text mining tasks. The introduced implementation takes advantage of the sparsity in



**Fig. 1.** CPU training time for **svmLin** and **svm** (from **e1071**) for different portions of the SpamAssassin data set starting with 1/10 of the data up to the whole data set. The **svmLin** implementation is faster for almost all configurations.



**Fig. 2.** CPU training time for **svmLin** and **svm** (from **e1071**) for different portions of the 20 newsgroups data starting with 1/10 of the data up to the whole data set. The **e1071** implementation scales super-linearly and is outperformed for larger portions of the data.



**Fig. 3.** CPU training time for `svmlin` and `svm` (from `e1071`) for different portions of the Reuters 21578 data starting with 1/10 of the data up to the whole data set. The `e1071` implementation is slower for almost all configurations.

the data to accelerate the optimization process. The computations are done in primal space thus no kernel is used. This is no problem in text mining tasks where linear methods have been shown to produce excellent results. We demonstrated the usefulness of the new implementation and the advantages it provides compared to previous implementations: in particular the linear scaling with the data size and the faster training time on larger data sets.

## References

- FEINERER, I. (2010): *tm: Text Mining Package*, 2010. URL <http://tm.r-forge.r-project.org/>. R package version 0.5-3.
- FEINERER, I., HORNIK, K. and MEYER, D. (2008): Text mining infrastructure in R. *Journal of Statistical Software* 25 (5), 1–54. ISSN 1548-7660. URL <http://www.jstatsoft.org/v25/i05>.
- HSU, C.-W. and LIN, C.-J. (2002): A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks* 13, 1045–1052. URL <http://www.csie.ntu.edu.tw/~cjlin/papers/multisvm.ps.gz>.
- KARATZOGLOU, A., MEYER, D. and HORNIK K. (2006): Support vector machines in R. *Journal of Statistical Software* 15 (9), 1–28. URL <http://www.jstatsoft.org/v15/i09/>.

- KEERTHI, S.S. and DECOSTE, D. (2005): A modified finite Newton method for fast solution of large scale linear SVMs. *Journal of Machine Learning Research* 6: 341–361.
- KNERR, S., PERSONNAZ, L. and DREYFUS G. (1990): Single-layer learning revisited: A stepwise procedure for building and training a neural network. In: J. Fogelman Soulié and J. Hérault (Eds.): *Neurocomputing: Algorithms, Architectures and Applications*, 41–50.
- KRESSEL, U. (1999): Pairwise classification and support vector machines. In: B. Schölkopf, C. J. C. Burges and A. J. Smola (Eds.): *Advances in Kernel Methods — Support Vector Learning*, 255–268.
- LEWIS, D. (1997): Reuters-21578 text categorization test collection. URL <http://www.daviddlewis.com/resources/testcollections/reuters21578/>.
- PLATT, J. (1999): Fast training of support vector machines using sequential minimal optimization. In: B. Schölkopf, C. J. C. Burges and A. J. Smola (Eds.): *Advances in Kernel Methods — Support Vector Learning*, MIT Press, Cambridge, MA.
- SINDHWANI, V. and KEERTHI, S.S. (2007): Large scale semi-supervised linear SVMs. In *SIGIR 2007*. ACM Press, 65–72.

# Computation of the Projection of the Inhabitants of the Czech Republic by sex, age and the highest education level<sup>\*</sup>

Tomáš Fiala and Jitka Langhamrová

Department of Demography, Faculty of Informatics and Statistics, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic,  
*fiala@vse.cz, langhamj@vse.cz*

**Abstract.** The computation is based on the classical component method of population projections computations. Four education levels: primary education, secondary lower education, secondary higher education and tertiary education are distinguished. The surviving probabilities are supposed to depend not only on the sex and age but also on the education level of the person. The projection has been computed for the population of the Czech Republic since 2001 until 2051.

**Keywords:** population projection, component method, education level

## 1 Introduction

The output of "classical" population projections is usually the sex and age structure of the population in each year of the projected period. This provides no information about the "quality" of the population, e.g. about expected development the professional qualification of the people. As a very simple (and of course very rough) measure of the professional qualification of a person can be regarded its education level.

This paper describes very briefly the methodology of a population projection not only by sex and age, but also by education level and provides the results of computation of such projection for the case of the Czech Republic.

## 2 Methodology

Not only sex and age but also education level of each person is taken into account. Only four following groups of education level are distinguished:

- A - primary education (including no education or incomplete education); each newborn child is supposed to belong to this group,

---

<sup>\*</sup> This article came into being within the framework of the long-term research project 2D06026, "Reproduction of Human Capital", financed by the Ministry of Education, Youth and Sport within the framework of National Research Program II.

B - secondary lower education (without the school leaving exam),  
 C - secondary higher education (finished with the school leaving exam),  
 D - tertiary education.

The computation of the population projection by sex, age and education level of each person is based on the classical component projection method see, e.g., Bogue et al. (1993), Koschin (2005) with simplified model of migration (only immigration at the level of net migration is assumed, emigration is supposed to be zero). The computation is carried out for each sex separately.

Let us denote (for each sex separately):

$S_{e,t,x}$  – the number of persons of the education level  $e$  at the age  $x$  at the beginning of the year  $t$ .

$I_{e,t,x}$  – the number of immigrants of the education level  $e$  at the age  $x$  during the year  $t$ . In the case of prevailing emigration these values are negative.

${}_{e_1}G_{e_2,t,x}$  – the number of persons increasing their education level from  $e_1$  to  $e_2$  at the age  $x$  during the year  $t$ . We accept only the following changes of education level:  $A \rightarrow B$ ,  $A \rightarrow C$ ,  $B \rightarrow C$ ,  $C \rightarrow D$ . (Relatively many young people having finished the primary school continue to study at a secondary school finished by the school leaving exam. Their education level will then increase from  $A$  directly to  $C$ .)

The projection formulas have been derived from the classical projection formula see, e.g. Koschin (2005). For each group of education level the number of persons belonging to this group at the beginning of the year  $t + 1$  equals to: the number of persons of this group at the beginning of the year  $t$  surviving the year  $t$  (1<sup>st</sup> term of the equation) and the number of immigrants in the year  $t$  having given education level and surviving until the end of the year (2<sup>nd</sup> term of the equation).

Moreover we must: add the number of persons reaching the given education level during the year  $t$  and surviving until the end of the year and subtract the number of persons having at the beginning of the year  $t$  given education level, reaching higher education level during the year and surviving until the end of the year (further terms of the equation). The projection formulas are the following (for each sex separately)

$$\begin{aligned}
 S_{A,t+1,x+1} = & S_{A,t,x} \cdot P_{A,t,x} + \frac{I_{A,t,x} \cdot P_{A,t,x}^{2/3} + I_{A,t,x+1} \cdot P_{A,t,x}^{1/3}}{2} - \\
 & - \frac{{}_AG_{B,t,x} + {}_AG_{B,t,x+1}}{2} \cdot P_{B,t,x}^{1/2} - \frac{{}_AG_{C,t,x} + {}_AG_{C,t,x+1}}{2} \cdot P_{C,t,x}^{1/2}, \\
 \\ 
 S_{B,t+1,x+1} = & S_{B,t,x} \cdot P_{B,t,x} + \frac{I_{B,t,x} \cdot P_{B,t,x}^{2/3} + I_{B,t,x+1} \cdot P_{B,t,x}^{1/3}}{2} + \\
 & + \frac{{}_AG_{B,t,x} + {}_AG_{B,t,x+1}}{2} \cdot P_{B,t,x}^{1/2} - \frac{{}_BG_{C,t,x} + {}_BG_{C,t,x+1}}{2} \cdot P_{C,t,x}^{1/2},
 \end{aligned}$$



$$\begin{aligned}
S_{C,t+1,x+1} = & S_{C,t,x} \cdot P_{C,t,x} + \frac{I_{C,t,x} \cdot P_{C,t,x}^{2/3} + I_{C,t,x+1} \cdot P_{C,t,x}^{1/3}}{2} + \\
& + \frac{A_{G_{C,t,x}} + A_{G_{C,t,x+1}}}{2} \cdot P_{C,t,x}^{1/2} + \frac{B_{G_{C,t,x}} + B_{G_{C,t,x+1}}}{2} \cdot P_{C,t,x}^{1/2} - \\
& - \frac{C_{G_{D,t,x}} + C_{G_{D,t,x+1}}}{2} \cdot P_{D,t,x}^{1/2}, \\
S_{D,t+1,x+1} = & S_{D,t,x} \cdot P_{D,t,x} + \frac{I_{D,t,x} \cdot P_{D,t,x}^{2/3} + I_{D,t,x+1} \cdot P_{D,t,x}^{1/3}}{2} + \\
& + \frac{C_{G_{D,t,x}} + C_{G_{D,t,x+1}}}{2} \cdot P_{D,t,x}^{1/2},
\end{aligned}$$

where  $P_{e,t,x}$  is the so called projection coefficient - the probability that a person of the education level  $e$  at the age  $x$  will survive the year  $t$  (mortality is supposed to be dependent not only on sex and age but also on the education level).

The number of youngest children (at the age 0) depends of course mainly on the number of live births computed by the formula

$$N_t^{(bs)} = \sum_{x=15}^{49} \frac{S_{t,x}^{(f)} + S_{t+1,x}^{(f)}}{2} \cdot f_{t,x}, N_t^{(m)} = 0.515 \cdot N_t^{(bs)}, N_t^{(f)} = 0.485 \cdot N_t^{(bs)},$$

where  $S_{t,x}^{(f)}$  is the number of females (regardless their education level) at the age  $x$  at the beginning of the year  $t$ ,

$f_{t,x}$  is the so called age specific fertility rate - the probability that a female (regardless of the education level) at the age  $x$  will bear a live child in the year  $t$ . Fertility is supposed to be zero for females younger then 15 and older then 50 years. The proportion of newborns according to sex (515 males to 485 females) is the very often used expert estimate.

$N_t^{(bs)}$ ,  $N_t^{(m)}$ ,  $N_t^{(f)}$  mean the number of live births in the year  $t$  of both sexes, males and females, respectively.

Then we have (for each sex separately) see, e.g. Koschin (2005)

$$S_{A,t+1,0} = N_t \cdot P_{A,t,*} + \frac{I_{A,t,0} \cdot P_{A,t,*}^{1/3}}{2}$$

and, of course,  $S_{B,t+1,0} = S_{C,t+1,0} = S_{D,t+1,0} = 0$ ,

$P_{A,t,*}$  is the so called projection coefficient for newborns - the probability that a child born during the year  $t$  will survive until the end of the year  $t$ .

### 3 Scenarios of the population projection

Computation of each population projection is based on the initial population structure (in our case also by education level) and the data describing the expected development of fertility, mortality and migration. In the case of our

projection taking into account also the education level of people also data of expected numbers of graduates of particular types of schools were necessary.

Latest available data of the population structure of the Czech Republic by sex, age and education level come from the population census in 2001. See ČSÚ (2003). Initial demographic structure for the projection has been so that of 1<sup>st</sup> January 2001 and the projection has then been computed until 1<sup>st</sup> January 2051.

Until the end of 2008 the computations were based on real data of mortality, fertility and migration. Since 2009 two variants of the future development of mortality, fertility and migration have been taken into account.

First variant is a slightly modified medium variant of the population prognosis computed by the Czech Statistical Office in 2009 (*variant CZSO*). See ČSÚ (2009). The second variant supposes that the fertility of the Czech females will (with several years "delay") follow the fertility of the Netherlands' females (*variant NL*). Netherlands' females fertility is very often used as a pattern of future fertility of Czech females because in this country the transition of the fertility to higher age of females has been finished and the fertility seems to be relatively stable.

The trends of population development are assumed to be the same in both variants but the rate of growths differs. In the variant NL higher increase in fertility, more rapid growth of the life expectancy and higher annual net migration have been assumed than in the variant CZSO. See the tables 1-3.

In both variants further increase of the total fertility rate (TFR) has been supposed, but not as rapid as in previous years. Linear increase in each decade has been supposed with subsequently diminishing increment.

In the variant CZSO we have assumed that

$$TFR_t = TFR_{2008} \text{ for } t = 2009 \text{ and } 2010,$$

$$TFR_t = TFR_{2010} + 0.01 \cdot (t - 2010) \text{ for } t = 2011, 2012, \dots, 2020,$$

$$TFR_t = TFR_{2020} + 0.006 \cdot (t - 2020) \text{ for } t = 2021, 2022, \dots, 2030,$$

$$TFR_t = TFR_{2030} + 0.003 \cdot (t - 2030) \text{ for } t = 2031, 2032, \dots, 2050.$$

In the variant NL we have had following assumptions

$$TFR_{2009} = TFR_{2008}, TFR_{2010} = TFR_{2009} + 0,01$$

$$TFR_t = TFR_{2010} + 0.019 \cdot (t - 2010) \text{ for } t = 2011, 2012, \dots, 2020,$$

$$TFR_t = TFR_{2020} + 0.01 \cdot (t - 2020) \text{ for } t = 2021, 2022, \dots, 2030,$$

$$TFR_t = TFR_{2030} + 0.005 \cdot (t - 2030) \text{ for } t = 2031, 2032, \dots, 2050,$$

(see Table 1).

**Table 1.** Assumed development of the total fertility rate

Variant	2008 <sup>1</sup>	2009	2010	2020	2030	2040	2050
CZSO	1.50	1.50	1.50	1.60	1.66	1.69	1.72
NL	1.50	1.50	1.51	1.70	1.80	1.85	1.90

The fertility structure in the CZSO variant has been expected to remain unchanged, in the NL variant the fertility structure has been supposed to converge to the structure of Netherlands until 2020 and then to be stable. The estimates of the age-specific fertility rates  $f_{t,x}$  have been based on these expected fertility scenarios.

Both variants of the projection have assumed continual increasing of the life expectancy ( $e_0$ ) as well. In the variant CZSO we have assumed linear increase in each decade with subsequently diminishing increment.

For males

$$e_{0,t} = e_{0,2008} + 0.27 \cdot (t - 2008) \text{ for } t = 2009 \text{ and } 2010,$$

$$e_{0,t} = e_{0,2008} + 0.25 \cdot (t - 2010) \text{ for } t = 2011, 2012, \dots, 2030,$$

$$e_{0,t} = e_{0,2008} + 0.2 \cdot (t - 2030) \text{ for } t = 2031, 2032, \dots, 2050.$$

For females

$$e_{0,t} = e_{0,2008} + 0.23 \cdot (t - 2008) \text{ for } t = 2009 \text{ and } 2010,$$

$$e_{0,t} = e_{0,2008} + 0.225 \cdot (t - 2010) \text{ for } t = 2011, 2012, \dots, 2030,$$

$$e_{0,t} = e_{0,2008} + 0.165 \cdot (t - 2030) \text{ for } t = 2031, 2032, \dots, 2050.$$

In the case of the variant NL we have supposed linear increase of the life expectancy based on the linear regression function.

For males

$$e_{0,t} = -589.906 + 0.330629t \text{ for all } t,$$

for females

$$e_{0,t} = -481.065 + 0.279503t \text{ for all } t,$$

the estimates of the parameters are based on the data of the period 2005-2008 (see Table 2).

The structure of mortality has been supposed unchanged. This common mortality scenario has been then differentiated according to education level. See, e.g., Mazouch and Fischer (2007). The estimates of the projection coefficients  $P_{e,t,x}$  (probabilities of surviving the given year) have been based on expected scenario of mortality, more precisely on expected values of survivors from life tables of the corresponding year.

**Table 2.** Assumed development of the life expectancy ( $e_0$ )

Variant	Sex	2008 <sup>1</sup>	2009	2010	2020	2030	2040	2050
CZSO	males	73.96	74.23	74.50	77.00	79.50	81.50	83.50
	females	80.14	80.37	80.60	82.85	85.10	86.75	88.40
NL	males	73.96	74.33	74.66	77.96	81.27	84.58	87.88
	females	80.14	80.46	80.74	83.53	86.33	89.12	91.92

We expect (in both variants) that the Czech Republic will remain to be the country of prevailing immigration but the annual net migration is expected to

<sup>1</sup> real value (Czech Statistical Office)

be much lower than in 2008. (See Table 3.) The preliminary net migration in 2009 has been about 28 000 persons. The sex and age structure of immigrants has been expected to converge to the structure of immigration to the EU. The expected numbers of immigrants by sex, age and education level  $I_{e,t,x}$  have been estimated by distributing the expected annual total net migration increment according to the expected sex and age structure of immigration. The structure of immigrants by education level has been assumed to be the same as of the Czech population.

**Table 3.** Assumed development of the annual net migration

Variant	2008 <sup>1</sup>	2009-2050				
CZSO	71 790	25 000				
Variant	2008	2009	2010	2015	2020	2022-2050
NL	71 790	30 000	31 000	35 000	38 000	40 000

And finally the estimate of numbers  ${}_eG_{e_2,t,x}$  of persons increasing their education level has been based on the data of Institute for Information in Education. See, e.g. Hulík (2009).

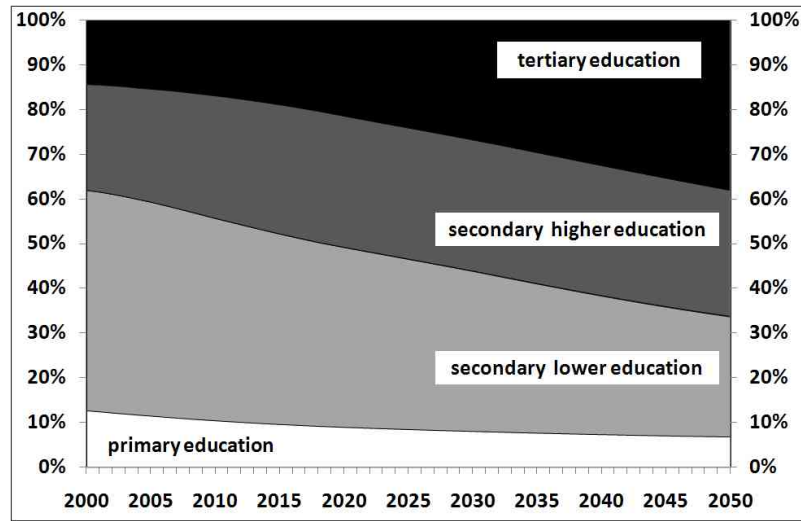
This projection is described in more detail e.g. in Langhamrová at al. (2009) and in Fiala and Langhamrová (2009), the assumptions concerning immigration are based on Kačerová (2008). More common information concerning the methodology of population projections can be found e.g. in Bogue at al. (1993).

## 4 Main results of the projection

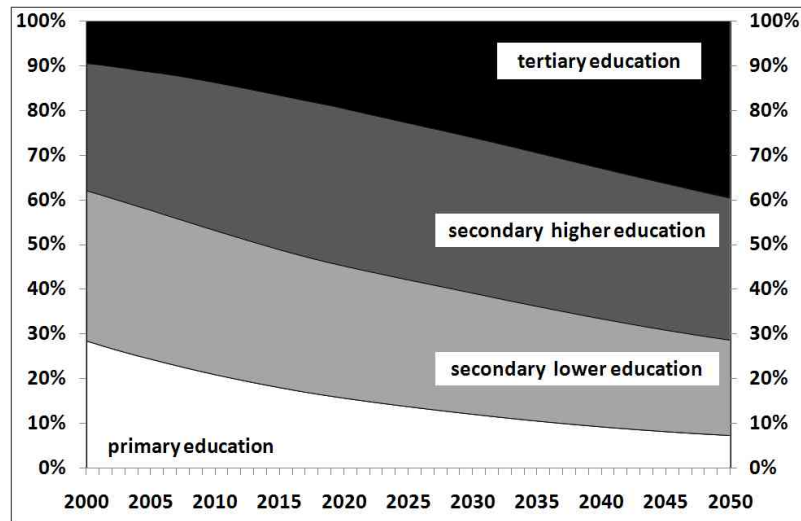
According to the variant CZSO the number of inhabitants of the Czech Republic in 2050 will be less than 11 millions, the variant NL gives more than 12 millions of inhabitants in 2050. Despite this relatively high difference in the population size the differences in the population structure by education level are negligible. Therefore we present the results for the latter variant (NL) only. The figures 1 and 2 show the development of the education level of the population older than 25 years of age.

At the end of the year 2000 the proportion of males with tertiary education has been only about 14 %, at the same time only about 10 % of females have had tertiary education level. Until 2050 the proportion of tertiary educated people is expected to grow to almost 38 % for males and even to 40 % for females. The proportion of males with secondary higher education will in the first half of this century slightly grow from 23 % to 28 %, for females the proportion will grow from 28 % to 32 %.

At the end of 2000 the most numerous education group has been the secondary lower education almost 50 % of males and 34 % of females. Until



**Fig. 1.** Development of the structure of Czech males older 25 years of age by the education level



**Fig. 2.** Development of the structure of Czech females older 25 years of age by the education level

2050 these proportions are supposed to radically drop to about 27 % for males and 21 % for females. The proportion of males having only primary education has been about 13 % at the end of 2000. It is assumed to drop to about 7 % until 2050. For the proportion of females with primary education we expect similar development: decrease from 28 % to 7 %.

The results confirm that the education level of Czech population will increase, at the same time the gap between males and females in the education level will diminish.

## References

- ARLT, J. and ARLTOVÁ, M. (2009): *Ekonomické časové řady*. Professional Publishing, Prague.
- BOGUE, D. J., ARRIAGA, E. E. and ANDERTON, D., L. (eds.). (1993): *Readings in Population Research Methodology* Vol. 5. Population Models, Projections and Estimates. United Nations Population Fund, Social Development Center, Chicago, Illinois.
- ČSÚ (Czech Statistical Office) (2009): *Projekce obyvatelstva České republiky do roku 2065*. <http://www.czso.cz/csu/2009edicniplan.nsf/p/4020-09>.
- ČSÚ (Czech Statistical Office) (2003): *Úroveň vzdělání obyvatelstva podle výsledku sčítání lidu*. <http://www.czso.cz/csu/2003edicniplan.nsf/p/4113-03>.
- FIALA, T. and LANGHAMROVÁ, J. (2009): Human resources in the Czech Republic 50 years ago and 50 years after. In: *IDIMT-2009 System and Humans A Complex Relationship*. Trauner Verlag universitat, Linz.
- HULÍK, V. and TESÁRKOVÁ, K. (2009): Vývoj přístupu terciárního vzdělávání v České republice v závislosti na demografickém vývoji. In: *Reprodukce lidského kapitálu Vzájemné vazby a souvislosti* [CD-ROM]. Oeconomica, Praha, 1-21.
- KAČEROVÁ, E. (2008): International migration and mobility of the EU citizens in the Visegrad group countries: Comparison and bilateral flows. In: *European Population Conference*. Barcelona. EPC, 142.
- KOSCHIN, F. (2005): *Kapitoly z ekonomické demografie*. Oeconomica, Praha.
- LANGHAMROVÁ, J. at al. (2009): *Prognóza lidského kapitálu obyvatelstva České republiky do roku 2050*. Oeconomica, Praha.
- MAZOUCH, P. and FISCHER, J. (2007): Střední délka života podle nejvyššího ukončeného vzdělání. In: *Firma a konkurenční prostředí*. MSD, Brno, 91-95.

# Two Kurtosis Measures in a Simulation Study

Anna Maria Fiori

Department of Quantitative Methods for Economics and Business Sciences  
University of Milano-Bicocca, Milano, Italy, *anna.fiori@unimib.it*

**Abstract.** We consider two measures of right/left/overall kurtosis which are based on a recent interpretation of kurtosis as inequality at either side of the median. We derive the symmetric influence functions of these measures and discuss their sampling properties by a simulation-based approach. Bootstrap confidence intervals are constructed for small and medium sample sizes. Compared to the standardized fourth moment coefficient (conventional kurtosis), the two measures are shown to provide both a more reliable and a more sophisticated picture of the kurtosis risk embedded in a dataset.

**Keywords:** right kurtosis, inequality, influence function, bootstrap

## 1 Introduction

Although the concept of kurtosis is resurgently playing a role in statistical applications, the conventional kurtosis coefficient:

$$\beta_2 = E \left( \frac{X - \mu}{\sigma} \right)^4 \quad (1)$$

suffers from various weaknesses. Owing to fourth power terms in (1), sample values of  $\beta_2$  can be arbitrarily large, especially when there are one or more tail outliers in the data (Schmid and Trede (2003)). Notwithstanding a central limit effect, the estimators of  $\beta_2$  in finite samples tend to normality very slowly. In addition, their sampling variance is related to the population moment of order eight, and can be significantly large even in large samples (Stuart and Ord (1994)). In this work we suggest various numerical experiments to evaluate the sampling distributions of two alternative kurtosis measures, introduced in Zenga (2006) and characterized in Fiori (2008). The relative importance of the tails in determining these measures is preliminarily discussed by an influence function approach.

## 2 Influence functions for kurtosis measures

Consider a continuous random variable  $X$  with cumulative distribution function (cdf)  $F$ . Define the conditional random variables:

$$\begin{aligned} D &= X - \gamma \mid X > \gamma \\ S &= \gamma - X \mid X \leq \gamma \end{aligned}$$

which describe right and left deviations of  $X$  from its median,  $\gamma = F^{-1}(0.5)$ . Denote by  $\delta_D(F)$  the expectation of  $D$  and by  $\delta_S(F)$  the expectation of  $S$ , so:  $\delta(F) = 0.5(\delta_D + \delta_S) = E(|X - \gamma|)$  is the mean deviation about the median of  $X$ . Without loss of generality, we will assume in the following that  $\gamma = 0$ . As argued in Zenga (2006) (cf. also Fiori (2008)), kurtosis increases as concentration (inequality) increases at either side of the median. As a consequence of this relationship, two kurtosis measures lying between 0 (minimum kurtosis) and 1 (maximum kurtosis) were defined in Zenga (2006) by averaging ratios of right and left scale functionals. The first measure is given by:

$$K_1(F) = \frac{C_D(F) + C_S(F)}{2} \quad (2)$$

where:

$$C_D(F) = 1 - \frac{\delta_D^2(F)}{\mu_{2D}'(F)}$$

is a right kurtosis index for  $\mu_{2D}'(F) = E(D^2)$  and:

$$C_S(F) = 1 - \frac{\delta_S^2(F)}{\mu_{2S}'(F)}$$

is a left kurtosis index, with  $\mu_{2S}'(F) = E(S^2)$ . The second measure is defined by:

$$K_2(F) = \frac{R_D(F) + R_S(F)}{2} \quad (3)$$

where  $R_D(F)$  (right kurtosis) and  $R_S(F)$  (left kurtosis) are the Gini indexes of  $D$  and  $S$ , respectively:

$$R_D(F) = \frac{\Delta_D(F)}{2\delta_D(F)}; \quad R_S(F) = \frac{\Delta_S(F)}{2\delta_S(F)}$$

and the symbol  $\Delta(F)$  stands for the Gini mean difference:

$$\Delta(F) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |x - y| dF(x) dF(y)$$

An intuitive way of understanding the functionals  $K_1$ ,  $K_2$  is through their influence functions, which characterize the limiting effect on the functionals when  $F$  undergoes a small perturbation of the form:

$$F_\varepsilon = (1 - \varepsilon)F + \varepsilon G \quad \text{for } 0 \leq \varepsilon \leq 1 \quad (4)$$

where  $G$  is a contaminating distribution and  $\varepsilon$  is the proportion of contamination. Following Ruppert (1987), we use a restricted notion of influence function called the “symmetric influence function”, which presumes:

$$G = \eta_x = 0.5(\delta_x + \delta_{-x}) \quad (5)$$



where  $\delta_x$  is point mass at the value  $x$ . Then  $\eta_x$  represents a contaminant at  $\pm x$  with equal probability, which leaves the position of the median unchanged. Based on (4) and (5), the symmetric influence function (SIF) of a functional  $T$  at  $F$  and  $x$  is defined by:

$$SIF(x; F, T) = \lim_{\varepsilon \downarrow 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon}$$

In some cases (Ruppert (1987)) it is easier (or more desirable) to compute the SIF of  $\log T$  rather than  $T$  itself. Then,

$$SIF(x; F, T) = T \cdot SIF(x; F, \log T)$$

We now assume that  $F$  is symmetric with finite second moment and discuss the SIF of  $K_1$ . In the symmetric case, both (2) and (3) are completely determined by their right (equivalently, left) components. For the right kurtosis measure  $C_D$  we have:

$$\begin{aligned} \log[1 - C_D(F_\varepsilon)] &= 2 \log \delta_D(F_\varepsilon) - \log \mu'_{2D}(F_\varepsilon) \\ &= 2 \log \{ \delta_D(F) + \varepsilon [x - \delta_D(F)] \} \\ &\quad - \log \left\{ \mu'_{2D}(F) + \varepsilon [x^2 - \mu'_{2D}(F)] \right\} \end{aligned} \quad (6)$$

and:

$$\frac{d}{d\varepsilon_+} \log[1 - C_D(F_\varepsilon)]|_{\varepsilon=0} = 2 \frac{x}{\delta_D(F)} - \frac{x^2}{\mu'_{2D}(F)} - 1 \quad (7)$$

Multiplying (7) by  $[1 - C_D(F)] = \frac{\delta_D^2(F)}{\mu'_{2D}(F)}$  and changing signs gives, after little computation and rearrangement:

$$SIF(x; F, C_D) = SIF(x; F, K_1) = \left( x - \frac{\mu'_{2D}}{\delta_D} \right)^2 - \left[ \left( \frac{\mu'_{2D}}{\delta_D} \right)^2 - \mu'_{2D} \right]$$

which is positive iff  $0 < x < \frac{\mu'_{2D}}{\delta_D} (1 - \sqrt{C_D})$  and  $x > \frac{\mu'_{2D}}{\delta_D} (1 + \sqrt{C_D})$ .

Now consider  $K_2$ . For a symmetric  $F$  with finite first moment we define:

$$SIF(x; F, K_2) = SIF(x; F, R_D) = R_D(F) \cdot \frac{d}{d\varepsilon_+} \log \left[ \frac{\Delta_D(F_\varepsilon)}{2\delta_D(F_\varepsilon)} \right] \Big|_{\varepsilon=0}$$

It can be shown that:

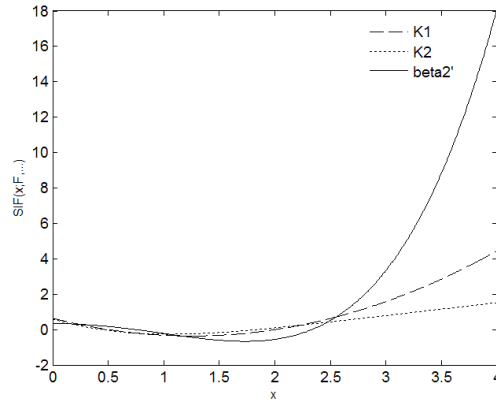
$$\Delta_D(F_\varepsilon) = (1 - \varepsilon)^2 \cdot \Delta_D(F) + 2\varepsilon(1 - \varepsilon) \cdot \delta_{D,x}(F) \quad (8)$$

where  $\delta_{D,x}(F) = E[|X - x| \mid X > 0]$  is the mean deviation of  $D$  about the point of right contamination  $x$ . Based on (6) and (8), the SIF of  $K_2$  is:

$$SIF(x; F, K_2) = \frac{1}{\delta_D} [\delta_{D,x} - R_D(x + \delta_D)]$$

whose roots depend on the value of  $\delta_{D,x}$  at the underlying cdf  $F$ .

Now let  $F$  be the standard normal cdf  $\Phi$ .  $\text{SIF}(x; \Phi, K_1)$  and  $\text{SIF}(x; \Phi, K_2)$  are graphed in Figure 1 over the range  $0 \leq x \leq 4$ . For a meaningful comparison with (1) we superpose  $\text{SIF}(x; \Phi, \beta'_2)$ , where  $\beta'_2 = 1 - 1/\beta_2$  is a normalized version of the conventional kurtosis coefficient. All the measures are increased by contamination in the tails and at the center and are decreased by contamination in the shoulders. Having unbounded SIF, they are sensitive to the location of tail outliers as well as their frequency. However,  $\beta'_2$  is much more sensitive than  $K_1$  and  $K_2$  because its SIF is a quartic function ( $\text{SIF}(x; \Phi, \beta_2)$  is discussed in detail by Fiori and Zenga (2005)).



**Fig. 1.** Symmetric influence functions: the new kurtosis measures  $K_1$  and  $K_2$  compared with a normalized version of the conventional kurtosis coefficient  $\beta'_2$ .

### 3 Inference for kurtosis measures

Consider an i.i.d. sample  $X_1, \dots, X_n$  from  $X$  and denote by  $F_n$  the empirical cdf. Then  $k_1 = K_1(F_n)$  and  $k_2 = K_2(F_n)$  are the natural estimators of the kurtosis measures  $K_1$  and  $K_2$ , both lying between 0 and 1. While a rigorous asymptotic inference for these measures is still under development, we perform here some numerical experiments aimed at describing how  $k_1$  and  $k_2$  behave in small and medium sized samples from various distributions. Again, for a meaningful comparison with  $\beta_2$ , we consider a normalized estimator of the conventional kurtosis coefficient:

$$b'_2 = 1 - \frac{1}{\beta_2(F_n)} \quad \in [0, 1]$$

Subject to the existence of the population moment of order eight,  $b'_2$  is asymptotically normally distributed. However, its finite sample behaviour may differ

significantly from the limiting normal distribution, as emerges from a simple Monte Carlo experiment.

We draw  $N = 20,000$  samples of various sizes ( $n = 20, 40, 80, 160, 320, 640$ ) from three symmetric distributions with different peak-and-tails structure:

- the standard normal;
- the standard Laplace (or double exponential);
- the symmetric Tukey lambda, with shape parameter  $\lambda = 0.089$  (this is regarded as a close approximation to the Student  $t$  distribution with 5 degrees of freedom, but has even moments up the tenth order).

In Table 1 we report the three kurtosis parameters  $K_1$ ,  $K_2$  and  $\beta'_2 = 1 - \frac{1}{\beta_2}$  evaluated at each parent distribution (see Fiori (2005) for computational details).

Distribution	$K_1$	$K_2$	$\beta'_2$
Normal	0.3634	0.4142	0.6667
Laplace	0.5	0.5	0.8333
Tukey	0.4614	0.4638	0.8418

**Table 1.** Kurtosis parameters of the three distributions considered in the Monte Carlo experiment

The natural estimators  $k_1$ ,  $k_2$  and  $b_2$  are computed for each distribution and sample size. Denoting by  $M_1$  the arithmetic mean in a sample of size  $n$ , we compare the relative bias:

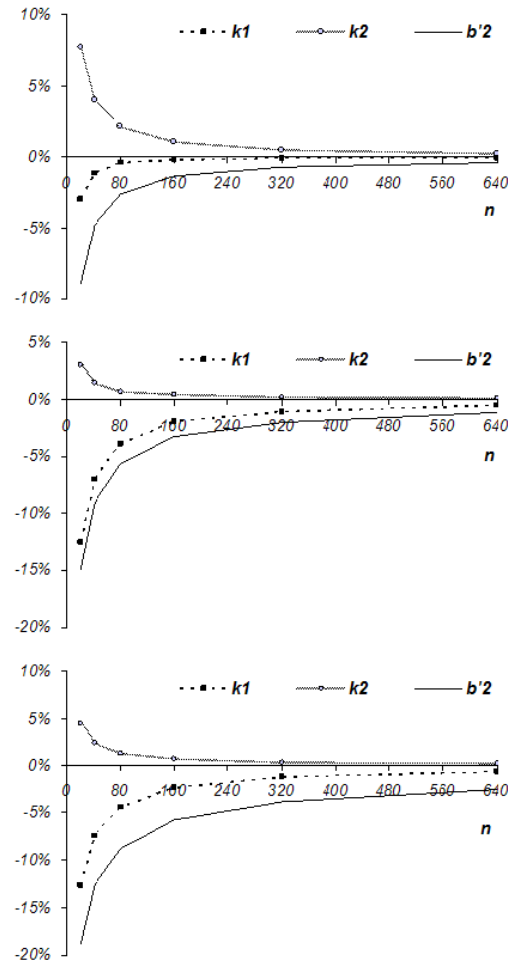
$$RB(T_n, \theta) = \frac{M_1(T_n) - \theta}{\theta}$$

and relative root mean squared error:

$$RRMSE(T_n, \theta) = \frac{\sqrt{M_1(T_n - \theta)^2}}{\theta}$$

of  $T_n = k_{1n}$ ,  $k_{2n}$ ,  $b'_{2n}$  for  $\theta = K_1$ ,  $K_2$ ,  $\beta'_2$ , respectively.

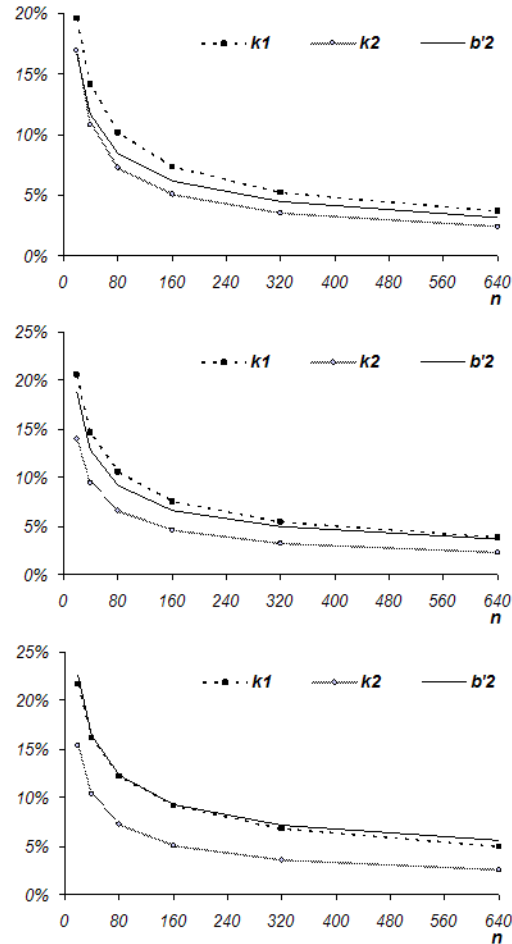
As shown in Figures 2 and 3, the most accurate measure of sample kurtosis is  $k_2$ , which has the lowest  $RB$  and the lowest  $RRMSE$  in all circumstances (with the only exception of the normal samples, for which  $RB(k_{1n}, K_1)$  is the lowest). It interestingly emerges that the sample performance of  $k_2$  improves when the underlying distribution becomes more peaked (i.e. in the Laplace case). We have also verified (Fiori (2005)) that sample histograms of  $k_1$  and  $k_2$  look very close to normality even in small samples. Conversely, the sample performance of  $b'_2$  deteriorates quickly as heavy tailedness of the parent distribution increases.



**Fig. 2.** Relative bias of the three kurtosis measures in samples from the normal distribution (top), the Laplace (center) and the Tukey lambda (bottom).

Using bootstrap techniques we derive nonparametric confidence intervals for the three kurtosis parameters  $K_1$ ,  $K_2$  and  $\beta'_2$  in finite samples and validate the results by another Monte Carlo experiment. We draw  $N = 10,000$  samples of  $n = 40$  (small size) and  $n = 160$  (medium size) from the same distributions considered previously (normal, Laplace and lambda). For each sample we compute the bootstrap percentile confidence intervals of  $K_1$ ,  $K_2$  and  $\beta'_2$  at 95% confidence level, based on  $B = 2000$  bootstrap resamples (see Efron and Tibshirani (1998) for a detailed description). The empirical coverage and average length of these confidence intervals are reported in Tables 2 and 3 (more sophisticated, e.g. BCa confidence intervals, were

also computed but they did not provide significantly different results). While confidence intervals for  $\beta'_2$  suffer from significant undercoverage, these Monte Carlo/bootstrap experiments confirm that the alternative kurtosis parameter  $K_2$  (and, partially,  $K_1$ ) can be estimated with higher accuracy in small and medium samples.



**Fig. 3.** Relative RMSE of the three kurtosis measures in samples from the normal distribution (top), the Laplace (center) and the Tukey lambda (bottom).

A promising direction for further research is therefore the asymptotic inference for the kurtosis estimator  $k_2$ , possibly in view of its practical application in financial contexts. Decomposing this measure into its right and left kurtosis components is likely to convey useful information to investors and

risk managers, whose risk perceptions are typically related to the left tail of return distributions. Empirical comparison with quantile based measures of kurtosis (with bounded influence function, e.g. Brys et al. (2006)) could be considered as well.

	$n = 40$			$n = 160$		
	Normal	Laplace	Lambda	Normal	Laplace	Lambda
$K_1$	0.9879	0.8420	0.8561	0.9700	0.8702	0.8710
$K_2$	0.9877	0.9770	0.9903	0.9760	0.9620	0.9597
$\beta'_2$	0.8825	0.5500	0.3573	0.8810	0.6520	0.4760

**Table 2.** Empirical coverage (fraction of 10,000 samples which included the true kurtosis parameter) of nonparametric bootstrap confidence intervals at 95% confidence level for the three measures  $K_1$ ,  $K_2$  and  $\beta'_2$ .

	$n = 40$			$n = 160$		
	Normal	Laplace	Lambda	Normal	Laplace	Lambda
$K_1$	0.2112	0.2302	0.2300	0.1036	0.1278	0.1331
$K_2$	0.1848	0.1901	0.1914	0.0826	0.0899	0.0914
$\beta'_2$	0.2591	0.2608	0.2756	0.1363	0.1416	0.1611

**Table 3.** Average length of nonparametric bootstrap confidence intervals at 95% confidence level for the three measures  $K_1$ ,  $K_2$  and  $\beta'_2$ .

## References

- BRYs, G., HUBERT, M. and STRUYF, A. (2006): Robust measures of tail weight. *Computational Statistics and Data Analysis* 50(3), 733–759.
- EFRON, B. and TIBSHIRANI, R. J. (1998): *An Introduction to the Bootstrap*. Chapman and Hall, London.
- FIORI, A. M. (2008): Measuring kurtosis by right and left inequality orders. *Communications in Statistics: Theory and Methods* 37 (17), 2665–2680.
- FIORI, A. M. and ZENGA, M. (2005): The meaning of kurtosis, the influence function and an early intuition by L. Faleschini. *Statistica* 65 (2), 135–144.
- FIORI, A. M. (2005): *Kurtosis: new theoretical results and inference issues*. Unpublished Ph.D. thesis. University of Milano-Bicocca.
- RUPPERT, D. (1987): What is kurtosis? An influence function approach. *The American Statistician* 41 (1), 1–5.
- STUART, A. AND ORD, J. K. (1994): *Kendall's Advanced Theory of Statistics. Volume I: Distribution Theory*. Edward Arnold, London.
- ZENGA, M. (2006): Kurtosis. In: S. Kotz, C. B. Read, N. Balakrishnan and B. Vidakovic (Eds.): *Encyclopedia of Statistical Sciences*. Wiley, New York, 2nd online edition.

# Clustering of Czech Household Incomes Over Very Short Time Period

Marie Forbelská<sup>1</sup> and Jitka Bartošová<sup>2</sup>

<sup>1</sup> Masaryk University, Department of Mathematics and Statistics of the Faculty of Science, Kotlářská 2, Brno, Czech Republic, *forbel@math.muni.cz*

<sup>2</sup> University of Economics Prague, Department of Management of Information of the Faculty of Management, Jarošovská 1117/II, Jindřichuv Hradec, Czech Republic, *bartosov@fm.vse.cz*

**Abstract.** The article deals with cluster analysis of household income dynamics based on the results of statistical survey EU SILC 2005, 2006 and 2007. We handle the problem of clustering many short time series. Mixed effects models offer a flexible framework for appropriate modeling of among trial correlations and individual trial variance heterogeneity. Consequently, we assume that random parameters are distributed according to a finite normal mixture and we use this mixture model for clustering short time series. The R environment (R Development Core Team, 2008) is used for both mixed model analysis and cluster analysis .

**Keywords:** household income, finite mixture model, clustering, mixed effects models

## 1 Introduction

Income distribution provides a basis for the evaluation of a country's living standards for the population as a whole. Additionally, it provides a comparison for the living standards of different social classes.

## 2 Model Specification

The distribution of household income in most populations is highly skewed, with a long right-hand-side tail and high density at the lower percentiles. The logarithm is the natural transformation for such data.

Let

$$\{Y_{it} = \log(\text{income})_{it}\}_{t=1, \dots, T}$$

be a panel of multiple time series observed for  $N$  units  $i = 1, \dots, N$ . Denote also by  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT})'$   $i$ -th univariate time series with the joint density  $f(\mathbf{y}; \boldsymbol{\theta}_i)$ , where  $\boldsymbol{\theta}_i$  is unknown parameters that need to be estimated from the data.

If  $T$  were large, the parameters  $\boldsymbol{\theta}_i$  could be estimated for each time series  $\mathbf{Y}_i$  individually. However, if  $T$  is relatively small one might use information from the other time series in the panel to estimate unknown parameters.

## 2.1 Simple Linear Mixed Model with Autoregressive Errors (LMM-AR)

For our longitudinal data we assume a very simple linear mixed model with autoregressive errors  $\varepsilon_i \sim AR(1)$  described by the structure

$$\mathbf{Y}_i = (\mathbf{1}, \mathbf{t})\beta_i^* + \varepsilon_i, \quad (1)$$

where  $\beta_i^* = (\alpha_i^*, \beta_i^*)' = (\alpha + a_i, \beta + b_i)'$ ,  $\mathbf{1}$  is vector of ones,  $\mathbf{t} = (1, \dots, T)'$ ,  $i = 1, \dots, N$ . In this model we call  $\beta = (\alpha, \beta)'$  the fixed effects (fixed intercept and fixed slope) and  $\mathbf{b}_i = (a_i, b_i)'$  the random effects (random intercept and random slope).

We assume that the random vectors  $\mathbf{b}_i$  and  $\varepsilon_i$  are independent and identically distributed with zero means and covariance matrices  $cov(\mathbf{b}_i) = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}$ ,

$$cov(\varepsilon_i) = \frac{\sigma^2}{1-\varphi^2} \mathbf{R} = \frac{\sigma^2}{1-\varphi^2} (\varphi^{|j-k|})_{j,k=1,\dots,T} \text{ (with } |\varphi| < 1).$$

Parameters of the mixed model can be estimated using Maximum Likelihood Estimation (MLE) or Restricted Maximum Likelihood Estimation (RMLE), while the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC) can be used as measures of goodness of fit for particular models, where smaller values for both are considered more preferable.

We use the *nlme* package (Pinheiro and Bates, 2000) of the R environment for fitting and examining linear mixed-effects models.

## 2.2 Mixture model of random coefficients

Consequently, we assume that random coefficients  $\beta_i^* = (\alpha_i^*, \beta_i^*)'$  are distributed according to a finite normal mixture. With the mixture approach to clustering  $\mathbf{X}_i = (X_{i1}, X_{i2})' = \beta_i^*$ , let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  are assumed to be an observed random sample from mixture of a finite number, say  $K$ , of groups in some unknown proportions  $p_1, \dots, p_K$ . For the bivariate case, the normal mixture density of  $\mathbf{x}_i$  is expressed as  $f(\mathbf{x}_i; \Psi) = \sum_{k=1}^K p_k f_k(\mathbf{x}_i; \theta_k) =$

$$\sum_{k=1}^K \frac{p_k}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_k^2}} \exp\left\{-\frac{1}{2(1-\rho_k^2)} \left[ \frac{(x_{i1}-\mu_{k1})^2}{\sigma_{k1}^2} - 2\rho_k \frac{x_{i1}-\mu_{k1}}{\sigma_{k1}} \frac{x_{i2}-\mu_{k2}}{\sigma_{k2}} + \frac{(x_{i2}-\mu_{k2})^2}{\sigma_{k2}^2} \right]\right\}.$$

Using an estimate of the vector of all unknown parameters  $\Psi$ , this approach gives a probabilistic clustering of the data into  $K$  clusters in terms of estimates of the posterior probabilities of component membership

$$\omega_k(\mathbf{x}_i) = \frac{p_k f_k(\mathbf{x}_i; \theta_k)}{f(\mathbf{x}_i; \Psi)},$$

where  $\omega_k(\mathbf{x}_i)$  is the posterior probability that  $\mathbf{x}_i$  (really the time series characterized by coefficients  $\mathbf{x}_i$ ) belongs to the  $k$ th component of the mixture ( $i = 1, \dots, N$ ,  $k = 1, \dots, K$ ).



In the Bayesian framework, we use the rule which assigns observation  $\mathbf{x}_i$  to the class for which  $\mathbf{x}_i$  has the highest posterior probability.

The parameter vector  $\boldsymbol{\Psi}$  can be estimated by maximum likelihood (MLE) and can be obtained via the Expectation–Maximization (EM) algorithm of Dempster et al. (1977).

In practice, the number of components  $K$  is unknown and can be chosen as that which minimizes some criterion, e.g. Bayesian Information Criterion BIC of Schwarz (1978), see also McLachlan and Peel (2000).

### 3 Data

A sample survey of household income in the Czech Republic is made by the Czech Statistical Office (CSO). From the fifties of the last century there was an irregular survey, which took place at intervals of 2 to 5 years under the name Microcensus.

After the entrance to the European Union, Microcensus was replaced by an annual survey of income and living conditions of households called EU - SILC. (European Union - Statistics on Income and Living Conditions). For the first time this investigation was carried out by the Czech Statistical Office in 2005 under the name Living Conditions 2005.

Investigation is carried out by the so-called rotating panel, where the same households were re-interviewed in the annual intervals for four years. After this time they are replaced by other households living in the newly visited homes that are added to the investigation file continuously by the random selection. Longer monitoring of a household permits building an image of their social situation, not only in the year, but also the changes and developments over time.

## 4 Clustering of household incomes over years 2005, 2006 and 2007

Clustering is typically used as a tool for understanding and exploring large data sets. The clustering algorithm discussed here consists of two phases: fitting simple linear mixed model LMM-AR and classification of random coefficients via finite mixture model.

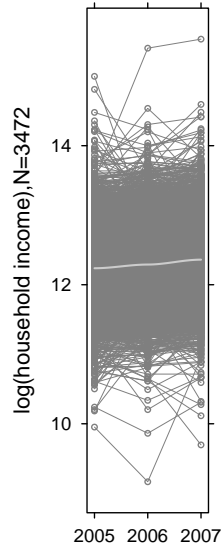
Our approach extracts random intercept and slope from a time series, and uses information about linear trend to clustering household incomes over years 2005, 2006 and 2007.

### 4.1 Fitting LMM-AR model

Plotting the profiles of log incomes over time as a trace for each household suggests that the simple linear mixed model denoted by LMM-AR can be

assumed (see Figure 1). The loess trend line indicates a gentle increase in the level of log incomes.

In Table 1, the results of the LMM-AR model are presented.

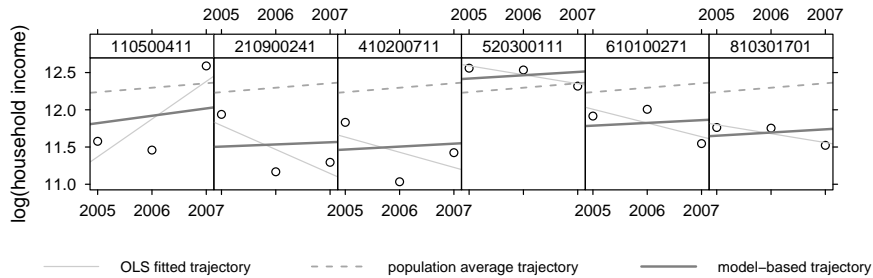


**Fig.1.** EU SILC data and loess curve.

Final estimation of fixed effects					
	Value	Std.Error	DF	t-value	p-value
$\alpha$	12.296	0.0095	6943	1293.54	0
$\beta$	0.059	0.0027	6943	21.47	0
Final estimation of variance components					
Description		Value			
$\sigma_a$		0.537			
$\sigma_b$		0.046			
$\rho_{ab} = \frac{\sigma_{ab}}{\sigma_a \sigma_b}$		0.061			
$\sigma$		0.230			
$\varphi$		0.309			
Model fitting information for responses					
Description		Value			
AIC		7225.97			
BIC		7276.73			
loglik		-3605.99			

**Table 1.** Results of LMM-AR model

Figure 2 provides a graphical illustration of the results for six randomly selected households. Each subject's data are shown in a separate panel. The subject number is given in the strip above the panel.

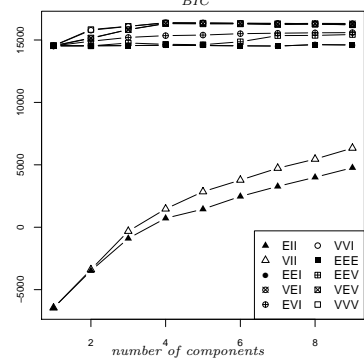


**Fig.2.** Log incomes of randomly selected households with OLS line (thin solid line), the subject-specific line with both fixed and random coefficients (thick solid line), and the population-average line with fixed coefficients (dashed line).

## 4.2 Model based clustering of random coefficients

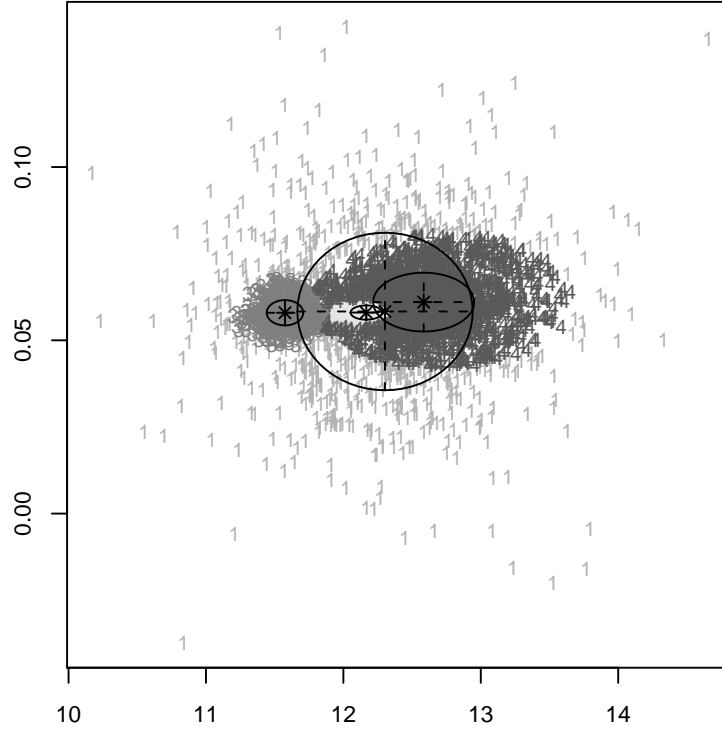
We now proceed to apply model-based clustering to the random coefficients using the *mclust* package (Fraley and Raftery, 2006). Software *mclust* is promising in that it uses a general multivariate Gaussian mixture model

to account for various possible covariance structures and automatically gives an estimate of the number of clusters by the Bayesian Information Criterion.



**Fig.3.** The Bayesian Information Criterion (BIC) for model-based methods applied to the random coefficients.

In this diagram three characters refer to different model assumptions about the shape of clusters. (*EII*: spherical, equal volume, *VII*: spherical, unequal volume, *EEI*: diagonal equal volume, equal shape, *VEI*: diagonal varying volume, equal shape, *EVI*: diagonal equal volume, varying shape, *VVI*: diagonal varying volume, varying shape, *EEE*: ellipsoidal, equal volume, shape and orientation, *EEV*: ellipsoidal, equal volume and shape, varying orientation, *VEV*: ellipsoidal, varying volume and orientation, equal shape, *VVV*: ellipsoidal, varying volume, shape and orientation). The model with the highest BIC value ( $BIC = 16390$ ) is a four component mixture with the *VVI* covariance structure.

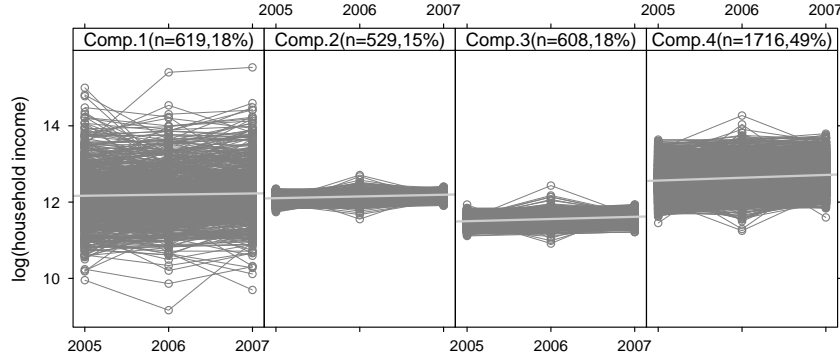


**Fig.4.** Classification plot of random parameters  $\beta_i^*$ . Ellipses superimposed on the plot correspond to the covariances of the components. In the classification plot, points in different classes are indicated by different symbols (1-4).

The fitted parameters of the mixture for each component are given in Table 2.

Component	$p_k$	$\mu_{1k}$	$\mu_{2k}$	$\sigma_{1k}$	$\sigma_{2k}$	$\rho_k$
1	0.267	12.303	0.0583	0.638	0.0223	0
2	0.126	12.167	0.0580	0.115	0.0021	0
3	0.159	11.575	0.0579	0.133	0.0037	0
4	0.448	12.585	0.0610	0.367	0.0085	0

**Table 2.** Resulting parameters of fitted mixtures for random coefficients  $\beta_i^*$ .



**Fig.5.** Resulting clusters of household log incomes over years 2005, 2006 and 2007.

Finally, we applied the above defined LMM-AR mixed model (1) separately on each of the four components. The fitted parameters for each component are given in Table 3.

Description	Comp.1	Comp.2	Comp.3	Comp.4
$\alpha$ - Value	12.1962	12.1472	11.5557	12.6383
$\alpha$ - Std.Error	0.0284	0.0048	0.0055	0.0085
$\alpha$ - p-value	0.0000	0.0000	0.0000	0.0000
$\beta$ - Value	0.0298	0.0450	0.0577	0.0736
$\beta$ - Std.Error	0.0136	0.0010	0.0018	0.0024
$\beta$ - p-value	0.0287	0.0000	0.0000	0.0000
$\sigma_a$	0.6832	0.1102	0.1338	0.3468
$\sigma_b$	0.2615	0.0016	0.0021	0.0098
$\rho_{ab}$	-0.0390	0.0420	0.0700	-0.2640
$\sigma$	0.3043	0.0540	0.0709	0.1437
$\varphi$	-0.0003	-0.7842	-0.4357	-0.2230

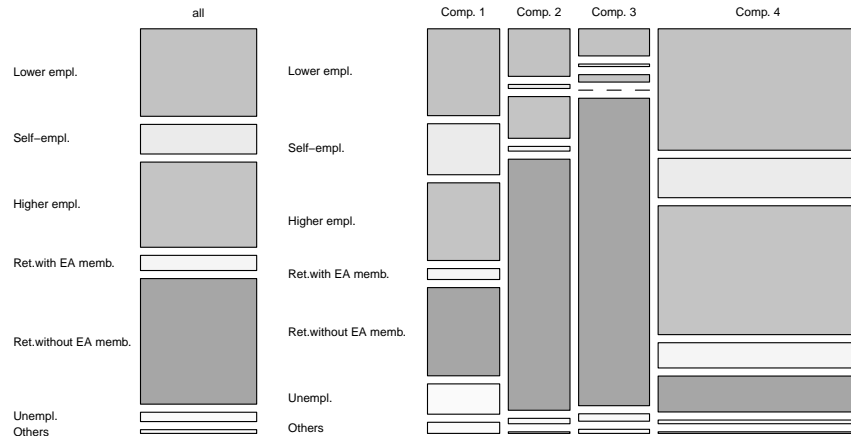
**Table 3.** Resulting parameters of LMM-AR for each component.

It can be seen from Table 3 and Figure 5 that for cluster *Comp.1*, which consisted of 18% of the sample, all variance components are the largest, but fixed slope is the lowest. Cluster *Comp.4* (49% of the sample) has the largest fixed intercept and slope. Clusters *Comp.2* (15% of the sample) and *Comp.3* (18% of the sample) have similar variance components, but *Comp.2* has a higher fixed intercept.

Performing more detailed analysis of the households, we find, for example (see Figure 6 and Figure 7), that the households in the cluster *Comp.4* are characterized by a higher proportion of male householders and lower proportion of pensioners. Households from the clusters *Comp.2* and *Comp.3* have a high proportion of pensioners, but households in the cluster *Comp.2* are characterized by a small proportion of female householders in contrast to households from the cluster *Comp.2* with a high proportion of female householders.



**Fig.6.** Mosaic plots for factors householder and components. Levels of *householder*: male, female.



**Fig.7.** Mosaic plots for factors social status of householder and components. Levels of *social status*: lower employee, self-employed, higher employee, retired (not working) with working members in household, retired (not working) without working members in household, unemployed, others.

Detailed economic analysis of the structure of each cluster was not the aim of this contribution.

## 5 Conclusion

In this paper, we address the problem of clustering very short time series of household incomes. Our proposed method extracts random intercept and slope from a time series, and uses information about the linear trend to clustering household incomes over a very short time period.

## Acknowledgments

The research was supported by project of Grant Agency of the Czech Republic no. 402/09/0515 with title: "Analysis and modeling of financial power of Czech and Slovak Households".

## References

- AIKAKE, H. (1973): Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium in Information Theory* (eds B. N. Petrov and F. Csaki). Budapest: Akademiai Kiado.
- DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. (1977): Maximum likelihood estimation from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, 39, 1-38
- FRALEY, C., RAFTERY, A. E. (2006): *MCLUST: Normal Mixture Modeling and Model-Based Clustering*. R package version 3.0-0.
- McCULLOCH, C.E., SEARLE, S.R. (2001): *Generalized, linear, and mixed models*, Wiley
- McLACHLAN, G. J., PEEL, D. (2000): *Finite Mixture Models*. Wiley, New York.
- PINHEIRO, J. C., BATES, D. M. (2000): *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag. New York.
- R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- SCHWARTZ, G. (1978): Estimating the Dimension of a Model. *The Annals of Statistics*, 6 (2), 461-464.

# Model-Based Nonparametric Variance Estimation for Systematic Sampling. An Application in a Forest Survey

Mario Francisco-Fernández<sup>1</sup>, Jean Opsomer<sup>2</sup>, and Xiaoxi Li<sup>3</sup>

<sup>1</sup> Universidad de A Coruña. Departamento de Matemáticas, Facultad de Informática, A Coruña, 15071, Spain, *mariofr@udc.es*

<sup>2</sup> Colorado State University. Department of Statistics, Fort Collins, CO 80523, USA, *jopsomer@stat.colostate.edu*

<sup>3</sup> Pfizer, Inc. Groton, CT 06340, USA, *xiaoxi.li@pfizer.com*

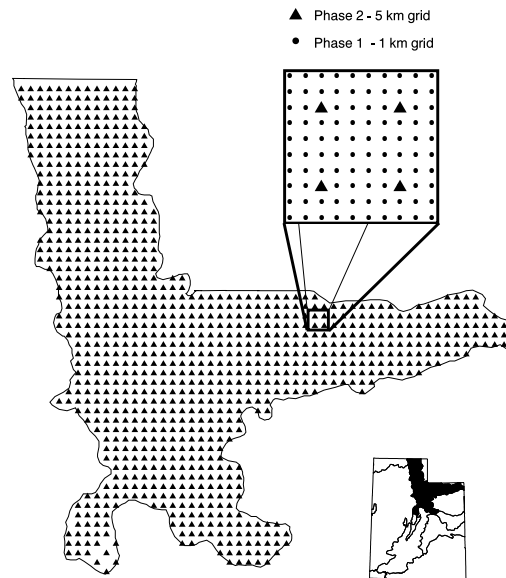
**Abstract.** Systematic sampling is frequently used in natural resource and other surveys, because of its ease of implementation and its design efficiency. An important drawback of systematic sampling, however, is that no direct estimator of the design variance is available. We describe a new estimator of the model-based expectation of the design variance, under a nonparametric model for the population. We prove the model consistency of the estimator for both the anticipated variance and the design variance. We compare the nonparametric variance estimators with several design-based estimators on data from a forestry survey.

**Keywords:** local polynomial regression, two-per-stratum variance approximation, smoothing

## 1 Introduction

The *Forest Inventory and Analysis* (FIA) is a program within the US Department of Agriculture Forest Service that conducts nationwide forest surveys. Sampling for the FIA surveys has traditionally followed a stratified systematic design. In these surveys, the population quantities of interest are, for example, total tree volume, growth and mortality, and area by forest type. Design-based estimates of such quantities are produced on a regular basis. In this work, we are considering survey data collected during the 1990's by the Forest Service within a 2.5 million ha ecological province in northern Utah, USA. Forest resource data are collected through field visits on sample plots located on a regular spatial grid. These field-level data are supplemented by remotely sensed data available on a much finer spatial grid. Figure 1 displays the study region and sample locations for the survey data and additional remote sensing data. In the current work, we will use the auxiliary information to construct an estimator for the variance of survey estimators.

A well-known and long-standing issue in surveys that follow a systematic sampling design is the lack of a theoretically justified, generally applicable



**Fig. 1.** Map of the study region in northern Utah. Each triangle represents a field-visited phase two sample point. Each dot in the magnified section represents a phase one sample point.

design-based variance estimator. A whole chapter of the recently reissued classic monograph by Wolter (Wolter (2007)) is devoted to this issue, and a number of possible estimation approaches are evaluated there. In particular, it considers a set of eight “model-free” estimators and outlines a model-based estimation approach. For the set of eight estimators, their statistical properties are evaluated for several model scenarios and through simulation experiments. None of these estimators is best overall, and there was a clear interaction between the behavior of the estimators and the underlying data model. On the other hand, in Bartolucci and Montanari (2006), an unbiased model-based variance estimator when the population follows a linear regression model is proposed.

In practice, despite its potential efficiency, wide applicability of the model-based method is viewed as being hampered by lack of robustness. However, this lack of robustness can be at least partly offset by the use of a nonparametric model specification. Compared to parametric models, this class of models makes much less restrictive assumptions on the shape of the relationship between variables. Hence, the risk of model misspecification is significantly reduced. This is particularly important in the survey context, because the same variance estimation method often needs to be applied to many survey variables collected in the same survey, and a single parametric model is much less likely to be correct for all these variables.



In the current work, we will consider a broadly applicable model for the data, in which both the mean and the variance are left unspecified subject only to smoothness assumptions. We propose a model-based nonparametric variance estimator, in which both the mean and the variance functions of the data are estimated nonparametrically. We will show that the proposed estimator is model consistent for the design variance of the survey estimator, subject only to the population smoothness assumptions. We also evaluate the practical properties of the estimator in a simulation study, and with the analysis for the northern Utah forestry data previously presented

## 2 Systematic sampling and design-based variance estimation

We will be sampling from a finite population  $U$  of size is  $N$ . We consider a single study variable  $Y_j \in \mathbb{R}$ ,  $j = 1, 2, \dots, N$  with population mean

$$\bar{Y}_N = \frac{1}{N} \sum_{j=1}^N Y_j.$$

Let  $n$  denote the sample size and  $k = N/n$  denote the *sampling interval*. For simplicity, we assume throughout this article that  $N$  is an integral multiple of  $n$ . The variable  $Y$  will only be observed on the sampled elements only.

Let  $\mathbf{x}_j \in \mathbb{R}^p$  ( $j = 1, 2, \dots, N$ ) be vectors of auxiliary variables available for all the elements in the population. To draw a systematic sample, the population is first sorted by some appropriate criterion. After sorting the population, drawing a systematic sample is done by randomly choosing an element among the first  $k$  with equal probability, say the  $b$ th one, after which the systematic sample, denoted by  $S_b$ , consists of the observations with labels  $\{b, b+k, \dots, b+(n-1)k\}$ . The random sample  $S$  can therefore only take on  $k$  values on the set of possible samples  $\{S_1, \dots, S_k\}$ .

The sample mean,

$$\bar{Y}_S = \frac{1}{n} \sum_{j \in S} Y_j,$$

is the Horvitz-Thompson estimator for the finite population mean. Its design-based variance is equal to

$$\text{Var}_p(\bar{Y}_S) = \frac{1}{k} \sum_{b=1}^k (\bar{Y}_{S_b} - \bar{Y}_N)^2. \quad (1)$$

It should be clear that, if only a single systematic sample is drawn and hence only one of the  $\bar{Y}_{S_b}$  is observed, no unbiased design-based estimator of

$\text{Var}_p(\bar{Y}_S)$  exists for general variable  $Y$ . A more formal way to state this is that the systematic sampling design is *not measurable*.

We describe the three main methods used in practice to estimate  $\text{Var}_p(\bar{Y}_S)$ . The simplest estimator is to treat the systematic sample as if it had been obtained by simple random sampling without replacement:

$$\hat{V}_{SRs} = \frac{1-f}{n} \frac{1}{n-1} \sum_{j \in S} (Y_j - \bar{Y}_S)^2, \quad (2)$$

where  $f = n/N$ . The two remaining estimators are based on pairwise differences and are recommended in Wolter (2007) as being the best general-purpose estimators of  $\text{Var}_p(\bar{Y}_S)$ . They are defined as

$$\hat{V}_{OL} = \frac{1-f}{n} \frac{1}{2(n-1)} \sum_{j=2}^n (Y_j - Y_{j-1})^2, \quad (3)$$

and

$$\hat{V}_{NO} = \frac{1-f}{n} \frac{1}{n} \sum_{j=1}^{n/2} (Y_{2j} - Y_{2j-1})^2. \quad (4)$$

Estimators (2), (3) and (4) are design biased for  $\text{Var}_p(\bar{Y}_S)$  in general.

### 3 Variance estimation under a nonparametric model

In the model-based context, the finite population is regarded as a random realization from a superpopulation model. In this section, we propose a model consistent variance estimator under the following nonparametric model:

$$Y_j = m(x_j) + v(x_j)^{1/2} e_j \quad 1 \leq j \leq N, \quad (5)$$

where  $m(\cdot)$  and  $v(\cdot)$  are continuous and bounded functions. The errors  $e_j$ ,  $1 \leq j \leq N$ , are independent random variables with model mean 0 and variance 1. Define  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_N)^T$ ,  $\mathbf{m} = (m(x_1), \dots, m(x_N))^T$  and  $\mathbf{\Sigma} = \text{diag}\{v(x_1), v(x_2), \dots, v(x_N)\}$ .

The design variance of  $\bar{Y}_S$  can be written as

$$\text{Var}_p(\bar{Y}_S) = \frac{1}{k} \sum_{b=1}^k (\bar{Y}_{S_b} - \bar{Y}_N)^2 = \frac{1}{kn^2} \mathbf{Y}^T \mathbf{D} \mathbf{Y}, \quad (6)$$

where  $\mathbf{D} = \mathbf{M}^T \mathbf{H} \mathbf{M}$ , with  $\mathbf{M} = \mathbf{1}_n^T \otimes \mathbf{I}_k$  and  $\mathbf{H} = \mathbf{I}_k - \frac{1}{k} \mathbf{1}_k \mathbf{1}_k^T$ , with  $\otimes$  denoting Kronecker product and  $\mathbf{1}_r$  a vector of 1's of length  $r$ . Then, the model anticipated variance of  $\bar{Y}_S$  under model (5) is

$$\text{E}[\text{Var}_p(\bar{Y}_S)] = \frac{1}{kn^2} \mathbf{m}^T \mathbf{D} \mathbf{m} + \frac{1}{kn^2} \text{tr}(\mathbf{D} \mathbf{\Sigma}). \quad (7)$$

To estimate  $E[\text{Var}_p(\bar{Y}_S)]$ , we propose the following estimator

$$\hat{V}_{NP} = \frac{1}{kn^2}(\hat{\mathbf{m}}^T \mathbf{D} \hat{\mathbf{m}}) + \frac{1}{kn^2} \text{tr}(\mathbf{D} \hat{\Sigma}), \quad (8)$$

where  $\hat{\mathbf{m}} = (\hat{m}(x_1), \dots, \hat{m}(x_N))^T$ , with  $\hat{m}(x_j)$  the local polynomial regression (LPR) estimator of  $m(x_j)$  computed on the observations in the sample  $S$  with kernel  $K$  and bandwidth  $h_m$ , given by

$$\hat{m}(x_j) = \mathbf{e}_1^T (\mathbf{X}_{Sj}^T \mathbf{W}_{Sj} \mathbf{X}_{Sj})^{-1} \mathbf{X}_{Sj}^T \mathbf{W}_{Sj} \mathbf{Y}_S,$$

with  $\mathbf{e}_1$  a vector of length  $(p+1)$  having 1 in the first entry and all other entries 0,  $\mathbf{Y}_S$  a vector containing the  $Y_j \in S$ ,  $\mathbf{X}_{Sj}$  a matrix with  $i$ th row equal to  $(1, (x_i - x_j), \dots, (x_i - x_j)^p)$ ,  $i \in S$ , and

$$\mathbf{W}_{Sj} = \text{diag} \left\{ K \left( \frac{x_i - x_j}{h_m} \right), \quad i \in S \right\},$$

and  $\hat{\Sigma} = \text{diag}\{\hat{v}(x_1), \hat{v}(x_2), \dots, \hat{v}(x_N)\}$ , with  $\hat{v}(x_j)$  the LPR estimator of  $v(x_j)$ . The expression of  $\hat{v}(x_j)$  is completely analogous to  $\hat{m}(x_j)$ , except that  $\mathbf{Y}_S$  is replaced by the vector of squared residuals  $\hat{\mathbf{r}}_S$  with elements  $\hat{r}_j = (Y_j - \hat{m}(x_j))^2$ ,  $j \in S$ , and a different bandwidth  $h_v$  is used instead of  $h_m$  in the weight matrix  $\mathbf{W}_{Sj}$ .

Under suitable regularity conditions, we have the following Theorem.

**Theorem 1.** *Assume that the degree  $p$  of the local polynomial is odd. Using superpopulation model (5) and under some regularity assumptions, the design variance is model consistent for the anticipated variance, in the sense that*

$$\text{Var}_p(\bar{Y}_S) - E[\text{Var}_p(\bar{Y}_S)] = O_p \left( \frac{1}{\sqrt{N}} \right), \quad (9)$$

*and the nonparametric variance estimator is model consistent for the anticipated variance and for the design variance, in the sense that*

$$\hat{V}_{NP} - E[\text{Var}_p(\bar{Y}_S)] = O_p(h_m^{p+1}) + O_p \left( \frac{1}{\sqrt{nh_m}} \right) \quad (10)$$

and

$$\hat{V}_{NP} - \text{Var}_p(\bar{Y}_S) = O_p(h_m^{p+1}) + O_p \left( \frac{1}{\sqrt{nh_m}} \right). \quad (11)$$

In Li (2006), a simpler nonparametric estimator is defined as

$$\hat{V}_{NP}^{ho} = \frac{1}{kn^2}(\hat{\mathbf{m}}_S^T \mathbf{D} \hat{\mathbf{m}}_S) + \frac{1}{kn^2} \text{tr}(\mathbf{D}) \hat{\sigma}_S^2 \quad (12)$$

with

$$\hat{\sigma}_S^2 = \frac{1}{n} \sum_{j \in S} (Y_j - \hat{m}(x_j))^2, \quad (13)$$

and its properties were studied under the special case of superpopulation model (5) with homoscedastic errors, i.e. when  $v(x_j) \equiv \sigma^2$ ,  $j = 1, \dots, N$ .

## 4 Simulation study

The performance of the estimators  $\hat{V}_{NP}$ ,  $\hat{V}_{NP}^{ho}$ ,  $\hat{V}_{OL}$ ,  $\hat{V}_{NO}$  and  $\hat{V}_{SRs}$  described in Sections 2 and 3 was compared in a comprehensive simulation study. Different mean and variance functions (including homocedastic errors), sample sizes and sorting criteria used in generating the systematic samples were considered. To compare, we used the relative bias  $\left( \text{RB} = \frac{E^*(\hat{V}) - E^*[\text{Var}_p(\bar{Y}_S)]}{E^*[\text{Var}_p(\bar{Y}_S)]} \right)$ , the mean squared error ( $\text{MSE} = E^*(\hat{V} - E^*[\text{Var}_p(\bar{Y}_S)])^2$ ) and the mean squared prediction error ( $\text{MSPE} = E^*(\hat{V} - \text{Var}_p(\bar{Y}_S))^2$ ), with  $\hat{V}$  one of the estimators above, and  $E^*$  indicates which expectations are obtained by averaging across the replicates.

In general, the results showed that nonparametric estimators  $\hat{V}_{NP}^{ho}$  and  $\hat{V}_{NP}$  perform well under all superpopulation models if proper bandwidths  $h_m$  and  $h_v$  values are chosen. The results also showed that the variance function specification generally has only a modest effect on the performance of the estimators. An interesting result is that the estimator  $\hat{V}_{NP}^{ho}$  appears to perform better than many of the more complicated  $\hat{V}_{NP}$  even when the errors were heteroscedastic. Full results are in Opsomer et al. (2009).

## 5 Application in Forest Inventory and Analysis

We now return to the FIA data collected in Northern Utah. Data are available for 24,980 remote sensing points and 968 field-visited points. It should be noted that the remote sensing data are available at essentially any desirable resolution, so this grid of points is somewhat arbitrary and can be used as an approximation for the underlying continuous population. We therefore treat these as the population of interest, and field-visited points as a sample drawn from that population, corresponding to 1-in-25 systematic sample. At the “population” level, we have auxiliary information such as location (**LOC**, bivariate scaled longitude and latitude) and elevation (**ELEV**). At the sample level, information is available for the field-collected forestry variables in addition to the population-level variables.

We consider here the following representative forestry variables:

- BIOMASS - total wood biomass per acre in tons
- CRCOV - percent crown cover
- BA - tree basal area per acre
- NVOLTOT - total cubic feet volume per acre
- FOREST - forest/nonforest indicator.

We are interested in estimating the population mean for these variables using the systematic sample mean  $\bar{Y}_S$ , and estimating its design-based variance  $\text{Var}_p(\bar{Y}_S)$ . We will consider two traditional design-based variance estimators,  $\hat{V}_{SRs}$  as in (2) and  $\hat{V}_{ST}$  (see below), and the model-based nonparametric

	$\bar{Y}_S$	$\hat{V}_{SRS}$	$\hat{V}_{ST}$	$\hat{V}_{NP0.5}^{ho}$	$\hat{V}_{NP0.2}^{ho}$	$\hat{V}_{NP0.1}^{ho}$
BIOMASS	14.5	0.46	0.36	0.40	0.38	0.37
CRCOV	22.5	0.71	0.62	0.64	0.62	0.59
BA	48.5	3.87	3.19	3.40	3.30	3.12
NVOLTOT	906.9	1886	1538	1645	1584	1511
FOREST (%)	54.8	2.46	1.89	2.16	2.05	1.91

**Table 1.** Mean and variance estimates for the five response variables, using estimators  $\hat{V}_{SRS}$ ,  $\hat{V}_{ST}$  and  $\hat{V}_{NP}^{ho}$  under model (14) with span = 0.5, 0.2 and 0.1.

variance estimator  $\hat{V}_{NP}$ . The stratified sampling variance estimator  $\hat{V}_{ST}$  is similar to the nonoverlapping differences estimator  $\hat{V}_{NO}$  in (4), generalized to a spatial setting by considering an approximate 4-per-stratum design obtained by overlaying a grid of equal-sized “cells” over the study region.

$$\hat{V}_{ST} = \frac{1-f}{n} \frac{1}{n} \sum_{h=1}^H \frac{n_h}{n_h-1} \sum_{j \in S_h} (Y_j - \bar{Y}_{S_h})^2,$$

where  $S_h$  denotes the sample in cell  $h$  and  $n_h$  the cell sample size.

For the purpose of constructing  $\hat{V}_{NP}$ , we consider the following model with location (**LOC**) as bivariate auxiliary variables:

$$Y_j = m(\mathbf{LOC}_j) + \varepsilon_j. \quad (14)$$

Because the homoscedastic version of the nonparametric estimator appeared to behave at least as well as the more complicated estimator that captures heteroscedasticity, we will assume here that the errors are independent with homogeneous variance. Full results for other model specifications are shown in Opsomer et al. (2009).

Under model (14), we implemented the nonparametric variance estimator  $\hat{V}_{NP}^{ho}$  given in (12) and (13) with  $x_j$  replaced by  $\mathbf{LOC}_j$ . Here  $m(\cdot)$  is estimated by bivariate local linear regression, and the estimator  $\hat{m}(\cdot)$  is obtained using `loess()` in R. In `loess()`, the bandwidth parameter  $h$  is replaced by the *span*, the fraction of the sample observations that have non-zero weight in the computation of  $\hat{m}(\mathbf{LOC}_j)$ . Since the samples points are approximated equally spaced ( $5 \times 5$  km grid), using `loess()` will produce similar results to those obtained using a fixed bandwidth in the interior of the estimation region. In order to evaluate the sensitivity of the results to the choice of the smoothing parameters, we choose three spans: 0.1, 0.2 and 0.5. After obtaining  $\hat{m}(\cdot)$ , we can calculate the nonparametric variance estimator  $\hat{V}_{NP}^{ho}$  for each response variable. Table 1 presents the sample means and the estimated variances using  $\hat{V}_{SRS}$ ,  $\hat{V}_{ST}$  and  $\hat{V}_{NP}^{ho}$ .

While we do not know the true variance, the estimator  $\hat{V}_{ST}$  is likely to be a reasonable approximation as long as the  $Y_j$  can be modeled as a spatial trend plus random errors. The naive estimator  $\hat{V}_{SRS}$  produces the largest

	$\hat{V}_{NP0.5}^{ho}$	$\hat{V}_{NP0.2}^{ho}$	$\hat{V}_{NP0.1}^{ho}$	$\hat{V}_{NP(0.1,0.3)}^{ho}$
BIOMASS	0.36	0.34	0.33	0.34
CRCOV	0.59	0.55	0.53	0.55
BA	3.11	2.96	2.78	2.87
NVOLTOT	1487	1417	1342	1396
FOREST (%)	1.92	1.77	1.65	1.71

**Table 2.** Variance estimates for five response variables for FIA data, using nonparametric estimator for additive model (15) with same span used for both variables (span = 0.5, 0.2 and 0.1), and span 0.1 for location and 0.3 for elevation.

values among the five variance estimators for all response variables and so is likely to be biased upwards for this survey. In contrast, the nonparametric variance estimator  $\hat{V}_{NP}^{ho}$  results in estimates that are close to those of  $\hat{V}_{ST}$ , with smaller spans leading to slightly smaller estimates.

Next, we consider more sophisticated models that also includes elevation (ELEV) in additive to **LOC**:

$$Y_j = m_1(\mathbf{LOC}_j) + m_2(ELEV_j) + \varepsilon_j. \quad (15)$$

We fit model (15) in R using the Generalized Additive Models (gam) package. We use the same span for both **LOC** and ELEV, as well as span = 0.1 for location and span = 0.3 for elevation. Table 2 shows that, relative to the simpler model without elevation, the estimated variances all decreased, by 8-14%. This decrease is due primarily to a reduction in the  $\hat{\sigma}_S^2$  component in (12), which accounts for the fact that the extended mean model in (15) captures more of the observed behavior of these forestry variables.

## Acknowledgments

This work was partially supported by MEC Grant MTM2008-00166 (ERDF included) and by Xunta de Galicia Grant PGIDIT07PXIB105259PR.

## References

- BARTOLUCCI, F. and MONTANARI, G. E. (2006): A new class of unbiased estimators for the variance of the systematic sample mean. *Journal of Statistical Planning and Inference* 136, 1512-1525.
- LI, X. (2006): *Application of Nonparametric Regression in Survey Statistics*. Ph. D. thesis, Department of Statistics, Iowa State University.
- OPSOMER, J., FRANCISCO-FERNANDEZ, M. AND LI, X. (2009): Additional results for model-based nonparametric variance estimation for systematic sampling in a forestry survey. Technical report, Department of Statistics, Colorado State University.
- WOLTER, K. M. (2007): *Introduction to Variance Estimation (2 ed.)*. Springer-Verlag Inc., New York.

# Thresholding-Wavelet-Based Functional Estimation of Spatiotemporal Strong-Dependence in the Spectral Domain

María Pilar Frías<sup>1</sup> and María Dolores Ruiz-Medina<sup>2</sup>

<sup>1</sup> University of Jaén  
Campus Las Lagunillas  
23071 Jaén, Spain  
(e-mail: [mpfrias@ujaen.es](mailto:mpfrias@ujaen.es))

<sup>2</sup> University of Granada  
Campus Fuente Nueva  
18071 Granada, Spain  
(e-mail: [mrui@ugr.es](mailto:mrui@ugr.es))

**Abstract.** Four functional parameter estimation algorithms are proposed for the statistical analysis of temporal and spatial long-range dependence models. Specifically, the class of strong-dependence spatiotemporal random fields studied in Frías et al. (2006a, 2008, 2009) is considered. The functional sample information is assumed to be collected in the spectral domain, and affected by additive measurement noise. In the estimation methodology proposed, a wavelet analysis of the spectral functional data is first performed. Compactly supported wavelet functions are considered in this analysis. Thresholding techniques are applied for removing the observation noise. The parameter estimators are then computed by applying linear regression in the log-thresholding wavelet domain. The performance of the estimation algorithms proposed is illustrated from simulated data.

**Keywords:** fractal spectral processes, long-range dependence parameters, spatiotemporal parametric models, wavelet thresholded transform

## 1 Introduction

Strong dependence constitutes a key feature in the analysis of complex systems, where large dimensional data sets are available. Since the classical spectral projection methods can not be applied in this context, several efforts have been made for the definition of suitable models (see, Frías et al. (2006a, 2006b, 2007, 2008, 2009), among others). Most of these models display global or local self-similarity which can be related with fractality (see, Kelbert et al. (2005)).

This paper deals with the problem of parameter estimation of strong-dependence spatiotemporal stationary random fields in the spectral domain. The singular character at the origin of the spectral density of these random fields motivates us to consider the wavelet domain for the analysis of such

a family of spectral densities. In several applications, for example, in Functional Magnetic Resonance Imaging (fMRI) the data are collected in the spectral domain by a suitable measurement transforming device (see, for example, Friston, 2007 and Lazar, 2008). This device introduces a measurement noise, which here is incorporated, in the functional observation model, as an additive observation spectral noise. This is the reason why we consider the application of thresholding techniques to removing the observation noise in the spectral functional data. The parameter estimation algorithms formulated in this paper are then based on the thresholded-wavelet transform of functional spectral information.

## 2 Preliminaries

The strong-dependence spatiotemporal model considered in this paper is given by

$$X(t, \mathbf{z}) = \int_{m.s} \int_{\mathbb{R}^{d+1}} r(t-s, \mathbf{z}-\mathbf{y}) Y(s, \mathbf{y}) ds d\mathbf{y}, \quad (1)$$

where

$$r(t, \mathbf{z}) = |t|^{-1+\nu} \prod_{i=1}^d |z_i|^{-1+\beta_i}, \quad (2)$$

with  $(\nu, \beta_1, \dots, \beta_d) \in (0, 1/2)^{d+1}$ ,  $t \in \mathbb{R}$ ,  $\mathbf{z} \in \mathbb{R}^d$ . Depending on the local regularity and the moment conditions satisfied by the sample-paths of the input random field  $Y$ , that ensure the integral (1) exists, random field  $X$  can be defined in the strong sense (i.e. pointwise) or in the weak sense (i.e. in terms of test functions, see Ruiz-Medina et al. (2003)). In this paper, we assume that  $Y$  satisfies the conditions needed for the pointwise definition of  $X$ . Specifically, the following conditions are assumed on the spectral density  $f_Y$  of  $Y$  (see, Leonenko(1999), Adler, (1981)).

**Condition 1.**  $|f_Y(\omega, \boldsymbol{\lambda})| \rightarrow C_1$ , when  $\omega \rightarrow 0$  and  $\lambda_i \rightarrow 0$ , for  $i = 1, \dots, d$ , with  $C_1$  being a positive constant.

**Condition 2.**  $\frac{|f_Y(\omega, \boldsymbol{\lambda})|}{(1+|(\omega, \boldsymbol{\lambda})|^2)^{-\tilde{\nu}-\sum_{i=1}^d \tilde{\beta}_i}} \rightarrow C_2$ , when  $\omega \rightarrow \infty$  and  $\lambda_i \rightarrow \infty$ , for  $i = 1, \dots, d$ , where  $C_2$  is a positive constant, and  $(\tilde{\nu}, \tilde{\beta}_1, \dots, \tilde{\beta}_d) \in (1/2, \infty)^{d+1}$ , and  $(\nu, \beta_1, \dots, \beta_d) \in (0, 1/2)^{d+1}$ .

*Remark 3.* Note that Condition 1 means that the integrability order of the spectral density of the spatiotemporal process  $X$  at zero frequency depends only of the behavior of the Fourier transform  $\hat{r}(\omega, \boldsymbol{\lambda}) = |\omega|^{-\nu} \prod_{i=1}^d |\lambda_i|^{-\beta_i}$  of kernel  $r$ , at a neighborhood of zero-frequency. This behavior is characterized in terms of the range of the parameter vector  $(\nu, \beta_1, \dots, \beta_d)$  (which determines the integrability of the spectral density  $f_X$  at the origin). Specifically, the integrability at zero of  $f_X$  holds for  $(\nu, \beta_1, \dots, \beta_d) \in (0, 1/2)^{d+1}$ . On the other



hand, the asymptotic order at infinity of the spectral density  $f_X$  of  $X$  depends on the considered ranges for vectors  $(\nu, \beta_1, \dots, \beta_d)$  and  $(\tilde{\nu}, \tilde{\beta}_1, \dots, \tilde{\beta}_d)$ . In this case, for  $\tilde{\nu} > (1/2) - \nu$  and  $\tilde{\beta}_i > (1/2) - \beta_i$ , for  $i = 1, \dots, d$ ,  $f_X$  is absolutely integrable at infinity. Consequently,  $X$  can be defined as an ordinary second-order random field. Moreover, from Ruiz-Medina et al. (2003) (see also Leonenko (1999) on random fields with singular spectra), random field  $X$  is continuous in the mean-square sense for the established ranges for  $(\nu, \beta_1, \dots, \beta_d)$  and  $(\tilde{\nu}, \tilde{\beta}_1, \dots, \tilde{\beta}_d)$ . Thus, from Adler (1981) results,  $X$  has continuous sample-paths, and it can be defined pointwise in the sample-path sense, in the Gaussian case.

### 3 Results and Methodology

In the implementation of the functional estimation algorithms described in this section, we consider the following functional spectral observation model:

$$Z(\omega, \boldsymbol{\lambda}) = X(\omega, \boldsymbol{\lambda}) + \sigma N(\omega, \boldsymbol{\lambda}), \quad (3)$$

where  $X$  is a Gaussian process satisfying equation (1), in terms of the input random field  $Y$ . Process  $N$  is a Gaussian spatiotemporal white noise with covariance function defined as  $E[N(\omega_1, \boldsymbol{\lambda}_1)N(\omega_2, \boldsymbol{\lambda}_2)] = \delta(\omega_1 - \omega_2)\delta(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)$ , with  $\delta$  representing the Dirac Delta distribution in the frequency domain. Parameter  $\sigma$  introduces the local variability due to the observation noise, in the functional spectral sample model. This noise is assumed to be uncorrelated with the output random field  $X$ .

The spectral density  $f_X$  of our spatiotemporal process  $X$  of interest presents the following asymptotic local fractal behavior, when  $|\omega| \rightarrow 0$ , and  $|\lambda_i| \rightarrow 0$ ,  $i = 1, \dots, d$ , (see, Leonenko (1999)),

$$f_X(\omega, \boldsymbol{\lambda}) \sim C_1 |\omega|^{-2\nu} \prod_{i=1}^d |\lambda_i|^{-2\beta_i}, \quad \nu \in (0, 1/2), \quad \beta_i \in (0, 1/2), \quad i = 1, \dots, d. \quad (4)$$

In the implementation of the functional estimation algorithms formulated below, compactly supported orthonormal wavelet bases are considered. The multiresolution analysis of  $L^2(\mathbb{R}^{d+1})$  is performed in terms of the tensorial product of  $d + 1$  one-dimensional orthonormal wavelet bases (see, Meyer (1992)). The one-dimensional wavelet transforms are defined in terms of a scaling basis  $\{\phi_k : k \in \Gamma_0 \subset \mathbb{Z}\}$  of a coarsest scale space  $V_0$ , and a sequence of wavelet bases  $\{\psi_{j:k} : k \in \Lambda_j \subset \mathbb{Z}, j \geq 0\}$  of the detail space sequence  $\{W_j, j \geq 0\}$ . Suitable examples of such bases can be constructed from the tensorial product of Haar and Daubechies systems. To remove the local variability reflected in parameter  $\sigma$ , due to the measurement noise, we apply thresholding techniques, eliminating the noise levels in the wavelet transform of the sample spectral curves.

The local asymptotic behavior (4) combined with the asymptotic, in the scale, properties of the variance of the wavelet coefficients lead to the following

identities for the temporal and spatial directional log-wavelet coefficients of the spectral density  $f_X$  of  $X$  (see, Frías et al. (2008) and Frías and Ruiz-Medina (2010)). That is,

$$\begin{aligned} f_{j:k}^1 &= \int_{\epsilon(\tilde{\omega})} f_X(\omega, \boldsymbol{\lambda}^0) \psi_{j:k}(\omega) d\omega \sim 2^{-j(-2\nu+1)} C(\psi, \boldsymbol{\lambda}^0), \quad |\tilde{\omega}| \rightarrow 0, \\ f_{j:k}^{1+i} &= \int_{\epsilon(\tilde{\lambda}_i)} f_X(\omega^0, \lambda_1^0, \dots, \lambda_i, \dots, \lambda_d^0) \psi_{j:k}(\lambda_i) d\lambda_i \\ &\sim 2^{-j(-2\beta_i+1)} C(\psi, \omega^0, \dots, \lambda_{i-1}^0, \lambda_{i+1}^0, \dots, \lambda_d^0), \quad |\tilde{\lambda}_i| \rightarrow 0, \end{aligned}$$

for  $\boldsymbol{\lambda}^0$  and  $(\omega^0, \dots, \lambda_{i-1}^0, \lambda_{i+1}^0, \dots, \lambda_d^0)$  fixed frequency values in a neighborhood of the zero frequency, and for  $\epsilon(\tilde{\omega})$  and  $\epsilon(\tilde{\lambda}_i)$  being neighborhoods of the temporal frequency  $\tilde{\omega}$ , with  $|\tilde{\omega}| \rightarrow 0$ , and of frequency  $\tilde{\lambda}_i$ , with  $|\tilde{\lambda}_i| \rightarrow 0$ ,  $i = 1, \dots, d$ , respectively. Here,  $C(\psi, \boldsymbol{\lambda}^0)$  and  $C(\psi, \omega^0, \dots, \lambda_{i-1}^0, \lambda_{i+1}^0, \dots, \lambda_d^0)$  represent constants depending on the wavelet basis chosen and on the fixed frequency values.

Therefore,

$$\log_2 f_{j:k}^1 \sim [-j(-2\nu+1)] + \log_2 C(\psi, \boldsymbol{\lambda}^0), \quad (5)$$

$$\log_2 f_{j:k}^{1+i} \sim [-j(-2\beta_i+1)] + \log_2 C(\psi, \omega^0, \dots, \lambda_{i-1}^0, \lambda_{i+1}^0, \dots, \lambda_d^0), \quad (6)$$

for  $i = 1, \dots, d$ . The temporal memory parameter  $\nu$  and spatial dependence parameters  $\beta_i$ ,  $i = 1, \dots, d$ , can then be estimated from the above expressions applying linear regression, after selecting the signal wavelet coefficients by applying thresholding rules. Specifically, the following estimators are derived

$$\hat{\nu} = \frac{-\hat{\theta}^1 + 1}{2}, \quad \hat{\beta}_i = \frac{-\hat{\theta}^{i+1} + 1}{2}, \quad i = 1, \dots, d,$$

where  $\hat{\theta}^1$  and  $\hat{\theta}^{i+1}$ ,  $i = 1, \dots, d$  are the least-squares estimates of the slope in equations (5) and (6), respectively (see, Frías and Ruiz-Medina (2010)).

The following fourth functional estimation algorithms are implemented in the simulation study developed in the next section:

*Estimation Algorithm 1:*

- Step 1: Select the zero frequency neighborhood sequences at the temporal and spatial directions.
- Step 2: At each element of the zero frequency neighborhood sequences considered in the previous step, average the temporal and spatial spectral curves, obtained by evaluation of the functional periodogram on lines parallel to the principal temporal and spatial axes.
- Step 3: Apply the one-dimensional wavelet transform to each element of the averaged temporal and spatial spectral curve sequence obtained in Step 2.

- Step 4: Universal wavelet threshold is considered for removing noise in the wavelet coefficients computed in Step 3.
- Step 5: At each zero frequency neighborhood, for each of the  $d+1$  directions considered, compute from equations (5) and (6), applying linear regression, estimates  $\hat{\nu}$  and  $\hat{\beta}_i$ ,  $i = 1, \dots, d$ , of the temporal and spatial long-range dependence parameters.
- Step 6: The arithmetic mean, for each one of the  $d+1$  parameter estimate sequences derived in the previous step, is obtained.

*Estimation Algorithm 2:*

- Step 1: Select the zero frequency neighborhood sequences at the temporal and spatial directions.
- Step 2: At each element of the zero frequency neighborhood sequences, apply the one-dimensional wavelet transform to the sample spectral curves, obtained by evaluation of the functional periodogram on lines parallel to the principal temporal and spatial axes.
- Step 3: Universal wavelet threshold is considered for removing noise in the wavelet coefficients computed in Step 2.
- Step 4: At each element of the zero frequency neighborhood sequences, compute from equations (5) and (6), applying linear regression, estimates of the temporal and spatial long-range dependence parameters, from the one-dimensional thresholded wavelet transforms calculated in the previous step.
- Step 5: At each element of the zero frequency neighborhood sequences, average the temporal and spatial long-range dependence parameter estimates obtained in Step 4.
- Step 6: Compute the arithmetic mean for each of the  $d+1$  temporal and spatial long-range dependence parameter estimate sequences obtained in the previous step.

Algorithm 3, for spatiotemporal long-range dependence parameter estimation, is implemented as Algorithm 2, but in terms of a smoothing version of the wavelet transform, with respect to the translation parameter at each multiresolution level. Finally, in Algorithm 4, the wavelet transform is applied to the averaged temporal and spatial spectral curves at each zero-frequency neighborhood, as in Algorithm 1, and also, smoothing is performed over the translation parameter of the wavelet coefficients at each scale. In all the cases, universal wavelet threshold is considered, for removing the observation noise from the functional spectral data (see, for example, Vidakovic, 1999).

## 4 Simulations

The implementation of the functional estimation methodologies proposed in the previous section is now illustrated, considering several scenarios within

the spatiotemporal model class subsequently described. Spatiotemporal process  $X$  is defined in two ways: as a Gaussian stationary process with spectral density given by

$$f_{X_1}(\omega, \lambda_1, \lambda_2) = \left[ \frac{1}{(1 + |\omega|^2)^{\frac{\alpha_1}{2}}} \right] \left[ \frac{1}{(1 + |\lambda_1|^2)^{\frac{\alpha_2}{2}}} \right] \left[ \frac{1}{(1 + |\lambda_2|^2)^{\frac{\alpha_3}{2}}} \right] \\ \times |\omega|^{-2\nu} |\lambda_1|^{-2\beta_1} |\lambda_2|^{-2\beta_2}, \quad (7)$$

with  $\alpha_i \in (0, 1)$ ,  $i=1,2,3$ , and as a Gaussian stationary process with spectral density given by

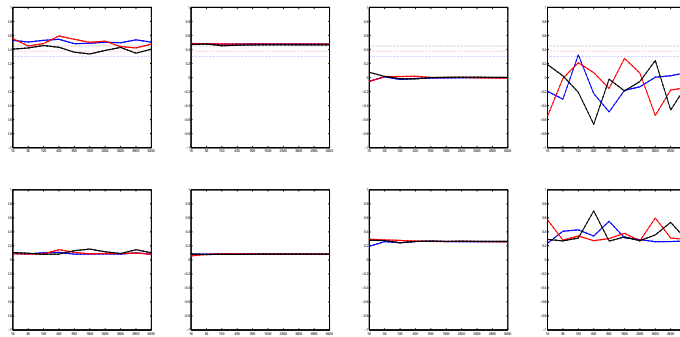
$$f_{X_2}(\omega, \lambda_1, \lambda_2) = \left[ \frac{1}{1 + |\omega|^{\alpha_1}} \right] \left[ \frac{1}{1 + |\lambda_1|^{\alpha_2}} \right] \left[ \frac{1}{1 + |\lambda_2|^{\alpha_3}} \right] |\omega|^{-2\nu} |\lambda_1|^{-2\beta_1} |\lambda_2|^{-2\beta_2}, \quad (8)$$

with  $\alpha_i \in (0, 1)$ ,  $i = 1, 2, 3$ . Note that, within the range given by the interval  $(0, 1)$ , for the parameters  $\alpha_i$ , for  $i = 1, 2, 3$ , the spectral density of the input  $Y$  displays the asymptotic behavior given in Condition 2. Functional spectral data are constructed from  $256 \times 256 \times 256$  frequency points belonging to the interval  $[-127.5 \times 10^{-8}, 127.5 \times 10^{-8}]$ , that is,  $(\omega, \lambda_1, \lambda_2) \in [-127.5 \times 10^{-8}, 127.5 \times 10^{-8}]^3$ , with discretization step size  $10^{-8}$ . The simulation study is developed considering the following two structural parameter scenarios, corresponding to two extreme cases in the two-above introduced spectral models, heavy and slight spectral singularity, in the range of strong dependence. (The parameter value  $\sigma_{\varepsilon_2} = 2 \times 10^2$  is considered as observation noise intensity in such cases, to obtain a reasonable signal to noise ratio, according to the truncated spectral density values on a zero-frequency spectral neighborhood).

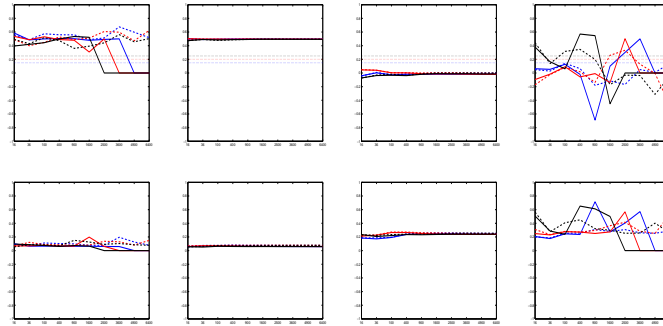
Case I:  $\nu = 0.3$ ,  $\beta_1 = 0.375$ ,  $\beta_2 = 0.45$ ,  $\alpha_1 = 0.2$ ,  $\alpha_2 = 0.3$ ,  $\alpha_3 = 0.4$ ,

Case II:  $\nu = 0.15$ ,  $\beta_1 = 0.2$ ,  $\beta_2 = 0.25$ ,  $\alpha_1 = 0.2$ ,  $\alpha_2 = 0.3$ ,  $\alpha_3 = 0.4$ .

After implementation of the fourth functional estimation algorithms proposed, in Figures 1 and 2, the three (temporal and spatial) long-range dependence parameter estimate sequences are represented, considering the above-referred models and cases. Functional estimation algorithms 1, 2, 3 and 4 are implemented from the following spectral curve sample sizes  $n = 16, 36, 100, 400, 900, 1600, 2500, 3600, 4900, 6400$ , at temporal and spatial directions. Specifically, Figures 1 and 2 show  $\hat{\nu}$ ,  $\hat{\beta}_1$ , and  $\hat{\beta}_2$  values on top and standard deviations on bottom, obtained without (dashed blue, red and black line, respectively) and with (blue, red and black line, respectively) thresholding techniques. Parameter values for  $\nu$ ,  $\beta_1$ ,  $\beta_2$  are displayed with dotted blue, red and black line, respectively. Hard thresholding-wavelet- methods are applied. In this high local singularity case, discrimination between the structural local variability and the noise local variability is needed. Specifically, the universal threshold, given by  $\sigma_{\varepsilon_2} \sqrt{2 \log n_i}$ ,  $n_i = 256$ ,  $i = 1, 2, 3$ , is considered (see, Donoho and Johnstone (1995)). The results displayed show that Algorithms 1 and 2 are more suitable for smooth (differentiable) input models like in



**Fig. 1.**  $\hat{\nu}$ ,  $\hat{\beta}_1$ , and  $\hat{\beta}_2$  values (top) and standard deviations (bottom), algorithm 1 (left), algorithm 2 (left-medium), algorithm 3 (right-medium), algorithm 4 (right), for case I and for model (7). The values on horizontal axis represent the spectral curve sample sizes considered.



**Fig. 2.**  $\hat{\nu}$ ,  $\hat{\beta}_1$ , and  $\hat{\beta}_2$  values (top) and standard deviations (bottom) algorithm 1 (left), algorithm 2 (left-medium), algorithm 3 (right-medium), algorithm 4 (right), for case II and for model (8). The values on horizontal axis represent the spectral curve sample sizes considered.

equation (7), and Algorithms 4 is most suitable for fractal input models like in (8).

## 5 Final comments

Long-range dependence is a key feature in the analysis of complex systems which can be equivalently studied, thanks to Tauberian-type theorems (see, for example, Leonenko (1999)), in terms of the local singularity level, in a neighborhood of the zero frequency, of the spectral density. This fact motivated the parameter estimation methodology proposed in this paper, in terms of compactly supported wavelet functions. The fixed-domain asymptotic properties of the designed functional parameter estimators will be de-

rived in a subsequent paper, from the weak-consistency of the functional spectral and wavelet periodograms.

**Acknowledgments.** This work has been supported in part by projects MTM2008-03903, MTM2009-13393 of the DGI, MEC, and P09-FQM-5052 of the Andalusian CICE, Spain.

## References

- ADLER, R. J. (1981): *The Geometry of Random Fields*. Wiley, London.
- DONOHU, D. L. and JOHNSTONE, I. M. (1995): Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* 90, 1200–1224.
- FRIAS, M. P. and RUIZ-MEDINA, M. D. (2010): Computing functional estimators of the long-range dependence parameters in the spectral-wavelet domain. *Journal of Statistics Planning and Inference (Submitted)*.
- FRIAS, M. P., ALONSO, F. J., RUIZ-MEDINA, M. D. and ANGULO, J. M. (2007): Semiparametric estimation of spatial long-range dependence. *Journal of Statistics Planning and Inference* 138, 1479–1495.
- FRIAS, M. P., RUIZ-MEDINA, M. D., ALONSO, F. J. and ANGULO, J. M. (2006a): Spatiotemporal generation of long-range dependence models and estimation. *Environmetrics* 17, 139–146.
- FRIAS, M. P., RUIZ-MEDINA, M. D., ALONSO, F. J. and ANGULO, J. M. (2008): Parameter estimation of self-similar spatial covariogram models. *Computation Statistics - Theory and Methods* 37, 1011–1023.
- FRIAS, M. P., RUIZ-MEDINA, M. D., ALONSO, F. J. and ANGULO, J. M. (2009): Spectral-marginal-based estimation of spatiotemporal long-range dependence. *Computation Statistics - Theory and Methods* 38, 103–114.
- FRIAS, M. P., RUIZ-MEDINA, M. D., ANGULO, J. M. and ALONSO F. J., (2006b): Semiparametric estimation of spatiotemporal anisotropic long-range dependence. In: A. Rizzi , M. Vichi (Eds.): *Proceedings in Computational Statistics (Contributed Paper)*. Physica-Verlag, Rome, 1201–1208.
- FRISTON, K. J. (2007): *Statistical parametric mapping: the analysis of functional brain images*. Academic Press, inc.
- KELBERT, M., LEONENKO, N., and RUIZ-MEDINA, M. D. (2005): Fractional Random Fields Associated with Stochastic Fractional Heat Equation. *Advances in Applied Probability* 37, 108–133.
- LAZAR, N. A. (2008): The statistical analysis of functional MRI data. *Statistics for biology and Health*. Springer.
- LEONENKO, N. (1999): *Limit Theorems for Random Fields with Singular Spectrum, Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht/Boston/London.
- MEYER, Y. (1992): *Wavelet and Operators*. Cambridge University Press, Cambridge.
- RUIZ-MEDINA, M. D., ANGULO, J. M. and ANH, V. V. (2003): Fractional generalized random fields on bounded domains. *Stochastics Analysis and Applications* 21, 465–492.
- VIDAKOVIC, B. (1999): *Statistical Modeling by Wavelets*. Wiley Series in Probability and Statistics.

# Boolean Factor Analysis by the Expectation-Maximization Algorithm

Alexander. A. Frolov<sup>1</sup>, Pavel. Y. Polyakov<sup>2</sup>, and Dusan Húsek<sup>3</sup>

<sup>1</sup> Institute of Higher Nervous Activity and Neurophysiology of the Russian Academy of Sciences, Butlerova 5a, Moscow, Russia *aafrolov@mail.ru*

<sup>2</sup> Scientific-Research Institute for System Studies of the Russian Academy of Science, Nakhim. prosp. 36/1, 117 218 Moscow, Russia *pavel.mipt@mail.ru*

<sup>3</sup> Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodarenskou vezi 2, 182 07 Prague, Czech Republic *dusan@cs.cas.cz*

**Abstract.** Compared are efficiencies of two methods for Boolean factor analysis based on expectation-maximization technique. First one is Maximal Causes Analysis proposed by Lücke and Sahani (2008). Second one is Expectation-Maximization Boolean Factor Analysis, introduced here. Last method is strictly based on the general Boolean factor analysis generative model. Comparison is based on so called bars problem benchmark (Földiák, 1990). Further informational theoretic measure of Boolean factor analysis efficiency is developed. Then it is shown that the efficiency of our Expectation-Maximization Boolean Factor Analysis method is higher then Maximal Causes Analysis in Boolean factor analysis model parameters entirety.

**Keywords:** boolean factor analysis, generative model, information gain, efficiency measure

## 1 Introduction

In Boolean factor analysis (BFA) it is supposed that components of original signals, factor loadings and factor scores are binary values. Each binary signal can be interpreted as a pattern describing the states of categorical attributes. The whole number of considered attributes is the dimensionality of the signal space, the state of each attribute is 1 or 0 depending on its presence or absence in the pattern. Thus, every pattern of the data set is a binary vector  $\mathbf{x}$  whose dimensionality  $N$  is equal to the total number of attributes in the current context. Every component of  $\mathbf{x}$  takes value 1 or 0 depending on the appearance of the corresponding attribute in the pattern. Each factor  $\mathbf{f}_i$ ,  $i = 1, \dots, L$  ( $L$  is the total number of factors) is a binary vector of dimensionality  $N$  in which the entries with value 1 correspond to highly correlated attributes appearing in the data set when the related factor is present in patterns of the data set. Although the probability of a factor's attribute to appear in a pattern with the factor is high, it is not obligatory equal to 1. Sometimes attribute could vanish. We take into account this property of objects by introducing the probabilities  $p_{ij}$ ,  $i = 1, \dots, L$ ,  $j = 1, \dots, N$  which are

assumed to be high for attributes constituting the factor, and for the other attributes we put them to be 0. As in linear factor analysis, we suppose that additionally to common factors  $\mathbf{f}_i$  every signal contains also specific factors. The contribution of specific factors is defined by binary vector  $\eta$ , which we call “specific noise”. Each specific factor is characterized by probabilities  $q_j$  that the  $j$ -th component of vector  $\eta$  takes 1. As a result, in general vector  $\mathbf{x}$  can be presented in the form

$$\mathbf{x} = \left[ \bigvee_{i=1}^L s_i \mathbf{f}'_i \right] \vee \eta, \quad (1)$$

where  $\mathbf{s}$  is a vector of factor scores of dimensionality  $L$ ,  $\mathbf{f}'_i$  is a corrupted version of the  $i$ -th factor and  $\eta$  is a specific noise. Factor corruption implies that some Ones of the  $i$ -th factor become Zeros with probability  $1 - p_{ij}$ . We suppose that each component of the common factor is corrupted independently of the presence of other factors in the pattern, and independently of specific noise. We also assume that factors appear in patterns (i.e., corresponding scores entries  $s_i$  take value 1) independently with probabilities  $\pi_i$ ,  $i = 1, \dots, L$ .

BFA is performed on the set  $\mathcal{X}$  of patterns  $\mathbf{x}^{(m)}$  containing  $M$  different representatives. The aim of Boolean factor analysis is to find the parameters of generative model  $\Theta = (p_{ij}, q_j, \pi_i, i = 1, \dots, L, j = 1, \dots, N)$  and factor scores  $s_i^{(m)}$ ,  $m = 1, \dots, M$  in each of the  $M$  patterns of the data set. However, it is supposed that found factors could be also detected in any arbitrary pattern outside  $\mathcal{X}$ , if the pattern is subjected to the same generative model.

Recently Lücke and Sahani (2008) proposed the method for non-linear component extraction based on expectation-maximization algorithm (EM, Dempster et al., 1977) which was called the Maximal Causes Analysis (MCA<sub>3</sub>). In MCA<sub>3</sub> generative model multiple active hidden causes (factors in terms of BFA) combine to determine the values of an observed variable through a max function. Each cause results in a set of followings given by a vector of generative influences (factor loadings in terms of BFA). If several causes result in the same following, then the strongest influence alone determines the value of the observed variable. If all influences have the same value, then a max function is equivalent to Boolean summation of the influences, and the generative model becomes equivalent to the above generative model of BFA.

The main goal of the present study is to apply the powerful EM-technique directly to the BFA generative model given by (1) and to compare its efficiency with MCA<sub>3</sub>. As a measure of methods efficiency we use the information gain, that is difference between the data set entropies when its hidden factor structure is unknown and when it is revealed by BFA.

As a benchmark test for comparison of different BFA methods we used the bars problem (Földiák, 1990)). In standard bars problem, each pattern of the data set is  $n$ -by- $n$  binary pixel image containing several of  $L = 2n$  possible (one-pixel wide) horizontal and vertical bars. Pixels belonging and



not belonging to the bar take value 1 and 0 respectively. For each image each bar could be chosen with a probability  $C/L$ , where  $C$  is the mean number of bars mixed in an image. In the point of intersection of vertical and horizontal bars pixel takes the value 1. The goal of the task is to recognize all bars as individual objects on the basis of a data set containing  $M$  complex images of bars mixture. In the most papers where the bars problem was tested  $C$  amounted to 2 and  $n \leq 8$ . In terms of BFA, bars are factors, each image is Boolean superposition of factors, and factors scores take values 1 or 0 dependently on bars presence in the image. Thus, bars problem is only a special case of BFA.

## 2 Information gain

If the factor structure of the signal space is unknown, then storing the  $j$ -th component of vector  $\mathbf{x}$  requires  $h(p_j)$  bits of information, where  $h(x) = -x \log_2 x - (1-x) \log_2 (1-x)$  is Shannon function and  $p_j$  is a probability of the  $j$ -th component to take 1. Storing  $M$  vectors  $\mathbf{x}$  requires

$$H_0 = M \sum_{j=1}^N h(p_j)$$

bits of information. If the hidden factor structure of the signal space is revealed and all factor scores and generative model parameters are found, then storing the  $j$ -th component of vector  $\mathbf{x}^{(m)}$  requires  $h_j^{(m)} = h(\mathcal{P}(x_j^{(m)}))$  bits of information where

$$\mathcal{P}(x_j^{(m)}) = x_j^{(m)} - (2x_j^{(m)} - 1)(1 - q_j) \prod_{i=1}^L (1 - p_{ij})^{s_i^{(m)}} \quad (2)$$

and scores  $s_i^{(m)}$  are assumed to be given. Storing the whole data set requires

$$H = M \sum_{i=1}^L h(\pi_i) + \sum_{m=1}^M \sum_{j=1}^N h_j^{(m)} \quad (3)$$

bits of information. The terms in (3) define the information that is required to store factor scores and all patterns of the data set when factor scores are given. The information gain is determined by the difference between  $H_0$  and  $H$ . We define the relative information gain as follows:

$$G = (H_0 - H)/H_0. \quad (4)$$

### 3 Expectation-maximization method

The EM method maximizes the likelihood of the observed data by maximizing the free energy

$$\mathcal{F}(\Theta, g) = \sum_{m=1}^M \sum_{\mathbf{s}} g_m(\mathbf{s}) [\log P(\mathbf{x}^{(m)} | \mathbf{s}, \Theta) + \log P(\mathbf{s} | \Theta)] + H(\mathbf{g}),$$

where  $g_m(\mathbf{s})$  is the expected distribution of factor scores for the  $m$ -th pattern and  $H(\mathbf{g}) = \sum_{\mathbf{s}} H(g_m(\mathbf{s}))$  is the Shannon entropy of  $\mathbf{g}$ . The iterations of EM alternatively increase  $\mathcal{F}$  with respect to the distributions  $g_m$  (practically factor scores), while keeping  $\Theta$  constant (the E-step), and with respect to  $\Theta$ , while keeping  $g_m$  constant (the M-step).

At the E-step, when  $\Theta$  is fixed, the distributions  $g_m$  which maximize  $\mathcal{F}(\Theta, \mathbf{g})$  are calculated from the equation

$$g_m(\mathbf{s} | \Theta) = \frac{P(\mathbf{s} | \Theta) P(\mathbf{x}^{(m)} | \mathbf{s}, \Theta)}{\sum_{\mathbf{s}} P(\mathbf{s} | \Theta) P(\mathbf{x}^{(m)} | \mathbf{s}, \Theta)},$$

where  $P(\mathbf{s} | \Theta) = \prod_{i=1, L} \pi_i^{s_i} (1 - \pi_i)^{1-s_i}$ ,  $P(\mathbf{x}^{(m)} | \mathbf{s}, \Theta) = \prod_{j=1, N} P(\mathbf{x}_j^{(m)})$  and  $P(\mathbf{x}_j^{(m)})$  is given by (2). The obtained distributions  $g_m$  provide the equality  $\mathcal{F}(\Theta, \mathbf{g}) = \mathcal{L}(\Theta)$  (Dempster et al., 1977), where  $\mathcal{L}(\Theta)$  is the likelihood of the observed data under the given parameters of the generative model.

At the M-step when distributions  $g_m$  are fixed,  $\pi_i$  can be found by  $\pi_i = (1/M) \sum_{m=1}^M s_i^{(m)}$ , where  $s_i^{(m)} = \sum_{\mathbf{s}} g_m(\mathbf{s} | \Theta) s_i$ . Respectively,  $p_{ij}$  and  $q_j$  can be found by maximization of  $\mathcal{F}(\Theta, \mathbf{g})$  according to the system of  $L \times N + N$  equations

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial p_{ij}} &= \sum_{m=1}^M \sum_{\mathbf{s}} g_m(\mathbf{s} | \Theta) P(x_j^{(m)} | \mathbf{s}, \Theta)^{-1} \frac{\partial P(x_j^{(m)} | \mathbf{s}, \Theta)}{\partial p_{ij}} = 0, \\ \frac{\partial \mathcal{F}}{\partial q_j} &= \sum_{m=1}^M \sum_{\mathbf{s}} g_m(\mathbf{s} | \Theta) P(x_j^{(m)} | \mathbf{s}, \Theta)^{-1} \frac{\partial P(x_j^{(m)} | \mathbf{s}, \Theta)}{\partial q_j} = 0 \\ p_{ij}(k+1) &= \frac{\sum_{m=1}^M \sum_{\mathbf{s}} g_m(\mathbf{s} | \Theta) s_i p_{ij}(k) x_j^{(m)} D_j^{-1}(k)}{\sum_{m=1}^M \sum_{\mathbf{s}} g_m(\mathbf{s} | \Theta) s_i} \\ q_j(k+1) &= \frac{1}{M} \sum_{m=1}^M \sum_{\mathbf{s}} g_m(\mathbf{s} | \Theta) q_j(k) x_j^{(m)} D_j^{-1}(k), \end{aligned} \tag{5}$$

where  $D_j(k) = 1 - (1 - q_j(k)) \prod_{i=1, L} (1 - p_{ij}(k))^{s_i}$ . The obtained values of  $p_{ij}$ ,  $q_j$  and  $\pi_i$  are used as an input for the next E-step. The convergence of the procedure provides the maximum of the likelihood function. When procedure

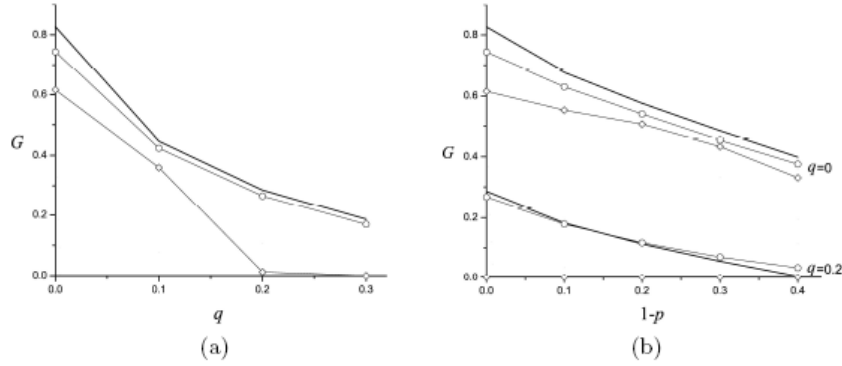
converged, the final values  $s_i^{(m)}$  are estimates of factor scores. As for MCA<sub>3</sub>, we restricted our study by the case of sparse scores when not more than three factors are supposed to be mixed in the observed patterns. Then, summation by  $\mathbf{s}$  in above formulas is reduced to

$$\sum_{\mathbf{s}} (\dots) = (\dots)_{\mathbf{s}=0} + \sum_i (\dots)_{\mathbf{s}=\mathbf{s}_i} + \sum_{i<j} (\dots)_{\mathbf{s}=\mathbf{s}_{ij}} + \sum_{i<j<k} (\dots)_{\mathbf{s}=\mathbf{s}_{ijk}}, \quad (6)$$

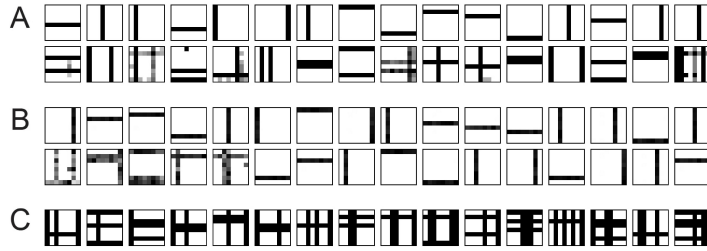
where  $\mathbf{s}_i$  is a vector of factor scores with all zeros except  $s_i$ ,  $\mathbf{s}_{ij}$  is a vector of factor scores with all zeros except  $s_i$  and  $s_j$ , and  $\mathbf{s}_{ijk}$  is a vector of factor scores with all zeros except  $s_i$ ,  $s_j$  and  $s_k$ .

#### 4 Comparison of EMBFA and MCA<sub>3</sub> in solving the bars problem

Patterns of the data set are 8-by-8 binary images, 16 vertical and horizontal bars (one pixel width) are factors, 2 bars are mixed in each image on average. For EMBFA as for MCA<sub>3</sub> the number of desired hidden factors has to be determined in advance. In the most computer experiments performed by Lücke and Sahani (2008), this number was taken twice as high than the actual number of factors. It ensured successful search of all the 16 true factors among 32 desired factors. In our experiments, we set the number of predefined hidden factors also twice as high as the number of actual factors. To finish EMBFA we used the same convergence criterion as Lücke and Sahani (2008). To start EM procedure we put  $\pi_i = 1/32$ ,  $q_j = 0$  and  $p_{ij}$  are random values uniformly distributed in the range from 0.3 to 0.8. In computer experiments performed with MCA<sub>3</sub> we used just the same parameters as recommended in the original paper (Lücke and Sahani (2008)). Figure 1(a) demonstrates the dependence of information gain  $G$  on the level of noise when the size of a data set  $M = 800$ . In this case,  $M$  is sufficiently large to provide the saturation of  $G$  when  $M$  increases. For shown examples  $p_{ij} = p$  for components constituting factors ( $f_{ij} = 1$ ) and  $q_j = q$  for any  $j$ . Note that  $p_{ij} = 0$  for components not constituting factors ( $f_{ij} = 0$ ). As shown in Figure 1(a), both EMBFA and MCA<sub>3</sub> provide an information gain smaller than that for “precise” solution when all scores and generative model parameters are found precisely. This gain decrease occurs due to the omission of scores. According to (6) the method is able to identify scores only in patterns of the data set containing not more than three mixed factors. For the used generative model, the number of mixed factors  $k$  has binomial distribution  $B(k, C/L, L)$ , where  $C = 2$ ,  $L = 16$ . In this case, about 30% of scores in the data set is expected to be missed. Information gain provided by MCA<sub>3</sub> corresponds to this estimation for the number of missed scores. However, the gain obtained by EMBFA is higher. The reason is the specificity of additional factors in EMBFA in having the form of a mixture of two bars, while for MCA<sub>3</sub> additional factors coincide with true ones (compare Fig. 2(A) and Fig. 2(B)). Due to this specificity,



**Fig. 1.** Information gain in dependence on  $q$  for  $p = 1$  (a) and on  $1 - p$  for  $q = 0$  and  $q = 0.2$  (b).  $\circ$  – EMBFA,  $\diamond$  – MCA<sub>3</sub>. Thick line – “precise” solution.



**Fig. 2.** Factors found by EMBFA (A) and by MCA<sub>3</sub> (B) in the absence of noise ( $q = 0$ ,  $p = 1$ ), C – examples of images where bars were identified by EMBFA, but not by MCA<sub>3</sub>.

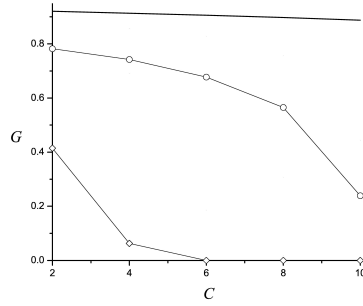
EMBFA is able to recognize a true factor even in a pattern containing the mixture of 5 bars, if this pattern is identified by the method as a mixture of 3 found factors: one of them is the actual true factor corresponding to a single bar, and the 2 others are false factors corresponding to a mixture of two bars. The examples of patterns of the data set, where true factors were recognized by EMBFA but not by MCA<sub>3</sub>, are shown in Fig. 2(C).

Particularly the first of the shown patterns is recognized as having been created by factors 4 and 7 in the upper row and by factor 13 in the lower row in Fig. 2(A). Another specificity of EMBFA is the competition between two kinds of noise. Due to this competition EMBFA is able to recognize three factors in an image even when these factors do not cover the whole image. In this case, not recognized factors are treated as specific noise. The example is the last image in Fig. 2(C), where EMBFA recognizes factor 13 in the upper row, and factors 8 and 12 in the lower row. EMBFA treats the two remaining bars as a specific noise. As a result, EMBFA missed less than 10% of scores.

Information gain obtained by  $\text{MCA}_3$  demonstrates strong sensitivity to  $q$  (Fig. 1(a)). It drops to almost zero when  $q$  increases to 0.2. In this case the solution of the bars problem by  $\text{MCA}_3$  becomes unstable, i.e. it drastically depends on the peculiarity of the data set or the choice of initial parameters required to start the iterative EM procedure. With one random realization of a data set  $\text{MCA}_3$  provides perfect solution, with another its realization chosen from the same distribution the iterative procedure converges to some random image (usually for 3-5 iteration steps as contrasted to the case of correct solution requiring about 300 steps). For  $q = 0.2$   $\text{MCA}_3$  provides successful search of bars only in 2 out of 50 trials.  $\text{MCA}_3$  is less sensitive to factors corruption than to specific noise (compare Fig. 1(a) and Fig. 1(b)). The reason is that factors corruption corresponds to  $\text{MCA}_3$  generative model. All factors are correctly found by this method and information gain is less than that for precise solution only due to omission of scores in images containing more than 3 bars, as explained above. EMBFA happened to be not sensitive to both specific noise and factors corruption (Fig. 1(a)). Information gain obtained by EMBFA is close to that for precise solution even when both kinds of noise are superimposed ( $q = 0.2$ ,  $p < 1$ , Fig. 1(b)). For small  $p$ , it is even paradoxically higher than for precise solution. This effect is explained by omission of some factor scores. EMBFA omitted factors in images where they were by chance considerably corrupted. As a result, the images that contained omitted factors were excluded from estimation of  $p_{ij}$ . Then, estimated values of  $p_{ij}$  increased, while estimated values of  $q_j$  remained almost unchanged. For example, in the case  $p = 0.7$ , estimated values of  $p_{ij}$  on average increased to 0.77, while  $q_j$  increased from 0.2 to only 0.22. The increase of  $p_{ij}$  dominates over the increase of  $q_j$  and thus gain increases.

## 5 Conclusion

$q_j$  and  $C$ . First,  $\text{MCA}_3$  is based on the generative model which does not take into account specific noise defined by  $q_j$ , but only a noise in the form of factors corruption. Second, it is restricted by the case of sparse scores and ignores signals, where more than three factors are mixed. In principle, it is easy to expand the method taking into account the signals with  $C > 3$ , but then its performance would take an incredible long time. Although EMBFA is also restricted by the case of sparse scores, it is less sensitive to  $C$  increase. It gives reasonable results even in the case when the averaged number of mixed bars in the image is six (Fig. 3). When the number of bars mixed in an image exceeds three EMBFA treats some bars in the image as specific noise. The peculiarity of EMBFA is the competition between two kinds of noise. When specific noise is considerable EMBFA prescribes additional pixels to bars, and when factors corruption is large, it deletes some pixels from bars, adjusting to the specific distribution of noise in the data set. Paradoxically due to this peculiarity information gain provided by EMBFA often exceeds



**Fig. 3.** Information gain in dependence on  $C$  for images of 16-by-16 pixels for  $M = 800$ . Results are marked as in Fig. 1

the gain given by precise solution (Fig. 1). High efficiency of EMBFA is not amazing because it is based on the powerful EM approach specially applied to BFA generative model. Thus, it has low sensitivity to both kinds of noise, while  $MCA_3$  also based on EM is very sensitive to specific noise, which is not included into its generative model. Note that the performance of EMBFA does not depend on any tuning parameters.

And finally it is interesting to compare their computational complexity. According to the formulas (5) and (6) the number of computer operations required for one iteration step of EMBFA is proportional to  $\Omega_{EMBFA} = MN(2L)^3 \langle p_j \rangle$ . For PC Core2 6400, 2.13 GHz execution time for one EMBFA step amounts to  $5 \cdot 10^{-9} \Omega_{EMBFA}$ . For  $MCA_3$  the dependence of computational time for one iteration step on data set parameters is the same but the proportionality coefficient for execution time is twice as high. For EMBFA the mean number of iteration steps required for its convergence amounts to about 100. Thus, to evaluate the execution time one must again multiply the proportionality coefficient by a factor of 100. For  $MCA_3$  it amounts to about 300. Thus, the total execution time for  $MCA_3$  is six as high as for EMBFA. For the bars problem  $n_f = \sqrt{N}$ ,  $L = 2\sqrt{N}$  and in the absence of noise  $\langle p_j \rangle \simeq C/\sqrt{N}$ . On the whole to perform BFA on a data set of 800 images of 16-by-16 pixels containing two not corrupted bars, takes about 9000 sec for  $MCA_3$  and 1500 sec for EMBFA.

**Acknowledgement.** This work was supported by projects AV0Z10300504, 1M10567, GACR P202/10/0262 and GACR 205/09/1079.

## References

- DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological* 39(1), 1–38.
- FÖLDIAK, P. (1990): Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics* 64, 165–170.
- LÜCKE, J. and SAHNI, M. (2008): Maximal causes for non-linear component extraction. *The Journal of Machine Learning Research* 9, 1227–1267.

# Modeling and Forecasting Electricity Prices and their Volatilities by Conditionally Heteroskedastic Seasonal Dynamic Factor Analysis

Carolina García-Martos<sup>1</sup>, Julio Rodríguez<sup>2</sup> and María Jesús Sánchez<sup>3</sup>

<sup>1</sup> Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, Spain, *garcia.martos@upm.es*

<sup>2</sup> Facultad de Ciencias Económicas y Empresariales, Universidad Autónoma de Madrid, Spain, *jr.puerta@uam.es*

<sup>3</sup> Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, Spain, *mjsan@etsii.upm.es*

**Abstract.** In this work we propose a new model, that allows to extract conditionally heteroskedastic common factors from a vector of series. These common factors, their relationship with the original vector of series, as well as the dynamics affecting both their conditional mean and variance are jointly estimated. Considering that ARCH and GARCH effects can be handled under the state-space formulation, the estimation of the model is carried out in this way. The new model proposed is applied to extract seasonal common dynamic factors as well as common volatility factors for electricity prices. Then, the estimation results are used to forecast electricity prices and their volatilities in the Spanish Market.

**Keywords:** forecasting, dimensionality reduction, electricity prices, conditional heteroskedasticity

## 1 Introduction

Electricity markets are liberalized in most developed countries for more than a decade. For every hour in a day, producers and users submit their hourly bids to the market operator. Thus, a 24-dimensional vector of prices is generated daily. There is a need of developing specific models that are able to forecast electricity prices. The high-dimensional vector of series of electricity prices present structure both in the conditional mean and conditional variance.

Concerning previous works, there are many authors that have developed methodology for modelling and forecasting electricity prices, and most of them deal with one-day-ahead forecasting (Conejo et al., 2005), useful for scheduling power generation units. But the problem of reducing the risk that every bilateral contract imply, is not usually faced. This can be done by long-term forecasting, covering at least the length of the contract, usually one year. Concerning joint modelling of conditional mean and variance, Koopman et al.

(2007) provided novel periodic extensions of dynamic long-memory regression models with autoregressive conditional heteroskedastic errors for the analysis of daily electricity spot prices in several European deregulated markets. Focusing on factor models, well known references are Stock and Watson (2002), Peña and Box (1987), Lee and Carter (1992) and Peña and Poncela (2004, 2006), who extended the Peña-Box model to the Non-Stationary case. Alonso et al. (2008) provided the Seasonal Dynamic Factor Analysis.

In this work we allow unobserved common factors extracted from the 24 hourly time series of prices to be conditionally heteroskedastic, following a seasonal VARIMA plus ARCH or GARCH processes, so not only the common structure in mean is extracted, but also the common volatility factors. The Seasonal Dynamic Factor Analysis proposed by Alonso et al. (2008), and the works by Diebold and Nerlove (1989) and Harvey et al. (1992) for dealing with conditionally heteroskedastic disturbances in state space models are applied and extended.

## 2 Formulation of the Model

### 2.1 Seasonal Dynamic Factor Analysis with Homoskedastic disturbances

Alonso et al. (2008) developed the Seasonal Dynamic Factor Analysis (SeaDFA), which is able to extract a  $r$ -dimensional vector of unobserved seasonal common factors from a  $m$ -dimensional observed vector of time series (where  $r < m$ ). They assume that vector  $\mathbf{y}_t$  can be written as a linear combination of the unobserved common factors,  $\mathbf{f}_t$ , plus  $\varepsilon_t$ , to which we will refer from now on as specific components or specific factors:

$$\mathbf{y}_t = \Omega_1 \mathbf{f}_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, S), \quad (1)$$

where  $NID$  means Normally and Identically Distributed. Besides, these common factors are allowed to follow a multiplicative seasonal Vector Autoregressive Integrated Moving Average, VARIMA( $p, d, q$ )  $\times$  ( $P, D, Q$ )<sub>s</sub> model:

$$(\mathbf{I} - B)^d (\mathbf{I} - B^s)^D \phi(B) \Phi(B^s) \mathbf{f}_t = \mathbf{c}_1 + \theta(B) \Theta(B^s) \mathbf{w}_t^1, \quad (2)$$

where  $\mathbf{c}_1$  is the  $r$ -dimensional constant of the model of the common factors,  $\phi(B) = (\mathbf{I} - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)$ ,  $\Phi(B^s) = (\mathbf{I} - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_p B^{Ps})$ ,  $\theta(B) = (\mathbf{I} - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q)$  and  $\Theta(B^s) = (\mathbf{I} - \Theta_1 B^s - \Theta_2 B^{2s} - \dots - \Theta_Q B^{Qs})$  are polynomial matrices  $r \times r$ ,  $B$  is the backshift operator such that  $B\mathbf{y}_t = \mathbf{y}_{t-1}$ , the zeros of the determinants  $|\phi(B)|$  and  $|\Phi(B^s)|$  are on or outside the unit circle, and the roots of  $|\theta(B)| = 0$  and  $|\Theta(B^s)| = 0$  are outside the unit circle and  $\mathbf{w}_t^1 \sim \mathbf{N}_r(\mathbf{0}, \mathbf{Q}_1)$  is serially uncorrelated for all leads and lags, where  $\mathbf{N}_r$  is an  $r$ -dimensional multivariate Gaussian distribution. The homoskedastic SeaDFA is given by equations (1) and (2), and



can be rewritten under the state-space formulation, just reformulating them as an observation and transition equation, (3) and (4), respectively:

$$\mathbf{y}_t = \Omega \mathbf{F}_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, S), \quad (3)$$

$$\mathbf{F}_t = \mathbf{c} + \Psi \mathbf{F}_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim NID(0, \mathbf{Q}). \quad (4)$$

Concerning the evolution of the common factors over time, in general, following Ansley and Kohn (1986), any multiplicative seasonal VARIMA  $(p, d, q) \times (P, D, Q)_s$  model as given by equation (2), can be easily rewritten as a transition equation like (4). Moreover, the model is unidentified under rotations, but this problem is solved imposing restrictions like  $\mathbf{Q}_1 = \mathbf{I}$  or  $\Omega_1' \Omega_1 = \mathbf{I}$ , as well as  $\omega_{ij} = 0$ , for  $j > i$ , where the  $\omega_{ij}$ 's are the elements in  $\Omega_1$  (Geweke and Singleton, 1981).

## 2.2 Conditionally Heteroskedastic Seasonal Dynamic Factor Analysis

In this work we introduce the possibility of the unobserved common factors having structure both in mean and variance, since we allow for seasonal VARIMA plus ARCH/GARCH (Generalized Autoregressive Conditionally Heteroskedastic, Engle, 1982 and Bollerslev, 1986) unobserved common factors. For this purpose the conditionally heteroskedastic disturbances  $\mathbf{w}_t^*$  are added in the transition equation:

$$\mathbf{y}_t = \Omega \mathbf{F}_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, S) \quad (5)$$

$$\mathbf{F}_t = \mathbf{c} + \Psi \mathbf{F}_{t-1} + \mathbf{w}_t^*. \quad (6)$$

Disturbances  $\varepsilon_t, \mathbf{w}_t^*$ , which appear in equations (5) and (6) are mutually independent, and  $\mathbf{w}_t^* = \begin{pmatrix} \mathbf{w}_t^{*1} \\ \mathbf{0}_{(b-r) \times 1} \end{pmatrix}_{b \times 1}$ , where  $b = r \cdot (s \cdot (D + P) + d + p)$ .

In the simplest case, we allow the disturbances of the transition equation to follow univariate ARCH(1) models:

$$\begin{aligned} \mathbf{w}_t^{*1} | I_{t-1} &\sim N(\mathbf{0}_{r \times 1}, \text{diag}(\sigma_{1t}^2, \sigma_{2t}^2, \dots, \sigma_{rt}^2)) = N(\mathbf{0}_{r \times 1}, \mathbf{Q}_{1t}), \\ \mathbf{w}_t^* | I_{t-1} &\sim N(\mathbf{0}_{b \times 1}, \text{diag}(\sigma_{1t}^2, \sigma_{2t}^2, \dots, \sigma_{rt}^2, \mathbf{0}_{(b-1)})) = N(\mathbf{0}_{b \times 1}, \mathbf{Q}_t) \\ w_{j,t}^* &= \sigma_{jt} a_{jt}, \quad a_{jt} \sim NID(0, 1), \quad \sigma_{jt}^2 = \alpha_{0j} + \alpha_j w_{j,t-1}^{*2}, \text{ for } j = 1, \dots, r \end{aligned} \quad (7)$$

where  $I_{t-1}$  refers to all the information available at time  $t - 1$ .

**Quasi-Maximum Likelihood Estimation and Augmented Kalman filter** For estimating the parameters involved in this model, we must maximize the log-likelihood function, and this function is calculated for models

expressed under state space formulation using the expression (Durbin and Koopman, 2001)  $\log L = -\frac{1}{2} \sum_{t=1}^T \log((2\pi)^n |\Sigma_t|) - \frac{1}{2} \sum_{t=1}^T v_t \Sigma_t^{-1} v_t'$ , where  $v_t$  are the innovations and  $\Sigma_t$  its variance-covariance matrix. For calculating  $v_t$  and  $\Sigma_t$ , the Kalman Filter (KF) must be run, and some difficulties arise when conditional heteroskedasticity is present in the disturbances. When running the KF, the computation of the matrix whose diagonal contains  $(\sigma_{1t}^2, \sigma_{2t}^2, \dots, \sigma_{rt}^2)$  is needed, and the terms  $w_{j,t-1}^{*2}$ , for  $j = 1, \dots, r$ , must be calculated. Since they are unobservable, the KF recursions cannot be operated directly. Here, we propose extending the Homoskedastic SeaDFA to the Conditionally Heteroskedastic case, by means of the idea introduced by Harvey et al. (1992), i.e., including conditionally heteroskedastic shocks of the common factors into the "original" state vector. This is needed because the expectation of  $w_{j,t-1}^{*2}$  conditional on the information available at time  $t-1$ , i.e.  $E(w_{j,t-1}^{*2} | I_{t-1})$ , must be calculated. This quantity is obtained as an output from the KF if  $w_{j,t-1}^*$  is a latent or state variable. For the transition equation, incorporating  $\mathbf{w}_t^*$  in the "original" state vector gives:

$$\mathbf{F}_t^A = \mathbf{c}^A + \Psi^A \mathbf{F}_{t-1}^A + \mathbf{G}^A \mathbf{v}_t^A, \quad (8)$$

$$\text{and } \mathbf{F}_t^A = \begin{pmatrix} \mathbf{F}_t \\ \mathbf{w}_t^{1*} \end{pmatrix}, \mathbf{c}^A = \begin{pmatrix} \mathbf{c} \\ \mathbf{0}_{r \times 1} \end{pmatrix}, \mathbf{G}^A = \begin{pmatrix} I_r \\ 0_{(b-r) \times r} \\ I_r \end{pmatrix} \text{ and } \Psi^A = \begin{pmatrix} \Psi & \mathbf{0}_{b \times r} \\ \mathbf{0}_{r \times b} & \mathbf{0}_{r \times r} \end{pmatrix}.$$

So,  $\mathbf{w}_t^{1*}$  are playing both roles of state-vector and disturbances,  $\mathbf{v}_t^A$  denotes  $\mathbf{w}_t^{1*}$  when it is a disturbance. The conditional expectation  $E(\mathbf{v}_t^A \mathbf{v}_t^{A'} | I_{t-1}) = Q_t^A = \text{diag}(\sigma_{1t}^2, \dots, \sigma_{rt}^2)$ . Moreover, the observation equation must be replaced by  $\mathbf{y}_t = \Omega^A \mathbf{F}_t^A + \varepsilon_t$ , and  $E(\varepsilon_t \varepsilon_t') = S$ , where  $\Omega^A = (\Omega \ \mathbf{0}_{m \times r})$ .

Once we have the formulation of the Conditionally Heteroskedastic SeaDFA, the KF can be run for this "augmented" formulation. KF recursions for the augmented formulation are given by:

$$\begin{aligned} \mathbf{F}_{t|t-1}^A &= \mathbf{c}^A + \Phi^A \mathbf{F}_{t-1|t-1}^A \\ P_{t|t-1}^A &= \Psi^A P_{t-1|t-1}^A \Psi^{A'} + \mathbf{G}^A Q_t^A \mathbf{G}^{A'} \\ v_t^A &= y_t - \Omega^A \mathbf{F}_{t|t-1}^A \\ \Sigma_t^A &= \Omega^A P_{t|t-1}^A \Omega^{A'} + S \\ \mathbf{F}_{t|t}^A &= \mathbf{F}_{t|t-1}^A + P_{t|t-1}^A \Omega^{A'} (\Sigma_t^A)^{-1} v_t^A \\ P_{t|t}^A &= P_{t|t-1}^A - P_{t|t-1}^A \Omega^{A'} (\Sigma_t^A)^{-1} \Omega^A P_{t|t-1}^A \end{aligned} \quad (9)$$

For the calculation of  $Q_t^A$ ,  $\sigma_{1t}^2, \dots, \sigma_{rt}^2$  must be computed, and the approximation proposed by Harvey et al. (1992) is used, and  $E(w_{j,t-1}^{*2} | I_{t-1})$  and  $E(w_{j,t-1}^{*2} | I_{t-1})$  must be calculated for  $j = 1, \dots, r$ . But now, in the "augmented" formulation, this is easy since  $\mathbf{w}_{t-1}^*$  can be expressed as  $\mathbf{w}_{t-1}^* = E(\mathbf{w}_{t-1}^* | I_{t-1}) - (E(\mathbf{w}_{t-1}^* | I_{t-1}) - \mathbf{w}_{t-1}^*)$ , and  $E(\mathbf{w}_{t-1}^* | I_{t-1})$  are given by the last  $r$  elements in  $\mathbf{F}_{t-1|t-1}^A$ .

Concerning the variances, they can be expressed as  $E((\mathbf{w}_{t-1}^*)^2 | I_{t-1}) = (E(\mathbf{w}_{t-1}^* | I_{t-1}))^2 + E[(\mathbf{w}_{t-1}^* - E(\mathbf{w}_{t-1}^* | I_{t-1}))^2]$ , where the last addend is given by the elements  $(b+1)$  to  $(b+r)$  in the diagonal of  $P_{t-1|t-1}^A$ . So the elements in the diagonal of  $Q_t^A$ , i.e.,  $\sigma_{1t}^2, \dots, \sigma_{rt}^2$  are easily computed by:

$$\sigma_{jt}^2 = \alpha_{0j} + \alpha_{1j} \left\{ (E(w_{j,t-1}^* | I_{t-1}))^2 + E[(w_{j,t-1}^* - E(w_{j,t-1}^* | I_{t-1}))^2] \right\}. \quad (10)$$

Finally, the parameters of the model are estimated maximizing the expression for the log-likelihood in the "augmented" formulation,  $\log L^A =$

$-\frac{1}{2} \sum_{t=1}^T \log((2\pi)^n |\Sigma_t^A|) - \frac{1}{2} \sum_{t=1}^T v_t^A (\Sigma_t^A)^{-1} v_t^{A'}$ . It is easy to let the model being more general, managing not only transitory disturbances following ARCH processes, but also GARCH ones, since in practice the conditional variance of most of the data can be sufficiently described by a GARCH(1,1) model (Bollerslev, 1986). The slight modification introduced consists in considering an additional term  $\beta_{1j} \sigma_{j,t-1}^2$ , and approximating it by  $\beta_{1j} E(\sigma_{j,t-1}^2 | I_{t-2})$ , where  $E(w_{j,t-1}^* | I_{t-1})$  is obtained in (10) and  $E(\sigma_{j,t-1}^2 | I_{t-2}) = \hat{\sigma}_{j,t-1}^2$ , for  $j = 1, \dots, r$ , and it is available from  $Q_{t-1}^A$ .

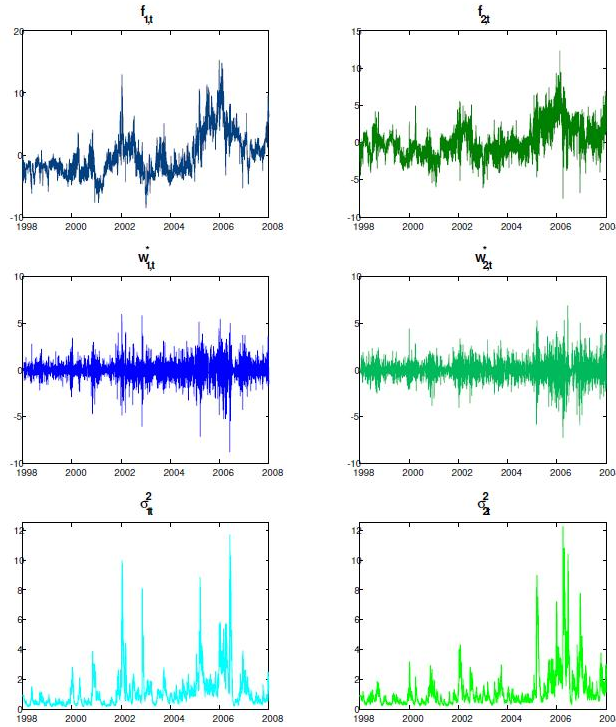
### 3 Application to electricity price data

There are two interesting problems to be solved concerning electricity price forecasting. On the one hand, power generation units must be scheduled for the forthcoming day (24-hour ahead) trying to maximize their profits. For this purpose the disposal of accurate one-step-ahead forecasts is crucial. On the other hand, the risk that bilateral contracts imply should be reduced. So, there is also a need of computing accurate year-ahead forecasts, the usual length of bilateral contracts.

#### 3.1 The model for electricity price data in the Spanish Market (1998-2007).

First of all, and using the test proposed by Peña and Poncela (2004) the number of common factors is selected. Concerning the data under study, vector of  $m = 24$  hourly series of electricity prices in the period 1998-2007,  $r = 2$  common factors are extracted. Electricity prices exhibit a weekly seasonal pattern since there is an instantaneous relationship between load and price and the consumption heavily depends on the day of the week (García-Martos et al., 2007), so seasonality of order  $s = 7$  is present. A VARIMA(1,0,0)  $\times$  (1,1,0)<sub>7</sub> with univariate GARCH(1,1) disturbances is finally chosen for the common factors. Simpler models for the conditional mean were not able to capture all features present in the data. GARCH(1,1) model (Bollerslev, 1986) is well-known for being able to capture the structure in

the vast majority of series whose conditional variance evolves over time. Thus, the GARCH-SeaDFA model is estimated for the the particular case of Spanish Market in the period 1998-2007 using QML via Augmented KF, as described in Section 2. Figure 1 shows the common factors, their estimated GARCH(1,1) disturbances,  $\mathbf{w}_t^*$ , and volatilities  $\sigma_{1t}^2$  and  $\sigma_{2t}^2$ , where  $\mathbf{w}_t^* | I_{t-1} \sim N(\mathbf{0}_{2 \times 1}, \text{diag}(\sigma_{1t}^2, \sigma_{2t}^2))$ .



**Fig. 1.** Conditionally heteroskedastic common factors and common volatility factors (1998-2007)

Estimation results are used to compute long and short-term forecasts.

### 3.2 Year-ahead forecasting

The conditionally heteroskedastic dynamic factor model estimated for the data in the period 1998-2007 is used to compute forecasts in the Spanish Market for the whole year 2008. The accuracy metrics used to check the performance of the model proposed in this paper, are those usually encountered

**Table 1.** Year-ahead forecasting errors.

Year 2008	Jan	Feb	Mar	Apr	May	Jun	
MAPE	20.25	18.00	13.06	13.31	12.82	11.10	
MAPE2	19.87	17.91	12.20	12.40	12.20	10.38	
	Jul	Aug	Sep	Oct	Nov	Dec	TOTAL
MAPE	16.63	18.30	20.66	17.31	14.98	17.39	16.15
MAPE2	15.94	17.73	18.93	15.60	14.17	15.27	15.22

in previous works (Conejo et al., 2005). These accuracy metrics are the Mean Average Percentage Error (MAPE) and the MAPE2, which are the average of the daily prediction errors in the period under study. This daily prediction error is calculated as the daily mean (for the MAPE) or median (for MAPE2) of the hourly relative forecasting error. In Table 1 the monthly MAPE and MAPE2 for the whole year 2008, using the model estimated with the data from 1998 up to 2007 are shown.

It should be pointed out, that the models usually handled to compute one-day-ahead forecasts are not valid for year ahead forecasting (long-term). For example, using the Mixed Model provided in García-Martos et al. (2007), which is the one with the best results in the Spanish market gets errors for the predictions in 2008 which are above 35%. Furthermore, the GARCH-SeaDFA model not only produce accurate forecasts in the long-run, but also in the short-term.

## 4 Conclusions

This paper faces the important problem of forecasting electricity prices and their volatilities in liberalized markets, both in the long and short term. Electricity prices present structure both in the conditional mean and conditional variance, so it is necessary to develop methodology that is able to capture jointly both dynamics. In this paper we develop the Conditionally Heteroskedastic SeaDFA, allowing for unobservable common factors that exhibit conditional heteroskedasticity. Detailed numerical results are provided for the Spanish Market, using the new GARCH-SeaDFA model. Moreover, concerning forecasting, the results for all the hourly prices in the year 2008 are calculated and validated, both for the short and long term.

The methods here proposed, in which the common trend of a vector of series is captured both for the conditional mean and variance, could be of application in the field of macroeconomic data.

## References

- ALONSO, A.M., GARCIA-MARTOS, C., RODRIGUEZ, J. and SANCHEZ, M.J.  
(2008): Seasonal Dynamic Factor Analysis and Bootstrap Inference: application

- to electricity market forecasting. *Working Paper 08-14, Statistics and Econometrics Series, Universidad Carlos III de Madrid.*
- ANSLEY, C.F. and KOHN, R. (1986): Estimation Prediction and Interpolation for ARIMA models with missing data. *Journal of the American Statistical Association*, 81, 395, 751-761.
- BOLLERSLEV, T. (1986): Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31, 307-327.
- CONEJO, A.J., CONTRERAS, J., ESPINOLA, R., PLAZAS, M.A. (2005): Forecasting electricity prices for a day-ahead pool-based electric energy market. *International Journal of Forecasting*, 21, 3, 435-462.
- CONEJO, A.J., CARRION, M., MORALES, J.M. and NOGALES, J. (2010): Electricity Pool Prices: Long-Term Uncertainty Characterization for Futures-Market Trading and Risk Management. *Journal of the Operational Research Society*, 61, 2, 235-245.
- CONNOR, G., KORAJCZYK, R., LINTON, O.B. (2006): The common and specific components of dynamic volatility. *Journal of Econometrics*, 132, 1, May 2006, 231-255.
- CONTRERAS, J., ESPINOLA, R., NOGALES, F.J. and CONEJO, A.J. (2003): ARIMA Models to Predict Next-Day Electricity Prices. *IEEE Transactions on Power Systems*, 18, 3, 1014-1020.
- DIEBOLD, F.X. and NERLOVE, M. (1989): The dynamics of exchange rate volatility: a multivariate latent factor ARCH model. *Journal of Applied Econometrics*, 4, 1, 1-21.
- DURBIN, J. and KOOPMAN, S.J. (2001): *Time series analysis by state space methods*. Oxford University Press.
- ENGLE, R.F. (1982): Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50, 4, 987-1007.
- GARCIA-MARTOS, C., RODRIGUEZ, J. and SANCHEZ, M.J. (2007): Mixed models for short-run forecasting of electricity prices: application for the Spanish market. *IEEE Transactions on Power Systems*, 2, 2, 544-552.
- GEWEKE, J. and SINGLETON, K.J. (1981): Maximum likelihood 'confirmatory' factor analysis of economic time series. *International Economic Review*, 22, 37-54.
- HARVEY, A., RUIZ, E. and SENTANA, E. (1992): Unobserved Component Time Series Models with ARCH Disturbances. *Journal of Econometrics*, 52, 129-158.
- KOOPMAN, S.J., OOMS, M. and CARNERO, M.A. (2007): Periodic Seasonal Reg-ARFIMA-GARCH Models for Daily Electricity Spot Prices. *Journal of the American Statistical Association*, 102, 477, 16-27.
- LEE, R.D. and CARTER, L.R. (1992): Modeling and Forecasting U. S. Mortality. *Journal of the American Statistical Association*.
- PEÑA, D. and BOX, G.E.P. (1987): Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, 87, 419, 659-671.
- PEÑA, D. and PONCELA, P. (2004): Forecasting with Nonstationary Dynamic Factor Models. *Journal of Econometrics*, 119, 2, 291-321.
- PEÑA, D. and PONCELA, P. (2006): Nonstationary Dynamic Factor Analysis. *Journal of Statistical Planning and Inference*.
- STOCK, J.H. and WATSON, M. (2002): Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 2002, 97, 460, 1167-1179.

# Consensus Analysis Through Modal Symbolic Objects

Jose M. Garcia-Santesmases<sup>1</sup> and M. Carmen Bravo<sup>2</sup>

<sup>1</sup> Universidad Complutense de Madrid, Departamento de Estadística e Investigación Operativa, Facultad de Ciencias Matemáticas, 28040 Madrid, Spain, *josemgar@mat.ucm.es*

<sup>2</sup> Universidad Complutense de Madrid, Servicio Informático de Apoyo a Docencia e Investigación, Edificio Real Jardín Botánico Alfonso XIII, 28040 Madrid, Spain, *mcbravo@pas.ucm.es*

**Abstract.** This paper addresses the problem of analyzing the existence of different patterns of consensus when data come from several observers who separately evaluated several issues on a rating scale of ordered categories.

We give a consensus measure for a group of individuals and we analyze its properties. A procedure that uses clustering techniques and modal symbolic objects is given to identify and describe groups of individuals with a high agreement on several questions, consensus groups. We use the concept of extension of modal symbolic objects to give the consensus solution: a set of consensus groups (symbolic objects) that cover all the individuals satisfying at least a fixed number of issues.

**Keywords:** symbolic objects, consensus analysis, cluster analysis, consensus measure

## 1 Introduction

One common meaning of consensus is a general agreement among the members of a given group and can be seen as a function of shared team feelings towards an issue. To analyze it, two main steps can be distinguished:

a) The use of consensus measures to evaluate the strength of consensus in a class of individuals

b) The evaluation of each individual to propose changes in his opinion in order to increase the strength of the consensus.

For step a) we give a consensus measure for a group of individuals that satisfies the requirements given by Tastle (2005) based on a single issue. We extend this measure to several issues and we define consensus measures for symbolic objects.

For step b) we extend the procedure described in Garcia-Santesmases and Bravo (2008) and we use modal symbolic objects to analyze the consensus on a class of individuals. A clustering based solution is proposed. Symbolic objects built from the obtained clusters are the consensus groups, which extensions of a fixed level satisfying at least a fixed number of issues cover the set of individuals.

## 2 Basic concepts and notation

Let  $E = \{u_1, u_2, \dots, u_n\}$  be the set of individuals or experts that answer  $p$  questions  $y_1, y_2, \dots, y_p$  on an ordinal scale  $Y_j = \{r_1^j, r_2^j, \dots, r_{k_j}^j\}$  that represents the ratings or rankings of each individual preference. We shall consider a symmetric ordinal level rating for example like the rating levels on a Likert-type scale. The rating measures the extent to which a person agrees or disagrees with the question. For example with five possible values: 1 strongly disagree, 2 somewhat disagree, 3 undecided, 4 somewhat agree, 5 strongly agree.

In a general way this order relation can be standardized and be represented by  $k_j \times k_j$  symmetric matrix  $\mathbf{A}^j = (a_{ls}^j)$  with  $a_{ls}^j = 1 - \frac{|r_l^j - r_s^j|}{|r_{k_j}^j - r_1^j|}$ . The value  $r_{ls}^j$  represents the similarity between  $r_l^j$  and  $r_s^j$  in a  $[0, 1]$  scale.

This kind of matrix satisfies:

$$a_{ll}^j = 1, a_{1k_j}^j = 0 \quad (1)$$

$$a_{1s}^j + a_{1(k_j-s+1)}^j = 1 \quad (2)$$

for  $s = 1, \dots, \frac{k_j}{2}$  when  $k_j$  is *even* and  $s = 1, \dots, \frac{k_j+1}{2}$  when  $k_j$  is *odd*

$$a_{ls}^j = a_{(l+1)(s+1)}^j \quad (3)$$

for  $l = 2, \dots, k_j - 1$  and  $s = 1, 2, \dots, k_j - 1$

For example, for a Likert type scale  $Y_j = \{1, 2, 3, 4, 5\}$  we have:

$$\mathbf{A}^j = \begin{pmatrix} 1 & 0.75 & 0.5 & 0.25 & 0 \\ 0.75 & 1 & 0.75 & 0.5 & 0.25 \\ 0.5 & 0.75 & 1 & 0.75 & 0.5 \\ 0.25 & 0.5 & 0.75 & 1 & 0.75 \\ 0 & 0.25 & 0.5 & 0.75 & 1 \end{pmatrix}$$

The  $p$  questions can be represented by a  $k \times k$  ( $k = \sum k_j$ ) matrix given by:

$$\mathbf{A} = \frac{1}{p} \begin{pmatrix} \mathbf{A}^1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^2 & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{A}^p \end{pmatrix}$$

Let  $s$  be a modal symbolic object (Bock and Diday (2000), Diday and Noirhomme-Fraiture (2008)) such that  $s = \bigwedge_{j=1, \dots, p} [y_j R_j D_j]$ ,  $D_j = (r_1^j(w_1^j), r_2^j(w_2^j), \dots, r_{k_j}^j(w_{k_j}^j))^t$ , where  $w_1^j, w_2^j, \dots, w_{k_j}^j$  are the frequency, probability or weight values assigned to the values  $r_1^j, r_2^j, \dots, r_{k_j}^j$  of variable  $y_j$  and  $R_j$  is a relation between descriptions  $y_j(u)$  and  $D_j$  given by the product through



$\mathbf{A}^j$ :  $[y_j(u)R_jD_j] = z^j(u)^t \mathbf{A}^j w^j$  with  $w^{jt} = (w_1^j, w_2^j, \dots, w_{k_j}^j)$  and  $z^j(u)^t = (0, 0, \dots, 0, 1^{(y_j(u))}, 0, \dots, 0)$ , where value 1 is placed in the position of the order of the  $y_j(u)$  category in  $\{1, \dots, k_j\}$ . The symbolic object  $s$  is the mapping  $s : E \rightarrow [\frac{1}{2}, 1]$  defined by:

$$s(u) := \mathbf{z}(u)^t \mathbf{A} \mathbf{w} \quad \text{for } u \in E \quad (4)$$

with  $\mathbf{w}^t = (w^{1t}, w^{2t}, \dots, w^{pt})$  and  $\mathbf{z}(u)^t = (z^1(u)^t, z^2(u)^t, \dots, z^p(u)^t)$ .

For any given threshold  $\alpha \in [\frac{1}{2}, 1]$ , the extent of level  $\alpha$  of  $s$ , is defined by:

$$g_\alpha(s) := EXT_\alpha(s) = \{u \in E | s(u) \geq \alpha\} \quad (5)$$

For  $R_j$  and  $\alpha$  we define the Boolean relation  $R_{j\alpha}$  as:

$$[y_j(u)R_{j\alpha}D_j] = 1 \text{ when } [y_j(u)R_jD_j] \geq \alpha \quad (6)$$

The value  $\sum_{j=1, \dots, p} [y_j(u)R_{j\alpha}D_j]$  is the number of issues or variables for which  $u$  verifies the descriptions of  $s$  at level  $\alpha$

For a given integer  $q$  with  $0 \leq q \leq p$ , we define the extension of level  $(\alpha, 1 - q/p)$  of  $s$  by:

$$g_\alpha^{1-q/p}(s) := \left\{ u \in E \mid \frac{1}{p} \left( \sum_{j=1, \dots, p} [y_j(u)R_{j\alpha}D_j] \right) \geq 1 - \frac{q}{p} \right\} \quad (7)$$

The set  $g_\alpha^{1-q/p}(s)$  is composed by the individuals that verifies the description of  $s$  at level  $\alpha$  at least in  $p - q$  issues. We have  $g_\alpha(s) = g_\alpha^1(s) = g_\alpha^{1-q/p}(s) = g_\alpha^0(s) = E$ .

As in Brito (2000) we define the object builder function  $f_m$  that associates for any subset  $C \subset E$ , a modal symbolic object  $f_m(C)$  defined by:

$$f_m(C) := \bigwedge_{j=1, \dots, p} [y_j R_j D_j], \quad w_l^j = p_l^j(C) \quad (8)$$

where  $p_l^j(C)$  are the relative frequencies of  $r_l^j$  on  $C$ .

### 3 Consensus measure

For  $C \subset E$ , we propose the following consensus measure of  $C$  for issue  $y_j$ :

$$ts(j, C) := p^j(C)^t \mathbf{A}^j p^j(C) \quad (9)$$

with  $p^j(C)^t = (p_1^j(C), p_2^j(C), \dots, p_{k_j}^j(C))$ .

This measure can be extended to all issues by :

$$\mathbf{ts}(C) := \mathbf{P}(C)^t \mathbf{A} \mathbf{P}(C) \quad (10)$$

with  $\mathbf{P}(C)^t = (p^1(C)^t, p^2(C)^t, \dots, p^p(C)^t)$ .

Due to properties (1), (2) and (3) of  $\mathbf{A}^j$  it is easy to prove that this definition satisfies the requirements given by Tastle et al. (2005) for a consensus measure:

- a. *For a given (even) number of individuals if an equal number of individuals, separate themselves into two disjoint teams, each centered on the strongly disagree and strongly agree categories, the team is considered to have no consensus. In this case, for any issue  $y_j$ ,  $p^j(C)^t = (\frac{1}{2}, 0, \dots, 0, \frac{1}{2})$  and  $ts(j, C) = p^j(C)^t \mathbf{A}^j p^j(C) = \frac{1}{2}$ , that is the minimum for (9) and  $\mathbf{ts}(C) = \mathbf{P}(C)^t \mathbf{A} \mathbf{P}(C) = \frac{1}{2}$ .*
- b. *If all the participants classify themselves in the same category, regardless of the category, then the consensus is complete. In this case,  $ts(j, C) = p^j(C)^t \mathbf{A}^j p^j(C) = 1$ , that is the maximum for (9) and  $\mathbf{ts}(C) = \mathbf{P}(C)^t \mathbf{A} \mathbf{P}(C) = 1$ .*
- c. *If the mix of participants is such that  $\frac{n}{2} + 1$  participants assign themselves to any one category the consensus must be greater than  $\frac{1}{2}$ . The balance in the team is no longer equal.*
- d. *As the number of categories to which each participant classifies himself/herself dismisses, the agreement/consensus must increase.*
- e. *The dispersion of the categorical values must be captured by the consensus to provide an indication of the variance of the data.*

Definition (10) can be easily generalized to deal with symbolic objects. Let  $s$  be a symbolic object we define the consensus of  $s$  as:

$$\mathbf{ts}(s) := \mathbf{w}^t \mathbf{A} \mathbf{w} \quad (11)$$

## 4 Consensus analysis

To analyze the consensus of the set  $E$  we can use the above consensus measures to evaluate the strength of the consensus through  $\mathbf{ts}(E)$  or  $\mathbf{ts}(f_m(E))$  and use a graphical representation of  $f_m(E)$ . We propose a graphic representation based on the 2D zoom star of Noirhomme-Fraiture and Rouard (2000) as an explanatory graphical representation of  $f_m(E)$ . In each axis the intervals are of length  $1 - ts(j, f_m(E))$  and are centered in the mean value of  $y_j(u)$ ,  $u \in E$ . This star representation gives a good idea of the type and strength of the consensus in  $E$ . The narrower the star is, the stronger the consensus is. Given  $u \in E$ , the value  $f_m(E)(u)$  given by (4) can be interpreted as the strength of the consensus of  $E$  with individual  $u$ . Any individual can be displayed on the same graphic as Figure 1 shows in section 5.

This generalization process of  $E$  into  $f_m(E)$  usually gives over-generalization, that is, low values for  $\mathbf{ts}(f_m(E))$ . A desirable property is that  $E$  is

covered by  $g_\alpha^{1-q/p}(f_m(E))$  with a high value of  $\alpha$  and a low value of  $q$ . We propose a clustering based solution to reduce this over-generalization by identifying a set of symbolic objects  $\mathbf{S} = \{s_1, s_2, \dots, s_L\}$  such that  $\{g_\alpha^{1-q/p}(s_1), g_\alpha^{1-q/p}(s_2), \dots, g_\alpha^{1-q/p}(s_L)\}$  is a clustering of  $E$ , not necessarily a partition and such that all individuals of  $E$  are covered satisfying at least p-q issues. Desirable properties of  $\mathbf{S}$  in order to increase the consensus measure are those of any clustering process, based in maximizing the values of  $\mathbf{ts}(s)$  of each  $s \in \mathbf{S}$  and minimizing  $L$ .

Each  $s \in \mathbf{S}$  shall be a consensus group and  $\mathbf{S}$  a solution of level  $(\alpha, 1-q/p)$ . We apply a method similar to Hartigan's (1975, p. 74-78) leader algorithm to obtain solutions of level  $(\alpha, 1-q/p)$ . The algorithm starts with an initial individual seed to which iteratively other individuals are added to get a maximal subset  $C_1$  in the sense that it is included in the  $\alpha$  level extension of the symbolic object  $s_1 = f_m(C_1)$ . In a second step, if  $E$  is not covered by the  $(\alpha, 1-q/p)$  extension of  $s_1$  then a new individual seed is selected in  $E - C_1$  and, in the same way as above, we get a new maximal subset  $C_2 \subset E - C_1$  and  $s_2 = f_m(C_2)$ . The process follows while  $E$  is not covered by the  $(\alpha, 1-q/p)$  extensions of the symbolic objects, selecting individual seeds and getting maximal subsets with individuals in the complementary of the union of the already obtained clusters.

#### Algorithm

Step 1 (initialization)

Fix  $\alpha \in [\frac{1}{2}, 1]$  and  $q \in \{0, 1, \dots, p\}$

Let  $u_1 \in E$ ,  $C_1 = \{u_1\}$ ,  $r = 1$ ,  $\mathbf{C} = \{C_1\}$

Step 2

While  $E \neq \bigcup_{C \in \mathbf{C}} g_\alpha^{1-q/p}(f_m(C))$

Do

$E' = E - \bigcup_{C \in \mathbf{C}} C$

If  $\{u \in E' | u \in g_\alpha(f_m(C_r \cup \{u\}))\} = \emptyset$  then:

Select any  $u \in E'$

$C_{r+1} = \{u\}$ ,  $\mathbf{C} \leftarrow \mathbf{C} \cup C_{r+1}$ ,  $r \leftarrow r + 1$

Else

Select  $u \in \{u \in E' | u \in g_\alpha(f_m(C_r \cup \{u\}))\}$

$C_r \leftarrow C_r \cup \{u\}$

End Do

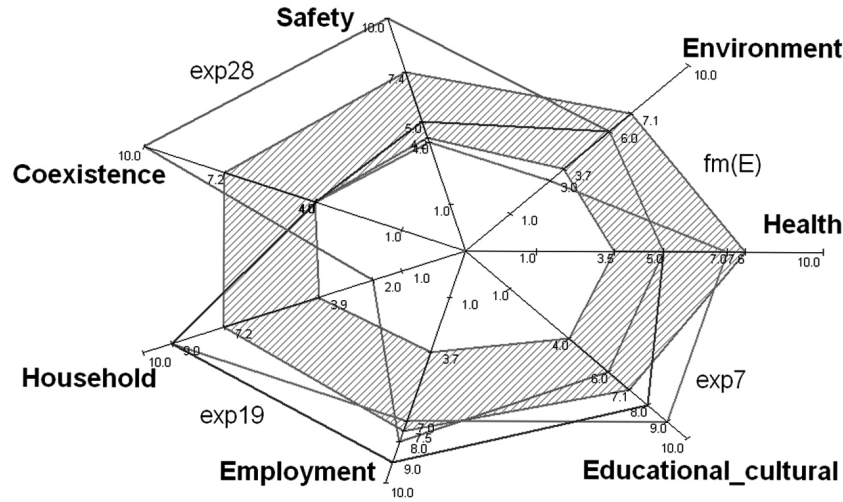
## 5 Example

We have applied the algorithm to a data set that represents the ratings given by 50 experts to 7 quality indexes. The ratings are on a 1 to 10 scale with 1 meaning very low quality and 10 meaning very strong quality ( $Y_j = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ ). The indexes are:  $y_1$ , health;  $y_2$ , educational and cultural;  $y_3$ , employment;  $y_4$ , household;  $y_5$ , economical resources;  $y_6$ , safety;

$y_7$ , environment. The data table is:

expert	$y_1$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$	$y_7$	expert	$y_1$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$	$y_7$	expert	$y_1$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$	$y_7$
exp1	6	7	8	6	6	6	5	exp19	5	8	9	9	4	5	6	exp37	5	8	5	5	4	6	8
exp2	4	9	6	8	3	1	8	exp20	3	10	7	9	3	5	7	exp38	3	5	8	6	3	4	10
exp3	6	10	10	10	1	1	10	exp21	8	2	4	5	8	7	2	exp39	2	6	4	5	4	8	9
exp4	9	10	9	9	2	1	6	exp22	10	6	7	6	7	9	3	exp40	5	7	7	7	6	4	3
exp5	8	10	10	10	4	2	5	exp23	9	1	4	2	10	7	6	exp41	7	3	7	4	8	9	1
exp6	3	4	7	7	6	5	3	exp24	5	2	3	4	7	7	3	exp42	4	7	9	7	7	6	4
exp7	7	9	7	9	4	4	3	exp25	8	4	6	6	8	7	3	exp43	6	5	3	7	6	5	3
exp8	4	6	4	3	4	6	10	exp26	6	4	2	4	7	8	3	exp44	5	4	4	5	4	4	6
exp9	2	3	6	1	5	8	6	exp27	8	9	6	8	6	5	3	exp45	7	2	1	5	7	7	7
exp10	1	1	6	1	6	9	7	exp28	5	6	8	2	10	10	6	exp46	5	2	3	5	6	6	7
exp11	2	7	4	3	4	7	9	exp29	8	5	6	6	9	7	5	exp47	5	6	3	5	4	4	2
exp12	6	3	2	4	7	7	6	exp30	7	7	3	3	7	6	5	exp48	5	3	1	5	7	7	4
exp13	8	3	4	5	7	6	4	exp31	4	6	5	5	8	8	4	exp49	6	5	2	6	4	7	5
exp14	3	6	7	7	4	4	5	exp32	4	5	5	4	4	7	6	exp50	7	4	7	5	6	5	5
exp15	4	4	7	6	6	4	5	exp33	6	7	7	5	7	6	4	exp18	6	6	9	7	5	6	5
exp16	5	4	5	6	5	4	6	exp34	7	2	3	2	8	7	5	exp36	6	6	5	4	3	5	7
exp17	5	7	8	7	4	3	6	exp35	7	9	6	6	5	8	9								

For the application of the algorithm we have considered  $\alpha = 0.8$  and  $q = 2$ .



**Fig. 1.** The 2D zoom star for  $f_m(E)$  and individuals exp7, exp19 and exp28 of expert data set.

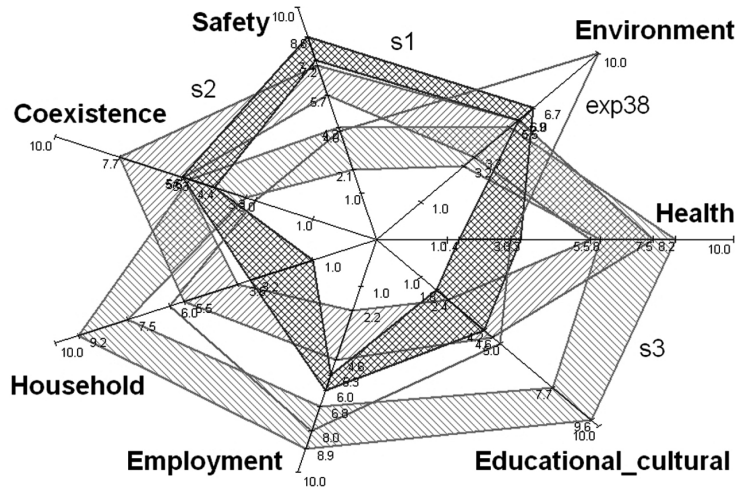
Initially we have  $\mathbf{ts}(E) = 0.78$ . The set  $g_{0.8}^{1-5/7}(E)$  does not cover  $E$  and it is not a  $(0.8, 1-5/7)$  consensus solution. It contains only 19 experts representing a relative weight of 0.38. The graphical representation of  $f_m(E)$  and the individuals *exp7*, *exp19* and *exp28* are shown in Figure 1. These individuals do not verify the description of  $f_m(E)$  at level 0.8 in at least five issues.

The output of the algorithm is  $\mathbf{S} = \{s_1, s_2, s_3, s_4\}$ ,  $s_i = f_m(C_i)$ ,  $i = 1, \dots, 4$ .  $C_1$  contains experts 8, 12, 13, 16, 21, 24, 26, 34 and 43 to 49,  $C_2$  contains experts 4, 5, 7, 18, 20, 27 and 35,  $C_3$  contains experts 9, 10 and 32 and  $C_4$  contains expert 38.

In the next table, consensus measures for  $s_i$  are shown. The values  $p_{C_i}$  and  $p_{g_{0.8}^{1-5/7}(s_i)}$  represent the relative weights of  $C_i$  and  $g_{0.8}^{1-5/7}(s_i)$  in  $E$ , respectively.

$s_i$	$\mathbf{ts}(s_i)$	$p_{C_i}$	$p_{g_{0.8}^{1-5/7}(s_i)}$
$s_1$	0.87	0.3	0.5
$s_2$	0.88	0.14	0.32
$s_3$	0.94	0.06	0.16
$s_4$	1	0.02	0.26

A global consensus measure of the solution could be a weighted average of  $\mathbf{ts}(s_i)$  with weights  $p_{C_i}$ , that is, 0.886.



**Fig. 2.** The 2D zoom star for the consensus solution  $s_i = f_m(C_i)$ ,  $i = 1, \dots, 4$

The 2D zoom star representations of  $s_i = f_m(C_i)$ ,  $i = 1, \dots, 4$  are shown in Figure 2.

A property of this algorithm is that experts in  $C_i$  verify the descriptions of  $s_i$  at level 0.8 in all issues. An expert can be covered by more than one

consensus group to level  $(0.8, 1 - 5/7)$ ). For example, *exp6* is covered by  $s_1$ ,  $s_2$  and  $s_4$  and *exp12* by  $s_1$  and  $s_3$ . In this data set, three experts are covered by three consensus groups and nine experts are covered by two.

## 6 Conclusion

We have described how modal symbolic objects can be used to analyze consensus over a set of individuals. We have given a consensus measure for a set of individuals and we have extended it to modal symbolic objects. The consensus solution is obtained applying a clustering algorithm, the symbolic object builder and symbolic object level extensions. We have used a graphical representation of the consensus groups based on the zoom star representation. An example is given to illustrate a complete analysis.

On going work is being done to maximize a global consensus measure of the solution and to compare consensus solutions using different consensus measures and different values of  $\alpha$  and  $q$ .

## References

- BOCK, H.H. and DIDAY, E. (Eds.) (2000): *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, Heidelberg.
- BRITO, P. (2000): Hierarchical and Pyramidal Clustering with Complete Symbolic Objects. In: H.H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data*. Springer Verlag, Heidelberg, 312-323.
- DIDAY, E. and NOIRHOMME-FRAITURE, M. (Eds.) (2008): *Symbolic Data Analysis and the SODAS Software*. Wiley & Sons, Chichester.
- GARCIA-SANTESMASES, J.M. and BRAVO, M.C. (2008): Analysis of Consensus Through Symbolic Objects. In: P. Brito (Ed.): *Proceedings in Computational Statistics 2008, vol II*. Physica Verlag, 481-489.
- HARTIGAN, J. A. (1975): *Clustering Algorithms*. Wiley & Sons, New York.
- NOIRHOMME-FRAITURE, M. and ROUARD, M. (2000): Visualizing and Editing Symbolic Objects. In: H.H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data*. Springer Verlag, Heidelberg, 125-138.
- STÉPHAN, V., HEBRAIL, H. and LECHEVALLIER, Y. (2000): Generation of Symbolic Objects from Relational Databases. In: H.H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data*. Springer Verlag, Heidelberg, 78-105.
- TASTLE, W. J., WIERMAN, M. J. and DUMDUM, U. R. (2005): Ranking Ordinal Scales Using the Consensus Measure. *Iss. in Information Systems VI (2)*, 96-1.

# Nonlinear Regression Model of Copper Bromide Laser Generation

Snezhana Georgieva Gocheva-Ilieva<sup>1</sup> and Iliycho Petkov Iliev<sup>2</sup>

<sup>1</sup> Department of Applied Mathematics and Modelling, University of Plovdiv  
24 Tzar Assen Str., 4000 Plovdiv, Bulgaria, *snow@uni-plovdiv.bg*

<sup>2</sup> Department of Physics, Technical University of Sofia, branch Plovdiv  
25 Tzanko Dzhushtabanov Str., 4000 Plovdiv, Bulgaria, *iliev55@abv.bg*

**Abstract.** The focus of this study is on the relationship between the output laser power and basic laser input variables in copper bromide vapour laser with wavelengths of 510.6 and 578.2 nm. Based on experimental data, a nonlinear regression model has been constructed. To deal with the multicollinearity the initial predictors were grouped in three PCA factors. The transformation of factors by the Yeo-Johnson transformation was applied. The model has been validated using independent evaluation data sets. The results obtained via the model allow for a more thorough analysis of relationship between the most important laser parameters in order to improve further experiments planning and laser production technology.

**Keywords:** Yeo-Johnson transformation, PCA factors, nonlinear regression, output laser power

## 1 Introduction

The use of multidimensional statistical methods for the study of the behavior of output laser characteristics of gas vapor lasers, and in particular those of a copper bromide vapor laser, was introduced in the last few years (see Iliev et al. (2008a, 2008b, 2007, 2009) and Gocheva-Ilieva and Iliev (2010)). In papers Iliev et al. (2008a, 2008b) we used factor analysis to study 10 independent laser variables, showing that only 6 of them are statistically significant. These variables were grouped in three factors derived by means of multiple factor analysis with Principal Component Analysis (PCA). After that, using the factors we constructed linear regression models for the estimation of the response variable – output laser power  $P_{out}$ . In Iliev et al. (2007, 2009) the same data population was examined through cluster analysis. The relevance of the observed data was confirmed both for their grouping and their level of influence on the dependent variable. Recently in Gocheva-Ilieva and Iliev (2010), these models were compared to nonparametric models, constructed using the multivariate adaptive regression splines technique (MARS), developed in Friedman (1991). It was established that nonparametric methods

provide better estimates and better prediction compared to standard methods for multidimensional linear regression (MLR).

Within this study, on the basis of an experimental data sample and the obtained factor variables, a nonlinear model is constructed using the least squares method. Nonlinear regression (NLR) has been applied to factor variables subject to Yeo-Johnson transformation. The model was tested using a cross-validation technique. It has been established that the nonlinear model provides better estimates for output laser power as compared to the respective parametric MLR. The model can be utilized when describing the relationship between independent input laser characteristics and output laser generation in order to improve the experiment.

Modeling was carried out based on experimental data obtained at the Laboratory of Metal Vapour Lasers with the Georgi Nadjakov Institute of Solid State Physics, Bulgarian Academy of Sciences. The models have been calculated using the statistical package SPSS and *Mathematica* software.

## 2 Problem setup

We are studying a copper bromide vapor laser with wavelength 510.6 nm and 578.2 nm. This is a metal vapor laser which operates under medium and low pressures. It is notably the laser with the highest efficiency in the visible spectrum and is easy to maintain due to its low gas temperature (around 600 °C) while at the same time it is capable of self-heating. The copper bromide vapor laser has a wide range of applications (see for instance Sabotinov (2006) and Foster (2005)).

In order to construct the nonlinear model and to carry out the statistical analysis we use the following independent laser variables:  $D$  (mm) - inside diameter of the laser tube;  $dr$  (mm) - inside diameter of the rings;  $L$  (cm) - length of the active area (electrode separation);  $P_{in}$  (kW) - input electrical power;  $PL$  (kWm<sup>-1</sup>) - input electrical power per unit length;  $PH_2$  (Torr) - hydrogen gas pressure.

The response variable is laser generation or output laser power,  $P_{out}$  (W).

All of the used experimental data has been published in Astadjov et al. (1985, 1994, 1997a, 1997b), Dimitrov and Sabotinov (1996), NATO (2000), and Stoilov et al. (2000). It includes different CuBr lasers, which can be divided into three main groups according to their geometry: small-bore lasers of inside diameter  $D < 20$  mm, medium-bore lasers of  $D = 20 - 40$  mm and large-bore lasers of  $D > 40$  mm. From the available data for about 300 experiments with the three general types of lasers, a random sample with size  $n = 109$  has been used. Since over 60% of all data is about small-bore lasers, the sample is partially stratified, in order to avoid the imbalance of the available data. The data is of historical type. Here we have to mention the complexity, long duration and high cost of each conducted experiment.



Typically, the studied data does not meet the requirement for multivariate normal distribution, although this can be assumed for the global population. Furthermore, as already shown in Iliev et al. (2008a, 2008b, 2007, 2009) the abovementioned six independent variables exhibit a strong multicollinearity. For this reason first we apply a multivariate factor analysis in order to determine the PCA factors, which later on act as predictors. The models constructed so far with the aid of MLR are not completely satisfactory and can be considered to be the first approximation for the description of the dependencies we are interested in.

In this study, our goal is to construct a nonlinear regression model which would provide a more accurate description of the relationship between the data and to study the predictive power of the model.

### 3 Application of Principal Components Factor Analysis

In order to avoid the multicollinearity phenomenon we are going to group the six independent variables using a classic multidimensional factor analysis. Normally this method can be applied without making any distributional assumption (e.g., Gaussian) (see for instance Izenman (2008, page 583)). Using the SPSS software for our data sample we obtained the Kaiser-Meyer-Olkin measure of sampling adequacy  $KMO=0.660$  and Bartlett's test of sphericity with significance level equal to 0.000. The respective measures of sampling adequacy (MSA) are also of significance for each variable. This indicates that the factor analysis of the sample is adequate and can be carried out. The factors have been extracted using PCA. Usually the number of factors chosen is equal to the number of eigenvalues of the correlation matrix greater than 1. However, as shown for instance in Jolliffe (1982) and Izenman (2008), the low-variance principal components may also be important. In our case, although there is only one eigenvalue greater than one, we have chosen the number of factors to be three. When variables are grouped in three factors the subsequent rotation using the Varimax method with Kaiser normalization clearly reveals the following orthogonal factors:  $F_1$  (including  $Pin, dr, L, D$ ),  $F_2$  (including  $PL$ ) and  $F_3$  (including  $PH2$ ). They account for 95.41% of the total variability of the data sample. The choice of three factors is justified as follows. When hydrogen is added this leads to a two-fold increase of  $P_{out}$ , which is an indisputable fact proven by experimental results (see Astadjov et al. (1985, 1997b)) and so the factor  $F_3$  must not be overlooked. The  $PL$  variable (factor  $F_2$ ) also plays a special role and during experiments it has been detected to noticeably affect laser generation. Omission of this variable leads to regression models which do not provide sufficiently good estimates.

For a sample with size  $n = 109$  at level of significance  $\alpha = 0.05$ , the statistically significant factor loadings are those with absolute value over 0.5 (see SOLO (1993)). The factor loadings of the observed six input variables are respectively: in factor  $F_1$  -  $Pin(0.913), dr(0.887), D(0.807), L(0.769)$ ; in

factor  $F_2$  -  $PL(-0.914)$ ; in factor  $F_3$  -  $PH2(0.929)$ . The good quality of the factor model is confirmed by the calculated reproduced correlations matrix, for which there is only one non-redundant residual with absolute value greater than 0.05 (actually it is equal to -0.059).

The factor scores which are used in this study have also been calculated at this stage of the statistical analysis.

## 4 Nonlinear regression model

### 4.1 Yeo-Johnson transformation of PCA factors

The careful examination of generated PCA factor variables shows that their relationships with the output laser power  $P_{out}$  are partially polynomial rather than linear.

Statistics utilizes various transformations in order to improve the mutual data distribution. In our case the standardized factor scores have both positive and negative values. For this reason, we are going to use the following transformation of Yeo-Johnson (Yeo and Johnson (2000)):

$$\psi_{Y-J}(\lambda, x) = \begin{cases} \{(x+1)^\lambda - 1\}/\lambda, & x \geq 0, \lambda \neq 0 \\ \log(x+1), & x \geq 0, \lambda = 0 \\ -\{(-x+1)^{2-\lambda} - 1\}/(2-\lambda), & x < 0, \lambda \neq 2 \\ -\log(-x+1), & x < 0, \lambda = 2 \end{cases}$$

Here  $x$  is the transformed variable, and  $\lambda$  is a parameter. The Yeo-Johnson transformation has a number of good properties, including continuous first and second derivatives with respect to  $\lambda$ , usually  $\lambda \in [-2, 2]$ .

### 4.2 Construction and estimation of the nonlinear model

We are looking for the nonlinear model for estimation of  $P_{out}$  in the following form:

$$\widehat{P_{out}}(\theta, \lambda) = \theta_0 + \theta_1 \psi_{Y-J}(\lambda_1, F_1) + \theta_2 \psi_{Y-J}(\lambda_2, F_2) + \theta_3 \psi_{Y-J}(\lambda_3, F_3), \quad (1)$$

where the parameters  $\theta_i$ , ( $i = 0, \dots, 3$ ) and  $\lambda_j$ , ( $j = 1, \dots, 3$ ) are determined using the least squares method.

In order to carry out the calculations we have compiled the *Mathematica* compact code shown in Figure 1.

The resulting parameters for the seven-dimensional model (1) are

$$\theta_0 = 39.735372, \theta_1 = 27.167573, \theta_2 = 4.456846, \theta_3 = 11.777153, \quad (2)$$

$$\lambda_1 = 1.290534, \lambda_2 = 0.381756, \lambda_3 = 0.767572.$$

```

 $\psi[\lambda\_ , y\_ ] := \text{If}[y \geq 0 \ \&\& \ \lambda \neq 0, \frac{(y+1)^\lambda - 1}{\lambda},$ 
 $\text{If}[y < 0 \ \&\& \ \lambda \neq 2, -\frac{(1-y)^{2-\lambda} - 1}{2-\lambda},$ 
 $\text{If}[y \geq 0 \ \&\& \ \lambda == 0, \text{Log}[y+1],$ 
 $\text{If}[y < 0 \ \&\& \ \lambda == 2, -\text{Log}[1-y]]]$ 

n = 109;
f1 = ReadList["f1-109.txt", Number] ;
f2 = ReadList["f2-109.txt", Number] ;
f3 = ReadList["f3-109.txt", Number] ;
pout = ReadList["pout-109.txt", Number] ;
data = Table[{f1[[i]], f2[[i]], f3[[i]], pout[[i]]}, {i, 1, n}];

<< Statistics`NonlinearFit`
NonlinearRegress[data, theta0 + theta1 *  $\psi[\lambda_1, x_1]$  +
theta2 *  $\psi[\lambda_2, x_2]$  + theta3 *  $\psi[\lambda_3, x_3]$ , {x1, x2, x3},
{theta0, theta1, theta2, theta3,  $\lambda_1, \lambda_2, \lambda_3$ }]

```

Fig. 1. The *Mathematica* code for calculating the nonlinear model (1)-(2).

Figure 2 shows a comparative graphic of experimental data for laser output power *Pout* versus those estimated by the model (1) - (2). In particular the estimated value of the highest experiment *Pout* = 120W obtained by the model (1)-(2) is 122.639.

The main results from ANOVA are shown in Table 1. Further diagnostics of the model give maximum parameter-effects twice greater than the critical value of the 95% confidence region of the fit curvature. This corroborates that more appropriate model for our data is a nonlinear regression model rather than a linear one. Finally, high asymptotic correlation between parameters for all pairs is not observed, so that the model (1)-(2) is correct.

The calculations were carried out using double precision arithmetic on a dual core personal computer and took approximately 30 minutes.

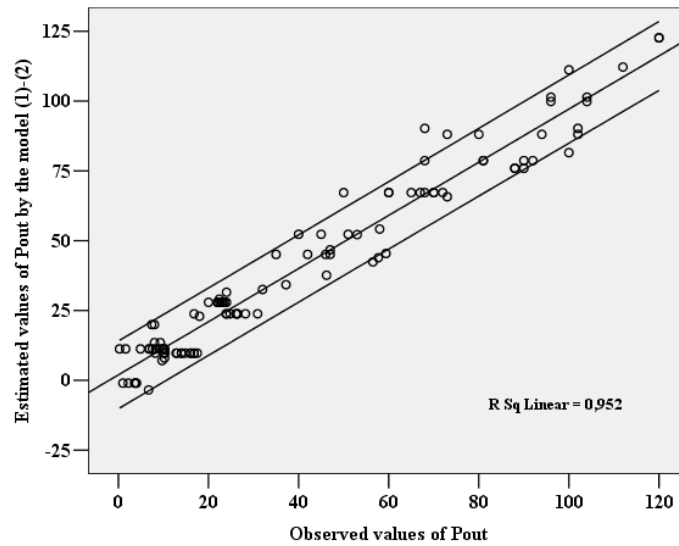
## 5 Assessment of the model predictive ability

In order to have a reliable estimate of the predictive ability of the nonlinear model (1), we apply the common practice to use data set independent of that used to fit the model. The initial data sample was split randomly into one training and one evaluation data set, containing approximately 70% and 30% of the total cases, respectively. The training data set was used to generate the model which was then tested with the independent evaluation data set.

The following are the parameters for the nonlinear regression model of type (1), for the randomly chosen 70% training data set from all 109 cases:

$$\theta_0 = 40.063269, \theta_1 = 26.973166, \theta_2 = 4.283957, \theta_3 = 11.859342, \quad (3)$$

$$\lambda_1 = 1.257025, \lambda_2 = 0.389093, \lambda_3 = 0.767572.$$



**Fig. 2.** The observed vs estimated values of laser generation *Pout*.

Predicted values for 30% evaluation data set compared to those already known for *Pout* are shown in Figure 3.

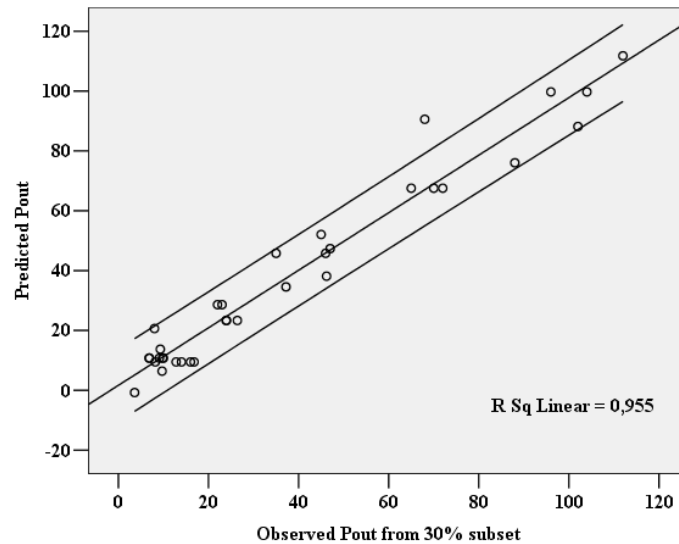
The basic statistics for the considered cases are given in Table 1.

Model	$R^2$ of the estimates	$R^2$ adj.	Std. Err.
Model (1), (2)	0.952	0.950	7.58979
Model (1), (3) for a 70% subset	0.950	0.948	7.94853
Model (1), (3) for a 30% subset	0.955	0.954	6.96404

**Table 1.** Basic statistics of the basic nonlinear regression model (1) and cross evaluation of 70% and 30% randomly extracted sets.

## 6 Discussion and conclusions

The comparison between constructed models is carried out on the basis of the quality of the calculated estimate values for laser output power and the results from the cross evaluation of the model. From the results given in Table 1 it is seen that the nonlinear model (1), (2) fits the data very well. Also, the indexes of model (1), (3) are relatively good and fall only a little behind those of (1), (2). The substituted in (1), (3) values from the 30% evaluation data set, which have not been included in the extraction of parameters (3) confirm the good quality of the constructed models. We can conclude that



**Fig. 3.** Predicted values for  $P_{out}$  compared to the initial observed values for a 30% evaluation data set.

nonlinear models of the suggested type are stable and fit the data well. The indexes of these estimates exceed the analogical statistics, obtained for the same data set using MLR. They are almost equal of the statistics from the second degree polynomial regression and fall behind the accuracy of the polynomial regression of the third degree and the MARS models based on linear regression splines and splines with first and second order interactions (see Gocheva-Ilieva and Iliev (2010)).

One can conclude that the obtained nonlinear regression model is fully applicable for estimation and prediction of the output laser power.

## Acknowledgements

This study was conducted with the financial support of the Scientific National Fund of Bulgarian Ministry of Education, Youth and Science, project number VU-MI-205/2006 and the Scientific Fund of Plovdiv University Paisii Hilendarski - NPD, projects RS2009-M-13 and IS-M-4.

## References

- ASTADJOV, D. N., DIMITROV, K. D., JONES, D. R., KIRKOV, V.K., LITTLE, C. E., LITTLE, N. and SABOTINOV, N. V. (1997a): Influence on operating characteristics of scaling sealed-off CuBr lasers in active length. *Optics Communications* 135, 289-294.

- ASTADJOV, D. N., DIMITROV, K. D., JONES, D. R., KIRKOV, V.K., LITTLE, C. E., LITTLE, N. and SABOTINOV, N. V. (1997b): Copper bromide laser of 120-W average output power. *IEEE Journal of Quantum Electronics* 33, 705-709.
- ASTADJOV, D. N., DIMITROV, K. D., LITTLE, C. E. and SABOTINOV, N. V. (1994): A CuBr laser with 1.4 W/cm<sup>3</sup> average output power. *IEEE Journal of Quantum Electronics* 30, 1358-1360.
- ASTADJOV, D. N., SABOTINOV, N. V., VUCHKOV, N. K. (1985): Effect of hydrogen on CuBr laser power and efficiency. *Optics Communications* 56, 279-282.
- DIMITROV, K. D. and SABOTINOV, N. V. (1996): High-power and high-efficiency copper bromide vapor laser. *SPIE* 3052, 126-130.
- FOSTER, P. G. (2005): *Industrial applications of copper bromide laser technology*. Ph.D. Thesis, University of Adelaide, School of Chemistry and Physics, Dept. of Physics and Mathematical Physics, Australia.
- FRIEDMAN, J. H. (1991): Multivariate adaptive regression splines (with discussion). *The Annals of Statistics* 19 (1), 1-141.
- GOCHEVA-ILIEVA, S. G. and ILIEV, I. P. (2010): Parametric and nonparametric empirical regression models: case study of copper bromide laser generation. *Mathematical problems in Engineering*, Article ID 582732, 16 p.
- ILIEV, I. P. and GOCHEVA-ILIEVA, S. G. (2007): Statistical techniques for examining copper bromide laser parameters. In: T. E. Simos, G. Psihoyios and Ch. Tsitouras (Eds.): *Proceedings of International Conf. of Numerical Analysis and Applied Mathematics, ICNAAM 2007, Corfu - Greece, Proc. AIP CP936*. Springer, New York, 267-270.
- ILIEV, I. P., GOCHEVA-ILIEVA, S. G. and SABOTINOV, N. V. (2008a): Statistical approach in planning experiments with a copper bromide vapor laser. *Quantum Electronics* 38 (5), 436-440.
- ILIEV, I. P., GOCHEVA-ILIEVA, S. G., ASTADJOV, D. N., DENEV, N. P. and SABOTINOV, N. V. (2008b): Statistical analysis of the CuBr laser efficiency improvement. *Optics and Laser Technology* 40 (4), 641-646.
- ILIEV, I. P., GOCHEVA-ILIEVA, S. G. and SABOTINOV, N. V. (2009): Classification analysis of CuBr laser parameters. *Quantum Electronics* 39 (2), 143-146.
- IZENMAN, A. J. (2008): *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, New York.
- JOLLIFFE, I. T. (1982): A note on the use of principal components in regression. *Journal of Royal Statistical Society, Series C (Applied Statistics)* 31, 300-303.
- NATO CONTRACT SFP (2000): 97 2685, 50W Copper Bromide laser.
- SABOTINOV, N. V. (2006): Metal vapor lasers. In: M. Endo and R.F. Walter (Eds.): *Gas Lasers*. CRC Press, Boca Raton, 449-494.
- SOLO (1993): *Computation with solo power analysis*. BMDP Statistical Software Inc., LA.
- STOILOV, V. M., ASTADJOV, D. N., VUCHKOV, N. K. and SABOTINOV, N. V. (2000): High spatial intensity 10 W- CuBr laser with hydrogen additives. *Optics and Quantum Electronics* 32, 1209-1217.
- YEO, I. K. and JOHNSON, R. A. (2000): A new family of power transformations to improve normality or symmetry. *Biometrika, Oxford Press* 87 (4), 954-959.

# Random Forests Based Feature Selection for Decoding fMRI Data

Robin Genuer<sup>1,2</sup>, Vincent Michel<sup>1,2,3,5</sup>, Evelyn Eger<sup>4,5</sup>, and Bertrand Thirion<sup>3,5</sup>

<sup>1</sup> Université Paris-Sud 11, Mathématiques, Orsay, France

<sup>2</sup> Select team, INRIA Saclay-Île-de-France, France

<sup>3</sup> Parietal team, INRIA Saclay-Île-de-France, France

<sup>4</sup> INSERM U562, Gif/Yvette, France

<sup>5</sup> CEA, DSV, I2BM, NeuroSpin, Gif/Yvette, France

**Abstract.** In this paper we present a new approach for the prediction of a behavioral variable from Functional Magnetic Resonance Imaging (fMRI) data. The difficulty in this problem comes from the huge number of image voxels that may provide relevant information with respect to the limited number of available images. A very common solution consists in using feature selection techniques, i.e. to evaluate the significance of each individual brain region with respect to the target information, and then to use the best ranked features as input to a classifier, such as linear Support Vector Machines (SVM; we take this as the *reference method*). However, this kind of scheme ignores the correlations between features, so that it is potentially suboptimal, and it does not generally provide an interpretable pattern of predictive voxels. Based on Random Forests, our approach provides an accurate auto-calibrated framework for selecting a set of very few jointly informative regions. Comparisons with the reference method on real data show that our approach yields a little bit higher classification performance, but the real gain comes from the sparsity of our variable selection.

**Keywords:** feature selection, variable importance, random forests, classification, fMRI

## 1 Introduction

A new way of analyzing neuroimaging data consists in assessing how well behavioral information or cognitive states can be predicted from brain activation images such as those obtained with functional magnetic resonance imaging (fMRI) (Cox and Savoy (2003)). This approach opens the way to understanding the mental representation of various perceptual and cognitive parameters. Indeed, certain neuronal populations are thought to activate specifically when a certain perceptual or cognitive parameter reaches a given value. The accuracy of the prediction of the target behavioral variable, as well as the spatial layout of predictive regions can provide valuable information about functional brain organization; in short, it helps to *decode* the brain system (Dayan and Abbott (2001)).

The main difficulty in this procedure is the huge dimensionality of the data, with far more features than samples. In this article, the samples will refer to the activation parameter maps resulting from a General Linear Model (GLM), the features being the voxel-based activation values. The large number of features leads to overfitting and thus to a dramatic decrease in prediction accuracy. Feature selection is thus mandatory, and is often performed by a mass-univariate selection based on F-test statistics. However, this classical approach is not well suited for neuroimaging as it does not cope with the multivariate structure of the data.

In order to improve the predictive framework, we introduce a new multivariate method of feature selection based on Random Forests (RF henceforth). RF is an increasingly used statistical method introduced in Breiman (2001). It gives outstanding results in prediction for lots of diverse applications. In addition, it computes a variable importance that can be used to select variables. Our RF-based algorithm uses the variable importance index in a feature selection framework. This variable selection procedure comes from Genuer et al. (2010), where one can find more information about RF variable importance.

After introducing the Random Forests and the RF-based algorithm, we show that our self-calibrated method performs an accurate feature selection, yielding a little bit better classification score than the reference technique, while keeping much less jointly informative variables. And this very sparse aspect of our variable selection method can help understanding functional brain organization. Let us finally emphasize that all along this paper, we distinguish two objectives: interpretation, which aims at selecting all the variables the most related to the response variable (even if they are correlated to each other); and prediction, which focuses on building a model involving the smallest subset of variables sufficient to make accurate predictions.

## 2 Methods

Let  $(Y_1, \dots, Y_n)$  represent the behavioural data to be fitted ( $\forall i, Y_i \in \{1, \dots, c\}$ , where  $c$  is the number of classes) related to a set of  $n$  parameter maps obtained with a GLM, where each image corresponds to one stimulus presentation;  $(X_1, \dots, X_n)$  are the  $m$ -dimensional activation maps ( $X \in \mathbb{R}^m$ ) and  $m$  is the number of features (voxels or parcels). In fMRI data, we have  $n \ll m$ , so that feature selection is mandatory.

### Random Forests

The principle of random forests is to aggregate many binary decision trees built on several bootstrap samples drawn from the learning set. The bootstrap samples are obtained by uniformly drawing  $n$  samples among the learning set with repetition. The decision trees are fully developed binary trees and the split rule is the following:

First, the whole dataset (also called the root of the tree) is split into two sub-



sets of data (called two children nodes). To do that, one randomly chooses a given number  $m_{try}$  of variables, and computes all the splits only for the previously selected variables. A split is of the form  $\{X^i \leq s\} \cup \{X^i > s\}$ , which means that data with the  $i$ -th variable value less than the threshold  $s$  go to the left child node and the others to the right one. Finally the selected split is the one leading to the most homogeneous children nodes (i.e. subsets associated to the same class).

Then, one restrains to one child node, randomly chooses another set of  $m_{try}$  variables and calculates the best split. And so on, until each node is a terminal node, i.e. it comprises observations associated with the same class.

A new data item  $X$ , starting in the root of the tree, goes down the tree following the splits and falls in a terminal node. Then the tree predicts for  $X$ , the common class  $\hat{Y}$  of the data in this terminal node. To finally get the RF classifier, one aggregates all the tree classifiers through a majority vote heuristic: for a new observation, each tree predicts a class and RF finally returns the most popular class.

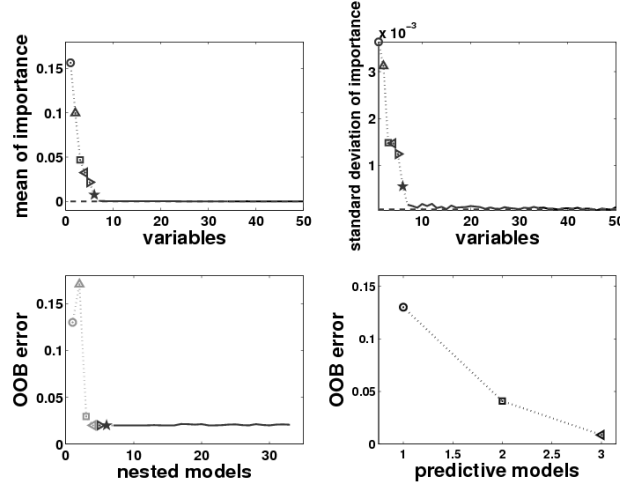
Inside the variable selection procedure, we use an estimation of the prediction error directly computed by the RF algorithm. This is the Out Of Bag (OOB) error and is calculated as follows. Fix one data in the learning sample, and consider all the bootstrap samples which do not contain this data (i.e. for which the data is “out of bag”). Now perform a majority vote only among trees built on these bootstrap samples. After doing this for all data, compare to the true classes and get an estimation of the prediction error (which is a cross-validated error estimate).

Let us now detail the computation of the RF variable importance for the first variable  $X^1$ . For each tree, one has a bootstrap sample associated with an OOB sample. Predict the OOB data with the tree classifier. Now, randomly permute the values of the first variable of the OOB observations, predict these modified OOB data with the tree classifier. The variable importance (VI henceforth) of  $X^1$  is defined as the mean increase of prediction errors after permutation. The more the error increases, the more important the variable is (note that it can be slightly negative, typically for irrelevant variables).

### Variable selection procedure

Let us give (following Genuer et al. (2010)) some details about the variable selection procedure that we use here. We apply it on a simulated learning set of size  $n = 100$  from the classification toys data model, introduced in Weston et al. (2003), with  $m = 200$ . It is an equiprobable two-class problem,  $Y \in \{-1, 1\}$ , with 6 true variables, the others being some noise.

The results are summarized in Figure 1. The true variables (1 to 6) are respectively represented by ( $\triangleright, \triangle, \circ, \star, \triangleleft, \square$ ). Based on to the learning set, we compute 50 forests with  $n_{tree} = 2000$  and  $m_{try} = 100$ , which are values of the main parameters considered as well adapted for VI estimation (for more details, see Genuer et al. (2010)).



**Fig. 1.** Variable selection procedure for a toy dataset. The top left graph shows the variable ranking. The curve of the top right graph is used to determine the threshold (represented by the horizontal dashed line) needed in Elimination step. OOB errors of the nested models are plotted in the bottom left graph to illustrate the Interpretation step. The bottom right graph stands for Prediction step.

Let us detail the four main steps of the procedure:

**Variable ranking:** First the variables are sorted according to the VI (averaged from the 50 runs) in descending order. Note that true variables are significantly more important than the noisy ones.

**Elimination step:** Keeping this order in mind, the corresponding standard deviations of VI are plotted. A threshold for importance is computed using this graph, and only the variables with an importance exceeding this level are kept. More precisely, the threshold is set as the minimum prediction value given by a CART model fitting this curve (for details, see Breiman et al. (1984)).

**Interpretation step:** Then, OOB error rates of the nested random forests models are computed; starting from the one with only the most important variables, and ending with the one involving all important variables kept previously. The set of variables leading to the smallest OOB error is selected.

**Prediction step:** Finally a sequential variable introduction with testing is performed: a variable is added only if the error gain exceeds a data-driven threshold (see Genuer et al. (2010)). The rationale is that the error decrease must be significantly greater than the average variation obtained by adding noisy variables.

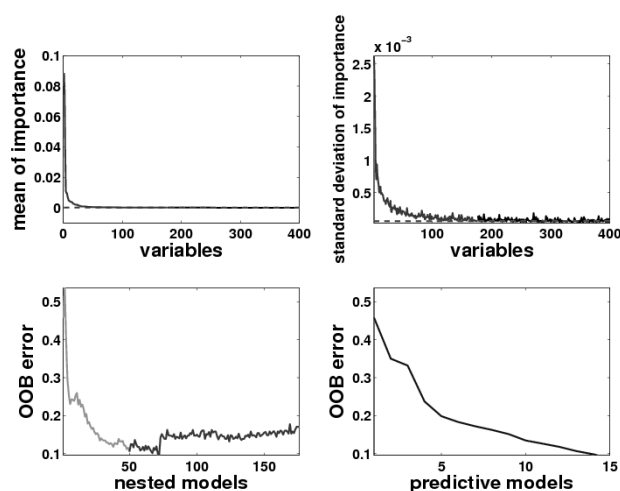
### 3 Experiments and Results

#### Real Data

We used a real dataset related to an experiment on the representation of objects Eger et al. (2008). During the experiment, twelve healthy volunteers

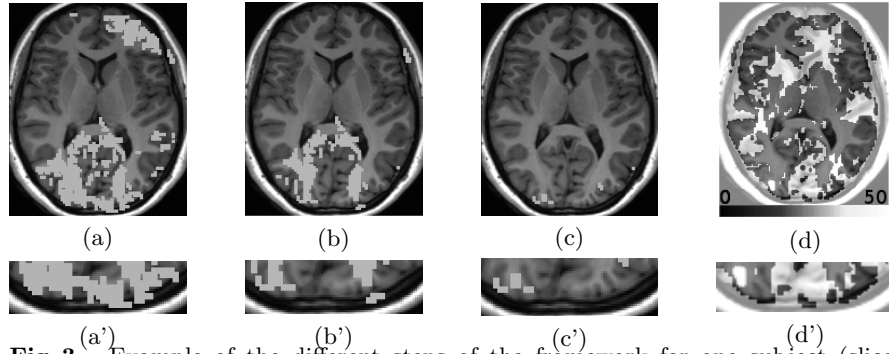
viewed objects of three different sizes and four different shapes, with 6 repetitions of each stimulus (referring to 6 sessions), resulting in a total of  $n = 72$  images by subject. Functional images were acquired on a 3-T MR system with eight-channel head coil (Siemens Trio, Erlangen, Germany) as T2\*-weighted echo-planar image (EPI) volumes. Twenty transverse slices were obtained with a repetition time of 2 s (echo time, 30 ms; flip angle, 70°;  $2 \times 2 \times 2$ -mm voxels; 0.5-mm gap). Realignment, normalization to MNI space and GLM fit were performed with the SPM5 software. For our analysis we used the resulting session-wise parameter estimate images. The four different shapes of objects are pooled across the three sizes, and we are interested in discrimination between shapes. We used parcellation as a preprocessing, which allows important unsupervised reduction of dimensions. Parcellation uses Ward's algorithm (hierarchical agglomerative clustering) to create groups of voxels which have similar activity across trials. Thus, the signal is averaged in each parcel. The number of parcels created is fixed to 1000 for the whole brain.

### Feature selection results for one subject



**Fig. 2.** Variable selection procedure for one subject. The graphs follow the exact same description as in Figure 1.

We apply the procedure described in Section 2 for the subject 2 of the study. The results are plotted in Figure 2. The horizontal dotted line of the top graphs indicates the threshold, computed using standard deviations of VI (see the top right graph) and used in the top left graph to eliminate variables of small importance. Starting with all the 1000 variables, this elimination step retains 176 variables. The minimum OOB error rate in the bottom left graph is obtained by the RF model involving 50 variables, which constitute the interpretation set. Finally, the prediction procedure, illustrated by the bottom right graph, selects 15 variables.



**Fig. 3.** Example of the different steps of the framework for one subject (slice  $z=6$  mm). (a) Selected parcels after Elimination Step. (b) Selected parcels after Interpretation Step. (c) Selected parcels after Prediction Step. (d) Shows the parcels selected by the reference method, and their F-test values. (a'), (b'), (c') and (d') are magnifications of the occipital part.

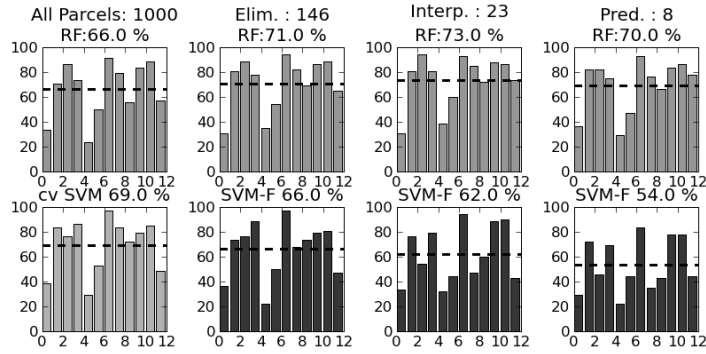
Figure 3 shows the selected parcels for the different steps of the algorithm in one axial slice for subject 2: sub-figures (a), (b) and (c) represent the variables selected in the Elimination step, Interpretation step and Prediction step, and (d) represent the variables selected by the reference method. Sub-figures (a'), (b'), (c') and (d') are magnifications of the occipital part. During the interpretation step, our algorithm keeps only three regions of the occipital cortex, reducing the features to a much smaller sets while keeping an accurate prediction (see Figure 4). In addition, the prediction step (c) allows to avoid redundancy in the features. The selected regions are different between the two hemispheres, while the interpretation step retained more symmetric regions. Finally, comparison with sub-figure (d) highlights the most beneficial aspect of our method: we select very localised informative regions, while the reference method keeps lots of regions distributed in all brain.

### Prediction results for the whole data

We perform a leave-one-session-out cross-validation: we successively train the classifier with all the sessions except one, and report the performance of the trained classifier on the left out session. Importance-based feature selection was applied independently on the twelve datasets. The results are shown in Figure 4. The first row represents the classification score of RF for each subject (from left to right: all parcels, after Elimination step, after Interpretation step and after Prediction step). The average number of selected parcels across subjects is noted above each histogram, with the average classification score across all subjects.

The first graph of the second row shows the results of a cross-validated linear SVM: the optimal number of parcels to be kept (from 50 to 1000 parcels with a step of 50) for the linear SVM is selected using the F-statistic, by leave-one-out validation on the training set. The average number of selected

parcels across subjects is equal to 350. The three last histograms of the second row show the results of a linear SVM: the parcels are selected by using a F-statistics, and the number of features used is equal to the number of parcels found by the three different steps of the RF-based algorithm. We can see that our algorithm gives better accuracy for the three steps of selection than the reference method (cross-validated linear SVM). And the three last histograms of the second row illustrate the fact that a linear SVM (coupled with F-test) do not manage to keep good accuracy with as few features as selected by our method.



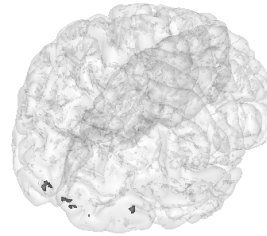
**Fig. 4.** Results on real data: rate of correct object identification, using the mean signal of 1000 parcels of the brain volume (chance level=25 %). The first row shows the prediction accuracy in each individual dataset, the mean classification score and the number of selected parcels for (from left to right) the whole brain, the Elimination step, the Interpretation step and the Prediction step. The first graph of the second row represents the results for the reference method. The three last histograms of the second row show the prediction accuracy of a linear SVM trained with the same number of parcels as above, but selected by F-statistics.

## 4 Discussion

This work presents the first application of a RF-based feature selection technique to brain state decoding. We show that it is competitive with state of the art method (univariate selection followed by linear SVM classifier). More importantly, the insensitivity of the correct classification rates along the different steps of feature reduction that is observed in Figure 4 for the RF model shows that our strategy manages to extract the statistical information of the data: it keeps much of the information while significantly reducing the dimension. This suggests that the multivariate RF variable importance index performs better than the classical univariate F-test score to detect the most predictive variables. Another noticeable aspect of the proposed procedure is that it is entirely data-driven: at each step of the procedure, thresholds are computed using only the data. So this procedure can adapt to lots of different applications, without the need of adding prior information (like e.g. a number of variables to be selected).

**Fig. 5.** Regions selected in at least 3 subjects among 12 by the last step of the RF-based selection. The MNI coordinates are :

[18, -102, 6]mm  
 [10, -100, 4]mm  
 [-12, -96, 0]mm  
 [50, -78, 6]mm.



From a neuroscientific point of view, we can notice that the spatial distribution of the selected parcels is quite informative: first, by avoiding redundancy, the algorithm is able to focus on few extremely precise regions of the brain without loss of accuracy. Moreover, starting from whole brain, the algorithm selects very few parcels in the occipital cortex, corresponding to visual areas. If we look at the regions selected for 3 subjects or more among the 12 subjects by the last step of the RF-based selection (see Figure 5), there are only few regions in the early visual cortex, and a slightly more anterior parcel. This is consistent with the fact that early visual cortex contains highly reliable signals discriminative of feature/shape differences between object exemplars, as long as no generalization across image changes is required (Cox and Savoy (2003) and Eger et al. (2008)).

**Conclusion** In this article, a multivariate and threshold-free feature selection algorithm based on Random Forests, yields an accurate selection for fMRI data analysis, and creates a highly informative set of very few features. Results on real data show the benefits of our approach for both interpretation and prediction, with higher accuracy and higher sparsity than the reference method.

## References

- BREIMAN, L. (2001): Random Forests. *Machine Learning* 45 ,5-32.
- BREIMAN, L., FRIEDMAN, J., OLSHEN, R.A. and STONE, C.J. (1984): *Classification and Regression Trees*. Chapman & Hall.
- COX, D.D. and SAVOY, R.L. (2003): Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* 19 (2),261-270.
- DAYAN, P. and ABBOTT, L.F. (2001): *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press.
- EGER, E., KELL, C. and KLEINSCHMIDT, A. (2008): Graded size sensitivity of object exemplar evoked activity patterns in human LOC subregions. *Journal of Neurophysiology* 100 (4) , 2038-47.
- GENUER, R., POGGI, J.-M. and TULEAU, C. (2010): Variable selection using random forests. *Pattern Recognition Lett.* doi:10.1016/j.patrec.2010.03.014
- WESTON, J., ELISSEEF, A., SCHOLKOPF, B., TIPPING, M., KAEHLING, P. (2003): Use of the Zero-Norm with Linear Models and Kernel Methods. *Journal of Machine Learning Research* 3 , 1439-1461.

# Differentiation Tests for the Mean Shape and the Mean Variance of Renal Tumours appearing in early Childhood

Stefan Markus Giebel<sup>1</sup>, Jens-Peter Schenk<sup>2</sup> and Jang Schiltz<sup>1</sup>

<sup>1</sup> Université du Luxembourg

4, rue Albert Borschette, L-1246 Luxembourg, *jang.schiltz@uni.lu*

<sup>2</sup> Uniklinikum Heidelberg

Im Neuenheimer Feld 430, D-69120 Heidelberg,

*jens-peter.schenk@med.uni-heidelberg.de*

**Abstract.** There are different kinds of renal tumours that can appear in childhood: nephroblastoma, clear cell sarcoma, neuroblastoma etc. The chosen therapy depends upon the diagnosis of the radiologist which is done with the help of MRI (Magnetic resonance images). We present a mathematical treatment of the MRI of renal tumours ( $n = 80$ ) to help the radiologist with his. We are using transversal, frontal and sagittal images and compare their potential for differentiation of the different kind of tumours by use of Statistical Shape Analysis. After a quick overview of the results of the the classical mean shape test from Ziezold (1994), we introduce a new test for the comparing of the mean variance in the sense of Fréchet of two groups of shapes and discuss the medical implications of our results.

**Keywords:** statistical shape analysis, renal tumours, mean shape, variance comparison

## 1 Introduction

In a wide variety of disciplines it is of great practical importance to measure, describe and compare the shapes of objects. In general terms, the shape of an object, data set, or image can be defined as all the information that is invariant under translation, rotation and isotropic rescaling. The field of shape analysis involves hence methods for the study of the shape of objects where location, rotation and scale can be removed. The two- or more dimensional objects are summarised according to key points called landmarks. This approach provides an objective methodology for classification whereas even today in many applications the decision for classifying according to the appearance seems at most intuitive.

Statistical shape analysis is concerned with methodology for analysing shapes in the presence of randomness. It is a mathematical procedure to get the information of two- or more dimensional objects with a possible correction of size and position of the object. So objects with different size and/or position can be compared with each other and classified. To get the shape of an object

without information about position and size, centralisation and standardisation procedures are used in some metric space.

Interest in shape analysis began in 1977. D.G. Kendall (1977) published a note in which he introduced a new representation of shapes as elements of complex projective spaces. K.V. Mardia (1977) on the other hand investigated the distribution of the shapes of triangles generated by certain point processes, and in particular considered whether towns in a plain are spread regularly with equal distances between neighbouring towns. The full details of this elegant theory which contains interesting areas of research for both probabilists and statisticians were published by D. Kendall (1984) and F. Bookstein (1986). The details of the theory and further developments can be found in the textbooks by C.G. Small (1996) and I.L. Dryden and K.V. Mardia (1998).

In this paper, we describe one interesting application of statistical shape analysis: the classification of renal tumours. In contrast to many applications called also “Shape Analysis” (Favero and Soatto (2007)) we have to determine a mean shape, representative for a group of objects, and not only to detect an already known shape. Since the renal tumour is limited by spleen or liver, the rest of the kidney, the spine and retroperitoneal vessels, and all our images are taken from the same direction, we do not have to bother here about rotation and can use the Euclidean distance. Giebel et al. (2010) showed that none of the landmarks has a special influence for the determining of the mean shape according to the independence test of Ziezold (2003). In Giebel et al. (2009) we presented the results of a study of 74 tumours and showed that statistical shape analysis can be a very useful tool for the differentiation of different kinds of tumour.

In this paper we introduce the concept of mean variance in the sense of Fréchet of a set of objects and we introduce a test that allows to compare the mean variance of two groups of shapes. So far, the concept of variance has never been used in statistical shape analysis and we think that it is quite useful to get more precise comparison results.

## 2 Renal tumours appearing during childhood

Nephroblastoma (Wilms (1889)) is the typical tumour of the kidneys appearing in childhood. Therapy is organized in therapy-optimizing studies of the Society of Paediatric Oncology and Haematology (SIOP). Indication of preoperative chemotherapy is based on radiological findings. The preferred radiological method is sonography and MRI. Both methods avoid radiation exposure, which is of great importance in childhood. Preoperative chemotherapy is performed without prior biopsy (Schenk (2006)).

Information of the images of magnetic resonance tomography, especially the renal origin of a tumour and the mass effect with displacement of other organs, is needed for diagnosis. Beside nephroblastomas other tumours of the



retro peritoneum exist, which are difficult to differentiate (Schenk (2008)). Renal tumours in childhood are classified in three stages of malignancy (I, II, III). Typical Wilms tumours mostly belong in stage II. In stage II different subtypes of nephroblastoma tissue exist (Graf(2003)).

In our sample of tumours in childhood, there are four different types of retroperitoneal tumours: nephroblastoma, neuroblastoma, clear cell carcinoma, and renal cell carcinoma. Renal cell carcinomas are very rare in childhood. They represent the typical tumours of adult patients. They have no high sensitivity for chemotherapy. Clear cell sarcomas are very rare in childhood and are characterized by high malignancy. Neuroblastoma are the typical tumours of the sympathetic nervous system and suprarenal glands. Infiltration of the kidney is possible.

The tumour grows with encasement of vessels. Because of the high importance of radiological diagnosis for therapy, it is of great interest to find markers for a good differentiation of tumours. MRI produces 2D images. We then compute a 3D image from the two dimensional data.

### 3 The mean shape

To compare the standardised and centred sets of landmarks, we have to define the mean shape of all the objects and a distance function which allows us to evaluate how far away every object is from this mean shape.

The term “mean” is here used in the sense of Fréchet (1948). If  $X$  denotes a random variable defined on a probability space  $(\Omega, \mathcal{F}, P)$  with values in a metric space  $(\Xi, d)$ , an element  $m \in \Xi$  is called a mean of  $x_1, x_2, \dots, x_k \in \Xi$  if

$$\sum_{j=1}^k d(x_j, m)^2 = \inf_{\alpha \in \Xi} \sum_{j=1}^k d(x_j, \alpha)^2. \quad (1)$$

That means that the “mean shape” is defined as the shape that guarantees the smallest possible variance for a group of objects. For computing the mean shape we use the algorithm of Ziezold (1994). It is however easy to show that in the case of the euclidian distance the mean shape of a set of objects is identical to their arithmetic mean. In our application, it is hence sufficient to compute it like that.

In the special case of oncology there is no theoretical medical reason to select a specific group of landmarks for differentiation. All landmarks in this research have thus to be selected by an explorative procedure.

The test of Ziezold (1994) is a statistical test which allows to determine if a given object belongs to a set of objects defined by their mean shape. We have used this test to see if given Wilm’s tumours can be differentiated from the mean shape of the neuroblastomas and vice versa.

## 4 Mean variance of a set of shapes

We define the mean variance in the sense of Fréchet of a set of objects as the average of the distances to the mean shape.

If  $X$  denotes a random variable defined on a probability space  $(\Omega, \mathcal{F}, P)$  with values in a metric space  $(\Xi, d)$  and  $m \in \Xi$  is the mean of  $x_1, x_2, \dots, x_k \in \Xi$ ,  $\sigma^2$  is the variance of  $x_1, x_2, \dots, x_k \in \Xi$  if

$$\sum_{j=1}^k d(d(x_j, m)^2, \sigma^2)^2 = \inf_{\alpha \in \Xi} \sum_{j=1}^k d(d(x_j, m)^2, \alpha)^2. \quad (2)$$

That means that the variance is defined as the mean of the distances between the "mean shape" and the objects.

## 5 The variance test

In this section we propose a test to compare the mean variance of two groups of objects. It functions analogously to the test of Ziezold (1994):

### step 1: Definition of the set of objects

There is one set  $M = \{o_1, \dots, o_N\}$  that can be divided into two subsets: objects with the characteristics A:  $A^{sample} = \{o_1, \dots, o_n\} = \{a_1, \dots, a_n\}$  and objects with the characteristics B:  $B^{sample} = \{o_{n+1}, \dots, o_N\} = \{b_1, \dots, b_{N-n}\}$ . The subset A is a realisation of a distribution  $P$  and the subset B is an independent realisation of a distribution  $Q$ .

$$\text{Hypothesis:} \quad H_0 : \sigma_1^2 = \sigma_2^2$$

$$\text{Alternative:} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

Define the *level of significance*  $\alpha$ . If the probability for  $H_0$  is smaller, we neglect  $H_0$  and assume  $H_1$ .

### step 2: Computing the variance

The variance is calculated by means of a straightforward generalisation of the algorithm of Ziezold (1994). Let  $\sigma_1^2$  denote the variance of the subset A.  $\sigma_2^2$  is then computed for the subset B.

### step 3: Computing the $F$ -value

$$F = \frac{|\hat{\sigma}_1^2|}{|\hat{\sigma}_2^2|}.$$

**step 4: Determination of all the possibilities of dividing the set into two subsets with given sizes**

**step 5: Comparing the  $F$ -value to all possible  $F$ -values. Computing the rank (small  $F$ -value mean a small rank).**

**step 6: Calculate the  $p$ -value for  $H_0$**

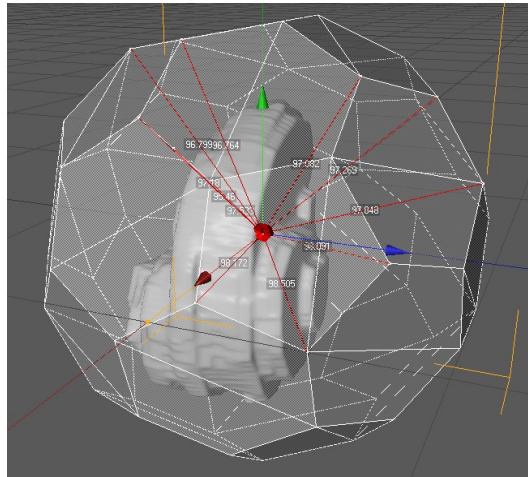
$p_{r=i} = 1 - \frac{1}{\binom{N}{n}}$  for  $i = 1, \dots, \binom{N}{n}$ , where  $r$  is the rank for which we assume a rectangular distribution on the right side and  $p_{r=i} = \frac{1}{\binom{N}{n}}$  on the left side.

## 6 Numerical application

There appear about 130 Wilms tumours and 84 neuroblastomas with abdominal origin every year in Germany. For about 75% of the patients MRI are used. The University of Heidelberg gets about 100 patients every year for reference radiology. In practices, more than half of the patients can not be considered because of bad image quality or an insufficient number of images to construct a useful 3D image. Our research sample consisted of 74 comparable tumours (69 nephroblastoma and 5 neuroblastoma). To get 3D landmarks we construct a three dimensional object of the tumour from the 2D MRI. Then we take the intersection between the surface of the tumour and the vectors going from the centre to the edges of the platonic body C60 as landmarks as is shown in figure 1.

The test of Ziezold (1994) for the differentiation of the mean shape gives the following result: We get an  $u_0$ -value of 72 when comparing the neuroblastomas to the mean shape of the nephroblastoma. According to the rank in a randomized sample ( $n = 1000$ ), this gives a  $p$ -value of 0.116. When comparing the nephroblastomas to the mean shape of neuroblastomas we get an  $u_0$  value of 112, which corresponds to a the  $p$ -value of 0.080. It is hence easier to distinguish a nephroblastoma from neuroblastoma than vice versa.

The result of the test of Ziezold (1994) could be a consequence of different variances in the two groups. Our variance test allows to test this. For the renal tumours, the  $F$ -value for the differentiation of the variance of the group of nephroblastomas to the group of neuroblastomas is 1.28128 and the rank is 315. So the corresponding  $p$ -value is  $1 - 0.315 = 0.685$  and we have to accept the null hypothesis that the variance is similar in both groups. So both kind of tumours seem to have more or less the same dispersion and a possible difference in the dispersion can be excluded as cause for difficulties in distinguishing the two kinds of tumours.



**Fig. 1.** 3D-Landmarks as cut points between the edge of a platonic body and the surface of the tumor.

## 7 Conclusion

We have introduced the concept of mean variance of a set of objects in shape analysis and exhibited a test to compare the mean variance of two groups of objects. Both the test of Ziezold and our variance test are permutation tests. They do not need any assumptions concerning the distributions and the size of the sample.

Our medical application shows that Wilms tumors can be clearly differentiated from neuroblastomas. It is moreover possible to differentiate the whole set of non-Wilms tumors from the mean shape of Wilms tumors. But we cannot use statistical shape analysis to say if a given general tumor is not a Wilms tumor.

For improving our results, non-Euclidean transformations could be considered. A possible approach is to use a supervised 1-layer neural network with weighted landmarks. Instead of the difference between output and reality we will consider the distance between the “mean shape” and the objects.

## References

- BOOKSTEIN, F.L. (1986): Size and shape spaces for landmark data in two dimensions. *Statistic Sciences* 1, 181-242.
- DRYDEN, I.L. and MARDIA, K.V. (1998): *Statistical Shape Analysis*. Wiley, Chichester.
- FAVORO, P. and SOATTO, S. (2007): *3D-Shape Estimation and Image Restoration*. Springer, Berlin.
- FRECHET, M. (1948): Les éléments aléatoires de nature quelconque dans un espace distancié, *Annales de l'Institut Henri Poincaré* 10, 215-310.

- GIEBEL, S. (2007): *Statistical Analysis of the shape of renal tumours in childhood*. Diploma thesis, University of Kassel.
- GIEBEL, S., SCHENK, J.-P. and SCHILTZ, J. (2009): *Application of Statistical Shape Analysis on renal tumours into the Classification of three dimensional renal tumours appearing in early childhood*. In: *Proceedings of the the 20th international congress of Jangjeon Mathematical Society*.
- GIEBEL, S., SCHENK, J.-P. and SCHILTZ, J. (2010): *Shape Analysis of retroperitoneal tumors in childhood in Magnetic Resonance Imaging*. In: *Bulletin de la Société des Sciences Médicales du Grand-Duché du Luxembourg*. to appear
- GRAF, N. (2003): *Urologe A* 43:421.
- KENDALL, D.G. (1977): The diffusion of shape. *Adv. Appl. Probab.* 9(3), 428-430.
- KENDALL, D.G. (1984): Shape manifolds, Procrustean metrics and complex projective spaces. *Bulletin of the London Mathematical Society* 16, 81-121.
- MARDIA, K.V. (1977): Mahalanobis distance and angles. In: P.R. Krishnaiah (Ed): *Multivariate Analysis IV*, Amsterdam: North Holland, 495-511.
- SCHENK, J.P (2006): *Fortschr. Rntgenstr.* 178:38.
- SCHENK, J.P (2008): *Eur. Radiol.* 18:683.
- SMALL, C.G. (1996): *The Statistical Theory of Shape*. Springer-Verlag, New York.
- WILMS, M. (1889): *Die Mischgeschwulste der Niere*. Verlag von Arthur Georgi, Leipzig.
- ZIEZOLD, H. (1994): Mean Figures and Mean Shapes Applied to Biological Figure and Shape Distributions in the Plane *Biometrical Journal* 36, 491-510.
- ZIEZOLD, H. (2003): Independence of Landmarks of Shapes. *Mathematische Schriften Kassel* 3.



# Local or Global Smoothing? A Bandwidth Selector for Dependent Data

Francesco Giordano<sup>1</sup> and Maria Lucia Parrella<sup>1</sup>

University of Salerno - Department of Economics and Statistics  
Via Ponte Don Melillo, 84084 Fisciano (SA), Italy  
*giordano@unisa.it, mparrella@unisa.it*

**Abstract.** The selection of the smoothing parameter represents a crucial step in local polynomial regression, due to the implications on the consistency of the non-parametric estimator and to the difficulties in the implementation of the selection procedure. In order to capture the complexity of the unknown regression curve, a local variable bandwidth may be used, but this may increase the variability of the estimates and the computational costs. This paper focuses on the problem of estimating the smoothing parameter adaptively on the support of the function, after evaluating the effective gain in using a local bandwidth rather than a global one.

**Keywords:** kernel regression, variable bandwidth selection, dependent data.

## 1 Context and aims

Kernel based estimators represent one of the most popular nonparametric tools for the estimation of a regression function. The good asymptotic properties of such estimators (see Fan & Gijbels (1996); Masry & Fan (1997)) are often challenged by the misspecification of the tuning parameter, *i.e.* the bandwidth of the kernel function. The difficulties in specifying such a tuning parameter may do vanish at all the advantages in using these nonparametric tools. Moreover, a local bandwidth may be preferred in order to take account of the local features of the unknown function. Anyway, the difficulties in estimating such local bandwidth may vanish, at least in part, the gain in the estimations. Therefore, if the unknown regression function has a simple structure, the use of a global bandwidth on the whole support of the function may represent the best solution. The aim of this work is to propose a new procedure for the automatic selection of the smoothing parameter. The latter is based on a preliminary evaluation of the opportunity to use a local bandwidth rather than a global bandwidth on a given subset. Just as an example of application, we focus on the problem of estimating the volatility function of dependent data, although our procedure can be adapted to other contexts.

In the following section we set the assumptions and the notation. Our procedure, which is organized in two stages, is described in detail in section 3. The theoretical properties of the procedure are still under investigation. A

simulation study has been performed, showing the good performance of the procedure. The results are reported in Giordano and Parrella (2009).

## 2 Setup of the local polynomial estimators

Consider the process  $\{Y_t, X_t\}$ , where  $X_t$  and  $Y_t$  are real valued observed processes. Define the following nonparametric regression model

$$Y_t = m(X_t) + \sigma(X_t)\varepsilon_t, \quad t = 1, 2, \dots, \quad (1)$$

where the errors  $\varepsilon_t$  are real random variables independent from  $X_t$ , for which  $E(\varepsilon_t) = 0$  and  $Var(\varepsilon_t) = 1$ , for all  $t$ . Given model (1), we consider the generic problem of estimating the conditional regression function

$$m_\phi(x) = E\{\phi(Y_t)|X_t = x\}, \quad \forall x \in \mathbb{R}, \quad (2)$$

which includes several special cases by suitable definition of the function  $\phi(\cdot)$  (conditional moment functions, conditional distribution functions, etc.). Given a realization of the process  $\{Y_t, X_t; t = 1, \dots, n\}$ , the unknown function  $m_\phi(x)$  and its first  $p$  derivatives can be estimated nonparametrically using the local polynomial estimators of degree  $p$ , assuming that the derivative of order  $p+1$  exists (Fan & Gijbels (1996)). Moreover, model (1) can be extended to the nonparametric ARCH model by putting  $Y_t = X_{t+1}$ . In this paper we consider in particular the last case of autoregressive data, for which we denote with  $\mu_X$  its stationary distribution and we consider these assumptions:

(a1) the  $\varepsilon_t$  have continuous and positive density function and, for some  $\delta > 4$ ,

$$E(\varepsilon_t^2) = 1, \quad E(\varepsilon_t) = E(\varepsilon_t^3) = 0, \quad E|\varepsilon_t|^\delta < \infty;$$

(a2) the functions  $m(\cdot)$  and  $\sigma(\cdot)$  have continuous second derivative. Moreover, the function  $\sigma(\cdot)$  is positive;

(a3) there exist the constants  $M_1 > 0$  and  $M_2 > 0$  such that, for all  $y \in \mathbb{R}$ ,

$$|m(y)| \leq M_1(1 + |y|), \quad |\sigma(y)| \leq M_2(1 + |y|), \quad M_1 + M_2 [E|\varepsilon_t|^\delta]^{1/\delta} < 1;$$

(a4) the density function  $f_X(\cdot)$  of  $\mu_X$  exists, it is bounded, continuous and positive on every compact set in  $\mathbb{R}$ .

Under the assumptions (a1)-(a4), it can be shown that the process is geometrically ergodic (see Härdle & Tsybakov (1997)). The value of  $\delta$  must be fixed also considering the particular function  $\phi(\cdot)$  in eq. (2), in order to guarantee that  $E|\phi(Y_t)|^2 < \infty$ . Let denote with  $\hat{m}_\phi^{(v)}(x; h)$  the LP estimator (of degree  $p$ ) for the function  $m_\phi^{(v)}(x)$ , where  $h$  is the smoothing parameter and  $v$  is the degree of the derivative ( $v = 0$  for the function itself,  $p = 1$  for local linear estimator, etc.). The asymptotic mean squared error of the estimator is

$$AMSE\{\hat{m}_\phi^{(v)}(x; h)\} = \mathbb{B}^2(x)h^{2(p+1-v)} + \mathbb{V}(x)h^{-(2v+1)}, \quad \forall x \in \mathbb{R}, \quad (3)$$



where

$$\mathbb{B}^2(x) = \left\{ C_1 m_\phi^{(p+1)}(x) \right\}^2 \quad \mathbb{V}(x) = \frac{C_2 \sigma_\phi^2(x)}{n f_X(x)}. \quad (4)$$

Note that the only unknown components in the (4) are the variance function  $\sigma_\phi^2(x) = \text{Var}\{\phi(Y_t)|X_t = x\}$ , the derivative function  $m_\phi^{(p+1)}(x)$  and the design density  $f_X(x)$ . The constant values  $C_1$  and  $C_2$  depend on known quantities, such as the kernel function,  $p$  and  $v$ . See Fan and Gijbels (1995) for the details and some examples. The *asymptotically optimal local bandwidth* is the bandwidth which minimizes the right-hand side of eq. (3). We denote such bandwidth with  $h_{AMSE}^{opt}(x)$ . It is given by

$$h_{AMSE}^{opt}(x) = \left\{ \frac{(2v+1)\mathbb{V}(x)}{2(p-v+1)\mathbb{B}^2(x)} \right\}^{1/(2p+3)} \quad \forall x \in \mathbb{R}. \quad (5)$$

The *plug-in* method derives an estimation of the optimal bandwidth by estimating the unknown functionals  $\mathbb{V}(x)$  and  $\mathbb{B}^2(x)$ , and plugging them into equation (5). Note that these functionals and the optimal bandwidth  $h_{AMSE}^{opt}(x)$  depend on the degree of the estimated derivative  $v$  by known constants, so the optimal bandwidth for the estimation of the function  $m_\phi(x)$  or for the estimation of some derivative  $m_\phi^{(v)}(x)$  can be computed in the same way. For this reason, we consider here only the case  $v = 0$ .

We may also consider an integrated measure of the AMSE. By minimizing it, we derive the *asymptotically optimal global bandwidth* as

$$h_{AMISE}^{opt} = \left\{ \frac{(2v+1)\mathbb{V}_\omega}{2(p-v+1)\mathbb{B}_\omega^2} \right\}^{1/(2p+3)}, \quad (6)$$

where, for a given weight function  $\omega(x)$ ,

$$\mathbb{B}_\omega^2 = C_1^2 R_f(m_\phi^{(p+1)}), \quad \mathbb{V}_\omega = \frac{C_2 R(\sigma_\phi)}{n} \quad (7)$$

$$R_f(m_\phi^{(p+1)}) = \int [m_\phi^{(p+1)}(x)]^2 \omega(x) d\mu_X \quad R(\sigma_\phi) = \int \sigma_\phi^2(x) \omega(x) dx. \quad (8)$$

The following relation holds,  $\forall x \in \mathbb{R}$ , for the asymptotic mean squared error

$$AMSE\{\hat{m}_\phi^{(v)}(x; h_{AMSE}^{opt}(x))\} \leq AMSE\{\hat{m}_\phi^{(v)}(x; h_{AMISE}^{opt})\}, \quad (9)$$

so the global bandwidth is suboptimal when used for local estimations. Anyway, considering the particular structure of model (1), the difference between the two terms in the (9) may be very little sometimes, and in such case the relation (9) may not hold anymore when we substitute the optimal bandwidths  $h_{AMSE}^{opt}(x)$  and  $h_{AMISE}^{opt}$  with the estimated ones. Moreover, the two bandwidth estimators may have different estimation rates, since the estimation of the (5) is generally more difficult than the estimation of the (6). Therefore,

notwithstanding the relation (9), one could ask if there is an effective gain in using a (estimated) local bandwidth  $h_{AMSE}^{opt}(x)$  instead of a (estimated) global bandwidth  $h_{AMSE}^{opt}$ . This consideration will be taken into account in the first stage of our procedure, basing on a preliminary evaluation of the structure of  $AMSE\{\hat{m}_\phi^{(v)}\}$ .

### 3 A two stage bandwidth selector for the estimation of the volatility function.

We propose a two stage procedure. In the first stage we evaluate the effective gain in using a local bandwidth for the estimation of the function  $m_\phi(x)$ , for  $x \in \mathbb{R}$ , instead of a global bandwidth. Eventually, we split the support of the function  $m_\phi(x)$  into a number of compact subsets. In the second stage we estimate the “*locally global bandwidths*”, that are global bandwidths which are estimated on each one of the *homogeneous* subsets through a global optimization.

#### 3.1 First stage: evaluating the “homogeneity” on the support.

Suppose for now that we want to estimate the function  $m_\phi(x)$ , for each  $x$  belonging to a subset  $I_X \subseteq \mathbb{R}$ . We suggest the following approach:

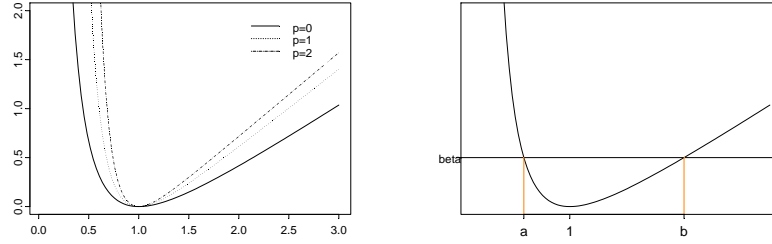
- consider the optimal global bandwidth  $h_{AMSE}^{opt}$  and the local bandwidth  $h_{AMSE}^{opt}(x)$ , both defined on the subset  $I_X$ . For simplicity of notation we denote the first one with  $h_{glob}$  and the second one with  $h_{loc}(x)$ ;
- use some relative indicator in order to evaluate the gain in using the local bandwidth  $h_{loc}$  instead of the global bandwidth  $h_{glob}$  on the subset  $I_X$ ;
- if there is an effective gain of using a more refined bandwidth, split the subset  $I_X$  into (two or more) subsets and repeat the first two steps on each one of the subsets.

The problem now is how to derive the relative indicator in step 2, how to evaluate and estimate it on the whole subset  $I_X$ , and then how to determine the stopping rule in step 3. Starting from the relation shown in (9), consider the relative variation of the  $AMSE$  observed when we use a global bandwidth on the subset  $I_X$  instead of a local bandwidth

$$\begin{aligned}\Delta_{AMSE}(x) &= \frac{AMSE\{\hat{m}_\phi(x; h_{glob})\} - AMSE\{\hat{m}_\phi(x; h_{loc}(x))\}}{AMSE\{\hat{m}_\phi(x; h_{loc}(x))\}} \\ &= \frac{AMSE\{\hat{m}_\phi(x; h_{glob})\}}{AMSE\{\hat{m}_\phi(x; h_{loc}(x))\}} - 1, \quad \forall x \in I_X.\end{aligned}\quad (10)$$

The minimum value for eq. (10) is zero, and it is observed when there is no gain in using the local bandwidth. Remembering (3), we can write

$$\Delta_{AMSE}(x) = \left[ \frac{\mathbb{B}^2(x)h_{glob}^{2(p+1)}}{\mathbb{V}(x)h_{loc}^{-1}(x)} + \frac{h_{glob}^{-1}}{h_{loc}^{-1}(x)} \right] \left[ \frac{\mathbb{B}^2(x)h_{loc}^{2(p+1)}(x)}{\mathbb{V}(x)h_{loc}^{-1}(x)} + 1 \right]^{-1} - 1 \quad (11)$$



**Fig. 1.** *Left:* plot of the function  $f(z) = 1/(2p+3)z^{-2(p+1)} + (2p+2)/(2p+3)z - 1$ , for different values of  $p$ . *Right:* the interval  $[a, b]$  for a fixed threshold  $\beta$ .

By (5), it can be shown that

$$\frac{\mathbb{B}^2(x)h_{loc}^{2(p+1)}(x)}{\mathbb{V}(x)h_{loc}^{-1}(x)} = \frac{1}{2(p+1)}. \quad (12)$$

Now using (12) and defining  $\pi_h(x) = \frac{h_{loc}(x)}{h_{glob}}$ , we may write eq. (11) as follows

$$\Delta_{AMSE}(x) = \frac{1}{2p+3} [\pi_h(x)]^{-2(p+1)} + \frac{2p+2}{2p+3} \pi_h(x) - 1. \quad (13)$$

Note that the equation (13) depends on the unknown functionals of the process only by means of  $\pi_h(x)$ . Also note that  $\pi_h(x) \geq 0$ . If we study the equation (13) as a function of  $z = \pi_h(x)$ , for  $z \geq 0$ , we can see that the unique solution for which there is no gain in using the local bandwidth is when  $\pi_h(x) = 1$ , as shown in Figure 1 (and this is true for each  $p$ ). The higher the deviation from 1, the higher the relative increment of  $\Delta_{AMSE}(x)$ . For example, for  $p = 0$ , a relative increment of about 50% of  $\Delta_{AMSE}(x)$  will be observed for those  $x \in I_X$  for which the local bandwidth is approximately the double or one half of the global bandwidth. A global measure of the (10) may be derived by considering some kind of mean value on the subset  $I_X$ . In order to get a robust measure we propose the following approach. Fix an initial threshold  $\beta$  for the  $\Delta_{AMSE}$  and derive the extreme values of the interval  $[a_\beta, b_\beta]$  where  $\Delta_{AMSE} < \beta$ , as indicated in Figure 1. Note that the values  $a_\beta$  and  $b_\beta$  may be derived immediately by solving the equation (13) for the desired value of  $p$ . Consider the set of points  $\mathbb{S}_\beta = \{x : x \in I_X, \pi(x) \in [a_\beta, b_\beta]\}$  and  $\bar{\mathbb{S}}_\beta = \{x : x \in I_X, \pi(x) \notin [a_\beta, b_\beta]\}$ . Given now the measure of the process  $\mu_X$ , note that

$$\int_{\mathbb{S}_\beta} \Delta_{AMSE}(x) f_X(x) dx \leq \beta \mu_X(\mathbb{S}_\beta), \quad \int_{\bar{\mathbb{S}}_\beta} \Delta_{AMSE}(x) f_X(x) dx \geq \beta \mu_X(\bar{\mathbb{S}}_\beta).$$

Let  $\beta^*$  denote the threshold value for which  $\mu_X(\mathbb{S}_{\beta^*}) = \mu_X(\bar{\mathbb{S}}_{\beta^*})$ . This means that  $\beta^*$  represents a median value of  $\Delta_{AMSE}(x)$  on the subset  $I_X$ . Note that

we do not need to calculate the integral of  $\Delta_{AMSE}$  in order to derive such median value, but we only need to search iteratively for  $\beta^*$ , basing on the estimation of  $\pi_h(\cdot)$  and on the relation  $\mu_X(\mathbb{S}_{\beta^*}) = \mu_X(\bar{\mathbb{S}}_{\beta^*})$ . Now we propose the algorithm which implement our first stage procedure:

- a) fix a threshold value  $\tau$ , which represents the max relative error tolerated for  $\Delta_{AMSE}$  when using a global bandwidth on  $I_X$  and derive the correspondent interval  $[a_\tau, b_\tau]$ ;
- b) consider a preliminary estimation of the bandwidth function  $h_{loc}(X_t)$ ,  $\forall X_t \in I_X$ . Given the purpose of this stage of the procedure, we should consider a fast and simple (although rough) estimator. We propose, for example, to use the ICI adaptive bandwidth selector of Zhang et al. (2008), starting from a grid of bandwidth values built around the interval  $[a_\tau, b_\tau]$ ;
- c) estimate the global bandwidth on  $I_X$ , as suggested in the next section;
- d) estimate  $\pi_h(X_t)$ , for all  $X_t \in I_X$ , by using the results in steps b) and c); then search iteratively for  $\beta^*$  such that

$$\sum_{X_t \in I_X} \mathbb{I}\{\hat{\pi}_h(X_t) \in [a_{\beta^*}, b_{\beta^*}]\} \approx \sum_{X_t \in I_X} \mathbb{I}\{\hat{\pi}_h(X_t) \notin [a_{\beta^*}, b_{\beta^*}]\}$$

where  $\mathbb{I}(\cdot)$  denote the indicator function.

- e) the subset  $I_X$  should be splitted if the value of  $\beta^*$  is greater of the prefixed threshold value  $\tau$ , not otherwise.

### 3.2 Second stage: deriving the locally global bandwidth

We are interested in the estimation of a global bandwidth which could be considered optimal on the subset  $I_X$ . As seen before, we have to estimate the two functionals in (8) conditionally on the subset  $I_X$ . Specific problems follow concerning the integration, since we have to consider a conditional probability measure. In particular, denoting with  $f_{X|I_X}$  the density of the process conditional on the event  $X \in I_X$ , we have

$$\begin{aligned} AMISE\{\hat{m}_\phi\} &= \int_{I_X} AMSE\{\hat{m}_\phi(x; h)\} f_{X|I_X}(x) dx \\ &= C_1^2 h^{2(p+1)} \int_{I_X} [m_\phi^{(p+1)}(x)]^2 \frac{f_X(x)}{\mu(I_X)} dx + \frac{C_2}{nh} \int_{I_X} \sigma_\phi^2(x) \frac{dx}{\mu(I_X)} \\ &= \mathbb{B}_{\omega_I} h^{2(p+1)} + \mathbb{V}_{\omega_I} h^{-1} \end{aligned}$$

where we dropped for simplicity the  $X$  from  $\mu_X$ , and

$$d\omega_I = \frac{dx}{\mu(I_X)}, \quad \mathbb{B}_{\omega_I} = C_1^2 R_f^I(m_\phi^{(p+1)}), \quad \mathbb{V}_{\omega_I} = \frac{C_2 R^I(\sigma_\phi)}{n} \quad (14)$$

$$R_f^I(m_\phi^{(p+1)}) = \int_{I_X} [m_\phi^{(p+1)}(x)]^2 f_X(x) d\omega_I \quad R^I(\sigma_\phi) = \int_{I_X} \sigma_\phi^2(x) d\omega_I. \quad (15)$$

As usual, the optimal bandwidth can be derived by estimating the functionals in (14) and (15) and plugging them into the (6). In order to do that, the problem is now how to estimate the functions  $\sigma_\phi^2(x)$  and  $m_\phi^{(p+1)}(x)$  on the subset  $I_X$ . We propose to use a global estimator based on the neural network technique (see FINE (1999))

Suppose we want to estimate the volatility function  $\sigma^2(x)$  for model (1) and suppose that  $m(x) = 0$ . This is a very typical setup in presence of financial data. It results that we have to consider  $\phi(Y_t) = Y_t^2$  in (2) and following equations. Let  $m_r(x)$  denote the conditional moment function  $E(Y^r|X_t = x)$ . Note that  $\sigma^2(x) \equiv m_2(x)$  and

$$\sigma_\phi^2(x) = \text{Var}\{Y_t^2|X_t = x\} = m_4(x) - m_2^2(x). \quad (16)$$

Generally, the nonparametric estimation of the (16) implies two simultaneous (but different) nonparametric estimations of the functions  $m_4(x)$  and  $m_2(x)$ , as for example in Härdle & Tsybakov (1997), Fan & Yao (1998) and Franke & Diagne (2006). In particular, denoting with  $\hat{\eta}_i$  a generic nonparametric estimator, we should have

$$\hat{\eta}_i = \arg \min_{\eta} \sum_{t=1}^n [g_i(Y_t) - q(X_t; \eta)]^2, \quad i = 1, 2 \quad (17)$$

where  $g_1(z) = z^4$  for the estimation of  $m_4(x)$ ,  $g_2(z) = z^2$  for the estimation of  $m_2(x)$  and  $q(X_t; \eta)$  is some approximation function (neural network function, local polynomial function, etc...).

This procedure is somehow inefficient. For example, when the estimation is (again) based on the use of a kernel estimator, it is necessary to select two different *pilot bandwidths*. To avoid the problem of the pilot bandwidths, we could use the neural network technique, as in Franke & Diagne (2006), but still we would have to consider the two estimations in (17), for  $i = 1, 2$ , which is particularly inefficient with neural networks. For this reason, we propose an alternative approach based on the following reparametrization:

$$\sigma_\phi^2(x) = m_4(x) - m_2^2(x) = m_2^2(x) [m_{4\varepsilon} - 1] \quad (18)$$

where  $m_{4\varepsilon} = E(\varepsilon_t^4)$ . Then we use a *Feedforward Neural Network* (FNN), with one input layer and one hidden layer, as approximation function in the (17). In this case, the quantity  $q(X_t; \eta)$  is defined as

$$q(X_t; \eta) = \sum_{k=1}^d c_k \Gamma(a_k X_t + b_k) + c_0, \quad (19)$$

where  $\eta = (c_0, c_1, \dots, c_d, a_1, \dots, a_d, b_1, \dots, b_d)$  is the vector of parameters of the FNN to be estimated;  $d$  is the number of nodes in the hidden layer such that  $d = O(\sqrt{n/\log n})$  and  $\Gamma(\cdot)$  is the *logistic activation function*.

Now using (17), (18) and (19), we define the following estimator of the function  $\sigma_\phi^2(x)$  based on the neural network approach:

$$\hat{\sigma}_\phi^2(x) = \hat{m}_2^2(x) [\hat{m}_{4\epsilon} - 1] \quad \hat{m}_2(x) = q(x, \hat{\eta}_2) \quad \hat{m}_{4\epsilon} = \frac{\sum_{t=1}^n X_t^4}{\sum_{t=1}^n [q(X_t, \hat{\eta}_2)]^2}. \quad (20)$$

The appeal of the estimator of  $\sigma_\phi^2(x)$  proposed in (20) is that we use only one neural network estimator  $\hat{\eta}_2$  in (17). Secondly, we need to estimate the derivative function  $m_\phi^{(p+1)}(x)$ . Considering that  $m_\phi^{(p+1)}(x) \equiv m_2^{(p+1)}(x)$ , we can use again the previous NN estimate in order to get:

$$\hat{m}_2^{(p+1)}(X_t) = q^{(p+1)}(X_t; \hat{\eta}_2), \quad \forall X_t \in I_X. \quad (21)$$

Finally, given (20), (21) and the ergodicity of the process, we propose the following two estimators for the functionals in (15):

$$\hat{R}_f^I(m_\phi^{(p+1)}) = \frac{\sum_{X_t \in I_X} [\hat{m}_2^{(p+1)}(X_t)]^2}{\sum_{t=1}^n \mathbb{I}(X_t \in I_X)}, \quad \hat{R}^I(\sigma_\phi) = \frac{\sum_{i=1}^{n^*} \hat{\sigma}_\phi^2(x_i)}{n^*}. \quad (22)$$

The points  $\{x_1, x_2, \dots, x_{n^*}\}$  in  $\hat{R}^I(\sigma_\phi)$  are uniformly spaced in the interval  $I_X$ , with  $n^* < n$ , such as  $n^* = \lfloor n/2 \rfloor$  ( $\lfloor x \rfloor$  is the integer part of  $x$ ). The global bandwidth on  $I_X$  can be estimated by using the (22) into the (7), and then replacing the results into the (6).

## References

- FAN, J. and GIJBELS, I. (1995): Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation, *Journal of the Royal Statistical Society, B*, 57, 371–394.
- FAN, J. and GIJBELS, I. (1996): *Local polynomial modelling and its applications*, Chapman and Hall, London.
- FAN, J. and YAO, Q. (1998): Efficient estimation of conditional variance functions in stochastic regression, *Biometrika*, 85, 645–660.
- FRANKE, J. and DIAGNE, M. (2006): Estimating market risk with neural networks, *Statistics & Decision*, 24, 233–253.
- GIORDANO, F. and PARRELLA, M.L. (2009): A locally adaptive bandwidth selector for kernel based regression, *Working paper n.3.209 - DiSES - University of Salerno*.
- HÄRDLE, W. and TSYBAKOV, A. (1997): Local polynomial estimation of the volatility function in nonparametric autoregression, *Econometrica*, 81, 223–242.
- MASRY, E. and FAN, J. (1997): Local polynomial estimation of regression functions for mixing processes, *Scandinavian Journal of Statistics*, 24, 165–179.
- FINE, T.L. (1999): *Feedforward Neural Network Methodology*, Springer.
- ZHANG, Z.G., CHAN, S.C., HO K.L. and HO, K.C. (2008): On bandwidth selection in local polynomial regression analysis and its application to multi-resolution analysis of non-uniform data, *Journal of Signal Process System*, 52, 263–280.

# Panel Data Models for Productivity Analysis

Luigi Grossi<sup>1</sup> and Giorgio Gozzi<sup>2</sup>

<sup>1</sup> Dipartimento di Economia, Università di Verona,  
Via dell'Artigliere 19,  
37129, Verona, Italy  
(e-mail: luigi.grossi@univr.it)

<sup>2</sup> Dipartimento di Economia, Università di Parma,  
Via Kennedy 6,  
43100, Parma, Italy  
(e-mail: giorgio.gozzi@unipr.it)

**Abstract.** In the present paper dynamic panel models for productivity analysis will be analyzed. Recent years have seen a relevant increase in studies on productivity. This is partly due to rising availability of longitudinal micro-level data. This paper is an attempt to give an answer to some questions about productivity dynamics and determinants, using the large data base of company accounts constructed by Research Center of Unioncamere. In our study we investigate the distribution of labor productivity in two important Italian manufacturing sectors. A new derivation of dynamic panel model starting from a Cobb-Douglas production function has been applied in this paper to discover the underlying generating process of productivity growth and to estimate the elasticities of productivity to personnel expenditure.

**Keywords:** italian manufacturing sector, panel data, productivity growth

## 1 Introduction

In the present paper the productivity dynamics of two Italian manufacturing sectors is analyzed. Productivity has been traditionally studied through aggregate - and/or industry - level data to put in evidence the sources and patterns of productivity growth. Theoretical industrial organization studies have suggested that aggregate productivity growth typically stems from behavior at the firm - or plant - level (Baily et al., 1996; Baldwin and Gu, 2006). In this work we will focus on labor productivity (LP from now on), which is only one component of Total Factor Productivity (TFP). The fact that LP varies widely between plants and companies is well-known (Bottazzi et al., 2002). Recent years have seen a relevant increase in studies on productivity. This is partly due to rising availability of longitudinal micro-level data (LMD). This paper is an attempt to give an answer to some questions about productivity dynamics and determinants, using the large data base of company accounts constructed by Research Center of Unioncamere. In our study we investigate the distribution of LP in two important Italian manufacturing

sectors: mechanical (NACE Rev. 1.1 code: DK29) and textile (DB17). These industries are very relevant in Italian economy because in 2004 they accounted for, respectively, 20.3% and 5.2% of total national exports. Furthermore, they reflect two neatly different type of production processes: textile is defined by OECD as a “low technology skill” industry, while mechanical is considered as a “high-medium technology skill” industry. Available data from the Unioncamere Database allow to compute LP measured by different output (value added, sales, gross output) per employee. The original contribution of the paper can be summarized in two main points. Firstly, a new formulation of a dynamic panel model has been obtained to estimate the elasticities of productivity to personnel expenditure and material assets, starting from a classical version of the Cobb-Douglas production function. Secondly, the data set provided by UnionCamere has never been analyzed before for productivity analysis purposes being a rare case of administrative file merging reliable and deeply checked labor data with company accounts data. The remainder of the paper is structured as follows. In section 2 the UnionCamere data-set is introduced putting in evidence its originality and main features of the analyzed industries. Section 3 is devoted to the development of a new dynamic panel model for elasticities estimates. Section 4 concludes.

## 2 Preliminary data analysis

Data for our research is taken from the administrative file of company accounts of Cerved, which is the largest and most accurate database of company accounts in Italy, suitably processed from the statistical point of view by Unioncamere. For more detailed information about the data, see (Ganugi et al., 2005). The source of data in our research is the universe of companies of the Italian mechanical sector (DK29) and textile sector (DB17) in operation in the period 1998-2004. The database contains firms which enter the set after transformation of their juridical type (individual firms or partnership becoming companies), companies which follow an opposite procedure and consequently exit from the set, and firms which enter and exit in the same period (mergers, wind-up, bankruptcy). Each company has a unique identification number. We identify entry and exit by linking successive years together. A unit with a new identification number (ID) in  $t$ , is called “incoming”. If an ID present in  $t - 1$  has disappeared in  $t$ , the corresponding company is an “outgoing”. However, it is also possible to be absent in  $t - 1$  and  $t + 1$ , i.e. an incoming that exits after one year. This latter category is thus both an entrant and an outgoing and is called “one-year-only”. If the ID is present in each year of the period, the unit is a “survivor”. A unit whose ID is present in  $t - 1, t, t + 1$  is called “stayer”. Consequently, the number of establishments in  $t$  consists of stayer plus incoming plus outgoing less one-year-only companies. LP estimates are derived as the ratio of a measure of output and inputs. In order to preserve the largest number of units in the



estimation process, we focused on LPGO which is a labor productivity index computed as gross output per employee. To take the inflationary dynamics into account, it is necessary deflating the nominal gross output at industry level. We deflated the gross output with an industry implicit price deflator of sales. For each company group we computed the above cited deflated productivity index called LPGO00 (constant 2000 prices). Weighted means for each year are reported in Table 1. In the whole period (1998-2004) the LPGO at prices of year 2000 of all companies is decreased of by 6.8% in the textile industry (yearly -1.2%), while is totally increased of 8% in the mechanical industry (yearly +1.3%). Thus, the variation of LP in Italian textile companies showed a strong slowdown while is substantially increased among survivors in the mechanical industry.

**Table 1.** Average labor productivity in terms of gross output for different types of companies (textile industry: DB17; mechanical industry: DK29). Euro  $\times$  1000.

DB17						DK29				
years	all	stay.	inc.	out.	surv.	all	stay.	inc.	out.	surv.
1998	157.7	-	-	140.4	167.1	189.9	-	-	161.9	192.8
1999	155.7	157.4	126.8	143.9	160.5	192.4	192.9	162.0	212.7	192.9
2000	166.0	168.8	112.4	137.3	172.6	196.7	199.0	137.6	203.6	202.4
2001	169.1	169.9	141.9	170.9	173.1	203.8	202.1	238.6	217.0	202.4
2002	159.9	160.7	155.6	146.9	163.8	203.0	205.2	163.5	197.2	206.9
2003	152.9	155.2	131.7	139.0	156.1	195.1	197.5	161.3	177.3	200.6
2004	154.5	-	152.8	-	155.7	199.9	-	159.1	-	208.3
Tot. $\Delta$	-2.0	-1.4	20.5	-1.0	-6.8	5.3	2.4	-1.8	9.5	8.0
Yearly $\Delta$	-0.3	-0.4	3.8	-0.2	-1.2	0.9	0.6	-0.4	1.8	1.3

To develop policies to raise productivity we have to understand the causes of productivity growth of survivors. To support this phase of the analysis, companies in the sector were assigned to four categories or “quadrants” defined as follows. Quadrant 1 consists of the “successful upsizers”, companies that were able to increase both LP and employment. Quadrant 2 are the “successful downsizers”, companies that raised productivity but did so by reducing employment. Quadrant 3 are the “unsuccessful downsizers”, companies that faced reductions in both productivity and employment. Finally, quadrant 4 includes companies called “unsuccessful upsizers”, firms that raised employment but at the expense of productivity.

Following a decomposition of productivity growth introduced by Grossi and Gozzi (2007) we found that the negative productivity growth in the textile sector has been mainly caused by unsuccessful upsizers which decreased their productivity at a rate very lower than the industry average, while the positive growth of the mechanical sector can be mainly attributed to successful downsizers which showed a very high productivity growth rate.

### 3 Dynamic panel models for productivity elasticities

In order to study the relationship between productivity, employment and personnel expenditure we estimated a dynamic panel models. Our analysis is based on the Cobb-Douglas production function homogeneous of degree  $d$  assumed to hold for a given industry, that is

$$Q = f(L, C) = a \times L^{\alpha_1} \times C^{\alpha_2}, \quad (1)$$

where  $\alpha_1 + \alpha_2 = d = 1$  under constant returns to scale;  $Q$  represents the production level given by the quantity of output,  $L$  represents labor, that is the number of workers,  $C$  represents the capital stock and  $a$  represents any force for which we have no quantitative data usually thought as the technological progress, which we normalize to one. By Euler's theorem, for a production function homogeneous of degree  $d$ , we have

$$dQ = \frac{\partial Q}{\partial C} C + \frac{\partial Q}{\partial L} L \quad (2)$$

If perfect competition in the factor markets is assumed, then factor of production are paid their marginal product. Let  $W$  be the nominal wage,  $P$  the output price and  $R$  the rental cost of capital. Then the real total factor payment to labor,  $(W/P)L$ , is given by

$$\frac{W}{P} L = \frac{\partial Q}{\partial L} L = (\alpha_1 a C^{\alpha_2} L^{\alpha_1-1}) L = \alpha_1 Q \quad (3)$$

and the total real factor payment to capital,  $(R/P)C$ , is given by

$$\frac{R}{P} C = \frac{\partial Q}{\partial C} C = (\alpha_2 a C^{\alpha_2-1} L^{\alpha_1}) C = \alpha_2 Q. \quad (4)$$

Substituting (3) and (4) into (2) yields

$$dQ = \frac{W}{P} L + \frac{R}{P} C = \alpha_1 Q + \alpha_2 Q \quad (5)$$

where  $\alpha_1$  is the labor share of nominal output ( $WL/PQ$ ) and  $\alpha_2$  is the capital share of nominal output ( $RK/PQ$ ). We have now to assume that output prices are marked-up by a factor,  $m$ , over total per-unit costs. This is given by

$$P = m \left( \frac{WL}{Q} + \frac{RC}{Q} \right) \quad (6)$$

Keeping in mind that, for a Cobb-Douglas function we have  $d = \alpha_1 + \alpha_2$ , equation (6) can be rewritten as

$$P = m \left[ \frac{WL}{Q} + \frac{\alpha_2}{\alpha_1} \frac{WL}{Q} \right] = m \left[ 1 + \left( \frac{\alpha_2}{\alpha_1} \right) \right] \frac{WL}{Q} \quad (7)$$

and rearranging we get

$$\frac{PQ}{L} = \frac{md}{\alpha_1} W. \quad (8)$$

We thus have the labor productivity being proportional to wages. Taking natural logs of equation (8) and denoting natural logs by lowercase letters, we can write

$$y_{it} = \beta_0 + \beta_1 w_{it} \quad (9)$$

where  $\beta_0 = \ln(md/\alpha_1)$  and  $y_{it} = p_{it} + q_{it} - l_{it}$ , which is the labor productivity of firm  $i$  at time  $t$ . The parameter  $\beta_1$  can be interpreted as an elasticity. This elasticity could be estimated using a pooling OLS model, but such a model would be inadequate because we are examining several companies over a multi-year time period. This model fails to account for interfirm differences which persist throughout the time period. Thus, this model is maximally susceptible to bias caused by spurious correlations. For example, firms with fewer employees may have both smaller wages and less productivity. If we did not control for such interfirm differences, and we found that  $\beta_1$  was positive, we might mistakenly attribute between-firm differences to within-firm differences, and conclude that decreases in the wages cause decreases in productivity levels.

The econometric model should explore the relation among real per capita personnel expenditure (from now on called “wages”), employment and productivity. Studying the role of firm real wage growth in the productivity change and employment change raise obviously endogeneity problems. For this reason a dynamic approach has been applied where endogenous and predetermined variables are instrumented by lagged values of the same variables, taken in levels, giving consistent estimates of parameters (see Arellano, 2003).

The general model we are interested in, is a dynamic panel model having the following form:

$$y_{it} = \sum_{k=1}^p \delta_k y_{i(t-k)} + \beta'(L)x_{it} + u_{it} \quad (10)$$

with

$$u_{it} = \mu_i + \eta_t + v_{it} \quad (11)$$

$t = q + 1, \dots, T; i = 1, \dots, N$ , where  $\delta_k$  are the coefficients of the lagged dependent variable,  $x_{it}$  is a  $r \times 1$  vector of explanatory variables,  $\beta'(L)$  is a  $r \times 1$  vector of associate polynomials in the lag operator and  $q$  is maximum lag length in the model. We will assume a two-way error component model where  $\mu_i$  and  $\eta_t$  are respectively individual and time specific effects. In order to identify the model the error term  $v_{it}$  is assumed to be serially uncorrelated. Furthermore  $v_{it}$  are assumed to be independently distributed across units with zero mean, while heteroskedasticity of disturbances across units and times is allowed. A sufficient general method of estimation is the

GMM method (see Arellano and Bover, 1995). For consistency, this estimator requires the regressors  $x_{it}$  to be strictly exogenous with respect to the disturbances  $v_{it}$ . Therefore, in dynamic panel models estimation, we shall usually apply transformations that allow the use of lagged endogenous and predetermined variables as instruments in the transformed equations. Now, the problem of correlation with individual effects comes to be important because when there are no instruments which are uncorrelated with individual effects  $\mu_i$  some transformations must be applied to wipe out this component of error term such as first difference.

The consistency of the GMM estimators which instrument the lagged dependent variable with further lags of the same variable, relies on the hypothesis that the disturbances  $v_{it}$  are not correlated. If the disturbances are not serially correlated these two conditions should be observed:

- a. negative first order serial correlation in differenced residuals, that is  $E(\Delta\hat{v}_{it}\Delta\hat{v}_{i(t-1)}) < 0$ ;
- b. no evidence of second order serial correlation in the differenced residuals, that is  $E(\Delta\hat{v}_{it}\Delta\hat{v}_{i(t-2)}) = 0$  which is trivially proved because all the terms of the product are null.

In this paper two tests are used for the null of absence of first ( $\rho_1$ ) and second ( $\rho_2$ ) order autocorrelation.

The dynamic panel model has been estimated for both sectors and for each category defined by quadrants (see section 2) using the LPGO (labor productivity measured as Gross Output per employee) as the dependent value and the cost of labor per employee as explanatory variable (average wage in firm  $i$  at time  $t$ ). In order to wipe out the correlation between the lagged values of the dependent variable and the error term, the first differences have been computed. Using notation of equation (10) we have:

$$\Delta LPGO_{it} = \delta \Delta LPGO_{i,t-1} + \beta_1 (CL_{it}/E_{it}) + u_{it} \quad (12)$$

where  $CL_{it}$  is the personnel expenditure of company  $i$  at year  $t$ ,  $E_{it}$  represents the number of employees and  $u_{it}$  is defined as in equation (11).

Estimated parameters (with standard errors in brackets) are reported in Table 2. Column titled  $\rho_1$  and  $\rho_2$  reports the statistic of first and second order autocorrelation tests with the correspondent  $p$ -values in brackets. The companies that increased productivity (successful upsizers and downsizers) had the highest real wage growth with the largest increases coming from the ones that decreased employment (+2.7% in DK29, +3.5% in DB17). The plants that experienced declines in productivity had reductions in real wages with the exception of the firms that decreased employment which show a light increase of real wages. The unsuccessful upsizers experienced instead the highest decrease of real wages in both sectors. The estimated parameters raise interesting questions. The first possibility is that some of the wage changes observed may be associated with changes in labor quality. Under this

interpretation, successful upsizers may be adding more skilled workers paying higher wages, successful downsizers retaining their higher skilled workers paying higher wages and having recourse to overtime work, unsuccessful downsizers retaining their less skilled workers and unsuccessful upsizers adding less skilled workers decreasing the per capite personnel expenditure. A second interpretation is that increases in the wages for certain types of workers may have led to capital/labor substitution and vice versa for the plant with wage declines. A third possibility is rent sharing. Those firms that increased productivity gave (or were forced to give) a fraction of that increase to their workers. This would also lead to a positive relationship between the change in the wages and the change in productivity. In the case of companies which increased their productivity the elasticity of productivity to real wages is positive in both industries and rather high for the successful downsizers. This result seems to support the first hypothesis that successful upsizers employed more skilled workers and successful downsizers fired the less skilled employees. Among mechanical successful downsizers, holding other condition constant (that is assets growth and autoregressive effect of productivity) more than 80% of the pay increase passed to productivity. Looking at Table 2 we observe that, in this group of firms, more skilled workers produced an average yearly increase of gross output and a very high increase of labor productivity measured as gross output per employee. In the textile industry, even if the elasticity is positive and high, the average wage increase is larger than in the mechanical and the average gross output increase is negative (but in absolute value less than the employment decrease, so that productivity increases). This can be probably interpreted as a larger recourse to overtime working hours which, holding worker's skills constant, caused a productivity increase. In this case the third hypothesis is likely to hold because a fraction of productivity increase has been shared with the workers who has been required to work better and more. Elasticities among unsuccessful upsizers are negative for both industries. This means that, holding other factors constant, when wages increase the productivity decrease on average. Firms belonging to this group added less skilled workers decreasing the average real wages but this policy unmotivated the more skilled workers (paid more) with a final decrease of productivity. Finally, the group of unsuccessful downsizers show positive elasticities, and for textile industry elasticity is greater than one so that the changes of productivity are more than proportional to wage variations. In this case the second hypothesis can be valid: the increase of the wages for a certain type of workers has lead to a substitution of labor with capital with a decrease of labor productivity. This hypothesis is supported also by the high elasticity of per capita assets. It is worth noticing that in this paper we used the number of employees as a denominator of labor productivity index. As suggested recently by OECD (OECD, 2001), quantifying the labor input by working hours could lead to more refined results. This possibility will be explored in a future paper conditionally to the availability of proper data.

**Table 2.** Panel models estimates: up- and down-sizers quadrants.

Quadrant	DK29				DB17			
	$\delta$	$\beta_1$	$\rho_1$	$\rho_2$	$\delta$	$\beta_1$	$\rho_1$	$\rho_2$
1. suc-up	0.273 (0.032)	0.654 (0.057)	-8.770 (0.000)	1.759 (0.039)	0.571 (0.039)	0.258 (0.094)	-4.102 (0.000)	0.222 (0.412)
2. suc-down	0.354 (0.030)	0.892 (0.074)	-8.152 (0.000)	-1.154 (0.124)	0.565 (0.038)	0.962 (0.126)	-5.855 (0.000)	-0.646 (0.259)
3. uns-up	0.370 (0.024)	-0.331 (0.060)	-8.444 (0.000)	1.503 (0.066)	0.638 (0.030)	-0.282 (0.155)	-4.933 (0.000)	1.504 (0.066)
4. uns-down	0.512 (0.031)	0.750 (0.093)	-12.346 (0.000)	0.276 (0.391)	0.427 (0.048)	1.891 (0.140)	-3.226 (0.00)	-1.384 (0.083)
Total	0.103 (0.020)	-0.004 (0.043)	-13.451 (0.000)	-2.047 (0.020)	0.277 (0.035)	0.004 (0.134)	-9.527 (0.000)	0.728 (0.233)

## 4 Concluding remarks

In this paper an original dynamic panel model has been obtained to study elasticities of productivity to personnel expenditure. The model perfectly fit to the literature about industrial organization because has been derived from a classical Cobb-Douglas production function. Elasticities have been estimated for two important Italian industries (textile and mechanical) using a very rich and reliable data set provided by the Italian Federation of Chambers of Commerce. Problems of endogeneity have been addressed exploiting the longitudinal structure of the data set and introducing a dynamic term in the model. Estimates have been obtained by the Generalized Method of Moments (GMM).

## References

- ARELLANO M. (2003): *Panel data econometrics*. Oxford University Press, Oxford.
- ARELLANO M. and BOVER O. (1995): Another look at the instrumental-variable estimation of error-components models. *Journal of Econometrics*, 68, 29-52.
- BAILY M.N., BARTELSMAN E. J. and HALTIWANGER J. (1996): Downsizing and productivity growth: myth or reality? *Small Business Economics*, 8, 259-278.
- BALDWIN J. R. and GU W. (2006): Plant Turnover and Productivity Growth in Canadian Manufacturing. *Industrial and Corporate Change*, 15, 417-465.
- BOTTAZZI G. , CEFIS E. and DOSI G. (2002): Corporate Growth and Industrial Structure. Some Evidence from the Italian Manufacturing Industry. *Industrial and Corporate Change*, 11, 705-723.
- GANUGI P., GROSSI L., GOZZI G. (2005): Testing Gibrat's law in italian macro-regions: analysis on a panel of mechanical companies. *Statistical Methods and Applications*, 14, 101-126.
- GROSSI L. and GOZZI G. (2007): Firm turnover and labor productivity growth in the Italian mechanical sector. In: Skiadas C. H. (ed.): *Recent advances in stochastic modeling and data analysis*, World Scientific, 382-389.
- OECD (2001): *Measuring Productivity. Measurement of Aggregate and Industry-Level Productivity Growth*, OECD Manual, Available online, <http://www.oecd.org/dataoecd/59/29/2352458.pdf>.

# A Stochastic Gamma Diffusion Model with Threshold Parameter. Computational Statistical Aspects and Application

Ramón Gutiérrez<sup>1</sup>, Ramón Gutiérrez-Sánchez<sup>1</sup>, Ahmed Nafidi<sup>2</sup>, and Eva Maria Ramos-Ábalos<sup>1</sup>

<sup>1</sup> Department of Statistics and Operational Research, University of Granada, Faculty of Sciences, Campus de Fuentenueva  
18071 Granada, Spain, *ramosa@ugr.es*

<sup>2</sup> Ecole Supérieure de Technologie de Berrechid, Université Hassan 1<sup>er</sup>, Quartier Tagadom, Passage d'Alger  
B.P: 218, Berrechid, Maroc

**Abstract.** In this paper, we propose a new study of a stochastic gamma diffusion process, with threshold parameter, which can be considered as an extension of the gamma diffusion process. The estimation of the threshold parameter requires the solution of a nonlinear equation. To do so, we propose the classical Newton-Raphson method. This methodology is applied to an example with simulated data.

**Keywords:** discrete sampling, statistical inference in diffusion process, application

## 1 Introduction

Diffusion processes, which play a fundamental role in stochastic modelling, are considered either from the standpoint of the corresponding Ito stochastic differential equations or from that of the associated Kolmogorov (Fokker-Planck and backward) partial differential equations. This role can be seen in applications in fields such as biology, physics, demography, economics, finance and environmental sciences.

Questions of statistical inference and parameter estimation in these processes have received considerable attention in recent years, both when the process is observed continuously and when discretely. In most cases, parameter estimation is based on approximating the maximum likelihood methodology. A large body of literature addresses this question, and important studies in general or in particular cases include Bibby and Sorensen (1995); Eugene (2000) and extensive review given in Prakasa-Rao (1999).

In this paper, therefore, we propose a new stochastic type gamma diffusion process. The paper is structured as follows: in Section 2, the proposed model is defined, we identify the main characteristics of the proposed process. In section 3, the parameter estimators are derived by the maximum likelihood

method, using discrete sampling of the process. Estimation of the threshold parameter requires us to resolve a non-linear equation, which is done by means of the Newton-Raphson method. Section 4 contains a simulation of the exact solution of Ito's stochastic differential equation which characterises the process. The simulated process data are used to estimate the parameters of the model using the proposed methodology and these are compared with the true values used for the simulation.

## 2 The model and its basic probabilistic characteristics

### 2.1 The model

The one-dimensional Gamma diffusion process with threshold parameter can be introduced by means of the Kolmogorov backward and forward equation as a Markov process  $\{X(t), t_1 \leq t \leq T, t_1 > 0\}$  with values in  $]\gamma, +\infty[$ , with almost-certainly continuous trajectories and with a distribution function for the process transition that is given by  $P(y, t|x, s) = P[X(t) \leq y|X(s) = x]$  where  $x > \gamma$ ,  $y > \gamma$  and  $\gamma \in \mathbb{R}$ .

By assuming the following conditions:

- $\lim_{h \rightarrow 0} \frac{1}{h} \int_{|y-x| > \epsilon} P(dy, t+h|x, t) = 0$
- $\lim_{h \rightarrow 0} \frac{1}{h} \int_{|y-x| \leq \epsilon} (y-x)P(dy, t+h|x, t) = A_1(x, t) = \left(\frac{\alpha}{t} - \beta\right)(x - \gamma)$
- $\lim_{h \rightarrow 0} \frac{1}{h} \int_{|y-x| \leq \epsilon} (y-x)^2 P(dy, t+h|x, t) = A_2(x, t) = \sigma^2(x - \gamma)^2 > 0$
- The higher-order infinitesimal moments are null.

and so the infinitesimal moments of the process are

$$\begin{aligned} A_1(x, t) &= \left(\frac{\alpha}{t} - \beta\right)(x - \gamma) \\ A_2(x, t) &= \sigma^2(x - \gamma)^2 \end{aligned}$$

The corresponding Kolmogorov backward and forward equations are

$$\begin{aligned} \frac{\partial p}{\partial s} + \frac{1}{2} A_2(x, s) \frac{\partial^2 p}{\partial x^2} + A_1(x, s) \frac{\partial p}{\partial x} &= 0 \\ -\frac{\partial p}{\partial t} + \frac{1}{2} \frac{\partial^2 A_2(y, t)p}{\partial y^2} - \frac{\partial A_1(y, t)p}{\partial y} &= 0 \end{aligned}$$

where  $p$  denotes the transition probability density function (t.p.d.f.),  $p(y, t | x, s)$ , corresponding to the transition distribution function  $P(y, t | x, s)$ .

Alternatively, the above-defined process can be considered as the solution to the following Itô's stochastic differential equation (SDE)

$$dX(t) = \left(\frac{\alpha}{t} - \beta\right)(X(t) - \gamma)dt + \sigma(X(t) - \gamma)dW(t), \quad X(t_1) = x_{t_1} \quad (1)$$



where  $W(t)$  represents the Wiener process with independent increments  $W(t) - W(s)$  distributed according to one-dimensional normal distribution  $\mathcal{N}(0, \sqrt{t-s})$  for  $t > s$ ,  $x_{t_1}$  is positive real (fixed).

It can be proved that the functionals  $A_1(x, t)$  and  $A_2(x, t)$  are non-anticipative and satisfy the Lipschitz and the growth conditions and consequently, that there exists a unique, strong solution to Eq. (1) (see, for example, Liptser and Shiriyayev (1978), Theorem 4.6).

Furthermore, it is straightforward to show that these functionals are Borel measurable and satisfy the uniform Lipschitz condition and the c-Holder, in particular order 1 Holder, conditions (see, for example, Wong and Hajeck (1985), Propositions 4.1 and 7.1). Consequently, there exists a separable, measurable and almost surely (a.s.) sample continuous diffusion process  $\{X(t); t \in [t_1, T]\}$  which is the unique (a.s.) solution to Ito's SDE Eq. (1) with infinitesimal moments (drift and diffusion coefficients) given, respectively, by  $A_1(x, t)$  and  $A_2(x, t)$ .

By applying Itô's formula to the time-independent transformation  $Y(t) = \log(X(t) - \gamma)$ , we have

$$dY(t) = \left( \frac{\alpha}{t} - \beta - \frac{\sigma^2}{2} \right) dt + \sigma dW(t) = \left[ \frac{\alpha}{t} - \left( \beta + \frac{\sigma^2}{2} \right) \right] dt + \sigma dW(t)$$

The solution to this is

$$\begin{aligned} Y(t) - Y(t_1) &= \int_{t_1}^t \left[ \frac{\alpha}{\tau} - \left( \beta + \frac{\sigma^2}{2} \right) \right] d\tau + \sigma(W(t) - W(t_1)) = \\ &= \alpha \log(t/t_1) - \left( \beta + \frac{\sigma^2}{2} \right) (t - t_1) + \sigma(W(t) - W(t_1)) \end{aligned}$$

from which we obtain the solution to the original SDE:

$$X(t) - \gamma = \exp \left\{ \log(X(t_1) - \gamma) + \alpha \log \left( \frac{t}{t_1} \right) - \left( \beta + \frac{\sigma^2}{2} \right) (t - t_1) + \sigma(W(t) - W(t_1)) \right\} \quad (2)$$

The distribution of the random variable  $(X(t) - \gamma) | X(s) = x_s$  is the one-dimensional three-parameter lognormal distribution  $A_1[\mu(s, t, x_s); \sigma^2(t-s)]$  with  $\mu(s, t, x_s) = \log(x_s - \gamma) + \alpha \log \left( \frac{t}{s} \right) - \left( \beta + \frac{\sigma^2}{2} \right) (t - s)$ .

Then, the transition probability density function (t.p.d.f.) for the considered process is:

$$p(y, t | x, s) = [2\pi\sigma^2(t-s)]^{-\frac{1}{2}} (y - \gamma)^{-1} \exp \left\{ -\frac{[\log(y - \gamma) - \mu(s, t, x_s)]^2}{2\sigma^2(t-s)} \right\} \quad (3)$$

with the initial condition  $p(y, s | x, s) = \delta(y - x)$ .

## 2.2 Moments of the process

The moments of the Gamma diffusion process with a threshold parameter are derived from those of the two-parameter lognormal diffusion process,  $X(t) - \gamma$ .

Thus, we have,

$$\mathbb{E} \left[ (X(t) - \gamma)^k | X(s) = x_s \right] = \exp \left\{ k\mu(s, t, x_s) + \frac{k^2 \sigma^2}{2} (t - s) \right\}.$$

Taking into account that  $X^k(t) = (X(t) - \gamma + \gamma)^k$  and applying Newton's binomial, we have:

$$X^k(t) = \sum_{j=0}^k \binom{k}{j} (X(t) - \gamma)^j \gamma^{k-j}.$$

Therefore,

$$\begin{aligned} \mathbb{E}[X^k(t) | X(s) = x_s] &= \sum_{j=0}^k \binom{k}{j} \gamma^{k-j} \exp \left\{ j\mu(s, t, x_s) + \frac{j^2 \sigma^2}{2} (t - s) \right\} = \\ &= \sum_{j=0}^k \binom{k}{j} \gamma^{k-j} \exp \left\{ j \log(x_s - \gamma) + j \log \left( \frac{t}{s} \right)^\alpha - \right. \\ &\quad \left. - j \left( \beta + \frac{\sigma^2}{2} \right) (t - s) + \frac{j^2 \sigma^2}{2} (t - s) \right\} \end{aligned}$$

Considering the initial condition,  $P[X(t_1) = x_{t_1}] = 1$  and  $k = 1$ , the trend function of the Gamma diffusion process with threshold parameter (TF) is given by

$$\mathbb{E}[X(t)] = \gamma + (x_{t_1} - \gamma) \left( \frac{t}{t_1} \right)^\alpha e^{-\beta(t-t_1)} \quad (4)$$

The conditional trend function (CTF) in this case is given by

$$\mathbb{E}[X(t) | X(s) = x_s] = \gamma + (x_s - \gamma) \left( \frac{t}{s} \right)^\alpha e^{-\beta(t-s)} \quad (5)$$

**Remark 1.**

Note that the expression of  $\mathbb{E}[X(t)]$ , takes the form of the model of the family of deterministic curves of gamma growth. In other words, the trend of the Gamma diffusion with a threshold parameter  $\gamma$  as introduced here has the same expression as the cited growth model. Moreover, except the constants, this trend takes the form of the Gamma density function, which justifies the denomination of stochastic diffusion which we are considering.

### 3 Statistical inference and computational aspects

We now estimate the parameters  $\alpha$ ,  $\beta$ ,  $\sigma^2$  and  $\gamma$  of the model using the maximum likelihood method. Let us consider a discrete sampling of the process  $\{X(t_i) = x_{t_i} = x_i, \quad 1 \leq i \leq n\}$  for the instants  $t_1, \dots, t_n$ , with the initial

condition  $P[X(t_1) = x_1] = 1$ . The associated maximum likelihood function is thus

$$\begin{aligned}\mathbb{L}(x_1, \dots, x_n; \alpha, \beta, \sigma^2, \gamma) &= \prod_{i=2}^n p(x_i, t_i | x_{i-1}, t_{i-1}) = \\ &= \prod_{i=2}^n \left[ [2\pi\sigma^2]^{-\frac{1}{2}} (x_i - \gamma)^{-1} \right] \exp \left\{ -\frac{[\log(x_i - \gamma) - \mu(t_{i-1}, t_i, x_{i-1})]^2}{2\sigma^2} \right\} = \\ &= [2\pi\sigma^2]^{-\frac{n-1}{2}} \prod_{i=2}^n (x_i - \gamma)^{-1} \times \\ &\quad \times \exp \left\{ -\frac{\sum_{i=2}^n \left[ \log(x_i - \gamma) - \log(x_{i-1} - \gamma) - \alpha \log\left(\frac{t_i}{t_{i-1}}\right) + \left(\beta + \frac{\sigma^2}{2}\right)(t_i - t_{i-1}) \right]^2}{2\sigma^2(t_i - t_{i-1})} \right\}\end{aligned}$$

This function tends to infinity when  $\gamma$  tends to  $x_{(1)}$ , where  $x_{(1)} = \inf_{0 \leq j \leq n} (x_j)$ .

In order to work with a known likelihood function and to calculate the estimators in the simplest possible way, the discrete sampling is transformed as follows:

$$\begin{aligned}\mathbf{u}_i &= (t_i - t_{i-1})^{-1/2} (\log(t_i/t_{i-1}), t_i - t_{i-1})' \\ \mathbf{a} &= \left( \alpha, -\left(\beta + \frac{\sigma^2}{2}\right) \right)' \\ v_{i,\gamma} &= (t_i - t_{i-1})^{-1/2} (\log(x_i - \gamma) - \log(x_{i-1} - \gamma)),\end{aligned}$$

for  $i = 2, \dots, n$  and thus, the likelihood function can be written as

$$\begin{aligned}\mathbb{L}_{v_2, \dots, v_n}(\mathbf{a}, \sigma^2, \gamma) &= [2\pi\sigma^2]^{-(n-1)/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=2}^n (v_{i,\gamma} - \mathbf{u}_i' \mathbf{a})^2 \right\} = \\ &= [2\pi\sigma^2]^{-(n-1)/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=2}^n (\mathbf{v}_\gamma - \mathbf{U}' \mathbf{a})' (\mathbf{v}_\gamma - \mathbf{U}' \mathbf{a}) \right\}\end{aligned}$$

where  $\mathbf{v}_\gamma = (v_{2,\gamma}, \dots, v_{n,\gamma})'$  and  $\mathbf{U} = (\mathbf{u}_2, \dots, \mathbf{u}_n)$ .

The logarithm of this function is

$$\log(\mathbb{L}_{v_2, \dots, v_n}(\mathbf{a}, \sigma^2, \gamma)) = -\frac{n-1}{2} \log(2\pi) - \frac{n-1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{v}_\gamma - \mathbf{U}' \mathbf{a})' (\mathbf{v}_\gamma - \mathbf{U}' \mathbf{a}).$$

By differentiating the log-likelihood function with respect to  $\mathbf{a}$ ,  $\sigma^2$  and  $\gamma$  and by equalling to zero, the following equations are obtained:

$$\mathbf{U}(\mathbf{v}_\gamma - \mathbf{U}' \mathbf{a}) = 0 \quad (6)$$

$$(n-1)\sigma^2 = (\mathbf{v}_\gamma - \mathbf{U}' \mathbf{a})' (\mathbf{v}_\gamma - \mathbf{U}' \mathbf{a}) \quad (7)$$

$$\frac{\partial \mathbf{v}_\gamma'}{\partial \gamma} (\mathbf{v}_\gamma - \mathbf{U}' \mathbf{a}) = 0 \quad (8)$$

After some algebraic rearrangement, the maximum likelihood estimators of  $\mathbf{a}$  and  $\sigma^2$  yield the equation (6), and (7) is

$$\hat{\mathbf{a}} = (\mathbf{U}\mathbf{U}')^{-1}\mathbf{U}\mathbf{v}_{\hat{\gamma}} \quad (9)$$

$$(n-1)\hat{\sigma}^2 = \mathbf{v}_{\hat{\gamma}}'\mathbf{H}_{\mathbf{U}}\mathbf{v}_{\hat{\gamma}} \quad (10)$$

where the matrix  $\mathbf{H}_{\mathbf{U}}$  is the symmetric and idempotent matrix given by  $\mathbf{H}_{\mathbf{U}} = \mathbf{I}_{n-1} - \mathbf{U}'(\mathbf{U}\mathbf{U}')^{-1}\mathbf{U}$ .

Taking into account equations (2) and (3), expression (8) can be expressed as

$$\frac{\partial \mathbf{v}_{\hat{\gamma}}'}{\partial \hat{\gamma}} \mathbf{H}_{\mathbf{U}} \mathbf{v}_{\hat{\gamma}} = 0 \quad (11)$$

Equation (11) can be resolved numerically using the Newton-Raphson method. This solution always exists, but may not be compatible with the data being considered.

**Remark 2.**

From Zehna's theorem, we can obtain the maximum likelihood estimated trend function (ETF) and the conditional estimated trend function (ECTF) of the process, by substituting the parameters by their estimators in expressions (4) and (5).

$$\hat{\mathbb{E}}[X(t)] = \hat{\gamma} + (x_{t_1} - \hat{\gamma}) \left( \frac{t}{t_1} \right)^{\hat{\alpha}} e^{-\hat{\beta}(t-t_1)} \quad (12)$$

$$\hat{\mathbb{E}}[X(t) \mid X(s) = x_s] = \hat{\gamma} + (x_s - \hat{\gamma}) \left( \frac{t}{s} \right)^{\hat{\alpha}} e^{-\hat{\beta}(t-s)} \quad (13)$$

## 4 Application to simulated data: statistical fit using Newton-Raphson

The stochastic differential equation (1) has a single continuous solution in the interval  $[t_1, T]$ , which corresponds to the three-parameter lognormal diffusion process. Equation (2) is the explicit expression of this solution, which can be obtained by means of Itô's formula, applied to the transform  $\ln(X(t) - \gamma)$  (see section 2.1).

From this explicit solution, the simulated trajectories of the process are obtained from the following discretizing time interval  $[t_1, T] : t_i = t_1 + (i-1)h$ , for  $i = 1, \dots, n$  ( $n$  is an integer and  $h$  is the discretization step), taking into account that the Wiener process is obtained as the sum of the distributions  $\mathcal{N}(0, h)$  with the initial condition  $W(t_1) = 0$ .

We consider an application of the estimation and simulation of the process studied previously; We then go on to analyze, using a simulated sample of observations of the proposed model, the problem of estimating the threshold

parameter. From this simulated process sample, by considering  $h = 1$ ,  $n = 25$ , and initial value  $x_1 = 1, 22139$ , we estimate the parameters by maximum likelihood, using the Newton-Raphson (NR) nonlinear approach to approximate the value of  $\hat{\gamma}$  (Eq. (11)), reserving the values observed for the time  $t = 25$  for comparison with the corresponding prediction by the model.

Table 1 shows the simulated data, and the estimated trend function (ETF). Table 2 shows the values used in the simulation and the results obtained by estimating the parameters, using the methods described above, implemented using the Mathematica packages. Figure 1 shows the fit and the prediction obtained for  $X(t)$  using the ETF and TF for different parameter values  $\alpha$  and  $\beta$ .

**Table 1.** Simulated data and Estimated Trend Function

Time	$X(t)$	ETF	Time	$X(t)$	ETF
1	1,2242	1,2242	13	14,0414	13,9930
2	2,0424	2,0434	14	13,9702	13,8648
3	3,3721	3,3589	15	13,6454	13,5769
4	4,9955	4,9752	16	13,2528	13,1598
5	6,7895	6,7039	17	12,6854	12,6426
6	8,2852	8,3933	18	11,9527	12,0517
7	9,8124	9,9341	19	11,1937	11,4103
8	11,1853	11,2559	20	10,4424	10,7387
9	12,1080	12,3207	21	9,8033	10,0538
10	12,8758	13,1156	22	9,3053	9,3697
11	13,5611	13,6461	23	8,6935	8,6975
12	13,9292	13,9300	24	8,0377	8,0461
			<b>Prediction</b>		
			25	7,4194	7,4219

**Table 2.** Starting values used in the simulation and estimation of the parameters

	$\gamma$	$\sigma$	$\alpha$	$\beta$
Simulation	1	0,01	2,5	0,2
	$\hat{\gamma}$	$\hat{\sigma}$	$\hat{\alpha}$	$\hat{\beta}$
Estimation NR	0,99785	0,01015	2,48812	0,194298

## 5 Conclusions

Other computation methods seeking to optimize (maximize) likelihood, other than the Newton-Raphson method described here, may be used in the future. For example, Simulated Annealing, as has already been used in diffusion processes other than Gamma, for example in the three-parameter lognormal process (see Gutiérrez et al. (2009)), and also in the particular case of the

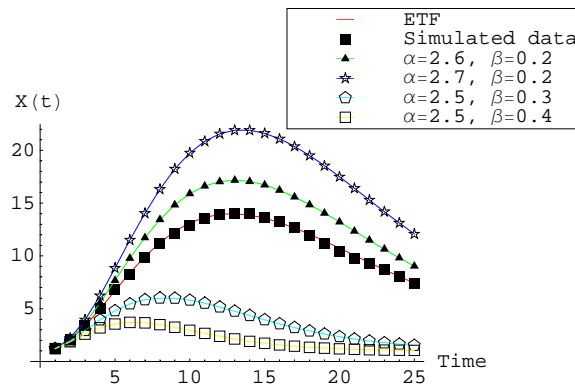


Fig. 1. Fit and prediction based on ETF and TF for different parameter values

distribution of three-parameter lognormal probability (see Vera and Díaz-García (2008)).

## References

- BIBBY, B.M. and SORENSEN, M. (1995): Martingale estimation functions for discretely observed diffusion processes. *Bernoulli* 1(1/2), 17–39.
- EUGENE, M.C. (2000), Maximum likelihood estimations of a class of one-dimensional stochastic differential equation models from discrete data. *Journal of Time Series Analysis* 22(5), 505–515.
- GUTIÉRREZ, R., GUTIÉRREZ-SÁNCHEZ, R., NAFIDI, A. and RAMOS, E. (2006): A new stochastic Gompertz diffusion process with threshold parameter: Computational aspects and applications. *Appl. Math. Comput.* 183, 738–747.
- GUTIÉRREZ, R., GUTIÉRREZ-SÁNCHEZ, R., NAFIDI, A. and RAMOS, E. (2009): Three-parameter stochastic lognormal diffusion model: statistical computation and simulating annealing. Application to real case. *J. Stat. Comput. Simul.* 79(1), 25–38.
- GUTIÉRREZ, R., GUTIÉRREZ-SÁNCHEZ, R. and NAFIDI, A. (2009): The trend of the total stocks of the private car-petrol in Spain: Stochastic modelling using a new Gamma Diffusion Process. *Appl. Energy* 86, 18–24.
- LIPTER, R.S. and SHIRYAYEV, A.N. (1978): *Statistics of Random Processes II. Applications*. Springer-Verlag, New York.
- PRAKASA RAO, B.S.L. (1999): *Statistical inference for diffusion type process*. Arnold, London and Oxford University Press, New York, 1999.
- VERA, J.F. and DÍAZ-GARCÍA J. (2008): A global simulated annealing heuristic for the three-parameter lognormal maximum likelihood estimation. *Comput. Stat. Data Anal.* 52(12), 5055–5065.
- WONG, E. and HAJEK, B. (2008): *Stochastic processes in engineering systems*. New York, 1985.
- ZEHN, P.W. (1966): Invariance of maximum likelihood estimators. *Ann. Math. Stat.* 37, 744.

# On the Correlated Gamma Frailty Model for Bivariate Current Status Data

Niel Hens<sup>1,2</sup> and Andreas Wienke<sup>3</sup>

<sup>1</sup> Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University

Campus Diepenbeek, Agoralaan 1, 3590 Diepenbeek, Belgium

*niel.hens@uhasselt.be*

<sup>2</sup> Centre for Health Economics Research and Modeling Infectious Diseases, Centre for the Evaluation of Vaccination, Vaccine & Infectious Disease Institute, University of Antwerp

Campus Drie Eiken, Universiteitsplein 1, 2610 Antwerpen, Belgium

*niel.hens@ua.ac.be*

<sup>3</sup> Institute of Medical Epidemiology, Biostatistics and Informatics, Medical Faculty, Martin Luther University Halle-Wittenberg

Magdeburger Strasse 8, 06097 Halle, Germany,

*andreas.wienke@medizin.uni-halle.de*

**Abstract.** Frailty models are often used to study the individual heterogeneity in multivariate survival analysis. Whereas the shared frailty model is widely applied, the correlated frailty model has gained attention because it elevates the restriction of unobserved factors to act similar within clusters. Estimating frailty models is not straightforward due to various types of censoring. In this manuscript focus is on Type II interval censored data commonly known as current status data for which we study the behavior of the correlated gamma frailty model. We show that the correlated frailty model for bivariate current status data is identifiable when using a parametric baseline hazard but that misspecification of the frailty model or the baseline hazard can lead to biased estimates.

**Keywords:** correlated frailty, bivariate binary data, Type II interval censored data

## 1 Introduction

We motivate our research from infectious disease epidemiology where a key parameter is the so-called force of infection or the per capita rate at which a susceptible person acquires the infection. In endemic equilibrium and under the assumption of lifelong immunity, the force of infection can be estimated from serological data providing information on whether or not an individual has been infected before, thus constituting current status data (Hens et al., 2010). These serological samples are often tested for more than one infection and as a result the heterogeneity in the acquisition of infections can be estimated using for example a shared gamma frailty (Farrington et al., 2001).

In a shared frailty model, the frailty (or random effect) is common to the individuals in the group, and is thus responsible for creating dependence. The shared frailty model abounds in the literature on frailty models and was extensively studied in the monographs by Hougaard (2000), Therneau and Grambsch (2000) and Duchateau and Janssen (2008).

The correlated frailty model is a natural extension of the shared frailty model. In the correlated frailty model, the frailties of individuals in a cluster are correlated, but not shared. It enables the explicit inclusion of additional correlation parameters, whereas in the shared frailty approach all correlations between group members are equal.

We propose the use of the correlated frailty model as an extension of the shared frailty model for the analysis of bivariate current status data.

## 2 Methodology

Denote by  $\lambda_i(t, Z_i)$  the hazard function at time  $t$  conditional on the frailty  $Z_i$  ( $i = 1, 2$ ). The corresponding conditional survival function  $S_i(t|Z_i)$  is then given by

$$S_i(t|Z_i) = e^{-\int_0^t \lambda_i(s, Z_i) ds}, \quad (1)$$

which we combine with the proportional hazards assumption  $\lambda_i(t, Z_i) = Z_i \lambda_{i0}(t)$  to obtain

$$S_i(t|Z_i) = e^{-Z_i \int_0^t \lambda_{i0}(s) ds}. \quad (2)$$

The unconditional survival function can be obtained by integrating out the random frailty  $Z_i$  by using the Laplace transform  $\mathbf{L}_i$  of  $Z_i$  ( $i = 1, 2$ ):

$$S_i(t) = \mathbf{E}S_i(t|Z_i) = \mathbf{L}_i \left( \int_0^t \lambda_{i0}(s) ds \right). \quad (3)$$

Assuming conditional independence, we can formulate the conditional bivariate survival function. Depending on the choice for the bivariate frailty distribution, either an explicit expression for the unconditional survival function can be given or numerical integration is required. In general, numerical integration with respect to the frailty, or random-effects, distribution is not straightforward but has become more accessible through the development of appropriate statistical software and reformulating non-normal random effects, as done by Nelson et al. (2006) and Liu and Yu (2007). In what follows, we will focus on the gamma frailty distribution as the most often used frailty distribution because of its explicit solution for the unconditional survival function (see e.g. Hougaard, 2000 and Duchateau and Janssen, 2008), which owes to conjugacy properties.

The explicit expression for the unconditional joint survival function for the correlated gamma frailty model is given by (Yashin et al., 1995):

$$S(t_1, t_2) = [S_1(t_1)]^{1-\frac{\sigma_1}{\sigma_2}\rho} [S_2(t_2)]^{1-\frac{\sigma_2}{\sigma_1}\rho} [S_1^{-\sigma_1^2}(t_1) + S_2^{-\sigma_2^2}(t_2) - 1]^{-\frac{\rho}{\sigma_1\sigma_2}}. \quad (4)$$



where  $\sigma_1, \sigma_2$  and  $\rho$  are the standard deviations of, and correlation between, both frailty variables. Note that if  $Z_1 = Z_2$ , and thus  $\sigma_1 = \sigma_2 = \sigma, \rho = 1$ , we end up with the shared gamma frailty model. Using expression (4) one can express the likelihood function under various censoring schemes (Hens et al., 2009).

Note that in case of current status data without any covariates, the model is not identifiable using a nonparametric baseline hazard (Chang et al., 2007), motivating the use of a parametric baseline hazard function such as, for example, the Gompertz baseline hazard where  $\lambda_{i0}(t) = a_i \exp(b_i t)$ , ( $i = 1, 2$ ). Although we do not investigate the sufficient conditions required for the model to be identifiable in this paper, we rely on the more general methodology of detecting parameter redundancy (Catchpole and Morgan, 1997, 2001).

### 3 Simulations

The results in this section are partly taken from Hens et al. (2009). Table 1 summarizes the results of a simulation study showing the unbiasedness of the correlated frailty model under various censoring schemes. Table 2 summarizes the results of a simulation study showing that a misspecification of the frailty model could lead to biased estimates.

**Table 1.** Averaged parameter estimates and empirical standard errors for the simulation study of the correlated gamma frailty model with uncensored time to event; right censored and current status data using a Gompertz baseline hazard ( $\lambda_{i0}(t) = a_i \exp(b_i t)$ ,  $i = 1, 2$ ).

parameter	true value	uncensored time to event		right censored data		current status data	
		mean	(e.s.e.)	mean	(e.s.e.)	mean	(e.s.e.)
$\sigma_1$	1.600	1.604	(0.113)	1.621	(0.466)	1.694	(1.854)
$\sigma_2$	1.000	0.999	(0.068)	1.056	(0.214)	1.179	(0.920)
$\rho$	0.500	0.501	(0.035)	0.540	(0.169)	0.636	(0.257)
$a_1$	0.006	0.006	(0.001)	0.006	(0.001)	0.006	(0.001)
$b_1$	0.020	0.020	(0.002)	0.022	(0.010)	0.045	(0.420)
$a_2$	0.008	0.008	(0.001)	0.008	(0.001)	0.008	(0.001)
$b_2$	0.030	0.030	(0.003)	0.032	(0.007)	0.048	(0.228)

We also studied the performance of the correlated frailty model when misspecifying the baseline hazard. We generated bivariate data sets based on Weibull and exponential baseline hazards rates with the same correlated gamma distributed frailty as before:  $\sigma_1 = 1.6, \sigma_2 = 1$  and  $\rho = 0.5$ ; and then fitted those data sets using the correlated gamma frailty models with Weibull, exponential and Gompertz baseline, respectively. Table 3 summarizes the results of this simulation study indicating that a correctly specified baseline hazard leads to the lowest AIC-value. Since both the Weibull and Gompertz

**Table 2.** Averaged parameter estimates and empirical standard errors for the simulation study on the misspecification of the frailty distribution for current status data using a Gompertz baseline hazard ( $\lambda_{i0}(t) = a_i \exp(b_i t)$ ,  $i = 1, 2$ ).

	true	correlated frailty	common variance	CF shared frailty	univariate frailty
par	value	mean (e.s.e.)	mean (e.s.e.)	mean (e.s.e.)	mean (e.s.e.)
$\sigma_1$	1.600	1.694 (1.854)	1.185 (0.429)	0.769 (0.051)	1.962 (2.219)
$\sigma_2$	1.000	1.179 (0.920)	1.185 (0.429)	0.769 (0.051)	1.107 (0.941)
$\rho$	0.500	0.636 (0.257)	0.679 (0.219)	1.000 (-)	0.000 (-)
$a_1$	0.006	0.006 (0.001)	0.006 (0.001)	0.006 (0.001)	0.006 (0.005)
$b_1$	0.020	0.045 (0.420)	0.013 (0.009)	0.007 (0.004)	0.062 (0.135)
$a_2$	0.008	0.008 (0.001)	0.008 (0.001)	0.008 (0.001)	0.008 (0.001)
$b_2$	0.030	0.048 (0.228)	0.039 (0.019)	0.024 (0.003)	0.047 (0.047)

are extensions of the exponential baseline hazard the difference in AIC refers to the number of parameters used. Note that the frailty parameters are consistently estimated for this setting. When data are generated according to a Weibull baseline hazard and fitted using an exponential or Gompertz baseline hazard, the fit worsens in terms of the AIC and moreover the estimates of the frailty parameters are biased.

**Table 3.** Simulation results for the correlated frailty model under misspecification of the baseline hazard. The first letter in the scheme denotes the generating function with W: Weibull and E: Exponential. The second letter in the scheme denotes the fitted function with W: Weibull, G: Gompertz and E: Exponential. The final two letters denote RC: right censored data and CS: current status data. Note that AIC values can only be compared for data generated under the same mechanism: i.e. the same baseline hazard and the same censoring scheme.

Scheme	AIC	$\sigma_1$	$\sigma_2$	$\rho$
EERC	24912	1.599 (0.019)	1.000 (0.014)	0.501 (0.019)
EGRC	24914	1.599 (0.023)	1.001 (0.017)	0.501 (0.019)
EWRC	24914	1.600 (0.034)	0.999 (0.024)	0.502 (0.019)
EECS	5045	1.601 (0.057)	0.998 (0.042)	0.500 (0.050)
EGCS	5049	1.601 (0.126)	1.010 (0.087)	0.508 (0.053)
EWCS	5049	1.619 (0.459)	1.078 (0.272)	0.559 (0.104)
WWRC	36413	1.601 (0.033)	1.000 (0.023)	0.501 (0.019)
WGRC	37733	1.063 (0.025)	1.566 (0.071)	0.448 (0.023)
WERC	37939	1.134 (0.016)	0.417 (0.016)	0.368 (0.016)
WWCS	3797	1.632 (0.307)	1.031 (0.145)	0.525 (0.081)
WGCS	3816	1.128 (0.113)	0.601 (0.078)	0.535 (0.072)
WECS	3815	1.241 (0.041)	0.581 (0.041)	0.468 (0.036)

## 4 Hepatitis A and B

We reanalyzed the bivariate current status data (whether or not past infection occurred) on hepatitis A and B presented by Hens et al. (2009) using a Gompertz and Weibull baseline hazard. Table 4 shows the standard deviation estimates and their standard error together with the AIC-value for the different correlated frailty models. The correlated frailty model with common variance and Gompertz baseline hazard has the lowest AIC-value among the models under consideration. The models using the Gompertz baseline hazard have a lower AIC-value compared to those using the Weibull baseline hazard. Note that among the Weibull models, the shared frailty model has the lowest AIC-value and that the correlation estimate changes from 0.17 (0.06) to 1.00 (0.10) when imposing a common variance, albeit the AIC-values are very similar. It is clear from these results that the choice of the baseline hazard is very important and that a model comparison using different baseline hazards is quintessential.

**Table 4.** Hepatitis A & B analysis. Variance estimates and standard errors for the different frailty distributions using a Gompertz and Weibull baseline hazard together with their AIC-values.

parameter	correlated frailty	common variance	CF shared frailty	univariate frailty
<i>Gompertz baseline hazard</i>				
$\sigma_1$	1.63 (0.17)	1.63 (0.18)	0.72 (0.08)	1.42 (0.18)
$\sigma_2$	1.19 (0.68)	1.63 (0.18)	0.72 (0.08)	2.74 (23.1)
$\rho$	0.66 (0.37)	0.49 (0.08)	1 (-)	0 (-)
AIC	5667.4	5665.6	5697	5694.8
<i>Weibull baseline hazard</i>				
$\sigma_1$	0.63 (0.11)	0.55 (0.09)	0.55 (0.08)	0.38 (0.19)
$\sigma_2$	3.70 (1.44)	0.55 (0.09)	0.55 (0.08)	4.37 (2.51)
$\rho$	0.17 (0.06)	1.00 (0.10)	1 (-)	0 (-)
AIC	5694.8	5694.5	5692.5	5705.8

## 5 Discussion

In this manuscript, we focused on the performance of the correlated gamma frailty model for bivariate data under various censoring schemes including Type II interval censored data, also known as current status data. We have shown that misspecification of both the frailty model and/or the baseline hazard leads to biased estimates in the case of current status data. These results shed a first light on the use of the correlated gamma frailty model for bivariate current status data. This situation typically applies in infectious disease epidemiology where multisera data constituting multivariate current

status data are studied to quantify the heterogeneity in acquisition of infections using a shared gamma frailty (Farrington et al. 2001). The use of correlated frailty model facilitates the separation of heterogeneity and correlation. Studying this correlation could indicate whether different infections are transmitted via the same routes. This could prove worthwhile for diseases for which the transmission route is unknown. We used data on hepatitis A and B, two infections transmitted through different routes, to illustrate the importance of selecting the baseline hazard.

### Acknowledgments

We thank Marc Aerts, Tom Cattaert, Jianxing Chen, Elasma Milanzi and Geert Molenberghs for many helpful discussions and contributions to our work. This work has been funded by “SIMID”, a strategic basic research project funded by the institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), project number 060081 and by the IAP research network nr P6/03 of the Belgian Government (Belgian Science Policy). Andreas Wienke was supported by the German Research Council, project number WI 3288/1-1.

### References

- CATCHPOLE, E. and MORGAN, B. (1997): Detecting parameter redundancy. *Biometrika* 84, 187-196.
- CATCHPOLE, E. and MORGAN, B. (2001): Deficiency of parameter redundant models. *Biometrika* 88, 593-598.
- CHANG, I., WEN, C. and WU, Y. (2007): A profile likelihood theory for the correlated gamma-frailty model with current status family data. *Statistica Sinica* 17, 1023-1046.
- DUCHATEAU, L. and JANSSEN, P. (2008): *The Frailty Model*. Springer, New York.
- FARRINGTON, C., KANAAN, M. and GAY, N. (2001): Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Applied Statistics* 50, 251-292.
- HENS, N., AERTS, M., FAES, C., SHKEDY, Z., LEJEUNE, O., VAN DAMME, P. and BEUTELS, P. (2010): Seventy five years of estimating the force of infection from current status data. *Epidemiology and Infection*, In press.
- HENS, N., WIENKE, A., AERTS, M. and MOLENBERGHS, G. (2009): The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data. *Statistics in Medicine* 28, 2785-2800.
- HOUGAARD, P. (2000): *Analysis of Multivariate Survival Data*. Springer, New York.
- LIU, L. and YU, Z. (2007): A likelihood reformulation method in non-normal random effects models. *Statistics in Medicine* 27, 3105-3124.

- NELSON, K., LIPSITZ, S., FITZMAURICE, G., IBRAHIM, J.G., PARZEN, M. and STRAWDERMAN, R. (2006): Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects. *Journal of Computational and Graphical Statistics* 15, 39-57.
- THERNEAU, T. and GRAMBSCH, P. (2000): *Modelling Survival Data*. Springer, New York.
- YASHIN, A., VAUPEL, J. and IACHINE, I. (1995): Correlated individual frailty: An advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies* 5, 145-159.



# Evolutionary Stochastic Portfolio Optimization and Probabilistic Constraints

Ronald Hochreiter<sup>1</sup>

Department of Finance, Accounting and Statistics, WU Vienna University of Economics and Business. Augasse 2-6, A-1090 Vienna, Austria.  
*ronald.hochreiter@wu.ac.at*

**Abstract.** In this paper, we extend an evolutionary stochastic portfolio optimization framework to include probabilistic constraints. Both the stochastic programming-based modeling environment as well as the evolutionary optimization environment are ideally suited for an integration of various types of probabilistic constraints. We show an approach on how to integrate these constraints. Numerical results using recent financial data substantiate the applicability of the presented approach.

**Keywords:** probabilistic constraints, portfolio optimization, stochastic optimization, evolutionary algorithms

## 1 Introduction

Stochastic programming is a powerful method to solve optimization problems under uncertainty, see [?] for theoretical properties and [?] for an overview of possible applications. One specific application area is portfolio optimization, which was pioneered by [?]. Given a set of financial assets with unknown returns, we like to create a portfolio of this assets, which solve the classical bi-criteria problem of an investor. That is to maximize the expected return, while minimizing the inherent risk. The risk is measured using some statistical functional over the expected loss distribution, see also [?].

The advantage of using the stochastic programming approach is that the optimization can be done without using a covariance matrix of the assets, which is on one hand hard to estimate and on the other hand does not capture the uncertainty in sufficient detail, especially if there are many assets. Instead of using the asset means and the covariance matrix, a stochastic programming approach uses a set of scenarios, each weighted by a certain probability. In the specific portfolio optimization context one scenario is a set of one possible asset return per asset under consideration - see below for more details or e.g. see Chapter 16 of [?].

Evolutionary algorithms have proven to be an optimal framework to solve portfolio optimization problems especially because of the rather direct genotype-phenotype representation scheme.

This paper is organized as follows: Section 2 summarizes the evolutionary approach, which was used to solve the stochastic portfolio optimization problems. Section 3 adds probabilistic constraints to the standard optimization problem and shows an approach on how to integrate these type of constraints, which represents the main contribution of this work. Section 4 summarizes numerical results using two different sets of recent financial data, while Section 4 concludes the paper.

## 2 Evolutionary Stochastic Portfolio Optimization

We follow the approach taken by [?] and [?], which builds a stochastic programming-based environment on top of the general evolutionary portfolio optimization findings reported by [?], [?], and [?].

### 2.1 Stochastic portfolio optimization

Let us define the stochastic portfolio optimization problem as follows. We consider a set of assets (or asset categories)  $\mathcal{A}$  with cardinality  $a$ . Furthermore, we base our decision on a scenario set  $\mathcal{S}$ , which contains a finite number  $s$  of scenarios each containing one uncertain return per asset. Each scenario is equipped with a non-negative probability  $p_s$ , such that  $\sum_{s \in \mathcal{S}} p_s = 1$ . The scenario set can be composed of historical data or might be the output of a scenario simulator.

For every portfolio  $x$  we can easily calculate the profit and loss distribution by multiplying the scenario matrix with the portfolio weighted by the respective probability. Let us denote the profit function of a certain portfolio  $x$  by  $\pi(x)$  and the loss function by  $\ell(x) = -\pi(x)$ .

Every portfolio optimization procedure is a multi-objective optimization. In the traditional case it is a trade-off between return and risk of the profit and loss function. We do not want to employ an multi-objective approach such that we use the classical Markowitz approach and use the Variance of the loss distribution for the risk dimension, which we want to minimize, and the expectation of the profit function for the reward dimension, on which we want to set a lower limit. Hence, the main optimization problem is shown in Eq. (1) below.

$$\begin{aligned} &\text{minimize} && \text{Variance}(\ell_x) \\ &\text{subject to} && \mathbb{E}(\pi_x) > \mu \end{aligned} \tag{1}$$

Furthermore we consider the classical constraints, i.e. budget normalization, as well as setting an upper and lower limit on each asset position, as shown in Eq. (2). These are naturally fulfilled by the evolutionary approach shown below, especially since we restrict short-selling in our case.

$$\begin{aligned} &\text{subject to} && \sum_{a \in \mathcal{A}} x_a = 1 \\ &&& l \leq x_a \leq u \quad \forall a \in \mathcal{A} \end{aligned} \tag{2}$$



## 2.2 Evolutionary stochastic portfolio optimization

The evolutionary algorithm chosen is based on the commonly agreed standard as surveyed by [?] and is based on the literature cited at the beginning of this Section.

We use the following genetic encoding of a portfolio. Each gene consists of two parts: One that determines the amount of budget to be distributed to each selected asset and one part which determines in which assets to invest. The first part  $g_1$  consists of a predefined number  $b$  of real values between 0 and 1 and the second part  $g_2$  is encoded as a bit-string of the size of the amount of assets.

The following evolutionary operators have been implemented and used:

- Elitist selection. A certain number  $o_1$  of the best chromosomes will be used within the next population.
- Crossover. A number  $o_2$  of crossovers will be added, both 1-point crossovers ( $g_1$  and  $g_2$ ) and intermediate crossovers (only  $g_1$ ).
- Mutation.  $o_3$  mutated chromosomes will be added
- Random addition. Furthermore  $o_4$  randomly sampled chromosomes are added, which are also used for creating the initial population.

The specific number of operators  $o = (o_1, o_2, o_3, o_4)$  has to be determined for each concrete number of assets  $a$  as well as the parameter  $b$ .

## 3 Probabilistic Constraints

The main advantage of the stochastic programming approach is that the complete distribution is naturally available and can be used for optimization purposes by integrating these directly into the constraint handling mechanism. In the most simplest case, we want to restrict that the probability that the loss exceeds a certain value  $\delta$  is lower than a specified probability  $\varepsilon$ . Given our profit function  $\pi_x$ , the constraint we want to add to our optimization model is given in Eq. (3).

$$\text{subject to } \mathbb{P}(\pi_x \leq \delta) \leq \varepsilon. \quad (3)$$

We will not treat probabilistic constraints as a hard constraint, but use the size of the violation for adding a penalty to the objective function. Let the fitness value which we aim to minimizing be  $f$ . We calculate the penalty  $p$  by

$$p = f \times (\mathbb{P}(\pi_x \leq \delta) - \varepsilon) \times \gamma,$$

where  $\gamma$  is a factor to control the penalization level. The fitness value used for evolutionary optimization purposes is thus given by  $f' = f + p$ . Such a constraint can be implemented and handled conveniently.

## 4 Numerical Results

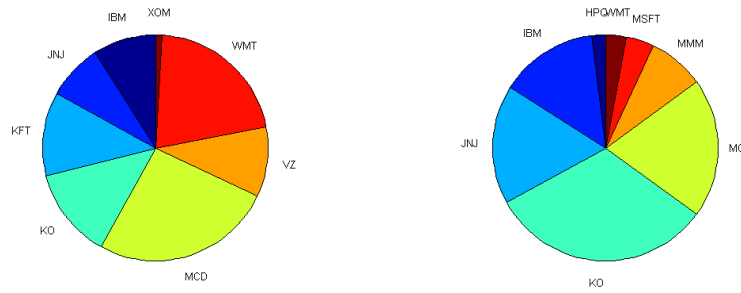
The program code was implemented using MatLab 2008b without using any further toolboxes.

We are using two different sets of data. The first set - DJIA - consists of data from the 30 stocks of the Dow Jones Industrial Average index (at the beginning of 2010), i.e. the ticker symbols AA, AXP, BA, BAC, CAT, CSCO, CVX, DD, DIS, GE, HD, HPQ, IBM, INTC, JNJ, JPM, KFT, KO, MCD, MMM, MRK, MSFT, PFE, PG, T, TRV, UTX, VZ, WMT, XOM. Daily data from each trading day in 2009 has been used. The second data set - labeled S&P100 - consists of 98 stocks out of the Standard & Poors 100 Index (constituents as of March 2010). Daily data from each trading day in 2008 and 2009 has been used. In both cases, weekly returns have been calculated.

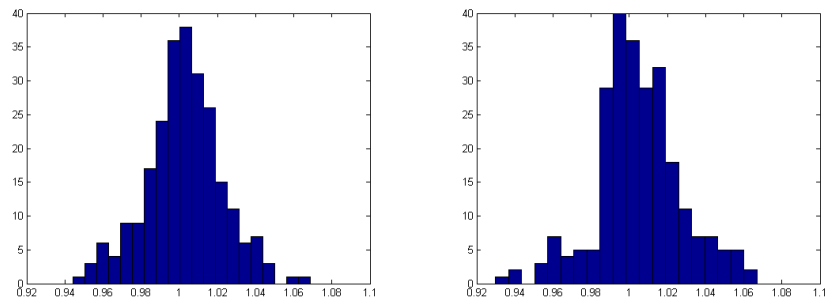
We are using  $b = 100$  buckets which are distributed to the respective asset picks, such that each chromosome has a length of  $b + a$ . The initial population consists of 1000 random chromosomes. The operator structure defined in Section 2.2 is  $o = (100, 420, 210, 100)$ , such that each follow-up population has a size of 830. This number is due to the different combinations between the crossovers and mutations of  $g_1$  and  $g_2$ .

First, we optimize without using the probabilistic constraints, i.e. the main optimization problem given Eq. 1 using  $\mu = 0.001$ . Then we add the probabilistic constraint shown in Eq. 3 with  $\delta = 0.975$  and  $\varepsilon = 0.1$ .

Fig. 1 shows the optimal portfolio without applying the probabilistic constraint  $P_1$  (left) and the optimal one using the probabilistic constraint (right) for the DJIA set, and Fig. 3 for the S&P100 set respectively. Analyzing e.g. the DJIA results one can see that the allocation changed considerably. The diversification has not changed, i.e. three assets (KFT, VZ, XOM) are dropped from the portfolio, and three others are picked (HPQ, MSFT, MMM). The resulting loss distributions are shown in Fig. 2, where the impact of the probabilistic constraint is immediately visible. Furthermore, the statistical properties of the portfolios are shown in Table 1. In this table, the naive  $1/N$  portfolio  $P_3$  has been added for comparison purposes. While  $P_3$  provides the highest expected return, it is also the most risky one in terms of both risk parameters - standard deviation and the probability to fall below the specified threshold. Another interesting fact is that the probabilistic constrained portfolio yields a higher expected return than the standard optimal portfolio. This is partly due to the fact that the lower level  $\mu$  has been set to a level below the expected return of the standard portfolio but is definitely another indicator that the plain classical Markowitz approach should not be used for contemporary portfolio optimization purposes. Further results for the S&P100 data set are shown in Fig. 4 and Table 2. Due to the much larger data set, the structure of the results changed, but the central idea of shaping the loss distribution given the probabilistic constraint was fulfilled as expected.



**Fig. 1.** Portfolio  $P_1$  without (left) and with  $P_2$  (right)  $\mathbb{P}$  constraint, DJIA.



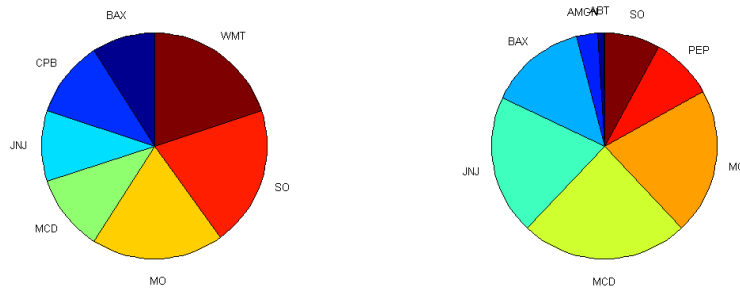
**Fig. 2.** Loss distribution of  $P_1$  (left) and  $P_2$  (right), DJIA.

	$P_1$ (no $\mathbb{P}$ )	$P_2$	$P_3(1/N)$
Mean	0.0024	0.0051	0.0062
Std.Dev.	0.02	0.0225	0.0398
Prob.	0.1774	0.1089	0.2702

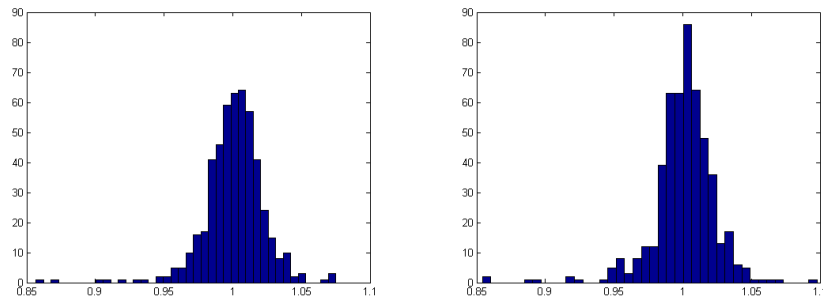
**Table 1.** Statistical properties of the portfolios, DJIA.

	$P_1$ (no $\mathbb{P}$ )	$P_2$	$P_3(1/N)$
Mean	0.0011	0.0009	0.0007
Std.Dev.	0.0220	0.0231	0.0451
Prob.	0.1677	0.1238	0.3134

**Table 2.** Statistical properties of the portfolios, S&P100



**Fig. 3.** Portfolio  $P_1$  without (left) and with  $P_2$  (right)  $\mathbb{P}$  constraint, S&P100.



**Fig. 4.** Loss distribution of  $P_1$  (left) and  $P_2$  (right), S&P100.

## 5 Conclusion

In this paper, an extension of an Evolutionary Stochastic Portfolio Optimization to include probabilistic constraints has been presented. It can be shown that the integration of such constraints is straightforward as the underlying probability space is the main object considered for the optimization evaluation. Numerical results visualized the impact of using such constraints in the area of financial engineering. Future extensions of this work include the integration of risk measures into the probabilistic constraint, e.g. constraining the maximum draw-down of the optimal portfolio.

## References

- P. ARTZNER, F. DELBAEN, J-M. EBER, and D. HEATH (1999): Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

- C. BLUM and A. ROLI (2003): Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys*, 35(3):268–308.
- G. CORNUEJOLS and R. TÜTÜNCÜ (2007): *Optimization methods in finance*. Mathematics, Finance and Risk. Cambridge University Press, Cambridge.
- R. HOCHREITER (2007): An evolutionary computation approach to scenario-based risk-return portfolio optimization for general risk measures. In *EvoWorkshops 2007*, volume 4448 of *Lecture Notes in Computer Science*, pages 199–207. Springer, 2007.
- R. HOCHREITER (2008): Evolutionary stochastic portfolio optimization. In A. Brabazon and M. O'Neill, editors, *Natural Computing in Computational Finance*, volume 100 of *Studies in Computational Intelligence*, pages 67–87. Springer, 2008.
- H. M. MARKOWITZ. (1952): Portfolio selection. *The Journal of Finance*, 7(1): 77–91.
- A. RUSZCZYŃSKI and A. SHAPIRO, eds. (2003): *Stochastic programming*, volume 10 of *Handbooks in Operations Research and Management Science*. Elsevier Science B.V., Amsterdam.
- F. STREICHERT, H. ULMER, and A. ZELL (2003): Evolutionary algorithms and the cardinality constrained portfolio selection problem. In *Selected Papers of the International Conference on Operations Research (OR 2003)*, pages 253–260. Springer.
- F. STREICHERT, H. ULMER, and A. ZELL (2004a): Comparing discrete and continuous genotypes on the constrained portfolio selection problem. In Kalyanmoy D. et al., editor, *Genetic and Evolutionary Computation (GECCO 2004) - Proceedings, Part II*, volume 3103 of *Lecture Notes in Computer Science*, pages 1239–1250. Springer .
- F. STREICHERT, H. ULMER, and A. ZELL (2004b) Evaluating a hybrid encoding and three crossover operators on the constrained portfolio selection problem. In *CEC2004. Congress on Evolutionary Computation, 2004*, volume 1, pages 932–939. IEEE Press.
- S.W. WALLACE and W.T. ZIEMBA, eds. (2005): *Applications of stochastic programming*, volume 5 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM).



# Boosting a Generalized Poisson Hurdle Model

Vera Hofer<sup>1</sup>

Department of Statistics and Operations Research, University of Graz,  
UniversitaetsstraÙe 15/E3, 8010 Graz, Austria, *vera.hofer@uni-graz.at*

**Abstract.** Common boosting techniques are based on estimating one ensemble by means of gradient descent. Count data regressions by means of a generalised Poisson hurdle model consist of three different parameters. Fitting regression functions to all three parameters raises the question how to use boosting techniques. Since a triple of inter-related ensembles ought to be determined, the gradient of the loss is a 3-component vector. A boosting method for this hurdle model using multivariate componentwise least squares is introduced.

**Keywords:** boosting, count data, triple of ensembles

## 1 Introduction

Traditionally, count data are analysed by means of Poisson or negative binomial regression (Hilbe, 2008; Winkelmann, 2008). For many real datasets the assumption on the variance structure of Poisson models does not hold. Negative binomial regression might improve the model (critical view: McCullagh and Nelder, 1997), but the generalised Poisson (GP) distribution (Consul and Jain (1970); Consul (1979)) is a reasonable alternative. Regression models based on GP were first developed in Consul and Famoye (1992) and Famoye (1993), and refined later (cf. Cazdo et al., 2007).

To address an excess number of zeros, zero-inflated models were introduced (Johnson and Kotz, 1969; Mullahy, 1986; Lambert, 1992). They are derived from mixing a count distribution and a point mass at zero. Due to different sources of zeros, interpretation may be complex. In contrast, hurdle models consist of two-components: a hurdle component to account for zeros, and a zero-truncated count component to account for non-zeros. The zero-truncated component follows any zero-truncated count distribution.

In the recent years ensemble methods have been developed to improve the predictive performance of fitting techniques. They are based on the idea of constructing multiple function predictions from the data by means of a “weak” base procedure and use a convex combination of them for final aggregated prediction (Buehlmann and Hothorn, 2007a). The pioneering work by Breiman (1998, 1999) on gradient descent approximation in function space (Lutz and Buehlmann, 2006), created an easy tool to use boosting in regression (Friedman, 2001; Lutz and Buehlmann, 2006).

Common boosting methods involve only one ensemble. For a GP hurdle model that consists of a GP count component and a binomial hurdle component the data might suggest fitting regression functions to all three parameters. This raises the questions how to apply boosting techniques. Using the negative loglikelihood function as loss function, its gradient consists of three components. In contrast to Borisov et al. (2009) who introduced a zero-inflated Poisson model, the three components of the gradient are not fit separately but by multivariate regression. Instead of regression trees (Borisov et al., 2009), componentwise linear least squares are used as “base learner” here.

The predictive performance of the boosting model presented here is investigated by means of two real datasets. The model is compared to a Poisson hurdle model, a negative binomial model, and a negative binomial hurdle model using 5-fold cross-validation and Vuong’s test for non-nested models. a triple of three inter-related ensembles have to be fit.

## 2 Generalised Poisson Regression

### 2.1 Generalised Poisson Distribution

The probability density function,  $p(y|\mu, \phi)$ , of a random variable  $Y$  that follows a generalised Poisson distribution,  $GP$ , with mean  $\mu$ , and dispersion parameter  $\phi$  is  $p(y|\mu, \phi) = \mu W^{y-1} (y!)^{-1} \phi^{-y} e^{-\frac{W}{\phi}}$ , where  $W = \mu + (\phi - 1)y$  and  $\mu > 0$  (Consul and Famoye, 1992). In the present work it is assumed that  $\phi > 1$ . Otherwise  $\phi$  must be restricted to guarantee that  $p(y|\mu, \phi) \geq 0$ . For  $\phi = 1$  the GP reduces to the Poisson distribution,  $\phi > 1$  indicated overdispersion, whereas  $\phi < 1$  indicates underdispersion. Mean and variance of the GP are:  $\mathbb{E}(Y) = \mu$  and  $\text{Var}(Y) = \phi^2 \mu$ .

### 2.2 Generalised Poisson Hurdle Model

A hurdle model is a two-component model consisting of a hurdle component that models zeros versus nonzeros, and a zero-truncated count component to account for the nonzeros. The hurdle at zero is assumed to be a Bernoulli variable  $B(\omega, 1)$  where  $\omega$  is the probability of a zero. The zero-truncated component  $Y_T$  follows any zero-truncated count distribution. In the present analysis this component is chosen to be zero-truncated GP distributed, i.e.  $Y_T \sim GP_T(\mu, \phi, p)$  such that the probability density function is

$$p_T(y|\mu, \phi) = p(y|\mu, \phi) (1 - p(0|\mu, \phi))^{-1} = p(y|\mu, \phi) (1 - e^{-\mu/\phi})^{-1}$$

The probability density function,  $p_H(z|\mu, \phi, \omega)$ , of a random variable  $Y$  following a generalised Poisson hurdle distribution (GPH) then is

$$p_H(y|\mu, \phi, \omega) = 1_{(y=0)} \cdot \omega + 1_{(y>0)} \cdot (1 - \omega) \frac{p(y|\mu, \phi)}{1 - e^{-\mu/\phi}},$$



where  $p(y | \mu, \phi)$  is the generalised Poisson probability density function. Mean and variance of  $Y$  are

$$\mathbb{E}(Y) = \frac{(1 - \omega) \mu}{1 - e^{-\mu/\phi}} \quad \text{Var}(Y) = \frac{\phi^2 \mu (1 - \omega)}{1 - e^{-\mu/\phi}} + \frac{\mu^2 (1 - \omega)(\omega - e^{-\mu/\phi})}{(1 - e^{-\mu/\phi})^2}.$$

In contrast to common regression models, the GPH regression considered below is based on models for the parameters and not for the moments. Similar to a Poisson regression model (McCullagh and Nelder, 1997), the GPH regression assumes

$$\begin{aligned} \text{a. } & Y_i \stackrel{iid}{\sim} GPH(\mu_i, \phi_i, \omega_i). \\ \text{b. } & f_1(\mu_i) = \log(\mu_i) = g(\mathbf{x}_i) \\ \text{c. } & f_2(\phi_i) = \log(\phi_i - 1) = h(\mathbf{z}_i) \\ \text{d. } & f_3(\omega_i) = \log\left(\frac{\omega_i}{1 - \omega_i}\right) = l(\boldsymbol{\xi}_i) \end{aligned} \tag{M1}$$

where  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ ,  $\mathbf{z}_i = (z_{i1}, \dots, z_{iq})$  and  $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iq})$  are vectors of predictor values. There is no need for using the same predictors for all three parameters. The loglikelihood function serves as a loss function for determining the predictors  $g$ ,  $h$ , and  $l$ . It takes the form

$$\begin{aligned} L(Y, g, h, l) = & -1_{(Y=0)} (-\log(1 + e^{-l})) - 1_{(Y>0)} (-\log(1 + e^l) + g + \\ & + (Y - 1) \log(e^g + e^h Y) - \log(Y!) - Y \log(1 + e^h) \\ & - \frac{e^g + e^h Y}{1 + e^h} - \log\left(1 - \exp\left(-\frac{e^g}{1 + e^h}\right)\right) \end{aligned} \tag{1}$$

Formulation (1) allows for nonparametric estimation of  $g$ ,  $h$ , and  $w$ . Thus, this formulation is used in boosting.

### 3 Boosting

Boosting attempts to find a regression function  $F(\mathbf{x}) = \sum_{i=0}^m f_m(\mathbf{x})$  from minimizing expected loss  $\mathbb{E}_{Y|\mathbf{x}} L(y, F)$ . The functions  $f_m$  are simple functions of  $\mathbf{x}$  (“base learners”). The choice of the loss function on the one hand and the base procedure on the other hand yield a variety of different boosted regression models.

#### 3.1 Gradient Descent Boosting

Starting from an initial function  $f_0(\mathbf{x})$ , the expected loss is minimized by following the steepest descent of the expected loss in a forward stagewise manner (Friedman, 2001). To reduce the expected loss in step  $m \geq 1$ , the current argument  $f_{m-1}$  is changed into the direction of its negative gradient

$$\begin{aligned} U_m(\mathbf{x}) &= -\frac{\partial}{\partial f} \mathbb{E}_{Y|\mathbf{x}} (L(Y, F(\mathbf{x})) | \mathbf{x} = \mathbf{x}) |_{f=f_{m-1}(\mathbf{x})} = \\ &= \mathbb{E}_{Y|\mathbf{x}} (-\nabla L(y, f)) |_{f=f_{m-1}(\mathbf{x})} \end{aligned} \tag{2}$$

such that  $f_m = f_{m-1} + \nu U_m$ , where  $\nabla L$  is the gradient of the loss function, i.e. its derivative with respect to  $f$ , and  $\nu$  is the shrinkage parameter. Sufficient regularity is assumed for interchanging integration and differentiation.

In the sample version,  $f_0$  might be chosen as  $f_0 = \arg \min_c \sum_{i=1}^N L(y_i, c)$ . The conditional mean in (2) suggests evaluating the negative gradient of the loss function  $V_i = -\nabla L(y_i, f_{m-1}(\mathbf{x}_i))$  at the given sample and fitting it as the “pseudo-response” to the predictors  $\mathbf{x}_i$  by the “base learner”  $u_m$  to get the direction  $\hat{U}_m(\mathbf{x}) = u_m(\mathbf{x})$ . The regression function then becomes  $f_m = f_{m-1} + \nu u_m$ . The process is iterated until  $m = M$ .

The main tuning parameter is the number of iterations  $M$  which can be determined by means of cross-validation or by using the degree of freedom of the boosting fit (Buehlmann and Hothorn, 2007a; Buehlmann and Hothorn, 2007b). As a rule of thumb, the size of  $\nu$  can be regarded to be minor important, as long as it is small such as  $\nu = 0.1$ . (Buehlmann and Hothorn, 2007a, p. 480; Buehlmann and Hothorn, 2007b, 521; Friedman, 2001).

### 3.2 Componentwise linear least square

As “base learner” simple models such as regression tree or componentwise linear least squares (CLLS) are used. CLLS are very fast in calculation, whereas tree can easily handle predictors measured at any scale (Buehlmann and Hothorn, 2007a), and can even cope with nonlinear structures.

In each boosting step CLLS selects only one predictor in the sense of ordinary least squares fitting (Buehlmann and Hothorn, 2007a). Let  $x^{(j)}$  be the  $j$ -th predictor variable,  $\mathbf{X}^{(j)}$  be the  $j$ -column of the centered design matrix, and let  $\mathbf{V}$  be the vector of pseudo responses obtained from evaluating the gradient at the current solution. The “base learner” has the form  $u_m(x) = \beta^{(s)} x^{(s)}$ , where  $\beta^{(j)} = \|\mathbf{X}^{(j)}\|^{-2} (\mathbf{X}^{(j)})^t \mathbf{V}$ , and

$$s = \arg \min_{1 \leq j \leq p} \sum_{i=1}^N \left( V_i - \beta^{(j)} X_i^{(j)} \right)^2.$$

The parameter selected is updated as  $\beta_m^{(s_m)} = \beta_{m-1}^{(s_m)} + \nu \beta^{(s)}$ , and the regression function becomes  $f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \nu \beta^{(s)} x^{(s_m)}$ .

## 4 Boosting the Generalised Poisson Hurdle Model

Common boosting methods are based on a loss function that involves only one ensemble. Thus, they can only be applied when a regression function is fit only for one parameter. However, the GPH model requires estimating a regression function on all three parameters. When using ensemble techniques, three ensembles must be fit simultaneously. Thus, the GPH boost uses the loglikelihood function in (1) as loss function. This loss function depends on three inter-related regression functions,  $g$ ,  $h$ , and  $l$ . In contrast to common

boosting techniques, the gradient of the GPH boost is a three component vector. At any step  $m > 0$  the pseudo-responses,  $(V_i^g, V_i^h, V_i^l)$  of the three ensembles, are obtained as the negative gradient of the loss function evaluated at the current values  $(g_{m-1}, h_{m-1}, l_{m-1})$  of  $g$ ,  $h$  and  $l$

$$(V_i^g, V_i^h, V_i^l) = \left( -\frac{\partial L}{\partial g}, -\frac{\partial L}{\partial h}, -\frac{\partial L}{\partial l} \right) \Big|_{(y_i, g_{m-1}, h_{m-1}, l_{m-1})} \quad (3)$$

where

$$\begin{aligned} -\frac{\partial L}{\partial g} &= 1_{(y>0)} \left( 1 + \frac{(y-1)e^g}{e^g + ye^h} - \frac{e^g}{1+e^h} - \frac{\exp\left(-\frac{e^g}{1+e^h}\right) \frac{e^g}{1+e^h}}{1 - \exp\left(-\frac{e^g}{1+e^h}\right)} \right) \\ -\frac{\partial L}{\partial h} &= 1_{(y>0)} \left( \frac{y(y-1)e^h}{e^g + ye^h} - \frac{ye^h}{1+e^h} - \frac{e^h(y-e^g)}{(1+e^h)^2} + \frac{\exp\left(-\frac{e^g}{1+e^h}\right) \frac{e^{g+h}}{(1+e^h)^2}}{1 - \exp\left(-\frac{e^g}{1+e^h}\right)} \right) \\ -\frac{\partial L}{\partial l} &= 1_{(y=0)} \left( \frac{1}{1+e^l} \right) - 1_{(y>0)} \left( \frac{1}{1+e^{-l}} \right) \end{aligned}$$

The three pseudo-responses are estimated by means of a multivariate version of CLLS, referred to as MCLLS (cf. Lutz and Buehlmann, 2006, p. 477, who introduce the term “row-boosting”). The method assumes that all three ensembles have the same predictors. In each boosting step only one predictor variable is selected in the sense of Wilks’ lambda (Rencher, 2002, p. 344). Let  $\mathbf{X}^{(j)}$  be the  $j$ -column of the design matrix, and let  $\mathbf{V}$  be the matrix with  $i$ th row  $(V_i^g, V_i^h, V_i^l)$ . The “base learner” has the form  $u_m(\mathbf{x}) = \beta^{(s)} x^{(s)}$ , where  $\beta^{(j)} = (\beta_g^{(s)}, \beta_h^{(s)}, \beta_l^{(s)}) = \|\mathbf{X}^{(j)}\|^{-2} (\mathbf{X}^{(j)})^t \mathbf{V}$ , and

$$s = \arg \min_{1 \leq j \leq p} \frac{\det(\mathbf{V}^t \mathbf{V} - (\beta^{(j)})^t (\mathbf{X}^{(j)})^t \mathbf{V})}{\det(\mathbf{V}^t \mathbf{V} - n \bar{\mathbf{V}}^t \bar{\mathbf{V}})},$$

where  $\bar{\mathbf{V}}$  is the mean gradient, and  $n$  stands for the sample size. This yields the coefficient  $\beta_g^{(s)}$  for the  $\mu$ -ensemble  $g$ ,  $\beta_h^{(s)}$  for the  $\phi$  ensemble  $h$ , and  $\beta_l^{(s)}$  for the  $\omega$  ensemble  $l$ . Then the ensembles are updated as

$$\begin{aligned} g_m(\mathbf{x}) &= g_{m-1}(\mathbf{x}) + \nu \beta_g^{(s)} x^{(s_m)}, & h_m(\mathbf{x}) &= h_{m-1}(\mathbf{x}) + \nu \beta_h^{(s)} x^{(s_m)}, \\ w_m(\mathbf{x}) &= w_{m-1}(\mathbf{x}) + \nu \beta_l^{(s)} x^{(s_m)}. \end{aligned}$$

After  $M$  iteration the parameters take the form

$$\hat{\mu}_i = e^{g_m(\mathbf{x}_i)} \quad \hat{\phi}_i = 1 + e^{h_m(\mathbf{x}_i)} \quad \hat{\omega}_i = \frac{e^{l_m(\mathbf{x}_i)}}{1 + e^{l_m(\mathbf{x}_i)}},$$

The number of iterations,  $M$ , is estimated by means of 5-fold cross-validation. Initial values  $g_0$ ,  $h_0$  and  $w_0$  are obtained from a nonlinear system of equations. The equations are obtained from moment estimators considering a zero-truncated GP distribution. More precisely: Mean and variance of a zero-truncated GP are,

$$\mathbb{E}(Y_T) = \mu_T = \frac{\mu}{1 - e^{-\frac{\mu}{\phi}}} \quad \text{Var}(Y_T) = \sigma_T^2 = \frac{\mu(\mu + \phi^2)}{1 - e^{-\frac{\mu}{\phi}}}.$$

Using moment estimators

$$\hat{\mu}_T = \frac{1}{n_T} \sum_{y_i > 0} y_i \quad \hat{\sigma}_T^2 = \frac{1}{n_T - 1} \sum_{y_i > 0} (y_i - \hat{\mu}_T)^2$$

where  $n_T$  is the number of nonzero observations. Let  $n_0$  be the number of zeros and  $n = n_0 + n_T$  the total sample size. Estimations for the parameters  $\mu$  and  $\phi$  are then obtained from the nonlinear systems of equations with respect to  $\hat{\mu}$  and  $\hat{\phi}$ :

$$\hat{\mu}_T = \frac{\hat{\mu}}{1 - e^{-\frac{\hat{\mu}}{\hat{\phi}}}} \quad \hat{\sigma}_T^2 = \frac{\hat{\mu} \left( \hat{\phi} \left( 1 - e^{-\frac{\hat{\mu}}{\hat{\phi}}} \right) - \hat{\mu} e^{-\frac{\hat{\mu}}{\hat{\phi}}} \right)}{\left( 1 - e^{-\frac{\hat{\mu}}{\hat{\phi}}} \right)^2}$$

Furthermore,  $\hat{\omega}_0 = \frac{n_0}{n}$ . Finally,  $g_0(\mathbf{x}) = \log(\hat{\mathbf{x}})$ ,  $h_0(\mathbf{x}) = \log(\hat{\phi} - 1)$ , and  $l(\mathbf{x}) = \log(\hat{\omega}) - \log(1 - \hat{\omega})$ .

## 5 Application and Results

The GP hurdle boost (GPH) is applied to two real datasets: the first data set is the US National Medical Expenditure Survey 19987/88 which was used by Deb and Trivedi (1997) to investigate the number of physician/non-physician office and hospital outpatient visits of individuals aged 66 and over, who are covered by a particular public insurance program. The data is available for download on the data archive website of the Journal of Applied Econometrics. The second data set is from Fair (1978) who studied extramarital affairs. The data is available in the AER library of R.

Table 1 compares GP hurdle boost (GPH), Poisson hurdle (P), negative binomial (nB), and negative binomial hurdle (nBH) by means of loglikelihood (LogLik) and average loglikelihood (Avg LogLik) for training (train) and testing (test) using a weighted mean due to different data size during 5-fold CV. Furthermore, standard deviation of the loglikelihood per sample unit (Std Avg LogLik) is shown for training and testing. To compare the models by their ability to estimate zeros, root mean squared error of the number of zeros (RMSE zeros) is given for training and testing. Due to the smallest loglikelihood the results of the nBH model are compared to the results of

**Table 1.** Comparison of GP hurdle boost (GPH), Poisson hurdle (P), negative binomial (nB), and negative binomial hurdle (nBH)

	GPH	P	nB	nBH
US National Medical Expenditure Survey with $M = 9308$ iterations				
LogLik train	-9776	-12897	-9735	-9668
LogLik test	-2452	-3250	-2437	-2423
Avg LogLik train	-2.7736	-3.6590	-2.7619	-2.7428
Avg LogLik test	-2.7823	-3.6883	-2.7657	-2.7502
Std Avg LogLik train	0.0027	0.0431	0.0056	0.0049
Std Avg LogLik test	0.0121	0.1776	0.0225	0.0200
RMSE zeros train	44.5515	0.0000	60.3376	0.0000
RMSE zeros test	12.5356	6.2778	16.0971	6.2778
Vuong test value (model versus nBH)	-1.3178	-13.3882*	-6.0281*	
Affairs with $M = 5000$ iterations				
LogLik train	-554.0638	-604.0786	-581.8527	-553.8644
LogLik test	-144.4555	-158.6399	-147.0998	-142.8394
Avg LogLik train	-1.1519	-1.2589	-1.2097	-1.1515
Avg LogLik test	-1.2038	-1.3220	-1.2258	-1.1903
Std Avg LogLik train	0.0060	0.0056	0.0017	0.0036
Std Avg LogLik test	0.0081	0.0332	0.0322	0.0120
RMSE zeros train	10.6489	0.0000	5.2154	0.0000
RMSE zeros test	6.1644	5.7096	5.7271	5.7096
Vuong test (model versus nBH)	-0.2800	-4.8098*	-3.9305*	

the other models by means of Vuong's test (Vuong, 1989). In contrast to the other models, the GP hurdle boost (GPH) does not differ significantly from the nBH.

## 6 Summary

A boosted version of the generalised Poisson hurdle model is introduced using the likelihood function as loss function. In contrast to common boosting techniques, the gradient of the loss function is a 3-component vector since for all three parameters an ensemble ought to be estimated. As base learner a multivariate version of componentwise linear least squares is used. Using two real datasets, the model is compared to a classic Poisson hurdle model, a negative binomial model and a negative binomial hurdle model by means of the loglikelihood as well as by means of Vuong's test. The results show that the boosted GP hurdle model can compete with a negative binomial hurdle model.

## References

- BORISOV, A., RUNGER, G., TUV, E. and LURPONGLUKANA-STRAND, N. (2009): Zero-inflated ensembles for rare event counts. In: N. Adams, C. Rønbardet, A. Siebes, and J.-F. Boulicaut (Eds.): *Advances in Intelligent Data Analysis VIII*. Springer, 225–236.
- BREIMAN, L. (1998): Arcing classifiers. *Annals of Statistics* 26, 801–824.
- BREIMAN, L. (1999): Prediction games and arcing algorithms. *Neural Computation* 11, 1493–1517.
- BUEHLMANN, P. and HOTHORN, T. (2007a): Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science* 22, 477–505.
- BUEHLMANN, P. and HOTHORN, T. (2007b): Rejoinder: Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science* 22 (4), 516–522.
- CONSUL, P.C. and JAIN, G.C. (1970): A generalization of the Poisson distribution. *Technometrics* 15 (4) 791–799.
- CONSUL, P.C. (1979): *Generalized Poisson distributions*. Marcel Dekker, Inc., New York.
- DEP, P. and TRIVEDI, P.K. (1997): Demand for medical care by elderly: a finite mixture approach. *Journal of Applied Econometrics* (12), 313–336.
- CONSUL, P.C. and FAMOYE, F. (1992): Generalized Poisson regression model. *Communications in Statistics, Theory and Methods* 21 (1) 89–109.
- CZADO C., ERHARDT V., MIN A. and WAGNER S. (2007): Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling* (7), 125–153.
- FAMOYE, F. (1993): Restricted generalized Poisson regression model. *Communications in Statistics, Theory and Methods* 22 (5) 1335–54.
- FRIEDMAN, J.H. (2001): Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29, 1189–1232.
- GREEN, W.H. (2003): *Econometric Analysis*. 5th edition. Upper Saddle River, NJ: Prentice Hall. 1004 p.
- HILBE, J.M. (2008): *Negative binomial regression*. Cambridge University Press.
- JOHNSON, N.L. and KOTZ, S. (1969): *Distributions in statistics: discrete distributions*. Wiley: New York.
- LAMBERT, D. (1992): Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics* (34) 1–14.
- LUTZ, R.W. and BUEHLMANN, P. (2006): Boosting for high-multivariate responses in high-dimensional linear regression. *Statistica Sinica* 16, 471–494.
- MCCULLAGH, P. and NELDER, J.A. (1997): *Generalized linear models*. 2nd ed., Chapman and Hall/CRC.
- MULLAHY, J., 1986. Specification and testing in some modified count data models. *Journal of Econometrics* (33), 341–365.
- RENCHE, A. (2002): *Methods of Multivariate Analysis*. 2nd ed. Wiley Series in Probability and Statistics.
- VUONG, Q. H., 1989. Ratio tests for model selection and non-nested hypotheses. *Econometrica* 57 (2), 307–333.
- WINKELMANN, R., 2008. *Economic analysis of count data*. 5th edition, Springer, 333 p.

# Fast and Robust Classifiers Adjusted for Skewness

Mia Hubert<sup>1</sup> and Stephan Van der Veeken<sup>2</sup>

<sup>1</sup> Department of Mathematics - LStat, Katholieke Universiteit Leuven  
Celestijnenlaan 200B, Leuven, Belgium, *Mia.Hubert@wis.kuleuven.be*

<sup>2</sup> Department of Mathematics - LStat, Katholieke Universiteit Leuven  
Celestijnenlaan 200B, Leuven, Belgium, *Stephan.Vanderveeken@wis.kuleuven.be*

**Abstract.** In this paper we propose two new classification rules for skewed distributions. They are based on the adjusted outlyingness (AO), as introduced in Brys et al. (2005) and applied to outlier detection in Hubert and Van der Veeken (2008). The new rules combine ideas of AO with the classification method proposed in Billor et al. (2008). We investigate their performance on simulated data, as well as on a real data example. Throughout we compare the classifiers with the recent approach of Hubert and Van der Veeken (2010) which assigns a new observation to the group to which it attains the minimal adjusted outlyingness. The results show that the new classification rules perform better when the group sizes are unequal.

**Keywords:** robustness, classification, outlyingness

## 1 Introduction

In a classification context, a random sample from a group of  $k$  populations is given and the aim is to construct a rule to classify a new observation into one of the  $k$  populations. Many of the classification methods proposed in the literature rely on quiet strict distributional assumptions such as multivariate normality, or at least elliptical symmetry. Moreover many are sensitive to outliers in the data. Recently we proposed a classification rule based on the adjusted outlyingness (Hubert and Van der Veeken (2010)). This classifier does not rely on any distributional assumption and is robust to outliers. Observations are classified in the group to which they attain minimal adjusted outlyingness (AO). This AO can be seen as a type of projection depth, and hence this approach corresponds to assigning an observation to the group for which it attains *maximal* depth. Consequently this method generalizes the maximum depth classifiers of Ghosh and Chaudhuri (2005). In Billor et al. (2008) a slight modification to the work of Ghosh and Chaudhuri (2005) has been proposed. Observations are classified according to the group for which the *rank* of their depth is maximal. In this paper we propose two modifications of our original rule based on the AO, in the line of Billor et al. (2008). Simulation results and an application to a real data set show that we obtain lower misclassification errors when the group sizes are unequal. In

Section 2 we define the different classification rules. Section 3 contains the results of a simulation study, whereas in Section 4 we apply our rules to a real data set.

## 2 Construction of the classification rules

We assume we have sampled observations from  $k$  different classes  $X^j$ ,  $j = 1, \dots, k$ . The data belonging to group  $X^j$  are denoted by  $\mathbf{x}_i^j$  for  $i = 1, \dots, n_j$ . The dimension of the data space is  $p$  and is assumed to be much smaller than the sample sizes. In Hubert and Van der Veeken (2010) the following classification rule was proposed: for each new observation  $\mathbf{y}$  to be classified, its *adjusted outlyingness* with respect to each group  $X^j$  is calculated. Then  $\mathbf{y}$  is assigned to the group for which its adjusted outlyingness is minimal. The adjusted outlyingness is introduced in Brys et al. (2005) and studied in detail in Hubert and Van der Veeken (2008). It generalizes the Stahel-Donoho outlyingness towards skewed data. The skewness is estimated by means of the medcouple (MC), a robust measure of skewness (Brys et al. (2004)). For univariate data, the adjusted outlyingness of an observation  $x_i^j$  w.r.t. its group  $X^j$  is defined as:

$$\text{AO}^{(1)}(x_i^j, X^j) = \begin{cases} \frac{x_i^j - \text{med}(X^j)}{c_2 - \text{med}(X^j)} & \text{if } x_i^j > \text{med}(X^j) \\ \frac{\text{med}(X^j) - x_i^j}{\text{med}(X^j) - c_1} & \text{if } x_i^j < \text{med}(X^j) \end{cases} \quad (1)$$

where  $c_1$  corresponds to the smallest observation greater than  $Q_1 - 1.5e^{-4}\text{MC IQR}$ , and  $c_2$  to the largest observation smaller than  $Q_3 + 1.5e^3\text{MC IQR}$ . The notations  $Q_1$  and  $Q_3$  stand for the first and third quartile of the data, and  $\text{IQR} = Q_3 - Q_1$  is the interquartile range. This definition assumes that the data are right skewed, which is concluded when  $\text{MC} > 0$ . If the medcouple is negative, the  $\text{AO}^{(1)}$  is computed on the inverted data  $-X^j$ . In order to define the adjusted outlyingness for a multivariate data point  $\mathbf{x}_i^j$ , the data are projected on all possible directions  $\mathbf{a}$  and the  $\text{AO}^{(1)}$  is computed. The overall  $\text{AO}_i^j$  is then defined as the supremum over all univariate  $\text{AO}$ 's:

$$\text{AO}_i^j = \text{AO}(\mathbf{x}_i^j, X^j) = \sup_{\mathbf{a} \in \mathbb{R}^p} \text{AO}^{(1)}(\mathbf{a}^t \mathbf{x}_i^j, X^j \mathbf{a}). \quad (2)$$

Since in practice it is impossible to consider all possible directions, we use  $m = 250p$  directions. Random directions are generated as the direction perpendicular to the subspace spanned by  $p$  observations, randomly drawn from  $X^j$ . As such, the  $\text{AO}$  is invariant to affine transformations of the data. Note that this procedure can only be applied in our classification setting when  $p < n_j$ , and when the dimension  $p$  is not too large (say  $p < 10$ ). Otherwise taking  $250p$  directions is insufficient and more directions are required



to achieve good estimates. We do not consider this as an important drawback of our rules as it is well known that skewness is only an issue in small dimensions (when the dimensionality increases, the data are more and more concentrated in an outer shell of the distribution). Of course, the algorithm can be easily adapted to search over more than  $m$  directions, but this will come at the cost of more computation time.

To apply our new classification rules, we have to define the outlyingness of a *new observation*  $\mathbf{y}$  w.r.t. each group  $X^j$ . One approach would be to compute this outlyingness  $\text{AO}(\mathbf{y}, \tilde{X}^j)$  according to (2), with  $\tilde{X}^j$  the augmented data set  $\tilde{X}^j = \{\mathbf{x}_1^j, \dots, \mathbf{x}_{n_j}^j, \mathbf{y}\}$ . This would of course become computationally very demanding when many new observations need to be classified, as then the augmented data set is modified for each new observation and the median, the IQR and the medcouple have to be recomputed each time on every projection. Hence, we compute the outlyingness of  $\mathbf{y}$  w.r.t. a fixed data set which does not include  $\mathbf{y}$ . A natural candidate is of course  $X^j$  itself. However, we obtain better results when we first remove the outliers from  $X^j$ . As explained in Hubert and Van der Veeken (2008) this can be easily performed by first computing the adjusted outlyingness of all observations  $\text{AO}_i^j$  in group  $X^j$ . Then the univariate outlyingness of every  $\text{AO}_i^j$  is computed. Formally we can define the *outlier score*  $\text{OS}_i = \text{AO}^{(1)}(\text{AO}_i^j, \{\text{AO}_i^j\})$ . Observations with a 'too large' outlyingness can be defined as those  $\mathbf{x}_i^j$  for which  $\text{AO}_i^j > \text{med}(\text{AO}_i^j)$  and  $\text{OS}_i > 1$  (or equivalently with  $\text{AO}_i^j > c_2$ ). Those observations are removed from  $X^j$ , yielding  $\tilde{X}^j$ . To compute the outlyingness of a new case  $\mathbf{y}$ , we then consider  $\text{AO}(\mathbf{y}, \tilde{X}^j)$ , so we fixed the median, IQR and medcouple of the projected outlier-free data from group  $j$ . Further we denote  $\{\tilde{\text{AO}}^j\}$  as the set of AO values of the outlier-free group  $\tilde{X}^j$  of size  $\tilde{n}_j$ .

We now consider the following classification rules:

- **Rule 1:** The observation  $\mathbf{y}$  is assigned to the group  $j$  for which  $\text{AO}(\mathbf{y}, \tilde{X}^j)$  is minimal. This is the classification method proposed in Hubert and Van der Veeken (2010).
- **Rule 2:** Let  $r_y^j$  be the 'rank' of  $\text{AO}(\mathbf{y}, \tilde{X}^j)$  with respect to the  $\{\tilde{\text{AO}}^j\}$ , formally defined as

$$r_y^j = \frac{1}{\tilde{n}_j} \sum_{i=1}^{\tilde{n}_j} I(\tilde{\text{AO}}_i^j \leq \text{AO}(\mathbf{y}, \tilde{X}^j)).$$

The observation  $\mathbf{y}$  is then assigned to the group  $j$  for which  $r_y^j$  is minimal. In case of ties, rule 1 is applied. This is the approach which follows closely Billor et al. (2008).

- **Rule 3:** To measure the position of  $\text{AO}(\mathbf{y}, \tilde{X}^j)$  within the  $\{\tilde{\text{AO}}^j\}$ , we do not use the rank, but a distance which is related to the definition of the univariate AO given in (1). Let in general

$$\text{SAO}^{(1)}(x, X) = \text{AO}^{(1)}(x, X) \text{sign}(x - \text{med}(X^j))$$

be the *signed* adjusted outlyingness of an observation  $x$  with respect to a univariate data set  $X$ . The observation  $\mathbf{y}$  is then assigned to the group  $j$  for which  $\text{SAO}^{(1)}(\text{AO}(\mathbf{y}, \tilde{X}^j), \{\tilde{\text{AO}}^j\})$  is minimal.

### 3 Simulation results

In this section, we compare the different classifiers on several simulated data sets. We consider the two-class problem ( $k = 2$ ). In all simulation settings, one uncontaminated group is generated from a normal distribution, while the other uncontaminated cases come from a skew-normal distribution (Azzalini and Dalla Valle (1996)). Using the notation  $\mathbf{0}_p = (0, 0, \dots, 0)^t \in \mathbb{R}^p$ , a  $p$ -dimensional random variable  $Z$  is said to be standard skew-normal distributed  $\text{SN}_p(\mathbf{0}_p, \tilde{\Omega}, \boldsymbol{\alpha})$  if its density function is of the form

$$f(\mathbf{x}) = 2\phi_p(\mathbf{x}; \tilde{\Omega})\Phi(\boldsymbol{\alpha}^t \mathbf{x})$$

where  $\phi_p(\mathbf{x}; \tilde{\Omega})$  is the  $p$ -dimensional normal density with zero mean and correlation matrix  $\tilde{\Omega}$ ,  $\Phi$  is the c.d.f. of the standard normal distribution and  $\boldsymbol{\alpha}$  is a  $p$ -dimensional vector that regulates the skewness. By adding location and scale parameters, we obtain  $X = \boldsymbol{\mu} + \boldsymbol{\omega}^t Z \sim \text{SN}_p(\boldsymbol{\mu}, \Omega, \boldsymbol{\alpha})$  with  $\boldsymbol{\omega} = (\omega_1, \dots, \omega_p)^t$  and  $\Omega = \boldsymbol{\omega}^t \tilde{\Omega} \boldsymbol{\omega}$ .

We also consider contaminated training data, in which 5% of the observations come from a normal distribution. With  $\mathbf{1}_p = (1, 1, \dots, 1)^t \in \mathbb{R}^p$ , we can describe the different simulation settings as follows:

- a. (a)  $p = 2, X_1 \sim N_2(\mathbf{0}_2, I_2), X_2 \sim \text{SN}_2(-2.1_2, I_2, 5.1_2)$   
 (b) 5% observations from the second group replaced with observations from  $N_2(-3.1_2, 0.2I_2)$
- b. (a)  $p = 3, X_1 \sim N_3(\mathbf{0}_3, I_3), X_2 \sim \text{SN}_3(-2.1_3, I_3, 5.1_3)$   
 (b) 5% observations from the first group replaced with observations from  $N_3(-3.1_3, 0.2I_3)$ .
- c. (a)  $p = 5, X_1 \sim N_5(\mathbf{0}_5, I_5), X_2 \sim \text{SN}_5(-1.5.1_5, I_5, 5.1_5)$   
 (b) 5% observations from the first group replaced with observations from  $N_5(-3.1_5, 0.2I_5)$

In the 'equal group size' setting, we use  $n_1 = n_2 = 500$ . We also perform the simulation with unequal group sizes, by taking  $n_1 = 100$  and  $n_2 = 500$ .

From each population we randomly generate  $n_j$  training observations and 100 test data. On the training data we apply the three different classifiers as defined in Section 2. The results of the simulations are summarized in terms of average misclassification errors. The misclassification error is defined as the overall proportion of wrongly classified observations in the test sets. The results listed in Table 1 and Table 2 are average misclassification errors with their respective standard errors over 100 simulations.

	Rule 1	Rule 2	Rule 3
2D, No Cont.	0.0737 (0.0018)	0.0751 (0.0019)	0.0758 (0.0019)
2D, 5% Cont.	0.0744 (0.0021)	0.0751 (0.0021)	0.0756 (0.0021)
3D, No Cont.	0.0440 (0.0015)	0.0449 (0.0016)	0.0451 (0.0016)
3D, 5% Cont.	0.0425 (0.0015)	0.0437 (0.0015)	0.0425 (0.0015)
5D, No Cont.	0.0737 (0.0015)	0.0749 (0.0017)	0.0758 (0.0018)
5D, 5% Cont.	0.0736 (0.0016)	0.0735 (0.0016)	0.0767 (0.0019)

**Table 1.** Simulation results for equal group sizes.

	Rule 1	Rule 2	Rule 3
2D, No Cont.	0.1047 (0.0033)	0.0882 (0.0026)	0.0876 (0.0026)
2D, 5% Cont.	0.0991 (0.0032)	0.0797 (0.0024)	0.0818 (0.0023)
3D, No Cont.	0.0986 (0.0032)	0.0527 (0.0015)	0.0534 (0.0015)
3D, 5% Cont.	0.0965 (0.0032)	0.0533 (0.0018)	0.0499 (0.0017)
5D, No Cont.	0.2298 (0.0042)	0.0930 (0.0026)	0.0909 (0.0028)
5D, 5% Cont.	0.2284 (0.0041)	0.0956 (0.0023)	0.0916 (0.0028)

**Table 2.** Simulation results for unequal group sizes.

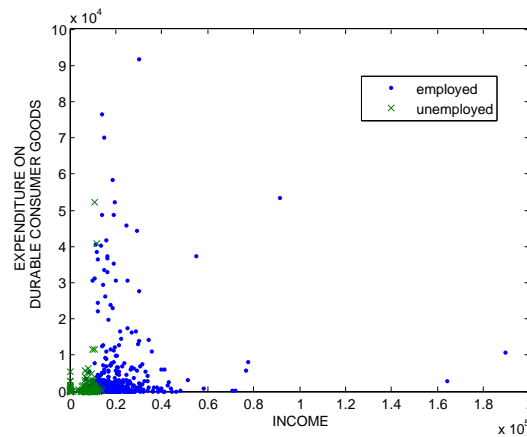
In case that the two groups are of equal size, we see that the three methods have a very comparable performance. Adding 5% outliers does not influence the results significantly. However, when the group sizes are unequal, the new rules 2 and 3 clearly outperform the first rule. This is due to the fact that the distribution of the outlyingnesses is different in both groups. The second and third rule adjust for this difference. The differences between the new rules are not significant (following *t*-test).

## 4 Example

The data used in this example come from the Belgian Household Survey of 2005. The Household Survey is a multi-purpose continuous survey carried out by the Social Survey Division of the Institute for National Statistics which collects information on people living in private households in Belgium.

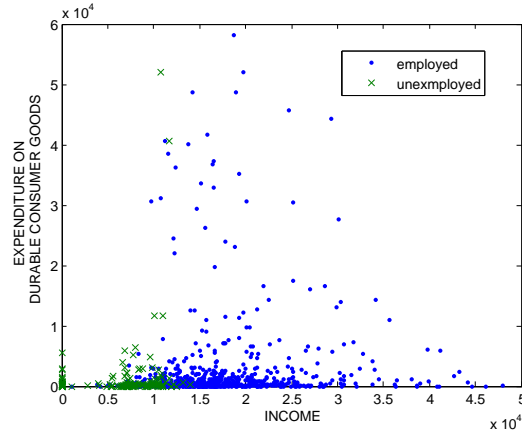
The main aim of the survey is to collect data on a range of topics such as education, welfare, family structure and health. We selected a subset of two variables from the data set: 'Income' and 'Expenditure on durable consumer goods'. In order to avoid correcting factors for family size, only single persons are considered. This group of single persons consists of 174 unemployed and 706 (at least partially) employed persons. Figure 1 is a scatterplot of the data for both groups. As the group of employed people is highly spread out, we have also plotted the lower left part of the data in Figure 2, in which both groups can be better distinguished. We notice the skewness in both classes, as well as some overlap between the groups.

Both groups are split into a training and a test set which contains 10 data points. This is done 100 times in a random way. Rule 1 results in an average misclassification error of 0.2580 with a standard error of 0.0099. Due to the fact that the group sizes are quite different, rules 2 and 3 clearly outperform this result. Rule 2 gives an average misclassification of 0.1655 (s.e. 0.0082) and rule 3 an average classification error of 0.1855 (s.e. 0.0086). This is in line with the simulation results.



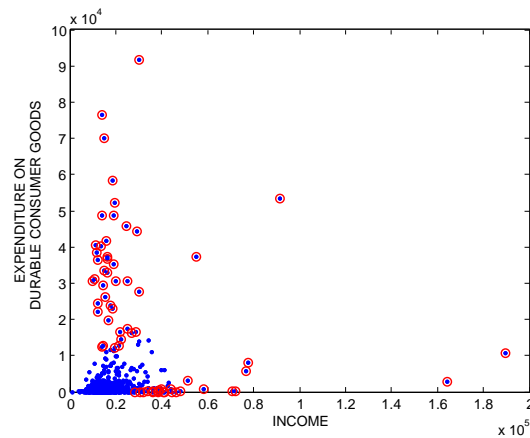
**Fig. 1.** Scatterplot of expenditure on durable consumer goods versus income (complete data set).

For illustrative purposes, we also show in Figures 3 and 4 the outliers in each group, as those observations flagged by their large AO. For the employed group (Figure 3) we find 67 outliers, whereas in the unemployed group (Figure 4) only the two most extreme observations are marked as outliers. For comparison, we also computed the Stahel-Donoho outlyingness of all observations in both groups. Then 179 of the employed and 40 of unemployed persons are flagged as outliers. This is because the skewness is not taken into account and the method searches for the most central elliptical part of the



**Fig. 2.** Scatterplot of expenditure on durable consumer goods versus income (reduced data set).

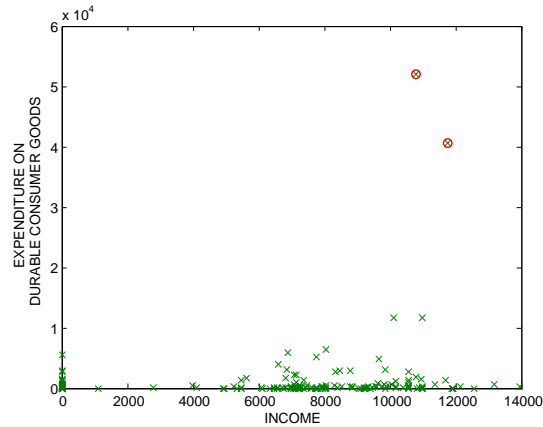
data. For skewed data, it is clearly more appropriate to use skewness-adjusted methods.



**Fig. 3.** Employed persons with outliers marked.

## References

- AZZALINI, A. and DALLA VALLE, A. (1996): The multivariate skew-normal distribution. *Biometrika* 83, 715–726.
- BILLOR, N., ABEBE, A., TURKMEN, A. and NUDURUPATI, S.V. (2008): Classification based on depth transvariations. *Journal of Classification* 25, 249–260.



**Fig. 4.** Unemployed persons with outliers marked.

- BRYN, G., HUBERT, M. and STRUYF, A. (2004): A robust measure of skewness. *Journal of Computational and Graphical Statistics* 13, 996–1017.
- BRYN, G., HUBERT, M., and ROUSSEEUW, P.J. (2005): A robustification of Independent Component Analysis. *Journal of Chemometrics* 19, 364–375.
- GHOSH, A.K., and CHAUDHURI, P. (2005): On maximum depth and related classifiers. *Scandinavian Journal of Statistics* 32, 327–350.
- HUBERT, M., and VAN DER VEEKEN, S. (2008): Outlier detection for skewed data. *Journal of Chemometrics* 22, 235–246.
- HUBERT, M., and VAN DER VEEKEN, S. (2010): Robust classification for skewed data. *Advances in Data Analysis and Classification*, in press.

# Modelling the Andalusian Population by Means of a non-Homogeneous Stochastic Gompertz Process

Maria Dolores Huete Morales<sup>1</sup> and Francisco Abad Montes<sup>2</sup>

<sup>1</sup> Department of Statistics & O.R. University of Granada  
C/Fuente Nueva, s/n, Faculty of Sciences, 18071-Granada, Spain  
*mdhuete@ugr.es*

<sup>2</sup> Department of Statistics & O.R. University of Granada  
C/Fuente Nueva, s/n, Faculty of Sciences, 18071-Granada, Spain  
*fabad@ugr.es*

**Abstract.** In this study, we examine the stochastic Gompertz non-homogeneous diffusion process, analysing its transition probability density function and conducting inferences on the process parameters using discrete sampling. All of the results are applied to the population of Andalusia (Spain), disaggregating the data by sex for the period 1981 to 2002, taking as exogenous factors only variables that are purely demographic, i.e. life expectancy at birth, foreign immigration to Andalusia and total fertility rate.

**Keywords:** Gompertz diffusion process, exogenous factors, demography, population

## 1 INTRODUCTION

It is a well known fact that growth studies are primarily concerned with human populations, although in many scientific fields (including biology and economics) growth models are currently being used to reflect the behaviour of diverse phenomena. These deterministic models are difficult to apply in real populations, since the size of a human population depends intrinsically on a large number of socio-economic variables, including changes in fertility patterns, improvements in living conditions, individual health factors which produce an increase or decrease in the number of years lived, the state of economic well-being, and changes in migratory flow patterns. As a result, it is necessary to use other frameworks in order to make provisions for population adjustment, including diffusion processes, which are widely used in growth models, Suddhendu (1988). The inclusion of exogenous factors in such models presents evident advantages, since this allows us to consider variables which influence population growth, which in turn produces considerable improvement in the modelling of phenomena. The Gompertz process has sometimes been used for modelling and forecasting economic information, but to the

best of our knowledge it has never been applied to the case of populational data. This process constitutes an innovative way of establishing or allowing for population growth, as the deterministic growth models normally used are highly dependent on population growth rates.

The Gompertz process was introduced by Ricciardi(1977), who considered applications in the field of biology, and by Crow and Shimizu (1998). In this respect, Gutiérrez et al. (2005) made inferences and examined discrete trajectories. In Ferrante (2000), continuous trajectories of the process were considered and applied to tumour growth. The non-homogenous case in the Gompertz process, involving the use of exogenous variables has been defined by Nafidi (1997) in a general context and was applied more recently by Gutiérrez in dealing with the problem of inference and considering discrete trajectories in the Gompertz process.

In this paper, we examine the non-homogenous univariate Gompertz process, which includes a series of exogenous variables within the trends. Initially, the likelihood function is obtained, which highlights the need to know the implicit expression of the functions related to the exogenous factors in order to be able to make inferences about the parameters. For this reason, a specific case is considered, thus facilitating estimates on these parameters. Finally, an exhaustive study is conducted into the application of previous theoretical results.

The Andalusian population is taken as an endogenous variable, while the exogenous variables are the number of foreign immigrants, life expectancy at birth and the synthetic indicator of fertility, taken in all cases for men and women separately (1981-2002). It was decided to perform this disaggregation because the behaviour of the exogenous variables with respect to men is very different from that corresponding to women. Male life expectancy is considerably less, and the migratory phenomenon also varies between the sexes. Moreover, it is standard practice in demographic analysis to distinguish between the sexes, and finally, the consideration of a single population value in this respect provides an inferior statistical fit.

## 2 THE NON-HOMOGENEOUS GOMPERTZ DIFFUSION PROCESS

### 2.1 Characterisation

Let  $\{X(t), t_0 \leq t \leq T\}$  be a one-dimensional diffusion process,  $\mathbb{R}$ - valued and with transition distribution function:

$$P(y, t|x, s) = P(X(t) = y|X(s) = x)$$

If we consider as infinitesimal moments (drift and diffusion coefficients) of the process respectively:



$$a(t, x) = g(t)x - h(t)x \log(x) \quad , \quad b(t, x) = \sigma^2 x^2$$

with  $h$  and  $g$  as two continuous and parametric functions, which, in other words, may depend on a certain number of parameters and  $\sigma > 0$ , we have the unidimensional Gompertz diffusion process with exogenous factors, with the following diffusion equations:

$$\begin{aligned} \frac{\partial p}{\partial t} &= -\frac{\partial}{\partial y} ((g(t)y - h(t)y \log(y))p) + \frac{1}{2} \frac{\partial^2}{\partial y^2} (\sigma^2 y^2 p) \\ \frac{\partial p}{\partial s} &= -((g(s)x - h(s)x \log(x))) \frac{\partial p}{\partial x} - \frac{1}{2} (\sigma^2 x^2) \frac{\partial^2 p}{\partial x^2} \end{aligned}$$

where  $p$  is the density of transition function. The transition distribution result:

$$\begin{aligned} P(y, t|x, s) &= \left( 2\pi\sigma^2 e^{-2 \int_s^t h(z) dz} \int_s^t e^{2 \int_s^\theta h(z) dz} d\theta \right)^{-\frac{1}{2}} y^{-1} \\ &\exp \left( -\frac{1}{2} \frac{\left[ \log(y) - \log(x) e^{-\int_s^t h(z) dz} - e^{-\int_s^t h(z) dz} \int_s^t k(\theta) e^{-\int_\theta^t h(z) dz} d\theta \right]^2}{\sigma^2 e^{-2 \int_s^t h(z) dz} \int_s^t e^{2 \int_s^\theta h(z) dz} d\theta} \right) \end{aligned}$$

With the above, we can deduce that the  $r$ -order conditional moments of the endogenous variable:

$$\begin{aligned} E(X^r(t)/X(s) = x) &= \exp \left\{ r \log(x) e^{-\int_s^t h(z) dz} + r e^{-\int_s^t h(z) dz} \int_s^t k(\theta) e^{-\int_\theta^t h(z) dz} d\theta + \right. \\ &\quad \left. + \frac{r^2 \sigma^2}{2} e^{-2 \int_s^t h(z) dz} \int_s^t e^{2 \int_s^\theta h(z) dz} d\theta \right\} \end{aligned}$$

and by taking  $r = 1$ , we immediately obtain the first-order conditional moment (conditioned trend function, CTF):

$$\begin{aligned} E(X(t)/X(s) = x) &= \exp \left\{ \log(x) e^{-\int_s^t h(z) dz} + e^{-\int_s^t h(z) dz} \int_s^t k(\theta) e^{-\int_\theta^t h(z) dz} d\theta + \right. \\ &\quad \left. + \frac{\sigma^2}{2} e^{-2 \int_s^t h(z) dz} \int_s^t e^{2 \int_s^\theta h(z) dz} d\theta \right\} \end{aligned}$$

## 2.2 Inference in the model

In order to find estimators of the process parameters, we use the maximum likelihood method and we consider discrete sampling, in other words, a realization of the same in the instants  $(t_0, t_1, \dots, t_n)$ ,  $X(t_0) = x_0$ ;  $X(t_1) =$

$x_1; \dots; X(t_n) = x_n$  with the initial condition  $P(X(t_0) = x_0) = 1$ . If we indicate:

$$m_\alpha = \log(x_{\alpha-1})e^{-\int_{t_{\alpha-1}}^{t_\alpha} h(z)dz} - e^{-\int_{t_{\alpha-1}}^{t_\alpha} h(z)dz} \int_{t_{\alpha-1}}^{t_\alpha} k(\theta)e^{-\int_{t_{\alpha-1}}^{\theta} h(z)dz} d\theta$$

the associated log-likelihood function will be:

$$\begin{aligned} \log(\mathbb{L})(x_0, x_1, \dots, x_n) &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \\ &- \frac{1}{2} \sum_{\alpha=1}^n \log \left( e^{-2 \int_{t_{\alpha-1}}^{t_\alpha} h(z)dz} \int_{t_{\alpha-1}}^{t_\alpha} e^{2 \int_{t_{\alpha-1}}^{\theta} h(z)dz} d\theta \right) - \sum_{\alpha=1}^n \log(x_\alpha) \\ &- \frac{1}{2\sigma^2} \sum_{\alpha=1}^n \frac{[\log(x_\alpha) - m_\alpha]^2}{e^{-2 \int_{t_{\alpha-1}}^{t_\alpha} h(z)dz} \int_{t_{\alpha-1}}^{t_\alpha} e^{2 \int_{t_{\alpha-1}}^{\theta} h(z)dz} d\theta} \end{aligned}$$

To be able to minimise this function with regard to the unknown parameters, we need to know the form of the functions  $h$  and  $g$ . However, this is not always possible, as it requires us to identify the function that best fits the exogenous factors introduced into the process and which, in turn, possess an integer that can be resolved. In this study, we took functions  $h$  and  $g$ , which enable us to work analytically with the associated likelihood function. It is assumed that the functions  $h$  and  $g$  are  $h(t) = \beta$  and  $g(t) = \alpha_0 + \sum_{i=1}^q \alpha_i g_i(t)$ , where the exogenous variables  $g_i(t)$  are continuous functions in  $[t_0, T]$ . Thus, the stochastic differential equation which characterizes the process is:

$$dx(t) = \{g(t)x(t) - \beta x(t) \log x(t)\} dt + \sigma x(t) dw(t) \quad (1)$$

On differentiating the log-likelihood in relation to  $\mathbf{a}$ ,  $\sigma^2$  y  $\beta$ , the following equations appear:

$$\begin{aligned} \mathbf{U}_\beta \mathbf{v}_\beta &= \mathbf{U}_\beta \mathbf{U}'_\beta \mathbf{a} \\ n\sigma^2 &= (\mathbf{v}_\beta - \mathbf{U}'_\beta \mathbf{a})' (\mathbf{v}_\beta - \mathbf{U}'_\beta \mathbf{a}) \\ \left( \nu_\beta^{-1} e^{-\beta} \mathbf{l}'_x - \mathbf{a}' \frac{\partial \mathbf{U}_\beta}{\partial \beta} \right) (\mathbf{v}_\beta - \mathbf{U}'_\beta \mathbf{a}) &= 0 \end{aligned}$$

where  $\mathbf{l}'_x = (\log(x_1), \dots, \log(x_n))'$  and  $\frac{\partial \mathbf{U}_\beta}{\partial \beta}$  the matrix formed by those derived from the elements of  $\mathbf{U}_\beta$  in relation to  $\beta$ . In this way, we obtain the maximum likelihood estimators of  $\mathbf{a}$  and  $\sigma^2$ :

$$\hat{\mathbf{a}} = (\mathbf{U}_\beta \mathbf{U}'_\beta)^{-1} \mathbf{U}_\beta \mathbf{v}_\beta \quad (2)$$

$$n\hat{\sigma}^2 = \mathbf{v}'_{\hat{\beta}} \mathbf{H}_{\mathbf{U}, \hat{\beta}} \mathbf{v}_{\hat{\beta}} \quad (3)$$

with  $\mathbf{H}_{\mathbf{U},\hat{\beta}} = \mathbf{I}_n - \mathbf{U}'_{\hat{\beta}}(\mathbf{U}_{\hat{\beta}}\mathbf{U}'_{\hat{\beta}})^{-1}\mathbf{U}_{\hat{\beta}}$  idempotent symmetric matrix. The estimator of  $\beta$  is obtained by substituting (2) and (3) in the third likelihood equation, leaving the following expression:

$$\left( \nu_{\beta}^{-1} e^{-\beta} \mathbf{I}'_x - \mathbf{v}_{\hat{\beta}} \mathbf{U}'_{\beta} (\mathbf{U}_{\hat{\beta}} \mathbf{U}'_{\hat{\beta}})^{-1} \frac{\partial \mathbf{U}_{\beta}}{\partial \beta} \right) \mathbf{H}_{\mathbf{U},\beta} \mathbf{v}_{\beta} = 0 \quad (4)$$

Given that the functions  $g_j(t)$  appear in (6), it is not possible to have an explicit estimator expression of  $\beta$ , since such functions may only be known as a result of discrete observations of the exogenous variables  $y_{i,j}$ ;  $i = 1, \dots, n$ ;  $j = 1, \dots, q$ . For this reason, exogenous factors are normally constructed from observed values of the variables through polygonal functions:

$$g_j(t) = y_{i-1,j} + (y_{i,j} - y_{i-1,j})(t - t_{i-1}) \quad (5)$$

thus, if we indicate  $z_{ij}(\beta) = y_{i-1,j} + (y_{i,j} - y_{i-1,j}) \frac{\beta - 1 + e^{-\beta}}{\beta(1 - e^{-\beta})}$ , we can state:

$$\int_{t_{i-1}}^{t_i} g_j(\xi) e^{-\beta(t_i - \xi)} d\xi = \gamma_{\beta} z_{ij}(\beta)$$

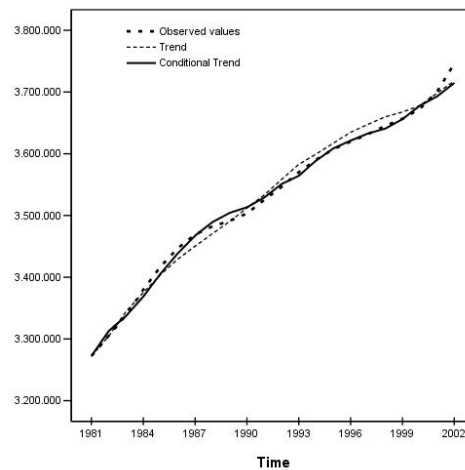
### 3 APPLICATION TO THE ANDALUSIAN POPULATION

The collected variables (or exogenous factors) used to fit the non-homogeneous model to the population of Andalusia, disaggregated by sex for the period 1981-2002, consist of foreign immigrants, life expectancy at birth ( $e_0$ ) or mean number of remaining years of life of new-born children, and a synthetic indicator of fertility (the total fertility rate ( $TFR$ ), or number of children per mother at a fertile age). The male and female population values were obtained from the Andalusian Institute of Statistics (Instituto de Estadística de Andalucía, IEA). The information on the number of foreign immigrants is provided by the National Institute of Statistics (Instituto Nacional de Estadística, INE) through the Residential Variations Statistics (note that until 1983, disaggregation by sex was an INE estimate, because the sex of immigrants was not previously recorded). The values for life expectancy at birth as used in this study were obtained by constructing biannual mortality tables, based on information provided by the IEA. In the following applications, the TFR indicator for the female and male populations was used jointly as an exogenous variable. It is true that the male TFR may be calculated, but this statistic is not frequently employed and there is no significant variation between the values for men and for women.

The female population of Andalusia was used as an endogenous variable and the number of foreign female immigrants, female life expectancy at birth and total fertility rate have been employed as exogenous variables. As has

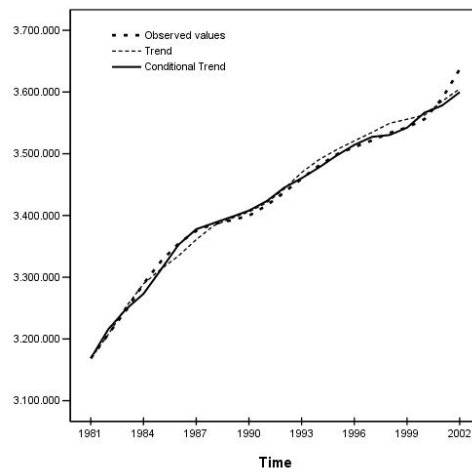
been previously mentioned, the period of observation is from 1981 to 2002, although the final two years have been reserved in order to carry out predictions and to test the goodness of the model. As mentioned above, the observation period was 1981 to 2002, although the final two years were reserved in order for forecasting and to test the goodness of the model. The exogenous factors were constructed from the above-mentioned observations, taking into account polygonal functions of the type (5). The  $\beta$  parameter was estimated by means of numerical procedures on (6) using "Mathematica" software; this value was used with expressions (2) and (3) to estimate the remaining parameters. These estimates are shown in Table 1. These values were used to derive the trend function and the conditional trend of the process, shown in Figure 1.

In relation to the male population of Andalusia, we took the number of foreign male immigrants, male life expectancy at birth and the total fertility rate as exogenous variables. The estimated parameters for the model are shown in Table 1 and the estimated trend functions are shown in Figure 6.



**Fig. 1.** Estimated trend functions (women)

The estimates obtained by means of discrete and continuous sampling in the homogeneous Gompertz process were compared, and the continuous form was found to be the most appropriate method. A further comparison was made of the lognormal and Gompertz processes, taking the total population as the analytical variable, while the exogenous factors assumed were foreign immigrants and the total fertility rate (the estimated trend functions in Figure 3). This figure shows that when the total population is used as the endogenous variable (with no disaggregation by sex), the estimations made are notably less accurate. In any case, the errors produced during the obser-

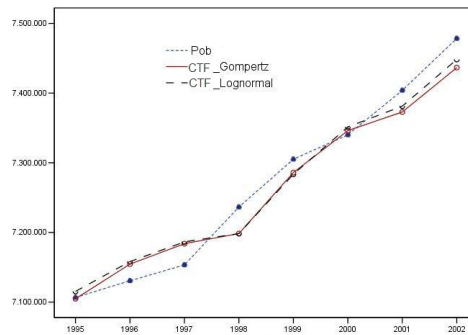


**Fig. 2.** Estimated trend functions (men)

Parameter	Estimated value (women)	Estimated value (men)
$\beta$	0.0378191	0.0797127
$\alpha_0 - \sigma^2/2$	0.0507695	0.1031180
$\alpha_1$	-0.0005712	-0.0029721
$\alpha_2$	1.2287000	1.0394100
$\alpha_3$	0.0066227	0.0271248
$\sigma^2$	$2.91204 \times 10^{-6}$	$3.32719 \times 10^{-6}$

**Table 1.** Estimated parameters

vation period are significantly fewer with the Gompertz process. This finding confirms the validity of the non-homogeneous Gompertz process as a tool for populational analysis.



**Fig. 3.** CTF: Gompertz and Lognormal

## 4 Conclusions

The conditioned trend function represents the population observations better than does the non-conditioned trend function (NCTF). However, the latter, the use of the NCTF, while it does not register small fluctuations, is more advantageous in terms of making future population predictions because as the NCTF is obtained for a single instant, it is not necessary to know the observed value of the previous instant. In order to do this, we need only establish a series of hypotheses for the values of the exogenous variables. In fact, in any population projection it is necessary to have previously established suppositions on the behaviour of the basic demographic indicators (life expectancy, TFR, migrations). Accordingly, a wide range of possibilities is opened up, since the behaviour of the population at each age (or age interval) can be studied, and predictions made. From the absolute value of the conditioned trend function and the observed population, the errors committed in each year of observation were calculated (%) as a coefficient of the difference between the observed and the estimated population values. The differences between the observed and estimated values over a small number of years amounted to 0.5% of the observed population (only in the year 2002, which is reserved for forecasts, does it exceed this value). This is an indicator that the non-homogeneous Gompertz model can acceptably represent population behaviour.

## References

- Crow, E.L. and Shimizu, K. (1988): *Lognormal distribution theory and application*. Ed. Marcel Dekker.
- Ferrante, L. and Bompadre, S. and Possati, L. and Leone, L. (2000): Parameter estimation in a gompertzian stochastics model for tumor growth. *Biometrics* 56, 10761081.
- Gutiérrez, R. and Gutiérrez-Sánchez, R. and Nafidi, A. and Román, P. and Torres, F. (2005): Inference in gompertz-type nonhomogeneous stochastic systems by means of discrete sampling. *Cybernetics and Systems* 36, 203-216.
- Huete, M.D.(2006): *El modelo estocástico de Gompertz. Modelización de datos sociodemográficos*. PhD. thesis, University of Granada.
- Nafidi A. (1997): *Difusiones Lognormales con factores exógenos. Extensiones a partir proceso de difusión de Gompertz*. PhD thesis, University of Granada.
- Ricciardi, L.M. (1977): Diffusion processes and related topics in biology. *Lect. Notes Biomath*, 14. Springer Verlag.
- Suddendun, B.(1988): *Stochastic Processes in Demography and Applications*. Ed. Wiley Eastern Limited, New Delhi.

# Neural Network Approach for Histopathological Diagnosis of Breast Diseases with Images

Yuichi Ishibashi<sup>1</sup>, Atsuko Hara<sup>2</sup>, Isao Okayasu<sup>2</sup>, and Koji Kurihara<sup>1</sup>

<sup>1</sup> Graduate School of Environmental Science, Okayama University, Okayama  
700-8530, Japan, *ishibashi@ems.okayama-u.ac.jp*

<sup>2</sup> Kitasato University School of Medicine, Sagamihara, Kanagawa, 228-8555,  
Japan

**Abstract.** Diagnosis of breast diseases relies on recognizing diseased tissue in histopathological images. The tissues studied will contain both diseased and normal areas. To insure a correct diagnosis a method is described here that is made up of three steps. The 1st step is to subdivide the histopathological image into sections. These subdivisions will then all be digitized (step 2). Several methods were tested and Wavelet transformation was found to be the best. The final step was evaluation by neural network analysis. The collective evaluation of subdivisions will increase the accuracy of diagnosis and help to avoid missing cancerous or inflamed tissue. In some studies (ref) malignancy of cancer was measured by support vector machine etc. in histopathology, but identification of the kind of cancer by pattern recognition is new. Our study attempts to digitize the features of tissue pattern for each kind of disease and to recognize the kind of disease by neural network.

**Keywords:** neural network, wavelet transformation, learning vector quantization, histopathological diagnosis, breast disease

## 1 Introduction

Doctors rely on accurate diagnosis of tissue by pathologists to help them decide on the right treatment for patients for example in the early stages of cancer. The tissue will be obtained from organs and will be observed and analyzed under the microscope. Patterns and defined features will be identified from amongst a large amount of information present in the specimen. This process of analysis by human eyes is subjective and poorly reproducible (Kumagai 2006). Findings in a single case may have different diagnosis and it may be difficult to make the differentiation between the benign and malignant state.

Diagnosis was attempted based on images with computers but it did not attain any practical use. In 1999 Mukai developed a prototype that digitizes the characteristics of cells and tissues with image analysis and identified whether it is benign or malignant. But the applied region is limited, such as intraductal proliferative diseases. When the system was applied to other

kinds of disease the concordance rate between the system and pathologists was about 90 percent. Therefore it is difficult to be applicable in practical use from the reliability point of view. Nippon Electrical Company (NEC) in 2009 developed a system that automatically extracts a cancerous region from histopathological images and features of cells with a high degree of accuracy. An extraction of cancerous candidate region is performed at low resolution followed by an evaluation of the feature values of cell nucleus at high resolution in the system. However there is no function that can differentiate which kind of cancer it is.

Histopathological diagnosis determines whether the lesion is a tumor or not, and whether the tumor is benign or malignant. If it is a tumor then it is important to decide what kind of treatment would be required if it is malignant, therefore diagnosing the kind of tumor is important. This study attempts to differentiate not only tumors but also inflammations and borderline lesions.

## 2 Image feature extraction and digitization

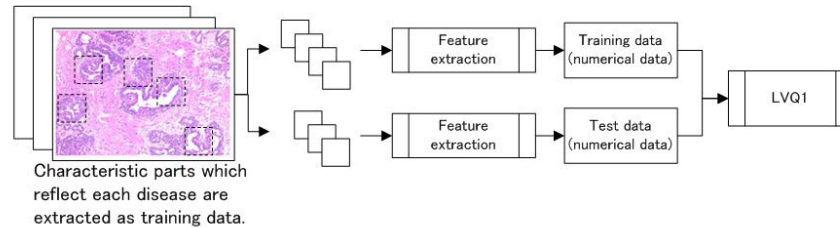
An optical microscope is regularly used in histopathological tests. A histopathological specimen is mounted on a glass slide and is normally observed at different resolutions, such as 4X, 10X and 40X. The first step of observation is recognition of the whole structural pattern at the lower resolutions, such as 4X or 10X. It enables probable histological diagnosis after confirmation of region, classification of epithelial/nonepithelial and classification of benign/malignant. The next step is a detailed deliberation on the cellular pattern which composes tissues, such as nuclear atypia, at higher resolution (40X) and final diagnosis with the consideration of both structural and cellular patterns. There are more than 50 kinds of breast diseases containing inflammations and borderline lesions other than tumor. In this study 9 kinds of breast diseases were recognized.

Texture analysis is one of the methods in image analysis. In this analysis the problem is how to numerically characterize the specific variation pattern of image element values in the picture image region (Tamura, 2002). We digitized the texture information of histopathological images in order to examine the structural patterns of specimens. There are many kinds of methods for calculating texture information. The following methods of analysis were tried in this study: mesh feature, co-occurrence matrix feature, Fourier feature and Wavelet feature.

Normal or interstitium tissue other than the tissue which reflects a specific disease is contained in a specimen on the glass slide. Therefore some small parts that are 128X128 pixels and contain a characteristic pattern were extracted from the original 1000X700 pixel image at low resolution. Color images were converted into 256 gray scale images and then feature values



were calculated (Figure 1). Learning vector quantization (LVQ) which is one of the neural network methods was adopted as image pattern recognition.



**Fig. 1.** Values of image feature extraction and image pattern recognition by LVQ.

In mesh analysis the numbers of pixels which are lower than a threshold of brightness in 64 16X16 pixel squares were counted. Two different thresholds were set in order to differentiate the deep and light gray patterns. In co-occurrence matrix analysis the energy, moment, entropy and correlation were calculated as feature values. In Fourier analysis the texture feature values were calculated from the frequency components after Fourier transformation of the images (Tamura, 2002). Wavelet transformation allows for lower resolution after dividing the patterns of sound or image into high and low frequency components. Multi-resolution analysis which repeats this method emphasizes the characteristics of pattern and distinguishes specific signals (Arai, 2000).

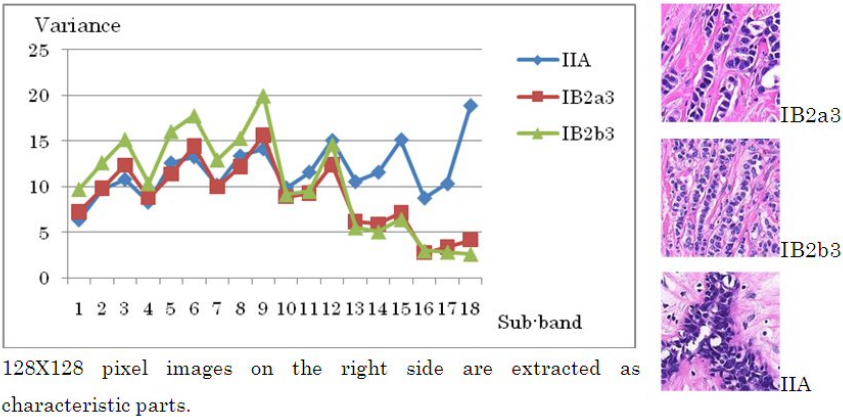
Principal component analysis was used for examining whether transformed numerical values extract the image information efficiently. Four kinds of feature values were analyzed using principal component analysis. Table 1 shows the error rate of neural network and the top 10 eigen values and cumulative contribution ratios as the results of principal component analysis. Image pattern recognition was performed by learning vector quantization (LVQ) method which is one of the methods of neural network. The error rate is high in the cases that 1st eigen value is extremely larger than other values or up to 4th or 5th principal components do not consolidate more than 80 percent of the information. In Wavelet feature the consolidation of information with small number of principal components is relatively large and the rate of recognition is the best, therefore Wavelet feature was chosen as the most appropriate method.

Wavelet transform for two-dimensional signals, such as images, is done as follows. First, in the horizontal direction one-dimensional Wavelet transform for each row divides the image into high and low frequency components. Then, for each column this converted signal is performed by one-dimensional trans-

	Error rate		1	2	3	4	5	6	7	8	9	10
Mesh	0.519	Eigen value	86.79	27.28	21.15	16.66	13.26	10.19	9.34	8.54	8.31	7.26
		Cumulative	0.19	0.25	0.3	0.34	0.37	0.39	0.41	0.43	0.45	0.46
co-occurrence matrix	0.719	Eigen value	15.87	1.89	1.34	0.43	0.27	0.08	0.04	0.04	0.02	0.01
		Cumulative	0.79	0.89	0.96	0.98	0.99	0.99	1	1	1	1
Fourier	0.5	Eigen value	37.9	9.63	4.77	3.5	2.46	1.38	1.02	0.62	0.52	0.39
		Cumulative	0.59	0.74	0.82	0.87	0.91	0.93	0.95	0.96	0.97	0.97
Wavelet	0.186	Eigen value	6.58	3.78	1.91	1.67	0.83	0.59	0.53	0.47	0.38	0.33
		Cumulative	0.37	0.58	0.68	0.77	0.82	0.85	0.88	0.91	0.93	0.95

**Table 1.** Recognition error rate of Neural network and the result of principal component analysis

formation in the vertical direction. One two-dimensional wavelet transform in horizontal and vertical directions divides the original signal into four components, such as LL, LH, HL and HH sub-bands. Two-dimensional Wavelet transformation is adapted to LL component recursively (Sakai, 2006). The values of the graph in Figure 2 are the variances in each sub-band. Figure 2 shows the results of Wavelet transformation and characteristic images of 3 types of cancer. Restibrachium is found in IB2a3(Scirrhou carcinoma)and IB2b3(Invasive lobular carcinoma) and the forms of changes in the graph are similar. But IIA(Fibroadenoma) is different from the others in the graph and image. As described above Wavelet feature reflects texture information, therefore classification and recognition using Wavelet feature is appropriate.



**Fig. 2.** Comparison of feature vectors by Wavelet transformation with images.

### 3 Pattern recognition using neural network

Accuracy of hierarchical neural network is worse in case of many variables and LVQ is more flexible and has an advantage in case of many variables (Jin, 2007). Therefore LVQ was adopted for the recognition of images. LVQ is a competitive learning type neural network model, and an input space is divided by finite number of codebook vectors in the model. The method of vector quantization divides a multi dimensional space into small regions which gather the points near the codebook vector (Kanamori, 2009). LVQ is an inclusive term of an algorithmic set which involves LVQ1, LVQ2, LVQ3 and so on. LVQ1 made the best result of recognition in this study. The algorithm of LVQ1 is as follows (Kohonen, 1995),

Let input data be  $x \in R^p$ , and label be  $y \in \{1, 2, \dots, G\}$ .  $n$  sets of input data and label  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  are given as training data. Assuming that  $k$  sets of codebook vector and label are expressed as (1).

$$\{(m_i, l_i), i = 1, \dots, k\} \quad (1)$$

LVQ divides an input space using a finite number of labeled codebook vectors and differentiates. In sequential type one data is selected at time  $t$  and the codebook vector is updated. In LVQ1 the codebook vector and the label of (1) are updated by expression (2).

$$m_c(t+1) = \begin{cases} m_c(t) + \alpha(t)(x(t) - m_c(t)), & y(t) = l_c(t) \\ m_c(t) - \alpha(t)(x(t) - m_c(t)), & y(t) \neq l_c(t) \end{cases} \quad (2)$$

In (2),  $c$  is the code of codebook vector which is nearest to the input data. Usually  $\alpha(t)$  is set smaller as time advances (Kohonen, 1995).

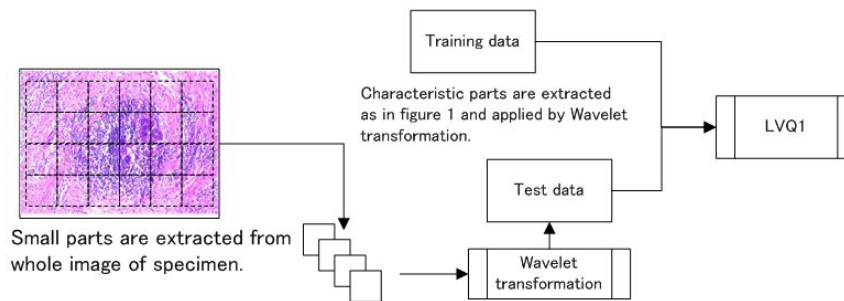
Low resolution (10X) images of specimen were used in order to differentiate the kind of breast disease based on the differences of structural pattern. Some characteristic parts of tumor or inflammation were extracted from an image of a specimen that was already diagnosed. There were 211 small images extracted from 9 kinds of diseases. Each disease contains 3 to 5 different cases. 211 images were divided into 141 training data and 70 test data. Table 2 denotes that misclassified rate of recognition is relatively low.

Figure 3 shows the method of diagnosis for a new case. Training data are transformed values by Wavelet transformation from 128X128 pixel areas of characteristic part of images which are already diagnosed as in Figure 1. Test data are transformed values by Wavelet transformation from the 128X128 pixel areas which are all over the image of a new case. Table 3 shows the results of recognition by LVQ1 for 9 different new cases. 2 different tumors, IB1b and IB2a3, are recognized for IB2a3. This method proposes 1 or more than 1 diseases after recognition, but the final decision should be made by a person e.g. a clinician or pathologist.

Training data are extracted from characteristic parts of each disease, but a specimen contains not only characteristic parts but also non characteristic parts, such as interstitium tissue etc. Neural network tries to recognize

Classification		IB1a	IB1b	IB2a1	IB2a2	IB2a3	IB2b3	IIA	IX	VIIA	Error*
IB1a	Noninvasive ductal carcinoma	8	0	1	0	0	0	0	0	1	0.200
IB1b	Lobular carcinoma in situ	0	10	0	0	0	0	0	0	0	0.000
IB2a1	Papillotubular carcinoma	0	1	3	0	1	0	0	1	1	0.571
IB2a2	Solid-tubular carcinoma	0	0	0	5	0	0	0	0	0	0.000
IB2a3	Scirrhus carcinoma	0	1	0	1	9	0	0	0	0	0.182
IB2b3	Invasive lobular carcinoma	0	0	0	0	0	5	0	0	0	0.000
IIA	Fibroadenoma	0	0	0	0	0	0	3	1	0	0.250
IX	Normal	0	1	0	0	0	0	1	5	0	0.286
VIIA	Atypical ductal hyperplasia	0	0	0	0	0	0	0	2	9	0.182
Total											0.186

Error\* means Error rate.

**Table 2.** Recognition results by LVQ**Fig. 3.** Wavelet transformation for a whole case image and the method of recognition by LVQ.

Classification		IB1a	IB1b	IB2a1	IB2a2	IB2a3	IB2b3	IIA	IX	VIIA	Error*
IB2a1	Papillotubular carcinoma	0	0	138	0	0	0	2	0	0	0.014
IB2a2	Solid-tubular carcinoma	5	4	4	69	39	4	0	1	0	0.452
IB2a3	Scirrhus carcinoma	0	61	0	0	61	0	0	1	2	0.512
IB1a	Noninvasive ductal carcinoma	55	0	0	0	15	0	0	28	29	0.567
IB1b	Lobular carcinoma in situ	0	122	0	0	3	0	0	0	2	0.039
IB2b3	Invasive lobular carcinoma	6	4	0	3	10	102	0	0	1	0.190
VIIA	Atypical ductal hyperplasia	0	0	0	0	0	0	0	0	127	0.000
IIA	Fibroadenoma	2	1	0	0	50	0	14	30	30	0.890
IX	Normal	1	44	0	3	5	0	0	36	37	0.714

**Table 3.** Recognition results by LVQ for a whole case image.

Classification		IB1a	IB1b	IB2a1	IB2a2	IB2a3	IB2b3	IIA	IIA_N	IX	IX_N	VIIA	Error*
IX	Normal	1	3	0	17	2	0	0	4	32	52	15	0.333
IIA	Fibroadenoma	5	0	0	1	5	0	18	51	30	0	17	0.457

**Table 4.** Recognition results of improved method by LVQ for a whole case image.

non-characteristic parts as some sort of disease. To avoid this situation we attempted to adopt non-characteristic part as training data (Table 4). "N" means "non-characteristic part". Table 5 shows that the cases of IX and IIA became better in recognition.

Classification	Error rate	
	Only characteristic parts	Including non-characteristic parts
IX Normal	0.714	0.333
IIA Fibroadenoma	0.890	0.457

**Table 5.** Comparison between before and after improvement

## 4 Conclusion

Learning vector quantization method with Wavelet transformation of different diseases as training data enables the diagnosis of breast disease. There are more than 50 types of breast disease and some types contain different patterns of lesion, such as atypical ductal hyperplasia. Many more kinds of image data should be accumulated in order to diagnose these diseases. We continue to accumulate the image data and to improve the method.

Referencing with similar cases previously studied is necessary especially for pathologists who are beginners at diagnosis. In this case, a database in which similar previously diagnosed images can be retrieved by using transformed numerical values and neural network technology, would be extremely valuable.

There are 4 methods of digitization including Wavelet transformation in this study, and then multiple view learning is applicable in recognition. In a multiple learning method there are different types of classifiers. Unlabeled data are predicted by using these classifiers and the results of majority voting for labeling are added. Table 1 shows that there are large differences in classification error rates between Wavelet transformation and the other methods, therefore majority decision by 4 methods may cause the reduction of recognition rate. In histopathology there are pathological diagnosis reports after diagnosing and we adopted text mining method for this text data. In pattern recognition both image and text data are used at training phase, and at prediction phase the accuracy of classification is improved if some key words are added by the user. We adapted statistical text mining to the reports in order to differentiate the kind of breast disease, hence key words were extracted for differentiation (Ishibashi, 2010). The next step of study is to combine image recognition and text mining.

## References

- ARAI, K. (2000): *Fundamental Theory on Wavelet Analysis*. Morikita Shuppan (in Japanese),70-71.
- ISHIBASHI, Y. et al. (2010): Statistical analysis of histopathological diagnosis reports with text mining, *Joint meeting of Japan-Korea Special Conference of Statistics*,227-230.
- JIN, M. (2007): *Data Science with R*. Morikita Shuppan (in Japanese),247-255.
- KANAMORI, T. et al. (2009): *Pattern Recognition*. Kyoritsu Shuppan (in Japanese),100-106.
- KOHONEN, T. et al. (1995): LVQ PAK:The Learning Vector Quantization Program Package Technical Report. *Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, FINLAND*,5-10.
- KUMAGAI, J. and ITO, T. (2006): Histopathological diagnosis by image processing. *Pathology and Clinic*, 24(4),387-391.
- MUKAI, K. (1999): Computer Application in Pathology -Possibilities and Problems- *Utilization of computer and internet in the pathological field, Saitama, Japan*.
- NEC (2009): Pathological image diagnosis support system, <http://www.nec.co.jp/solution/bio/products/>.
- SAKAI, K. (2006): *Introduction to the Image Processing and Pattern Recognition*. Morikita Shuppan (in Japanese),145-165.
- TAMURA, H. (2002): *Computer Image Processing*. Ohmsha (in Japanese),214-224.

# Detection of Spatial Cluster for Suicide Data using Echelon Analysis

Fumio Ishioka<sup>1</sup>, Makoto Tomita<sup>2</sup>, and Toshiharu Fujita<sup>3</sup>

<sup>1</sup> School of Law, Okayama University, Okayama 700-8530, Japan,  
*fishioka@law.okayama-u.ac.jp*

<sup>2</sup> Clinical Research Center, Faculty of Medicine, Tokyo Medical and Dental  
University, Tokyo 113-8519, Japan, *tomita.crc@tmd.ac.jp*

<sup>3</sup> The Institute of Statistical Mathematics, Research Organization of Information  
and Systems, Tokyo 106-8569, Japan, *fujita-t@ism.ac.jp*

**Abstract.** Recently, the number of suicides in Japan increases rapidly. For this problem, it is clear that a statistical implication is important. Our data consists of male suicide data from Kanto area in central part of Japan during 1973-2007. In this paper, we investigate the transition and the tendency of male suicides by detecting geographical spatial cluster. It is performed by echelon scan which we have proposed as one of the cluster detection. Furthermore, the performance of the cluster detection based on echelon analysis is compared to the cluster based on previous study.

**Keywords:** spatial data, spatial scan statistic, geographical clusters, echelon analysis

## 1 Introduction

The number of suicides in Japan is around 25,000 per year until 1997 had remained, in 1998 it was suddenly more than three million people (Fujita *et al.*, 2003), which has remained at that level until now. For the number of suicides in Japan by the vital statistics of the Ministry of Health, Labour and Welfare, 30,827 people in 2007 is number two after in 2003, which is a major social problem (Fujita *et al.*, 2003). It is a very high level look at the world by World Health Organization (WHO), in high suicide rate according to the countries, Japan 23.7 is number eight, furthermore it has been reported with the highest in Japan, among seven major countries (France 17.6, Germany 13.0, Canada 11.3, USA 11.0, Italy 7.1, UK 6.7) (Cabinet Office, 2008). For this serious problem, it is clear that a statistical implication is important.

Fujita (2009) has updated the Ministry of Health, Labour and Welfare demographic survey of death “local statistics about suicide” in 2009, and used secondary medical care zone in 2008, he organizes the situation goes back to 1973. In addition, on the secondary medical care zone, it is summarized the national ratio of mortality of suicide in different age groups. (age 10 or older.) We focus six time periods which are during 1973-1982, 1983-1987, 1988-1992,

1993-1997, 1998-2002 and 2003-2007 years in secondary medical care zone. So, there are six time periods and 350 areas (secondary medical area) for the space-time analyses. In these massive and large quantities of data, we use male suicides in each time period. As an analysis area, we use 70 regions at Kanto area in central part of Japan. In this paper, we investigate the transition and the tendency of suicides among men in 1973-2007 by detecting geographical spatial cluster. Additionally, we compare two spatial clusters obtained by a previous study and our proposing method.

## 2 Spatial cluster for the suicide data

### 2.1 Background

The importance of statistical analyses for spatial data has grown in various scientific fields. A statistical technique for the spatial data have been established by the radical advances of the computing power, for example GIS (geographical information system). Especially, the study for disease clustering in spatial epidemiology is of interest to statisticians. Regarding the detection of cluster areas, several methods have been proposed. Recently, the cluster detection by scan statistics has been a popular method. The scan statistics is a statistical method to detect clusters by scanning within the whole study area and testing whether such an excess have occurred by chance or not. Now, a spatial scan statistic (Kulldorff, 1997) is a very popular and useful method. The spatial scan statistic is a method of detection and inference for areas of markedly high or low rates based on the likelihood ratio. Kulldorff detected a most likely cluster, significant cluster (zone) for cellular data, based on spatial scan statistic with binomial and Poisson models. He proposed using a circular window to detect areas with high log-likelihood ratio. In this section, we apply the Kulldorff's spatial scan statistic to the male suicide data by using SaTScan program.

### 2.2 Spatial scan statistic

Suppose a geographical cluster candidate area  $Z$  are within whole area  $G$ . Each individual within the area  $Z$  has population probability  $p_1$  of the attribute, while the population probability for individual outside of the area  $Z$  is  $p_2$ . The probability for any individual is independent respectively. The null hypothesis is  $H_0 : p_1 = p_2 = p$ . The alternative hypothesis to detect high rate is  $H_1 : p_1 > p_2$ . Let  $n(G)$  be the total population in the whole area  $G$ , and  $n(Z)$  be the population within the area  $Z$ . The  $c(G)$  is the total number of the attribute in the whole area  $G$  and  $c(Z)$  is the number of the attribute within the area  $Z$ . Here, we consider the model based on the Poisson distribution. We can hence write the likelihood function as

$$L(Z, p_1, p_2) = \frac{\exp[-p_1 n(Z) - p_2 (n(G) - n(Z))]}{c(G)!} p_1^{c(Z)} p_2^{c(G) - c(Z)} \prod_{x_i} n(x_i)$$



In order to maximize the likelihood function, we calculate the maximize likelihood function conditioned the area  $Z$ . The maximum likelihood estimator  $\hat{p}_1 = c(Z)/n(Z)$  and  $\hat{p}_2 = (c(G) - c(Z))/(n(G) - n(Z))$  are substituted.

$$L(Z) = \frac{\exp[-c(G)]}{c(G)!} \left(\frac{c(Z)}{n(Z)}\right)^{c(Z)} \left(\frac{c(G) - c(Z)}{n(G) - n(Z)}\right)^{c(G) - c(Z)} \prod_{x_i} n(x_i)$$

The likelihood ratio  $\lambda$  is maximized over all subset area of whole area to detect a cluster.

$$\lambda = \frac{\max_z L(Z)}{L_0} = \frac{\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)} \left(\frac{c(G) - c(Z)}{n(G) - n(Z)}\right)^{c(G) - c(Z)}}{\left(\frac{c(G)}{n(G)}\right)^{c(G)}}$$

Here,  $L_0$  is the following likelihood function under the null hypothesis.

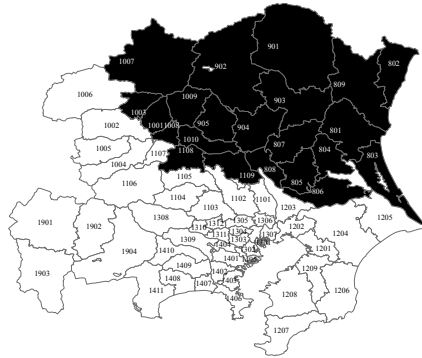
$$L_0 \stackrel{\text{def}}{=} \sup_p \frac{\exp[-pn(G)]}{c(G)!} p^{c(G)} \prod_{x_i} n(x_i) = \frac{\exp[-c(G)]}{c(G)!} \left(\frac{c(G)}{n(G)}\right)^{c(G)} \prod_{x_i} n(x_i).$$

The region  $Z$  that attains the maximum likelihood ratio is regarded as the most likely cluster.

### 2.3 Application to suicide data

As a study region, we use 70 secondary medical care zones at Kanto area in central part of Japan. The data comprise the total number of population and the number of suicides for each region. Here, we use the male data at 6th time period (2003-2007). The total numbers of suicides are 32,572 out of 95,850,805 population. We apply the Kulldorff's circular scan method by using SaTScan program. We detect a most likely cluster which based on 999 Monte Carlo replications for  $p$ -value estimation. Here, we set maximum cluster size to 50% of the population. The result is shown as Figure 1. The number in Figure 1 indicates the region number. A most likely cluster includes 22 regions,  $\hat{Z} = \{901, 903, 902, 809, 802, 904, 801, 807, 905, 1009, 1007, 804, 1010, 808, 805, 1109, 1001, 1008, 1003, 1108, 806, 803\}$  with log likelihood ratio = 87.28 and  $p$ -value = 0.001. We can see that there is the most likely cluster in northeast regions where a little outside from big cities such as Tokyo. The results of every time period are presented in Table 1.

It is useful to find circular shape's cluster because Kulldorff's scan method uses a circular window. However, it is difficult to detect clusters such as following the shape of a river or a road. To overcome this problem, several non-circular scan techniques were proposed (Patil and Taillie, 2004; Duczmal and Assunção, 2004; Tango and Takahashi, 2005). In addition to these methods, we have proposed a non-circular shape's cluster detection, using echelon analysis. So, we apply our method to these suicide data and compare it to Kulldorff's circular scan.



**Fig. 1.** Most likely cluster of male suicide in the secondary medical care zone of Kanto area at 6th time period, using circular scan.

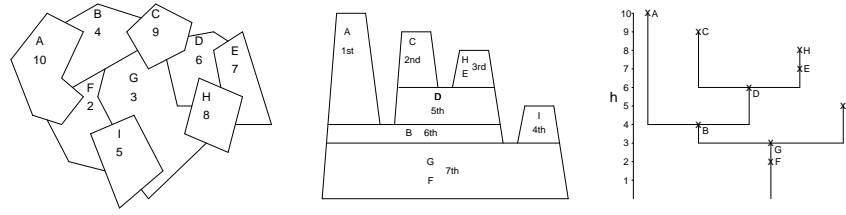
	No. of regions	No. of cases	No. of expected	Incidence rate	Log likelihood ratio	<i>p</i> -value
1st. (1973-1982)	21	5507	4459.52	1.23	134.70	< 0.001
2nd. (1983-1987)	22	3884	3081.51	1.26	114.25	< 0.001
3rd. (1988-1992)	22	3183	2589.65	1.23	74.87	< 0.001
4th. (1993-1997)	23	3822	3298.84	1.16	47.06	< 0.001
5th. (1998-2002)	22	5149	4593.74	1.12	37.95	< 0.001
6th. (2003-2007)	22	6531	5612.04	1.16	87.28	< 0.001

**Table 1.** Most likely clusters of male suicide for every time period, using circular scan.

### 3 Echelon approach for the suicide data

#### 3.1 Echelon analysis

Echelon analysis (Myers *et al.*, 1997) is a useful technique for investigating the phase-structure of spatial lattice data systematically and objectively. The echelons are derived from changes in topological connectivity. A regional data has areal referenced values  $h_i$  within spatial region  $D_i, i = 1, 2, \dots, n$ . Then, the data are expressed as the forms of  $(i, h)$ . As an example, Figure 2 (left) shows nine regions named from A to I and their values  $h$ . This spatial data is divided to the same structured area like Figure 2 (center). These parts are called echelons. The 1st, 2nd, 3rd and 4th echelons are peak, 5th echelon is foundation of peaks, and 6th and 7th echelons are foundation of peak and foundation. Each region is included in each echelon. For example, the 1st peak consists of region {A} and the 3rd peak consists of regions {H,E}. This spatial structure is given by the echelon dendrogram shown in Figure 2 (right). Some extended approaches using echelon analysis have been proposed for health and environmental data (Ishioka *et al.* (2007), Kurihara *et al.* (2000), Tomita *et al.* (2008)).



**Fig. 2.** A regional data (left), division to the same echelon (center) and Echelon dendrogram (right).

### 3.2 Bayesian estimates

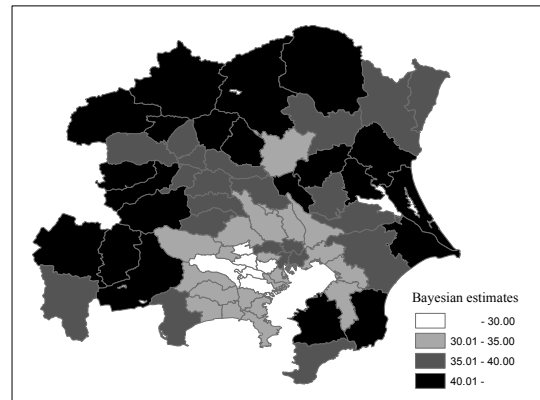
In this study, we use a Bayesian estimates as  $h$  for echelon analysis. When a group observed small population size, mortality rates vary greatly with a slight decrease in the number of deaths from suicide. In other words, the numbers become unstable because the effect of chance variation, small population size of population for suicide be used to calculate the comparison is often not suitable. Above mortality data, therefore, following age-adjusted death rate applied empirical Bayes estimates (Bayesian estimates) are used. (Fujita *et al.*, 2003)

$$\begin{aligned} & \text{Age - adjusted death rate (Bayesian estimation)} \\ &= \sum \left( \frac{\text{No. of death by age class for observation} + \hat{\beta}_i}{\text{Population by age class for observation} + \hat{\alpha}_i} \right. \\ & \quad \times \left. \frac{\text{Population by age class for base population}}{\text{Population for base population}} \right) \end{aligned}$$

where,  $\hat{\alpha}_i$  and  $\hat{\beta}_i$  are the prior distribution of the suicide situation in the country ( $\Gamma$  distribution selection), and they are gotten by the first and second moments as the weight per the size of population of the secondary health care areas. By Bayesian estimation, we can diminish the effect of chance fluctuation caused by the population size.

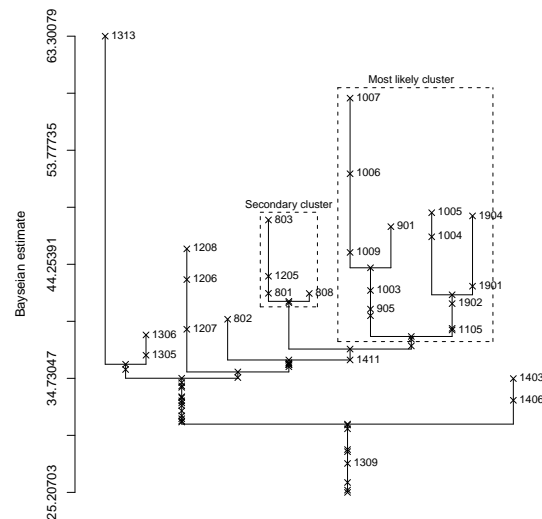
### 3.3 Application to suicide data

A spatial structure of male suicide based on Bayesian estimates at 6th time period and their bordering (connection) is given by an echelon dendrogram of Figure 4. We find most likely cluster by scanning from the regions included in upper echelon to the regions included in bottom echelon. Under this procedure, the most likely cluster is regarded as regions which take the maximum likelihood ratio. Generically, the larger cluster size is, the higher its maximum likelihood ratio is. Therefore, we detect the number of cluster regions as under



**Fig. 3.** Choropleth map of Bayesian estimation for male suicide at 6th time period.

23 regions because we compare to the result of circular scan. The most likely cluster is the regions enclosed by dot-line of the dendrogram shown as Figure 4. It includes 17 regions  $\hat{Z} = \{902, 1003, 905, 1001, 1106, 1902, 1903, 1105, 1007, 1006, 1009, 901, 1005, 1004, 1904, 1901, 1002\}$  with log likelihood ratio = 107.60 and  $p$ -value = 0.001. These regions are shown in Figure 5 (lower right). The

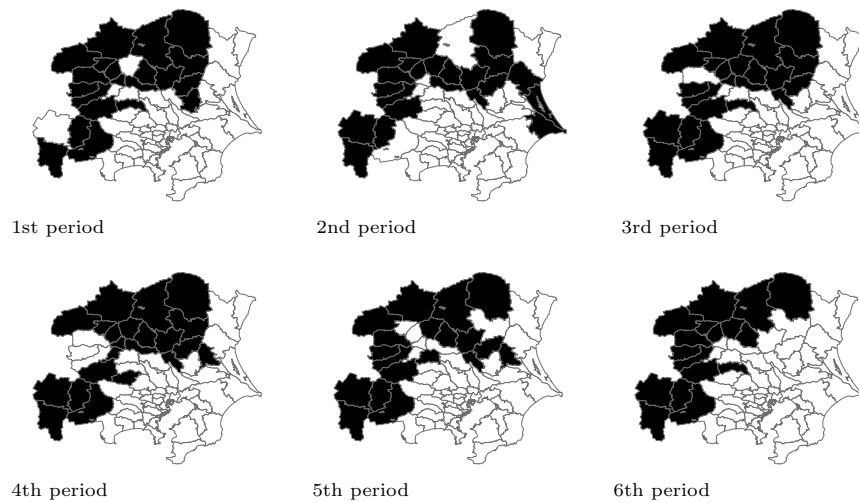


**Fig. 4.** Echelon dendrogram based on Bayesian estimates for male suicide at 6th time period.

shape of cluster does not be circularity, and its log likelihood ratio is higher

	No. of regions	No. of cases	No. of expected	Incidence rate	Log likelihood ratio	<i>p</i> -value
1st. (1973-1982)	21	5123	3081.51	1.66	151.45	< 0.001
2nd. (1983-1987)	22	4078	3138.59	1.30	152.88	< 0.001
3rd. (1988-1992)	22	3042	2323.06	1.31	118.00	< 0.001
4th. (1993-1997)	22	3576	2963.94	1.21	69.49	< 0.001
5th. (1998-2002)	19	4089	3404.84	1.20	72.80	< 0.001
6th. (2003-2007)	17	3386	2634.98	1.29	107.60	< 0.001

**Table 2.** Most likely clusters of male suicide for every time period, using echelon scan based on Bayesian estimates.



**Fig. 5.** Most likely clusters of male suicides for the every time periods, using echelon scan based on Bayesian estimates.

than the result of circular scan. The results of every time period are presented in Table 2. The echelon analysis based on Bayesian estimates provides the clusters with the high likelihood ratio than the circular scan in every period. There maps are shown in Figure 5. For the male suicide data, the most likely cluster exists northwest in all periods. However, there are little changes by periods. We can see that the most likely cluster is located on a little outside from big cities such as Tokyo, similar to the result of the circular scan.

## 4 Conclusion

In this paper, we investigated the spatial cluster of male suicide in Kanto area, by using the circular scan and the echelon scan. Additionally, we investigated the transition and the tendency for six time periods. We showed that the

most likely cluster is located on a little outside from big cities such as Tokyo. The result of echelon scan based on Bayesian estimates is shown to obtain higher likelihood clusters than the result of circular scan. The echelon scan is useful tool to detect spatial cluster because 1) it is not limited to the shape of circularly, 2) it is efficient because of scanning from regions which create the peak structure having a high value, 3) thus it helps a reduction of computation time.

## Acknowledgement

This work was supported by KAKENHI 21700305 and KAKENHI 21700317.

## References

- Cabinet Office. (2008): White Book for Strategy to Prevent Suicide. Saiki Printing Co.
- DUCZMAL, L. and ASSUNÇÃO, R.A. (2004): A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis*, 45, 269-286.
- FUJITA, T., TANIHATA, T. and MIURA Y. (2003): Geographical Features of the Increasing Number of Suicide After 1998 in Japan. *Journal of Health and Welfare Statistics*, 50(10), 27-34.
- FUJITA, T. (2009): Statistics of Community for the Death from Suicide. National Institute of Mental Health, National Center of Neurology and Psychiatry, Japan.
- ISHIOKA, F., KURIHARA, K., SUITO, H., HORIKAWA, Y. and ONO, Y. (2007): Detection of Hotspots for 3-dimensional Spatial Data and Its Application to Environmental Pollution Data. *Journal of Environmental Science for Sustainable Society*, 1, 15-24.
- KULLDORFF, M. (1997): A spatial scan statistics. *Communications in Statistics, Theory and Methods*, 26, 1481-1496.
- KULLDORFF, M. (2006): Information Management Services Inc: SaTScan v7.0: Software for the spatial and space time scan statistics, <http://www.satscan.org/>.
- MYERS, W.L., PATIL, G.P. and JOLY, K. (1997): Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics*, 4, 131-152.
- KURIHARA, K., MYERS, W.L. and PATIL, G.P. (2000): Echelon analysis of the relationship between population and land cover patterns based on remote sensing data. *Community Ecology*, 1, 103-122.
- PATIL, G.P. and TAILLIE, C. (2004): Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11, 183-197.
- TANGO, T. and TAKAHASHI, K. (2005). A flexible spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4, 11.
- TOMITA, M., HATSUMICHI, M. and KURIHARA, K. (2008). Identify LD blocks based on hierarchical spatial data. *Computational Statistics & Data Analysis*, 52(4), 1806-1820.

# Time-Varying Coefficient Model with Linear Smoothing Function for Longitudinal Data in Clinical Trial

Masanori Ito<sup>1</sup>, Toshihiro Misumi<sup>2</sup> and Hideki Hirooka<sup>3</sup>

<sup>1</sup> Astellas Pharma Inc.  
3-17-1, Hasune, Itabashi-ku, Tokyo, 174-8612, Japan  
*masanori.ito@jp.astellas.com*

<sup>2</sup> Astellas Pharma Inc.  
3-17-1, Hasune, Itabashi-ku, Tokyo, 174-8612, Japan  
*toshihiro.misumi@jp.astellas.com*

<sup>3</sup> Astellas Pharma Inc.  
3-17-1, Hasune, Itabashi-ku, Tokyo, 174-8612, Japan  
*hideki.hirooka@jp.astellas.com*

**Abstract.** In clinical trials, more than one visit for efficacy evaluation are scheduled and the analysis for longitudinal data is required. LOCF ANCOVA model is usually chosen. The LOCF approach assumes the missing data MCAR. But since the assumptions are often unrealistic and thus it is not the best choice. TVCM is applied to clinical trial data for evaluation of the drug treatment varying with time. The inference on the model is conducted by a simple linear smoothing function. The knots of the smoothing function are identified according to the scheduled patient's visits. The inference on the model can be conducted in the context of the mixed model methodology and software. From the results of the case study for sample clinical trial data, TVCM was superior to LOCF ANCOVA and MMRM approaches in terms of evaluating the treatment effect coupled with time variation in the early phase of the treatment in particular.

**Keywords:** time-varying coefficient model, linear smoothing, last observation carried forward, repeated measures mixed-effects model

## 1 Introduction

In clinical trials, the treatment period and the number of scheduled visits for efficacy evaluation are predetermined by study design. For example, a trial may be designed to treat patients for eight weeks with evaluations at baseline and at the end of each week. In this case, nine observations would be available for one patient by the end of the treatment period. In most trials, the primary end time point is taken as the last time point of the predetermined treatment period. If a patient withdrew from the trial before completion, some observations posterior to the discontinuation would be missed. Assessment of mean change from baseline to endpoint via analysis of (co)

variance using simple linear model is often conducted. Missing data is often stored into carrying the last observation forward (LOCF). However, the LOCF approach assumes that missing data are MCAR (missing completely at random) and that subject's responses are constant from the last observed value to the endpoint of the trial. Both of the assumptions are often unrealistic in clinical trials, so these conditions are seldom seen (Verbeke and Molenberghs, 2000).

Several authors including Laird and Ware (1982) proposed likelihood-based mixed-effect model to analyze incomplete data of longitudinal clinical trials. In general, when dropouts are negligible, the parameters of dropout and outcome processes are assumed to be distinct, and hence likelihood-based methods can be used on the marginal distribution of the observed data for statistical inferences. Such a mixed model is named Mixed-effect Model Repeated Measures (MMRM) analysis by Mallinckrodt *et al.* (2001). MMRM is quite flexible and powerful approach for a longitudinal data in clinical trial. Most analyses with longitudinal data are based on parametric models, such as multivariate linear regression, generalized linear regression, and non-linear regression models. References of the parametric approaches are written by Ware (1985), Diggle (1988), Davidian and Giltinan (1995). Likelihood-based estimation methods in linear models are discussed by Diggle (1988). Liang and Zeger (1986) proposed generalized linear models for longitudinal data.

While parametric approaches are useful, questions will always arise about the adequacy of the model assumptions and the potential impact of model misspecifications on the analysis (Hoover *et al.*, 1998). The most useful model for studying the association between the covariates and response for the longitudinal data in clinical trial is the Time-Varying Coefficient Model (TVCM). Hastie and Tibshirani (1993) proposed the smoothing spline method for the estimation of time-varying coefficients. Hoover *et al.* (1998) studied two types of nonparametric estimators of time-varying coefficients, smoothing spline and locally weighted polynomial. They investigated the cross validation criteria for selecting smoothing parameters. In this paper, we used TVCM with the simple linear smoothing spline function. The selection of the knots is usually troublesome. Speaking of the most of the clinical trial data, however, patients are planned to visit hospitals several times at scheduled timings and thus data are concentrated at the time of each visit. Therefore, we selected the knots for the scheduled timing of visits. We represented the linear spline as a Best Linear Unbiased Prediction (BLUP) in a mixed model to estimate the varying coefficients in the contexts of the mixed model methodology same as the MMRM approach. Through the case study of the clinical trial data, we investigated the performance of LOCF, MMRM and TVCM.



## 2 LOCF ANCOVA and MMRM approach

### 2.1 LOCF ANCOVA model

For subjects  $i = 1, \dots, I$  and repeated observations per visit  $j = 1, \dots, J_i$  (end of study visit), LOCF ANCOVA model is

$$Y_{iJ_i} = \beta_0 + \beta_1 Y_{i0} + \beta_2 x_i + \epsilon_{ij}, \quad (1)$$

where  $Y_{ij}$  is a change from baseline ( $Y_{i0}$ ) of outcome measurement at the  $j$ th time point for the  $i$ th subject,  $\beta_0$  is intercept,  $\beta_1$  is effect of baseline measurement ( $Y_{i0}$ ),  $\beta_2$  is effect size at time  $J$ ,  $x_i$  is dummy coded covariate for subject  $i$  (e.g.  $x_i = 0$  for placebo group and  $x_i = 1$  for treatment group) and  $\epsilon_{ij}$ s are assumed to be independently distributed from a univariate normal distribution. If  $Y_{iJ}$  is missing, the formula would be  $Y_{iJ} = Y_{ij}$  (where  $j = 1, \dots, J-1$ ). That is, if an endpoint measurement is missing, it would be filled in by the previously observed measurement.

### 2.2 MMRM approach

For subjects  $i = 1, \dots, I$  and repeated observations per visit  $j = 1, \dots, J_i$ , MMRM model can be described as

$$\mathbf{Y}_i = X_i \boldsymbol{\beta} + Z_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad (2)$$

where  $\mathbf{Y}_i$  is the  $J_i$  dimensional vector of outcome measurement for the  $i$ th subject,  $\boldsymbol{\beta}$  is the  $p$  dimensional vector containing the fixed effects (e. g. baseline, treatment effect and time),  $X_i$  and  $Z_i$  are the  $(J_i \times p)$  and  $(J_i \times q)$  dimensional design matrices of known covariates,  $\mathbf{b}_i$  is the  $q$  dimensional vector containing the random effects ( $\mathbf{b}_i \sim N(0, D)$ ),  $\boldsymbol{\epsilon}_i$  is the  $J_i$  dimensional vector of residual components ( $\boldsymbol{\epsilon}_i \sim N(0, \Sigma_i)$ ),  $D$  is the general  $(q \times q)$  covariance matrix with  $(i, j)$  element  $d_{ij} = d_{ji}$  and  $\Sigma_i$  is the  $(J_i \times J_i)$  covariance matrix which depends on  $i$  only through its dimension  $J_i$ . From the equation (2), it can be derived that  $\mathbf{Y}_i$  are distributed as independent normal, with mean  $X_i \boldsymbol{\beta}$ , and variance-covariance matrix  $Z_i D Z_i' + \Sigma_i$ . In the MMRM model, time  $t_{ij}$  is considered as a factor variable and treatment  $\times$  time effects is considered as an unstructured interaction effect instead of considering treatment  $\times$  time effect as a slope difference of treatment groups over the study time period. The advantage of considering the effect of treatment  $\times$  time as unstructured interaction effect is that it provides the direct estimates and statistical test of least square mean differences of the treatment groups at the study endpoint, as well as at each scheduled study time point with respect to the primary efficacy measure (Siddiqui, Hung and O' Neill, 2009).

### 2.3 LOCF vs. MMRM

Although a misspecification of wrong covariance structure in MMRM analysis inflates Type I error and alters power, a specification of unstructured covariance structure in MMRM analysis, regardless of the true variance-covariance structure, is reasonable and provides better control of Type I error rate and power than LOCF analysis (Mallinckrodt *et al.*, 2004).

In the paper of Siddiqui *et al.* (2009), a total of 48 clinical study datasets in FDA that were submitted to the division of neurological and psychiatric drug products were reanalyzed to compare the efficacy decisions at the study endpoint based on MMRM and LOCF ANCOVA endpoint analysis. As a result, the LOCF ANCOVA analysis underestimated a standard error of the treatment difference, as compared to the corresponding estimate in MMRM analysis. Hence, the MMRM approach appeared to be a superior approach in evaluating the efficacy of a study drug.

## 3 TVCM with Linear Smoothing

### 3.1 Brief Overview

A useful model for studying the association between the covariates and response for the longitudinal data in clinical trial is TVCM

$$Y_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta}(t_{ij}) + \epsilon_i(t_{ij}), \quad (3)$$

where  $\boldsymbol{\beta}(t) = (\beta_0(t), \dots, \beta_p(t))'$  are smooth functions of  $t$  and  $\epsilon_i(t)$  is zero mean stochastic process. Hastie and Tibshirani (1993) proposed a smoothing spline method for the estimation of  $\boldsymbol{\beta}(t)$ . Hoover *et al.* (1998) studied two types of nonparametric estimators of  $\boldsymbol{\beta}(t)$ , smoothing spline and locally weighted polynomial, and investigated the cross validation criteria for selecting smoothing parameters. In this paper, we focused on the analysis for the clinical trial data of chronic condition. In general, the subjects visit the hospital according to the scheduled time for a chronic disease study, therefore subject data are concentrated visit by visit. Accordingly, a simple regression structure as the linear smoothing spline function with visits as knots is enough to express the longitudinal variation of treatment effects.

### 3.2 Representation of Mixed-Effect Model

In this paper, we represented TVCM as a framework of the mixed-effects model. This representation is useful because it allows smoothing to be done using mixed model methodology and software. TVCM allows the intercept and slope coefficients to be arbitrary smooth functions of  $t_{ij}$ . The penalized linear spline version of this model is

$$Y_{ij} = \alpha_0 + \alpha_1 t_{ij} + \sum_{k=1}^K b_k^\alpha(t_{ij} - \kappa_k)_+ + (\beta_0 + \beta_1 t_{ij} + \sum_{k=1}^K b_k^\beta(t_{ij} - \kappa_k)_+) x_i + \epsilon_{ij}, \quad (4)$$

where  $\kappa_1, \dots, \kappa_K$  are knots (visits) over the range of the  $t_{ij}$  values,  $K$  is the number of the knots.  $\alpha_0$  and  $\alpha_1$  are the parameters of the intercept, and  $b_k^\alpha$  ( $k = 1, \dots, K$ ) shows the random effects of the intercept. In the same manner,  $\beta_0$  and  $\beta_1$  are the parameters of a slope coefficients, and  $b_k^\beta$  ( $k = 1, \dots, K$ ) shows a random effects of a slope coefficients.  $(t_{ij} - \kappa_k)_+$  shows a positive part of the function  $t_{ij} - \kappa_k$ , that is, it is zero for those values of  $t_{ij}$  where  $t_{ij} - \kappa_k$  is negative. A function such as  $(t_{ij} - \kappa_k)_+$  is also referred to as a truncated line. From the equation (2) and (4), the mixed-effects model representation is written as

$$\mathbf{Y}_i = X_i \boldsymbol{\beta} + Z_i \mathbf{b} + \boldsymbol{\epsilon}_i. \quad (5)$$

It is obtained by setting

$$\begin{aligned} \mathbf{X}_i &= [1 \ t_{ij} \ x_i \ t_{ij} x_i]_{1 \leq j \leq J_i}, \quad \boldsymbol{\beta} = [\alpha_0 \ \alpha_1 \ \beta_0 \ \beta_1]^T, \\ \mathbf{Z}_i &= [(t_{ij} - \kappa_k)_+ \ x_i(t_{ij} - \kappa_k)_+]_{\substack{1 \leq k \leq K \\ 1 \leq j \leq J_i}}, \end{aligned}$$

$$\mathbf{b} = [b_1^\alpha, \dots, b_K^\alpha, b_1^\beta, \dots, b_K^\beta]^T, \text{ and } \text{Cov}(\mathbf{b}) = \text{diag}\{\sigma_\alpha^2 \mathbf{1}_{K \times 1}, \sigma_\beta^2 \mathbf{1}_{K \times 1}\}.$$

#### 4 Application to Longitudinal Data: Example of CNS Drug Development

We used the sample data of the clinical trial for CNS (Central Nervous System) disease. The main objective of this randomized clinical trial was to compare the efficacy of test drug to placebo in patients suffering from CNS disease. This was a multicenter, randomized, double-blind, placebo-controlled, parallel-group study, comparing 3 doses of test drug with placebo given once daily to patients with CNS disease. Eligible patients were randomized in equal numbers into 1 of 4 treatment groups (placebo, low dose, middle dose or high dose) for 12 weeks of treatment. The primary efficacy variable was a change from baseline in the disease score at the end of treatment (12 weeks), after 1, 2, 4, 6, 8 and 10 weeks of treatment. A sample size of about 100 patients per group was chosen.

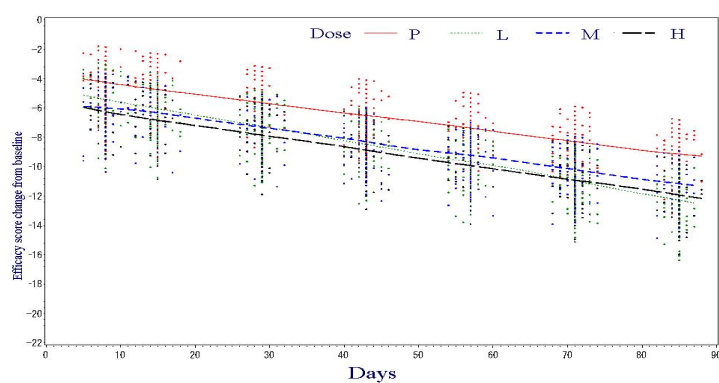
LOCF ANCOVA model, MMRM and TVCM were applied to this data to compare the results of these approaches. In the LOCF ANCOVA model (1), we set  $i = 1, \dots, 100$ ,  $j = 1, \dots, J_i$  ( $J = 7$ ),  $\beta_0$  as intercept,  $\beta_1$  as effect of baseline for the disease score,  $\beta_2$  as effect at time  $J$  ( $J = 7$ ) and  $x_i$  as dummy coded covariate for patients  $i$  ( $x_i = 0$  for placebo,  $x_i = 1$  for low dose,  $x_i = 2$  for middle dose,  $x_i = 3$  for high dose). LOCF ANCOVA model was conducted for the last evaluation data change from baseline in the disease score with dose and baseline score as exploratory variables. In the same manner, we applied MMRM (2) and TVCM (5) for this sample data. We conducted two kinds of MMRM analyses, one had only the first order variable as a time effect, and the other one had also the second order variable as time effects.

**Table 1** shows the least square means of efficacy score change from baseline difference between placebo and each dose group and P-values adjusted by Dunnett test. The superiority of high dose to placebo was confirmed by all approaches. MMRM and TVCM also showed the superiority of middle dose to placebo. As for the results of the least square means, only TVCM showed the clear monotone increase as a dose-response. **Fig. 1.** and **Fig. 2.** show the prediction values of MMRM and TVCM. For the first several weeks in the clinical trial, it seemed that the low dose was not effective. **Fig. 3.** shows the results of the estimated time-varying coefficients at each time. Clearly, the trend of the coefficients for low dose was different from other doses in early days. With regard to this case study, we concluded that TVCM is superior to LOCF ANCOVA and MMRM approaches in terms of evaluating the treatment effect coupled with time variation in the early phase of the treatment in particular.

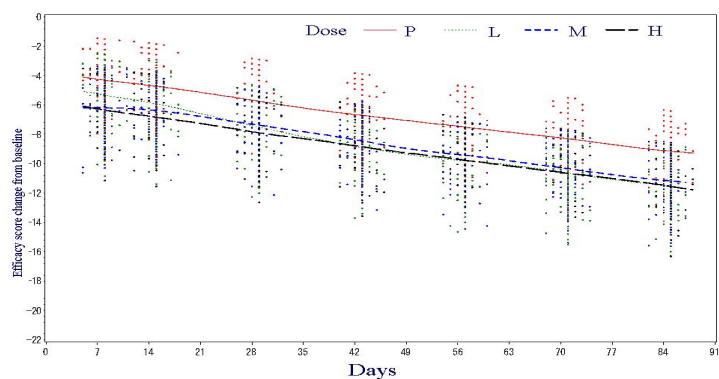
**Table 1:** Results for least square means difference between placebo and each dose group and P-values (adjusted by Dunnett test)

Approaches	Low dose	Middle dose	High dose
LOCF ANCOVA	-2.17 $P = 0.0738$	-1.42 $P = 0.3411$	-2.44 $P = 0.0413^*$
MMRM (first order time effect)	-2.04 $P = 0.0147^*$	-1.93 $P = 0.0236^*$	-2.33 $P = 0.0049^*$
MMRM (second order time effect)	-2.31 $P = 0.0161^*$	-1.92 $P = 0.0246^*$	-2.31 $P = 0.0054^*$
TVCM	-1.01 $P = 0.269$	-1.94 $P = 0.0078^*$	-2.07 $P = 0.0044^*$

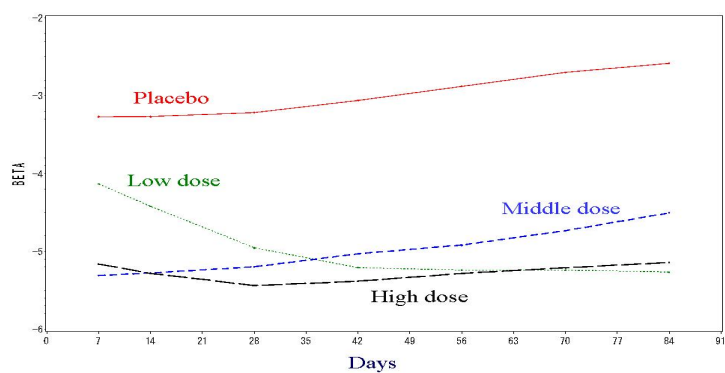
1. Dunnett test is two-sided test at significance level of 5%.
2. “\*” shows “P-value < 0.05”.



**Fig. 1.** Mean response prediction of MMRM (linear model)



**Fig. 2.** Mean response prediction of TVCM (linear smoothing function)



**Fig. 3.** Plots of the predictions for the time-varying coefficient

## References

- Davidian, M. and Giltinan, D. M. (1995): *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, London.
- Diggle, P. J. (1988): An Approach to the Analysis of Repeated Measurements. *Biometrics*, 44, 959-971.
- Hastie, T. and Tibshirani, R. (1993): Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series, B* 55(4), 757-796.
- Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998): Nonparametric Smoothing Estimates of Time-Varying Coefficient Models with Longitudinal Data. *Biometrika*, 85, 809-822.
- Laird, N. M. and Ware, J. H. (1982): Random Effects Models for Longitudinal Data. *Biometrics*, 38, 963-974.
- Liang, K. and Zeger, S. L. (1986): Longitudinal Data Analysis using Generalized Linear Models. *Biometrika*, 73, 13-22.
- Mallinckrodt, C. H., Clark, W. S. and David, S. R. (2001): Accounting for Dropout Bias using Mixed-Effects Models. *Journal of Biopharmaceutical Statistics*, 11, 9-21.
- Mallinckrodt, C. H., Kaiser, C. J., Watkin, J. G., Molenberghs, G., Carroll, R. J. (2004): The Effect of Correlation Structure on Treatment Contrasts Estimated from Incomplete Clinical Trial Data with Likelihood-based Repeated Measures Compared with Last Observation Carried Forward ANOVA. *Clinical Trials*, 1(6), 477-489.
- Ruppert, D., Wand, M. P. and Carroll, R. J. (2003): *Semiparametric Regression*. Cambridge University Press.
- Siddiqui, O., Hung, H. M. and O'Neill, R. (2009): MMRM vs. LOCF: A Comprehensive Comparison Based on Simulation Study and 25 NDA Datasets. *Journal of Biopharmaceutical Statistics*, 19, 227-246.
- Verbeke, G. and Molenberghs (2000): *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York.

# Metropolis-Hastings Algorithm for Mixture Model and its Weak Convergence

Kengo Kamatani

Graduate School of Mathematical Sciences, The University of Tokyo,  
3-8-1 Komaba, Meguro-ku, Tokyo 153-0041, Japan *kengok@ms.u-tokyo.ac.jp*

**Abstract.** This paper describes an application of the weak convergence framework of the Markov chain Monte Carlo (MCMC) method. It is well known that for the mixture model, when some of the mixture proportion parameters are 0, the Gibbs sampler behaves poorly. In this paper, we propose a simple Metropolis-Hastings (MH) algorithm and study its convergence property. In a usual Harris recurrence framework, both the MH algorithm and the Gibbs sampler are geometrically ergodic in probability 1. However, in the weak convergence framework, the former is consistent in a certain sense, but the latter is not. We present some numerical results.

**Keywords:** Gibbs sampler, finite mixture model, geometrical ergodicity, local asymptotic normality, Bernstein von-Mises's theorem

## 1 Introduction

Markov chain Monte Carlo (MCMC) method has become an essential tool in any study that has a complicated posterior calculation problem. Various new MCMC methods have been developed in the last decades. This research focuses on the efficiency of those MCMC methods. One of a useful measure of efficiency is the ergodicity of a transition kernel of a MCMC method. There are many studies related to sufficient conditions for ergodicity including Mengersen and Tweedie (1996), Roberts and Tweedie (1996) and Roberts and Polson (1994). In a recent paper Kamatani (2010) took a completely different approach to study the efficiency of MCMC methods. He considered a weak convergence of a sequence of MCMC methods and define its consistency. We apply and extend his results to a simple mixture model. We propose a Metropolis-Hastings (MH) algorithm for this problem and validate its good behavior.

First we review some results in Kamatani (2010). Let  $(X, \mathcal{X}, \mathcal{P})$  be a probability space and  $(S, d)$  be a separable and complete metric space. Let  $\mathcal{P}(\Omega, \mathcal{F}) = \mathcal{P}(\Omega)$  be a totality of probability measures on a measurable space  $(\Omega, \mathcal{F})$ . Set  $S^\infty = S \times S \times \cdots$ . We call a random variable  $M : X \rightarrow \mathcal{P}(S^\infty)$  Markov chain Monte Carlo (MCMC) if  $M(x)$  is a Markov probability measure on  $S$  for  $P$ -a.s.  $x$ . We call  $M$  strictly stationary if each  $M(x)$  is strictly

stationary as a Markov probability measure for  $P$ -a.s.  $x$ . We say that  $M$  is ergodic if  $M(x)$  is ergodic as a Markov probability measure for  $P$ -a.s.  $x$ .

*Example 3 (Gibbs sampler).* Let  $(Y, \mathcal{Y})$  be a measurable space and let  $\nu_{S|X}$  be a probability transition kernel from  $S$  to  $X$ . Let  $p$  be a  $\sigma$ -finite measure on  $S \times X \times Y$  such that

$$\begin{aligned} p(ds, dx, dy) &= p_{X,Y}(dx, dy)p_{S|X,Y}(ds|x, y) = p_{S,X}(ds, dx)p_{Y|S,X}(dy|s, x), \\ p_{S,X}(ds, dx) &= p_{S|X}(ds|x)p_X(dx) \end{aligned}$$

where for  $A, B, C = S, X, Y$ ,  $p_{C|A,B}$  and  $p_{B|A}$  are a probability transition kernel from  $A \times B$  to  $C$  and  $A$  to  $B$  with respectively. The **Gibbs sampler**  $M(x)$  (on  $(X, \mathcal{X})$  and  $(S, d)$  with initial guess  $\nu_{S|X}$  and underling measure  $p$ ) is a MCMC where  $M(x)$  is defined by the initial probability measure  $\nu_{S|X}(\cdot|x)$  and the transition kernel  $p_{S|S,X}(ds(1)|s(0), x)$  such that

$$p_{S|S,X}(A|s, x) := \int_Y p_{Y|X,S}(dy|x, s)p_{S|X,Y}(A|x, y).$$

For any  $s = (s(1), s(2), \dots) \in S^\infty$  and  $m \in \mathbf{N}$ , we define the empirical measure  $\hat{\nu}_m(s) \in \mathcal{P}(S)$  by

$$\hat{\nu}_m(s)(B) = \frac{1}{m} \sum_{i=1}^m 1_B(s(i))$$

where  $B$  is any Borel set of  $S$ . We denote the Prohorov metric of the measure on  $(S, d)$  by  $\rho_S$ . Consider a sequence of random variables  $(M_n; n \in \mathbf{N})$  where each  $M_n$  is a MCMC on a probability space  $(X_n, \mathcal{X}_n, P_n)$ . The following is a special case (strictly stationary case) of the definition of consistency in Kamatani (2010). For  $x \in \mathbf{R}$ , we write  $[x]$  for the largest integer  $a$  such that  $a \leq x$ .

**Definition 10.** Let  $(M_n; n \in \mathbf{N})$  be a sequence of strictly stationary MCMC and write  $\nu_n(x_n)$  for the invariant probability measure for  $M_n(x_n)$ . Let  $\delta_n$  be a positive sequence. We say that  $M_n$  is **weakly consistent** with rate  $\delta_n$  if for any  $\epsilon > 0$  and  $m(n) \rightarrow \infty$  and

$$A_n := \{s \in S^\infty; \rho_S(\hat{\nu}_{[m(n)\delta_n]}(s), \nu_n(x_n)) > \epsilon\},$$

it satisfies  $M_n(x_n)(A_n) \rightarrow 1$  in probability. If  $(M_n; n \in \mathbf{N})$  is weakly consistent with rate  $\delta_n \equiv 1$ , we simply call  $(M_n; n \in \mathbf{N})$  **consistent**.

We call  $(M_n; n \in \mathbf{N})$  is tight if it is tight as a sequence of random variables on  $\mathcal{P}(S^\infty)$ . We have the following result, which gives a sufficient condition for the empirical measure tending to the stationary measure.

**Theorem 1 (Kamatani, 2010).** *Let  $M_n$  be strictly stationary, tight sequence and any limit point be ergodic. Then  $M_n$  is consistent.*



Here we assumed that  $M_n$  is strictly stationary. For the Gibbs sampler, it means that the initial probability measure is the posterior distribution. Therefore the above theorem seems useless in practice. However the assumption can be weakened. We can use any initial measure which is close to the posterior distribution; details can be found in Theorem 2.9 of Kamatani (2010).

In a certain regular setting, MCMC methods such as the Gibbs sampler, the MH algorithm or the Stochastic EM algorithm all have good convergence properties. To see differences between these Monte Carlo methods, we suppose a non-regular setting. We consider a very simple, independent and identically distributed model on a state space  $(X, \mathcal{X})$  with distribution  $(p_{X|\Theta}(dx|\theta); \theta \in \Theta)$ :

$$p_{X|\Theta}(dx|\theta) = (1 - \theta)F_0(dx) + \theta F_1(dx) \quad (1)$$

where  $F_i$  is a probability measure on  $X$  for each  $i = 0, 1$  and  $F_0 \neq F_1$ . We can construct a natural Gibbs sampler for this model introducing a new variable  $y$  taking a value in  $\{0, 1\}$  which is the indicator for the mixture coefficients. We study behaviors of MCMC methods when the true parameter  $\theta_0 = 0$ .

Using the the weak convergence framework of the MCMC method, we see that MH algorithm is consistent for any  $\theta_0 \in [0, 1]$ , but the Gibbs sampler is not consistent when  $\theta_0$  is on the edge of the parameter space. We show the difference by numerical simulations. This may seem strange, because the Gibbs sampler is uniformly ergodic and the MH algorithm may not have that property. This indicates that in some cases, the “local” approach can explain the behavior of the MCMC algorithm better than the “global” (Harris recurrence) approach.

The main aim of the paper is twofold:

- to develop and apply the weak convergence framework in Kamatani (2010) for the MCMC method. In particular, we define degeneracy of MCMC. We show consistency property of the MH algorithm and degeneracy property of the Gibbs sampler.
- to find a good MCMC method for the mixture model. We propose a MH algorithm which uses an approximation of the posterior distribution. This method may be applicable to general mixture problems.

## 2 Bayesian approximation with Metropolis-Hastings algorithm

### 2.1 Gibbs sampler approach

Consider the model (1) with a prior distribution  $p_\Theta$ . A missing model structure is constructed by setting  $p_{X|Y, \Theta}(\cdot|y, \theta) = F_y$  ( $y = 0, 1$ ) and  $p_{Y|\Theta}(\{y\}|\theta) = (1 - \theta)^{1-y}\theta^y$ . We define  $p^n(d\theta, dx_n, dy_n) = p_\Theta(d\theta) \prod_{i=1}^n p_{X,Y|\Theta}(dx^i, dy^i|\theta)$

where  $p_{X,Y|\Theta}(dx, dy|\theta) = p_{X|\Theta}(dx|\theta)p_{Y|X,\Theta}(dy|x, \theta)$  and  $z_n = (z^1, \dots, z^n)$  for  $z = x, y$ . For simplicity, set a (an assigned) prior  $p_\Theta = \text{Beta}(\alpha_0, \beta_0)$ .

We write  $f_y = dF_y/d(F_0 + F_1)$  for  $y = 0, 1$  and  $g(x) = f_1(x)/f_0(x) - 1$  when  $f_0(x) > 0$  and  $g(x) = +\infty$  otherwise. According to Diebolt and Roberts (1994), the posterior distribution of  $\theta$  given  $x_n = (x^1, \dots, x^n)$  can be written as follows:

$$p_{\Theta|X_n}^n(d\theta|x_n) = \sum_{i=0}^n w_i^n(x_n) \frac{\theta^i p_\Theta(d\theta)}{\int_\Theta \vartheta^i p_\Theta(d\vartheta)}$$

where  $w_i^n(x_n)$  is a certain weight function.

The model  $(p_{X|\Theta}(\cdot|\theta); \theta \in \Theta)$  is a simple mixture model. This parametric family is quadratic mean differentiable at any  $\theta \in (0, 1)$  by Hajek (1972) since the Fisher information matrix is always positive and finite if  $F_0 \neq F_1$ .

The construction of  $p^n$  yields a natural Gibbs sampler. We consider properties of the Gibbs sampler when  $\theta_0 = 0$  (or more generally,  $n^{1/2}\theta_0 = O(1)$ ), that is, the true prior is  $\delta_{\{0\}}(d\theta)$  which is not equivalent to the assigned prior. Note that when the true prior is equivalent to  $p_\Theta$ , then the Gibbs sampler is consistent by Theorem 3.7 of Kamatani (2010). We also note that the transition kernel of the Gibbs sampler is uniformly ergodic for any  $n \in \mathbf{N}$  and  $x_n \in X_n$ . Although with this good convergence property, we will see behaviors of the Gibbs sampler may not be good for  $n^{1/2}\theta_0 = O(1)$ .

We assume

$$\alpha = F_1((\text{supp} F_0)^c) = \int g(x) F_0(dx) = 0.$$

The behavior of the Gibbs sampler is different whether  $\alpha > 0$  or  $\alpha = 0$ . In the former case, it has a Poisson-Gamma limit. When  $\alpha = 0$ , the limiting behavior is degenerate in the limit. Let

$$I := \int g(x)^2 F_0(dx).$$

We will write  $M_n$  for a scaled Gibbs sampler, where the scaling is  $\theta$  to  $n^{1/2}(\theta - \tilde{\theta}_n)$  for a Bayes estimator  $\tilde{\theta}_n$ . That is, if  $\theta(0), \theta(1), \dots$  are a path of the Gibbs sampler,  $M_n(x_n)$  is the law of  $(n^{1/2}(\theta(0) - \tilde{\theta}_n), n^{1/2}(\theta(1) - \tilde{\theta}_n), \dots)$  given  $x_n$ .

We call a Markov probability measure  $\omega \in \mathcal{P}(S^\infty)$  is degenerate if

$$\omega(\{s = (s(0), s(1), \dots); s(0) = s(1) = s(2) = \dots\}) = 1.$$

We say that a MCMC  $M$  is degenerate when each  $M(x)$  is degenerate in  $P$ -a.s.  $x$ . We say that  $M_n$  is degenerate in the limit when  $M_n$  is tight and each limit point is degenerate.

**Proposition 14.** For  $p^n \in \mathcal{P}(\Theta \times X_n \times Y_n)$  we assume the following:

- $\alpha = F_1((\text{supp} F_0)^c) = 0$  and  $I \in (0, \infty)$ .
- The prior distribution  $p_\Theta$  is  $\text{Beta}(\alpha_0, \beta_0)$ .

Then  $M_n$  is degenerate in the limit when  $n^{1/2}\theta_0 = O(1)$ . In particular,  $M_n$  is not consistent.

## 2.2 Bayesian approximation technique

We propose an independent type MH algorithm for the mixture model. For the independent type MH algorithm, the candidate probability distribution should be close to the target distribution (see ex. Mengersen and Tweedie (1996), Theorem 2.1). In our case, the target distribution is a posterior distribution. Therefore, a good approximation of a posterior distribution yields good MH algorithm. There are several papers related to a good approximation  $\bar{p}_{\Theta|X_n}^n$  such as quasi-Bayes approach or Approximate Bayes approach (see ex. Smith and Makov (1978) and Humphreys and Titterton (2000)).

Contrary to such a “direct” approximation method, we consider another “indirect” approach. We first choose a parametric family  $\{\bar{p}_{X|\Theta}(\cdot|\theta); \theta \in \Theta\}$  which is close to the original parametric family. Then set  $\bar{p}_{\Theta|X_n}^n(d\theta|x_n)$  as a posterior of the new model. The following is a procedure for choosing a parametric family  $\{\bar{p}_{X|\Theta}(\cdot|\theta); \theta \in \Theta\}$ .

- Choose a set of probability measures  $\bar{\mathcal{P}} \subset \mathcal{P}(X)$  which is wide enough.
- For each  $\theta$ , compute the Kullback-Leibler distance

$$\mathcal{K}(p_{X|\Theta}(\cdot|\theta), \mu) = \int_{\mathcal{X}} \log \frac{d\mu}{dp_{X|\Theta}(\cdot|\theta)}(x) p_{X|\Theta}(dx|\theta)$$

and find its minimizer  $\mu \in \bar{\mathcal{P}}$ .

- Set  $\mu$  as  $\bar{p}_{X|\Theta}(\cdot|\theta)$ .

*Remark 4.* We do not have to take the prior distribution  $\bar{p}_{\Theta}$  for  $\bar{p}_{X|\Theta}(\cdot|\theta)$  same as  $p_{\Theta}$ . However, we should choose  $\bar{\mathcal{P}}$  and  $\bar{p}_{\Theta}$  so as  $\bar{p}_{\Theta|X_n}^n(\cdot|x_n)$  is easy to compute.

Under mild conditions, the independent type MH algorithm using the above candidate distribution has consistency. Therefore, for the mixture model, we have the following results:

- If the true parameter  $\theta_0$  is small ( $\theta_0 = O(n^{-1/2})$ ) or close to 1, then the Gibbs sampler is degenerate but the MH algorithm has consistency.
- If the true parameter  $\theta_0$  is fixed in  $(0, 1)$ , both MCMC methods have consistency.

We consider the following example. Let  $F_i \sim N(\mu_i, \sigma_i^2)$  for  $i = 0, 1$  for  $\mu_1 \neq \mu_0$  and  $2\sigma_0^2 > \sigma_1^2$ . For a fixed  $\sigma > 0$ , take  $\bar{\mathcal{P}} = \{N(\mu, \sigma^2); \mu \in \mathbf{R}\}$ . Then  $\bar{p}_{X|\Theta}(\cdot|\theta) = N(\mu(\theta), \sigma^2)$  where  $\mu(\theta) = (1 - \theta)\mu_0 + \theta\mu_1$ . Note that  $\int x p_{X|\Theta}(dx|\theta) = \int x \bar{p}_{X|\Theta}(dx|\theta)$ . For the independent and identically distributed observation  $x_n$  of  $p_{X_n|\Theta}^n(dx_n|\theta_0)$ , the posterior  $\bar{p}_{\Theta|X_n}^n(\cdot|x_n)$  tends to  $N(\bar{\theta}_n, \sigma^2/n(\mu_1 - \mu_0)^2)$  tempered at 0 and 1 where  $\bar{\theta}_n = (\sum x_i/n - \mu_0)/(\mu_1 - \mu_0)$ . When  $\bar{p}_{\Theta}$  is the uniform prior,  $\bar{p}_{\Theta|X_n}^n(\cdot|x_n)$  is  $N(\bar{\theta}_n, \sigma^2/n(\mu_1 - \mu_0)^2)$  tempered at 0 and 1.

In this case, if  $\sigma_0 \neq \sigma_1$ ,

$$g(x) = \frac{f_1(x)}{f_0(x)} - 1 = \exp\left(-\frac{\sigma_0^2 \sigma_1^2}{2(\sigma_0^2 - \sigma_1^2)}\left(x - \frac{\mu_1 \sigma_0^2 - \mu_0 \sigma_1^2}{\sigma_0^2 - \sigma_1^2}\right)^2 + \frac{(\mu_0 - \mu_1)^2}{2(\sigma_0^2 - \sigma_1^2)}\right) - 1$$

and

$$I = \int g(x)^2 F_0(dx) = \frac{\sigma_1}{(2\sigma_0^2 - \sigma_1^2)^{1/2}} \exp\left(\frac{(\mu_1 - \mu_0)^2}{2\sigma_0^2 - \sigma_1^2}\right) - 1.$$

If  $\sigma_0 = \sigma_1$ , then

$$g(x) = \exp(\sigma^{-2}(\mu_1 - \mu_0)(x - \frac{\mu_1 + \mu_0}{2})) - 1,$$

and  $I = \exp(\sigma^{-2}(\mu_1 - \mu_0)^2) - 1$ .

The following is one possibility of adaptive choice of  $\sigma^2$ . Let  $\sigma = \sigma(\theta) = \theta(1 - \theta)(\mu_0 - \mu_1)^2 + (1 - \theta)\sigma_0^2 + \theta\sigma_1^2$ . Let  $\bar{\theta}_n$  be an initial guess which is a  $\sqrt{n}$ -consistent estimator of  $\theta_0$ . Fix  $\sigma$  to  $\sigma(\bar{\theta}_n)$  throughout the iteration. Then the posterior  $\bar{p}_{\Theta|X_n}^n(\cdot|x_n)$  tends to  $N(\bar{\theta}_n, \sigma(\theta_0)^2/n(\mu_1 - \mu_0)^2)$  tempered at 0 and 1.

*Remark 5.* The models considered here are too simple for real application. However, it is possible to implement for a mixture model with three or more components (Extension A) or each components may have own parameters (Extension B). For Extension A, we use the MH algorithm  $d - 1$  times for each generation of  $\theta$  where  $d$  is the number of components. For Extension B, we use the Metropolis-within-Gibbs algorithm. Although the theoretical validation is straightforward, it is not finished yet and simulations for those MCMC methods are not well studied. Further research should be done.

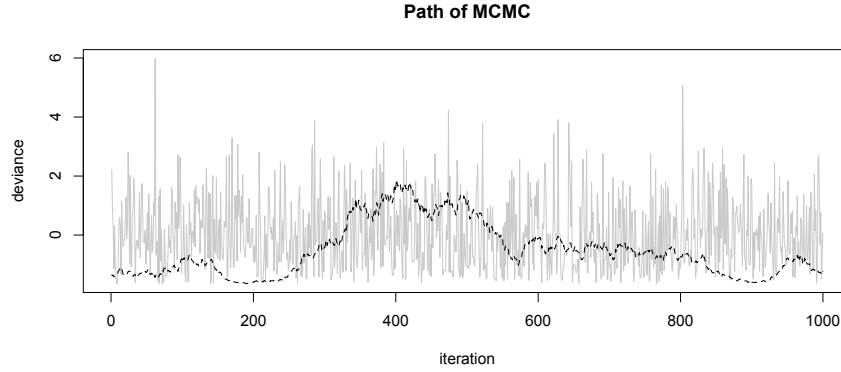
### 3 Simulation results

We compare the Gibbs sampler and the MH algorithm through a numerical simulation. Consider a normal mixture model for  $\sigma_0 = \sigma_1 = 2$ ,  $\mu_0 = 0$  and  $\mu_1 = 1$  and the true parameter is  $\theta_0 = 0$ . See the end of the section for other choice of parameters. We denote  $\theta(0), \theta(1), \dots$  for a path of a MCMC. First we see paths of

$$n^{1/2}(\theta(i) - \tilde{\theta}_n)$$

for  $i = 0, 1, 2, \dots$  of two MCMC methods for one observation  $x_n$  where  $\tilde{\theta}_n$  is the exact value of the Bayes estimator. The initial guess is the moment estimator. Even for a relatively small sample size ( $n = 50$ ), the path of the Gibbs sampler has much weaker mixing than that of the MH algorithm (the Figure is not shown here). For a large sample size ( $n = 10^4$ ), unlike the MH algorithm, the Gibbs sampler behaves like a path of a stochastic diffusion process (Figure 1). Consider the Gibbs sampler as a time series with time interval  $n^{-1/2}$ . Then it tends to a strictly stationary stochastic process

satisfying  $dX_t = S(\vartheta, X_t)dt + \sigma(X_t)dB_t$  where  $S(\vartheta, x) = x\vartheta - x^2I + \alpha_0$  and  $\sigma(x)^2 = 2x$  and  $\vartheta$  has a certain probability distribution. It implies that the Gibbs sampler is not consistent but weakly consistent with rate  $n^{1/2}$ . Therefore the Gibbs sampler has relatively slow convergence property.



**Fig. 1.** Plot of paths of MCMC methods for  $n = 10^4$ . The dashed line is a path from the Gibbs sampler and the solid line is the MH algorithm.

Figure 2 show the estimated values of the standard error of

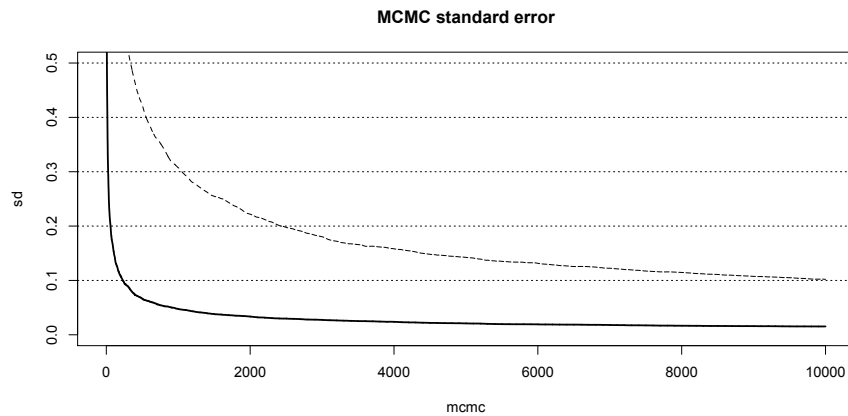
$$n^{1/2}(\tilde{\theta}_n^{(m)} - \tilde{\theta}_n) \quad (2)$$

starting from the moment estimator where  $m$  is the iteration number of MCMC methods. We denote  $\tilde{\theta}_n^{(m)} = m^{-1} \sum_{i=0}^{m-1} \theta(i)$  for an approximation of the Bayes estimator. The MH algorithm behaves better than the Gibbs sampler in terms of a smaller standard deviation.

For other choice of parameters, we have the following results:

- If  $F_0$  and  $F_1$  are similar, the MH algorithm show better performances for small  $p$  (case A) and for large  $p$  (case B).
- If  $F_0$  and  $F_1$  are very different, the MH algorithm is better for small  $p$  (case C). Two MCMC methods are similar for large  $p$  (case D).

The behaviors for cases A, C and D are theoretically validated. However, the behavior for case B is not validated. It may be possible to study by asymptotics for  $F_1 = F_\epsilon \rightarrow F_0$ . Note that for all cases, the Gibbs sampler and the MH algorithm are uniformly ergodic. Therefore, in Harris recurrence approach, to compare these MCMC methods, we need precise estimates for convergence rates. However, the estimates for the Gibbs sampler and for the MH algorithm are technically difficult in general.



**Fig. 2.** Plot of the standard deviation of MCMC methods for  $n = 10^3$  with MCMC iteration up to  $m = 10^4$ . The dashed line is the Gibbs sampler and the solid line is the MH algorithm.

## References

- DIEBOLT, J. and ROBERT, C. (1994): Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)* 56, 363–375.
- HAJEK, J. (1972): Local asymptotic minimax and admissibility in estimation. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. Vol.1. 175–194.
- HUMPHREYS, K. TITTERINGTON, D. M. (2000): Approximate Bayesian inference for simple mixtures. *COMPSTAT2000* 331–336.
- KAMATANI, K. (2010): On Some Asymptotic Properties of the Gibbs Sampler (in preparation).
- MENGERSEN, K. L. and TWEEDIE, R. L. (1996): Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics* 24, 101–121.
- ROBERTS, G. O. and POLSON, N. G. (1994): On the geometric convergence of the gibbs sampler. *Journal of the Royal Statistical Society. Series B (Methodological)* 56(2), 377–384.
- ROBERTS, G. O. and TWEEDIE, R. L. (1996): Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* 83, 95–110.
- SMITH, A. F. M. and MAKOV, U. E. (1978): A Quasi-Bayes Sequential Procedure for Mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)* 40 (1), 106–112.

# A Method for Time Series Analysis Using Probability Distribution of Local Standard Fractal Dimension

Kenichi Kamijo<sup>1</sup> and Akiko Yamanouchi<sup>2</sup>

<sup>1</sup> Graduate School of Life Sciences, Toyo University,  
1-1-1 Izumino, Itakura, Gunma, 374-0193, Japan, *kamijo@toyonet.toyo.ac.jp*

<sup>2</sup> Izu Oceanics Research Institute,  
3-12-23 Nishiochiai, Shinjuku, Tokyo, 161-0031, Japan,  
*gx0400018@toyonet.toyo.ac.jp*

**Abstract.** The moving local standard fractal dimension (LSFD) on the uniform or standard normal random process belongs to a non-symmetric normal distribution with a long tail towards the right hand side. These results can be applied to the statistical quality control, especially to the so-called control charts. Also the proposed method can be applied to the difference time series of seawater temperatures as a function of depth and the probability distribution of the moving LSFD was shown to generally conform to a power-law distribution. That is, prediction of abnormal phenomena in the global monitoring system may be possible using the moving LSFD and observing when it increases past the upper 5% significance level.

**Keywords:** local standard fractal dimension, statistical quality control, random process, difference time series of seawater temperatures

## 1 Introduction

We have already proposed the use of local fractal dimension (LFD) in discrete dynamical systems, and several examples of applications have been presented (Kamijo and Yamanouchi(2005), Kamijo and Yamanouchi(2007)). Based on our experience of having applied this index to various time series, when interpreting the index as being prognostic indicator for predicting the occurrence of a certain physical phenomenon, we consider that it is necessary to not only pay attention to the change patterns of these time series, but also the concept of absolute amount. We decided to adopt a type of standardization for LFD in recent study that we previously proposed to solve this kind of problem. This indicator is denoted by LSFD, and it can standardize the degree of complexity in time series fluctuations using the standard deviation as an absolute amount(Kamijo and Yamanouchi(2008), Kamijo and Yamanouchi(2009)). The moving LSFD on the uniform or standard normal random process belongs to a non-symmetric normal distribution with a long

tail towards the right hand side. This will open up applications in statistical quality control in the future. Also in this paper, as an example of this, we applied LSFD to the difference time series of seawater temperatures as a function of depth in the area around the Izu Peninsula, Japan.

Moreover Ayache et al.(2007) and Benassi et al.(2000) give theoretical result for the same model using their own method.

## 2 Local standard fractal dimension (LSFD)

### 2.1 Definition of local fractal dimension (LFD)

For a discrete time series, which can be considered as an objective vector,

$$\mathbf{x}_k = \{x_k, x_{k+1}, \dots, x_{k+L-1}\}, \quad (1)$$

the accumulated change  $N(r, k, L)$  can be defined as

$$N(r, k, L) = \frac{1}{r} \sum_{i=1}^r \sum_{j=0}^{\lfloor \frac{L}{r} \rfloor - 2} |x_{k+jr+r+i-1} - x_{k+jr+i-1}| \quad (2)$$

where

$L$ : length of the discrete time series,  $r$ : sampling interval.

The accumulated change  $N(r, k, L)$  can also be redefined as

$$N(r, k, L) = Ar^{-D_k} \quad (3)$$

where

$D_k$ : the  $k$ -th local fractal dimension,  $A$ : proportion constant.

Therefore,

$$\log N(r, k, L) = -D_k \log r + \log A \quad (4)$$

Here  $N(r, k, L)$  can be replaced by  $Y$ , and also  $\log r$  by  $X$ , then we have

$$Y = -D_k X + \log A \quad (5)$$

Then,  $D_k$  can be obtained as the local fractal dimension based on a regression analysis, and in this paper we will refer to  $LFD_k$  instead of  $D_k$ , as the  $k$ -th local fractal dimension. That is,  $D_k$  is the so-called regression coefficient, and can be obtained by the following equation.

$$D_k = -\frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2} \equiv LFD_k \quad (6)$$



A method for obtaining the LFD from the information on six plots has already been proposed as the Six-Point Evaluation Method (Kamijo and Yamanouchi(2005), Kamijo and Yamanouchi(2007)), where  $N(r, k, L)$  can be obtained varying  $r$  from 1 to 6 in fixed condition of  $L=30$ . In this case, we have 6 pairs for  $(X, Y)$  on the so-called  $X$ - $Y$  plane.

## 2.2 Moving measurement for LSFD

The finite time series  $\{x_k, x_{k+1}, \dots, x_{k+L-1}\}$  can be extracted from the infinite time series  $\{x_0, x_1, x_2, \dots, x_k, x_{k+1}, \dots, x_{k+L-1}, \dots\}$  for calculation with the length  $L$ , and the  $k$ -th standard deviation  $STDEV_k$  can be obtained by the following equation.

$$STDEV_k = \sqrt{\frac{\sum_{i=k}^{k+L-1} (x_i - \bar{x}_k)^2}{L-1}} \quad (7)$$

where

$$\bar{x}_k = \frac{\sum_{i=k}^{k+L-1} x_i}{L} \quad (8)$$

Next, the  $k$ -th local standard fractal dimension  $LSFD_k$  can be defined as follows:

$$LSFD_k = \frac{LFD_k}{STDEV_k} \quad (9)$$

Then we can have the discrete time series for moving  $LSFD$  by increasing the suffix  $k$  one by one.

## 3 Probability distribution of moving LSFD and LFD

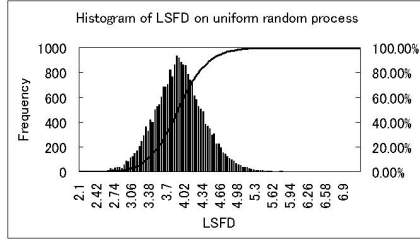
### 3.1 Application to uniform random process

In case of  $L=30$ , the histograms of moving LSFD and LFD on 30,000 uniform random numbers, which are generated by computer simulation, are shown in Fig. 1 and Fig. 2 respectively. Judging from the shape of the histogram, the probability distribution of moving LSFD can be considered as a non-symmetric normal distribution with a long tail towards the right hand side. We have also found that the LFD distribution exhibits reverse features, having a long tail towards the left hand side.

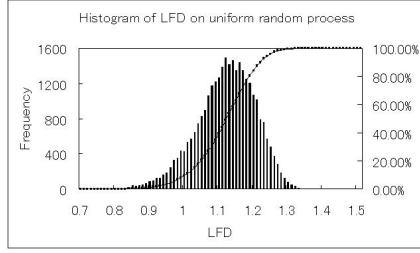
Then we obtain the following 95% confidence intervals for LSFD and LFD.

$$3.02 \leq LSFD \leq 4.84 \quad (10)$$

$$0.951 \leq LFD \leq 1.281 \quad (11)$$



**Fig. 1.** Histogram of moving LSF on 30,000 uniform random numbers.

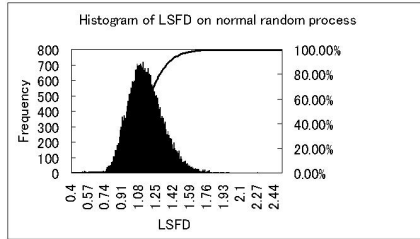


**Fig. 2.** Histogram of moving LFD on 30,000 uniform random numbers.

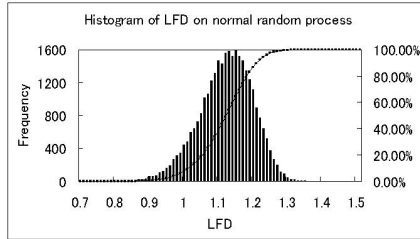
### 3.2 Application to normal random process

Similarly in case of  $L=30$ , the histograms of moving LSF and LFD on 30,000 standard normal random numbers, which are generated by computer simulation, are shown in Fig. 3 and Fig. 4 respectively. Judging from the shape of the histogram, the probability distribution of moving LSF can be considered as a non-symmetric normal distribution with a long tail towards the right hand side. We have also found that the LFD distribution exhibits reverse features, having a long tail towards the left hand side.

These characteristics are just the same as the simulation using the uniform random process with the contrast in Fig. 1 and Fig. 2.



**Fig. 3.** Histogram of moving LSF on 30,000 standard normal random numbers.



**Fig. 4.** Histogram of moving LFD on 30,000 standard normal random numbers.

Then we have the following 95% confidence intervals for LSF and LFD.

$$0.83 \leq LSF \leq 1.54 \quad (12)$$

$$0.967 \leq LFD \leq 1.267 \quad (13)$$

Accordingly the 95% confidence interval of the moving LSFD is considered to be applicable to methods such as control charts, which are often used in statistical quality control. When the observed moving LSFD falls outside this kind of confidence interval, the process in question is judged to be abnormal.

## 4 Applications in scientific fields

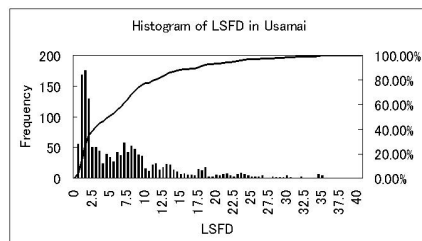
In this section, applications to difference time series of seawater temperatures as a function of depth are shown as examples for the earth environment in a kind of quality monitoring system.

### 4.1 Method for application to difference time series of seawater temperatures as a function of depth

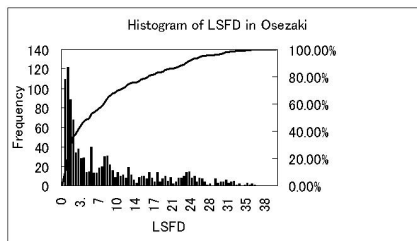
The meteorological characteristics of the near-surface layer were obtained at observation stations around the Izu Peninsula (Usami and Osezaki, Japan) by placing temperature data loggers at three depths. The observation periods and the difference time series of seawater temperatures as a function of depth at Usami and Osezaki are 2003/6/5-2007/8/24; 15m-5m and 2004/3/29-2007/5/4; 25m-5m respectively.

### 4.2 Probability distribution

This method was applied to the difference time series of seawater temperatures as a function of depth. Judging from the results of Fig. 5 and Fig. 6, the histogram of the moving LSFD, on the time series at both observation stations, was shown to generally conform to a power-law distribution respectively.



**Fig. 5.** Histogram of moving LSFD in Usami.



**Fig. 6.** Histogram of moving LSFD in Osezaki.

Then we have the following 95% confidence intervals for LSFD.

Usami;

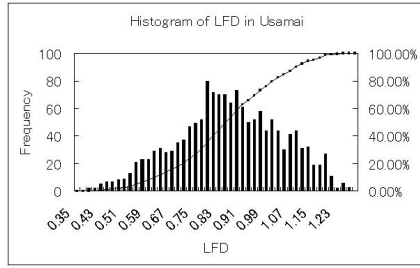
$$0 \leq LSF D \leq 22.5 \quad (14)$$

Osezaki;

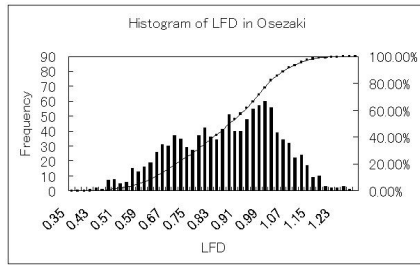
$$0 \leq LSF D \leq 26.5 \quad (15)$$

That is, the prediction of abnormal phenomena, for example, warnings of phase transition etc., in the global monitoring system may be possible in case a moving LSF D as observed statistic increases past the upper 5% significance level, similar to the case of general quality control.

For example, Fig. 7 and Fig. 8 show the histograms of moving LFD on difference time series of seawater temperatures as a function of depth at Usami and Osezaki respectively, where the length of the "processing window"  $L=30$  in these cases. Then the moving LSF D at both observation stations may actually belong to the so-called triangular distribution.



**Fig. 7.** Histogram of moving LFD in Usami.



**Fig. 8.** Histogram of moving LFD in Osezaki.

We also obtain the following 95% confidence intervals for LFD.

Usami;

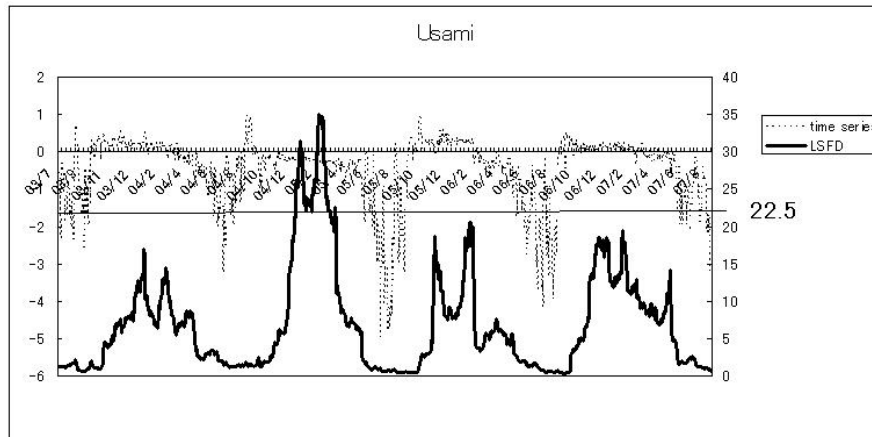
$$0.500 \leq LFD \leq 1.180 \quad (16)$$

Osezaki;

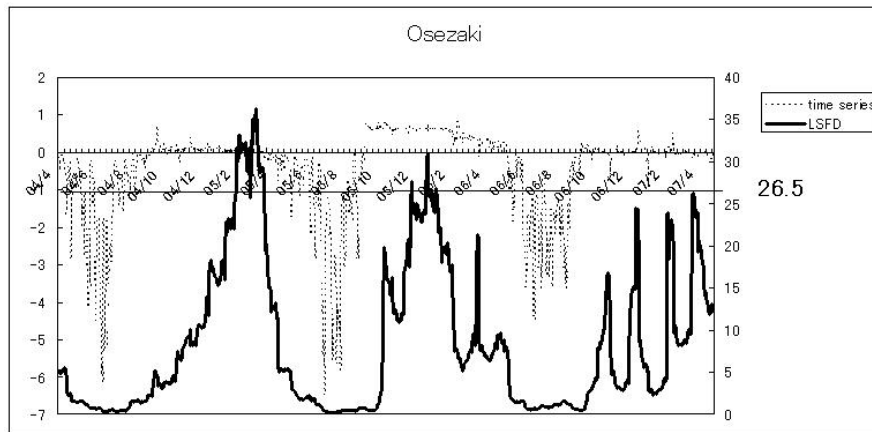
$$0.520 \leq LFD \leq 1.140 \quad (17)$$

#### 4.3 Approach to quality control relating to the global environment

As indicated in the preceding paragraph, because the range of values of individual moving LSF Ds can be inferred at these observation stations, this method can be applied to quality control relating to the global environment. Specifically, Fig. 9 and Fig. 10 show the moving LSF D and the difference time series of seawater temperatures as a function of depth at Usami and



**Fig. 9.** Moving LSF change on difference time series of seawater temperatures as a function of depth(Usami, Japan).



**Fig. 10.** Moving LSF change on difference time series of seawater temperatures as a function of depth(Osezaki, Japan).

Osezaki, respectively. Each graph shows the respective LSF upper limits, values above which are judged to represent an "abnormality".

Fig. 9 shows that abnormal values exceeding the upper limit were observed at Usami from the beginning of 2005; in this year there was a "warm winter", which is considered to be abnormal weather.

Also, Fig. 10 shows that abnormal values exceeding the upper limit were observed at Osezaki as prognostic indices relating to the start of the "Large Black Stream Meander" in July 2004 and its end in August 2005.

Based on a statistical quality control approach, if statistical abnormalities are observed in the LSF, that is, when the observed LSF exceeds its

upper limit, it is natural to assume that there are some kinds of "abnormalities" in the relevant physical systems. Further study is required to determine whether prognostic indices actually exist that can identify such abnormalities relating to the physical quantities in question, and whether or not the "Law of Increasing Local Standard Fractal Dimension" is valid for abnormal phenomena in a variety of actual cases.

## 5 Conclusions and future outlook

The following points summarize the results of this paper:

- 1) As a result of computer simulations, we found that the moving local standard fractal dimension (LSFD) on a uniform or standard normal random process belongs to a non-symmetric normal distribution with a long tail towards the right hand side. We have also found that the LFD distribution exhibits reverse features, having a long tail towards the left hand side.
- 2) The 95% confidence interval of the moving LSFD is considered to be applicable to methods such as control charts, which are often used in statistical quality control. When the moving LSFD falls outside this confidence interval, the process in question is judged to be abnormal.
- 3) This method was applied to the difference time series of seawater temperatures as a function of depth, which is one index of the condition of the global environment. By considering these time series to be a discrete dynamical orbit, the probability distribution of the moving LSFD was shown to generally conform to a power-law distribution.
- 4) Also, prediction of abnormal phenomena, such as warnings of phase transitions etc., in the global monitoring system may be possible using the moving LSFD as an indicator and observing when it increases past the upper 5% significance level, similar to the case of general quality control.

## References

- AYACHE, A. et al. (2007): A central limit theorem for the quadratic variations of the step fractional Brownian motion, *Stat. Inference for Stoch. Processes*, 10, 1-27.
- BENASSI, A. et al. (2000): Identification of the Hurst index of a Step Fractional Brownian Motion, *Stat. Inference for Stoch. Processes*, 3, 101-111.
- KAMIJO, K. and YAMANOUCHI, A. (2008): Time Series Analysis Using Local Standard Fractal Dimension -Application to Fluctuations in Seawater Temperature-, *International Conference on Computational Statistics (COMPSTAT2008)*, Porto, Portugal.
- KAMIJO, K. and YAMANOUCHI, A. (2009): Numerical and Practical Method for Statistical Quality Control Using Local Fractal Dimension in Discrete Time Series, *European Conference on Numerical Mathematics and Advanced Applications (ENUMATH 2009)*, Uppsala, Sweden.

# Assessment of Scoring Models Using Information Value

Jan Koláček<sup>1</sup> and Martin Řezáč<sup>1</sup>

Department of Mathematics and Statistics, Masaryk University  
Kotlářská 2, 611 37 Brno, Czech Republic, *kolacek@math.muni.cz*

**Abstract.** It is impossible to use a scoring model effectively without knowing how good it is. Quality indexes like Gini, Kolmogorov-Smirnov statistics and Information value are therefore used to assess quality of given scoring model.

The paper deals mainly with Information value. Commonly it is computed by discretisation of data into bins using deciles. One constraint is required to be met in this case. Number of cases have to be nonzero for all bins. If this constraint is not fulfilled there are numerous practical procedures for preserving finite results. As an alternative method to empirical estimates we can use the kernel smoothing theory.

**Keywords:** credit scoring, quality indexes, information value, quantiles, kernel smoothing

## 1 Introduction

Credit scoring, it is a term for a wide spectrum of predictive models and their underlying techniques that aid financial institutions in granting credits. These methods decide who will get credit, how much credit they should get, and what further strategies will enhance the profitability of the borrowers to the lenders. Credit scoring techniques assess the risk in lending to a particular consumer.

Methodology of credit scoring models and some measures of their quality were discussed in works like Hand and Henley (1997) or Thomas (2000) and books like Anderson (2007), Siddiqi (2006), Thomas et al. (2002) and Thomas (2009). Further remarks connected to credit scoring issues can be found there as well.

Once a scoring model is available, it is natural to ask how good it is. For a measurement of partial processes of a financial institution, especially their components like scoring models or other predictive models, it is possible to use quantitative indexes such as Gini index, K-S statistics, Lift, Information statistics and so forth. They can be used for comparison of several developed models at the moment of development. It is possible to use them for monitoring the quality of models after the deployment into real business as well.

The aim of this paper is to give an overview of quantitative indexes, to show some theoretical properties, their interactions and especially show the relationship to the Lorenz curve. The paper is focused on the Information value, which is one of the most used indexes in practice. One of scopes of this paper is to show computational problems for this index and propose an alternative method.

## 2 Some quality indexes

Assume the realization  $s \in \mathbb{R}$  of random value  $S$  (score) is available for each client. Let  $D$  be the indicator of good and bad client

$$D = \begin{cases} 1, & \text{client is good} \\ 0, & \text{client is bad} \end{cases}$$

and let  $F_0, F_1$  denote cumulative distribution functions of score of bad and good clients, i.e.

$$\begin{aligned} F_0(a) &= P(S \leq a \mid D = 0), \\ F_1(a) &= P(S \leq a \mid D = 1), \quad a \in \mathbb{R}. \end{aligned}$$

We assume  $F_0, F_1$  and their corresponding densities  $f_0, f_1$  are continuous on  $\mathbb{R}$ .

In practice, the empirical distribution functions as their nonparametric estimates are used

$$\begin{aligned} \hat{F}_0(a) &= \frac{1}{m} \sum_{i=1}^N I(s_i \leq a \wedge D = 0) \\ \hat{F}_1(a) &= \frac{1}{n} \sum_{i=1}^N I(s_i \leq a \wedge D = 1), \quad a \in [L, H], \end{aligned}$$

where  $s_i$  is the score of  $i$ -th client,  $n, m$  are the number of good and bad clients, respectively and  $N = n + m$ .  $L$  is the minimum value of given score,  $H$  is the maximum value.

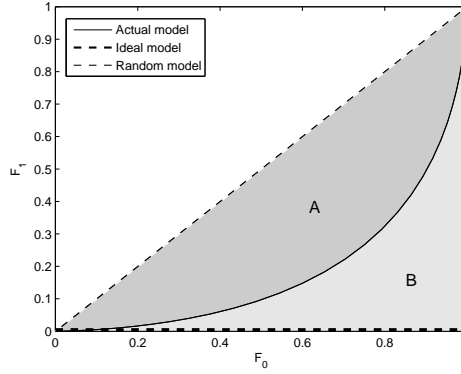
### 2.1 The Lorenz curve

The Lorenz curve (LC) can be successfully used to show the discriminatory power of scoring function, i.e. the ability to identify good and bad clients. The curve is given parametrically by

$$\begin{aligned} x &= F_0(a) \\ y &= F_1(a), \quad a \in [L, H]. \end{aligned}$$

Using the notation  $x = F_0(a)$ ,  $R(x) = F_1(F_0^{-1}(x))$  we can write the Lorenz curve as  $R(x)$ ,  $x \in [0, 1]$ . For a better idea see Figure 1. It illustrates the Lorenz curve for any given model and for two extreme situations; an ideal model (the best) and a random model (the worst).





**Fig. 1.** Lorenz curve, Gini index

## 2.2 Kolmogorov-Smirnov statistics

An often-used characteristic in describing the quality of the model (scoring function) is Kolmogorov-Smirnov statistics (K-S or KS). It is defined as

$$KS = \max_{a \in [L, H]} |F_0(a) - F_1(a)|.$$

In context with notation  $R(x)$  for the Lorenz curve we can express K-S statistics as

$$KS = \max_{x \in [0, 1]} |x - R(x)|,$$

so we can see the K-S statistics is the maximum distance between the Lorenz curve and the diagonal line (represents the random model).

## 2.3 Gini index

In connection to LC we assume next quality measure, Gini index. This index describes a global quality of scoring function. It takes values from 0 to 1. The ideal model, i.e. scoring function that perfectly separate good and bad clients, has Gini index equal to 1. On the other hand, model that assigns a random score to the client has this index equal to 0. Using Figure 1 it can be defined as

$$Gini = \frac{A}{A + B} = 2A.$$

The numerator  $A$  represents the area between the diagonal (random model) and Lorenz curve (actual model) and the denominator  $A + B = 1/2$  represents the area between the diagonal and Lorenz curve for an ideal model. For further details see Anderson (2007) or Xu (2003).

## 2.4 The cumulative Lift

Another possible indicator of the quality of scoring model can be *cumulative Lift*, which says, how many times, at a given level of rejection, is the scoring model better than random selection (random model). More precisely, the ratio indicates the proportion of bad clients with less than a score  $a$ ,  $a \in [L, H]$ , to the proportion of bad clients in the general population. Formally, it is defined as

$$Lift(a) = \frac{P(D = 0 | S \leq a)}{P(D = 0)} = \frac{P(S \leq a | D = 0)}{P(S \leq a)} = \frac{F_0(a)}{F(a)},$$

where

$$F(a) = P(S \leq a) = P(S \leq a \wedge D = 0) + P(S \leq a \wedge D = 1).$$

If we denote  $k = P(D = 0)$  the proportion of bad clients, we can write the Lift function as

$$Lift(a) = \frac{F_0(a)}{kF_0(a) + (1 - k)F_1(a)}, \quad a \in \mathbb{R}.$$

In context with reparametrization  $R(x)$  for the Lorenz curve we can express the Lift by

$$Lift(x) = \frac{x}{kx + (1 - k)R(x)}, \quad x \in [0, 1].$$

In practice, distribution functions  $F_0$  and  $F$  are replaced by their empirical estimates

$$\widehat{Lift}(a) = \frac{\widehat{F}_0(a)}{\widehat{F}_N(a)}, \quad a \in [L, H].$$

## 3 The information value

The quality index based on densities is the information statistics (value)  $I_{val}$ , defined in Hand and Henley (1997) as

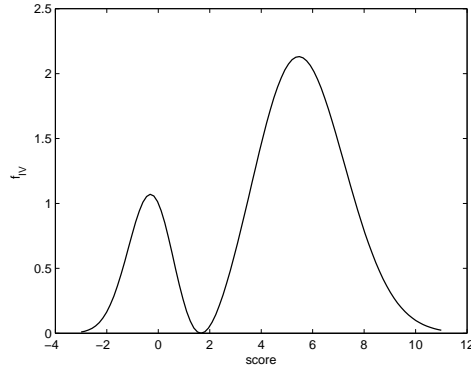
$$I_{val} = \int_{-\infty}^{\infty} f_{IV}(x) dx,$$

where

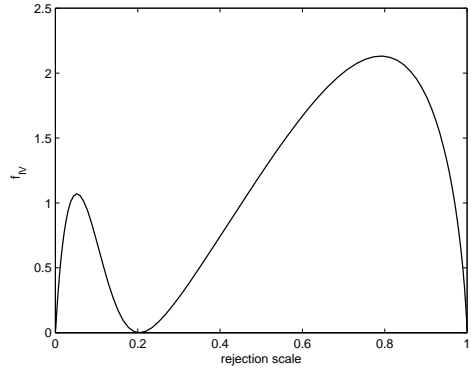
$$f_{IV}(x) = (f_1(x) - f_0(x)) \ln \left( \frac{f_1(x)}{f_0(x)} \right).$$

The example of  $f_{IV}(x)$  for 10% of bad clients with  $f_0 \sim N(0, 1)$  and 90% of good clients with  $f_1 \sim N(4, 2)$  is illustrated in Figure 2.

In practice, it is useful to examine  $f_{IV}$  ( $y$ -label) versus the cumulative distribution  $F$  for score of all clients ( $x$ -label). This is illustrated in Figure 3.



**Fig. 2.** Information value



**Fig. 3.**  $I_{val}$  vs. rejection scale

The  $x$ -label represents so-called “rejection scale”. It means, that we can see the value of  $f_{IV}$  (resp.  $I_{val}$ ) for each percentile of rejected clients.

According to reparametrization  $R(x)$  for the Lorenz curve we can express the  $I_{val}$  by

$$I_{val} = \int_0^1 (R'(t) - 1) \ln(R'(t)) dt.$$

However, in practice, the procedure of computation of the Information value index can be a little bit complicate. As the first, we don't know the right form of densities  $f_0, f_1$  generally and as the second, mostly we don't know how to compute the integral. We show some approaches to solve these computational problems.

### 3.1 Empirical estimates

The main idea of this approach is to replace unknown densities by their empirical estimates. Let's have  $m$  score values  $s_{0_i}$ ,  $i = 1, \dots, m$  for bad clients and  $n$  score values  $s_{1_j}$ ,  $j = 1, \dots, n$  for good clients and denote  $L$  (resp.  $H$ ) as the minimum (resp. maximum) of all values. Let's divide the interval  $[L, H]$  up to  $r$  equal subintervals  $[q_0, q_1], (q_1, q_2], \dots, (q_{r-1}, q_r]$ , where  $q_0 = L, q_r = H$ . Set

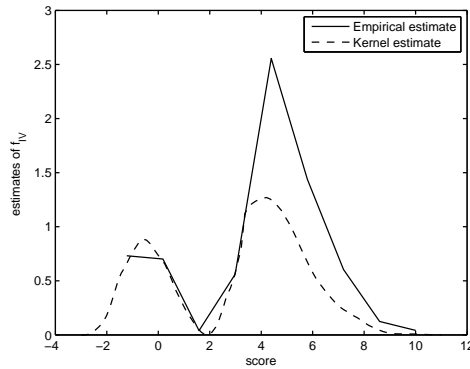
$$n_{0_j} = \sum_{i=1}^m I(s_{0_i} \in (q_{j-1}, q_j])$$

$$n_{1_j} = \sum_{i=1}^n I(s_{1_i} \in (q_{j-1}, q_j]), \quad j = 1, \dots, r$$

observed counts of bad or good clients in each interval. Then the empirical information value is calculated by

$$\hat{I}_{val} = \sum_{j=1}^r \left( \frac{n_{1_j}}{n} - \frac{n_{0_j}}{m} \right) \ln \left( \frac{n_{1_j} m}{n_{0_j} n} \right).$$

However in practice, there could occur computational problems. The Information value index becomes infinite in cases when some of  $n_{0_i}$  or  $n_{1_i}$  are equal to 0. When this arises there are numerous practical procedures for preserving finite results. For example we replace the zero entry of numbers of goods or bads by a minimum constant of say 0.0001. Choosing of the number of bins is also very important. In the literature and also in many applications in credit scoring, the value  $r = 10$  is preferred. Figure 4 illustrates the estimation of  $f_{IV}$  for 1000 randomly generated values from the example mentioned above.



**Fig. 4.** Comparison of empirical and kernel estimate

### 3.2 Kernel estimates

In the previous subsection, we described some difficulties arisen by computing the Information value. To avoid them we propose another approach of density estimation. We can use the kernel smoothing theory to obtain estimates of unknown densities  $f_0, f_1$ . The kernel density estimates are defined by

$$\begin{aligned}\tilde{f}_0(x, h_0) &= \frac{1}{m} \sum_{i=1}^m K_{h_0}(x - s_{0_i}), \\ \tilde{f}_1(x, h_1) &= \frac{1}{n} \sum_{i=1}^n K_{h_1}(x - s_{1_i}),\end{aligned}$$

where  $K_{h_i}(x) = \frac{1}{h_i} K\left(\frac{x}{h_i}\right)$ ,  $i = 0, 1$  and  $K$  is the Epanechnikov kernel

$$K(x) = \frac{3}{4}(1 - x^2)I_{[-1,1]}.$$

For further details see Wand and Jones (1995). The quality of kernel density estimates is affected mainly by smoothing parameters  $h_0$  and  $h_1$ . The estimation of optimal bandwidths  $h_i$  can be given by maximal smoothing principal approach, i.e.

$$\begin{aligned}\tilde{h}_{0,m} &= 2,5324 \tilde{\sigma}_0 m^{-\frac{1}{5}} \\ \tilde{h}_{1,n} &= 2,5324 \tilde{\sigma}_1 n^{-\frac{1}{5}}\end{aligned}$$

where  $\tilde{\sigma}_i, i = 0, 1$  are appropriate estimations of standard deviation for bad and good clients. For more details see Terrell (1990).

As the next step we need to estimate the final integral. We use the composite trapezoidal rule. Set

$$\tilde{f}_{IV}(x) = (\tilde{f}_1(x, h_1) - \tilde{f}_0(x, h_0)) \ln \left( \frac{\tilde{f}_1(x, h_1)}{\tilde{f}_0(x, h_0)} \right).$$

Then, for given  $M + 1$  equidistant points  $L = x_0, x_1, \dots, x_M = H$  we obtain

$$\tilde{I}_{val} = \frac{H - L}{2M} \left( \tilde{f}_{IV}(L) + \sum_{i=1}^{M-1} \tilde{f}_{IV}(x_i) + \tilde{f}_{IV}(H) \right).$$

See Figure 4 for a comparison of the kernel estimate  $\tilde{f}_{IV}(x)$  with  $\tilde{h}_{0,m} = 0,9266, \tilde{h}_{1,n} = 1,3047$  and the empirical estimate.

## 4 Conclusions

All described quantitative indexes are widely used in practice to assess quality of given scoring model. They all are based on some conditional probabilities

and from the paper it can be seen their interactions and the relationship to the Lorenz curve.

Although theoretical definitions and formulas for mentioned indicators are quite easy, there could arise some computational problems. We have focused on the Information value and described difficulties of its estimation. The most popular method is the empirical estimator. But it can lead to infinite values of  $I_{val}$  and so a remedy is necessary. To avoid these difficulties the kernel method was proposed. The advantage of this approach is the smoothness of the contribution and easy implementation with a polynomial kernel.

## 5 Acknowledgments

The research was supported by our department and by The Jaroslav Hájek center for theoretical and applied statistics (grant No. LC 06024).

## References

- ANDERSON, R. (2007): *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, Oxford.
- HAND, D.J. and HENLEY, W.E. (1997): Statistical Classification Methods in Consumer Credit Scoring: a review. *Journal of the Royal Statistical Society, Series A*. 160 (3), 523-541.
- SIDDIQI, N. (2006): *Credit Risk Scorecards: developing and implementing intelligent credit scoring*. Wiley, New Jersey.
- TERRELL, G.R. (1990): The Maximal Smoothing Principle in Density Estimation. *Journal of the American Statistical Association* 85, 470-477.
- THOMAS, L.C. (2000): A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting* 16 (2), 149-172.
- THOMAS, L.C. (2009): *Consumer Credit Models: Pricing, Profit, and Portfolio*. Oxford University Press, Oxford.
- THOMAS, L.C., EDELMAN, D.B., CROOK, J.N. (2002): *Credit Scoring and Its Applications*. SIAM Monographs on Mathematical Modeling and Computation, Philadelphia.
- WAND, M.P. and JONES, M.C. (1995): *Kernel smoothing*. Chapman and Hall, London.
- XU, K. (2003): How has the literature on Gini's index evolved in past 80 years?. *economics.dal.ca/RePEc/dal/wparch/howgini.pdf*. Accessed on 1 December 2009.

# The Moving Average Control Chart Based on the Sequence of Permutation Tests

Grzegorz Konczak<sup>1</sup>

Karol Adamiecki University of Economics in Katowice  
40-226 Katowice, Bogucicka 14, Poland, *grzegorz.konczak@ae.katowice.pl*

**Abstract.** Shewhart control charts are widely accepted as useful tools for monitoring manufacturing processes. The construction of control chart is based on the sequence of parametric tests. The classical methods for monitoring the process mean in quality control procedures are based on the normality assumption.

Permutation tests could be used even if the normality assumption is not fulfill. In the paper there is presented a modification of moving average control chart for monitoring process mean based on the sequence of permutation tests. This control chart could be used even if the distribution of the process is non-normal. The properties of the proposed control chart are considered in the Monte Carlo study.

**Keywords:** moving average, control charts, permutation tests, bootstrap, Monte Carlo

## 1 Introduction

The classical methods for monitoring the process mean in quality control procedures are based on the normality assumption. The Shewhart control charts are graphical representations of the sequence of parametric tests. It is important to remember assumptions such as normality and independence.

Permutation tests could be used even if the normality assumption is not fulfill. The control chart based on the sequence of permutation tests for two samples is presented in the paper. The permutation control chart can be used in short production run situation. The properties of the proposed control chart are considered in the Monte Carlo study.

## 2 The moving average control chart

Moving Average (MA) Control Charts are used for detecting shifts in the process mean. They will detect small shifts much faster than Shewhart charts with the same sample size. The MA chart is a time weighted control chart based on a unweighted moving average. This control chart is more effective than the Shewhart chart in detecting small process shift.

Suppose that the process  $X_1, X_2, \dots, X_n$  is monitored. Let us denote the collected observations by  $x_1, x_2, \dots, x_n$ . The moving average at time  $i$  is defined as (Montgomery (1996))

$$MA_i = \frac{x_{i-w+1} + x_{i-w+2} + \dots + x_i}{w} \quad (1)$$

where  $i \geq w$ .

Let us assume that  $X_1, X_2, \dots, X_n$  are independent and normal distributed with expected value  $\mu$  and variance  $\sigma^2$ . The expected value of the  $MA_i$  can be written as follows:

$$E(MA_i) = E\left(\frac{x_{i-w+1} + x_{i-w+2} + \dots + x_i}{w}\right) = \frac{1}{w} \sum_{j=i-w+1}^i E(X_j) = \frac{1}{w} \mu = \mu \quad (2)$$

and the variance

$$V(MA_i) = \frac{1}{w^2} \sum_{j=i-w+1}^i V(X_j) = \frac{1}{w^2} \sum_{j=i-w+1}^i \sigma^2 = \frac{\sigma^2}{w} \quad (3)$$

If  $\mu_0$  denotes the target value of the mean then the control lines could be written as follows

$$\begin{aligned} UCL &= \mu_0 + 3 \frac{\sigma}{\sqrt{w}} \\ CL &= \mu_0 \\ LCL &= \mu_0 - 3 \frac{\sigma}{\sqrt{w}} \end{aligned} \quad (4)$$

The control limits could be established for unknown parameters. In this case the parameters should be estimated based on the sample taken when process is thought to be in control. The parameters are estimated using formulas:  $\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  and  $\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

The control lines could be written as follows (Montgomery (1996))

$$\begin{aligned} UCL &= \bar{x} + 3 \frac{s}{c_4(w)\sqrt{w}} \\ CL &= \bar{x} \\ LCL &= \bar{x} - 3 \frac{s}{c_4(w)\sqrt{w}} \end{aligned} \quad (5)$$

where  $c_4(w) = \frac{E(s)}{\sigma}$ .

The moving average chart gives equal weight to the  $w$  most recent observations and zero weights to all other observations. These control charts are useful when the single observations themselves are used. The  $MA$  control chart could be used for monitoring short run processes but the distribution in the small sample sizes cases have to be normal. Some authors considered non-parametric control charts for monitoring non-normal processes (eg. Bertrand and Fleury (2008) or Chakraborti et al. (2004)).



### 3 Monitoring non-normal processes

There is a close connection between control charts and hypothesis testing. The parametric tests are used in Shewhart control charts constructing. The main assumptions in Shewhart control charts are independence and normality distribution of a quality characteristic. In many situations we may have reason to doubt the validity of normality assumption. To establish classical control chart we have to know the process parameters  $\mu$  and  $\sigma$ . In practice  $\mu$  and  $\sigma$  are usually not known. In these cases we have to estimate the parameters from the preliminary samples taken when process is thought to be in control. The type of the distribution should be determined. In the computer simulation the results of testing normality for three non-normal distribution are considered. The following distributions (see Table 1) were analyzed in this study:

- log-normal distribution with parameters  $\mu$  and  $\sigma$ ,
- beta distribution with the  $shape(s_1)$  and the  $scale(s_2)$  parameters,
- gamma distribution with the  $shape(s)$  parameter.

To test the normality Lilliefors test and Shapiro-Wilk test were used. The Monte Carlo study for samples of sizes  $n = 5$  and  $n = 15$  have been done. Table 1 shows the estimated probabilities of failing to reject the null hypothesis (normality) for these non-normal distributions. The estimated probabilities are based on the 10 000 simulations in each case. For considered non-normal distributions we get very often the decision "fail to reject". It could be said that the experimental data does not decisively reject the null hypothesis. In such cases the classical control charts, designed for use in normality assumption, are often used, but the normality assumption is not fulfill.

**Table 1.** The estimated probabilities of failing to reject the normality hypothesis

Distribution		Lilliefors test		Shapiro-Wilk test	
Type	Parameters	$n = 5$	$n = 15$	$n = 5$	$n = 15$
Log-normal	LN(0;1)	0.8104	0.3443	0.7667	0.1760
	LN(0.5;1)	0.8075	0.3494	0.7652	0.1756
	LN( $\mu; \sigma$ )	0.8070	0.3488	0.7657	0.1776
Beta	B(2;5)	0.9463	0.9084	0.9378	0.8736
	B(2;4)	0.9535	0.9287	0.9491	0.9065
	B( $s_1; s_2$ )	0.9541	0.9453	0.9521	0.9403
Gamma	$\Gamma(2)$	0.9118	0.7538	0.9062	0.6071
	$\Gamma(3)$	0.9273	0.8145	0.9225	0.7328
	$\Gamma(s)$	0.9398	0.8591	0.9340	0.7930

Source: Monte Carlo study

#### 4 Parametric tests versus permutation tests

Permutation tests were introduced by R.A. Fisher in the early 1930's. These tests are a computer-intensive statistical methods. The main application of these tests is two sample problem (Efron and Tibshirani (1993)). The Shewhart control charts are based on the sequence of parametric tests. The sequence of permutation tests will be used for the construction permutation control chart. The main idea of permutation tests is attractively simple and free of mathematical assumptions. Let us assume that the samples  $S_1 = \{x_{11}, x_{12}, \dots, x_{1n_1}\}$  of size  $n_1$  and  $S_2 = \{x_{21}, x_{22}, \dots, x_{2n_2}\}$  of size  $n_2$  were taken from possibly different continuous distributions  $F$  and  $G$ . The null hypothesis that the samples were taken from identically distribution will be tested (Sheskin (2004)). This hypothesis can be written  $H_0 : F = G$  against the alternative hypothesis  $H_1 : F \neq G$ . Let  $\alpha$  be the significance level ( $\alpha$  usually in hypothesis testing is equal to 0.05, 0.01 or 0.1 and in constructing control charts is equal to 0.0027). Let  $\bar{X}_1$  and  $\bar{X}_2$  are the samples means  $\bar{X}_k = \frac{1}{n_k} \sum_i^{n_k} x_{ki}$ , for  $k = 1, 2$ .

Let us consider the statistic:

$$T = \bar{X}_1 - \bar{X}_2 \quad (6)$$

Let us denote by  $T_0$  the value of this statistic calculated for the samples  $S_1$  and  $S_2$ . The very big or the very small values are against the null hypothesis. Having observed  $T_0$ , the achieved significance level (*ASL*) of the test is defined to be the probability of observing at least that large or at most as small value when the null hypothesis is true. *ASL* could be written as follows:

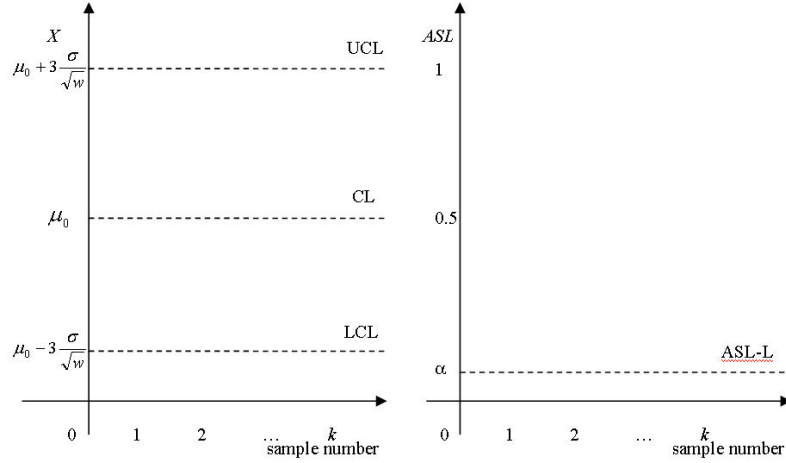
$$ASL = P_{H_0}(|T^*| \geq |T_0|) \quad (7)$$

where the statistic  $T^*$  has the null hypothesis distribution, the distribution of  $T$  if  $H_0$  is true. The small value of *ASL*, the stronger the evidence against  $H_0$ . If the value of *ASL* is less or equal to assumed significance level  $\alpha$  then the null hypothesis is rejected. If the null hypothesis is true, then the distribution of the statistic  $T^*$  can be estimated. The set  $S = S_1 \cup S_2$  is randomly divided into two samples of sizes  $n_1$  and  $n_2$ . This step is repeated  $N$  times ( $N$  should be at least 1000). For each case the value  $T_i (i = 1, 2, \dots, N)$  using formula (6) is calculated. Then the value of *ASL* is estimated following

$$ASL \approx \frac{\text{card}(T_i : |T_i| \geq |T_0|)}{N} \quad (8)$$

#### 5 MA control chart based on permutation tests

Let us assume that samples  $S_{11}, S_{12}, \dots, S_{1k}$  each of size  $w$  were taken from process, with distribution  $F$ , which is thought to be in control. Let  $S_1 = S_{11} \cup S_{12} \cup \dots \cup S_{1k}$ . Suppose that individual observations  $x_1, x_2, \dots, x_{n+w-1}$



**Fig. 1.** Moving Average control chart (left) and permutation control chart (right).

from distribution  $G$  have been collected. Let us consider the sets  $S_{2i} = \{x_i, x_{i+1}, \dots, x_{i+w-1}\}$ , for  $i = 1, 2, \dots, n$ . The hypothesis that the samples  $S_1$  and  $S_{2i}$  have the same distributions will be tested. To test the  $H_0$  hypothesis permutation test instead of parametric test will be used. In the  $MA$  chart the values of  $MA_i$  are pointed on the chart. There are three horizontal control lines in this control charts: upper control line ( $UCL$ ), central line ( $CL$ ) and lower control line ( $LCL$ ). The levels of upper and lower control lines are related to the critical values in parametric test. A point plotting outside the control limits is equivalent to reject the null hypothesis and the point plotting between control lines is equivalent to failing to reject this hypothesis in parametric test. On the permutation control chart (MA-P) there is only one control line connected to the significance level  $\alpha$  (Figure 1). There are plotted calculated  $ASL_i$  values for a sample  $t$  versus the sample number  $i$  ( $i = 1, 2, \dots, n$ ). A point plotting below the  $ASL$  line ( $ASL - L$ ) is equivalent to reject the null hypothesis in permutation test.

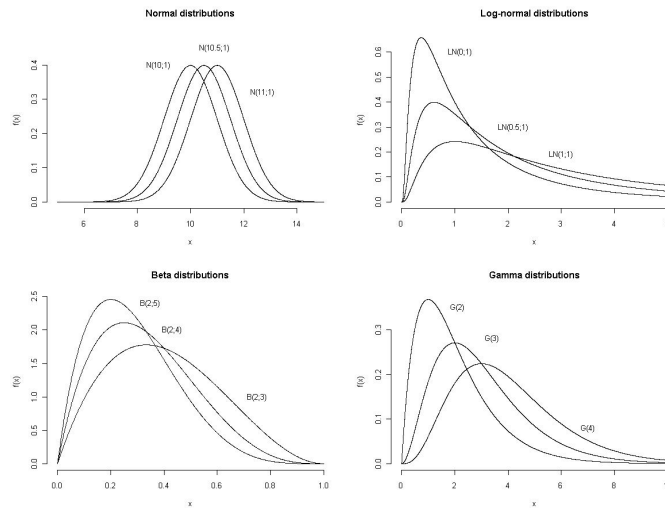
## 6 Monte Carlo study

To compare the properties of the moving average permutation control chart and the MA control chart a series of computer simulations have been done. The procedures were written using R language (<http://www.r-project.org> and Crawley (2005)). There were four types of distributions analyzed in the Monte Carlo study. For each of them the values from distributions  $G_X$ ,  $G_Y$  and  $G_Z$  were generated. The details of random variables under the study are presented in Table 2. There four in control processes (one normal and three non-normal) are considered in the study. The graphical view of densities of random variables used in the Monte Carlo study is presented in the Figure 2. The steps of the simulation study were following:

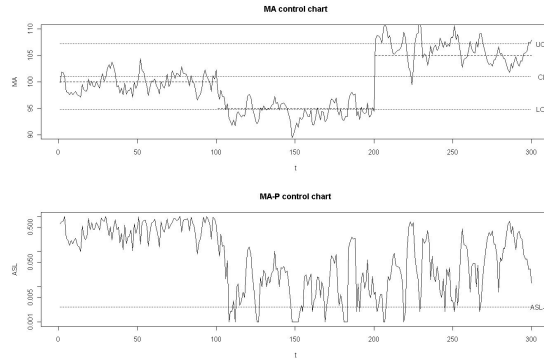
**Table 2.** The random variables details considered in the Monte Carlo study

The type of distribution	In control process	Distribution (G)		
		$G_X$	$G_Y$	$G_Z$
Normal	N(10.0;1)	N(10.0;1)	N(10.5;1)	N(11.0;1)
Log-normal	LN(0;1)	LN(0;1)	LN(0.5;1)	LN(1;1)
Beta	B(2;5)	B(2;5)	B(2;4)	B(2;3)
Gamma	$\Gamma(2)$	$\Gamma(2)$	$\Gamma(3)$	$\Gamma(4)$

Source: Monte Carlo study

**Fig. 2.** The densities of random variables under study.

1.  $k$  samples of size  $w$  ( $k = 3, w = 5$ ) from distribution  $F_X$  (in-control process) were generated. The process parameters were estimated.
2. 300 values were generated (100 from  $G_X$  distribution, 100 from  $G_Y$  distribution and 100 from  $G_Z$  distribution). One of the generated samples is presented in Figure 3 where the dashed line represents the process mean.
3. For each  $i$  ( $i = 1, 2, \dots, 300$ ) two tests was performed
  - a)  $t$  test for one sample where the tested value was estimated from the sample from  $F_X$  distribution
  - b) permutation test for two samples
4. Steps 1-3 were repeated  $N = 1000$  times.



**Fig. 3.** Moving Average (MA) control chart and Permutation control chart (MA-P).

5. The empirical probabilities of rejection  $H_0$  for  $t$  test and permutation test were calculated.

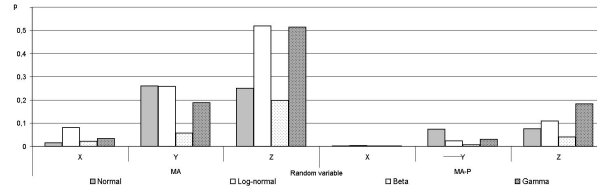
There was assumed the significance level  $\alpha = 0.0027$  in the Monte Carlo study. The estimated probabilities of rejection of the null hypothesis and the estimated values of ARL for MA and  $MA - P$  control charts are presented in Table 3 and in Figure 4. The standard error of the fraction estimation is less than 0.0016. It can be noticed, that probabilities of rejection of the null hypothesis for MA chart is no close to  $\alpha$  under  $H_0$ .

**Table 3.** The estimated probabilities of the rejection of the null hypothesis

The type of distribution	In control process $F_X$	Estimated probability (estimated ARL)					
		MA			MA-P		
		Distribution			Distribution		
		$G_X$	$G_Y$	$G_Z$	$G_X$	$G_Y$	$G_Z$
Normal	N(10;1)	0.0157 (63.7)	0.2604 (3.8)	0.2499 (4.0)	0.0017 (588.2)	0.0742 (13.5)	0.0753 (13.3)
Log-normal	LN(0;1)	0.0799 (12.5)	0.2586 (3.9)	0.5196 (1.9)	0.0027 (370.4)	0.0233 (42.9)	0.1104 (9.1)
Beta	B(2;5)	0.0218 (45.9)	0.0579 (17.3)	0.1958 (5.1)	0.0024 (416.7)	0.0063 (158.7)	0.0411 (24.3)
Gamma	$\Gamma(2)$	0.0334 (29.9)	0.1884 (5.3)	0.5151 (1.9)	0.0021 (476.2)	0.0307 (32.6)	0.1839 (5.4)

Source: Monte Carlo study

It can be noticed that for analyzed samples of sizes  $n = 5$  the permutation control chart is more accuracy than MA chart even for normal processes. For



**Fig. 4.** The estimated probabilities of the rejection of the null hypothesis.

the permutation control chart an incorrect out-of-control signal or false alarm is generated with probability close to the assumed  $\alpha = 0.0027$ . For the MA control chart an incorrect out-of-control signal or false alarm is generated with probability much greater than  $\alpha$  even for normal processes (due to the estimation process parameters based on small sizes of samples).

## 7 Concluding Remarks

The classical control charts may be used under the normality assumption. The construction of the control chart based on the sequence of permutation tests is presented in the paper. Permutation tests are free of mathematical assumptions, especially completely removes the normality condition. The properties of the proposed control chart are considered in the Monte Carlo study. The simulation study has shown that the permutation control chart could be used for monitoring process mean in the short production run situation. This control chart is useful for monitoring non-normal processes. The permutation control chart gives accurate probabilities of the incorrect out-of-control signals even for non-normal processes.

## References

- BERTRAND, P.R. and FLEURY, G. (2008): Detecting Small Shift on the Mean by Finite Moving Average. *International Journal of Statistics and Management System*, vol. 3 no.1-2, 56-73.
- CHAKRABORTI, S., van der Laan, P. and van de Wiel, M.A. (2004): A Class of Distribution-free Control Charts. *Applied Statistics* 53 part 3, 443-462.
- CRAWLEY, M.J. (2005): *Statistics. An Introduction Using R*. John Wiley & Sons, Ltd., London.
- EFRON, B. and TIBSHIRANI, R. (1993): *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- MONTGOMERY, D. C. (1996): *Introduction to Statistical Quality Control*. John Wiley & Sons, Inc.
- SHEKIN, D. J. (2004): *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, Boca Raton.

# Depth Based Procedures for Estimation ARMA and GARCH Models

Daniel Kosiorowski<sup>1</sup>

Department of Statistics, Cracow University of Economics  
ul. Rakowicka 27, Cracow, Poland, *daniel.kosiorowski@uek.krakow.pl*

**Abstract.** In this paper we propose two strategies for robust estimation of ARMA and GARCH models. The propositions are based on two statistical depth functions namely famous regression depth introduced by Rousseeuw and Hubert (1998) and general band depth function introduced by Pintado-Lopez and Romo (2006). We study a performance of the propositions on various time series simulated from ARMA(1,1) and GARCH(1,1) models containing additive outliers.

**Keywords:** depth function, robust estimation, ARMA, GARCH

## 1 Outliers in time series

Often data considered in a broad range of economic applications contains one or more atypical observations called outliers. This refers to observations that are well separated from the majority or a center of the data cloud, or in some way deviate from the general pattern of the data. Outliers in financial or macro-economical time series are more complex than in the other situations, where there is no temporal dependence in the data (for details see Marona et al. (2006)). Time series outliers can have an arbitrarily negative influence on parameter estimates for time series models, and the nature of this influence depends on the type of outlier. In the time series setting we encounter several different types of outliers. From a model - based point of view we have among others additive outliers (AO), replacement outliers (RO) and innovation outliers (IO). The AO model is a special case of the RO model, IO outliers refers to a special type of a process ex. ARMA process with a heavy - tailed distribution of the innovations (for details see Marona et al. (2006)). Further on we use a probability model for time series outliers including additive outliers (AO). Let  $x_t$  be a wide sense stationary core process of interest, and let  $v_t$  be a stationary outlier process which is a non zero fraction  $\varepsilon$  of time i.e.  $P(v_t = 0) = 1 - \varepsilon$ . Under an AO model, instead of  $x_t$  one actually observes

$$y_t = x_t + v_t, \tag{1}$$

where the processes  $x_t$  and  $v_t$  are assumed to be independent of one another.

The AO model will generate mostly isolated outliers if  $v_t$  is i.i.d. process, with scale much larger than that of  $x_t$ . In the presence of the AO outliers

in economical time series classical estimators of ARMA and GARCH models generally became useless.

## 2 Depth function in robust time series analysis

A statistical depth function expresses the centrality or "outlyingness" of an observation within a set of data (or with respect to a probability distribution) and provides a criterion to order observations from center - outwards. For a detailed overview see Serfling (2006) and references therein. For other applications of the statistical depth functions in a robust economical analysis see e.g. Kosiorowski (2007) or Kosiorowski (2008).

We use the notion introduced by Rousseeuw and Hubert (1998) notion of regression depth in order to propose a robust procedure for the  $ARMA(p)$  parameters estimation. The regression depth measures the quality of any candidate fit in a linear regression setting. This fits with higher regression depth fit the data better than does fits with lower regression depth. Hence, the regression depth ranks all possible fits from worst (depth=0) to best (maximal depth) case. The errors in the underlying regression model are assumed to be independent, each having zero median. These are very weak conditions, e.g. the error distribution does not have to be symmetrical, nor does it have to stay the same across different values of predictors. A maximal depth estimator (MDE) is a fit which maximizes regression depth. This is one of the best robust estimators for linear regression (for details see Van Aelst and Rousseeuw (2000)).

We also use a a generalized band depth function introduced by Lopez-Pintado and Romo (2006). Lopez-Pintado and Romo (2006) has extended the notion of statistical depth function to deal with functional observations. They proposed robust graph - based methods for supervised classification of curves. We incorporate their concepts in robust estimation for ARMA and GARCH models.

## 3 Robust estimation for ARMA models

An important class of models for describing the single time series  $z_t$  is the class of autoregressive - moving average models referred to as  $ARMA(p, q)$  models.

$$(z_t - \mu) = \phi_1(z_{t-1} - \mu) + \dots + \phi_p(z_t - \mu) + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} \quad (2)$$

where  $z_t$  is a stationary time series with a fixed mean  $\mu$ ,  $a_t$  is a random residual series,  $\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \mu$  are parameters to be estimated from the data.

ARMA models may be fitted to data, using an iterative cycle of identification, estimation and checking. Classical statistical procedures for estimating



ARMA models based on maximum likelihood, least squares or autocovariance estimates are not robust in the presence of AOs or ROs. Proposed diagnostic procedures generally suffer from the masking problem. Estimates based on regular residuals (M-, S- estimates) are not very robust. This is due to the fact that an outlier in one period, does not only affect the residual corresponding to this period, but it may also affect all the subsequent residuals (for detail see Maronna et al. (2006)). Recently Muller et al. (2009) proposed a generalization of the MM- estimates introduced by Yohai for regression which are robust when the series contain outliers. They showed also several asymptotic properties of the propositions.

In this paper we propose an alternative procedure for robust estimation ARMA models. Our regression depth based proposition is user-friendly and performs well due to the very good statistical properties of the regression depth.

**Proposition 1:** Let  $\mathbb{X}_T = \{y_1, y_2, \dots, y_T\}$ ,  $2 < T$ , denote a time series under consideration. We obtain estimates of the parameters of  $ARMA(p, q)$ ,  $0 < p + q \ll T$  in a two step procedure :

*STEP 1:* We calculate MDE estimates of the  $AR(p)$  part of the underlying process by choosing  $\phi_1, \dots, \phi_p$  as the maximal regression depth estimates applied to a data set  $\mathbb{Y} = \{y_1, \dots, y_{T-p}\}$ ,  $\mathbb{X}_1 = \{y_2, \dots, y_{T-p+1}\}$ , ...,  $\mathbb{X}_p = \{y_{p+1}, \dots, y_T\}$  constructed from  $\mathbb{X}_T$ .

*STEP 2:* We add the  $MA(q)$  part to the estimated in the step 1  $AR(p)$  part by minimizing a robust measure of a dispersion between observed and generated by the model values e.g. MAD (median absolute deviation).

## 4 Robust estimation for GARCH models

Many time series display time - varying dispersion, or uncertainty, in the sense that large (small) absolute innovations tend to be followed by other large (small) absolute innovations. Let  $y_t$  denote the observable univariate discrete-time stochastic process of interest. Denote the corresponding innovation process by  $\epsilon_t$ , where  $\epsilon_t \equiv y_t - E_{t-1}(y_t)$ , and  $E_{t-1}(\cdot)$  refers to the expectation conditional on time  $(t - 1)$  information. A general specification for the innovation process that takes account of the time-varying uncertainty would then be given by

$$\epsilon_t = z_t \sigma_t \quad (3)$$

where  $z_t$  is an i.i.d. mean - zero, unit - variance stochastic process, and  $\sigma_t$  represents the time-t latent volatility; i.e.  $E(\epsilon_t^2 | \sigma_t) = \sigma_t^2$ . In the  $GARCH(p, q)$  model, the conditional variance is parametrized as a distributed lag of past squared innovations and past conditional variances,

$$\sigma_t^2 = c + \sum_{i=1}^q \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^p \sigma_{t-j}^2 \quad (4)$$

$$\equiv \omega + \alpha(B)\epsilon_t^2 + \beta(B)\sigma_t^2 \quad (5)$$

where  $B$  denotes the backshift (lag) operator, i.e.,  $B^i y_t \equiv y_{t-i}$ .

This model is usually estimated by maximum likelihood (QML) assuming that the distribution of one observation conditionally to the past is normal. The QML estimate based on a normal likelihood is very sensitive to the presence of a few outliers in the sample. Several authors proposed robust estimates for  $GARCH(p, q)$  models (see Muller and Yohai (2007)). The main part of the propositions however is based on the predictors of the conditional variance which are very sensitive to large outliers. We propose a strategy based on two statistical depth functions and standard ARMA-based method of identification of the  $GARCH(p, q)$  model. We can use our proposition in the case of analysing several time series generated by the same model. Some part of the time series may be outliers and/or each time series may contain AO outliers. In our opinion our simple depth based proposition could be an alternative to latent variables approaches, which generally need very long time series or to a BM-estimators proposed by Muller and Yohai (2007).

First it is worth noticing that rearranging the terms in (5), we obtain

$$[1 - \alpha(B) - \beta(B)]\epsilon_t^2 = \omega + [1 - \beta(B)]\nu_t \quad (6)$$

where  $\nu_t \equiv \epsilon_t^2 - \sigma_t^2$ . Since  $E_{t-1}(\nu_t) = 0$ , and the  $GARCH(p, q)$  formulation in (6) we can estimate process as an  $ARMA(\max\{p, q\}, p)$  model for the squared innovation process  $\{\epsilon_t^2\}$ .

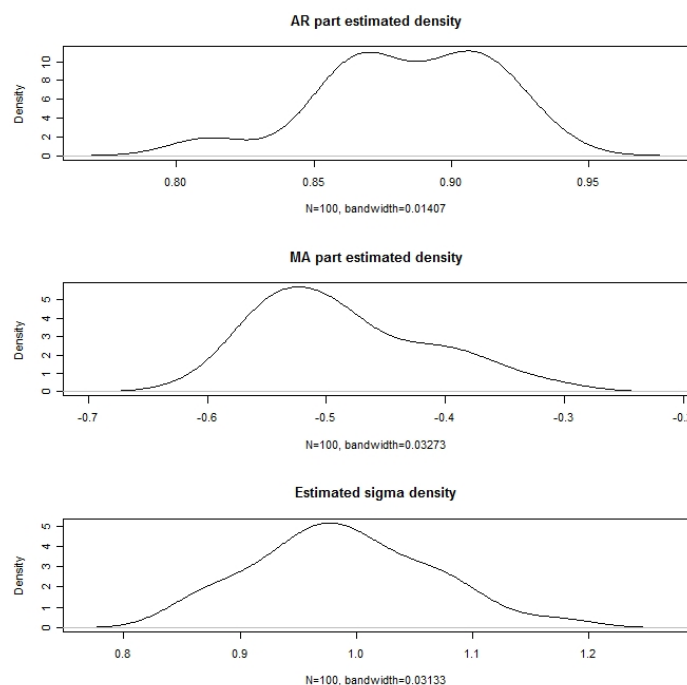
**Proposition 2:** Let  $\mathbf{y}_1 = \{y_1^{(1)}, \dots, y_n^{(1)}\}, \mathbf{y}_2 = \{y_1^{(2)}, \dots, y_n^{(2)}\}, \dots, \mathbf{y}_k = \{y_1^{(k)}, \dots, y_n^{(k)}\}$  denote  $k$  time series generated by  $GARCH(p, q)$  processes containing AO outliers. We obtain a median type estimate of the underlying processes parameters in a two-step procedure.

*STEP 1:* We choose the deepest time series from the  $k$  considered time series as a sample median induced by Pintado-Lopez and Romo generalized band depth.

*STEP 2:* We apply the first proposition to the standard ARMA-based identification of the  $GARCH(p, q)$  process.

## 5 Simulation evidence

To investigate the behaviour of our first proposition we ran simulations with 100 samples of sizes 200 generated by a stationary normal one-dimensional  $AR(1)$  with  $\phi_1 = 0.7$  and two-dimensional  $AR(2)$  with  $\phi_1 = 0.7, \phi_2 = -0.5$  models, both with  $\sigma_{u=1}$  and  $\gamma = 0$ , (scale of the innovations and the intercept). We considered situations when instead of data point one actually observes (AO model)  $y_t = x_t + v_t$ , where the processes  $x_t$  and  $v_t$  are assumed to be independent of one another. We assumed  $v_t$  has a normal mixture distribution  $v_t \sim (1 - \epsilon)N(0, 1) + \epsilon N(0, 100)$  where  $\epsilon = 0\%, 5\%, 10\%, 20\%$ . Table 1 shows



**Fig. 1.** Kernel density estimation of the proposed parameters estimators of the  $ARMA(1,1)$  with  $\phi_1 = 0.7$ ,  $\theta_1 = -0.5$ ,  $\sigma = 1$ . Each of the simulated trajectories contained 5% of the additive outliers.

the results for the maximal regression depth regression estimate for  $AR(1)$  parameters where  $\epsilon = 0\%, 5\%, 10\%, 20\%$  of the AO outliers. Table 2 shows the results the maximal regression depth estimate for  $AR(1,1)$  parameters with  $\epsilon = 0\%, 5\%, 10\%, 20\%$  of the AO outliers. It is easily seen that the conditional least squares (similarly LAD, LTS) estimate is much affected by AO outliers in both situations. Note that the proposed MRE estimate performs better than classical M-estimate for AR parameters. We ran also simulations of sizes  $n = 500$  from  $ARMA(1,1)$  model with  $\phi_1 = 0.9$ ,  $\theta_1 = -0.5$ ,  $\mu = 0$ ,  $\sigma_u = 1$ . Figure 1 shows kernel density estimation of the proposed parameters estimators of the  $ARMA(1,1)$  with  $\phi_1 = 0.7$ ,  $\theta_1 = -0.5$ ,  $\sigma = 1$ . Each of the simulated trajectories contained 5% of the additive outliers. These results seem to be rather promising.

In order to examine the performance of the second proposition we ran simulations of five time series generated by  $GARCH(1,1)$  model with  $\alpha = 0.2$ ,  $\beta = 0.7$ . Each of the simulated five time series was of size  $n = 500$  observations. Two of the simulated five time series contained 10% of the AO outliers (for an example see fig.2). Figure 3 shows results of kernel estimation of the densities of parameters proposed estimators. The results show very

good properties of the proposition in terms of robustness to the AO outliers. Note that the first part of the second proposition (Pintado-Lopez & Romo median) also performs well also in cases of several ARMA time series or real data examples.

Outliers %	$\hat{\phi}_1$
0	0,692
5	0,615
10	0,575
15	0,478

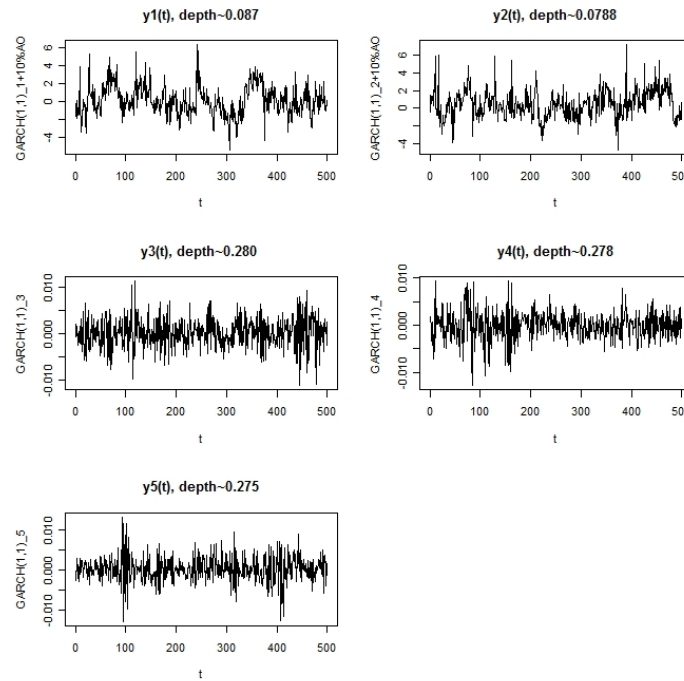
**Table 1.** Percent of the additive outliers in a simulated trajectories from  $AR(1)$  model with  $\phi_1 = 0.7$  vs. mean of the first proposition estimates.

Outliers %	$\hat{\phi}_1$	$\hat{\phi}_2$
0	0.776	-0.492
5	0.631	-0.335
10	0.489	-0.234
15	0.268	-0.075

**Table 2.** Percent of the additive outliers in a simulated trajectories from  $AR(2)$  model with  $\phi_1 = 0.7$ ,  $\phi_2 = 0.2$  vs. mean of the first proposition estimates.

## 6 Conclusions

In our opinion the proposed strategy for  $ARMA(p, q)$  model estimation is an attractive approach to robust estimation of the real economic processes' parameters. Simulation studies show that our approach is not only more robust than conditional least squares or least absolute deviations estimators but also than some new promising propositions as M- or S-estimators (see Marona et al. (2006)), and procedures based on robust filters. Note that our proposition performs well also in situations where data does not contain outliers. We also examined also our proposition on an empirical data set consisting of 200 observations of average exchange rates PLN/USD first differences in 2007 on the basis of the National Bank of Poland's data (the data set without outliers). In the case of  $AR(1)$  we estimated  $\hat{\phi}_1 = 0.0792$  for LS estimate and  $\hat{\phi}_1 = 0.1266$  for the our first proposition regression estimate, in the case of  $AR(2)$  we estimated  $\hat{\phi}_1 = 0.077$ ,  $\hat{\phi}_2 = 0.023$  for LS estimate and  $\hat{\phi}_1 = 0.07$ ,

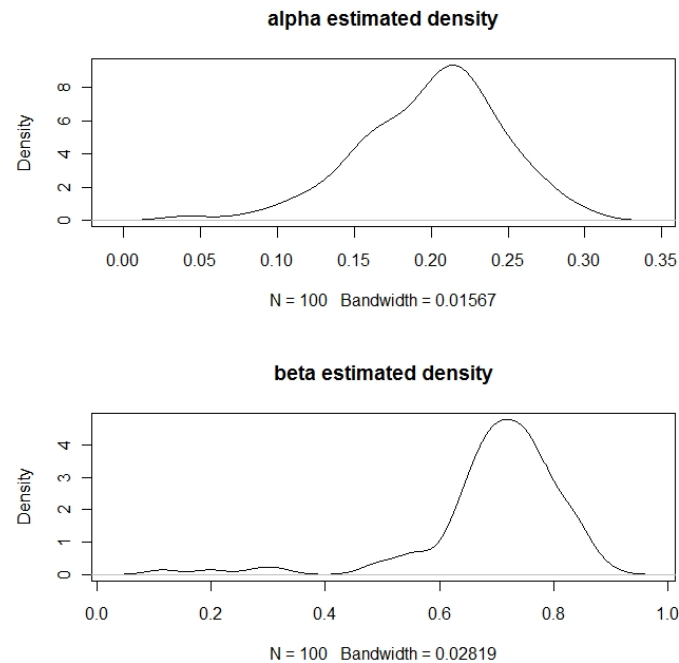


**Fig. 2.** General band depth of five independent trajectories of length 500 simulated from GARCH(1,1) model with  $\alpha = 0.2$  and  $\beta = 0.7$ . Two of the trajectories contain 10% of the additive outliers.

$\hat{\phi}_2 = 0.025$  for the our first proposition estimate. We can see that these differences are not significant. Our second proposition performs well in the case of the model estimation on base of several trajectories generated by the same process  $GARCH(p, q)$ . The trajectories may concern several stock exchange companies, districts, and goods of the same kind. The proposition could be also incorporated to a panel data analysis. We are currently working on the further development of the proposed methods i.e. among others on the robust identification of an ARMA and GARCH model order (we study the possibilities of introducing a depth based information criterion) and we focus our attention on an application the regression depth concept into a general VARIMA framework.

## References

MARONA, R. A., MARTIN R. D., YOHAI V. J. (2006): *Robust Statistics Theory and Methods*. John Wiley & Sons, Chichester.



**Fig. 3.** Kernel density estimation of the proposed parameters estimators of the GARCH(1,1) with  $\alpha_1 = 0.2$ ,  $\beta_1 = 0.7$ . Two of each five of the simulated trajectories contained 10% of the additive outliers.

- LOPEZ-PINTADO, S. and ROMO J. (2006): Depth-based classification for functional data : In: Liu R.Y., Serfling R., Souvaine D. L. (Eds.): *Series in Discrete Mathematics and Theoretical Computer Science*, AMS, vol. 72, 103 - 119.
- KOSIOROWSKI, D. (2007), Nonparametric Equity of Two Shapes Test Based on Multivariate Quantile Functional, *Bulletin of the ISI 56th Session*.
- KOSIOROWSKI, D. (2008), Robust Classification and Clustering Based on the Projection Depth Function, In: Brito P. (Eds.): *Proceedings in Computational Statistics 2008 (COMPSTAT 2008)*, vol. II, p. 209 - 216, Physica - Verlag.
- MULER, N., YOHAI, V. J. (2007): Robust estimates for GARCH models, Technical Report Instituto de Calculo Facultad de Ciencias Exactas y Naturales Universidad de Buenos Aires.
- MULER, N., PENA, D. and YOHAI V. J. (2009) : Robust estimation for ARMA models, *Annals of Statistics* 37 (2), 816-840.
- ROUSSEEUW, P. J., HUBERT, M. (1998) : Regression Depth, *Journal of the American Statistical Association*, 94, 388 - 433.
- SERFLING, R. (2006): Depth Functions in Nonparametric Multivariate Inference: In: Liu R.Y., Serfling R., Souvaine D. L. (Eds.): *Series in Discrete Mathematics and Theoretical Computer Science*, AMS, vol. 72, 1 - 15.
- VAN AELST, S., ROUSSEEUW, P. J. (2000): Robustness Properties of Deepest Regression, *J. Multiv. Analysis*, 73, 82-106.

# Half-Taxi Metric in Compositional Data Geometry Rcomp

Katarina Košmelj<sup>1</sup> and Vesna Žabkar<sup>2</sup>

<sup>1</sup> Biotechnical Faculty, University of Ljubljana  
Ljubljana, Slovenia, *katarina.kosmelj@bf.uni-lj.si*

<sup>2</sup> Faculty of Economics, University of Ljubljana  
Ljubljana, Slovenia, *vesna.zabkar@ef.uni-lj.si*

**Abstract.** Miller (2002) presents the half-taxi metric applicable to compositional data without suggesting how it might be applied. We believe that the half-taxi metric is preferable to other metrics in compositional data geometry rcomp because it takes into account the fact that compositions are closed to one and it has a simple geometric representation on the ternary graph. In an application on advertising expenditure components (Electronic, Print and Online) for 17 European countries in the period 2001-2008 we use the half-taxi metric to detect the structural changes in time, in particular in view of the newer Online component. The results are satisfactory and can be explained in the subject-matter context in view of Hofstede's theory.

**Keywords:** compositional data, R package compositions, online advertising

## 1 Introduction

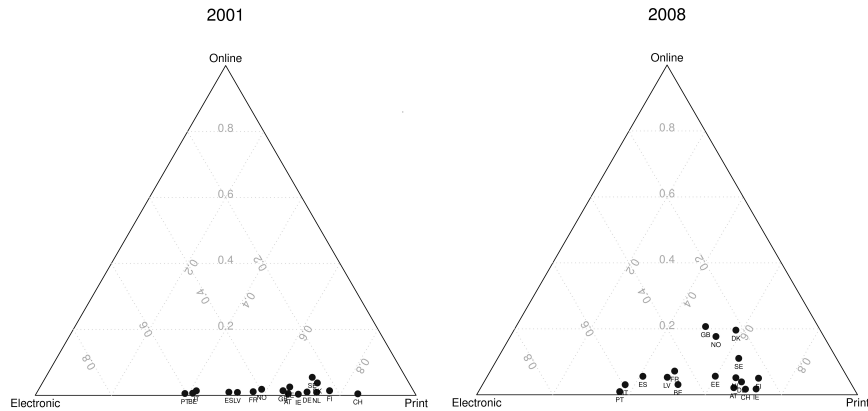
Let us first present the background of the problem. Advertising expenditure (ADSPEND) at the country level includes expenses for the following components: press (newspapers and magazines), television, radio, and outdoor advertising expenditures. In the last decade, however, a new advertising medium-online-has evolved. This newer medium offers different forms of advertising, all supported by the Internet. The first European country with a reported value for the online ADSPEND in the Euromonitor database was Finland in 1996, followed by France, Great Britain, and Sweden in 1997.

Our present analysis focuses on the cluster of 17 stable European countries in the period 2001 to 2008 for the following two reasons. In our previous study (Košmelj and Žabkar, 2008), we detected no significant growth in ADSPEND/GDP in this cluster; that is, on average ADSPEND represented about 0.7 percent of GDP. It can be anticipated that no new money was allocated to ADSPEND within the period under study. Values for Online ADSPEND were not collected/reported before 2001 for all the countries under study. For the period 2001-2008, the Euromonitor database (2009) is complete.

We define three ADSPEND components: 1) Electronic, which summarizes radio and television; 2) Print, which includes press and outdoor; and 3) Online. The data are presented in local currency, therefore not comparable between countries. For that reason we convert the data into proportions: for each country, we calculate the proportion reporting the relative magnitude of a particular component of ADSPEND, in each of the years studied. For each country, we obtain a compositional time series. In Table 1, the data for Denmark are presented.

Table 1: Compositional time series for Denmark for the period 2001-2008; percentage values for Electronic, Print and Online component.

Year	2001	2002	2003	2004	2005	2006	2007	2008
Electronic	24.1	24.4	25.8	26.1	25.8	23.5	22.7	22.1
Print	72.2	70.2	68.3	67.4	66.7	61.2	59.2	58.3
Online	3.8	5.4	5.9	6.5	7.6	15.3	18.1	19.6



**Fig. 1.** Ternary graph for the ADSPEND components Electronic, Print and Online for 17 European countries, for 2001 and for 2008.

In Figure 1 we present the 17 countries in 2001 and in 2008 in a ternary (simplex) graph, a standard graphical presentation for three dimensional compositional data. The three vertices of the equilateral triangle present the components Electronic, Print, and Online, which sum up to 1. In a ternary graph, the points represent units in a particular way. For example, the triplet (0.221, 0.583, 0.196) for Denmark (DK) in 2008 will plot a distance 0.221 from the opposite side of vertex Electronic, a distance 0.583 from the opposite side of vertex Print, and a distance 0.196 from the opposite side of vertex Online.

Figure 1 reveals increased impact of the Online component in time. Our main objective is the assessment of structural changes in ADSPEND compo-



nents in time. Our basic questions are: to what extent is Online advertising substituting advertising in other media? For which countries is an increase in Online made on the account of Print, on the account of Electronic or on the account of both?

The structure of the paper is as follows: in the next section we briefly describe the methodology we used to address the objectives, the third section presents the results obtained by several statistical methods, the last section is devoted to conclusions.

## 2 Methodology

### 2.1 Rcomp geometry

Compositional data analysis can be based on different geometries. This concept is very successfully implemented in the R package, called *compositions* (van den Boogaart and Tolosana-Delgado, 2008) which provides an excellent tool to analyze amount data sets and compositional data sets in four different geometries. The first two geometries are on amounts: *rplus* in a real, classical geometry, and *aplus* in a logarithmic geometry. The last two geometries are on proportions: *rcomp* for closed compositions in a real geometry, and *acom* (Aitchison composition) for closed compositions in a logistic geometry.

As presented before, our input data are in proportions, thus *rcomp* and *acom* geometries are applicable. The main difference in these two geometries is in the evaluation of the distance: for *rcomp* geometry the distance is based on an absolute scale difference (for example, the distance from 1 to 2 is equal to the distance from 51 and 52); for *acom* geometry the distance is based on a relative scale (the distance from 1 to 2 is equal to the distance from 10 to 20).

For the components in our data and for the assessment of their time trend, the time difference in a particular component is meaningful and important from the subject-matter point of view. This difference is expressed in percentage points; for example, for Denmark the increase in the Online component from 2001 to 2008 is 15.8 percentage points (Table 1). Therefore we choose the *rcomp* geometry.

### 2.2 Distance in *rcomp* geometry

For several multivariate methods, e.g., cluster analysis and multidimensional scaling, the first step is the calculation of a distance matrix between units; in our case between countries at a particular time point. The question is: which distance is natural for compositional data in *rcomp* geometry?

Let us assume we have units  $a = (a_1, a_2, a_3)$  and  $b = (b_1, b_2, b_3)$  in *rcomp* geometry; their values are positive and closed to 1:  $\sum a_i = 1, \sum b_i = 1$ . Units

can be represented as points in a ternary graph. Let us define a similarity between two points within the ternary graph as follows:

$$s(a, b) = \min(a_1, b_1) + \min(a_2, b_2) + \min(a_3, b_3)$$

Similarity is defined as the sum of the minimal components; for two identical points similarity equals one, for two vertex points it is zero. A simple mathematical expression for the minimal value of two positive  $x$  and  $y$ ,

$$\min(x, y) = \frac{1}{2}(|x + y| - |x - y|)$$

allows us to simplify the expression taking into account that compositions are closed to 1:

$$s(a, b) = 1 - \frac{1}{2}[|a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3|]$$

Measure of dissimilarity is presented in a complementary form, i.e., subtracted from 1:

$$d(a, b) = 1 - s(a, b) = \frac{1}{2}[|a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3|]$$

Generalization to more than three dimensions is based on mathematical induction.

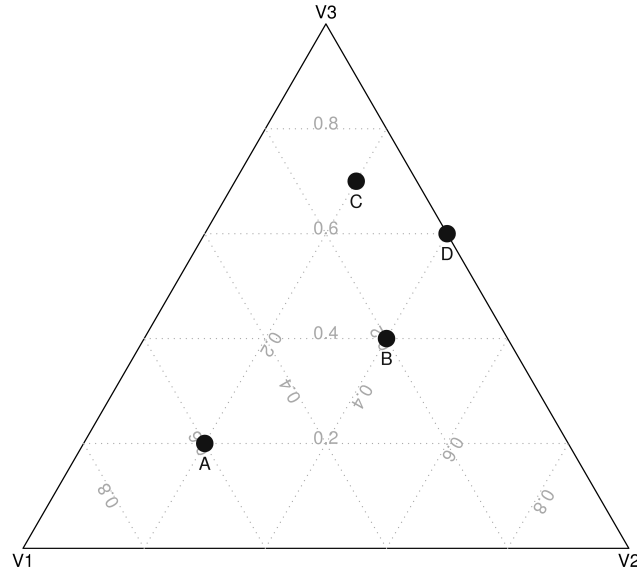
Recently, Miller (2002) named this measure *half-taxi* metric, because it represents one half of the standard taxi (Manhattan) distance. It is proportional to the taxi metric, therefore it is also a metric, and it is natural to assume that other properties are transferred to it. However, its unit circle is a hexagon within the ternary graph. The taxi and the half-taxi metric also have similar geometric interpretations: the first presents the shortest path between two points on the rectangular grid, the second on a triangular grid (see Figure 2).

This measure was presented in the literature long time ago. It is referred to as SIM7 in the paper by Hajdu (1981) which is on resemblance measures in phytosociology. Hajdu presents its alternative names (percentage similarity of distribution, relativized Czekanowski coefficient, relative absolute value function) and the older literature sources (Renkonen, 1938; Whittaker, 1952, Orloci, 1973).

In rcomp geometry Euclidean distance and other distances may also be used. However, the half-taxi distance is preferable due to the following reasons: it takes into account the fact that compositions are closed to 1; it has a simple geometric representation on the ternary graph.

### 3 Results

We define the distance between two time trajectories  $D$  as a simple linear combination of time distances  $d_t$  and weights  $w_t$ :  $D = \sum w_t \cdot d_t$ . A preferable choice for the rcomp distance  $d_t$  is the half-taxi metric. The choice for



**Fig. 2.** Half-taxi distance between two units is the shortest path between the corresponding two points on the triangular coordinate system. For example,  $d(A,B) = 0.4$ ,  $d(A,D) = 0.6$ ,  $d(C,D) = 0.2$ .

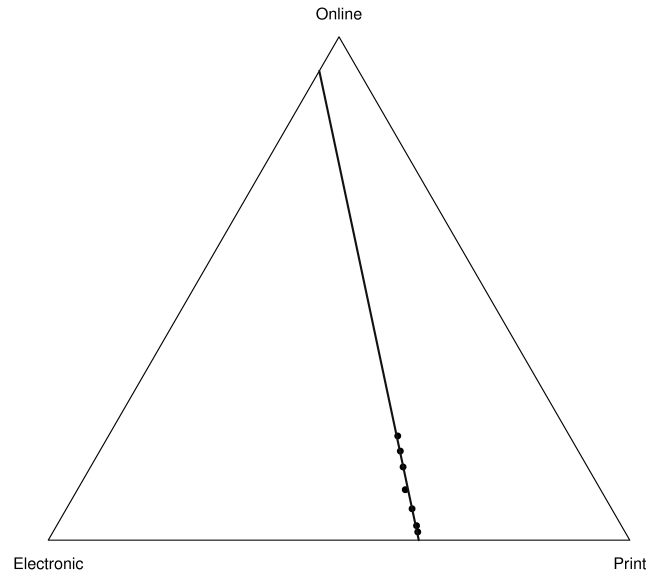
the weights  $w_t$  should reflect the relative importance of a particular time distance; an important variable for the Online component is the internet coverage. Table 2 presents the number of internet users per thousand inhabitants in Europe for the period 2001 -2008; internet coverage in these countries increased from 30% in 2001 to 66% in 2008. Cluster analysis and metric scaling were applied using the distance between time trajectories  $D$ .

Table 2. Number of Internet users per thousand inhabitants in Europe in the period 2001-2008 (Euromonitor, 2009).

Year	2001	2002	2003	2004	2005	2006	2007	2008
Internet coverage	304.9	422.2	485.1	531.9	564.1	601.7	634.5	663.9

Other statistical methods were also applied to these data. We fitted the data for each country with a linear regression model. Linearity was satisfied for all of the countries from 2002 onwards. Figure 3 presents the results for Great Britain: the direction of the regression line and the slope values show that Online is increasing on the account of Electronic and Print.

Principal component analysis was undertaken on the regression slopes. The results are graphically presented in Figure 4; the countries and the original variables are presented on the biplot. In the direction of slope Online (b-O) there are three Pro Online countries: GB, NO and DK; in the direc-



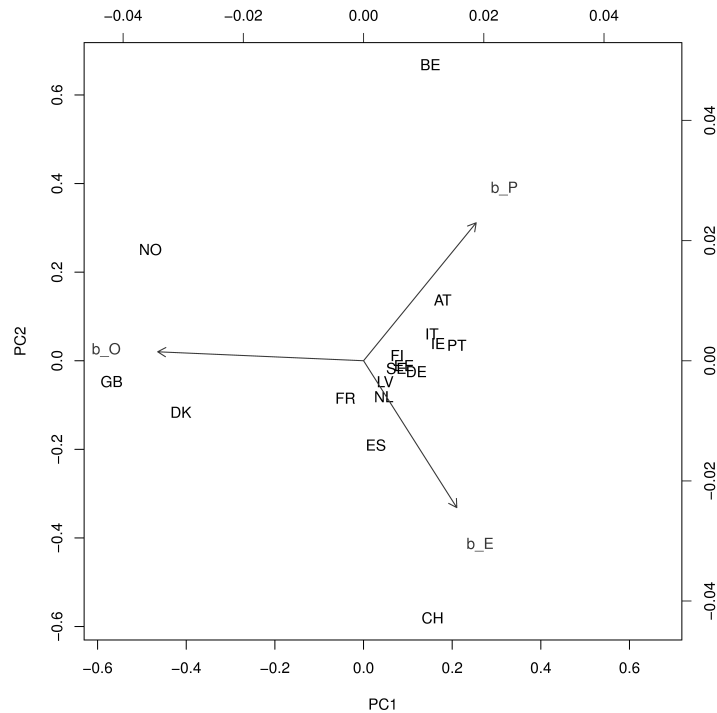
**Fig. 3.** Points for Great Britain for 2002-2008 and the linear regression line. Slope Electronic ( $b_E = -0.011$ ), slope Print ( $b_P = -0.023$ ) and slope Online ( $b_O = 0.034$ ) sum to zero.

tion of slope Electronic ( $b_E$ ) there is CH, in the direction of slope Print ( $b_P$ ) is BE. The majority of countries are clustered near the coordinate origin, revealing a stationary process.

Metric scaling results are in Figure 5. Detailed inspection of all the results allows us to interpret the "metric scaling map" in the subject-matter context in view of Hofstede's (2001) theory.

Left on the x-axis are countries which are less individualistic, more "high context cultures" (cluster: IT, PT and cluster: ES, BE, LV, FR). For these cultures, closer contacts among members are typical; their preferred mode of communication is informal, indirect, often based on symbols and pictures. Consistent with this, these are Electronic dominant countries. Right on the x-axis are more individualistic, "low-context countries", where communication is formal, explicit, often by the way of written text. These countries are Print dominant.

Below on the y-axis, countries reveal no change in time in the Online component in advertising spending. Since Online component is connected to changes in technology, these are more traditional countries ("high on uncertainty avoidance"). Above on the y-axis we follow an increase in Online component on the account of Print and Electronic, with Online around 20



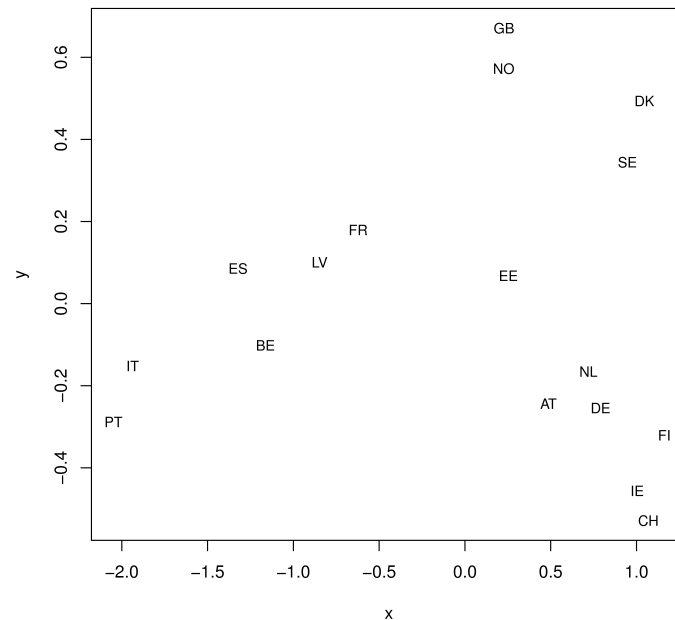
**Fig. 4.** Biplot obtained by PCA on the slopes for Electronic (b\_E), Print (b\_P) and Online (b\_O).

percent in 2008. These are more risk taking countries ("low on uncertainty avoidance"): GB, NO, DK and SE.

## 4 Conclusions

Miller (2002) presents the half-taxi metric applicable to compositional data without suggesting how it might be applied. Literature search reveals that the corresponding similarity measure was used long before, in particular in the field of phytosociology.

Compositional data analysis can be based on different geometries; for amounts in rplus and aplus geometry, for proportions in rcomp and acomp geometry. Rcomp geometry is applicable only when the absolute difference on proportions is meaningful. We believe that the half-taxi metric is preferable to other metrics in rcomp geometry because it takes into account the fact that compositions are closed to one and it has a simple geometric representation on the ternary graph.



**Fig. 5.** Metric scaling results based on half-taxi metric. Weights are defined as the number of internet users per thousand inhabitants in Europe.

In an application on advertising expenditure components (Electronic, Print and Online) for 17 European countries in the period 2001-2008 we use the half-taxi metric to detect the structural changes in time, in particular in view of the newer Online component. The results are satisfactory and can be explained in the subject-matter context in view of Hofstede's theory.

## References

- van den BOOGAART, K. G., TOLOSANA-DELGADO, R. (2008). "compositions": A unified R package to analyze compositional data. *Computers and Geosciences*, 34(4), 320-338.
- EUROMONITOR (2009): World Marketing Data and Statistics. ([www.euromonitor.com/womdas](http://www.euromonitor.com/womdas))
- HAJDU, L.J. (1981): Graphical Comparison of Resemblance Measures in Phytosociology. *Vegetatio*, v. 48, 47-59.
- HOFSTEDE, G.E. (2001): *Culture's consequences: comparing values, behaviors, institutions, and organizations across nations*, Sage, London.
- KOŠMELJ, K., ŽABKAR, V. (2008): A Methodology for Identifying Time-Trend Patterns: an Application to the Advertising Expenditure of 28 European Countries in the 1994-2004 Period. *Metodološki zvezki*, 5 (2), 161-171.
- MILLER, W. E. (2002): Revisiting the geometry of a ternary diagram with the half-taxi metric. *Mathematical Geology*, 34(3), 275-290.

# LTPD Plans by Variables when the Remainder of Rejected Lots is Inspected

J. Klufa<sup>1</sup> and L. Marek<sup>2</sup>

<sup>1</sup> University of Economics, Department of Mathematics

W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic, *klufa@vse.cz*

<sup>2</sup> University of Economics, Department of Statistics and Probability

W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic, *marek@vse.cz*

**Abstract.** In this paper we shall consider two types of the LTPD single sampling plans - for inspection by variables and for inspection by variables and attributes (all items from the sample are inspected by variables, remainder of rejected lots is inspected by attributes) - see Klufa (1994). These plans we shall compare with the corresponding Dodge-Romig LTPD plans by attributes. We shall report on an algorithm allowing the exact calculation of these plans when the non-central  $t$  distribution is used for the operating characteristic. The calculation is considerably difficult, we shall use an original method and software Mathematica. From the results of numerical investigations it follows that under the same protection of consumer the LTPD plans for inspection by variables are in many situations more economical than the corresponding Dodge-Romig attribute sampling plans.

**Keywords:** acceptance sampling, LTPD plans, inspection by variables, software Mathematica

## 1 Introduction

Under the assumption that each inspected item is classified as either good or defective (acceptance sampling by attributes) Dodge and Romig (1998) introduced sampling plans which minimize the mean number of items inspected per lot of process average quality

$$I_s = N - (N - n) \cdot L(\bar{p}; n, c) \quad (1)$$

under the condition

$$L(p_t; n, c) = 0.10 \quad (2)$$

(LTPD single sampling plans), where  $N$  is the number of items in the lot (the given parameter),  $\bar{p}$  is the process average fraction defective (the given parameter),  $p_t$  is the lot tolerance fraction defective (the given parameter,  $P_t = 100p_t$  is the lot tolerance per cent defective – denoted LTPD),  $n$  is the number of items in the sample ( $n < N$ ),  $c$  is the acceptance number (the lot is rejected when the number of defective items in the sample is greater

than  $c$ ),  $L(p)$  is the operating characteristic (the probability of accepting a submitted lot with fraction defective  $p$ ). Condition (2) protects the consumer against the acceptance of a bad lot – the probability of accepting a submitted lot of tolerance quality  $p_t$  (consumer's risk) shall be 0.10.

The corresponding LTPD plans for inspection by variables are described in Klufa (1994). In second part we shall repeat this problem, in third part we shall calculate one of these LTPD plans using software Mathematica. The aim of the paper is to find an algorithm allowing the exact calculation of these plans.

## 2 LTPD plans by variables and attributes

The problem to find LTPD plans for inspection by variables has been solved by Klufa (1994) under the following assumptions:

Measurements of a single quality characteristic  $X$  are independent, identically distributed normal random variables with unknown parameters  $\mu$  and  $\sigma^2$ . For the quality characteristic  $X$  is given either an upper specification limit  $U$  (the item is defective if its measurement exceeds  $U$ ), or a lower specification limit  $L$  (the item is defective if its measurement is smaller than  $L$ ). It is further assumed that the unknown parameter  $\sigma$  is estimated from the sample standard deviation  $s$ .

The inspection procedure is as follows (e.g. Klufa (1999)): Draw a random sample of  $n$  items and compute

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Accept the lot if

$$\frac{U - \bar{x}}{s} \geq k, \quad \text{or} \quad \frac{\bar{x} - L}{s} \geq k. \quad (3)$$

The problem is to determine the sample size  $n$  and the critical value  $k$ . There are different solutions of this problem. Klufa (1994) used for determination  $n$  and  $k$  a similar conditions as Dodge and Romig.

Now we shall formulate this problem. Let us consider *LTPD plans for inspection by variables and attributes* - all items from the sample are inspected by variables, but the remainder of rejected lots is inspected only by attributes. Let us denote

$c_s^*$  - the cost of inspection of one item by attributes,

$c_m^*$  - the cost of inspection of one item by variables.

Inspection cost per lot, assuming that the remainder of rejected lots is inspected by attributes (the inspection by variables and attributes), is  $n \cdot c_m^*$  with probability  $L(p; n, k)$ , and  $[n \cdot c_m^* + (N - n) \cdot c_s^*]$  with probability  $[1 - L(p; n, k)]$ . The mean inspection cost per lot of process average quality is



therefore

$$C_{ms} = n \cdot c_m^* + (N - n) \cdot c_s^* \cdot [1 - L(\bar{p}; n, k)]. \quad (4)$$

Now we shall look for the acceptance plan  $(n, k)$  minimizing the mean inspection cost per lot of process average quality  $C_{ms}$  under the condition

$$L(p_t; n, k) = 0.10. \quad (5)$$

The condition (5) is the same one as used for protection the consumer Dodge and Romig (1998). Let us introduce a function

$$I_{ms} = n \cdot c_m + (N - n) \cdot [1 - L(\bar{p}; n, k)], \quad (6)$$

where

$$c_m = c_m^* / c_s^*. \quad (7)$$

Since

$$C_{ms} = I_{ms} \cdot c_s^*, \quad (8)$$

both functions  $C_{ms}$  and  $I_{ms}$  have a minimum for the same acceptance plan  $(n, k)$ . Therefore, we shall look for the acceptance plan  $(n, k)$  minimizing (6) instead of (4) under the condition (5).

For these LTPD plans for inspection by variables and attributes *the new parameter*  $c_m$  was defined - see (7). This parameter must be estimated in each real situation. Usually is

$$c_m > 1. \quad (9)$$

Putting formally  $c_m = 1$  into (6) ( $I_{ms}$  in this case is denoted  $I_m$ ) we obtain

$$I_m = N - (N - n) \cdot L(\bar{p}; n, k), \quad (10)$$

i.e. the mean number of items inspected per lot of process average quality, assuming that both the sample and the remainder of rejected lots is inspected by variables. Consequently *the LTPD plans for inspection by variables* are a special case of *the LTPD plans by variables and attributes* for  $c_m = 1$ . From (10) is evident that for the determination LTPD plans by variables it is not necessary to estimate  $c_m$  ( $c_m = 1$  is not real value of this parameter).

Summary: For the given parameters  $p_t$ ,  $N$ ,  $\bar{p}$  and  $c_m$  we must determine the acceptance plan  $(n, k)$  for inspection by variables and attributes, minimizing  $I_{ms}$  in (6) under the condition (5).

In the first place we shall deal with the solution of the equation (5). The operating characteristic is (e.g. Klufa (1999))

$$L(p; n, k) = \int_{k\sqrt{n}}^{\infty} g(t; n - 1, u_{1-p}\sqrt{n}) dt, \quad (11)$$

where  $g(t; n - 1, u_{1-p}\sqrt{n})$  is probability density function of non-central  $t$  distribution with  $(n - 1)$  degrees of freedom and noncentrality parameter  $\lambda = u_{1-p}\sqrt{n}$ .

Instead of (11), using the normal distribution as an approximation of the non-central  $t$  distribution (Johnson and Welch (1940)), we have

$$L(p; n, k) = \Phi \left( \frac{u_{1-p} - k}{A} \right), \quad (12)$$

where

$$A = \sqrt{\frac{1}{n} + \frac{k^2}{2(n-1)}}. \quad (13)$$

The function  $\Phi$  in (12) is a standard normal distribution function and  $u_{1-p}$  is a quantile of order  $1-p$  (the unique root of the equation  $\Phi(u) = 1-p$ ). The approximation (12) holds both for an upper specification limit  $U$ , and for a lower specification limit  $L$ .

If we use (12) for operating characteristic, the equation  $L(p_t; n, k) = 0.10$  has one and only one solution (see Klufa (1994))

$$k = \frac{u_{1-p_t} - u_{0.10} \cdot h}{g}, \quad (14)$$

where

$$g = 1 - \frac{u_{0.10}^2}{2(n-1)}, \quad h = \sqrt{\frac{g}{n} + \frac{u_{1-p_t}^2}{2(n-1)}}. \quad (15)$$

This is an approximate solution of the equation (5). Exact solution of the equation (5) is

$$k = \frac{t_{0.9}(n-1, u_{1-p_t}\sqrt{n})}{\sqrt{n}}, \quad (16)$$

where  $t_{0.9}(n-1, u_{1-p_t}\sqrt{n})$  is a quantile of order 0.9 of non-central  $t$  distribution with  $(n-1)$  degrees of freedom and noncentrality parameter  $\lambda = u_{1-p_t}\sqrt{n}$ .

Inserting (14) or (16) into (6) we obtain a function of one variable  $n$

$$I_{ms}(n) = n \cdot c_m + (N - n) \cdot \alpha(n), \quad (17)$$

where  $\alpha(n)$  is the producer's risk<sup>1</sup> (the probability of rejecting a lot of process average quality). Now we shall look for the sample size  $n$  minimizing (17).

**Theorem 1.** (*Relation between lot size and sample size*)

Let  $\bar{p}$ ,  $p_t$  and  $c_m$  be given parameters,  $0 < \bar{p} < p_t < \frac{1}{2}$ ,  $c_m \geq 1$ . Let us denote

$$F(n) = \frac{c_m - \alpha(n+1)}{\alpha(n) - \alpha(n+1)} + n. \quad (18)$$

---

<sup>1</sup> Producer's risk is not given for these plans. Klufa (1994) proved that the producer's risk is nonincreasing function of lot size  $N$ .

If the lot size  $N > F(6)$ , then there is one and only one  $n \in \{7, 8, \dots, N - 1\}$  for which holds

$$F(n - 1) < N \leq F(n). \quad (19)$$

For this sample size  $n$  the function (17) has an absolute minimum.

*Proof:* See Klufa (1994)

*Remark.* From (19) it follows (the inverse function  $F^{-1}$  to the function  $F$  is for  $N \geq F(6)$  increasing – see Klufa (1994)) that

$$n - 1 < F^{-1}(N) \leq n, \quad (20)$$

i.e. when  $N$  increases, then  $n$  does not vary or increases (the sample size  $n$  is nondecreasing function of the lot size  $N$ ).

For the comparison of these plans with the corresponding Dodge-Romig LTPD attribute sampling plans from an economical point of view we used parameters  $E$  (inspection by variables) and  $e$  (inspection by variables and attributes), defined by relations

$$E = \frac{I_m}{I_s} \cdot 100, \quad e = \frac{I_{ms}}{I_s} \cdot 100. \quad (21)$$

The LTPD plans for inspection by variables and attributes are more economical than the corresponding Dodge-Romig plans when

$$e < 100, \quad (22)$$

similarly, if  $c_m$  is statistically estimated and the following inequality holds

$$E \cdot c_m < 100, \quad (23)$$

then the LTPD plans for inspection by variables are more economical than the corresponding Dodge-Romig LTPD plans.

It was shown that under the same protection of consumer *the LTPD plans for inspection by variables and attributes are in many situations **more economical** than the corresponding Dodge-Romig LTPD attribute sampling plans.* This conclusion is valid especially for the large lots and for the small values of the lot tolerance fraction defective – see Klufa (1999).

Similar conclusions were obtained also for the comparison of the LTPD plans for inspection by variables<sup>2</sup> with the Dodge-Romig LTPD plans.

### 3 Calculation of the LTPD plans by variables and attributes

For calculation of the LTPD plans by variables and attributes we shall use software Mathematica – see Wolfram (1991).

<sup>2</sup> The LTPD plans for inspection by variables and attributes are always more economical than the corresponding LTPD plans for inspection by variables.

*Example.* Let  $N = 450$ ,  $p_t = 0.01$ ,  $\bar{p} = 0.0015$  and  $c_m = 1.7$  (the cost of inspection of one item by variables is higher by 70% than the cost of inspection of one item by attributes). We shall look for the LTPD plan for inspection by variables and attributes. Furthermore we shall compare this plan and the corresponding Dodge-Romig LTPD plan for inspection by attributes.

Given parameters ( $N = \text{nbig}$ ,  $\bar{p} = \text{pbar}$ ):

```
In[1]:=nbig=450
In[2]:=pt=0.01
In[3]:=pbar=0.0015
In[4]:=cm=1.7
```

Approximate solution (according to (15), (14), (12) and Theorem 1):

```
In[5]:=<< Statistics`ContinuousDistributions`;
ndist = NormalDistribution[0, 1];
g[n_]:=1-(Quantile[ndist, 0.10]^2/(2n-2));
h[n_]:=Sqrt[(g[n]/n)+(Quantile[ndist,1-pt]^2/(2n-2))];
k0[n_]:= (Quantile[ndist,1-pt]-Quantile[ndist,0.10]*
h[n])/g[n];
alpha0[n_]:=CDF[ndist,(k0[n]-Quantile[ndist,1-pbar])/
Sqrt[(1/n)+(k0[n]^2/(2n-2))]];
F[n_]:=((cm-alpha0[n+1])/(alpha0[n]-alpha0[n+1])) + n;
FR:=FindRoot[F[n]==nbig,{n,7}];
n0:=n /. FR;
nAPPROX = Ceiling[n0];
```

Exact<sup>3</sup> solution (half-intervals method where we use the approximate solution)<sup>4</sup>:

```
lambda[n_,p_]:=Quantile[ndist,1-p]*Sqrt[n];
nonctdist[n_,p_]:=NoncentralStudentTDistribution[n-1,
lambda[n,p]];
k[n_]:=Quantile[nonctdist[n,pt],0.9]/Sqrt[n];
alpha[n_]:=CDF[nonctdist[n,pbar],k[n]*Sqrt[n]];
Ims[n_]:=n cm + (nbig-n)*alpha[n];
FMinSearch[nl_,nu_]:=nl /; nl==nu;
FMinSearch[nl_,nu_]:=FMinSearch[nl,nl+Floor[(nu-nl)/2]] /;
Ims[nl+Floor[(nu-nl)/2]]<=Ims[nl+Floor[(nu-nl)/2]+1];
FMinSearch[nl_,nu_]:=FMinSearch[nl+Floor[(nu-nl)/2]+1,nu];
n=FMinSearch[nAPPROX-2,nAPPROX+2]
Out[23]=67
In[24]:=k=k[%]
Out[24]=2.67084
```

<sup>3</sup> Approximate solution is  $n = 67$ ,  $k = 2.66203$  ( $n\text{APPROX}=67$ ,  $k_0[67]=2.66203$ ).

For this plan consumer's risk is only approximately 0.10.

<sup>4</sup> See (11), (16) and (17)

0.0015	0.878356	0.708472
0.0032	0.602687	0.478998
0.0049	0.386768	0.323568
0.0066	0.24508	0.218382
0.0083	0.155851	0.147261
0.01	0.1	0.0992141
0.0117	0.0648491	0.0667842
0.0134	0.0425131	0.0449145
0.0151	0.0281648	0.0301794
0.0168	0.0188455	0.0202602
0.0185	0.0127277	0.0135889
0.0202	0.00867083	0.00910601

**Fig. 1.** Out[30]: 1st column-fraction defective  $p$ ,  
2nd column-values of OC for inspection by variables and attributes  $L_1(p)$ ,  
3rd column-values of OC for inspection by attributes  $L_2(p)$

The LTPD plan for inspection by variables and attributes<sup>5</sup> is  $n = 67, k = 2.67084$ . The corresponding LTPD plan for inspection by attributes we find in a book written by Dodge and Romig (1998). For given parameters  $N, p_t$  and  $\bar{p}$  we have  $n_2 = 180, c = 0$ .

Comparison of OC (operating characteristics) of these plans (see (11) and e.g. Hald (1981) - OC for inspection by attributes):

```
In[25] := n2=180
In[26] := c=0
In[27] := L1[p_] := 1 - CDF[nonctdist[n,p], k*Sqrt[n]];
        L2[p_] := Sum[Binomial[nbig*p, i]*
        Binomial[nbig-nbig*p, n2-i]/Binomial[nbig, n2], {i,0,c}];
        Table[{p,N[L1[p],6],N[L2[p],6]}, {p,0.0015,0.0202,
        0.0017}];
        TableForm[%]
Out[30]//TableForm= see Figure 1

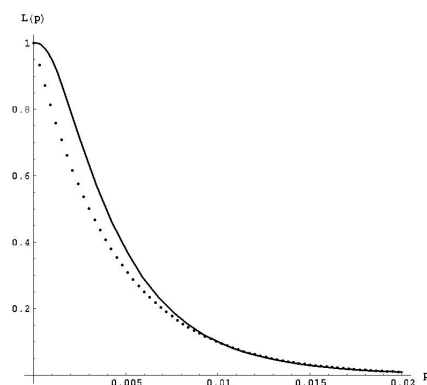
For example (see Figure 1) we get  $L_1(\bar{p}) = L_1(0.0015) = 0.878356$ , i.e. the
producer's risk6 for the LTPD plan for inspection by variables and attributes
is  $\alpha = 1 - L_1(\bar{p}) = 0.121644$ . The producer's risk for the corresponding
Dodge-Romig plan is  $\alpha = 1 - L_2(\bar{p}) = 1 - 0.708472 = 0.291528$ .

Finally graphic comparison of the operating characteristics:
In[31] := oc1 := Plot[L1[p], {p, 0, 0.02}, AspectRatio -> 0.9,
        AxesLabel->{"p", "L(p)"}, PlotStyle->Thickness[0.0045]];
        oc2:=ListPlot[Table[{p,L2[p]}, {p,0,0.02,0.0003}]];
        Show[oc1, oc2]
Out[33]= -Graphic- see Figure 2
```

Comparison from an economical point of view (see (21), (6) and (1)):

<sup>5</sup> Calculation of this plan takes about 11 minutes.

<sup>6</sup> The consumer's risk is exactly 0.10 ( $L_1(p_t) = L_1(0.01) = 0.1$ ).



**Fig. 2.** Out [33]: operating characteristic of LTPD plans

- for inspection by variables and attributes (67, 2.67084) —————
- for inspection by attributes (180, 0) .....

```
In[34] := e = 100 * (n * cm + (nbig - n) * (1 - L1[pbar])) / (nbig - (nbig - n2) * L2[pbar])
Out[34] = 62.034
```

*Conclusion.* From these results it follows that under the same protection of consumer the LTPD plan for inspection by variables and attributes (67, 2.67084) is more economical than the corresponding Dodge-Romig LTPD attribute sampling plan (180, 0). Since  $e = 62.034$  (see Out[34]), it can be expected approximately **38% saving of the inspection cost.**

Furthermore the OC curve for the LTPD plan by variables and attributes (67, 2.67084) is better than corresponding OC curve for the LTPD plan by attributes - see Out[33]. For example (see Out[30]) the producer's risk for the LTPD plan by variables and attributes  $\alpha = 0.12$  is considerably less than for the corresponding Dodge-Romig plan  $\alpha = 0.29$ .

## References

- DODGE, H. F. and ROMIG, H. G. (1998): *Sampling Inspection Tables: Single and Double Sampling*. John Wiley.
- HALD, A. (1981): *Stat Theory of Sampling Inspection by Attributes*. Academic Press.
- JOHNSON, N. L. and WELCH, B. L. (1940): Applications of the Non-central t distribution. *Biometrika* 31, 362 - 389.
- KLUFU, J. (1994): Acceptance Sampling by Variables when the Remainder of Rejected Lots is Inspected. *Statistical Papers* 35, 337 - 349.
- KLUFU, J. (1999): *Economical Aspects of Acceptance Sampling*. Ekopress.
- KLUFU, J. (2008): Dodge-Romig AOQL plans for inspection by variables from numerical point of view. *Statistical Papers* 49, 1 - 13.
- KLUFU, J. (in print): Exact calculation of the Dodge-Romig LTPD single sampling plans for inspection by variables. *Statistical Papers*
- WOLFRAM, S. (1991): *Mathematica*. Addison-Wesley.

# A Comparison between Two Computing Methods for an Empirical Variogram in Geostatistical Data

Takafumi Kubota<sup>1</sup> and Tomoyuki Tarumi<sup>2</sup>

<sup>1</sup> Okayama University, Graduate school of humanities and social sciences  
Tsushimanaka 3-1-1 Okayama, Japan, *kubota@law.okayama-u.ac.jp*

<sup>2</sup> Okayama University, Admission Centre  
Tsushimanaka 3-1-1 Okayama, Japan, *t2@ems.okayama-u.ac.jp*

**Abstract.** In this paper, we propose a new calculation method for an empirical variogram, which the range of distance of points are divided to equal number of observation pairs. Then, we do both simulation study and application for Meuse river data set (Burrough and McDonnell(1998)) in order to compare our proposal calculation method with traditional calculation method for an empirical variogram, which the range of points are divided to equal distance.

**Keywords:** geostatistics, empirical variogram, kriging, cross-validation

## 1 Introduction

Environmental data, such as data on weather, air pollution or water quality, have several parameters which provide not only their observed characteristic values but also geometric information. These data are called geostatistical data. One purpose for geostatistical data analyses is to predict characteristic values at unobserved points. The most famous prediction method is Kriging, and for this method we have to determine parameters of fitted model of variogram which measures variance of data. (This variogram is called theoretical variogram.)

To determine parameters, we can use some methods such as an ad hock method and a least square method. However, there are some problems in each method. For the ad hock method, because we determine parameters by our eyes using the graph of an empirical variogram, there is our arbitrariness. For least square method, because we estimate parameters by minimizing square error, we meet singular cases then we cannot estimate parameters. One reason of singular model in variogram fit is instability of empirical variogram, and it derives from following calculation problem.

- a. To use border observation points; an empirical variogram becomes instable in longer part of distance.
- b. To use equal distance to classify an empirical variogram. The class which consists in small number of observation pairs becomes instable.

To solve the former problem, Kubota et al. (2005) proposed the cutoff value as half of the maximum distance between every pair. In this paper we use this value for calculation of an empirical variogram. To solve the latter problem, we propose a new calculation method for an empirical variogram, which the range of distance of points are divided to equal number of observation pairs. To compare our proposal calculation method and traditional calculation method for an empirical variogram, we firstly do a simulation study. However, the latter problem frequently occurs when we calculate directional variogram. Therefore, we apply these two methods to Meuse river data set which has geometric anisotropy, thus we have to use directional variogram to collect it. In the process of application we use cross validation (leave-one-out) to predict characteristic values of observation points, and to obtain the parameters of geometric anisotropy we use the way of Kubota and Tarumi (2008); Ellipse model is fitted to points which coordinates are calculated by range values of directional variogram in all directions.

We describe two calculation methods of an empirical variogram in Section 2, we show how these methods can be applied to a simulation study in Section 3, to Meuse river data set in Section 4 and, finally, we present our concluding remarks in Section 5.

## 2 Variogram cloud and empirical variogram

### 2.1 Calculate variogram cloud and empirical variogram

To obtain the variogram cloud, we measured the difference between pairs of characteristic values  $z(\mathbf{x}_{p_1})$  and  $z(\mathbf{x}_{p_2})$  located at the points  $\mathbf{x}_{p_1}$  and  $\mathbf{x}_{p_2}$ . Dissimilarity (squared difference) is given by

$$\gamma(\mathbf{h})^* = \frac{1}{2}(z(\mathbf{x}_{p_1} + \mathbf{h}) - z(\mathbf{x}_{p_1}))^2 \quad (1)$$

where  $\mathbf{h} = \mathbf{x}_{p_1} - \mathbf{x}_{p_2}$ . Variogram cloud is a graph with plots of dissimilarity  $\gamma_l^*$  calculated by (1) versus distance  $d_l = |\mathbf{x}_{p_1} - \mathbf{x}_{p_2}|$  in all pairs of observation  $((\mathbf{x}_{p_1}, \mathbf{x}_{p_2}) \in \{1, 2, \dots, n\}, l = 1, 2, \dots, n(n+1)/2$ , where  $n$  is the number of observation). Empirical variogram calculated as follow order. We describe two calculation methods of an empirical variogram ((1);classify by distance, (2);classify by the number of pairs).

#### a. Separation

We define Intervals  $(I^{(1)}, I^{(2)})$  as

$$\begin{aligned} (1) \quad & 0 < R_1^{(1)} < R_2^{(1)} < \dots < R_K^{(1)} \\ & \text{where } R_1^{(1)} = \alpha, R_2^{(1)} = 2\alpha, \dots, R_K^{(1)} = K\alpha, \\ & I_1^{(1)} = (0, R_1], I_2^{(1)} = (R_1, R_2], \dots, I_K^{(1)} = (R_{K-1}, R_K] \end{aligned}$$



- (2)  $(d_1, \gamma_1^*), (d_2, \gamma_2^*), \dots, (d_{n(n+1)/2}, \gamma_{n(n+1)/2}^*)$  is sort to  $(d_{(1)}, \gamma_{(1)}^*), (d_{(2)}, \gamma_{(2)}^*), \dots, (d_{(n(n+1)/2)}, \gamma_{(n(n+1)/2)}^*)$ .  
 We define that  $\beta$  is the number of pairs in each class.  
 $0 < R_1^{(2)} < R_2^{(2)} < \dots < R_{K'}^{(2)}$   
 where  $I_{k'}^{(2)} = (d_{(\beta(k'-1)+1)}, d_{(\beta k')}]$ ,  $\#\{d_{(c')} \in I_{k'}^{(2)}\} = \beta$ ,  $c' = \beta(k' - 1) + 1, \dots, \beta k'$ ,  $k' = 1, \dots, K'$ .  
 It means  $\{d_{(1)}, \dots, d_{(\beta)}\} \in I_1^{(2)}$ ,  $\{d_{(\beta+1)}, \dots, d_{(2\beta)}\} \in I_2^{(2)}$ ,  $\dots$ ,  
 $\{d_{(L(\beta-1)+1)}, \dots, d_{(n(n+1)/2)}\} \in I_{K'}^{(2)}$

b. Classify

We define classes as follows

- (1)  $N_k^{(1)}$  is the class where  $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\| \in I_k^{(1)}$   
 $|N_k^{(1)}|$  is the number of element of  $N_k^{(1)}$   
 (2)  $N_k^{(2)}$  is the class where  $d_{ij} \in I_k^{(2)}$   
 $|N_k^{(2)}|$  is the number of element of  $N_k^{(2)}$

c. Average distances

We define average distances as

$$h_k^{(m)} = \frac{1}{|N_k^{(m)}|} \sum_{N_k} d_{ij} \quad (2)$$

where  $m$  corresponds to method of classify ( $m = 1, 2$ ).

d. Average dissimilarities

We define average dissimilarities as

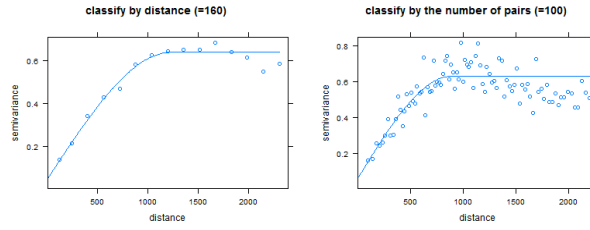
$$\hat{\gamma}^{(m)} = \frac{1}{|N_k^{(m)}|} \sum_{N_k} \tilde{\gamma}_{ij} \quad (3)$$

where  $\tilde{\gamma}_{ij} = \frac{(z(\mathbf{x}_i) - z(\mathbf{j}))^2}{2}$  and  $m$  corresponds to method of classify ( $m = 1, 2$ ).

We use the spherical model for calculating the theoretical variogram:

$$\gamma(\mathbf{h}; \xi_0, \xi_1, \xi_2) = \begin{cases} \xi_0 + \xi_1 \left( \frac{3}{2} \|\mathbf{h}\| / \xi_2 - \frac{1}{2} [\|\mathbf{h}\| / \xi_2]^3 \right), & 0 < \|\mathbf{h}\| \leq \xi_2 \\ \xi_0 + \xi_1, & \|\mathbf{h}\| > \xi_2 \\ 0, & \|\mathbf{h}\| = 0 \end{cases} \quad (4)$$

where  $\xi_0$  is the nugget effect value,  $\xi_1$  is the sill value, and  $\xi_2$  is the range value. We use least square method to estimate parameters of theoretical variogram. Figure 1 shows empirical variogram (points) and fitted theoretical variogram (solid line) in each classify method.



**Fig. 1.** Empirical variogram with theoretical variogram which is classified by distance (left) and by the number of pairs (right)

## 2.2 Kriging

Because a motivation of this study is to compare two calculation methods for an empirical variogram, we calculate prediction errors to check efficacy of calculation methods. Characteristic value  $\hat{z}_0$  at point  $\mathbf{y}_0$  is predicted by ordinary kriging

$$\hat{z}_0 = \hat{z}(\mathbf{y}_0) = \sum_{\alpha=1}^n \lambda_{\alpha}(\mathbf{y}_0) z(\mathbf{y}_{\alpha}), \quad \alpha = 1, \dots, n \quad (5)$$

$$\begin{cases} \sum_{\alpha=1}^n \lambda_{\alpha}(\mathbf{y}) C(\mathbf{y}_{\alpha} - \mathbf{y}_{\beta}) + \mu(\mathbf{y}_0) = C(\mathbf{y}_0 - \mathbf{y}_{\beta}), & \alpha = 1, \dots, n \\ \sum_{\alpha=1}^n \lambda_{\alpha}(\mathbf{y}_0) = 1 \end{cases} \quad (6)$$

where  $\mu$  is Lagrange multiplier and  $\lambda$  is weight term which is defined by estimated variogram parameters  $\gamma(\mathbf{h}) = (\xi_0, \xi_1, \xi_2)$  as follows

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}). \quad (7)$$

## 3 Simulation Study

### 3.1 Calculation

We do a simulation study in the following order to compare our proposal calculation method with traditional calculation method.

#### Step 1 Determine parameters

In this paper we use parameters as follows; mean= 0, variance= 1, spherical variogram  $(\gamma_{(0)})(\xi_0, \xi_1, \xi_2) = (0.2, 0.8, 5)$  and data area= 10x10.

#### Step 2 Simulate data

We generate a random point pattern and simulate dependence of its spatial Gaussian random field data by the parameters of Step 1.

#### Step 3 Calculate and estimate variogram

We calculate empirical variogram by traditional calculation method, and then we estimate theoretical variogram  $(\gamma_{(1)})$  by it. We also calculate by our proposal method and then we estimate theoretical variogram  $(\gamma_{(2)})$ .

**Step 4** Calculate difference

We calculate differences ( $D^{(1)}$  and  $D^{(2)}$ ) of area of theoretical variograms(given variogram and estimated variograms by two methods) as,

$$D^{(m)} = \sum_{x \in X} |\gamma_{(0)} - \gamma_{(m)}| \delta x \quad (8)$$

where  $\delta x = \text{cutoff}/100$  ( cutoff is half of the maximum distance between every pair),  $X = \{(i + \frac{1}{2})\delta x\}, i = 0, 1, \dots, 99$  and  $m = 1, 2$ .

**Step 5** Iteration

We iterate 100 times from Step 1. to Step 4 by two calculation methods changing parameter of lambda which is the number of points per unit area.

**3.2 Result**

Table 1 shows the result of simulations; left column is by distance, right 3 columns are by the number of pairs 10, 30 and 50 ."# pairs" means the number of pairs in each class of empirical variogram.

**Table 1.** the D values of results

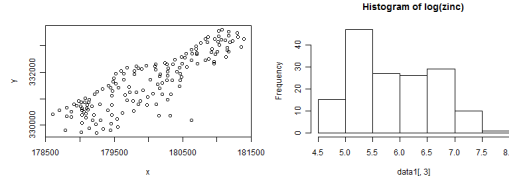
lambda	distance	# pairs=10	# pairs=30	# pairs=50
0.5	1.590	1.479	1.483	1.508
1.0	1.270	1.483	1.303	1.296
2.0	1.346	1.508	1.374	1.372

**4 Applying to Meuse river data set**

We apply two calculation methods for an empirical variogram to Meuse river data set. Then we check efficacy of these two methods. In this section, we explain data, calculations and results of these applications.

**4.1 Data**

We use natural logarithm of zinc ( $\log(\text{zinc})$ ) of Meuse river data set. Figure 2 shows observation points of Meuse river data set and histogram of characteristic value ( $\log(\text{zinc})$ ).



**Fig. 2.** Observation points(left) and histogram of  $\log(\text{zinc})$  (right)

## 4.2 Calculation

To evaluate the efficacy of two methods of calculation we use cross-validation methods, which are calculated as follows:

### Step 1 Detect and correct anisotropy

We detect and correct anisotropy that the way Kubota and Tarumi(2008) proposed from  $(n-1)$  data that  $i$ th data  $(z_i, \mathbf{x}_i)$  is removed for testing data, where  $\mathbf{x}$  is observation point and  $z$  is characteristic value ( $\log(\text{zinc})$ ). (Observation point  $\mathbf{x}$  is linear transformed to  $\mathbf{y}$ .)

### Step 2 Estimate parameters of variogram

We estimate parameters  $(\xi_0, \xi_1, \xi_2)$  by  $(n-1)$  data  $(z_k, \mathbf{y}_k)$  ( $k = 1, \dots, i-1, i+1, \dots, n$ ).

### Step 3 Kriging and prediction square error

We predict characteristic value  $\hat{z}_i$  at the point of  $\mathbf{y}_i$  by the way of section 2.2. Then we calculate prediction square error as

$$\{z_i - \hat{z}_i\}^2. \quad (9)$$

### Step 4 Calculate prediction mean square error of Cross-Validation(CV)

We repeat from Step 1 to 3 in all  $\alpha \in \{1, \dots, n\}$ . Then calculate CV as

$$CV = \frac{1}{n} \sum_{i=1}^n \{z_i - \hat{z}_i\}^2. \quad (10)$$

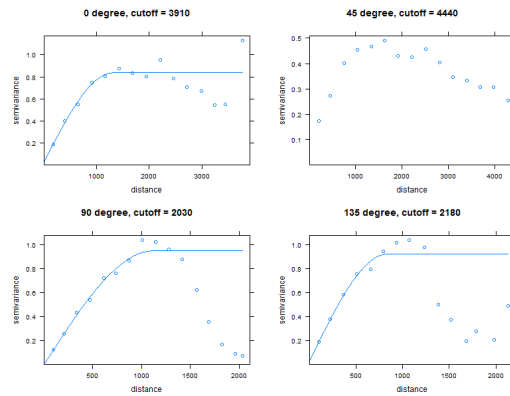
to compare its value in each case.

Figure 3 shows four directional empirical variogram with fitted theoretical variogram in the case of classify by equal distance, and Figure 4 shows corresponding fitted ellipse.

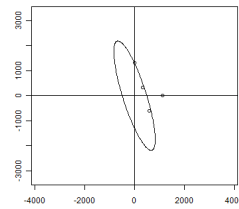
Figure 5 shows four directional empirical variogram with fitted theoretical variogram in the case of classify by the number of observation points, and Figure 6 shows corresponding fitted ellipse.

## 4.3 Result

Table 2 shows the result of cross-validation in the case of classify by the same distance (upper two lines) and the number of pairs (lower two lines). ”# class” means the number of class in each empirical variogram and ”# pairs” means the number of pairs in each class of empirical variogram.



**Fig. 3.** Four directional empirical variogram with fitted theoretical variogram (by distance); 0 degree (upper left), 45 degree (upper right), 90 degree (lower left) and 135 degree (lower right) from North direction by clockwise



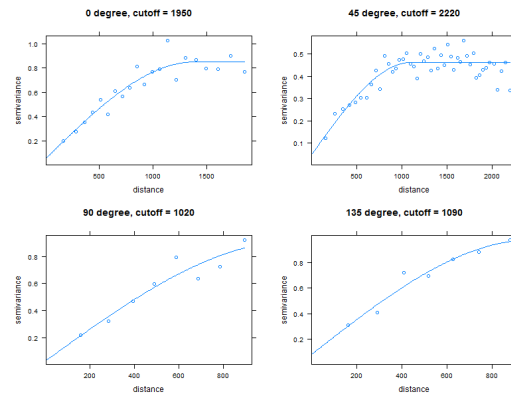
**Fig. 4.** Four range (points) and fitted ellipse (solid line) (by distance)

**Table 2.** the result of cross-validation

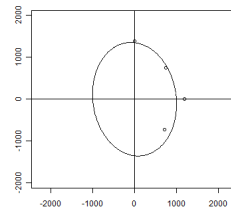
# class	20	15	10	5
CV	0.829	0.792	0.713	0.731
# pairs	10	30	50	100
CV	0.651	0.710	0.651	0.671

## 5 Concluding remarks and future studies

In this paper, we proposed a new method to calculate an empirical variogram, and compared two methods by simulation study and cross-validation using Meuse river data set. Regarding  $D^{(m)}$  in Section 3, in the case of small number of pairs ( $\lambda = 0.5$ ), our proposal method showed good result, because every  $D^{(2)}$  is smaller than  $D^{(1)}$ . On the other hand, in the cases of large number of pairs ( $\lambda = 1$  and  $2$ ) are upside down. Regarding CV criteria of cross-validation in Section 4, our proposal method showed good result, because each CV values of calculation method by the number of pairs



**Fig. 5.** Four directional empirical variogram with fitted theoretical variogram (by the number of pairs); 0 degree (upper left), 45 degree (upper right), 90 degree (lower left) and 135 degree (lower right) from North direction by clockwise



**Fig. 6.** Four range (points) and fitted ellipse (solid line) (by the number of pairs)

is smaller than method by equal distance. These results may contribute to the case of an empirical variogram which has the small number of pairs of observation points, especially to detect and correct a geometric anisotropy. However cross-validation was only applied to restricted data in this study; in future research, we will apply it to other data to check the validity of our method.

## References

- Burrough, P.A. and McDonnell, R.A. (1998): Principles of Geographical Information Systems. *Oxford University Press*.
- Kubota, T. Iizuka, M. Fueda, K. and Tarumi, T. (2005): The Selection of the Cutoff in Estimating Variogram Model. *The 5th IASC Asian Conference on Statistical Computing*. 97–100
- Kubota, T. and Tarumi, T. (2008): Using Geometric Anisotropy in Variogram Modeling. *COMPSTAT2008 Proceedings in Computational Statistics*. 793–801

# Improvement of Acceleration of the ALS Algorithm Using the Vector $\varepsilon$ Algorithm

Masahiro Kuroda<sup>1</sup>, Yuchi Mori<sup>2</sup>, Masaya Iizuka<sup>3</sup>, and Michio Sakakihara<sup>4</sup>

<sup>1</sup> Department of Socio-Information, Okayama University of Science

1-1 Ridaicho, Kita-ku, Okayama, Japan, *kuroda@soci.ous.ac.jp*

<sup>2</sup> Department of Socio-Information, Okayama University of Science

1-1 Ridaicho, Kita-ku, Okayama, Japan, *mori@soci.ous.ac.jp*

<sup>3</sup> Graduate School of Environmental Science, Okayama University

1-1-1 Tsushima-naka, Kita-ku, Okayama, Japan, *iizuka@ems.okayama-u.ac.jp*

<sup>4</sup> Department of Information Science, Okayama University of Science

1-1 Ridaicho, Kita-ku, Okayama, Japan, *sakaki@mis.ous.ac.jp*

**Abstract.** In principal components analysis dealing with qualitative data and mixed measurement levels data, the alternating least squares (ALS) algorithm is utilized. This type of algorithm may require many iterations in its application to very large data sets and thus take a long time to converge. Kuroda et al. (2008) proposed an iterative algorithm for accelerating the convergence of the ALS algorithm using the vector  $\varepsilon$  algorithm of Wynn (1962). In this paper, we derive a new algorithm which does not modify the original acceleration algorithm but includes additional restarting criteria for reducing both the number of iterations and the computational time.

**Keywords:** principal components analysis, alternating least squares algorithm, vector  $\varepsilon$  algorithm, restarting criteria, acceleration

## 1 Introduction

Principal components analysis (PCA) is a popular descriptive multivariate method for handling quantitative data and is extended to deal with qualitative data and mixed measurement levels data. The existing algorithms for extended PCA are PRINCIPALS of Young et al. (1978) and PRINCALS of Gifi (1989) in which the alternating least squares (ALS) algorithm is utilized. Both algorithms alternate between optimal scaling for quantifying qualitative data and the analysis of the optimal scaled data using the ordinary PCA approach. We will refer to PRINCIPALS and PRINCALS as PCA.ALS when not distinguishing between them.

In application of extended PCA for very large data sets and variable selection problems, many iterations and much computational time may be required for convergence of PCA.ALS. For example, for PCA based on a subset of variables for qualitative data, the PRINCIPALS approach taken by Mori et al. (1997) obtained estimates only after a large number of iterations.

Kuroda et al. (2008) proposed an iterative algorithm for speeding up the convergence of PCA.ALS using the vector  $\varepsilon$  algorithm of Wynn (1962) that enables the acceleration of convergence of a slowly convergent vector sequence and is very effective for linearly converging sequences. In this paper, we derive a new version of the vector  $\varepsilon$  accelerated PCA.ALS ( $v\varepsilon$ -PCA.ALS) algorithm which does not modify the original acceleration algorithm but includes additional restarting criteria for reducing both the number of iterations and computational time.

The paper is organized as follows. We briefly describe extended PCA for a mixture of quantitative and qualitative data in Section 2, and show PRINCIPALS and PRINCALS for finding least squares estimates of the model and optimal scaling parameters in Section 3. In Section 4, we present the procedure of  $v\varepsilon$ -PCA.ALS and propose the new version of the algorithm with restarting criteria in Section 5. Numerical experiments in Section 6 examine the performance of the  $v\varepsilon$  acceleration algorithms for PRINCIPALS with/without restarting criteria in terms of the number of iterations required for convergence.

## 2 Principal components analysis with variables measured with a variety of scaled levels

PCA transforms linearly an original data set of variables into a substantially smaller set of uncorrelated variables that contains much of the information in the original data set. The original data matrix is then replaced by an estimate constructed by forming the product of matrices of component scores and eigenvectors.

Let  $\mathbf{X} = (\mathbf{X}_1 \mathbf{X}_2 \cdots \mathbf{X}_p)$  be an  $n \times p$  matrix of  $n$  observations on  $p$  variables and be columnwise standardized. In PCA, we postulate that  $\mathbf{X}$  is approximated by the following bilinear form:

$$\hat{\mathbf{X}} = \mathbf{Z}\mathbf{A}^\top, \quad (1)$$

where  $\mathbf{Z} = (\mathbf{Z}_1 \mathbf{Z}_2 \cdots \mathbf{Z}_r)$  is an  $n \times r$  matrix of  $n$  component scores on  $r$  ( $1 \leq r \leq p$ ) components, and  $\mathbf{A} = (\mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_r)$  is a  $p \times r$  matrix consisting of the eigenvectors of  $\mathbf{X}^\top \mathbf{X}/n$  and  $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_r$ . Then we determine model parameters  $\mathbf{Z}$  and  $\mathbf{A}$  such that

$$\theta = \text{tr}(\mathbf{X} - \hat{\mathbf{X}})^\top (\mathbf{X} - \hat{\mathbf{X}}) = \text{tr}(\mathbf{X} - \mathbf{Z}\mathbf{A}^\top)^\top (\mathbf{X} - \mathbf{Z}\mathbf{A}^\top) \quad (2)$$

is minimized for the prescribed  $r$  components.

Ordinary PCA assumes that all variables are measured with interval and ratio scales and can be applied only to quantitative data. When the observed data contain several different types of variables with nominal, ordinal, interval and ratio scales, ordinary PCA can not be directly applied to such data. In



such situations, optimal scaling is used to quantify the observed qualitative data and then ordinary PCA can be applied to the optimal scaled data.

To quantify  $\mathbf{X}_j$  of qualitative variable  $j$  with  $K_j$  categories, the vector is coded by using an  $n \times K_j$  indicator matrix  $\mathbf{G}_j$  with entries  $g_{(j)ik} = 1$  if object  $i$  belongs to category  $k$ , and  $g_{(j)ik'} = 0$  if object  $i$  belongs to some other category  $k' (\neq k)$ ,  $i = 1, \dots, n$  and  $k = 1, \dots, K_j$ . Then the optimally scaled vector  $\mathbf{X}_j^*$  of  $\mathbf{X}_j$  is given by  $\mathbf{X}_j^* = \mathbf{G}_j \alpha_j$ , where  $\alpha_j$  is a score vector for categories of  $\mathbf{X}_j$ . Let  $\mathbf{X}^*$  be an  $n \times p$  matrix of optimally scaled observations to satisfy restrictions

$$\mathbf{X}^{*\top} \mathbf{1}_n = \mathbf{0}_p \quad \text{and} \quad \text{diag} \left[ \frac{\mathbf{X}^{*\top} \mathbf{X}^*}{n} \right] = \mathbf{I}_p, \quad (3)$$

where  $\mathbf{1}_n$  and  $\mathbf{0}_p$  are vectors of ones and zeros of length  $n$  and  $p$  respectively. In the presence of nominal and/or ordinal variables, the optimal criterion (2) is replaced by

$$\theta^* = \text{tr}(\mathbf{X}^* - \hat{\mathbf{X}})^\top (\mathbf{X}^* - \hat{\mathbf{X}}) = \text{tr}(\mathbf{X}^* - \mathbf{Z}\mathbf{A}^\top)^\top (\mathbf{X}^* - \mathbf{Z}\mathbf{A}^\top). \quad (4)$$

To apply PCA to data with mixed measurement levels, we determine the optimal scaling parameter  $\mathbf{X}^*$ , in addition to estimating  $\mathbf{Z}$  and  $\mathbf{A}$ .

### 3 Alternating least squares algorithms for principal components analysis

#### 3.1 PRINCIPALS

PRINCIPALS proposed by Young et al. (1978) is a method for utilizing the ALS algorithm for PCA of data with mixed measurement levels of single discrete and continuous, single nominal, ordinal and numerical variables. PRINCIPALS alternates between ordinary PCA and optimal scaling, and minimizes  $\theta^*$  defined by Equation (4) under the restriction (3). Then  $\theta^*$  is to be determined by model parameters  $\mathbf{Z}$  and  $\mathbf{A}$  and optimal scaling parameter  $\mathbf{X}^*$ , by updating each of the parameters in turn, keeping the others fixed.

For the initialization of PRINCIPALS, we determine initial data  $\mathbf{X}^{*(0)}$ . The observed data  $\mathbf{X}$  may be used as  $\mathbf{X}^{*(0)}$  after it is standardized to satisfy the restriction (3). For given initial data  $\mathbf{X}^{*(0)}$  with the restriction (3), PRINCIPALS iterates the following two steps:

- *Model parameter estimation step:* Obtain  $\mathbf{A}^{(t)}$  by solving

$$\left[ \frac{\mathbf{X}^{*(t)\top} \mathbf{X}^{*(t)}}{n} \right] \mathbf{A} = \mathbf{A} \mathbf{D}_r, \quad (5)$$

where  $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_r$  and  $\mathbf{D}_r$  is an  $r \times r$  diagonal eigenvalue matrix, and the superscript  $(t)$  indicates the  $t$ -th iteration. Compute  $\mathbf{Z}^{(t)}$  from  $\mathbf{Z}^{(t)} = \mathbf{X}^{*(t)} \mathbf{A}^{(t)}$ .

- *Optimal scaling step*: Calculate  $\hat{\mathbf{X}}^{(t+1)} = \mathbf{Z}^{(t)} \mathbf{A}^{(t)\top}$  from Equation (1). Find  $\mathbf{X}^{*(t+1)}$  such that

$$\mathbf{X}^{*(t+1)} = \arg \min_{\mathbf{X}^*} \text{tr}(\mathbf{X}^* - \hat{\mathbf{X}}^{(t+1)})^\top (\mathbf{X}^* - \hat{\mathbf{X}}^{(t+1)})$$

for fixed  $\hat{\mathbf{X}}^{(t+1)}$  under measurement restrictions on each variables. Scale  $\mathbf{X}^{*(t+1)}$  by columnwise normalizing and centering.

### 3.2 PRINCALS

PRINCALS of Gifi (1989) can handle multiple nominal variables in addition to single nominal, ordinal and numerical variables accepted in PRINCIPALS. We denote the set of multiple variables by  $\mathcal{J}_M$  and the set of single variables with single nominal and ordinal scales and numerical measurements by  $\mathcal{J}_S$ . For  $\mathbf{X}$  consisting of a mixture of multiple and single variables, the algorithm finds  $\mathbf{Z}$ ,  $\mathbf{A}$  and  $\mathbf{X}^*$  by alternating between ordinary PCA and optimal scaling subject to minimizing

$$\theta^* = \text{tr}(\mathbf{Z} - \mathbf{X}^* \mathbf{A})^\top (\mathbf{Z} - \mathbf{X}^* \mathbf{A}) \quad (6)$$

under the restriction

$$\mathbf{Z}^\top \mathbf{1}_n = \mathbf{0}_r \quad \text{and} \quad \mathbf{Z}^\top \mathbf{Z} = n \mathbf{I}_p. \quad (7)$$

For the initialization of PRINCALS, we determine initial data  $\mathbf{Z}^{(0)}$ ,  $\mathbf{A}^{(0)}$  and  $\mathbf{X}^{*(0)}$ . The values of  $\mathbf{Z}^{(0)}$  are initialized with random numbers under the restriction (7). For multiple variable  $j$ , the initial value of  $\mathbf{X}_j^*$  is obtained as  $\mathbf{X}_j^{*(0)} = \mathbf{G}_j (\mathbf{G}_j^\top \mathbf{G}_j)^{-1} \mathbf{G}_j^\top \mathbf{Z}^{(0)}$ . For single variables measured by single nominal and ordinal scales,  $\mathbf{X}_j^{*(0)}$  is defined as the first  $K_j$  successive integers under the normalization restriction, and the initial value of  $\mathbf{A}_j$  is calculated as the vector  $\mathbf{A}_j^{(0)} = \mathbf{Z}^{(0)\top} \mathbf{X}_j^{*(0)}$ . For given these initial values, PRINCALS provided in Michailidis and de Leeuw (1998) iterates the following two steps:

- *Model parameter estimation step*: Calculate  $\mathbf{Z}^{(t+1)}$  by

$$\mathbf{Z}^{(t+1)} = p^{-1} \left( \sum_{j \in \mathcal{J}_M} \mathbf{X}_j^{*(t)} + \sum_{j \in \mathcal{J}_S} \mathbf{X}_j^{*(t)} \mathbf{A}_j^{(t)} \right)$$

Columnwise center and orthonormalize  $\mathbf{Z}^{(t+1)}$ . Estimate  $\mathbf{A}_j^{(t+1)}$  for single variable  $j$  by  $\mathbf{A}_j^{(t+1)} = \mathbf{Z}^{(t+1)\top} \mathbf{X}_j^{*(t)} / \mathbf{X}_j^{*(t)\top} \mathbf{X}_j^{*(t)}$ .

- *Optimal scaling step*: Estimate the optimally scaled vector  $\mathbf{X}^{*(t)}$  by for  $j \in \mathcal{J}_M$  by

$$\mathbf{X}_j^{*(t+1)} = \mathbf{G}_j (\mathbf{G}_j^\top \mathbf{G}_j)^{-1} \mathbf{G}_j^\top \mathbf{Z}^{(t+1)}$$

and for  $j \in \mathcal{J}_S$  by

$$\mathbf{X}_j^{*(t+1)} = \mathbf{G}_j (\mathbf{G}_j^\top \mathbf{G}_j)^{-1} \mathbf{G}_j^\top \mathbf{Z}^{(t+1)} \mathbf{A}_j^{(t+1)} / \mathbf{A}_j^{(t+1)\top} \mathbf{A}_j^{(t+1)}$$

under measurement restrictions on each variables.

#### 4 The $v\varepsilon$ acceleration of the ALS algorithm

We briefly introduce the  $v\varepsilon$  algorithm of Wynn (1962) used in the acceleration of PCA.ALS. The  $v\varepsilon$  algorithm is utilized to speed up the convergence of a slowly convergent vector sequence and is very effective for linearly converging sequences. Kuroda and Sakakihara (2006) proposed the  $\varepsilon$ -accelerated EM algorithm that speeds up the convergence of the EM sequence via the  $v\varepsilon$  algorithm and demonstrated that its speed of convergence is significantly faster than that of the EM algorithm. Wang et al. (2008) studied the convergence properties of the  $\varepsilon$ -accelerated EM algorithm.

Let  $\{\mathbf{Y}^{(t)}\}_{t \geq 0}$  be a linear convergent sequence generated by an iterative computational procedure and let  $\{\dot{\mathbf{Y}}^{(t)}\}_{t \geq 0}$  be the accelerated sequence of  $\{\mathbf{Y}^{*(t)}\}_{t \geq 0}$ . Then the  $v\varepsilon$  algorithm generates  $\{\dot{\mathbf{Y}}^{(t)}\}_{t \geq 0}$  by using

$$\dot{\mathbf{Y}}^{(t-1)} = \mathbf{Y}^{(t)} + \left[ \left[ (\mathbf{Y}^{(t-1)} - \mathbf{Y}^{(t)}) \right]^{-1} + \left[ (\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}) \right]^{-1} \right]^{-1}, \quad (8)$$

where  $[\mathbf{Y}]^{-1} = \mathbf{Y} / \|\mathbf{Y}\|^2$  and  $\|\mathbf{Y}\|$  is the Euclidean norm of  $\mathbf{Y}$ . For the detailed derivation of Equation (8), see Appendix. When  $\{\mathbf{Y}^{(t)}\}_{t \geq 0}$  converges to a stationary point  $\mathbf{Y}^{(\infty)}$  of  $\{\mathbf{Y}^{(t)}\}_{t \geq 0}$ , it is known that, in many cases,  $\{\dot{\mathbf{Y}}^{(t)}\}_{t \geq 0}$  generated by the  $v\varepsilon$  algorithm converges to  $\mathbf{Y}^{(\infty)}$  faster than  $\{\mathbf{Y}^{(t)}\}_{t \geq 0}$ .

We assume that  $\{\mathbf{X}^{*(t)}\}_{t \geq 0}$  generated by PCA.ALS converges to a limit point  $\mathbf{X}^{*(\infty)}$ . Then  $v\varepsilon$ -PCA.ALS produces a faster convergent sequence  $\{\dot{\mathbf{X}}^{*(t)}\}_{t \geq 0}$  of  $\{\mathbf{X}^{*(t)}\}_{t \geq 0}$  by using the  $v\varepsilon$  algorithm and enables the acceleration of convergence of PCA.ALS. The general procedure of  $v\varepsilon$ -PCA.ALS iterates the following two steps:

- *PCA.ALS step*: Compute model parameters  $\mathbf{A}^{(t)}$  and  $\mathbf{Z}^{(t)}$  and determine optimal scaling parameter  $\mathbf{X}^{*(t+1)}$ .
- *Acceleration step*: Calculate  $\dot{\mathbf{X}}^{*(t-1)}$  using  $\{\mathbf{X}^{*(t-1)}, \mathbf{X}^{*(t)}, \mathbf{X}^{*(t+1)}\}$  from the  $v\varepsilon$  algorithm:

$$\begin{aligned} \text{vec} \dot{\mathbf{X}}^{*(t-1)} &= \text{vec} \mathbf{X}^{*(t)} \\ &+ \left[ \left[ \text{vec}(\mathbf{X}^{*(t-1)} - \mathbf{X}^{*(t)}) \right]^{-1} + \left[ \text{vec}(\mathbf{X}^{*(t+1)} - \mathbf{X}^{*(t)}) \right]^{-1} \right]^{-1}, \end{aligned} \quad (9)$$

where  $\text{vec} \mathbf{X}^* = (\mathbf{X}_1^{*\top} \mathbf{X}_2^{*\top} \dots \mathbf{X}_p^{*\top})^\top$  stands for the vectors of columns of  $\mathbf{X}^*$ , and check the convergence by  $\|\text{vec}(\dot{\mathbf{X}}^{*(t-1)} - \dot{\mathbf{X}}^{*(t-2)})\|^2 < \delta$ , where  $\delta$  is a desired accuracy.

$v\varepsilon$ -PCA.ALS is designed to generate  $\{\dot{\mathbf{X}}^{*(t)}\}_{t \geq 0}$  converging to  $\mathbf{X}^{*(\infty)}$ . Thus the estimate of  $\mathbf{X}^*$  can be obtained from the final value of  $\{\dot{\mathbf{X}}^{*(t)}\}_{t \geq 0}$  when  $v\varepsilon$ -PCA.ALS terminates. The estimates of  $\mathbf{Z}$  and  $\mathbf{A}$  can then be calculated immediately from the estimate of  $\mathbf{X}^*$  in the *Model parameter estimation step* of PCA.ALS.

Note that  $\dot{\mathbf{X}}^{*(t-1)}$  obtained at the  $t$ -th iteration of the *Acceleration step* is not used as the estimate  $\mathbf{X}^{*(t+1)}$  at the  $(t+1)$ -th iteration of the *PCA.ALS step*. Thus  $v\varepsilon$ -PCA.ALS speeds up the convergence of  $\{\mathbf{X}^{*(t)}\}_{t \geq 0}$  without affecting the convergence properties of the ordinary PCA.ALS procedure.

## 5 Improvement of the $v\varepsilon$ accelerated ALS algorithm by using a restarting strategy

We derive  $v\varepsilon$ -PCA.ALS with a restarting strategy. During the computation of  $v\varepsilon$ -PCA.ALS, the *PCA.ALS* and *Acceleration steps* are alternated until attaining convergence. However, it may not be needed to calculate  $\dot{\mathbf{X}}^{*(t)}$  in the *Acceleration step* within the first several iterations. We present  $v\varepsilon$ -PCA.ALS with the restarting strategy, such that it continues PCA.ALS iterations till achieving restarting criteria and starts  $v\varepsilon$ -PCA.ALS by using a new initial value of  $\mathbf{X}^*$ . Thus we decide the starting point of iteration of the *Acceleration step* and give the new initial value of  $\mathbf{X}^*$ .

Given an initial value  $\mathbf{X}^{*(0)}$ , we continue taking PCA.ALS as long as  $|\theta^{*(t+1)} - \theta^{*(t)}|$  is greater than restarting criteria  $\delta_0$ . When this condition is violated, we compute a new initial value of  $\mathbf{X}^*$  from Equation (9) and start  $v\varepsilon$ -PCA.ALS. We provide the new acceleration algorithm:

- *Single PCA.ALS step*: Repeat the following computation till  $|\theta^{*(t+1)} - \theta^{*(t)}| < \delta_0$ .
  - Estimate model parameters  $\mathbf{A}^{(t)}$  and  $\mathbf{Z}^{(t)}$  and determine optimal scaling parameter  $\mathbf{X}^{*(t+1)}$ . Calculate  $\theta^{*(t+1)}$ .
- *New initial value computation*: Compute  $\dot{\mathbf{X}}^{*(T-2)}$  from Equation (9) using  $\{\mathbf{X}^{*(T-2)}, \mathbf{X}^{*(T-1)}, \mathbf{X}^{*(T)}\}$  and set  $\mathbf{X}^{*(T+0)} = \dot{\mathbf{X}}^{*(T-2)}$ , where  $T$  is the number of iterations of the *Single PCA.ALS step*.
- *$v\varepsilon$ -PCA.ALS step*: Set  $t = 0$ . Alternate the following two steps by using  $\mathbf{X}^{*(T+t)}$  as the starting value.
  - Obtain  $\mathbf{X}^{*(T+t+1)}$  from the *PCA.ALS step*.
  - Compute  $\dot{\mathbf{X}}^{*(T+t-1)}$  using  $\{\mathbf{X}^{*(T+t-1)}, \mathbf{X}^{*(T+t)}, \mathbf{X}^{*(T+t+1)}\}$  in the *Acceleration step* and check the convergence by  $\|\text{vec}(\dot{\mathbf{X}}^{*(T+t-1)} - \dot{\mathbf{X}}^{*(T+t-2)})\|^2 < \delta$ .

## 6 Numerical experiments

We use data obtained in teacher evaluation by students. The data are obtained from 56 students and consisted of 13 categorical variables with 5 levels each; the lowest evaluation level is 1 and the highest 5.

In this experiments, we examine the performance of the  $v\varepsilon$  acceleration algorithm with/without restarting criteria for PRINCIPALS in terms of the number of iterations required for convergence. We denote  $v\varepsilon$ -PRINCIPALS

**Table 1.** The numbers of iterations and CPU times of PRINCIPALS,  $v\varepsilon$ -PRINCIPALS and  $r$ - $v\varepsilon$ -PRINCIPALS

$r$	PRINCIPALS		$v\varepsilon$ -PRINCIPALS		$r$ - $v\varepsilon$ -PRINCIPALS	
	Iter.	Time	Iter.	Time	Iter.	Time
1	9	0.25	4	0.167	2 (4)	0.222
2	92	2.52	23	0.704	9 (6)	0.469
3	28	0.59	9	0.231	4 (3)	0.194
4	25	0.74	7	0.276	3 (5)	0.210
5	28	0.58	10	0.248	5 (3)	0.207
6	29	0.61	9	0.251	4 (4)	0.210
7	28	0.79	9	0.330	3 (4)	0.254
8	47	1.07	14	0.373	7 (5)	0.324
9	45	1.30	13	0.433	6 (5)	0.380
10	45	0.88	14	0.323	7 (5)	0.279
11	33	0.65	10	0.236	5 (3)	0.200
12	40	1.11	10	0.333	6 (3)	0.309

with restarting criteria as  $r$ - $v\varepsilon$ -PRINCIPALS. All computations are performed with the statistical package R executing on a Core Duo 1.5GHz computer with 1 GB of memory.

Table 1 presents the numbers of iterations and CPU times of ordinary PRINCIPALS and two acceleration algorithms for  $\delta = 10^{-8}$ . CPU times taken (in second) are typically available to 10 msec by the function `proc.time`. Each value in the parenthesis of the sixth column is the number of iterations of the *Single PRINCIPALS step* under the restarting criteria  $\delta_0 = 1$ . Both two accelerated algorithms converge 3 to 4 times faster than PRINCIPALS. Thus the new algorithm has the same performance of  $v\varepsilon$ -PRINCIPAL in terms of the numbers of iterations. The computational times of  $r$ - $v\varepsilon$ -PRINCIPALS are shorter than those of  $v\varepsilon$ -PRINCIPALS except  $r = 1$ . We can see that the restating strategy works to reduce the computational time of  $v\varepsilon$ -PRINCIPALS.

In the experiments, the value of  $\delta_0$  was decided roughly and thus it may not be optimal. When requiring the larger number of iterations for analysis of large data sets, it is a serious problem to find a optimal value of  $\delta_0$ . We intend to deduce criteria for  $\delta_0$  systematically but not ad hoc.

## Acknowledgement

The authors would like to thank the editor and two referees whose valuable comments and kindly suggestions that led to an improvement of this paper. This research is supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (C), No 20500263.

## Appendix: The $v\varepsilon$ algorithm

Let  $\mathbf{Y}^{(t)}$  denote a vector of dimension  $d$  that converges to a vector  $\mathbf{Y}^{(\infty)}$  as  $t \rightarrow \infty$ . Let the inverse  $[\mathbf{Y}]^{-1}$  of a vector  $\mathbf{Y}$  be defined by  $[\mathbf{Y}]^{-1} = \mathbf{Y}/\|\mathbf{Y}\|^2$ , where  $\|\mathbf{Y}\|$  is the Euclidean norm of  $\mathbf{Y}$ .

In general, the  $v\varepsilon$  algorithm for a sequence  $\{\mathbf{Y}^{(t)}\}_{t \geq 0}$  starts with

$$\varepsilon^{(t,-1)} = 0, \quad \varepsilon^{(t,0)} = \mathbf{Y}^{(t)},$$

and then generates a vector  $\varepsilon^{(t,k+1)}$  by

$$\varepsilon^{(t,k+1)} = \varepsilon^{(t+1,k-1)} + \left[ \varepsilon^{(t+1,k)} - \varepsilon^{(t,k)} \right]^{-1}, \quad k = 0, 1, 2, \dots \quad (10)$$

For practical implementation, we apply the  $v\varepsilon$  algorithm for  $k = 1$  to accelerate the convergence of  $\{\mathbf{Y}^{(t)}\}_{t \geq 0}$ . Then the vector  $\varepsilon^{(t,2)}$  from Equation (10) becomes as follows:

$$\begin{aligned} \varepsilon^{(t,2)} &= \varepsilon^{(t+1,0)} + \left[ \left[ \varepsilon^{(t,0)} - \varepsilon^{(t+1,0)} \right]^{-1} + \left[ \varepsilon^{(t+2,0)} - \varepsilon^{(t+1,0)} \right]^{-1} \right]^{-1} \\ &= \mathbf{Y}^{(t+1)} + \left[ \left[ \mathbf{Y}^{(t)} - \mathbf{Y}^{(t+1)} \right]^{-1} + \left[ \mathbf{Y}^{(t+2)} - \mathbf{Y}^{(t+1)} \right]^{-1} \right]^{-1}. \end{aligned}$$

## References

- GIFI, A. (1989): Algorithm descriptions for ANACOR, HOMALS, PRINCIPALS, and OVERALS. *Report RR 89-01. Leiden: Department of Data Theory, University of Leiden.*
- KURODA, M. and SAKAKIHARA, M. (2006): Accelerating the convergence of the EM algorithm using the vector epsilon algorithm. *Computational Statistics and Data Analysis* 51, 1549-1561.
- KURODA, M., MORI, Y., IIZUKA, M. and SAKAKIHARA, M. (2008): Acceleration of convergence of the alternating least squares algorithm for principal component analysis. *Program & Abstracts IASC 2008*, 172-172.
- MICHAILIDIS, G. and DE LEEUW, J. (1998): The Gifi system of descriptive multivariate analysis. *Statistical Science* 13, 307-336.
- MORI, Y., TANAKA, Y. and TARUMI, T. (1997): Principal component analysis based on a subset of variables for qualitative data. In: C. Hayashi, K. Yajima, H. Bock, N. Ohsumi, Y. Tanaka, Y. Baba (Eds.): *Data Science, Classification, and Related Methods (Proceedings of IFCS-96)*. Springer-Verlag, 547-554.
- YOUNG, F.W., TAKANE, Y., and DE LEEUW, J. (1978): Principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika* 43, 279-281.
- WANG, M., KURODA, M., SAKAKIHARA, M. and GENG, Z. (2008): Acceleration of the EM algorithm using the vector epsilon algorithm. *Computational Statistics* 23, 469-486.
- WYNN, P. (1962): Acceleration techniques for iterated vector and matrix problems. *Mathematics of Computation* 16, 301-322.

# Unsupervised Recall and Precision Measures: a Step towards New Efficient Clustering Quality Indexes

Jean-Charles Lamirel<sup>1</sup>, Maha Ghribi<sup>2</sup>, and Pascal Cuxac<sup>2</sup>

<sup>1</sup> LORIA - Campus Scientifique BP 239  
54506 Vandœuvre-lès-Nancy, France, *lamirel@loria.fr*

<sup>2</sup> INIST-CNRS  
2 allée du Parc de Brabois, 54500-Vandœuvre-lès-Nancy, France,  
*maha.ghribi@inist.fr*, *pascal.cuxac@inist.fr*

**Abstract.** Traditional quality indexes do not allow to properly estimate the quality of the unsupervised classification results in several cases, as in that one of the textual data. We thus present an alternative approach for clustering quality evaluation based on unsupervised measures of Recall, Precision and F-measure exploiting the descriptors of the data associated with the obtained clusters. The experimental comparison of the behavior of the classical indexes with our new approach on a dataset of bibliographical references issued from the PASCAL database clearly highlights that our method is the only one that can distinguish between homogeneous and heterogeneous clustering results.

**Keywords:** clustering, quality indexes, text mining, heterogeneous data

## 1 Introduction

The use of the methods of classification of information became current to analyze large corpus of data as it is the case in the domain of scientific survey or in that of strategic analyses of research. While carrying out a classification, the aim is to build homogeneous groups of data sharing a certain number of identical characteristics. Furthermore, the clustering, or unsupervised classification, makes it possible to highlight these groups without prior knowledge on the processed data. If those data are scientific publications and one regards the starting corpus as representative of a research field, the obtained clusters can be viewed as research topics related to this field. A central problem that then arises is to qualify the obtained results in terms of quality: a quality index is a criterion which indeed makes it possible all together to decide which clustering method to use, to fix an optimal number of clusters, and to evaluate or to develop a new method. The classical indexes used for the evaluation of the quality of clustering are mainly distance-based indexes relying on the concepts of intra cluster inertia and inter-cluster inertia (Lebart et al. (1982)).

- Intra-cluster inertia measures the degree of homogeneity between the data associated with a cluster. It calculates their distances relatively to the point representing the profile of the cluster.
- Inter-clusters inertia measures the degree of heterogeneity between the clusters. It calculates the distances between the points representing the profiles of the various clusters of the partition.

Thanks to these two quality indexes or to their adaptations, like the Dunn index (Dunn (1974)), the Davies Bouldin index (Davies and Bouldin (2000)), or the Silhouettes index (Rousseeuw (1987)), a clustering result is considered as good if it possesses low intra-clusters distances as compared to its inter-clusters distances. As it has been shown in (Lamirel and Al Shehaby (2004)), the distance-based indexes are often strongly biased<sup>1</sup> and highly dependent on the clustering method. Thus, they cannot be easily used for comparing different methods. Moreover, as it has been also shown in (Kassab and Lamirel (2008)), they are often properly unable to identify an optimal clustering model whenever the dataset is constituted by complex data that must be represented in a both highly multidimensional and sparse description space, as it is often the case with textual data. To cope with such problems, our own approach takes its inspiration both from the behavior of symbolic classifiers and from the evaluation principles used in Information Retrieval.

Our Recall/Precision and F-measures indexes exploit the properties of the data associated to each cluster after the clustering process without prior consideration of clusters profiles (Lamirel and al. (2004)). Their main advantage is thus to be independent of the clustering methods and of their operating mode. However, our last experiments highlighted that these new quality indexes did not make it possible to clearly distinguish between homogeneous results of clustering and heterogeneous ones (Ghribi and al. (2010)). After presenting our original quality indexes, we thus present hereafter some of their extensions which make it possible to solve the said problem. We then experimentally show the effectiveness of our extended approach, as compared to a classical distance-based approach, for discriminating between the results provided by two different clustering methods which have been applied on a documentary corpus containing multi-topics bibliographic records issued from the PASCAL CNRS scientific database.

## 2 Unsupervised recall precision f-measure indexes

In IR, the **Recall R** represents the ratio between the number of relevant documents which have been returned by an IR system for a given query and

---

<sup>1</sup> A bias can occur when the intrinsic dimensions of the obtained clusters (number of non-zero components in the reference vectors describing the clusters) are not of the same order of magnitude than the intrinsic dimensions of the data profiles (see (Lamirel and Al Shehaby (2004)) for more details).



the total number of relevant documents which should have been found in the documentary database. The **Precision P** represents the ratio between the number of relevant documents which have been returned by an IR system for a given query and the total number of documents returned for the said query. **Recall** and **Precision** generally behave in an antagonist way: as **Recall** increases, **Precision** decreases, and conversely. The **F** function has thus been proposed in order to highlight the best compromise between these two values. It is given by:

$$F = \frac{2(R * P)}{R + P} \quad (1)$$

Based on the same principles, the *Recall* and *Precision* indexes which we introduce hereafter evaluate the quality of a clustering method in an unsupervised way<sup>2</sup> by measuring the relevance of the clusters content in terms of shared properties. In our further descriptions, a cluster content is supposed to be represented by the data associated with this latter after the clustering process and the descriptors (i.e. the properties) of the data are supposed to be weighted by values within the range [0,1].

Let us consider a set of clusters  $C$  resulting from a clustering method applied on a set of data  $D$ , the local *Recall* ( $Rec$ ) and *Precision* ( $Prec$ ) indexes for a given property  $p$  of the cluster  $c$  can be expressed as:

$$Rec_c(p) = \frac{|c_p^*|}{|D_p^*|}, Prec_c(p) = \frac{|c_p^*|}{|c|}$$

where the notation  $X_p^*$  represents the restriction of the set  $X$  to the set members having the property  $p$ .

Then, for estimating the overall clustering quality, the averaged *Macro-Recall* ( $R$ ) and *Macro-Precision* ( $P$ ) indexes can be expressed as:

$$R = \frac{1}{|\overline{C}|} \sum_{c \in \overline{C}} \frac{1}{S_c} \sum_{p \in S_c} Rec_c(p), P = \frac{1}{|\overline{C}|} \sum_{c \in \overline{C}} \frac{1}{S_c} \sum_{p \in S_c} Prec_c(p) \quad (2)$$

where  $S_c$  is the set of properties which are peculiar to the cluster  $c$  that is described as:

$$S_c = \left\{ p \in d, d \in c \mid \overline{W_c^p} = \max_{c' \in C} \left( \overline{W_{c'}^p} \right) \right\} \quad (3)$$

where  $\overline{C}$  represents the peculiar set of clusters extracted from the clusters of  $C$ , which verifies:

$$\overline{C} = \{c \in C \mid S_c \neq \emptyset\} \text{ and } \overline{W_c^p} = \frac{\sum_{d \in c} W_d^p}{\sum_{c' \in C} \sum_{d \in c'} W_d^p} \quad (4)$$

where  $W_x^p$  represents the weight of the property  $p$  for element  $x$ .

<sup>2</sup> Conversely to classical Recall and Precision indexes that are supervised.

Similarly to IR, the *F-measure* could be used to combine averaged *Recall* and *Precision* results. Moreover, it can be demonstrated (Lamirel and al. (2004)) that if both values of averaged *Recall* and *Precision* reach the unity value, the peculiar set of clusters  $\bar{C}$  represents a Galois lattice. Therefore, the combination of this two measures enables to evaluate to what extent a numerical clustering model can be assimilated to a Galois lattice natural classifier.

*Macro-Recall* and *Macro-Precision* indexes defined by the (Eq. 2) can be considered as cluster-oriented measures because they provide average values of *Recall* and *Precision* for each cluster. They have opposite behaviors according to the number of clusters. Thus, these indexes permit to estimate in a global way an optimal number of clusters for a given method and a given dataset. The best data partition, or clustering result, is in this case the one which minimizes the difference between their values. An example of the behavior of these indexes is given at the Figure 2B. However, similarly to the classical distance-based indexes, their main defect is that they are not enough sensitive to the presence of small number of heterogeneous clusters of large size, especially in the case of the joint existence of a big number of clusters of small size (Ghribi and al. (2010)). To correct that, we propose to construct complementary property-oriented indexes (i.e. micro-measures) of *Micro-Recall* and *Micro-Precision* by averaging the *Recall/Precision* values of the peculiar properties independently of the structure of the clusters:

$$R_m = \frac{1}{|L|} \sum_{c \in \bar{C}, p \in S_c} Rec_c(p), P_m = \frac{1}{|L|} \sum_{c \in \bar{C}, p \in S_c} Prec_c(p)$$

where  $L$  represents the size of the data description space.

In a complementary way, the role of clusters labeling is to highlight the peculiar characteristics or properties of the clusters associated to a cluster model at a given time. Labeling can thus be used both for visualizing or synthesizing clustering results and for validating or optimizing learning of a clustering method (Lamirel and al. (2008)). Labeling can both rely on endogenous data properties or on exogenous ones. Endogenous data properties represent the ones being used during the clustering process. Exogenous data properties represent either complementary properties or specific validation properties. Some label relevance indexes can be derivated from our former quality indexes using a probabilistic approach.

The *Label Recall L-R* derives directly from the Eq. 4. The *Label Precision L-P* can be expressed as:

$$L - P_c(p) = \frac{\sum_{d \in c} W_d^p}{\sum_{p' \in d, d \in c} W_d^p} \quad (5)$$

Consequently, the set of labels  $L_c$  that can be attributed to a cluster  $c$  can be expressed as the set of endogenous or exogenous cluster data properties

which verifies:

$$L_c = \left\{ p \in d, d \in c \mid L_c - F(p) = \max_{c' \in C} \left( L_{c'} - F(p) \right) \right\} \quad (6)$$

where the *Label F-measure*  $L_{c'} - F(p)$  can be defined as:

$$L_{c'} - F(p) = \frac{2 \left( L_{c'} - R(p) \times L_{c'} - P(p) \right)}{L_{c'} - R(p) + L_{c'} - P(p)} \quad (7)$$

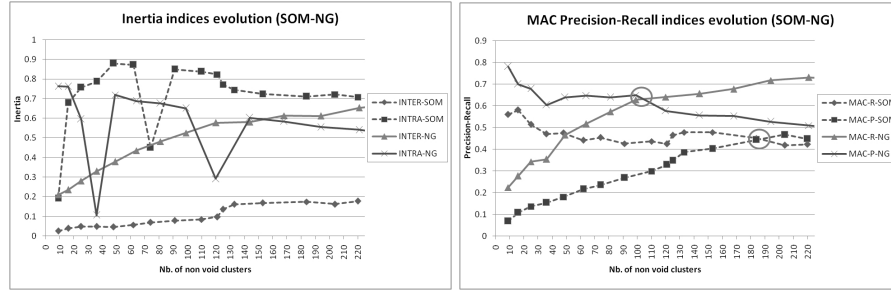
As soon as *Label Recall* is equivalent to the conditional probability  $P(c|p)$  and *Label Precision* is equivalent to the conditional probability  $P(p|c)$ , this former labeling strategy can be classified as an expectation maximization approach with respect to the original definition given by Dempster and al. (1977).

### 3 Experimentation and results

To illustrate the behavior of our new quality indexes, and to compare it to the one of the classical inertia indexes, our test dataset is build up from a set of bibliographic records drawn from the INIST PASCAL database and covering one year of research performed in the Lorraine area. The structure of the records makes it possible to distinguish the titles, the summaries, the indexing keywords and the authors as representatives of the contents of the information published in the corresponding article. In our experiment, the research topics associated with the French keywords field are solely considered. Our test dataset represents a dataset of 1341 records. A frequency threshold of 3 being finally applied on the index terms, it resulted in a data description set of 889 indexing keywords. These keywords cover themselves a large set of different topics (as far one to another as medicine from structural physics or forest cultivation, ...). Moreover, they comprise a high ratio of polysemic forms, like age, stress, pressure ... that are used in the context of many different topics. The resulting experimental dataset can thus be considered as a complex dataset for clustering.

A set of pre-processing steps is applied to the resulting index of the dataset records in order to obtain a weighted vector representation of the information it contains. The resulting index vectors associated to each record are finally weighted using the IDF-weighting scheme in order to decrease the respective influence of both more widespread indexing keywords and polysemic keywords.

To carry out the clustering we exploited in parallel the SOM fixed topology neural method (Kohonen (1982)) and the Neural Gas (NG) free topology neural method (Martinetz and al. (1994)). For each method, we do many different experiments letting vary the number of clusters from 9 to 324 clusters,



**Fig. 1.** Inertia (1A) and Macro Recall Precision (1B) indexes evolution as regards to the number of clusters.

employing the size of an increasing square SOM grid as a basic stepping strategy.

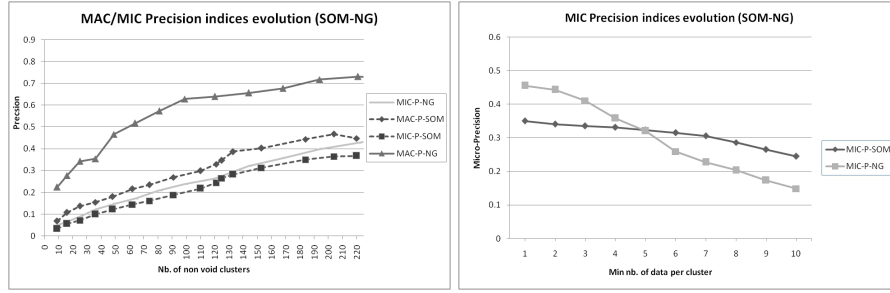
The analysis of the results carried out by an expert showed that only the SOM method provided homogeneous clustering results on this corpus. Hence, in the case of the NG method, the analyst highlighted the presence of garbage clusters attracting most of the data in parallel with chunks clusters representing either marginal groups or unformed topics. This behavior can be confirmed when one looks more precisely to the cluster content and to the cluster size distribution for the said methods, or even to the labels that can be extracted from the clusters in an unsupervised way using our expectation maximization methodology such as described in (Eq. 7).

The results presented in Figure 1A illustrate the fact that the classical indexes of inertia have an unstable behavior which does not make it possible to clearly identify an optimal number of clusters in both contexts of SOM and NG methods<sup>3</sup>. On the other hand, it also appears, in Figure 1B, that the behavior of the *Macro Recall/Precision* indexes is stable and makes it possible to identify an optimal number of clusters in all cases. Indeed, this optimal clusters number can be found out at the break-even point between the *Macro-Recall* and the *Macro-Precision* values (i.e. 99 clusters for NG and 184 clusters for SOM in Figure 1A).

However, none of these former groups of indexes makes it possible by itself to correctly estimate the quality of the results. In particular, they cannot discriminate between homogeneous results of clustering (SOM) and heterogeneous ones (NG). In both cases, they even present the important defect to give the favour to this last family of results.

In the context of our approach, the detection of heterogeneous clustering results can however be achieved by the joint exploitation of the values pro-

<sup>3</sup> The results presented in Figure 1A also highlight that in the case of Kohonen's SOM clustering method the constraints generated by the topography building principle tend to make artificially decreasing the inter-clusters inertia as compared to other classification methods.



**Fig. 2.** Evolution of the values of the Micro-Precision (MIC-P) and Macro-Precision (MAC-P) indexes according to the number of clusters (2A) and their size (2B).

vided by our *Macro-* and *Micro-Precision* indexes, as it is shown in Figure 2A and Figure 2B. The *Micro-Recall/Precision* indexes have general characteristics similar to the *Macro Recall/Precision*. However, by mixing them with these last indexes, they make it possible to identify heterogeneous results of clustering. Indeed, in this last case, the *Precisions* of the clusters of small size will not compensate for any more those of the clusters of big size, and the imprecise properties present in the latter, if they prove to be heterogeneous, will have a considerable effect on the *Micro-Precision*. Consequently, even if the *Macro-* and the *Micro-Precision* measure both the degree of homogeneity of the clusters, the difference between these two measures makes it possible to confirm the presence of heterogeneous clusters of important size.

In the case of NG, the differences between the values of *Micro-* and *Macro-Precision* are increasingly more important than in the case of SOM, whatever the considered number of clusters (Figure 2A). It illustrates the fact that the peculiar properties of the clusters in the partitions generated by NG are largely less precise than those of the clusters produced by SOM. The analysis of the evolution of the *Micro-Precision* curves of the two methods according to the size of the clusters (Figure 2B) permits to clearly highlight that this phenomenon concerns more particularly the NG clusters of big size.

## 4 Conclusion

We proposed a new approach for the evaluation of the quality of the clustering based on the exploitation of the properties associated with the clusters via the indexes of *Macro-* and *Micro- Recall/Precision*. We have shown the advantage of this approach with respect to the traditional methods of evaluation of clustering quality based on distances, at the same time by justifying its theoretical basis via its relationships with the approaches of symbolic classification, but also by illustrating its practical results for the optimization of the number of clusters of a given method. In the context of analysis of complex data, we moreover showed that the single use *Macro-indexes* did not make it

possible to identify heterogeneous results of clustering, which is on the other hand made possible by their joint exploitation with the *Micro-indexes* that we defined within the framework of this new approach. We have also highlighted the additional capabilities of our approach for synthesizing and labeling the clusters content. These capabilities have proved to be useful to the analyst to better understand the nature of clustering results. We more specifically tried out our methodology on textual data, but it is sufficiently general to be naturally applicable on any other type of data, whatever is their nature. In a near future, we thus plan to do some complementary validation experiments on simulated data before applying it at a larger scale in the challenging field of genomics data analysis.

## References

- DAVIES D.L. and BOULDIN D.W.(2000): A cluster separation measure. *IEEE Trans. Pattern Anal. Machine Intell*, 1(4), 224-227.
- DEMPSTER A.P., LAIRD N.M. and RUBIN D.B. (1977): Maximum likelihood for incomplete data via the em algorithm. *Journal of the Royal Statistical Society, vol. B-39: 1-38*.
- DUNN J. (1974): Well Separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4, 95-104.
- GHRIBI M., CUXAC P., LAMIREL J.C. and LELU A. (2010): Mesures de qualité de clustering de documents : Prise en compte de la distribution des mots-clés. *Atelier EvalECD2010, Hamamet, Tunisie*.
- KASSAB R. and LAMIREL J.-C. (2008) : Feature Based Cluster Validation for High Dimensional Data. *IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, 97-103, Innsbruck, Austria.
- KOHONEN T. (2001): *Self-Organising Maps*, 3rd ed., Springer-Verlag, Berlin.
- LAMIREL J.C., FRANÇOIS C., AL SHEHABI S. and HOFFMANN M. (2004): New classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping. *Scientometrics*, 60(3), 445-462.
- LAMIREL J.C. and AL SHEHABI S. (2004): Comparison of unsupervised neural clustering methods for mining Web and textual data. *SCI 2004, Orlando, FL, USA*.
- LAMIREL J.C., TA A.P. and ATTIK M. (2008): Novel Labeling Strategies for Hierarchical Representation of Multidimensional Data Analysis Results. *IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, Innsbruck, Austria.
- LEBART L., MAURINEAU A. and PIRON M. (1982): *Traitement des données statistiques*, Dunod, Paris.
- MARTINETZ T. and SCHULTEN K. (1994): Topology representing networks. *Neural Network.*, 7(3), 507-522.
- ROUSSEEUW P.J. (1987): Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

# Performance Assessment of Optimal Allocation for Large Portfolios

Fabrizio Laurini<sup>1</sup> and Luigi Grossi<sup>2</sup>

<sup>1</sup> Università di Parma, Dipartimento di Economia  
Via Kennedy 6, 43100, Parma, Italy, *fabrizio.laurini@unipr.it*

<sup>2</sup> Università di Verona, Dipartimento di Economia  
Via Dell'Artigliere 19, 37129, Verona, Italy, *luigi.grossi@univr.it*

**Abstract.** We consider the problem of optimal asset allocation for portfolio with a large number of shares. The numerical solution relies on the estimation of the covariance matrix between the assets. Such estimation, typically obtained with maximum likelihood, is affected by the so-called “maximization estimation error”, which grows with the dimension of the covariance matrix. The use of a robust estimator of the covariance matrix can reduce such estimation error considerably, even when data are outlier free and outperform the standard approaches when data have marked heavy tails or affected by the presence of outliers. The performance of our new robust estimator is studied with simulations, and real data.

**Keywords:** financial asset allocation, influential data, robust estimators

## 1 Introduction

To obtain the optimal mix of financial assets, looking at the mean and the risk of a static portfolio, an investor should hold a portfolio on the efficient frontier, as showed by Markowitz (1952). Hence, the investor must estimate the means and the covariances of asset returns. Because of sample estimation error, both the in-sample and out-of-sample performance of this method have theoretically and empirically revealed to perform badly (Broadie (1993)).

The optimal allocation of shares into a portfolio of  $N$  risky assets can be formulated as follows. Let  $Y = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ , be composed by log-returns  $\mathbf{y}_i = (y_{i1}, \dots, y_{it}, \dots, y_{iT})'$  with expected mean given by a  $N \times 1$  vector  $\mu$  and  $N \times N$  covariance matrix  $\Sigma$ , and let  $x = (x_1, \dots, x_N)'$  be the vector of portfolio weights. The portfolio expected return and variance will be  $\mu_p = x' \mu$  and  $\sigma_p^2 = x' \Sigma x$ , respectively.

For a given level of risk-aversion  $\gamma$ , the optimization problem is to solve numerically with quadratic programming

$$\arg \max_x \left( x' \mu - \gamma \frac{1}{2} x' \Sigma x \right), \quad (1)$$

subject to the constraints  $x \geq 0$  (meaning that all the weights are strictly non negative) and  $x' \iota_N = 1$ , where  $\iota_N$  is a  $N \times 1$  vector of ones. When

constraint of no short-selling ( $x \geq 0$ ) is removed, it is easily proved that, using the Lagrange method, for any  $\gamma \geq 0$ , the maximization problem has an analytical solution. However, through all our optimizations experiments, we impose  $x \geq 0$  since, in practice, funds managers and institutional investors are not allowed to sell stocks short. Increasing  $\gamma$  from zero to  $\infty$ , and for each instance solve the optimization problem, we end up calculating each portfolio along the efficient frontier. It is a common practice to calibrate  $\gamma$  such that a particular portfolio has the desired risk profile, with values of  $\gamma \in (1, 4)$  being the most common range of choices.

The expected means, variances and covariances are usually estimated by maximum likelihood estimators (MLE) which are:

$$\hat{\mu} = \frac{1}{T} \iota_T' Y, \quad \hat{\Sigma} = \frac{1}{T} (Y - \iota_T \hat{\mu})' (Y - \iota_T \hat{\mu}).$$

The extent of the bad performance mostly relies on the instability of the mean-variance approach to accommodate for sampling errors in the estimation of the mean of the asset returns (Broadie (1993)).

Here we propose a method to estimate the mean vector by weighting the multivariate portfolio with a robust, but efficient, algorithm. The weights are based on the forward search (FS) approach of Atkinson et al. (2009). The FS is capable to overcome the effect of masking when multiple influential observations are into the sample. The approach is more efficient than existing robust estimators of the covariance matrix, so that when data are “clean” it reveals to have better in-sample performance than, e.g., the minimum covariance determinant (MCD) estimator of Rousseeuw (1985). It also shows performance which are often superior to the classical maximum likelihood estimates (MLE), still when data do not have any influential observations.

In this work we consider only risky assets and compare our method with standard MLE and fast MCD estimator (with 50% breakdown point). We assume cross sectional correlated normally distributed time series, each of those without temporal heteroscedasticity. These assumptions are consistent with monthly data used by practitioners. When data are contaminated, we assume that outliers are drawn from a Gaussian distribution with larger variance than the variance used to generate clean data. In what follows we prove the performance testing some simulations and real data analysis.

## 2 Improving estimates to reduce the error maximization

### 2.1 The problem: error maximization of portfolios

The effect of the error maximization for asset allocation was firstly discussed by Michaud (1989). The main result is that according to the Markowitz model the efficient frontier, derived by mean and covariance estimates, tend to create



weights of securities which over-weights the assets with positive bias in the mean estimate and negative bias in the covariance matrix. As a resulting combination of these biases, the estimated frontier is “too optimistic” and the final outcome is to set up a portfolio where the expected return is too high and the risk too low.

The error maximization problem is exacerbated when the number of assets hold in a portfolio is of some dozens. Large portfolio, on the other hand, are commonly managed by practitioners, with hedge-fund institutions being probably the most notorious example.

## 2.2 A new covariance estimator

The covariance matrix is estimated commonly with MLE, but such estimation is problematic when the number of variables is large, no matter whether or not influential observations (which might mask each other) are included into the data. Here we propose a weighted estimator of the covariance based on the forward search method. Essentially the weights are computed after a forward sequential search of influential observation, and then a vector of weights is created. Such weights are attached to each entry in the original matrix of data, and then a simple weighted estimate of mean and covariance is carried. The details are the following.

We derive weights  $w_t \in [0, 1]$ , for each observation in the multiple time series  $y_t = (y_{1t}, \dots, y_{Nt})'$ ,  $t = 1, \dots, T$ , with a procedure similar to that of Grossi and Laurini (2009). The weights are such that the most influential observations get small weight.

We start partitioning the initial data by robustly selecting a small set of multivariate data with  $m_0$  rows (the complementary set would be of size  $T - m_0$ ), and with such small subset estimate the mean vector and the covariance matrix. Then, we compute squared Mahalanobis distance for all  $T$  observations as  $d_t^2 = (y_t - \hat{\mu})' \hat{\Sigma}^{-1} (y_t - \hat{\mu})$ . We build a new subsample of size  $m_0 + 1$  with the units having the smallest  $d_t^2$ . So, for a given subset  $S_*^{(m)}$  of dimension  $m \geq m_0$ , we calculate a set of  $T$  squared Mahalanobis distances, defined as

$$d_{t, S_*^{(m)}}^{*2} = (y_t - \hat{\mu}_m^*)' (\hat{\Sigma}_m^*)^{-1} (y_t - \hat{\mu}_m^*), \quad t = 1, \dots, T, \quad (2)$$

where  $\hat{\mu}_m^*$  and  $\hat{\Sigma}_m^*$  are the mean and covariance matrix estimated on the  $m$ -sized subset. The  $m + 1$  units with smallest  $d_{t, S_*^{(m)}}^{*2}$  will be included in the new subset and algorithm iterates until all units are included. When all units are included we have the special case of MLE estimates of  $\hat{\mu}_T^*$  and  $\hat{\Sigma}_T^*$ .

During the iterative inclusion of data, we exploit a result from Atkinson et al. (2009), that for a random set of observations with  $N$  columns, the squared Mahalanobis distance has distribution

$$d_t^{*2} \sim [T/(T-1)][N(m-1)/(m-N)]F_{N, T-N}.$$

So, we compare the trajectories of  $d_{(t)}^{*2}$  during the forward search with confidence bands from the  $F$  distribution. Alternatively, simulated envelopes can be adopted. At each step of the forward search, we measure the degree of outlyingness of each observation  $t = 1, \dots, T$ , as the squared Euclidean distance,  $\pi$ , between the distance (2) lying outside a confidence band and the boundaries of the band itself by considering the  $F$  distribution with  $N, T - N$  degrees of freedom, and the percentile  $F_\delta$  at the nominal level  $1 - \delta$ . For a fixed step of the forward search  $m$ , we record the distance of the  $t$ -th trajectory from the percentile of the confidence band, provided that the trajectory is over the  $1 - \delta$  nominal level. If  $d_{(t)}^{*2}$  lies under the  $F_\delta$  percentile, then, at the  $m$ -th step, it will get zero distance. At the next step  $m + 1$ , the weight of the  $t$ -th observation will be increased by an amount which is induced by the squared Euclidean distance from the  $t$ -th trajectory and the percentile of the confidence band, provided that at step  $m + 1$  the  $t$ -th trajectory exceeds the nominal level  $1 - \delta$ . If at step  $(m + 1)$ -th the  $t$ -th trajectory lies under the  $F_\delta$  quantile, then a zero will be added to the distance computed at the step  $m$ .

The overall degree of outlyingness for the  $t$ -th observation is given by the sum of all squared Euclidean distances, computed only when the trajectory exceeds the confidence band. Formally, letting  $\pi_m^{(t)}$  be the distance between  $d_{(t)}^{*2}$  and the percentile  $F_\delta$ , for the unit  $t$ -th at step  $m$ , we define the squared Euclidean distance as:  $\pi_m^{(t)} = 0$  if  $d_{(t)}^{*2} \in [0, F_\delta]$ ,  $\pi_m^{(t)} = (d_{(t)}^{*2} - F_\delta)^2$  if  $d_{(t)}^{*2} > F_\delta$  and we consider the overall distance of the  $t$ -th observation as the sum of such distances during the forward search, i.e.

$$\pi_t = \frac{\sum_{m=m_0}^T \pi_m^{(t)}}{T - m_0 + 1}.$$

Then, the influence index for time  $t$  must be distributed among the single asset return at the same time. This problem is addressed scaling the absolute value  $|y_{it}|$  of asset  $i$  at time  $t$  of the  $i$ -th asset return with the interquartile difference ( $DI$ ), that is:

$$\theta_{it} = \frac{|y_{it}|}{DI(y_{it})}.$$

The influence index  $\pi_{it}$  for each return  $y_{it}$  is obtained as:

$$\pi_{it} = \pi_t \frac{\theta_{it}}{\sum_{i=1}^N \theta_{it}}.$$

The computation of a weight, in the interval  $[0, 1]$ , is achieved using the mapping  $w_{it} = \exp(-\pi_{it})$ . The weights are computed for each observation at the end of the forward search.

We build a weighted mean vector  $\tilde{\mu}$  and a weighted covariance matrix  $\tilde{\Sigma}$  to be used in the optimization procedure. The weighted mean vector is trivially obtained as  $\tilde{\mu} = [W_1' y_t^{(1)}, \dots, W_N' y_t^{(N)}]$ , where  $W_i$  is a  $T \times 1$  weight vector whose generic element is  $\{w_{it}\}$ . For the covariance matrix, we consider

the weighted return  $y_{it}^* = y_{it}w_{it}^{1/2}$ . The weighted covariance matrix will be obtained as  $\tilde{\Sigma} = 1/T(Y^* - \iota_T\tilde{\mu})'(Y^* - \iota_T\tilde{\mu})$ , where  $Y^* = (y_1^*, \dots, y_N^*)$  and  $y_i^* = (y_{i1}^*, \dots, y_{iT}^*)'$ . In the last step of the procedure, the weighted covariance matrix is corrected to take into account the reduction of variability induced by weighting some observations to zero. This operation produce a trimming in the data set which we correct following Maronna *et al.* (2006, p. 186). Our scale parameter, which will be denoted as  $\hat{c}$ , and it is given by the inverse of the total weights over the total number of observations, that is  $\hat{c} = [\sum_{i=1}^N \sum_{t=1}^T w_{it}/NT]^{(-1)}$ . The final estimates of the covariance matrix will be then  $\hat{\Sigma} = \hat{c}\tilde{\Sigma}$ , and referred to as FWD estimator.

### 3 Simulations: the in-sample performance

We simulate a portfolio with 40 shares, assumed to be observed monthly. The marginal means  $\mu$  and variances  $\Sigma_{ii}$  are randomly selected from uniform distributions with range lying in the typical range of monthly observed asset returns. The correlation matrix is block diagonal with the first block of size 20 having constant value  $\rho = 0.3$ , and the second block with negative correlations  $\rho = -0.3$ . The two blocks are uncorrelated each other. In order to have a definite positive matrix, we use the built-in algorithm `nearPD` in the Matrix library of the freely available package R. As a result of the correction, the smallest negative correlation is approximately equal to  $-0.12$ . These parameters are tuned so that they are consistent with the empirical values commonly observed in financial monthly returns. The length of the simulated series are of  $T = 81$  observations (about 7-year data), so that there are no problems of identification and estimation of the covariance matrix from simulated data.

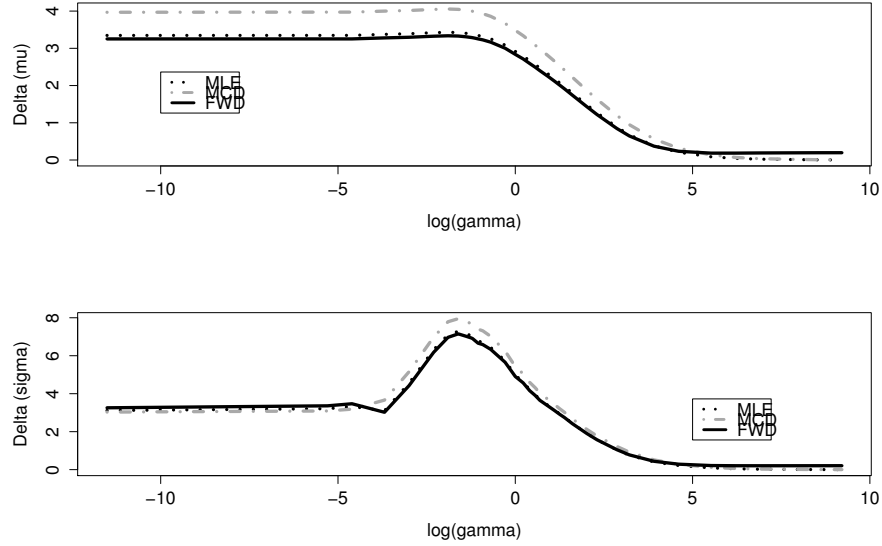
We assess the distance between true and estimated frontiers by exploring the errors onto both the “mean” and the “standard deviation” projections. This means that, for a fixed value of  $\gamma$ , we compare the value of the estimated mean  $\hat{\mu}_P(\gamma)$  and the estimated standard deviation  $\hat{\sigma}_P(\gamma)$  with the true parameters. For each simulation  $s = 1, \dots, S$  we average the squared distance by computing the root mean squared error (RMSE) in the  $\mu$  direction and  $\sigma$  direction, defined respectively as

$$\Delta_{\mu}(\gamma) = 100 \times \sqrt{\frac{1}{S} \sum_{s=1}^S [\hat{\mu}_P(\gamma; s) - \mu_P(\gamma; s)]^2}$$

and

$$\Delta_{\sigma}(\gamma) = 100 \times \sqrt{\frac{1}{S} \sum_{s=1}^S [\hat{\sigma}_P(\gamma; s) - \sigma_P(\gamma; s)]^2}$$

The estimation of  $\hat{\mu}_P$  and  $\hat{\sigma}_P$  requires to find the optimal weights for a given covariance matrix and mean returns vector. For the large portfolio with



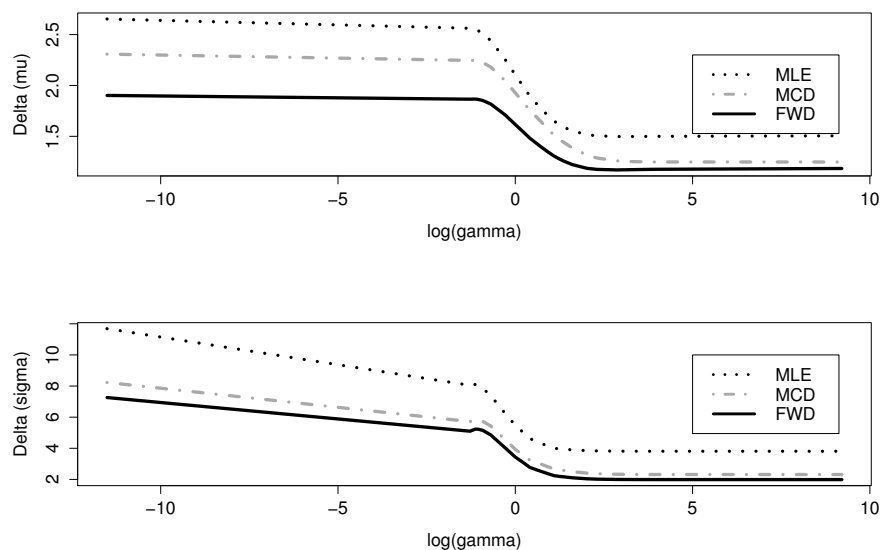
**Fig. 1.** RMSE for MLE (dotted line), FWD (solid line) and MCD (gray dashed line) for a large portfolio

parameters introduced above we have obtained RMSE as shown in Fig. 1, with FWD estimator performing very well.

We also report a result from a simulation (with the same parameters as before) where some “extreme” observations were introduced at random position so that the effect of error maximization is even bigger than with our result with “clean” data. The effect of adding influential observation is to have values very much away from the minimum variance portfolio (i.e. data are very far away from the case of  $\gamma \rightarrow \infty$ ). The better performance of the FWD method is illustrated from RMSE of Fig. 2. From such contaminated simulation it seems that the effect of outliers is higher when there is negative correlation, i.e. when there is higher diversification and risk reduction in the portfolio.

#### 4 Real data analysis: the out-of-sample performance

We consider the monthly returns of six stocks of the US market with data from January 1973 to March 2009 included. Data come from Datastream. We have computed the efficient frontier according to the so called “tangency” optimality, i.e. using the Sharpe ratio for the optimal weights allocation of each asset into the final portfolio. The Sharpe ratio corresponds, typically, to a value of  $\gamma \in (1, 4)$ . We show the out-of-sample performance by comparing the



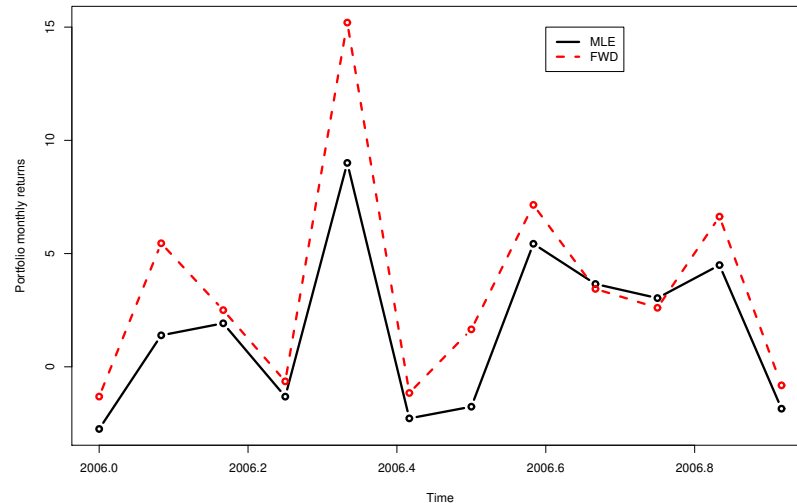
**Fig. 2.** RMSE for MLE (dotted line), FWD (solid line) and MCD (gray dashed line) for a large portfolio with influential observations added at random.

allocations derived from the analyzed estimators and using a rolling windows technique which consists of estimating weights (MLE and FWD) using data for  $t = 1, \dots, K - 1$  and get the average portfolio return in time  $K$ . Fig 3 reports the output of the rolling windows procedure estimating parameters on data until the end of 2005 and taking year 2006 as a forecast period (the time is in the horizontal axis of the figure). The dashed line is for FWD and solid for MLE. Portfolio performances are generally better when the forward search weights are applied. To summarize the performances of the two trajectories, the Sharpe ratio could be used and provide a single statistic with practical relevance.

## 5 Discussion: limitations and further work

In this preliminary work we have studied the performance of standard MLE and new robust but efficient estimator (FWD) for optimal asset allocation for a large portfolio of shares. Under the assumptions made in the paper, we conclude that the robust FWD estimator can be a useful tool for portfolio management. However, there are some drawbacks that require further work to be done.

Relaxing the normality and homoscedasticity assumptions can lead to different conclusions. The behaviour of the FWD estimator for multiple time



**Fig. 3.** Portfolio monthly performances in 2006 using a rolling windows technique

series with heteroscedasticity has still to be studied carefully. Moreover, accounting for ARCH/GARCH effects will increase significantly the overall computational load of the FWD estimator (which is already substantial, when compared to the fast MCD). These seem to be important avenues for future research.

## References

- ATKINSON, A.C., RIANI, M. and CERIOLO, A. (2009): Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society B* 71, 447-466.
- BROADIE M. (1993): Computing efficient frontiers using estimated parameters. *Annals of Operations Research* 45, 21-58.
- GROSSI, L. and LAURINI, F. (2009): A robust forward weighted Lagrange multiplier test for conditional heteroscedasticity. *Computational Statistics and Data Analysis* 53, 2251-2263.
- MARKOWITZ, H. M. (1952): Mean-variance analysis in portfolio choice and capital markets. *Journal of Finance* 7, 77-91.
- MARONNA, R.A., MARTIN, R.D. and YOHAI, V.J. (2006): *Robust Statistics*, Wiley, New York.
- MICHAUD, R. O. (1989): The Markowitz optimization enigma: is "optimized" optimal? *Financial Analyst Journal* 45, 31-42.
- ROUSSEEUW, P.J. (1985): Multivariate Estimation With High Breakdown Point. In: W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (Eds.): *Mathematical Statistics and Applications*, Vol. B. Dordrecht, Reidel, 283-297.

# Clustering of Multiple Dissimilarity Data Tables for Documents Categorization

Yves Lechevallier<sup>1</sup>, Francisco de A. T. de Carvalho<sup>2</sup>, Thierry Despeyroux<sup>1</sup>,  
and Filipe M. de Melo<sup>2</sup>

<sup>1</sup> INRIA, Paris-Rocquencourt  
78153 Le Chesnay cedex, France,  
{*Yves.Lechevallier,Thierry.Despeyroux*}@inria.fr

<sup>2</sup> Centro de Informatica -CIn/UFPE  
Av. Prof. Luiz Freire, s/n -Cidade Universitaria, CEP 50740-540, Recife-PE,  
Brasil, {*fatc, fmm*}@cin.ufpe.br

**Abstract.** This paper introduces a clustering algorithm that is able to partition objects taking into account simultaneously their relational descriptions given by multiple dissimilarity matrices. These matrices could have been generated using different sets of variables and a fixed dissimilarity function, using a fixed set of variables and different dissimilarity functions or using different sets of variables and dissimilarity functions. This method, which is based on the dynamic hard clustering algorithm for relational data, is designed to provided a partition and a prototype for each cluster as well as to learn a relevance weight for each dissimilarity matrix by optimizing an adequacy criterion that measures the fit between clusters and their representatives. These relevance weights change at each algorithm iteration and are different from one cluster to another. Experiments aiming at obtaining a categorization of a document data base demonstrate the usefulness of this partitionnal clustering method.

**Keywords:** clustering analysis, relational data, documents categorization

## 1 Introduction

Clustering is a popular task in knowledge discovering and it is applied in various fields including data mining, pattern recognition, computer vision, etc (Gordon (1999), Jain et al (1999)). Clustering methods aims at organizing a set of objects into clusters such that items within a given cluster have a high degree of similarity, while items belonging to different clusters have a high degree of dissimilarity. The most popular clustering techniques are hierarchical and partitioning methods. Partitioning methods seek to obtain a single partition of the input data into a fixed number of clusters. Such methods often look for a partition that optimizes (locally) an adequacy criterion function.

There are two common representations of the objects upon which clustering can be based : (usual or symbolic) feature data and relational data.

When each object is described by a vector of quantitative or qualitative values the set of vectors describing the objects is called a feature data. When each (complex) object is described by a vector of sets of categories, intervals or weight histograms, the set of vectors describing the objects is called a symbolic feature data. Symbolic data has been mainly studied in Symbolic Data Analysis (SDA) (Bock and Diday (2000)). Alternatively, when each pair of objects is represented by a relationship, then we have relational data. The most common case of relational data is when we have (a matrix of) dissimilarity data, say  $R = [r_{il}]$ , where  $r_{il}$  is the pairwise dissimilarity (often a distance) between objects  $i$  and  $l$ .

This paper gives a clustering algorithm that is able to partition objects taking simultaneously into account their relational descriptions given by multiple dissimilarity matrices. The main idea is to obtain a collaborative role of the different dissimilarity matrices (Pedrycz (2002)) in order to obtain a final consensus partition (Leclerc and Cucumel (1987)). These dissimilarity matrices could have been generated using different sets of variables and a fixed dissimilarity function (the final partition gives a consensus between different views (sets of variables) describing the objects), using a fixed set of variables and different dissimilarity functions (the final partition gives the consensus between different dissimilarity functions) or using different sets of variables and dissimilarity functions. Moreover, the influence of the different dissimilarity matrices is not equally important in the definition of the clusters in the final consensus partition. Thus, in order to obtain a meaningful partition from all dissimilarity matrices, it is necessary to learn cluster-dependent relevance weights for each dissimilarity matrix.

Frigui et al (2007) proposed CARD, a clustering algorithm that is able to partition objects taking into account multiple dissimilarity matrices and that learns a relevance weight for each dissimilarity matrix in each cluster. CARD is mainly based on the well know fuzzy clustering algorithms for relational data RFCM (Hathaway et al (1989)) and FANNY (Kaufman and Rousseeuw (1990)).

The clustering algorithm given in this paper is designed to give a partition and a prototype for each cluster as well as to learn a relevance weight for each dissimilarity matrix by optimizing an adequacy criterion that measures the fit between clusters and their representatives. These relevance weights change at each algorithm iteration and are different from one cluster to another. However, this method is based on the dynamic hard clustering algorithm for relational data (Lechevallier (1974), De Carvalho et al (2008), De Carvalho et al (2009)) as well as on the dynamic clustering method based on adaptive distances (Diday and Govaert (1977), De Carvalho and Lechevallier(2009)). The adaptive dynamic clustering method gives a partition as well as a prototype for each cluster and learns a relevance weight for each variable in each cluster. In order to demonstrate the usefulness of this clustering algorithm,



experiments were designed in order to obtain a categorization of a document data base.

This paper is organized as follows. Section 2 presents a partitioning clustering algorithm based on multiple dissimilarity matrices. In order to illustrate the usefulness of this clustering method, section 3 shows the application of this algorithm in order to obtain a categorization of a document data base. Finally, section 4 presents the conclusions.

## 2 A Dynamic Clustering Algorithm Based on Multiple Dissimilarity Matrices

In this section, we introduce an extension of the dynamic clustering algorithm for relational data (De Carvalho et al (2008)) which is able to partition objects taking simultaneously into account their relational descriptions given by multiple dissimilarity matrices.

Let  $E = \{e_1, \dots, e_n\}$  be a set of  $n$  examples and let  $p$  dissimilarity  $n \times n$  matrices  $(\mathbf{D}_1, \dots, \mathbf{D}_j, \dots, \mathbf{D}_p)$  where  $\mathbf{D}_j[i, l] = d_j(e_i, e_l)$  gives the dissimilarity between objects  $e_i$  and  $e_l$  on dissimilarity matrix  $\mathbf{D}_j$ . Assume that the prototype  $g_k$  of cluster  $C_k$  belongs to the set of examples  $E$ , *i.e.*,  $g_k \in E \forall k = 1, \dots, K$ .

The dynamic hard clustering algorithm with relevance weight for each dissimilarity matrix looks for a partition  $P = (C_1, \dots, C_K)$  of  $E$  into  $K$  clusters and the corresponding prototype  $g_k \in E$  representing the cluster  $C_k$  in  $P$  such that an adequacy criterion (objective function) measuring the fit between the clusters and their prototypes is locally optimized. The adequacy criterion is defined as

$$J = \sum_{k=1}^K \sum_{e_i \in C_k} d^{(k)}(e_i, g_k) = \sum_{k=1}^K \sum_{e_i \in C_k} \sum_{j=1}^p \lambda_k^j d_j(e_i, g_k) \quad (1)$$

in which

$$d^{(k)}(e_i, g_k) = \sum_{j=1}^p \lambda_k^j d_j(e_i, g_k) \quad (2)$$

is the dissimilarity between an example  $e_i \in C_k$  and the cluster prototype  $g_k \in E$  parameterized by relevance weight vector  $\lambda_k = (\lambda_k^1, \dots, \lambda_k^j, \dots, \lambda_k^p)$  where  $\lambda_k^j$  is the weight between the dissimilarity matrix  $\mathbf{D}_j$  and the clusters  $C_k$ , and  $d_j(e_i, g_k)$  is the local dissimilarity  $d_j$  between an example  $e_i \in C_k$  and the cluster prototype  $g_k \in E$ .

The relevance weight matrix  $\lambda$  composed by  $K$  relevance weight vectors  $\lambda_k = (\lambda_k^1, \dots, \lambda_k^j, \dots, \lambda_k^p)$  changes at each iteration, *i.e.*, they are not determined absolutely, and are different from one cluster to another. Our clustering algorithm alternates the three following steps:

**Step 1: Definition of the Best Prototypes**

In this step, the partition  $P = (C_1, \dots, C_K)$  of  $E$  into  $K$  clusters and the relevance weight matrix  $\lambda$  are fixed.

**Proposition 15.** *The prototype  $g_k = e_l \in E$  of cluster  $C_k$ , which minimizes the clustering criterion  $J$ , is computed according to:*

$$l = \arg \min_{1 \leq h \leq n} \sum_{e_i \in C_k} \sum_{j=1}^p \lambda_h^j d_j(e_i, e_h) \quad (3)$$

**Step 2: Definition of the Best Relevance Weight Matrix**

In this step, the partition  $P = (C_1, \dots, C_K)$  of  $E$  and the vector of prototypes  $\mathbf{g} = (g_1, \dots, g_K)$  are fixed.

**Proposition 16.** *The element  $j$  of the relevance weight vector  $\lambda_k = (\lambda_k^1, \dots, \lambda_k^p)$ , which minimizes the clustering criterion  $J$  under  $\lambda_k^j > 0$  and  $\prod_{j=1}^p \lambda_k^j = 1$ , is calculated by the following expression:*

$$\lambda_k^j = \frac{\{\prod_{h=1}^p [\sum_{e_i \in C_k} d_h(e_i, g_k)]\}^{\frac{1}{p}}}{[\sum_{e_i \in C_k} d_j(e_i, g_k)]} \quad (4)$$

**Step 3: Definition of the Best Partition**

In this step, the vector of prototypes  $\mathbf{g} = (g_1, \dots, g_K)$  and the relevance weight matrix  $\lambda$  are fixed.

**Proposition 17.** *The cluster  $C_k$ , which minimize the criterion  $J$ , is updated according to the following allocation rule:*

$$C_k = \{e_i \in E : d^{(k)}(e_i, g_k) < d^{(h)}(e_i, g_h) \forall h \neq k\} \quad (5)$$

*If the minimum is not unique,  $e_i$  is assigned to the class having the smallest index*

It's easy to demonstrate that each preview step decreases the criterion  $J$ . The dynamic hard clustering algorithm with relevance weight for each dissimilarity matrix sets an initial partition and alternates three steps until convergence, when the criterion  $J(P, \lambda, \mathbf{g})$  reaches a stationary value representing a local minimum. This algorithm is summarized below.

**The Dynamic Hard Clustering Algorithm with Relevance Weight Matrix****a. Initialization.**

Fix the number  $K$  of clusters;

Randomly select  $K$  distinct objects  $g_k \in E$ ;

Set the Relevance Weight Matrix  $\lambda$  where  $\lambda_k = (\lambda_k^1, \dots, \lambda_k^p) = (1, \dots, 1)$ ;

Assign each object  $e_i$  to the closest prototype in order to obtain the partition  $P = (C_1, \dots, C_K)$  where  $C_k$  is constructed by the rule (5).

- b. *Step 1: definition of the best prototypes.*  
The partition  $P = (C_1, \dots, C_K)$  and the relevance weight matrix  $\lambda$  are fixed.  
Compute the prototype  $g_k \in E$  of cluster  $C_k$  according to equation (3)
- c. *Step 2: definition of the best relevance weight matrix.*  
The vector of prototypes  $\mathbf{g}$  and the partition  $P = (C_1, \dots, C_K)$  are fixed.  
For each  $k$  compute the component of the weight vector  $\lambda_k$  according to equation (4)
- d. *Step 3: definition of the best partition.*  
The vector of prototypes  $\mathbf{g}$  and the relevance weight matrix  $\lambda$  are fixed.  
Construct the new partition  $P' = (C'_1, \dots, C'_K)$  with the rule given by (5) and control the convergence by:  
 $test \leftarrow 0$ ;  
 for  $i = 1$  to  $n$  do  
    $e_i$  belonged to the class  $C_m$  and belongs to the winning cluster  $C'_k$   
   if  $k \neq m$  then  $test \leftarrow 1$ ;  
 $P \leftarrow P'$ ;
- e. *Stopping criterion.* If  $test = 0$  then STOP, otherwise go to 2 (Step 1).

### 3 Application: document data base categorization

To illustrate the usefulness of the proposed clustering algorithm, we use it to categorize a document data base. The document data base is a collection of reports produced by every Inria (The French National Institute for Research in Computer Science and Control) research team in 2007. Research teams are grouped into scientific *themes* that do not correspond to an organizational structure (such as departments or divisions), but act as a virtual structure for the purpose of presentation, communication and evaluation. Choice of themes and team allocation are mostly related to strategic objectives and scientific closeness between existing teams, but also take in account some geographical constraints, such as the desire for a theme to be representative of most Inria centers. Our aim is to compare the categorization given automatically by the clustering algorithm introduced in this paper with the *a priori* expert categorization given by INRIA.

These reports are written in English. The sources are LaTeX documents, and are automatically translated into XML, then to HTML to be published on the Web. In the rest of the paper we implicitly refer to the XML version of the Activity Report. The logical structure of the RA is defined by an XML DTD with a few mandatory sections and some optional parts.

In this application we considered activity reports from 164 INRIA research teams in 2007. On each activity report, 4 sections have been selected to describe a research team: *overall objectives*, *scientific foundations*, *dissemination* and *new results*. The *overall objectives* part defines the research objectives and *scientific foundations* provides the scientific background followed

- 
- ▼ **APPLIED MATHEMATICS, COMPUTATION AND SIMULATION**
    - ▶ Computational models and simulation
    - ▶ Stochastic Methods and Models
    - ▶ Optimization, Learning and Statistical Methods
    - ▶ Modeling, Optimization, and Control of Dynamic Systems
  - ▼ **ALGORITHMS, PROGRAMMING, SOFTWARE AND ARCHITECTURE**
    - ▶ Programs, Verification and Proofs
    - ▶ Algorithms, Certification, and Cryptography
    - ▶ Embedded and Real Time Systems
    - ▶ Architecture and Compiling
  - ▼ **NETWORKS, SYSTEMS AND SERVICES, DISTRIBUTED COMPUTING**
    - ▶ Networks and Telecommunications
    - ▶ Distributed Systems and Services
    - ▶ Distributed and High Performance Computing
  - ▼ **PERCEPTION, COGNITION, INTERACTION**
    - ▶ Vision, Perception and Multimedia Understanding
    - ▶ Interaction and Visualization
    - ▶ Knowledge and Data Representation and Management
    - ▶ Robotics
    - ▶ Audio, Speech, and Language Processing
  - ▼ **COMPUTATIONAL SCIENCES FOR BIOLOGY, MEDICINE AND THE ENVIRONMENT**
    - ▶ Observation and Modeling for Environmental Sciences
    - ▶ Observation, Modeling, and Control for Life Sciences
    - ▶ Computational Biology and Bioinformatics
    - ▶ Computational Medicine and Neurosciences
- Members
  - Overall Objectives
    - Introduction
    - Highlights of the year
  - Scientific Foundations
    - Introduction
    - Modeling Interfaces and Contacts
    - Modeling the Flexibility of Macro-molecules
  - Software
    - Web services
    - CGAL and Ipe
  - New Results
    - Modeling Interfaces and Contacts
    - Modeling the flexibility of macro-molecules
    - Algorithmic foundations
  - Other Grants and Activities
    - International initiatives
  - Dissemination
    - Animation of the scientific community
    - Teaching
    - Participation to conferences, seminars, invitations
  - Bibliography
    - Major publications
    - Publications of the year
    - References in notes

**Fig. 1.** INRIA research categorization and example of the Activity Report summary

by potential applications of the research domain. *Dissemination* includes any teaching activity, involvement with the research community (program committees, editorial boards, conference and workshop organization) and seminars. The *new results* includes the principal results obtained during this year.

From these activity reports we initially obtained 4 feature data tables, each table with 164 individuals (INRIA research team) described by the frequent words (categories) present in one of 4 variables. The number of frequent words in *overall objectives* section is 220, 210 for *scientific foundations*, 404 for *dissemination* and 547 for *new results* sections. Each cell on a data table gives the frequency of a word for the considered activity report section and research team.

Then, 4 relational data tables have been obtained from the 4 feature data tables through a dissimilarity measure derived from the affinity coefficient (Barcelar-Nicolau (2000)). We assume that each individual is described by one set-valued variable (“presentation”, etc.) which has  $m_j$  modalities (or categories)  $\{1, \dots, m\}$ . An individual  $e_i$  is described by  $\mathbf{x}_i = (n_{i1}, \dots, n_{im})$  where  $n_{ij}$  is the frequency of modality  $j$ . The dissimilarity between a pair of individuals  $e_i$  and  $e_{i'}$  is given by:

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = 1 - \sum_{j=1}^m \sqrt{\frac{n_{ij}}{n_{i\bullet}} \frac{n_{i'j}}{n_{i'\bullet}}} \quad \text{where} \quad n_{i\bullet} = \sum_{j=1}^m n_{ij}.$$

All these relational data tables were normalized according to their overall dispersion (Chavent (2005)) to have the same dispersion. This means that each dissimilarity  $d(\mathbf{x}_i, \mathbf{x}_{i'})$  in a given relation data table has been normalized as  $\frac{d(\mathbf{x}_i, \mathbf{x}_{i'})}{T}$  where  $T = \sum_{i=1}^n d(e_i, g)$  is the overall dispersion and  $g = e_l \in E = \{e_1, \dots, e_n\}$  is the overall prototype, which is computed according to  $l = \operatorname{argmin}_{1 \leq h \leq n} \sum_{i=1}^n d(e_i, e_h)$ .

### 3.1 Results

The clustering algorithm has been performed simultaneously on these 4 relational data tables (“presentation”, “foundation”, “dissemination” and “bibliography”) in order to obtain a partition in  $K \in \{1, \dots, 15\}$ . For a fixed number of clusters  $K$ , the clustering algorithm is run 100 times and the best result according to the adequacy criterion is selected.

In order to determine the number of clusters, we used the approach described by Da Silva (2009), which consists on the choice of the peaks on the graph of the “second order differences” of the clustering criterion (equation (1)):  $J^{(K-1)} + J^{(K+1)} - 2J^{(K)}$ ,  $K = 2, \dots, 14$ . According to this approach, we fixed the number of clusters in 4 and 9.

The 4-cluster and the 9-cluster partitions obtained with this clustering algorithm were compared with the INRIA research team categorization 5-class partition known a priori. The 5-class a priori categorization is as follows: “Applied Mathematics, Computation and Simulation (M)”, “Algorithmics, Programming, Software and Architecture (A)”, “Networks, Systems and Services, Distributed Computing (N)”, “Perception, Cognition, Interaction (P)” and “Computational Sciences for Biology, Medicine and the Environment (C)”. These 5 categories are themselves divided into several sub-categories. In many points we retrieve in the 9-cluster partition the categorization done a priori by INRIA. For example the sub-category “Networks and Telecommunications” of N fits exactly in one cluster. The two sub-categories “Distributed Systems and Services” and “Distributed and High Performance Computing” are merged into a unique cluster, indicating that from the language used point of view the distinction between these two categories is artificial. Some teams have also migrate. For example it seems that the language used in Cryptography (that is part of A in the a priori categorization) is closer to the language used in math (M). Looking at the 4-cluster partition, some migrations are also clearly detected, which have a political sense, in particular when the concerned team is found in a cluster corresponding to the “right” category in a former categorization done by INRIA. Finally, it seems that teams in the C category share the same language as teams in M or in P stressing the fact that in the activity report the weight of the scientific foundations is important, and this fact showing up in both partitions is however clearer in the 4-cluster partition than in the 9-cluster one.

## 4 Concluding Remarks

This paper introduced a clustering algorithm that is able to partition objects taking into account simultaneously their relational descriptions given by multiple dissimilarity matrices. These matrices could have been generated using different sets of variables and dissimilarity functions. This algorithm provides a partition and a prototype for each cluster as well as a relevance weight for each dissimilarity matrix by optimizing an adequacy criterion that

measures the fit between clusters and their representatives. These relevance weights change at each algorithm iteration and are different from one cluster to another. The usefulness of this algorithm was illustrated comparing the categorization of INRIA research teams given by the clustering algorithm with the a priori expert categorization given by INRIA. The clustering algorithm was able to retrieve the a priori categorization, the observed minor divergences being explained by political choices of INRIA.

## References

- BACELAR-NICOLAU, H. (2000): The affinity coefficient. In: H.H Bock and E. Diday (Eds.): *Analysis of Symbolic Data*. Springer, Heidelberg, 160–165.
- BOCK, H.H. and DIDAY, E. (2000): *Analysis of Symbolic Data*. Springer, Heidelberg.
- CHAVENT, M. (2005): Normalized k-means clustering of hyper-rectangles. In: *Proceedings of the XIth International Symposium of Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France*, 670–677.
- DE CARVALHO, F. A. T. and LECHEVALLIER, Y. and VERDE, R. (2008): Clustering methods in symbolic data analysis. In: Edwin Diday; Monique Noirhomme-Fraiture. (Eds.): *Symbolic Data Analysis and the SODAS Software*. Wiley-Interscience, San Francisco, 181–204.
- DE CARVALHO, F. A. T. and CSERNEL, M. and LECHEVALLIER, Y. (2009): Clustering constrained symbolic data *Pattern Recognition Letters*, 30 (11), 1037–1045.
- DE CARVALHO, F. A. T., LECHEVALLIER, Y. (2009): Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recognition*, 42 (7), 1223–1236.
- DA SILVA, A. (2009): Analyse de données évolutives: application aux données d'usage Web. *Thèse de Doctorat. Université Paris-IX Dauphine*.
- DIDAY, E., GOVAERT, G. (1977): Classification automatique avec distances adaptatives. *R.A.I.R.O. Informatique Computer Science* 11 (4), 329–349.
- FRIGUI, H., HWANG, C. and RHEE, F. C. (2007): Clustering and aggregation of relational data with applications to image database categorization. *Pattern Recog.*, 40 (11), 3053–3068.
- GORDON, A.D. (1999): *Classification*. Chapman and Hall/CRC, Boca Raton, Florida.
- HATHAWAY, R. J., DAVENPORT, J. W. and BEZDEK, J. C. (1989): Relational duals of the c-means algorithms. *Pattern Recog.*, 22, 205–212.
- JAIN, A.K., MURTY, M.N. and FLYN, P.J. (1999): Data clustering: A review. *ACM Comput. Surv.* 31 (3), 264–323.
- KAUFMAN, L. and ROUSSEEUW, P. J. (1990): *Finding Groups in Data*. NewYork: Wiley.
- LECHEVALLIER, Y. (1974): Optimisation de quelques critères en classification automatique et application a l'étude des modifications des protéines sériques en pathologie clinique. *Thèse de 3eme cycle. Université Paris-VI*.
- LECLERC, B. and CUCUMEL, G. (1987): Consensus en classification : une revue bibliographique. *Mathématique et sciences humaines*, 100, 109–128
- PEDRYCZ, W. (2002): Collaborative fuzzy clustering. *Pattern Recognition Lett.*, 23, 675–686.

# Slimming Down a High-Dimensional Binary Datatable: relevant Eigen-Subspace and Substantial Content

Alain Lelu

Université de Franche-Comté, LASELDI & LORIA  
30 rue Mégevand, 25030 Besançon cedex, France, [alain.lelu@univ-fcomte.fr](mailto:alain.lelu@univ-fcomte.fr)

**Abstract.** Determining the number of relevant dimensions in the eigen-space of a data matrix is a central issue in many data-mining applications. We tackle here the sub-problem of finding the “right” dimensionality of a type of data matrices often encountered in the domains of text or usage mining: large, sparse, high-dimensional binary datatables. We present here the application of a randomization test to this problem. We validate our approach first on artificial datasets, then on a real documentary data collection, i.e. 1900 documents described in a 3600 keywords dataspace, where the actual, intrinsic dimension appears to be 28 times less than the number of keywords - an important information when preparing to cluster or discriminate such data. We also present preliminary results on the problem of clearing the datatable from non-essential information bits.

**Keywords:** randomization test, dimensionality reduction, data reconstitution, power-law distribution

## 1 Introduction

Determining the number of relevant dimensions in the eigen-space of a data matrix is a central issue in many data-mining applications. We tackle here the sub-problem of finding the “right” dimensionality of a type of data matrices often encountered in the domains of text or usage mining, or in a number of biological applications, generally displaying “Zipfian” power-law distributions (Newman (2005)): large, sparse, high-dimensional binary datatables, for which the assumptions underlying the state-of-the-art techniques such as the Catell’s scree-break heuristics (Cattell (1966)) or more recent model-based parametric tests (Bouveyron et al. (2009)) do not hold. Resampling tests, such as bootstrap (Efron (1981)) are akin to delineate the variability of a feature of interest, e.g. the positions of projected datapoints in a chosen factor plane (Lebart (2007)). Our problem is different, in that we try to determine which eigen-subspace of a binary data matrix bears the relevant information, and which extra eigen-dimension does not, due to the sole effect of noise, or distributions of the marginal sums. In this prospect, the general non-parametric solution we are interested in has to rest on comparing the

successive major eigenvalues of the original matrix to their counterpart in (at best) all the possible binary matrices endowed with the same row and column marginal sums (i.e. generalizing the *exact test* of Fisher (1936)), or, as it proves generally unfeasible, in a sample of these matrices (*randomization test*, Manly (1997)). Cadot (2005, 2006) has set up such a solution for any measure issued from a binary datatable, in her Tournebool randomization test, e.g. extracting significant graph edges between variables or between individuals (Lelu & Cadot (2010)).

In section 2 we will briefly recall the TourneBool process for generating randomized versions of the original datatable, and apply it to test its successive dominant eigenvalues against the null hypothesis - not being greater than expected from randomness. In section 3 we will describe how to generate artificial binary data endowed with two major characteristics of real-world binary data: Zipfian distribution of the variables, and intermingled clusters. We will successfully apply our test to an instance of such datatables. In section 4 we will describe a set of real-life bibliographic data, and will test it, resulting in 125 significant eigenvalues in this 1920 documents and 3600 keywords dataset, at the 99% significance threshold. In section 5 we will present an early empirical insight into the problem of the optimal binary reconstruction of a binary datatable, starting from its sole significant eigen-elements, which suggests a filtering process for “denoising”, “slimming down” such table, or strongly filtering the variables, while keeping the meaningful substance of the table unaltered.

## 2 Randomization process and test

The comparison with full-scale random simulations is now feasible, and is an alternative to the traditional comparisons with asymptotic theoretical statistic distributions. Noise may be added to the original datatable (bootstrap and Jackknife methods), or purely random tables may be generated, submitted to the same structural constraints as the original one. In this way, one may generate the random versions starting from the original database itself, by a sequence of elementary transformations keeping the row and column margins constant. This is the direction taken by the TourneBool method and test: a method for generating random versions of a binary datatable with prescribed margins, and the ensuing test for any measurement operated on the original matrix against the null hypothesis.

*Generating the Randomized Matrices.* Cadot (2005) presented a permutation algorithm based on rectangular flip-flops, incorporating a monitored convergence of the algorithm. Its theoretical legitimation can be found in Cadot (2006), based on the original notion of cascading flip-flops: the author has shown that any Boolean matrix can be converted into any other one with the same margins in a finite number of such cascades. These cascading flip-flops are themselves compositions of elementary rectangular exchanges, or



*flip-flops.* These flip-flops preserve the irreducible background structure of the datatable, but break up the meaningful links specific to a real-life data table. Getting rid of the background structure enables the method to process any type of binary data, both (1) taking into account the marginal distributions, (2) doing this without any need to specify the statistical models of these distributions. The number of rectangular flip-flops is controlled by two Hamming distance measures between matrices (i.e. number of cells with opposite values): 1) between the current random matrix and the one generated at the previous step, 2) between the current random matrix and the original one. The initial number of flip-flops is increased as long as these distances are growing. The value of this parameter is deemed optimal when they stabilize - in practice, about several times the number of ones in the original matrix. No bias, i.e. residual remnant of the original matrix, can be attributed then to the randomization process.

*Establishing the sequence of significant eigenvalues.* A nested test is needed, the principles of which are the following:

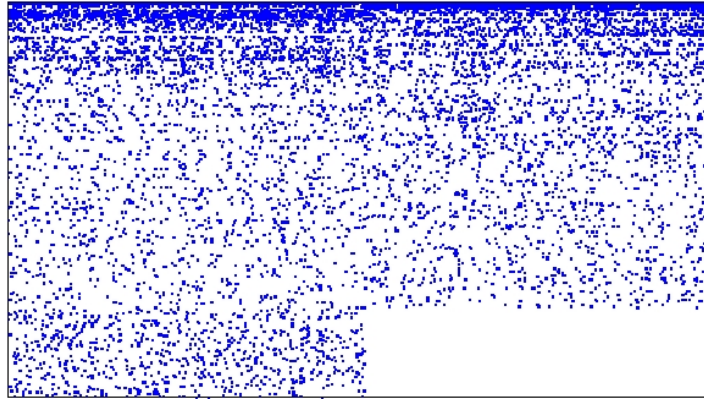
- Generate a sufficient sample ( $X_1, X_2, \dots, X_p$ ) of randomized versions of the original matrix  $X_0$  (e.g. 200 matrices).
- Extract the full sequence of singular values of  $X_0$ , in decreasing order.
- For each  $k$ -order eigen-space, starting from  $k = 1$ , compare the  $k$ -th singular value of  $X_0$  to the set of corresponding  $k$ -th singular values in the sample: if the current singular value  $\lambda_k(0)$  is greater than or equal to the randomized one located at the significance threshold (e.g. than the third one at the 99% threshold), it is deemed significantly diverging from randomness, and the algorithm goes on with  $k = k + 1$ .

When the algorithm stops, the value  $k$  is the dimension of the relevant eigenspace.

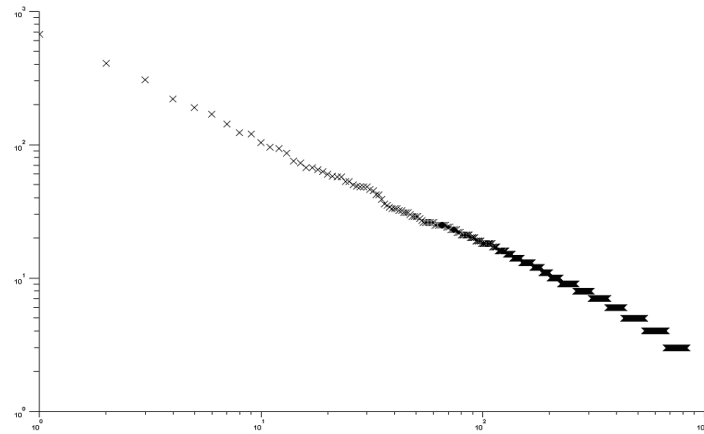
### 3 Validating on artificial data

Generating artificial datasets being a somehow unsubstantial and arbitrary task, we will focus on trying to reproduce two characteristics that stand out from our experience: 1) large-scale sparse datatables with binary features tend to exhibit a power-law distribution of their feature counts, as has been observed in many application domains, such as text mining ; 2) cluster structures are by no way all-or-none phenomena: they rather amount to progressive, fuzzy memberships around dense data-cores. In other words, clusters are generally intricate, entangled, and by no way orthogonal.

*Data generation:* We will first build such intermingled clusters in the simplest case of two clusters, by generating a one-cluster table, e.g. appending

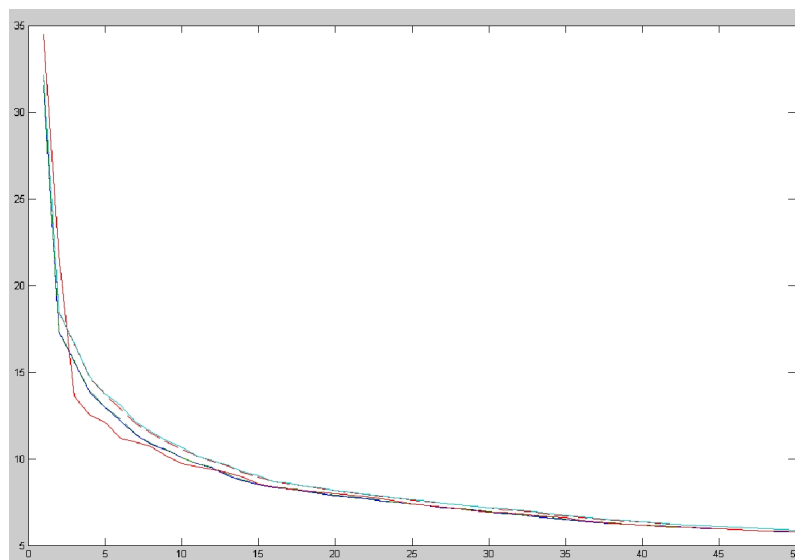


**Fig. 1.** A plot of the 2-cluster artificial data. Horizontally: the 1500 “documents” split into two clusters. Vertically: the 836 “keywords”.



**Fig. 2.** Characterizing the records in the 2-cluster artificial data. Vertically: the frequency count of each “keyword”. Horizontally: their ranks. The coordinates are log-log.

a full  $(750, 800)$  “ones” matrix and a full  $(750, 660)$  “zeros” one, then creating another  $(750, 1460)$  matrix by randomly permuting the columns, and eventually stacking the two matrices into a  $(1500, 1466)$  one. The second step consists in “morphing” this matrix so as to fit into prescribed relative column and row sum profiles (e.g. a power-law distribution for the column sums, and a binomial one for the row sums): the process of alternating a global stretching or expanding for each column vector so as to fit to the corresponding prescribed sum profile, with the same for the row vectors, lets the transformed datatable converge to a real positive matrix embedding a (distorted) memory of the initial structure.



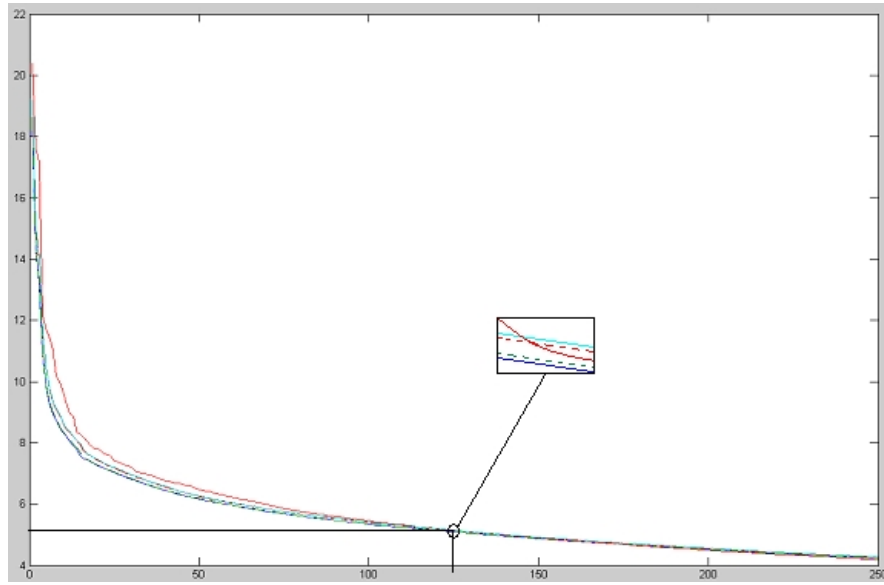
**Fig. 3.** Scree-plot of the singular values for the 2-cluster artificial data (in red). The dashed lines delimit the confidence interval, the solid lines delimit the minimum to maximum variation interval.

The last step consists in turning this table binary, first by normalizing it (i.e. dividing by its maximum value), then by considering each value as a probability for drawing a value “one”; the resulting (1500, 1460) table comprises many empty columns, or columns summing to 1 or 2; in a final cleansing process, we remove these columns for the sake of preventing side effects, and we now yield a (1500, 836) binary matrix  $X_0$  (see figure 1) with a visually convincing power-law distribution of the column sums (see figure 2).

*Eigenspace test:* Figure 3 shows the “scree-plot” of the 50 first eigenvalues of  $X_0$ , compared to the plot of the 99% confidence interval of its 200 randomized clones. As jumps out from the figure, the only two first singular values dominate their confidence intervals, emphasizing the 2-cluster intertwined structure.

#### 4 Relevant eigen-subspace of a real-world binary datatable

*Origin and characteristics of the data:* An excerpt of the Pascal bibliographic database, edited by CNRS/INIST, and spanning one year of research activity in the french Lorraine region, has been set up for diverse methodological



**Fig. 4.** Scree-plot of the singular values for the Lorraine data (in red). The dashed lines delimit the 99% confidence interval, the solid lines delimit the minimum to maximum observed variation interval.

evaluation tasks (Ghribi et al. (2010)), and will soon be publicly available<sup>1</sup>. We have chosen these data as a “not too large, but sufficient” sample of the very common documentary or text type of data. It consists of 1920 records manually indexed with 3557 keywords of frequency greater than one, resulting in a mean value of 5.6 keywords per document. As could be expected, the keywords’ occurrences follow a typical power-law distribution.

*Intrinsic dimension of the datatable:* Having generated 200 randomized versions of the original matrix with the Tournebool algorithm, we have applied the above-described test for assessing the 250 first singular values. As can be noticed in the scree-plot of Figure 4, it appears that, at the 99% threshold, the 125 first singular values significantly depart from the confidence interval due to randomness - thus establishing to 125 the dimension of the significant eigen-subspace, and suggesting further operations in this reduced dataspace without any loss of relevant information: e.g. similarity measures, as those implied in Latent Semantic Analysis (Deerwester et al. (1990)), or cluster axoids seeking<sup>2</sup> (Lelu(1994)). As a subsidiary observation, one may also no-

<sup>1</sup> We are indebted to INIST and Pascal Cuxac for having put these data at our disposal.

<sup>2</sup> In this case, as no cluster axoids can be colinear to another one by definition, the number of clusters cannot be lesser than the intrinsic dimension of the data matrix.

tice in figure 4 that the visual “scree-break” criterion of Cattell (1966) seems inoperative in such high-dimensional data, though effective in the case of our artificial example in section 3.

On the computational side, the two most time-demanding phases, i.e. the creation of the randomized matrices and the extraction of a significant proportion of their singular values (250 chosen here), are not far from proportional to their number and to the number of ones in each one. This is no problem for our 200 matrices filled with 10,700 ones, as the total running time of these modules has not exceeded ten minutes on a 2.7 GHz CPU, 4 Gb RAM, desktop computer.

## 5 Substantial content of a binary datatable: an empirical approach

The SVD reconstitution of the data writes:

$$X_0 = UDV'$$

where  $U$  and  $V$  are the matrices gathering respectively the row and column singular vectors, and  $D$  is the diagonal matrix of the singular values.

The rank- $k$  reconstitution of the data writes:

$$X_0^k = U_k D_k V_k'$$

We have computed the  $X_0^{125}$  reconstitution of the data in the relevant eigenspace. The distribution of the values in the cells is very assymetric, with more than 3 million values in the  $]0; .1]$  interval, 2600 values in the  $]0.9; 1]$  interval, and a clear minimum in the  $] .5; .9]$  range; hence, the empirical idea of thresholding these values for reconstructing a binary matrix. And for each value of the threshold, a coefficient of fit between the real data and the reconstructed ones can be computed. We have chosen the well-known *f-score* coefficient, i.e. the harmonic mean between the *precision* and *recall* of the reconstitution: the maximum value 0.803 corresponds to the .3 threshold. The resulting binary table has lost  $10,754 - 10,138 = 616$  *one* values, compared to the original matrix. We may conclude that these 616 values are pure noise and might be discarded from any further analysis. The same thresholding process might be applied for discarding more and more values, depending on the desired sharpness of this analysis. A progressive filtering of the binary features may also ensue.

## 6 Conclusions, perspectives

The use of the Tournebool randomization test appears to offer a satisfactory, if not rigorous, solution for establishing the intrinsic dimension of a large,

sparse, binary matrix, useful e.g. for fixing the relevant number of components in a Latent Semantic Analysis, or a lower bound to the “real” number of clusters to be pulled out. More has to be worked out on the subject of reconstructing the “core bits” of the data matrix, on which subject we hope to have brought a first contribution.

## References

- BAVAUD, F. (1998): *Modèles et données : une introduction à la Statistique uni-, bi- et trivariée*. L'Harmattan.
- BOUYEYRON C., CELEUX G. and GIRARD S. (2009): Intrinsic Dimension Estimation by Maximum Likelihood in Probabilistic PCA. In: *PREPRINT - December 10, 2009 1 (HAL 00440372)*
- CADOT, M. (2005): A Simulation Technique for Extracting Robust Association Rules. *CSDA 2005*.
- CADOT, M. (2006): *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Ph.D. thesis, Université de Franche-Comté.
- CATTELL, R. B. (1966). "The Scree Test for the Number of Factors." *Multivariate Behavioral Research* 1(2), 245-276.
- DEERWESTER S., DUMAIS S., FURNAS G. W., LANDAUER T. K., HARSHMAN R. (1990): Indexing by Latent Semantic Analysis. In: *Journal of the American Society for Information Science* 41 (6) 391-407.
- EFRON, B. (1981): Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* 68, 589-599.
- FISHER, R. (1936): The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 179-188.
- JENSEN, D. and COHEN, P. (2000): Multiple Comparisons in Induction Algorithms. *Machine Learning*, 309-338.
- LEBART, L. (2007): Which bootstrap for principal axes methods ? In: P. Brito et al. (eds): *Selected Contributions in Data Analysis and Classification*., Springer, 581-588.
- LELU, A. (1994): Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets. In: DIDAY E., LECHEVALIER Y. & al. (eds): *New Approaches in Classification and Data Analysis*, 241-248 Springer-Verlag , Berlin.
- LELU A., CADOT M. (2010): Statistically valid links and anti-links between words and between documents: applying TourneBool randomization test to a Reuters collection. In: Ritschard G. & Studer M. (eds). *Advances in Knowledge Discovery and Management (AKDM)*, 327-344 Springer-Verlag , Berlin, in press.
- MANLY, B. (1997): *Randomization, Bootstrap and Monte Carlo methods in Biology*. Chapman and Hall & CRC.
- NEWMAN, M. (2005): Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 323-351.
- PRESS, J. (2004): The role of Bayesian and frequentist multivariate modeling in statistical Data Mining. *Statistical Data Mining and Knowledge Discovery*, 1-14.

# Comparing Two Approaches to Testing Linearity against Markov-switching Type Non-linearity

Jana Lenčuchová, Anna Petričková and Magdaléna Komorníková

Department of Mathematics, Faculty of Civil Engineering, Slovak University of  
Technology Bratislava  
Radlinského 11, 813 68 Bratislava, Slovakia,  
*lencuchova@math.sk, petrickova@math.sk and magda@math.sk*

**Abstract.** In this paper we discuss an alternative approach to testing linearity against Markov-switching type non-linearity. We aim to avoid the classic testing via the likelihood ratio test, which doesn't have a standard distribution. That's why time-consuming simulations must be carried out. We suggest the classical test to be substituted by using Hamilton's dynamic specification test for validity of Markov assumptions. The same idea will be applied to testing the remaining non-linearity to compare 2-regime with 3-regime models. These two approaches will be confronted by being demonstrated on some selected time series, e.g. Slovak macro-economic indicators and some exchange rates.

**Keywords:** Markov-switching model, Markov assumptions, dynamic specification test, testing non-linearity, testing remaining non-linearity

## 1 Introduction

In two last decades we have noticed a great progress in the Markov-switching modeling. It has been shown that these models have excellent description properties. Model parameters can attain different values depending on a regime they are in. Simply, one regime can represent an expansion and another one recession, for instance economic expansion or recession. Such dynamic behavior is typical for macroeconomic and financial time series. Changes in model parameters are caused by occasional and rare events like financial crisis, wars, political developments, natural disasters and so on.

The Markov-switching model belongs to the class of the regime switching models, where the regime switching is determined by unobservable variables. So we suppose that the regime or "the state" which occurs at time  $t$  is unobserved and determined by a random variable  $s_t$ . If there are  $N$  possible regimes in the model, then the random variable  $s_t$  can attain values from set  $\{1, 2, 3, \dots, N\}$ . Hamilton (1989) describe a stochastic process  $s_t$  as following

a first-order Markov process. It means that the probability of the regime at time  $t$  depends only on the regime at time  $t - 1$ :

$$Pr(s_t = j | s_{t-1} = i, s_{t-2} = k, \dots) = Pr(s_t = j | s_{t-1} = i) = p_{ij}, \quad (1)$$

$t = 1, \dots, T$ , where  $T$  is length of time series and  $i, j, k = 1, \dots, N$ , where  $N$  is number of regimes. Transition probabilities  $(p_{ij})_{i,j=1,\dots,N}$  represent the probability that the regime  $i$  will be followed by the regime  $j$ . We assume:

$$p_{i1} + p_{i2} + \dots + p_{iN} = 1, \quad i = 1, \dots, N. \quad (2)$$

Because of the Markov property, the complete probability distribution of the state of a Markov chain is defined by the initial distribution  $\pi_i = Pr(s_t = i)$  and the state transition probability matrix  $P = (p_{ij})_{i,j=1,\dots,N}$ . We will consider further an autoregressive Markov-switching model of the form

$$y_t = \phi_{0,s_t} + \phi_{1,s_t}y_{t-1} + \dots + \phi_{q,s_t}y_{t-q} + \epsilon_t, \quad s_t = 1, \dots, N, \quad (3)$$

where  $\epsilon_t \sim N(0, \sigma^2)$ . More details about the parameter estimation are in Hamilton (1990,1994).

## 2 Classical testing linearity against Markov-switching type non-linearity

The general non-linear modeling procedure was described by Granger(1993) and one of the steps we should follow in the modeling of time series by non-linear models is to check whether our examined time series has a non-linear character at all.

The classical approach to the testing linearity is proceeded by the likelihood ratio test (Hansen(1992)). The null hypothesis stands for a suitability of a linear model and alternative stands for a 2-regime Markov-switching model:

$$H_0 : \varphi_1 = \varphi_2,$$

where  $\varphi_1, \varphi_2$  represents AR coefficients of a Markov-switching model for  $i = 1, 2$  against

$$H_1 : \phi_{i,1} \neq \phi_{i,2}$$

for at least one  $i \in \{0, 1, 2, \dots, q\}$ . The statistic of the likelihood ratio test has the form:

$$L = L_{MSW} - L_{AR}, \quad (4)$$

where  $L_{MSW}$  and  $L_{AR}$  are loglikelihood functions for the corresponding Markov-switching model and AR model.

Calculating this test statistic is a problem in the sense of the time difficulty. Indeed, Hansen (1992) proved that (4) has non-standard probabilistic distribution. So we have to carry out simulations to gain critical values to



determine the significance of the test statistic. To realize simulation experiment one needs to generate a large number (at least 5000) of artificial time series  $y^*$  according to the model that holds under the null hypothesis. Then one needs to estimate parameters by AR and Markov-switching model for each artificial time series, calculate relevant loglikelihood functions and finally critical values from (4). Simulations must be done for each examined time series and for each model order  $q$  distinctly.

### 3 Alternative approach to testing

Due to time-consuming classical test we suggest to use a test proposed by White (1987) using conditional moment tests of Newey (1985) and Tauchen (1985), derived score functions for Markov-switching model and a dynamic misspecification test for validity of Markov assumptions proposed by Hamilton (1996).

#### 3.1 Score function

Score function of  $t$ th observation  $\mathbf{h}_t(\boldsymbol{\theta})$  is defined as a derivation of the logarithm of the conditional probability likelihood function with respect to the parameter vector  $\boldsymbol{\theta}$ :

$$\mathbf{h}_t(\boldsymbol{\theta}) \equiv \frac{\partial \log f(y_t | \Omega_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \quad (5)$$

where  $\boldsymbol{\theta}$  represents parameter vector of the model,  $\Omega_{t-1}$  represents observation history and  $f(y_t | \Omega_{t-1}; \boldsymbol{\theta})$  is the probabilistic density of  $y_t$  conditional on the  $\Omega_{t-1}$ .

Suppose that the density of an observable variable  $y_t$  conditional on the random variable  $s_t$  and the history of observations for the basic Markov-switching model (3) is normal, i.e., it has the form:

$$f(y_t | s_t = j, \Omega_{t-1}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_t - \boldsymbol{\varphi}_j' \mathbf{X}_t)^2}{2\sigma^2}\right\}, \quad (6)$$

where  $\boldsymbol{\varphi}_j = (\phi_{0,j}, \phi_{1,j}, \dots, \phi_{q,j})'$  is the vector of AR coefficients for the regime  $j$ ,  $\mathbf{X}_t = (1, y_{t-1}, \dots, y_{t-q})'$ ,  $\Omega_{t-1} = (y_{t-1}, y_{t-2}, \dots)$  and  $\sigma^2$  is the residual variance of the model. The parameter vector  $\boldsymbol{\theta}$  consists of:

$$\boldsymbol{\theta}' = (\boldsymbol{\alpha}', \mathbf{p}'), \quad (7)$$

where  $\boldsymbol{\alpha}' = (\boldsymbol{\varphi}_1', \boldsymbol{\varphi}_2', \dots, \boldsymbol{\varphi}_N', \sigma^2)$  represents parameter vector in conditional density for all regimes. The vector  $\mathbf{p}$  is the vector of transition probabilities  $(p_{ij})_{i,j=1,\dots,N}$  with omitting redundant parameters  $p_{iN}$  for  $i = 1, \dots, N$ , which can be expressed by the remaining parameters, see (2).

Score function for the Markov-switching model (3) was derived by Hamilton(1996):

$$\begin{aligned} \frac{\partial \ln f(y_t | \Omega_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} &= \sum_{j=1}^N \frac{\partial \ln f(y_t | \mathbf{X}_t, s_t = j; \boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} Pr(s_t = j | \Omega_t) + \\ &+ \sum_{\tau=1}^{t-1} \sum_{s_\tau=1}^N \frac{\partial \ln f(y_\tau | \mathbf{X}_\tau, s_\tau = j; \boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} \{Pr(s_\tau | \Omega_t) - Pr(s_\tau | \Omega_{t-1})\} \end{aligned} \quad (8)$$

for  $t = 1, 2, \dots, T$ , where  $T$  is time series length.

$$\begin{aligned} \frac{\partial \ln f(y_t | \Omega_{t-1}; \boldsymbol{\theta})}{\partial p_{ij}} &= p_{ij}^{-1} Pr(s_t = j, s_{t-1} = i | \Omega_t) - p_{iN}^{-1} Pr(s_t = N, s_{t-1} = i | \Omega_t) + \\ &+ p_{ij}^{-1} \left\{ \sum_{\tau=2}^{t-1} [Pr(s_\tau = j, s_{\tau-1} = i | \Omega_t) - Pr(s_\tau = j, s_{\tau-1} = i | \Omega_{t-1})] \right\} - \\ &- p_{iN}^{-1} \left\{ \sum_{\tau=2}^{t-1} [Pr(s_\tau = N, s_{\tau-1} = i | \Omega_t) - Pr(s_\tau = N, s_{\tau-1} = i | \Omega_{t-1})] \right\} + \\ &+ \sum_{s_1=1}^N \frac{\partial \ln Pr(s_1; \mathbf{p})}{\partial p_{ij}} [Pr(s_1 | \Omega_t) - Pr(s_1 | \Omega_{t-1})] \end{aligned} \quad (9)$$

for  $i = 1, 2, \dots, N$ ,  $j = 1, 2, \dots, N - 1$  and  $t = 2, \dots, T$ . For  $t = 1$

$$\frac{\partial \ln f(y_1 | \Omega_0; \boldsymbol{\theta})}{\partial p_{ij}} = \sum_{s_1=1}^N \frac{\partial \ln Pr(s_1; \mathbf{p})}{\partial p_{ij}} Pr(s_1 | \Omega_1). \quad (10)$$

For more details about calculating (8), (9) and (10) see Hamilton (1996).

### 3.2 Newey-Tauchen-White test

Tests for serial correlation of the scores were proposed by White(1987). He applied conditional moment tests of Newey(1985) and Tauchen(1985). To carry out this test we need to construct  $(k \times 1)$  vector  $\mathbf{c}_t(\boldsymbol{\theta})$  consisting of elements of  $(m \times m)$  matrix  $[\mathbf{h}_t(\boldsymbol{\theta})][\mathbf{h}_{t-1}(\boldsymbol{\theta})]'$ , where  $m$  is number of parameters. We try to confirm zero mean when we evaluated this matrix with true parameter  $\boldsymbol{\theta}_0$ , in other words we try to confirm independence of the score at date  $t$  for the score at the  $t - 1$ , so we cannot predict  $\mathbf{h}_t(\boldsymbol{\theta})$  by help of  $\mathbf{h}_{t-1}(\boldsymbol{\theta})$ . If the model is specified with correct parameter, then test statistic has  $\chi^2(k)$  asymptotic distribution and the following form:

$$\left[ T^{-\frac{1}{2}} \sum_{t=1}^T \mathbf{c}_t(\hat{\boldsymbol{\theta}}) \right] \cdot \left[ T^{-1} \sum_{t=1}^T \mathbf{c}_t(\hat{\boldsymbol{\theta}}) \cdot \mathbf{c}_t(\hat{\boldsymbol{\theta}})' \right]^{-1} \cdot \left[ T^{-\frac{1}{2}} \sum_{t=1}^T \mathbf{c}_t(\hat{\boldsymbol{\theta}}) \right] \rightarrow \chi^2(k). \quad (11)$$

### 3.3 Dynamic specification test for validity of Markov assumptions

Hamilton(1996) described several specification tests. One of them is testing validity of Markov assumptions means verifying these assumptions:

$$Pr(s_t = j | s_{t-1} = i) = Pr(s_t = j | s_{t-1} = i, y_{t-1}), \quad i, j = 1, 2, \dots, N, \quad (12)$$

$$Pr(s_t = j | s_{t-1} = i) = Pr(s_t = j | s_{t-1} = i, s_{t-2} = k), \quad i, j, k = 1, 2, \dots, N. \quad (13)$$

We collect such elements to the vector  $\mathbf{c}_t(\boldsymbol{\theta})$  from the matrix  $[\mathbf{h}_t(\boldsymbol{\theta})] \cdot [\mathbf{h}_{t-1}(\boldsymbol{\theta})]'$ , which are corresponding to examined properties of the model, in this case to assumptions (12) and (13):

$$\frac{\partial \ln f(y_t | \Omega_{t-1}; \boldsymbol{\theta})}{\partial p_{ij}} \cdot \frac{\partial \ln f(y_{t-1} | \Omega_{t-2}; \boldsymbol{\theta})}{\partial \phi_{0,i}}, \quad i, j = 1, \dots, N, \quad (14)$$

$$\frac{\partial \ln f(y_t | \Omega_{t-1}; \boldsymbol{\theta})}{\partial p_{ij}} \cdot \frac{\partial \ln f(y_{t-1} | \Omega_{t-2}; \boldsymbol{\theta})}{\partial p_{ij}}, \quad i, j = 1, \dots, N. \quad (15)$$

Because of omitting redundant parameters, the vector  $\mathbf{c}_t(\boldsymbol{\theta})$  involves  $2N(N-1)$  elements and then the test statistic has  $\chi^2(2N(N-1))$  distribution, where  $N$  is number of regimes.

### 3.4 Testing linearity against Markov-switching type non-linearity

For the new testing linearity, the null hypothesis represents validity of Markov assumptions (12) and (13). When the null hypothesis is rejected, then the corresponding Markov-switching model isn't appropriate for the modeling of examined time series and this time series doesn't show Markov-switching type non-linear character. We calculate the test statistic (11) for the 2-regime model and we find out the  $p$ -value from  $\chi^2(2N(N-1))$  distribution. If  $p$ -value  $< \alpha$ , then we reject the null hypothesis and a linear model is better to use in this case.

### 3.5 Testing remaining non-linearity

If we go through the testing validity of Markov assumptions for the 2-regime model, we can test remaining non-linearity by the similar manner. We can compare the 2-regime model with a 3-regime model and find out appropriateness of the 2-regime model against alternative hypothesis about the 3-regime model. Here the test statistic (11) has different number of elements of the vector  $\mathbf{c}_t(\boldsymbol{\theta})$ , what means different number of degrees of freedom in distribution  $\chi^2(2N(N-1))$ , because  $N = 3$ . The following alternatives can arise:

- Non-rejecting the null hypothesis for a 2-regime model, but rejecting for a 3-regime model, what means appropriateness of 2-regime model.

- Non-rejecting the null hypothesis for a 2 and 3-regime model, then the model with greater  $p$ -value from testing validity of Markov assumptions is more suitable. This case we can check by other criterions, for example BIC (Bayesian Information Criterion) - better model has lower BIC or by residual dispersion, forecasting error values, results for testing autocorrelation and so on.

#### 4 Comparing results of simulations and the new testing

We try to support our theory about new testing by classical testing via simulations, which computation is described in part 2 of this paper. We are modeling 10 time series (5 selected Slovak macroeconomic indicators - GDP, consumption, real wages, inflation and unemployment; and 5 selected exchange rates - GBP, USD, PLN, HUF, CZK - all to EUR)

The biggest advantage of new testing is much lower time complexity comparing with simulations. For illustration, until new testing linearity against Markov-switching type non-linearity needed only 33.281  $s$  for the model order  $q = 5$  (length of time series was 77), the simulation was computed in 15860,28  $s$  for the same time series and order of model. Also the new testing remaining non-linearity needed 471.781  $s$  for the model order  $q = 5$  (length of time series was 77), but the simulation experiment took 43 477.3  $s$  for the same inputs. It is a really significant difference.

Concerning to new testing linearity, we've got interesting results. There was only one time series from 10, where simulation didn't confirm our suggestion for 4 orders of 5. For other time series it finished much more positively. The test via the validity of Markov assumptions was confirmed by simulation in 82%.

Since it is very time-consuming to test remaining non-linearity by simulations, we computed it only for that orders, where corresponding models has the lower BIC. The results of testing remaining non-linearity by both ways for the models with lower BIC are in Table 1.

Data	GDP	Infl.	Unempl.	Cons.	Wages	CZK	PLN	HUF	GBP	USD
New test	0.204	0.431	0.071	0.436	0.029	0.174	0.017	0.274	0.014	0.055
Simulation	0.03	0.293	0.02	0.019	0.057	0.002	0.003	0.108	0.097	0.601

**Table 1.** Comparing two approaches of testing remaining non-linearity

We can see that in 7 of 10 models was confirmed the same conclusion by both tests with significance level  $\alpha = 0.05$ . The same conclusion means that the hypotheses about a 2-regime model or a 3-regime model was confirmed. In other two cases the same conclusion was confirmed but with changed signifi-

cance level. We've got an opposite results in Inflation time series exactly the same as in the first testing linearity of Markov-switching type non-linearity.

## 5 Conclusion

In this paper we have analyzed alternative approach to testing linearity against Markov-switching type non-linearity as well as testing for remaining non-linearity for Markov-switching model. We have suggested the classic tests via simulations be substituted by using Hamilton's dynamic specification test for validity of Markov assumptions.

The new testing linearity against Markov-switching type non-linearity was confirmed in 82% cases by simulations and the new testing remaining non-linearity in 90% cases.

The biggest contribution of this testing is a significant time saving. But the new testing was verified only with 10 time series, therefore it should be continued verifying with bigger number of time series in spite of time-consuming simulations. That could confirm whether the new test is working correctly or not and with what reliability.

**Acknowledgement** The support of the grant APVV No. LPP-0111-09 is kindly announced.

## References

- HAMILTON, J.D. (1989): A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357-384.
- HAMILTON, J.D. (1990): Analysis of time series subject to changes in regime. *Journal of Econometrics* 45, 39-70.
- HAMILTON, J. D. (1994): *Time series analysis*. Princeton University Press, Princeton.
- HAMILTON, J.D. (1996): Specification testing in Markov-switching time series models. *Journal of Econometrics* 70, 127-157.
- HANSEN, B.E. (1992): The likelihood ratio test under nonstandard assumptions: testing the Markov switching model of GNP. *Journal of Applied Econometrics* 7, 61-82.
- NEWKEY, W.K. (1985): Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53, 1047-1070.
- TAUCHEN, G. (1985): Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics* 30, 415-443.
- WHITE, H. (1987): Specification testing in dynamic models. In: T. F. Bewley (Eds.): *Advances in econometrics*. Fifth world congress, Cambridge University Press, Cambridge, Vol. 2.



# Numerical Error Analysis for Statistical Software on Multi-Core Systems

Wenbin Li and Sven Simon

SimTech & IPVS, Stuttgart University  
Universitätsstr. 38, Stuttgart, Germany, (*liwn,simon*)@ipvs.uni-stuttgart.de

**Abstract.** In statistical software packages, usually no information about the numerical accuracy of the computed results is available. This leads to a risk of misinterpretation of inaccurate results. The Discrete Stochastic Arithmetic (DSA) provides estimation of numerical accuracy with respect to rounding error propagation. In this paper, the DSA is applied to a statistical package, benchmark results are presented to illustrate the effectiveness of the DSA, and parallelization approaches of the proposed method on a multi-core system are investigated for performance improvement.

**Keywords:** numerical accuracy, DSA, multi-core, parallelization

## 1 Introduction

With the advances in computer technology, the complexity of statistical problems which can be solved numerically is increasing. Due to the use of finite precision arithmetic for Floating Point (FP) numbers, each elementary FP operation (including assignment) may induce a round-off error. Consequently, the computed results are affected by the round-off error propagation.

There are many research works related to the reliability of commonly used statistical software packages with respect to numerical errors. McCullough and Heiser (2008) applied a standard set of intermediate-level accuracy tests to Excel 2007, and found it failed in three areas: statistical distributions, random number generation, and estimation. McCullough (2000) evaluated the performance of statistical software using the (American) National Institute of Standards and Technology (NIST) data sets, and mentioned that the effect of cumulative round-off error quickly degrades the numerical quality of the result. Round-off error analysis of the numerical result is necessary to reduce the risk of misinterpretation of the data and inaccurate result (Keeling and Pavur (2007), McCullough (1998), McCullough (1999)). The Discrete Stochastic Arithmetic (DSA) (Vignes (1993)) provides an approach to estimate the accuracy of the computed result during the execution of a program. The DSA is based on CESTAC method (Vignes and La Porte (1974), Vignes (1988)), and is implemented in a sequential architecture in the library CADNA (Control of Accuracy and Debugging for Numerical Applications) (Jezequel and Chesneaux (2008)).

Although the DSA is more efficient compared to traditional interval arithmetic as well as the arbitrary-high-precision approach, the multiple runs of the code are still intensive in computation time. Changing from single-core to multi-core CPUs, parallelization approaches will be investigated in this paper for performance improvement.

The rest of the paper is organized as follows: in Section 2, the basics of the DSA are briefly reviewed; in Section 3, the DSA is applied to a statistical software package, and numerical accuracy of the benchmark results is presented; in Section 4, two parallelization approaches of the DSA using multi-threading technology are discussed, and the performance as well as the scalability of the proposed parallelization methods is presented.

## 2 Principle of DSA

The DSA is based on the CESTAC method proposed in 1970s (Vignes and La Porte (1974)). The basic idea is to run the same code  $N$  times with random rounding arithmetic, which consists in choosing rounding mode at each step of FP operation either towards  $+\infty$  or  $-\infty$  with the same probability. Thereby,  $N$  samples,  $R_i$  ( $i = 1, \dots, N$ ), of the computed result are obtained with different round-off error propagations. The computed result  $\bar{R}$  is taken as the average of all samples  $R_i$ :

$$\bar{R} = \frac{1}{N} \sum_{i=1}^N R_i. \quad (1)$$

The number of significant digits of the computed result  $\bar{R}$  is defined as:

$$C_{\bar{R}} = \log_{10} \frac{\sqrt{N} \cdot |\bar{R}|}{\sigma \cdot \tau_{\beta}}, \quad (2)$$

where  $\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (R_i - \bar{R})^2$  and  $\tau_{\beta} = t_{\beta, N-1}$ , which is the critical value of  $t$ -distribution with  $N-1$  degrees of freedom. If  $C_{\bar{R}} \leq 0$ , then  $\bar{R}$  is defined as nonsignificant.

The CESTAC method is based on a first order probabilistic model:

$$R \approx r + \sum_{i=1}^n g_i(d) 2^{-p} \alpha_i, \quad (3)$$

where  $r$  is the mathematical result,  $n$  is the number of elementary FP operations,  $g_i(d)$  is coefficient depending on datapath and data,  $p$  is the wordlength of the mantissa, and  $\alpha_i$  is normalized round-off error.

The CESTAC method is proposed under two hypotheses. Hypothesis 1 states that the elementary round-off errors  $\alpha_i$ 's of the FP operations are independently centered and uniformly distributed variables. Hypothesis 2 states that the approximation of  $R$  by the first order in  $2^{-p}$  is legitimate. If both hypotheses hold then  $R_i$ 's are samples of Gaussian distribution, centered on the



exact result  $r$ . The CESTAC method consists in applying Student's test on  $\{R_i\}$ , and the number of significant digits can be obtained by calculating the confidence interval. In practice, Hypothesis 1 is of little importance because of the robustness of Student's test (Vignes (2004)). Concerning Hypothesis 2, exceptions only happen when (1) both operands of multiplication are non-significant, or (2) the divisor of division is nonsignificant. To maintain the reliability, such exceptions must be detected, which is called Self Validation (SV), and warning messages should be reported to the user.

### 3 Numerical accuracy of statistical package

In this section, the DSA is applied to a statistical package, R (v2.10.1)(Ligges (2009)), to assess its numerical accuracy. The NIST Statistics Reference Database (StRD) is used as benchmark to illustrate the effectiveness of the DSA. The StRD is a collection of data sets and certified values for assessing the accuracy of software for univariate statistics (UNIV), analysis of variance (ANOVA), linear regression (LINR), and nonlinear regression (NLINR).

With numerical error analysis based on DSA, the number of significant digits  $C_{\bar{R}}$  is estimated using Equ.(2) with  $N = 3$ . For comparison, the Logarithm of Relative Error (LRE) is taken as reference, which is defined as:

$$\lambda_{\bar{R}} = LRE = -\log_{10} (|\bar{R} - r|/|r|), \quad (4)$$

where  $\bar{R}$  is the computed result and  $r$  is the true mathematic result (the certified value from *StRD* is used). When  $r = 0$ , the Logarithm of Absolute Error is used instead:  $\lambda_{\bar{R}} = LAE = -\log_{10} (|\bar{R}|)$ .

The numerical error analysis is performed on all the computed results, while only parts of them are presented in Table 1 due to space limitation. For linear and nonlinear regression, usually multiple coefficients exist. In Table 1, under the name *coefficient*, only the numerical accuracy of coefficients which have the minimum LRE is shown.

In the experiments, all computations are in double precision with 52 mantissa bits which are approximately 15 decimal digits. However, due to round-off errors, the accuracy of the computed results comes down to only 2 digits in the worst scenario. As can be seen from Table 1, the DSA provides an exact numerical accuracy estimation in 67% of the cases. If 1 digit underestimation ( $0 \leq \lambda_{\bar{R}} - C_{\bar{R}} \leq 1$ ) is tolerable, the DSA provides a reliable estimation in 96% of the cases. Overestimation is rare ( $< 4\%$  in 412 experiments), and only overestimated by 1 digit (and a warning message concerning deviation of the hypotheses of DSA is generated). The rate of overestimation can be further reduced by increasing  $N$  (Vignes (2004)), at the expense of computation time.

Unlike other methods of numerical accuracy analysis for statistical packages (Keeling and Pavur (2007), McCullough (1998)), the DSA does not require any reference solution. So it is not restricted to the benchmark problems, but also applicable to numerical error estimation of any computed result in user's application.

UNIV	Name of benchmark	Mean		Standard Deviation		Auto-correlation		—	
		$C_{\bar{\mu}}$	$\lambda_{\bar{\mu}}$	$C_{\bar{\sigma}}$	$\lambda_{\bar{\sigma}}$	$C_{\bar{\rho}}$	$\lambda_{\bar{\rho}}$	—	—
(command in R: <i>mean, sd, acf.</i> )	Pidigits	15	15	14	14	14	14	-	-
	Mavro	15	15	13	13	13	14	-	-
	Numacc2	15	15	15	14	13	14	-	-
	Numacc3	15	15	10	10	11	10	-	-
	Numacc4	15	15	9	9	9	9	-	-
ANOVA	Name of benchmark	$SS_T$		$SS_E$		$MS_E$		F-statistic	
		$C_{SS_T}$	$\lambda_{SS_T}$	$C_{SS_E}$	$\lambda_{SS_E}$	$C_{MS_E}$	$\lambda_{MS_E}$	$C_F$	$\lambda_F$
(command in R: <i>aov.</i> )	SiRstv	12	12	12	13	12	13	12	12
	SmLs01	15	15	15	15	15	15	15	15
	SmLs04	10	10	10	11	10	11	10	10
	SmLs05	10	10	10	11	10	11	10	10
	SmLs08	4	4	2	2	2	2	2	2
LINR	Name of benchmark	Coefficient		RSD		$R^2$		F-statistic	
		$C_{\bar{c}}$	$\lambda_{\bar{c}}$	$C_{RSD}$	$\lambda_{RSD}$	$C_{R^2}$	$\lambda_{R^2}$	$C_F$	$\lambda_F$
(command in R: <i>lm.</i> )	Norris	12	13	14	14	15	15	13	14
	Pontius	12	12	12	13	15	15	12	13
	Noint1	15	15	14	14	15	15	13	14
	Wampler3	9	9	14	14	15	15	14	13
	Wampler5	5	6	15	15	14	14	14	15
NLINR	Name of benchmark	Coefficient(i)		Coefficient(j)		RSS		RSD	
		$C_{c(i)}$	$\lambda_{c(i)}$	$C_{c(j)}$	$\lambda_{c(j)}$	$C_{RSS}$	$\lambda_{RSS}$	$C_{RSD}$	$\lambda_{RSD}$
(command in R: <i>nls.</i> )	Chwirut1	7	7	7	8	11	11	11	11
	Lanczos3	5	5	5	6	9	10	10	10
	Lanczos2	5	5	6	6	7	8	8	8
	Bennett5	4	4	4	5	9	9	9	9
	Thurber	5	6	6	6	10	11	11	11
overestimation ( $C_{\bar{R}} > \lambda_{\bar{R}}$ )		3 cases out of 75 (4%) (only overestimated by 1 digit)							

$SS$ : Sum of Squares;

$SS_T$ : Between Treatment Sum of Squares;

$SS_E$ : Within Treatment Sum of Squares;  $MS_E$ : Within Treatment Mean Square;

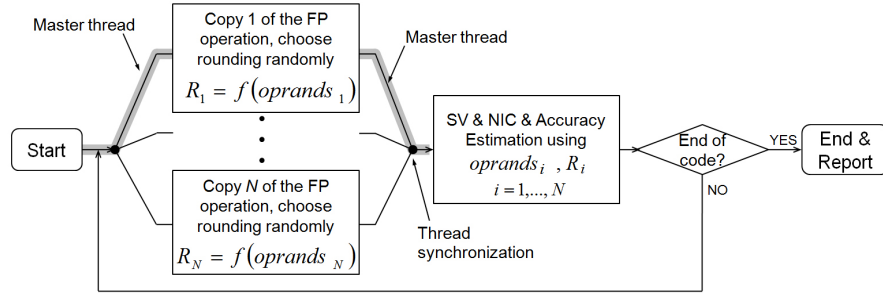
$RSS$ : Residual Sum of Squares;

$RSD$ : Residual Standard Deviation.

**Table 1.** Numerical accuracy of computed results using StRD datasets

## 4 Parallelization of DSA

Present implementations of DSA are mostly sequential in execution (e.g. CADNA), and suffer from computational bottlenecks. With the introduction of multi-core CPUs, it is possible to take advantage of this multi-core architecture to accelerate DSA. In this section, two parallelization approaches on multi-core systems are proposed, and all functionalities as those in CADNA (Jezequel and Chesneaux (2008)) are implemented: (1) to estimate the accuracy of any intermediate variable or final result; (2) to perform SV of the DSA; and (3) to perform Numerical Instabilities Checking (NIC), includ-



**Fig. 1.** Direct parallelization of the DSA

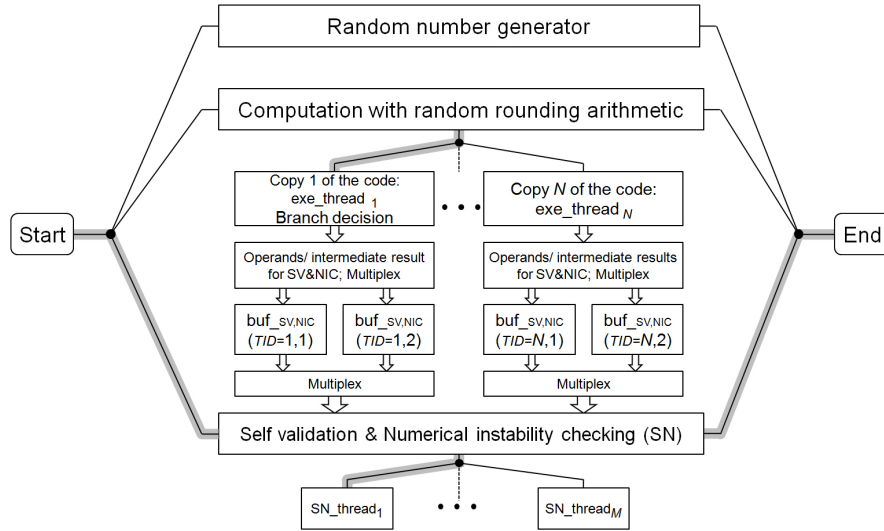
ing sudden accuracy loss (e.g. cancellation in '+' or '-') and instabilities in mathematical functions (such as EXP, LOG, etc.).

#### 4.1 Parallelization approaches

*Direct parallelization:* the direct parallelization (Fig. 1) is straightforward: each thread executes one copy of code with random rounding, and all threads are synchronized after each FP operation which requires SV or NIC. The master thread collects necessary operands as well as intermediate results, and is responsible for SV&NIC. The synchronization is also necessary before each branch-splitting statement, such as "IF(condition) THEN", in order to make a unitive decision for all the threads. However, in this architecture the synchronization overhead dramatically slows down the system performance. Figure 3 shows the time spending on threads synchronization.

*Parallelization with asynchronous SV and NIC:* actually the SV&NIC do not need to be carried out immediately after each FP operation. The necessary operands and intermediate results can be stored in a buffer, and the SV&NIC can be performed asynchronously with the execution of the code. Let us consider a parallel architecture shown in Figure 2. This architecture has 3 teams of threads.

- Thread-team 0 generates random numbers which are used before each FP operation to change the rounding mode.
- Thread-team 1 is composed of  $N$  execution-threads, each executes one copy of the code with random rounding. When SV&NIC are required, every execution-thread writes the necessary operands or intermediate results into a temporary buffer  $BUF_{SV\&NIC,TID,mux}$ , where  $TID$  is the ID of the execution-thread inside the current team, and  $mux$  is either 1 or 2. If  $BUF_{SV\&NIC,TID,1}$  (resp.  $BUF_{SV\&NIC,TID,2}$ ) is full, execution-thread writes to  $BUF_{SV\&NIC,TID,2}$  (resp.  $BUF_{SV\&NIC,TID,1}$ ) instead. When all  $BUF_{SV\&NIC,i,1}$  (resp.  $BUF_{SV\&NIC,i,2}$ ),  $i = 1, \dots, N$ , are full, thread-team 2 starts to perform SV&NIC using data stored in  $BUF_{SV\&NIC,i,1}$  (resp.  $BUF_{SV\&NIC,i,2}$ ),  $i = 1, \dots, N$ .



**Fig. 2.** Parallelization with asynchronous SV&NIC

- Thread-team 2 is responsible for SV&NIC, and generating warning messages when an exception is detected. It can be splitted into  $M$  parallel threads, where thread  $SN_{id}$  is responsible for SV&NIC using data in  $BUF_{SV\&NIC,i,mux} \left( SN_{id} \cdot \frac{BUF\_SIZE}{M} : (SN_{id} + 1) \cdot \frac{BUF\_SIZE}{M} - 1 \right)$ .

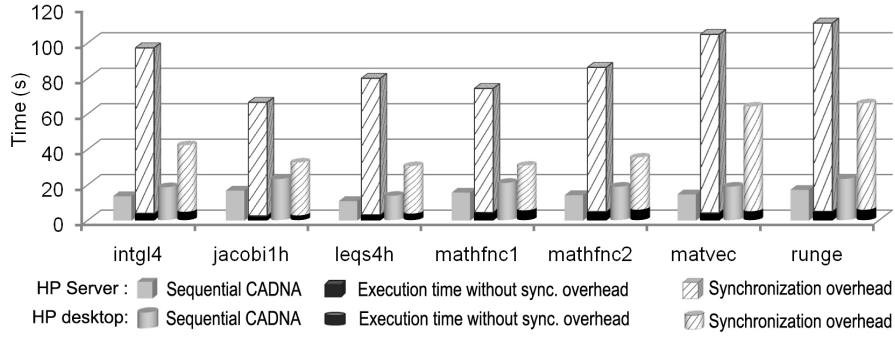
Every execution-thread must go through exactly the same datapath, so before branching splitting statement such as "IF(condition) THEN", all the execution-threads in thread-team 1 should be synchronized, and a unitive decision is made. This unitive decision is then broadcasted to all execution-threads, so that every execution-thread switches to the same branch, and continues the execution.

## 4.2 Benchmark results

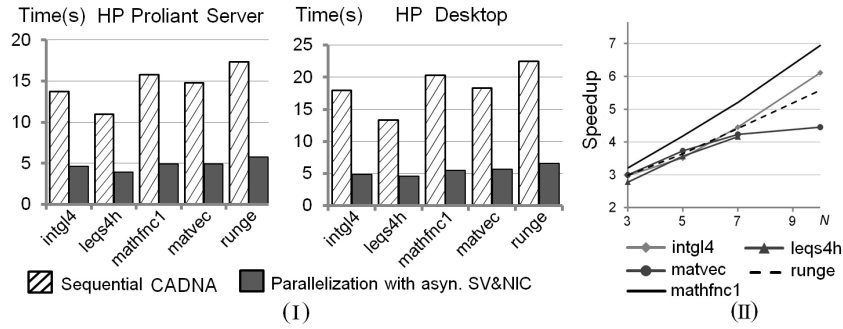
The fortran benchmarks from N.Tajima's collection (Tajima (1992)) are used to explore the efficiency of the two parallelization methods. A *HP* quad-core desktop and a *HP* Proliant server (Table. 2) are chosen as test bench.

	HP Quad-core Desktop	HP Proliant DL785GS Server
CPU	Quad-core Q6600@2.4GHz	8 Quad-core Opteron8300@2.3GHz
Memory	2 DIMM slots, DDR2 3.24GB of total memory @ 800 MHz	48 DIMM slots, DDR2 272GB of total memory @ 533 MHz

**Table 2.** Hardware platforms



**Fig. 3.** Performance of the direct parallelization.



**Fig. 4.** Performance of the parallel DSA with asynchronous SV&NIC.

In the first experiment, the performance of the direct parallelization is measured. Figure 3 plots the execution time when  $N = 3$ , on both of the two hardware platforms. As can be seen from the results, the synchronization overhead decreases the overall performance dramatically. With this architecture, acceleration is only possible when SV&NIC are not required.

In the second experiment, the parallelization method with asynchronous SV&NIC is used. Figure 4(I) plots the execution time of the benchmark problems when  $N = 3$ . The choice of  $M$  depends on  $N$  as well as hardware platforms. In this experiment, it is chosen as the minimum number so that it won't bring significant performance gain (e.g.  $> 5\%$ ) by continually increasing  $M$ . On *HP* desktop,  $M = 1$  is used, while on *HP* server,  $M = 2$ . To investigate the scalability of the proposed parallelization method, speed-up is measured at different values of  $N$ . The results are plotted in Figure 4(II).

## 5 Conclusion

In this work, the numerical accuracy of a statistical package is assessed using the Discrete Stochastic Arithmetic (DSA). Experiments show that even with

double precision, the accuracy of the computed results comes down to 2 digits in several problems, which is unacceptable in most applications. With the DSA, the reliability checking based on numerical error analysis is no longer restricted to certain benchmark problems. It is possible to estimate the accuracy of any computed result of user's application without the requirement of reference solutions, and thus avoid the risk of misinterpretation of inaccurate results. However, along with its effectiveness and reliability in numerical error analysis, the DSA suffers from computational bottlenecks due to multiple runs of the code with random rounding. For acceleration of the DSA, parallelization approaches on multi-core systems are investigated. The proposed method takes benefit from the increasing parallel computational power of multi-core CPUs, and shows an almost linear scalability. With parallelization, a speedup up to 3.3 for  $N = 3$ , 4.2 for  $N = 5$ , and 7 for  $N = 10$  is achieved, compared to the sequential implementation running on the same platform.

## References

- Jezequel, F. and Chesneaux, J.-M. (2008): CADNA: a library for estimating round-off error propagation. *Computer Physics Communications*, 178(12), 933-955.
- Keeling, K. B. and Pavur, R. J. (2007): A comparative study of the reliability of nine statistical software packages. *Computational Statistics & Data Analysis* 51(8), 3811-3831.
- Ligges, U. (2009): *Programmieren mit R*, 3rd edition. Springer-Verlag, Heidelberg.
- McCullough, B. D. (1998): Assessing the reliability of statistical software: part I. *The American Statistician* 52(4), 358-366.
- McCullough, B. D. (1999): Assessing the reliability of statistical software: part II. *The American Statistician* 53(2), 149-159.
- McCullough, B.D. (2000): Experience with the StRD: application and interpretation. *Computing Science and Statistics* 31, 16-21.
- McCullough, B. D. and Heiser, D. A. (2008): On the accuracy of statistical procedures in Microsoft Excel 2007. *Computational Statistics & Data Analysis* 52(10), 4570-4578.
- Vignes, J. and La Porte, M. (1974): Error analysis in computing. *Proceedings IFIP Congress*, 610-614.
- Vignes, J. (1988): Review on stochastic approach to round-off error analysis and its applications. *Mathematics and Computers in Simulation* 30(6), 481-491.
- Vignes, J. (1993): A stochastic arithmetic for reliable scientific computation. *Mathematics and Computers in Simulation* 35(3), 233-261.
- Vignes, J. (2004): Discrete stochastic arithmetic for validating results of numerical software. *Numerical Algorithms* 37 (1-4), 377-390.
- Tajima, N. (1992) : FORTRAN benchmark tests.  
<http://serv.apphy.fukui-u.ac.jp/~tajima/bench/index.e.html>

# Sparse Bayesian Hierarchical Model for Clustering Problems

Heng Lian<sup>1</sup>

Division of Mathematical Sciences  
School of Physical and Mathematical Sciences  
Singapore 637371  
Singapore, *henglian@ntu.edu.sg*

**Abstract.** Clustering is one of the most widely used procedures in the analysis of microarray data, with the goal of discovering cancer subtypes based on observed heterogeneity of genetic marks between different tissues. It is well-known that in such high-dimensional settings, the existence of many noise variables can overwhelm the few signals embedded in the high-dimensional space. We propose a novel Bayesian approach based on Dirichlet process with a sparsity prior that simultaneously performs variable selection and clustering, and also discover variables that only distinguish a subset of the cluster components. Unlike previous Bayesian formulations, we use Dirichlet process (DP) for both clustering of samples as well as for regularizing the high-dimensional mean/variance structure. To solve the computational challenge brought by this double usage of DP, we propose to make use of a sequential sampling scheme embedded within Markov chain Monte Carlo (MCMC) updates to improve the naive implementation of existing algorithms for DP mixture models. Our method is demonstrated on a simulation study and illustrated with the leukemia gene expression dataset.

**Keywords:** Dirichlet process, Markov chain Monte Carlo, sequential sampling, sparsity prior

## 1 Introduction

Clustering is one of the most widely used procedures in the analysis of microarray data. It has been used, for example, for cancer subtype discovery (Golub et.al. 1999). Technological advances over the last decade on microarrays have made possible simultaneous investigation of thousands of genes that potentially characterize and distinguish previously known or unknown cancer subtypes. Although obviously not all the arrayed genes possess discriminative power for different cancer subtypes, if fewer genes are used, the procedure might fail to distinguish between some of the subtypes. In this context, we generally treat the majority of genes that do not have differential expressions for different samples as noise variables and the genes that are informative about the cancer subtypes will be singled out for further biological investigations. Also, because of the cost of arraying the transcripts, this is a typical “large  $p$ , small  $n$ ” problem that has attracted much attention recently.

Among many classes of clustering procedures, the model-based approach (Banfield and Raftery 1993; Fraley and Raftery 2002), assuming the data come from a mixture of distributions, has the advantage of permitting principled statistical inferences compared to other procedures based largely on heuristics, such as k-means. This is especially important in our case where inferences should be made on the selected variables as well as on clustering structure.

In this paper, we propose a Bayesian model for simultaneous clustering and variable selection via DP mixture as well. Our formulation is based on the mean shift model (Hoff 2006). However, we use a novel hierarchical sparsity prior similar to that of Lukas et.al. (2006) which can improve separation of significant signals from noise variables and thus can lead to reduced false discoveries of uninformative noise variables. Also, we use a Dirichlet process shrinkage approach for both high-dimensional mean and variance that outperforms shrinkage using a non-DP prior, typically with normal distribution for mean and inverse-Gamma distribution for variance. Because of this double usage of Dirichlet process, both for sample clustering and for covariate shrinkage, the direct implementation of standard DP algorithms available in the literature becomes very inefficient. We solve this problem by utilizing an embedded sequential sampling step as the proposal distribution in the Markov chain Monte Carlo (MCMC) iterations. In the next section, we formulate our model using the sparsity prior. Section 3 includes a simulation study as well as an application to the leukemia gene expression data. We conclude the article with a brief discussion in Section 4.

## 2 Model Formulation

We consider the case where the clusters differ from each other only in terms of their respective means for some of the attributes. In our model we start by expressing the samples  $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})$ ,  $i = 1, \dots, n$ , as

$$y_{ij} = m_j + \mu_{ij} + \sigma_j \epsilon_{ij}, \quad \epsilon_{ij} \stackrel{i.i.d.}{\sim} N(0, 1).$$

In this formulation,  $m_j$  and  $\sigma_j$  are attribute-specific mean and standard deviations shared by all samples. We put the following priors for them:

$$m_j \stackrel{i.i.d.}{\sim} DP(\alpha N(m_0, \sigma_0)),$$

$$\sigma_j^2 \stackrel{i.i.d.}{\sim} DP(\beta IG(\alpha_0, \beta_0)),$$

where  $DP(\alpha H)$  is the Dirichlet process with concentration parameter  $\alpha > 0$  and base probability measure  $H$ , and  $IG()$  represents the inverse-Gamma distribution. In this paper, we use the notation  $\theta_i \stackrel{i.i.d.}{\sim} DP(\alpha H)$  as a short form for the more rigorous  $\theta_i \stackrel{i.i.d.}{\sim} G, G \sim DP(\alpha H)$ . This might be a misuse but simplifies our notation since DP appears multiple times at different



places within our model. When  $\alpha \rightarrow \infty$ , the first expression above reduces to  $m_j \sim N(m_0, \sigma_0)$ , for example. The use of Dirichlet process can be motivated from at least two point of views. First, it relaxes the normality assumption imposed on the components of the mean vector. Second, since the DP is a discrete measure, it provides a regularization mechanism by shrinking different parameters towards each other.

Since the attribute specific  $m_j$  and  $\sigma_j$  are shared by all samples, the clustering structure can only derive from appropriate specification on  $\mu_{ij}$ . As in Hoff (2006); Kim et.al. (2006), the clustering of samples will be determined by an infinite mixture of distributions via Dirichlet process mixture. Denote  $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{ip})$ . When it is intended that  $\boldsymbol{\mu}_i$  is the mean for cluster  $c$ , i.e. sample  $i$  is assigned to cluster  $c$ , we also use  $\boldsymbol{\mu}_c$  to denote the same mean vector. Although there might be some concern over misuse of notation, this can hardly cause any confusion in the context. The sample means are generated from an infinite mixture specified as the following:

$$\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{ip}) \stackrel{i.i.d.}{\sim} DP(\tau H),$$

where the concentration parameter  $\tau$  controls the a priori number of clusters and the base measure  $H$  on  $\boldsymbol{\mu}_i$  can be defined through the following hierarchical “point-mass mixture” prior:

$$\begin{aligned}\mu_{ij} &\sim (1 - \pi_{ij})\delta_0 + \pi_{ij}DP(\gamma N(0, \eta_i^2)), \\ \pi_{ij} &\sim (1 - \rho_j)\delta_0 + \rho_j Beta(a, b), \\ \rho_j &\sim Beta(c, d),\end{aligned}$$

where  $\delta_0$  is the point-mass distribution at the single point zero. Thus in our model, not only are samples assigned to different groups (i.e.,  $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j, 1 \leq i \neq j \leq n$  with positive probability), but the nonzero components of the mean specific to a cluster are also clustered (i.e.,  $\mu_{ij} = \mu_{ik}, 1 \leq j \neq k \leq p$  with positive probability). In this paper, we choose to use a more parsimonious model  $\eta_i \equiv \eta$ . The prior structure presented above has individual probability  $\pi_{ij}$  that attribute  $j$  has a nonzero effect for cluster  $c$  to which the  $i$ -th sample is assigned, while the attribute specific parameter  $\rho_j$  indicates the sparsity propensity of the covariate  $j$ . Marginalization over  $\pi_{ij}$  gives the more traditional point-mixture prior

$$\mu_{ij} \sim (1 - \frac{a}{a+b}\rho_j)\delta_0 + \frac{a}{a+b}\rho_j DP(\gamma N(0, \eta^2)).$$

Similar structure has been used in the regression context in Lucas et.al (2006); Seo et.al. (2007); Carvalho et.al. (2008). As discussed in those papers, the extended model is able to more adequately shrink towards zero through the induction of zeros for  $\pi_{ij}$  and thus can better separate real signals from noise and reduce false discovery of uninformative variables.

From the structure of the specified prior, one can see that the identifiability of our model is enforced by the assumed sparsity of  $\mu_{ij}$ . For example,

in a problem where all covariates are uninformative (in other words, there is only one single cluster), our formulation will shrink all  $\mu_{ij}$  to zero while  $m_j$  will assume the value of  $j$ -th covariate mean. Also, the effects of  $\sigma_j$  and the mean can be separated because of the clustering structure on the sample so the number of unique values among  $\mu_{ij}, i = 1, \dots, n$  usually is much smaller than  $n$ .

Next, we describe the choice of hyperpriors and the setting of hyperparameters. The base measure of the DP prior for  $m_j$  is set as a normal distribution with  $m_0 = y_{.j}, \sigma_0^2 = \sum_{j=1}^p (y_{.j} - \bar{y})^2 / p$  where  $y_{.j} = \sum_i y_{ij} / n$  is the observation mean for attribute  $j$  and  $\bar{y} = \sum_j y_{.j} / n$  is the overall mean of all observations. For the base measure of DP prior for  $\sigma_j^2$ , we use the vague prior  $IG(0.5, 0.5)$ . Similarly, the standard vague conjugate prior  $IG(0.5, 0.5)$  is also used as prior for  $\eta^2$ . For the four concentration parameters in the DPs,  $\tau, \alpha, \beta, \gamma$ ,  $\text{Gamma}(0.5, 0.5)$  is used as the prior. In the point-mass mixture prior, we follow Lucas et.al. (2006) and set  $a = 9, b = 1, c = 0.2, d = 199.8$ . Thus in our prior specification, we use vague prior distribution when appropriate, and also provided guide values for other hyperparameters. Using guide values raises some concerns on sensitivity to these choices, as advocated in Ibrahim et.al. (2002); Chipman et.al. (2009). Computation of the posterior distribution in our model is somewhat challenging. This is due to the double usage of the Dirichlet process. We have adopted a sequential sampling approach and successfully constructed a proposal distribution for the high-dimensional means. The details of our MCMC algorithm is not detailed here due to space constraint.

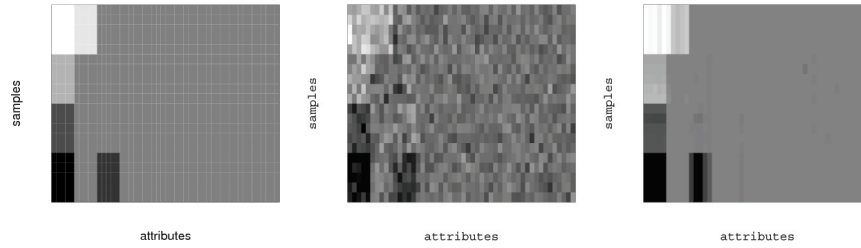
### 3 Simulation and Application

#### 3.1 Simulation Study

We investigate the performance of our estimation method in a simulation study. A dataset containing 20 samples and 200 covariates is generated as follows.

$$\begin{aligned}
 y_{ij} &= \mu_{ij} + \sigma_j \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, 1), \\
 \mu_{ij} &= 0.25, 1 \leq i \leq 5, 1 \leq j \leq 5, \\
 \mu_{ij} &= 0.1, 6 \leq i \leq 10, 1 \leq j \leq 5, \\
 \mu_{ij} &= -0.1, 11 \leq i \leq 15, 1 \leq j \leq 5, \\
 \mu_{ij} &= -0.25, 16 \leq i \leq 20, 1 \leq j \leq 5, \\
 \mu_{ij} &= 0.2, 1 \leq i \leq 5, 6 \leq j \leq 10, \\
 \mu_{ij} &= -0.15, 16 \leq i \leq 20, 11 \leq j \leq 15, \\
 \mu_{ij} &= 0 \text{ otherwise,} \\
 \sigma_j &= 0.1, 1 \leq j \leq 15, \\
 \sigma_j &= 0.05, \text{ otherwise.}
 \end{aligned}$$

The structure of  $\mu_{ij}$  is shown in Figure 1(a) where different values for  $\mu_{ij}$  show up as different gray levels. Each row in the image represents a sample and each column represents an attribute. Only the first 50 attributes are shown. We use the model described in Section 2 to fit the simulated dataset. Figure 1(b) shows the observed data in the same format as Figure 1(a). The posterior gives strong support for four clusters, with support for five clusters comes next. In simulation as well as real data application that follows, we used a burn-in period of 10,000 updates and 40,000 iterations after burn-in for inferences. The posterior estimates of  $\mu_{ij}$  is shown in Figure 1(c) as a matrix for the first 50 attributes only. Four clusters and the zero structures are clearly identified.



**Fig. 1.** (a) Mean structure  $\mu_{ij}$  for the simulated data plotted as an image. (b) Noisy observed data. (c) Estimated mean under our approach.

We have also examined the posterior estimate of  $\rho_j$  for  $1 \leq j \leq 50$ , which indicates the contribution of the  $j$ -th attribute to cluster discrimination. The results are quite encouraging, with the first 15 attributes clearly identified as signal variables and the first 5 attributes estimated to be associated with larger values of  $\rho_j$ , consistent with the simulation scheme.

Finally, for this simulated example, using DP for  $m_j$  and  $\mu_{ij}$ ,  $1 \leq j \leq p$  performs better than a normal prior (corresponding to the case with  $\alpha \rightarrow \infty$  and  $\gamma \rightarrow \infty$ ). The mean squared error of  $m_j + \mu_{ij}$ ,  $1 \leq i \leq 20, 1 \leq j \leq 15$ , under our model is 0.006, in contrast with 0.011 when  $\alpha, \gamma \rightarrow \infty$ . This is consistent with the results reported in Nott (2008).

### 3.2 Leukemia Gene Expression Data Example

We use the leukemia gene expression dataset (Golub et.al. 1999) to demonstrate the utility of our proposed method. The training dataset contains 38 tissue samples, among which 11 samples are acute myeloid leukemia (AML) and the rest are acute lymphoblastic leukemia (ALL). The 27 ALL samples are further divided into two subgroups: 8 T-cell and 19 B-cell samples. The samples were arrayed with a total of 7129 genes in a microarray experiment.

Following the standard preprocessing steps in Dudoit et.al (2002), we truncate the expression values to within the interval  $[1, 16000]$ , and delete those genes whose maximum and minimum expression across all samples satisfies  $\max / \min \leq 5$  and  $\max - \min \leq 500$ . Finally, we select the top 2000 genes with the largest variances across all samples so that at the end we have for this dataset  $n = 38, p = 2000$ .

We apply our proposed method to the dataset with the hyperparameters set exactly as discussed in Section 2. Convergence of the MCMC updates is invariably a concern in high-dimensional problem with variable selection. As a simple diagnostic, two MCMC runs of 50,000 iterations with the first 10,000 as burn-in are implemented, with different initialization. In particular, we start one Markov chain with initially all samples assigned to one cluster, and another chain where each sample is assigned to its own separate cluster. The posterior estimates of various unknown quantities for the two runs shows good agreement which indicates the chains mixed well in our implementation.

The posterior for this dataset put most of the support for the number of clusters between 3 and 9, with 6 clusters receiving the highest score. Conditional on  $K = 6$ , setting the threshold 0.5 for the posterior estimates of  $\pi_{cj}$  returns 872 genes. This is much larger than the 120 genes reported in Kim et.al (2006). Previous studies, such as Thomas et.al (2001), also demonstrated that there were a large number of genes differentially expressed between different tissue samples.

For unsupervised clustering problems, it is generally difficult to assess the performance of any procedure when the underlying truth is unknown. In this example, we use the known tissue subtypes for this dataset as the proxy and inferences about the cluster structure is compared to the known tissue subtypes. We estimate the posterior probability of  $c_i = c, 1 \leq c \leq 6$  ( $c_i$  is the cluster index of sample  $i$ , which is sampled within our MCMC algorithm) from posterior samples conditioned on  $K = 6$  with the help of the procedure that deals with label switching. Each sample is allocated to the cluster with the largest posterior probability. The relationship between this allocation and known tissue types are shown in Table 1. We see that the known AML and ALL-B tissue types might further consist of some subtypes. Using our method, we can also discover genes that distinguish only some subgroups. For example, among those 872 genes relevant for clustering only 64 of them can distinguish between ALL and AML samples without discriminative power for different subtypes.

## 4 Conclusion

In this article, we propose a novel Bayesian approach to high-dimensional clustering with variable selection. The distinguishing features of our method include a separate Dirichlet process for shrinkage estimation of cluster mean, as well as a hierarchical point-mass structure that improves the separation

**Table 1.** Clustering results for leukemia expression data conditional on  $K = 6$ .

samples	cluster from the proposed method					
	1	2	3	4	5	6
ALL-T(8)	0	0	8	0	0	0
ALL-B(19)	0	1	0	6	4	8
AML(11)	7	3	0	0	1	0

of significant signal from noise variables. We propose a sequential sampling approach in one of the updating iterations of the MCMC algorithm to solve the computational problem associated with the high dimensionality of the mean vector.

## References

- BANFIELD, J.D. AND RAFTERY, A.E. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821 (1993).
- CARVALHO, C., CHANG, J., LUCAS, J., NEVINS, J., WANG, Q., AND WEST, M. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456 (2008).
- CHIPMAN, H.A., GEORGE, E.I., AND MCCULLOCH, R.E. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, to appear (2009).
- DUDOIT, S., FRIDLAND, J., AND SPEED, T.P. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87 (2002).
- FRALEY, C. AND RAFTERY, A.E. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631 (2002).
- GOLUB, T.R., SLONIM, D.K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J.P., COLLIER, H., LOH, M.L., DOWNING, J.R., CALIGIURI, M.A., BLOOMFIELD, C.D., AND LANDER, E.S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537 (1999).
- HOFF, P. Model-based subspace clustering. *Bayesian analysis*, 1:321–344 (2006).
- IBRAHIM, J., CHEN, M.H., AND GRAY, R.J. Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, 97(457):88–99 (2002).
- KIM, S., TADESSE, M.G., AND VANNUCCI, M. Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4):877–893 (2006).
- LUCAS, J., CARVALHO, C., WANG, Q., BILD, A., NEVINS, J.R., AND WEST, M. Sparse statistical modelling in gene expression genomics. In *Bayesian Inference for Gene Expression and Proteomics*, 155–176. Cambridge University Press (2006).

- NOTT, D.J. Predictive performance of Dirichlet process shrinkage methods in linear regression. *Computational Statistics & Data Analysis*, 52(7):3658–3669 (2008).
- SEO, D.M., GOLDSCHMIDT-CLERMONT, P.J., AND WEST, M. Of mice and men: Sparse statistical modeling in cardiovascular genomics. *Annals of Applied Statistics*, 1(1):152–178 (2007).
- THOMAS, J.G., OLSON, J.M., TAPSCOTT, S.J., AND ZHAO, L.P. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, 11(7):1227–1236 (2001).

# Data Mining and Multiple Correspondence Analysis via Polynomial Transformations

Rosaria Lombardo

Economics Faculty, Second University of Naples,  
Gran Priorato di Malta, 81043 Capua (CE), Italy, *rosaria.lombardo@unina2.it*

**Abstract.** In the framework of the Total Quality Management, earlier studies have suggested that enterprises could harness the predictive power of Learning Management System data to develop reporting tools that identify at-risk customers/consumers and allow for more timely interventions. To support decision making in customer-centric planning tasks, exploratory multivariate data analysis is an important part of corporate data mining. To monitor the overall (dis)satisfaction with respect to the service aspects, among different exploratory tools, we focus on Multiple Correspondence Analysis via polynomial transformations to deal with ordered categorical variables and nominal ones too.

**Keywords:** data mining, customer satisfaction, multiple correspondence analysis, polynomial transformations, customer classification

## 1 Introduction

The necessity of systems of evaluation and assessment in many socio-economic fields together with the Learning Management System data (LMSD) have sparked a need in building early warning system (EWS) which produces signal for possible risks. Accordingly various EWSs have been established (Kim et al. (2004)): for detecting fraud, for credit-risk evaluation in the domain of financial analysis (Fawcett and Provost (1997)), for detection of risks potentially existing in medical organizations (risk aversion of nurse incidents, infection control and hospital management), to support decision making in customer-centric planning tasks (Lessman and Vob (2009), Macfadyen and Dawson (2010)). In customer satisfaction research and in general in service evaluations, the amount of dissatisfied customers can represent an early warning data useful to detect the risk to lose customers. A very common approach to analyze large survey data is to perform Multiple Correspondence Analysis (MCA, Benzecri (1973), Greenacre (1984), Lebart et al. (1984), Gifi (1990)) where a question/variable is regarded as a set of category points. MCA can be introduced in many different ways, which is probably the reason why it was reinvented many times over the years. If we confine ourselves to categorical variables with ordered categories numbered (Likert items), recently, a confirmatory approach to MCA analysis (Lombardo and Meulman (2010), Lombardo and Beh (2010)) has been shown to be applicable to take into

account the information in ordered categorical variables. This approach is called Ordered MCA (OMCA, Lombardo and Meulman (2010)) where explanatory tools are combined with inference ones. Ordered MCA maintains all the features of MCA, and allows for additional information about the structure and association of the ordered categories and about the individual representations than MCA ones. The main aim of this project is to face with customer satisfaction studies in services to persons of public utility, by OMCA analysis, proposing a new strategy to deal not only with ordered variables, but also with nominal ones. After reviewing the confirmatory approach to multiple correspondence analysis, section 2, focuses on performing OMCA when both nominal and ordinal variables are considered. In section 3, we illustrate the applicability of OMCA, using data obtained from the SERVPERF questionnaire (Parasuraman et al. (1985)) to monitor the level of patient satisfaction in health care services. Finally section 4 contains final remarks and future work perspectives.

## 2 Multiple Correspondence Analysis via Polynomial Transformations

At first, we briefly summarize Multiple Correspondence Analysis via Polynomial transformations for evaluating survey data. Questionnaires often result in responses to a large number of questions with a limited number of answer categories. In a graphical representation, the association between the variables is represented by the closeness of the categories of different variables. The responses to these  $p$  questions, coded in complete disjunctive form, lead to different ways of classifying all the individuals in the sample. Let  $\mathbf{X}=[\mathbf{X}_1|\dots|\mathbf{X}_p]$  be the indicator super-matrix of  $p$  ordered categorical variables observed on the same set of  $n$  individuals, with  $J = \sum_{k=1}^p j_k$  the total number of categories, i.e. the number of columns in  $\mathbf{X}$ . Let  $\mathbf{X}_k$  be the indicator matrix of the  $k$ th variable with margins  $x_{.j_k} = \sum_{i=1}^n x_{ij_k}$ . Define  $\mathbf{D}$  as the diagonal super-matrix of dimension  $J \times J$ , whose generic diagonal elements are given by the diagonal elements of the  $k$  different matrices  $\mathbf{D}_k = x_{(.j_k)}$ . To studying the relationships among the categories, we can perform a correspondence analysis on the  $n \times J$  indicator super-matrix  $\mathbf{X}$ , by the hybrid decomposition (HD) which implies computing the singular vectors for the individuals and orthogonal polynomials for the ordered categorical variables (see for details Lombardo and Meulman (2010)). Therefore, the total inertia of the contingency table is not only partitioned into polynomial components, but can also be partitioned into  $m$  singular values and singular vectors. At the heart of the analysis lies the matrix  $\mathbf{Z} = 1/(p\sqrt{n})\mathbf{\Phi}'\mathbf{X}\mathbf{D}^{-1/2}\tilde{\mathbf{\Psi}}$ , by means of it the total inertia can be expressed as  $\text{trace}(\mathbf{Z}'\mathbf{Z}) = \text{trace}(\mathbf{Z}\mathbf{Z}') = \text{trace}(\mathbf{\Lambda}_{\mathbf{Z}}^2)$ . Differently from the matrix of singular values, the matrix  $\mathbf{Z}'\mathbf{Z}$  is not diagonal. The non-zero off-diagonal associations between the row and column categories allow us to identify important structures in the data not otherwise



detected. To test for statistically significant components in the decomposition of the total inertia using  $\mathbf{Z}$ , the mathematical equivalence between the inertia and the Pearson chi-squared statistic is considered. In fact, Lombardo and Meulman (2010) show that the element of the  $\mathbf{Z}$  matrix is asymptotically chi-squared distributed, due to the relationship between the bivariate moment  $z_{m,v_k}^2$  and the eigenvalue  $\lambda_{X_m^2}$ . Through the use of orthogonal polynomials that are associated with the ordered categorical variables, the partition allows to analyse and decompose the total inertia in terms of linear, quadratic and higher order polynomial components. Not only the total inertia, but also the contribution to the inertia by each singular vector can be partitioned into orthogonal polynomial components, showing the contribution of the dominant ones (linear, quadratic, cubic, etc). The  $(m, v_k)$ th value of  $\mathbf{Z}$  defines the contribution of the  $v_k$ th-order bivariate moment between the categories of the  $k$ th ordered variable to the  $m$ th principal axis. When  $v = 1$ , the element  $z_{m,1_k}$  describes the importance of the location component for the  $k$ th variable on the  $m$ th axis of a classical MCA plot. Therefore, the overall location component of the categories of the  $k$ th variable can be determined by calculating  $\sum_{m=1}^M z_{m,1_k}^2$ . If this component is significant, then there is a significant variation in the location of those categories (explaining the so-called horse-shoe that points to a very dominant first dimension). In practice, when the linear component is dominant, then the representation using polynomials allows us to visualize the linear trend in the categories along the first polynomial axis. Unlike classical factorial analysis, the first and second polynomial axis are not necessarily the most important. The quadratic component of the categories can be calculated by  $\sum_{m=1}^M z_{m,2_k}^2$  which reflects the spread of the categories of the  $k$ th variable. To display the association among variable categories and enhance the interpretation of the graphical display, a plotting system based on the orthogonal polynomials is employed. Looking at the representation of the categories, we can say that the computed coordinates for the categories in OMCA are identical to the coordinates obtained from a classical MCA. Concerning the units, unlike classical MCA coordinates which lead to a scatter of points with very often no apparent pattern, the unit plot obtained by using orthogonal polynomials is very informative because units are automatically arranged in distinct clusters, thereby giving a simple structure and classification of the individuals. Assuming that all variables consist of the same number of ordered categories (as in data consisting of Likert items) such that  $j_k = j$  for all  $k = 1, 2, \dots, p$ , the OMCA unit plot will consist of  $j$  clusters of objects. This particular feature makes very attractive OMCA to monitor (dis)satisfaction of each customer cluster in different time or spaces. Furthermore, to take into account the different information given by nominal variables (example: sex, professions, residence, etc.) often included in questionnaires, we propose a new strategy which consists in splitting data in so many sets as the number of nominal categories and applying the hybrid de-

composition to each data set. An example of this strategy has been illustrated in the following section.

### 3 An Application: Customer Satisfaction in Health Care Services

The data concern a survey on the perception of various aspects of quality in an hospital in Naples (survey of the Second University of Naples, Italy, June 2008). Patient satisfaction was measured by the so-called SERVPERF instrument, using the questionnaire version presented by Babakus and Mangold (1992). The SERVPERF questionnaire measures the perceived quality

<i>Variable</i>	<i>Component</i>	$z_{1(v_k)}^2$	$X^2$	$z_{2(v_k)}^2$	$X^2$	<i>df</i>
<i>Tangibility</i>	Location	0.104	73.230***	0.030	2.093	8
	Dispersion	0.000	0.328	0.051	35.956***	8
	Skewness	0.001	0.362	0.008	2.398	8
	Kurtosis	0.002	1.567	0.000	5.936	8
<i>Reliability</i>	Location	0.140	98.781***	0.000	0.282	8
	Dispersion	0.000	0.219	0.099	69.999***	8
	Skewness	0.001	0.368	0.003	2.217	8
	Kurtosis	0.000	0.038	0.000	0.033	8
<i>Capability of Response</i>	Location	0.153	107.539***	0.002	1.154	8
	Dispersion	0.003	1.950	0.131	92.568***	8
	Skewness	0.001	0.523	0.008	5.806	8
	Kurtosis	0.000	0.027	0.002	1.748	8
<i>Capability of Assurance</i>	Location	0.151	106.328***	0.002	1.106	8
	Dispersion	0.005	3.313	0.119	84.106***	8
	Skewness	0.001	0.529	0.013	9.315	8
	Kurtosis	0.001	0.454	0.000	0.011	8
<i>Empathy</i>	Location	0.143	101.009***	0.003	2.094	8
	Dispersion	0.003	2.242	0.093	65.398***	8
	Skewness	0.001	0.615	0.016	11.082	8
	Kurtosis	0.002	1.665	0.000	0.020	8
Total		0.711	501.088***	0.558	393.320***	160

**Table 1.** Decomposition of the first two non-trivial eigenvalues and chi-square tests

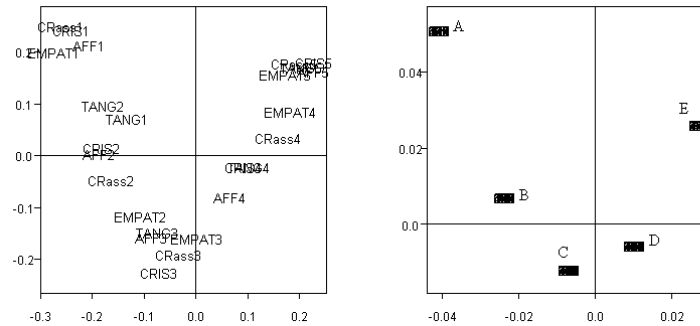
of some important aspects of the quality, using a response scale with ordered categories numbered from 1 to 5. In the health care context, *tangibility* refers to the structural aspects, *reliability* to trust and precision, *capacity of response* to emergency ready ward, *capacity of assurance* to competence and courtesy, and *empathy* to personal attention towards the patient. The data set consists of 705 patients, and 15 variables (Likert items), each having five

ordered categories. Three items measure *tangibility* (*Tang*), three measure *reliability* (*Rel*), four items measure *response capacity* (*CRes*), three *capacity of assurance* (*CRas*), and two *empathy* (*Emp*). As a composite measure for each of the five quality aspects, the respective medians for *Tang*, *Rel*, *CRes*, *CRas*, *Emp* were computed across subsets. Categories of these composite variables will be indicated by the label number, for example, the composite responses *tangibility* on a five point scale are denoted by *Tang1*, *Tang2*, *Tang3*, *Tang4* and *Tang5*. As previously discussed, to reflect the ordinal structure natural scores have been transformed in four orthogonal polynomials. Using multiple correspondence analysis, the total inertia is 4.0, after removing the redundancy. The first two eigenvalues are  $\lambda_X^2 = 0.711$  and  $\lambda_X^2 = 0.558$ , which are decomposed and tested in Table 1. The graphical results obtained by OMCA are given in Figure 1a,b. The display consists of two panels. Figure 1a depicts the association across the 25 response categories of the five quality aspects. This representation is the same for MCA and OMCA. The plot shows that a low satisfaction level (response category 1) for one quality aspect is associated with a low level of satisfaction for the other quality aspects. Reversely, an aspect being judged as excellent (with response category 5) is associated with excellent ratings for the other aspects. Figure 1b shows the five classes of points for the 705 hospital patients who participated in the study. By partitioning the total inertia on the basis of these polynomials, it is possible to determine the dominant variables in the plot for the category points, and also the dominant sources in each variable. Table 1 provides a summary of the components that reflect the first four moments (location, dispersion, skewness and kurtosis) of each variable, and their contribution to the first and second principal axis of the MCA and OMCA plot. The statistically significant components are identified at three levels of significance: 0.01 (\*\*\*), 0.05 (\*\*) and 0.10 (\*), respectively. The total degrees of freedom are equal to d.f.=160. Table 1 shows that the variation between the categories of each variable is best explained in terms of the differences in their location and dispersion: the location component explains 97.1% of the inertia accounted for by the first dimension, while the dispersion component accounts for 88.4% of inertia accounted for by the second dimension.

These values of inertia accounted for can be further partitioned to identify those variables that dominate the solution for each dimension. If one considers the chi-squared inertia's for each dimension, in Table 1 *tangibility* accounts for 15% of the first principal inertia. Similarly, *reliability*, *capability of response*, *capability of assurance* and *empathy* contribute to 19.8%, 22.1%, 22.2% and 20.9% to the first principal axis, respectively. Similarly, for the second principal axis of Figure 1a, these variables contribute to 14.3%, 22.0%, 31.5%, 21.7% and 10.5%, respectively, to the inertia. For this two-dimensional representation of the association among the categories, we can therefore determine that *tangibility*, *reliability*, *capability of response*, *capability of assurance* and *empathy* account for 15.9%, 18.3%, 25.6%, 24.6% and 20.1%, respectively, of

the total variation between the variables.

This result indicates that the patients consider that the most important



**Fig. 1.** Figure 1a: graphical display of response categories in overall hospital - Figure 1b: graphical display of patients in overall hospital

aspect of quality is the hospital staffs *capability to respond* to the patients' needs, while *tangibility* is the least important factor. But now consider Fig-

<i>Cluster % of Patients in Cluster</i>	
A	13.6%
B	41.7%
C	30.6%
D	4.7%
E	9.4%

**Table 2.** Percentage of patients that lie in each of the 5 clusters in Fig. 1b

ure 1b, the OMCA representation for the patients. It is clear that there are five distinct clusters of patients, each associated with one of the five categories that form an ordered variable. It is evident that those in cluster **A** have an overall poor judgement of the quality of their hospital, while cluster **E** clearly shows those patients who gave an overall excellent rating to the hospital services. In fact, to better understand how dominant each of the response categories is, Table 2 shows the distribution of the patients over the five clusters representing an overall judgement from poor to excellent. This distribution shows that about the 72.3% of the patients evaluate the services in the hospital of middle-bad quality, the 9.4% of the respondents qualified the services in the hospital of very-high quality. At the other end

of the scale, the 13.6% of the patients responded that the quality of the services was poor. Such an overall rating of the hospital services is not easily obtained when performing classical MCA. Thus the use of OMCA has a major advantage for this particular data set: we can easily monitor the overall (dis)satisfaction with the health care services the hospital provides. To take

<i>Cluster % of Patients in Cluster</i>	
A	15.3%
B	36.1%
C	36.1%
D	2.8%
E	9.7%

**Table 3.** Percentage of patients that lie in each of the 5 clusters in Fig. 1b

into account of nominal variables concerning different patient characteristics like the division department, a further analysis is considered. We investigate the perception of the same five quality aspects in gynaecology division. The patients of gynaecology are  $n = 216$ . For sake of brevity we do not display the graphical results, but we report in Table 3 the five clusters of gynaecology patients who were automatically classified with respect to their rating of the quality aspects from excellent (class **E**) to poor (class **A**). The different percentages of the subgroup of gynaecology patients that lie in each of the five clusters are reported in Table 3. In the gynaecology division we observe that the percentage of patients who found that the hospital offered excellent quality services is 9.7% (Table 3), similar than that from the overall patients. Furthermore, about 15.3% of patients of gynaecology thought that the hospital division offered very poor quality services, this percentage is greater than that from the overall patients, and only 12.5% of the patients rated the hospital division services as high/middle-high.

## 4 Conclusion

In literature, different works face with correspondence analysis by orthogonal polynomials for two-way and three-way contingency tables (Beh (1997), Beh and Davy (1999), Best and Rayner (1996), D'Ambra et al. (2005), Lombardo et al. (2007), Beh et al. (2007)). Recently a confirmatory approach to multiple CA (Lombardo and Meulman (2010), Lombardo and Beh (2010)) has been shown to be useful to enrich data representation of the categorical variables with ordered categories and of the individuals participating to the survey. In this paper we propose a new strategy based on OMCA to deal with nominal and ordered variables. In customer satisfaction studies where Likert items for the evaluation of quality aspects together with personal information can be

considered, the splitting of individual set with respect to the nominal variable categories and the automatic aggregation of individuals in so many clusters as the number of the ordered categories can provide early warning system data that help to identify at-risk customers/consumers and suggest for more timely interventions to improve service quality in department division. In perspectives the proof of stability of individual clusters, will be object of further study.

## References

- BABAKUS, E. and MANGOLD, G. (1992): Adapting the Servqual scale to hospital services: an empirical investigation. *Health Services Research Journal* 26, 767-786.
- BEH, E. J. (1997): Simple Correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical Journal* 39 (5), 589-613.
- BEH, E. J. and DAVY, P. J. (1999): Partitioning Pearson's chi-squared statistic for a partially ordered three-way contingency table. *The Australian and New Zealand Journal of Statistics* 41, 233-246.
- BENZÉCRI J. P. (1973): *Analyse des données* (2 vols). Paris: Dunod.
- BEST, D. J. and RAYNER, J. C. W. (1996): Nonparametric analysis for doubly ordered two-way contingency tables. *Biometrics* 52, 1153-1156.
- FAWCETT, T. and PROVOST, F. (1997): Adaptive fraud detection. *Data Mining and Knowledge Discovery* 1 (3), 291-316.
- GIFI, A. (1990): *Non-linear Multivariate Analysis*. Chichester: Wiley.
- GREENACRE, M. (1984): *Theory and Applications of Correspondence Analysis*. Academic Press, London.
- KIM, T.Y., OH, K.J., SOHN, I. and HWANG, C. (2004): Usefulness of artificial neural networks for early warning system of economic crisis. *Expert Systems with Applications* 26, 583-590.
- LEBART, L., MORINEAU, A. and WARWICK, K.M. (1984): *Multivariate Descriptive Statistical Analysis*. Wiley series, 3.
- LESSMAN, S. and VOB, S. (2009): A reference model for customer-centric data mining with support vector machines. *European Journal of Operational Research* 199 (2), 520-530.
- LOMBARDO, R. and BEH, E.J. (2010): Simple and Multiple Correspondence Analysis for Ordinal-scale Variables using Orthogonal Polynomials. *Journal of Applied Statistics*, in press.
- LOMBARDO, R. and MEULMAN, J. (2010): Multiple Correspondence Analysis via Polynomial Transformations of Ordered Categorical Variables. *Journal of Classification* 10, 32-48.
- LOMBARDO, R., BEH, E. and D'AMBRA, L. (2007): Non-symmetric Correspondence Analysis with ordinal variables using orthogonal polynomials. *Computational Statistics & Data Analysis* 52, 566-577.
- MACFADYEN, L. P. and DAWSON, S. (2010): Mining LMS data to develop an early warning system for educators: A proof of concept. *Computers & Education* 54 (2), 588-599.
- PARASURAMAN, A., ZEITHAML, V.A. and BERRY, L.L. (1985): A conceptual model of service quality and its implications for future research. *Journal of Marketing* 49, 41-50.

# Structural Modelling of Nonlinear Exposure-Response Relationships for Longitudinal Data

Xiaoshu Lu and Esa-Pekka Takala

Finnish Institute of Occupational Health  
Topeliuksenkatu 41 a A, FIN-00250 Helsinki, Finland, *xiaoshu@cc.hut.fi*

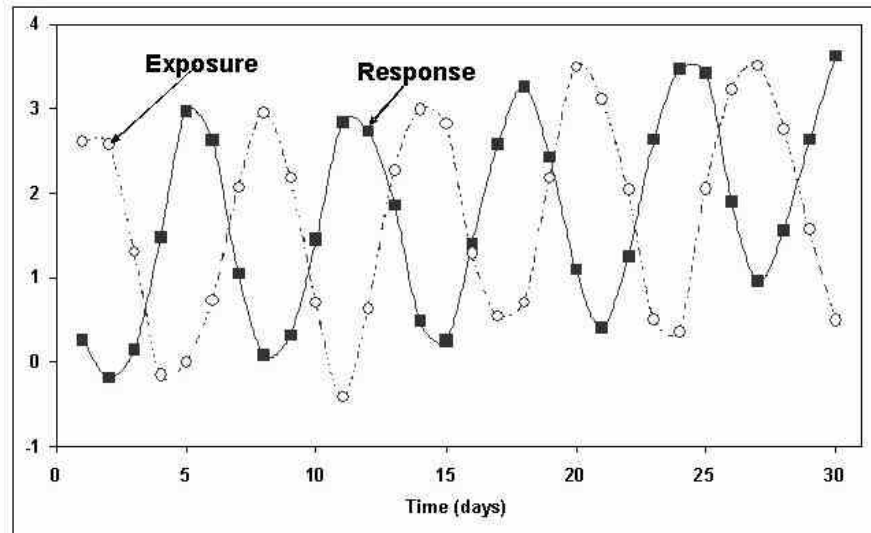
**Abstract.** Exposure-response relationships are of interest in many epidemiological, medical and other applications. Most commonly, linear relationships are examined. However, many longitudinal data show a remarkable dynamic and nonlinear characteristic, which requires a structure-based approach to elucidate the nonlinear exposure-response relationship behind the data. Exposure and response can have strong nonlinear association and no linear correlation. In this paper, we develop a new model for longitudinal data to address these challenges. The methodology includes time series analysis to estimate unobserved components for exposure and response, and to model their dynamic and structural relationship in a fixed-effects form for each subject. An extension of the fixed-effects form to mixed-effects model for all subjects is proposed and the relevant methods for estimating variance-covariance and correlation matrices are presented. The model-building procedure is explained. The performance of the model is demonstrated using the hypothetical data.

**Keywords:** structural modelling, exposure-response relationship, nonlinear, longitudinal data

## 1 Introduction

Epidemiological, medical and many other research is largely grounded on exposure and risk assessments. Once exposure information is obtained, following health outcomes over time would provide the information needed to complete the risk assessment. The widespread use of exposure and risk assessment procedures has produced a diversification and specialisation of different methodologies, depending on the case under consideration. However, a critical review reveals that linear model has been widely used in examining exposure-response relationships (Bassey and Effanga 2008).

In many applications, data often show complex and nonlinear exposure-response relationships (Davidian and Giltinan 1995). Standard statistical methods provide few theories on how to study nonlinear patterns (Wang 2002). As an example, Fig. 1 illustrates hypothetical data where the response responds to the exposure only by a phase shift with random errors. Hence



**Fig. 1.** The plot of hypothetical exposure-response data. Only a phase shift is introduced in response to exposure with random errors.

the response is positively correlated with the exposure. A longitudinal analysis with a linear mixed-effects model, for example through SAS's PROC MIXED procedure, shows that the parameter estimate of -0.61 is statistically discernible at 5% level, indicating that response is negatively associated with exposure which is obviously incorrect. Inspection of the graph in Fig. 1 indicates nonlinearity with the data. A linear analysis can lead to wrong results for such kind of data.

The hypothetical data are devised to mimic and simplify the supposed structure of our real data in a study of health effects of whole body vibration (data not shown here due to the space limitation). In this paper, we present a new approach to address these methodological challenges by focusing on the hypothetical data to find out possible nonlinear exposure-response relationships. The methodology includes time series analysis to estimate unobserved components for exposure and response, and to model their dynamic and structural relationship in a fixed-effects form for each subject. An extension of the fixed-effects form to mixed-effects model for all subjects is proposed and the relevant methods for estimating variance-covariance and correlation



matrices are presented. The model-building procedure is explained through the illustrated data.

## 2 Mathematical model

In this section, we present the methodological framework for building the model equations, and then we show that the proposed model is equivalent to a mixed-effects model.

### 2.1 Model equations

Let  $\{x\}_t$  and  $\{y\}_t$  be exposure and response measures for any subject. Here the subject index is omitted for brevity and convenience. We employ the structural Hodrick-Prescott (HP) filter technique to extract the trend-cycle component (Kydland and Prescott 1990, Hodrick and Prescott 1997, Proietti 2007). For response time series  $\{y\}_t$ , the HP filter defines its trend and cycle as

$$y_t = y_t^{trend} + \epsilon_{y_t^{trend}}. \quad (1)$$

$$y_{t+1}^{trend} = 2y_t^{trend} - y_{t-1}^{trend} + \epsilon_{y_t^{cycle}}. \quad (2)$$

with starting values  $y_1^{trend} = a_0 + a_1$  and  $y_2^{trend} = a_0 + 2a_1$ . The error components are  $\epsilon_{y_t^{trend}} \sim N(0, \sigma_{y_t^{trend}}^2)$  and  $\epsilon_{y_t^{cycle}} \sim N(0, \sigma_{y_t^{cycle}}^2)$ . Similarly, for exposure time series  $\{x\}_t$ , we have

$$x_t = x_t^{trend} + \epsilon_{x_t^{trend}}. \quad (3)$$

$$x_{t+1}^{trend} = 2x_t^{trend} - x_{t-1}^{trend} + \epsilon_{x_t^{cycle}}. \quad (4)$$

with starting values  $x_1^{trend} = b_0 + b_1$  and  $x_2^{trend} = b_0 + 2b_1$ . The error components are  $\epsilon_{x_t^{trend}} \sim N(0, \sigma_{x_t^{trend}}^2)$  and  $\epsilon_{x_t^{cycle}} \sim N(0, \sigma_{x_t^{cycle}}^2)$ . We assume  $\epsilon$ 's are independent for the exposure and response time series.

Combining equation 2 and the starting values gives

$$\begin{aligned} y_t^{trend} &= 2y_{t-1}^{trend} - y_{t-2}^{trend} + \epsilon_{y_{t-1}^{cycle}} \\ &= 2(2y_{t-2}^{trend} - y_{t-3}^{trend} + \epsilon_{y_{t-2}^{cycle}}) \\ &\quad - (2y_{t-3}^{trend} - y_{t-4}^{trend} + \epsilon_{y_{t-3}^{cycle}}) + \epsilon_{y_{t-1}^{cycle}} \\ &= \dots = a_0 + a_1 t + \sum_{j=2}^{t-1} (t-j) \epsilon_{y_j^{cycle}} \end{aligned} \quad (5)$$

Similarly,

$$x_t^{trend} = b_0 + b_1 t + \sum_{j=2}^{t-1} (t-j) \epsilon_{x_j^{cycle}} \quad (6)$$

Multiplying equation 5 by  $b_1$  and equation 6 by  $a_1$  and subtracting the two equations we get

$$\begin{aligned} y_t^{trend} &= a_0 - \frac{a_1 b_0}{b_1} - \frac{a_1}{b_1} x_t^{trend} + \sum_{j=2}^{t-1} (t-j) (\epsilon_{y_j^{cycle}} - \frac{a_1}{b_1} \epsilon_{x_j^{cycle}}) \\ &= a_0 - \frac{a_1 b_0}{b_1} - \frac{a_1}{b_1} x_t^{trend} + \sum_{j=2}^{t-1} (t-j) \eta_j \end{aligned} \quad (7)$$

where  $\eta_t = (\epsilon_{y_t^{cycle}} - \frac{a_1}{b_1} \epsilon_{x_t^{cycle}})$  and its variance is

$$Var(\eta_t) = Var(\epsilon_{y_t^{cycle}}) + \frac{a_1^2}{b_1^2} Var(\epsilon_{x_t^{cycle}}) = \sigma_{y^{cycle}}^2 + \frac{a_1^2}{b_1^2} \sigma_{x^{cycle}}^2 = \sigma_\eta^2 \quad (8)$$

Therefore,  $\eta_t \sim N(0, \sigma_\eta^2)$ .

$$y_t = x_t \alpha + \sum_{j=2}^{t-1} (t-j) \eta_j \quad (9)$$

Equation 9 represents a structural response to exposure over time for individual subject data. The time specific effect term  $\sum_{j=2}^{t-1} (t-j) \eta_j$  presents an autoregressive moving average process.

Suppose there are  $n$  subjects in the study. The exposure and response of the  $i$ th individual are  $x_i(t) = x_{it}$  and  $y_i(t) = y_{it}$  at the time  $t = 1 \cdots n_i$ . Therefore, for each individual  $i$ , we have the map as the model of equation 9. When the model is estimated in a group of subjects, the parameters will differ because of random between subjects' variation. This variation can be included in the model by adding random subject effects to improve the model power. In such a context, equation 9, by modification, can be extended as

$$y_{it} = x_{it} \alpha + x_{it} u_i + \sum_{j=2}^{t-1} (t-j) \eta_{ij} + \epsilon_{it} \quad (10)$$

Here the random effect  $u_i$  is inserted to account for the subject-specific variation from the group mean. We assume that  $u_i$  has a Gaussian distribution with identical variance  $\sigma_u^2$  and is independent of  $\eta_{it}$  and  $\epsilon_{it}$ . The uncontrollable errors are denoted as  $\epsilon_{it}$  which are independent across subjects:  $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ . The random effects  $\eta_{it}$  and  $\epsilon_{it}$  are independent.

In the matrix forms of  $\mathbf{y}_i = (y_{i1} \cdots y_{it} \cdots y_{in_i})^T$ ,  $\alpha = \alpha$ ,  
 $\eta_i = (\eta_{i2} \cdots \eta_{it} \cdots \eta_{in_i-1})^T$ ,  $\epsilon_i = (\epsilon_{i1} \cdots \epsilon_{it} \cdots \epsilon_{in_i})^T$

$$\begin{aligned} \mathbf{x}_i &= \begin{pmatrix} 1 & x_t \\ \cdots & \cdots \\ 1 & x_t \end{pmatrix}_{n_i \times 2}, \mathbf{z}_{ui} = \begin{pmatrix} x_t \\ \cdots \\ x_t \end{pmatrix}_{n_i \times 1}, \\ \mathbf{z}_{\eta i} &= \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ n_i - 3 & n_i - 4 & \cdots & 1 & 0 \\ n_i - 2 & n_i - 3 & \cdots & 2 & 1 \end{pmatrix}_{n_i \times (n_i - 2)} \end{aligned} \quad (11)$$

Equation 10 becomes

$$\mathbf{y}_i = \mathbf{x}_i \alpha + \mathbf{z}_{ui} u_i + \mathbf{z}_{\eta i} \eta_i + \epsilon_i \quad (12)$$

for  $t = 1, \dots, n_i$ . Equation 12 can be further generalised as

$$\mathbf{Y}_i = \mathbf{X}_i \alpha + \mathbf{Z}_{ui} \mathbf{u}_i + \mathbf{Z}_{\eta i} \eta_i + \epsilon_i \quad (13)$$

where  $\mathbf{X}_i$  is the vector of covariates with  $\alpha$  as the corresponding vector of fixed-effects parameters,  $\mathbf{u}_i$  is the vector of subject-specific terms with  $\mathbf{Z}_{ui}$  as the corresponding vector of covariates,  $\eta_i$  is the vector of subject- and time-specific terms with  $\mathbf{Z}_{\eta i}$  as the corresponding vector of covariates. Note that we have kept the same notations for  $\alpha$ ,  $\mathbf{u}_i$ ,  $\eta_i$  and  $\epsilon_i$  in equation 13 if there is no danger of confusion. Also note that  $\mathbf{X}_i$  contains more than one exposure/predictor/covariate.

Further denote the matrices  $\mathbf{Y} = (\mathbf{Y}_1^T \cdots \mathbf{Y}_n^T)^T$ ,  $\mathbf{X} = (\mathbf{X}_1^T \cdots \mathbf{X}_n^T)^T$ ,  $\mathbf{Z}_u = (\mathbf{Z}_{u1}^T \cdots \mathbf{Z}_{un}^T)^T$ ,  $\mathbf{Z}_\eta = (\mathbf{Z}_{\eta 1}^T \cdots \mathbf{Z}_{\eta n}^T)^T$ ,  $\mathbf{u} = (\mathbf{u}_1^T \cdots \mathbf{u}_n^T)^T$ ,  $\eta = (\eta_1^T \cdots \eta_n^T)^T$ ,  $\epsilon = (\epsilon_1^T \cdots \epsilon_n^T)^T$

Equation 13 can be written as the following mixed-effects model as

$$\mathbf{Y} = \mathbf{X} \alpha + \mathbf{Z}_u \mathbf{u} + \mathbf{Z}_\eta \eta + \epsilon \quad (14)$$

where

$$\begin{pmatrix} u \\ \eta \\ \epsilon \end{pmatrix} = N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_u & 0 & 0 \\ 0 & \Sigma_\eta & 0 \\ 0 & 0 & \Sigma_\epsilon \end{pmatrix} \right), \Sigma_u = \sigma_u^2 I, \Sigma_\epsilon = \sigma_\epsilon^2 I \quad (15)$$

and  $I$  is an identity matrix of dimension  $N = \sum_{i=1}^n n_i$

Now we need to specify the subject- and time-specific variance structure  $\Sigma_\eta$  of  $\eta$ . The choice of a model for the covariance structure depends on the system. Akaike Information Criterion (AIC) (Akaike 1973) and the Bayesian Information Criterion (BIC) (Schwarz 1978) can be used to compare the models and choose the best variance-covariance structure (Littell et al. 1996).

Both statistics are based on log-likelihood values penalised for the number of parameters which are considered as "parsimony" criteria. Very often, a first order autoregressive correlation structure is specified: Within subjects, measurements close together in time are more correlated than those farther apart:  $\text{corr}(\eta_{it}\eta_{is}) = \sigma_\eta^2 \rho^{|t-s|}$ . Then the variance-covariance matrix can be written as

$$\mathbf{V} = \text{Var}(\mathbf{Y}) = \sigma_u^2 \mathbf{Z}_u \mathbf{Z}_u^T + \mathbf{Z}_\eta \Sigma_\eta \mathbf{Z}_\eta^T + \sigma_\epsilon^2 \mathbf{I} \quad (16)$$

## 2.2 Model estimation

Then the fixed-effects and random-effects parameters can be obtained (Laird and Ware 1982). These estimates are the best linear unbiased predictors (BLUPs) of the model parameters if  $\mathbf{V}$  is known. However, in many data sets, the structure of the variancecovariance structure is unknown. The most common choice is to estimate  $\mathbf{V}$  and the parameters jointly using iterative methods. In SAS's MIXED procedure for example, a modified Newton-Raphson method was adopted to numerically search the optimum values of  $\Sigma_u$  and  $\Sigma_\eta$  (Wolfinger 1993), which are restricted maximum likelihood (REML) for the fixed-effects parameters and empirical Bayes estimates for the random-effects parameters.

## 3 Model validity and illustration

Consider the hypothetical data introduced previously in Fig. 1. Define  $y_t$  as the response and  $x_t$  the time-varying exposure at the  $t$ th day,  $t = 1 \cdots n$ . The proposed model has the following exposure-response form:

$$y_t^{\text{trend}} = a_0 + a_1 x_t^{\text{trend}} + \epsilon_t \quad (17)$$

where  $y_t^{\text{trend}}$  and  $x_t^{\text{trend}}$  are calculated according to HP decomposition. The fixed-effects parameters are  $a_0$  and  $a_1$  and the random error  $\epsilon_t$ .

Table 1 provides a summary of the model estimates calculated by applying the proposed model. Results show that response is positively associated with exposure ( $p < 0.001$ ), which demonstrates that the proposed model can indeed uncover the real association between exposure and response (see also Fig.1). Table 1 also shows the fit statistics of the comparison with the tradition linear mixed-effects model. The results indicate that the proposed model has substantially better fit than the linear mixed-effects model.

## 4 Conclusion

Exposure measures are common in many fields of epidemiology as well as other branches of health sciences such as laboratory experiments of physiology and psychology. Often the mechanisms or the structural parameters

**Table 1.** Results and comparison of model fit to the hypothetical data

	Response		
	Proposed model	Linear mixed-effects model	$Pr > \chi^2$
Exposure ( $\alpha_1$ )	$1.86(p < 0.01)$	$-0.61(p < 0.01)$	
AIC (smaller is better)	-6.3	89.7	(p 0.01)

which generate the data cannot be observed directly. If the data show a remarkable dynamic and nonlinear characteristic a structure-based approach is needed to elucidate the nonlinear exposure-response relationship and explore the mechanism behind in the data. This paper presents some ways to structural modelling of such longitudinal data that can not easily be modeled by traditional statistical methods, as demonstrated by the hypothetical data. It is important to note the proposed approach includes the deseasoning method (Chatfield 2004) as a special case which is often limited to a time series only. The developed model is computationally attractive as various software packages and routines exist to perform the final obtained mixed-effects model with no extra programming effort. The model has a logical structural interpretation for the relationship between exposure and response over time. In a general framework for multivariate analysis, such relevant exposure-response patterns are common to different longitudinal data, which represent the driving forces or mechanism of the study systems. Therefore, this approach has strong relevance for the interpretation of structures of cyclic systems. Most importantly, the model parameters correspond to characterising the dependence patterns of exposure and response as the complexity of the data or pattern curves increase, as demonstrated in this study.

### Acknowledgements

This study was supported by a grant from The Finnish Work Environment Fund and Farmers Social Insurance Institution.

### References

- AKAIKE, H. (1973): Information theory and an extension of the maximum likelihood principle. In: BN. Petrov and F. Csake (Eds.): *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267-281.
- BASSEY, UN and EFFANGA, EO (2008): A linear goal programming model for the linear absolute value regression problem. *Journal of Modern Mathematics and Statistics* 2(3), 123-125.
- CHATFIELD, C. (2004): *The Analysis of Time Series: An Introduction*. CRC Press, Boca Raton, FL.
- DAVIDIAN, M. and GILTINAN, DM. (1995): *Nonlinear models for repeated measurement data*. Monographs on Statistics and Applied Probability 62. Hapman & Hall/CRC.

- HODRICK, R. and PRESCOTT, EC. (1997): Post war business cycles: An empirical investigation. *Journal of Money Credit and Banking* 24, 1-16.
- KYDLAND, FE. and PRESCOTT, EC. (1990): Business cycles: Real facts and a monetary myth. *Federal Reserve Bank of Minneapolis Quarterly Review* 14, 318.
- LAIRD, NM and WARE, JH. (1982): Random-effects models for longitudinal data. *Biometrics* 38(4), 963-974.
- LITTLE, RC. and MILLIKEN, GA and STROUP, WW. (1996): *SAS System for Mixed Models*. SAS Institute Inc., Cary.
- PROIETTI, T. (2007): Signal extraction and filtering by linear semiparametric methods. *Computational Statistics & Data Analysis* 52, 935-958.
- SCHWARZ, G. (1978): Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.
- WANG, S. (2002): Nonlinear pattern hypothesis generation for data mining. *Data & Knowledge Engineering* 40(3), 273-283.
- WOLFINGER, RD. (1993): Covariance structure selection in general mixed models. *Communications in Statistics. Simulation and Computation* 22, 1079-1106.

# On Empirical Composite Likelihoods

Nicola Lunardon, Francesco Pauli, and Laura Ventura

Department of Statistics

Via C. Battisti 241, Padova, Italy

*lunardon@stat.unipd.it, fpauli@stat.unipd.it, ventura@stat.unipd.it*

**Abstract.** Composite likelihood functions are convenient surrogates for the ordinary likelihood, when the latter is too difficult or even impractical to compute, and they may be more robust to model misspecification. One drawback of composite likelihood methods is that the composite likelihood analogue of the likelihood ratio statistic does not have the standard  $\chi^2$  asymptotic distribution.

Invoking the theory of unbiased estimating equations, this paper proposes and discusses the computation of the empirical likelihood function from the unbiased composite scores. Two Monte Carlo studies are performed in order to assess the finite-sample performance of the proposed empirical composite likelihood procedures.

**Keywords:** empirical likelihood, estimating function, likelihood methods, pairwise likelihood, pseudo-likelihood

## 1 Introduction

In various modern applications, such as models with a complex dependence structure, classical likelihood-based methods may encounter both theoretical and computational problems. These are due to the difficulty – or even impracticability – of specifying the full likelihood function. In these situations, it is possible to resort to alternative inferential methods that are based on approximate likelihoods derived by combining marginal distributions (see e.g. Cox and Reid (2004) and Varin (2008)).

The notational setup is the following. Let  $Y$  be a  $q$ -dimensional random vector with probability  $f(y; \theta)$ , with  $\theta \in \Theta \subseteq \mathbb{R}^d$ ,  $d \geq 1$ . Let  $y = (y^{(1)}, \dots, y^{(n)})$  be a random sample of size  $n$  from  $Y$ . Suppose that there is a significant difficulty in evaluating  $f(y; \theta)$ , and the corresponding likelihood  $L(\theta)$ , but that we may compute likelihoods for a set of measurable events  $\{A_i; i = 1, \dots, m\}$  in the sample space. In this situation, we may derive a composite likelihood defined as the weighted product of the likelihoods corresponding to each  $A_i$ ,

$$L_c(\theta) = L_c(\theta; y) = \prod_{i=1}^m f(y \in A_i; \theta)^{w_i}, \quad (1)$$

where  $w_i, i = 1, \dots, m$ , are non-negative weights, which do not depend on the parameter  $\theta$  or on  $Y$ . The term composite likelihood for “product of likelihood” constructions has been coined by Lindsay (1988). Composite likelihood contains, and thus generalizes, the usual ordinary likelihood, as well as many other alternatives, such as the pseudo-likelihood of Besag (1974), the partial likelihood of Cox (1975) and the pairwise likelihood (Cox and Reid (2004)). The validity of using the composite likelihood to perform inference about  $\theta$  can be justified from the standpoint of unbiased estimating functions or the Kullback-Leibler criterion (for details, see Lindsay (1988), Cox and Reid (2004) and Varin (2008)).

One drawback with composite likelihood methods is that the null distribution of the composite analogue of the likelihood ratio statistic does not converge to the standard  $\chi_d^2$  distribution. Adjustments of the composite likelihood ratio to approximate the usual  $\chi^2$  distribution have been proposed (see e.g. Geys et al. (1999)), but the quality of these approximations may be poor. By invoking the theory of unbiased estimating equations, we investigate an alternative approach to recover the standard asymptotic distribution, which is based upon the composite score function. In particular, we propose and discuss how to compute the empirical likelihood from composite scores. The main appeal of this approach is that only unbiasedness of the composite score is required to obtain a standard chi-squared distribution for the empirical composite likelihood ratio statistic. Two Monte Carlo studies are performed in order to assess the finite-sample performance of the proposed empirical composite likelihood procedures.

The paper is organized as follows. Section 2 reviews briefly definitions and properties of composite likelihoods. The proposed empirical composite likelihood is discussed in Section 3. Section 4 presents the results of two Monte Carlo studies, which were designed to assess the behaviour of our approach. Some final remarks conclude the paper.

## 2 Background on composite likelihood methods

The validity of inference about  $\theta$  using the composite likelihood can be justified invoking the theory of unbiased estimating functions.

Under broad assumptions, the maximum composite likelihood estimator  $\hat{\theta}_c$  is the solution of the composite score equation

$$s(\theta) = \nabla \log L_c(\theta) = \sum_{i=1}^m s_i(\theta) , \quad (2)$$

with  $s_i(\theta) = w_i \partial \log f(y \in A_i; \theta) / \partial \theta$ ,  $i = 1, \dots, m$ . The composite score  $s(\theta)$  is unbiased, i.e.  $E[s(\theta)] = 0$ , since it is a linear combination of valid score functions. Moreover, the maximum composite likelihood estimator  $\hat{\theta}_c$  is consistent and approximately normal with mean  $\theta$  and variance  $V(\theta) =$



$H(\theta)^{-1}J(\theta)(H(\theta)^T)^{-1}$ , where  $H(\theta) = E[-\partial s(\theta)/\partial \theta^T]$  and  $J(\theta) = \text{var}[s(\theta)]$ . The matrix  $G(\theta) = V(\theta)^{-1}$  is known as Godambe information (Godambe (1960)). The form of  $V(\theta)$  is due to the failure of the second Bartlett identity since, in general,  $H(\theta) \neq J(\theta)$ .

Composite Wald-type or score-type test statistics based on  $L_c(\theta)$  are straightforward to derive using consistent estimates of the matrices  $H(\theta)$  and  $J(\theta)$ , and present the standard  $\chi^2$  asymptotic null distribution (see Varin (2008) for a detailed discussion). On the contrary, the composite analogue of the likelihood ratio statistic given by  $W_c(\theta) = 2(\ell_c(\hat{\theta}_c) - \ell_c(\theta))$ , with  $\ell_c(\theta) = \log L_c(\theta)$ , does not follow asymptotically the standard  $\chi_d^2$  null distribution. Indeed, the null distribution of  $W_c(\theta)$  converges to a linear combination of independent  $\chi_1^2$  distributions

$$W_c(\theta) \xrightarrow{d} \sum_{i=1}^d \lambda_i Z_i^2, \quad (3)$$

where the  $Z_i^2$  ( $i = 1, \dots, d$ ) are independent  $\chi_1^2$  random variables and the coefficients  $\lambda_1, \dots, \lambda_d$  are the eigenvalues of the matrix  $H(\theta)^{-1}J(\theta)$ . Geys et al. (1999) propose to use  $W_c(\theta)/\bar{\lambda}$ , where  $\bar{\lambda}$  is the average of the eigenvalues, in order to approximate the usual asymptotic  $\chi_d^2$  distribution. Hanfelt and Liang (1995) suggest to substitute the average of the eigenvalues with the consistent estimator  $\bar{\lambda} = \text{tr}(H(\hat{\theta}_c)^{-1}J(\hat{\theta}_c))/d$ . However, these simple corrections may be inaccurate because they correct only the mean of  $W_c(\theta)$ .

Note that in the special case where inference focuses on a scalar parameter  $\theta$  (or on a scalar component  $\psi$  of  $\theta$ ), asymptotically correct  $\chi_1^2$  inference is possible since the adjusted composite likelihood ratio statistic (or the profile adjusted composite likelihood ratio statistic)

$$W_c^\dagger(\theta) = W_c(\theta)/\bar{\lambda} \quad (4)$$

converges in distribution to a  $\chi_1^2$  (see e.g. Hanfelt and Liang (1995)). When  $d > 1$  the parametric bootstrap is a more valid alternative to evaluate the distribution of  $W_c(\theta)$  (Aerts and Claeskens (1999)).

### 3 Empirical composite likelihood ratio

In many situations of practical interest, we may derive a pseudo-likelihood function for inference about  $\theta$  from a very general estimating function, by computing the empirical likelihood function (see Owen (2001) as a general reference). The empirical likelihood is a nonparametric tool which allows to obtain pseudo-likelihoods in several contexts, which include inference for dependent data. The main appeal of the empirical likelihood is that only unbiasedness of the estimating equations is required to obtain a standard asymptotic  $\chi^2$  distribution for the empirical likelihood ratio statistic or for its profile counterpart.

The empirical composite likelihood ratio statistic for  $\theta$ , derived from the composite score equation (2), is

$$W_e(\theta) = 2 \sum_{i=1}^m \log \{1 + \lambda^T s_i(\theta)\} , \quad (5)$$

where the Lagrangian multiplier  $\lambda = \lambda(\theta)$  satisfies

$$\frac{1}{m} \sum_{i=1}^m \frac{s_i(\theta)}{(1 + \lambda^T s_i(\theta))} = 0 .$$

When inference about a scalar component  $\psi$  of  $\theta = (\psi, \omega)$  is desired, where  $\omega$  is a  $(d-1)$ -dimensional nuisance parameter, a profile version of  $W_e(\theta)$  can be computed,

$$W_{ep}(\psi) = \inf_{\omega} W_e(\psi, \omega) . \quad (6)$$

Under suitable regularity conditions (see for instance Adimari and Guolo (2010)), it can be shown that:

- ◇ when  $d = 1$ , we have  $c W_e(\theta) \xrightarrow{d} \chi_1^2$ , with  $c = H(\theta)^{-1} J(\theta)$ ;
- ◇ when  $d > 1$ , we have  $c_p W_{ep}(\psi) \xrightarrow{d} \chi_1^2$ , where  $c_p = H(\theta)^{jj} J_{jj}(\theta)$ , where  $H(\theta)^{jj}$  and  $J_{jj}(\theta)$  denote the  $(\theta_j, \theta_j)$ -components of  $H(\theta)^{-1}$  and  $J(\theta)$ , respectively.

In view of these results, in the presence of dependent data, the empirical composite likelihood ratio (5) can be adjusted so that the  $\chi^2$  approximation holds. The solution is to find consistent estimates of  $H(\theta)$  and  $J(\theta)$ . When  $d > 1$ ,  $W_e(\theta)$  may be adjusted using  $\text{tr}(H(\hat{\theta}_c)^{-1} J(\hat{\theta}_c))/d$ .

From the computational point of view, the calculation of  $W_e(\theta)$  involves both numerical methods to solve equations and optimization algorithms, or nested optimization algorithms (see Owen (2001)).

## 4 Monte Carlo studies

We now discuss two examples in order to compare the finite-sample behaviour of the inferential procedures based on  $W_c^\dagger(\theta)$  and  $W_e(\theta)$ . In both examples, we focus on a particular composite likelihood, i.e. on the pairwise likelihood (see e.g. Cox and Reid (2004)),

$$L_p(\theta) = \prod_{i=1}^n \prod_{h=1}^{q-1} \prod_{k=h+1}^q f(y_h^{(i)}, y_k^{(i)}; \theta) , \quad (7)$$

where  $f(\cdot, \cdot; \theta)$  are the bivariate marginal densities. Thus, the pairwise likelihood (7) is the weighted product of the likelihoods corresponding to each single bivariate contribution, and is a particular case of (1). The associated pairwise score function is of the form (2) with  $m = q(q-1)n/2$ .

#### 4.1 Multivariate normal distribution

Let us focus on the correlation coefficient  $\rho$  for a multivariate normal distribution, as in Cox and Reid (2004). In this case, the full likelihood function  $L(\rho)$  is available and it is possible to compare the complete likelihood ratio statistic  $W(\rho)$ , based on  $L(\rho)$ , with  $W_c^\dagger(\rho)$  and  $W_e(\rho)$ , based on the pairwise likelihood (7).

Let  $Y$  be a  $q$ -variate normal random variable, with standard normal margins and  $\text{corr}(Y_r, Y_s) = \rho$ , for  $r, s = 1, \dots, m$ , with  $r \neq s$ . Following Cox and Reid (2004), given the sample  $y = (y^{(1)}, \dots, y^{(n)})$  from  $Y$ , the pairwise loglikelihood  $\ell_p(\rho) = \log L_p(\rho)$  is

$$\ell_p(\rho) = -\frac{nq(q-1)}{4} \log(1-\rho^2) - \frac{q-1+\rho}{2(1-\rho^2)} SS_W - \frac{(q-1)(1-\rho)}{2(1-\rho^2)} \frac{SS_B}{m},$$

where

$$SS_W = \sum_{i=1}^n \sum_{r=1}^q (y_r^{(i)} - \bar{y}^{(i)})^2, \quad SS_B = \sum_{i=1}^n \bar{y}^{(i)2}$$

and  $\bar{y}^{(i)} = \sum_{r=1}^q y_r^{(i)} / q$ . The associated score function is

$$s(\rho) = \frac{nq(q-1)\rho}{2(1-\rho^2)} - \frac{1+\rho^2+2(q-1)\rho}{2(1-\rho^2)^2} SS_W + \frac{(q-1)(1-\rho)^2}{2(1-\rho^2)^2} \frac{SS_B}{q}.$$

Figure 1 gives the empirical coverages for equitailed confidence intervals for  $\rho$  based on  $W(\rho)$ ,  $W_c^\dagger(\rho)$  and  $W_e(\rho)$  (based on 20000 Monte Carlo runs). Note that both  $W_c^\dagger(\rho)$  and  $W_e(\rho)$  are multiplied by the same scale factor  $\tilde{\lambda}$ , which is evaluated at the composite maximum likelihood estimate. The results in Figure 1 show that the proposed empirical likelihood statistic  $W_e(\rho)$  performs quite well and is close to  $W(\rho)$  for moderate and large sample sizes. For  $n \leq 50$ ,  $W_e(\rho)$  outperforms on  $W_c^\dagger(\rho)$ .

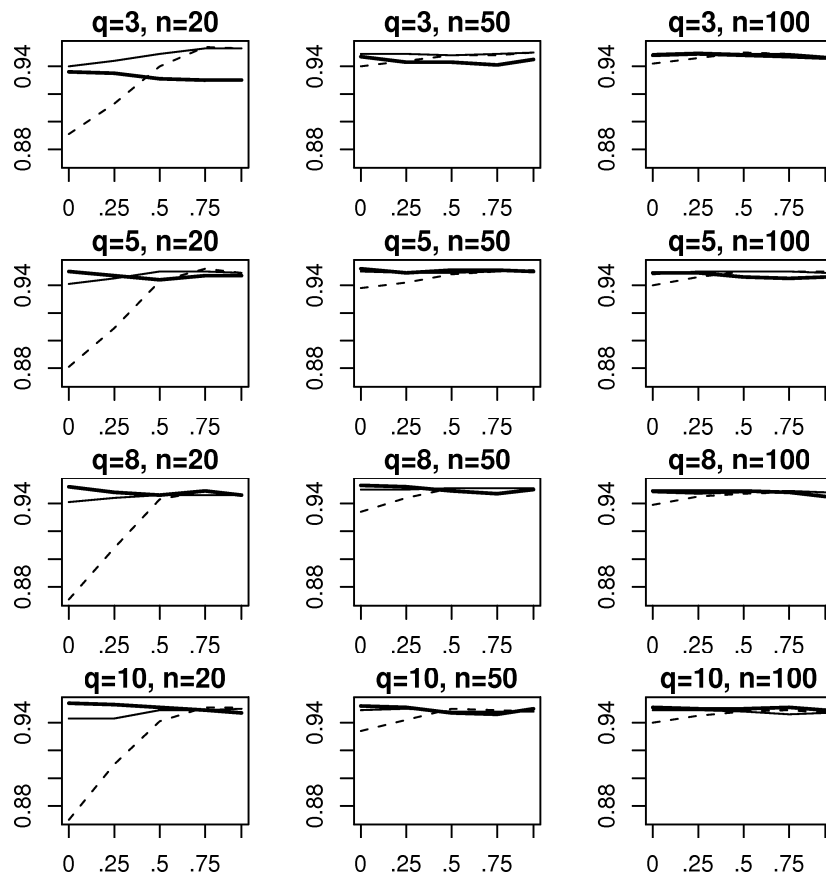
#### 4.2 Binary data

The pairwise likelihood (7) is particularly useful for modelling correlated binary outcomes, as discussed in Le Cessie and Van Houwelingen (1994). This kind of data arises, e.g., in the context of repeated measurements on the same subject, where maximum likelihood analysis involves multivariate integrals whose dimension equals the cluster sizes.

Let us focus on a multivariate probit model with logistic marginals and constant cluster sizes. In this case, the pairwise loglikelihood is

$$\ell_p(\theta) = \sum_{i=1}^n \prod_{h=1}^{q-1} \prod_{k=h+1}^q \log \Pr(Y_j^{(i)} = y_j^{(i)}, Y_k^{(i)} = y_k^{(i)}; \theta) \quad (8)$$

(see Le Cessie and Van Houwelingen (1994) and Kuk and Nott (2000)). Pairwise likelihood inference is much simpler than using the full likelihood since it



**Fig. 1.** Correlation coefficient: empirical coverages of equitailed confidence intervals with 95% nominal level based on  $W(\rho)$  (solid lines),  $W_c^\dagger(\rho)$  (dashed lines) and  $W_e(\rho)$  (thick lines).

involves only bivariate normal integrals. For instance (see also Renard et al., 2004), we have  $\Pr(Y_{ij} = 1, Y_{ik} = 1; \theta) = \Phi_2(\xi_{ij}, \xi_{ik}; \rho)$ , where  $\Phi_2(\cdot; \rho)$  denotes the standard bivariate normal distribution function with correlation coefficient  $\rho$  and  $\xi_{ij} = x_{ij}^\top \beta / \sigma$ , with  $\beta$  unknown regression coefficient,  $\sigma$  unknown scale parameter and  $x_{ij}$  fixed constants.

Table 2 gives the empirical coverages for equitailed confidence intervals for  $\theta = (\beta_0, \beta_1, \rho)$  based on  $W_c(\theta)$ , given in (3),  $W_c^\dagger(\theta)$  and  $W_e(\theta)$  (based on 20000 Monte Carlo trials). The derivatives of (8) are not available in closed form and numerical evaluation of all the likelihood quantities involved in the simulation study was used. The results in Table 2 show that also in this example the proposed empirical likelihood statistic  $W_e(\theta)$  performs similarly to  $W_c^\dagger(\theta)$ .

**Table 1.** Binary data: empirical coverages of equitailed confidence intervals with 95% nominal level.

$n$	30	50	80	30	50	80	30	50	80
	$q = 3$			$q = 6$			$q = 10$		
				$\rho = 0.25$					
$W_c(\theta)$	0.924	0.941	0.946	0.930	0.943	0.944	0.938	0.943	0.942
$W_c^\dagger(\theta)$	0.908	0.934	0.940	0.910	0.932	0.935	0.917	0.922	0.924
$W_e(\theta)$	0.897	0.932	0.940	0.908	0.937	0.937	0.924	0.927	0.930
				$\rho = 0.50$					
$W_c(\theta)$	0.922	0.941	0.938	0.929	0.942	0.946	0.942	0.943	0.945
$W_c^\dagger(\theta)$	0.904	0.935	0.932	0.910	0.932	0.938	0.920	0.925	0.927
$W_e(\theta)$	0.891	0.931	0.932	0.908	0.940	0.935	0.928	0.930	0.930
				$\rho = 0.75$					
$W_c(\theta)$	0.919	0.937	0.942	0.923	0.944	0.944	0.939	0.940	0.943
$W_c^\dagger(\theta)$	0.898	0.929	0.936	0.902	0.932	0.935	0.914	0.917	0.927
$W_e(\theta)$	0.871	0.922	0.935	0.898	0.935	0.938	0.918	0.921	0.930

## 5 Final remarks

In this paper we explored an empirical likelihood function derived from a composite score function, by investigating the performance of the corresponding empirical likelihood ratio test on two specific examples. The simulation results presented in Section 4 show that the proposed  $W_e(\theta)$  statistic proves useful to make inference in complex models. It furthermore represents a valid alternative to the adjusted composite likelihood ratio statistic  $W_c^\dagger(\theta)$ , whose approximation by the usual  $\chi^2$  distribution may be unsatisfactory. For higher values of  $q$ , the proposed empirical composite likelihood ratio  $W_e(\theta)$  leads to more accurate confidence intervals than  $W_c^\dagger(\theta)$ . In order to generalize these

results beyond the two specific examples discussed here, we plan to investigate the theoretical properties of  $W_e$  and to perform a more comprehensive simulation study.

## References

- ADIMARI, G. and GUOLO, A. (2010): A note on the asymptotic behaviour of empirical likelihood statistics. *Statistical Methods & Applications*, to appear.
- AERTS, M. and CLAESKENS, G. (1999): Bootstrapping pseudolikelihood models for clustered binary data. *Annals of the Institute of Statistical Mathematics* 51, 515–530.
- BESAG, J.E. (1974): Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B* 4, 192–236.
- COX, D.R. (1975): Partial likelihood. *Biometrika* 62, 269–276.
- COX, D.R. and REID, N. (2004): A note on pseudolikelihood constructed from marginal densities. *Biometrika* 91, 729–737.
- GEYS, H., MOLENBERGHS, G. and RYAN, L.M. (1999): Pseudolikelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association* 94, 734–745.
- GODAMBE, V.P. (1960): An optimum property of regular maximum likelihood equation. *The Annals of Statistics* 31, 1208–1211.
- HANFELT, J.J. and LIANG, K.Y. (1995): Approximate likelihood ratios for general estimating functions. *Biometrika* 82, 461–477.
- KUK, A.Y.C. and NOTT, D.J. (2000): A pairwise likelihood approach to analyzing correlated binary data. *Statistics and Probability Letters* 47, 329–335.
- LE CESSIE, S. and VAN HOUWELINGEN, J.C. (1994): Logistic regression for correlated binary data. *Applied Statistics* 43, 95–108.
- LINDSAY, B.G. (1988): Composite likelihood methods. *Contemporary Mathematics* 80, 221–240.
- OWEN, A.B. (2001): *Empirical likelihood*. Chapman and Hall, London.
- RENARD, D., MOLENBERGHS, G., GEYS, H. (2004): A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis* 44, 649–667.
- VARIN, C. (2008): On composite marginal likelihoods. *Advances in Statistical Analysis* 92, 1–28.

# A Fast Parsimonious Maximum Likelihood Approach for Predicting Outcome Variables from a Large Number of Predictors

Jay Magidson

Statistical Innovations Inc.  
Belmont, Massachusetts, United States *jay@statisticalinnovations.com*

**Abstract.** A new model with  $K$  correlated components is presented for predicting outcome variables where the number of predictors  $G$  may exceed the total sample size  $N$ . A fast maximum likelihood algorithm provides closed-form expressions for the model parameters and statistical tests for determining the number of components. We also propose a fully integrated step-down variable selection algorithm, at each step eliminating the least important predictor based on a new measure of importance. Validated results from 2 examples suggest that the methods can provide good predictions outside the sample, especially with  $K = 3$  or  $4$ .

**Keywords:** correlated component regression, variable selection, gene expression, high dimensional data, latent class analysis

## 1 Background and General approach

Traditional regression models with  $G$  predictors become unstable due to multicollinearity and predict new cases poorly as  $G$  approaches the sample size  $N$ , and can not predict at all for  $G > N$ . A new model with  $K$  correlated components is proposed that is asymptotically equivalent to the traditional model when  $K = G$ , and for  $K$  small (typically  $K < 5$ ), it can handle  $G > N$  despite the large number of total model parameters. Variables loading significantly on the first component ('prime predictors') have direct effects on the outcome variable  $Z$ . Variables loading significantly on the second component are primarily those having no direct effect on  $Z$  but improve prediction by suppressing irrelevant variation in the prime predictors ('proxy predictors'). The approach, called Correlated Component Regression (CCR) also incorporates a variable reduction procedure. CCR is illustrated on two data sets:

- a. Publicly available colon cancer vs. normals data containing  $G = 2000$  continuous gene expression predictors - validated results show that a model based on only  $G^* = 5$  genes provides strong prediction.
- b. Survival data for 2 groups of melanoma patients.

We begin with  $Z$  dichotomous and then describe CCR variants and extensions where  $Z$  is ordinal, continuous, and nominal,  $Z$  indicating occurrence

of an event (e.g., death), and  $Z$  denoting multiple outcome variables. Let  $Y_1, Y_2, \dots, Y_G$  denote  $G$  continuous predictors. The  $K$ -component model is a generalized linear model (GLZ), where the linear portion (ignoring the intercept) is obtained as a weighted sum of  $K \leq G$  components  $S_1, S_2, \dots, S_K$ , each component itself being an exact linear combination of the predictors,  $S_k = \sum_{g=1}^G \lambda_{kg} Y_g$ . For concreteness, we initially assume that  $Z$  is dichotomous, and denote the predicted logit obtained from the  $K$ -component model as  $\text{Logit.K}(Z)$ .

$$\text{Logit.K}(Z) = \alpha + \sum_{k=1}^K b_k^{(K)} S_k \quad (1)$$

$$= \alpha + \sum_{g=1}^G \beta_g^{(K)} Y_g \quad \text{where} \quad \beta_g^{(K)} = \sum_{k=1}^K b_k^{(K)} \lambda_{kg} \quad (2)$$

More generally, in a survival analysis a log-hazards rate might be used in the left-hand side of equation (1) with a time varying intercept, or the conditional expectation of a continuous outcome variable as a linear regression extension. Estimation of the loadings,  $\lambda_{kg}$ , and weights,  $b_k^{(K)}$ , proceeds as follows.

**Step 1:** Estimate loadings  $\lambda_{1g}$  defining the first component,  $S_1 = \sum_{g=1}^G \lambda_{1g} Y_g$ , using a maximum likelihood method, each term,  $\lambda_{1g} Y_g$ , corresponding to a GLZ prediction of  $Z$  obtained from the  $g$ th predictor (omitting the intercept). When  $Z$  is dichotomous, each  $\lambda_{1g}$  corresponds to a simple log-odds ratio, the odds of  $Z=1$  vs.  $Z=0$  being  $\exp(\lambda_{1g})$  times as high for a case having a 1 unit higher value on  $Y_g$  than another case. The CCR algorithm estimates the loadings, one at a time, as follows: Perform  $G$  separate linear regressions, the  $g$ th of which is the regression of  $Y_g$  on  $Z$ ,

$$Y_g = \alpha_{0g} + \lambda'_{0g} Z + \varepsilon_{0g} \quad (3)$$

Obtain an initial estimate for the loading  $\lambda_{1g}$ , denoted  $\hat{\lambda}_{0g}$ , by dividing the estimate for  $\hat{\lambda}_{0g}$  by the mean squared error,  $MSE(\varepsilon_{0g})$ , obtained from the  $g$ th regression in (3):

$$\hat{\lambda}_{0g} = \hat{\lambda}'_{0g} / MSE(\varepsilon_{0g}) \quad (4)$$

Under the assumption that the error  $\varepsilon_{0g}$  is normally distributed with constant variance,  $\hat{\lambda}_{0g}$  is a maximum likelihood estimate for the log-odds ratio  $\lambda_{0g}$  in the simple logistic regression model  $\text{Logit}(Z|Y_g) = \alpha_g + \lambda_{0g} Y_g$  (Lyles, et. al. 2009). Using these  $G$  loadings,  $\hat{\lambda}_{0g}$ ,  $g=1, 2, \dots, G$  the Naïve Bayes estimator for the component is  $S_0 = \sum_{g=1}^G \hat{\lambda}_{0g} Y_g$ , which is  $G$  times the average (or sum) of  $G$  simple logistic regression model predictions (ignoring intercepts). The first



component  $S_1$  is then obtained as a standardized version of  $S_0$  as follows: Perform a linear regression of  $S_0$  on  $Z$ :  $S_0 = \alpha_0 + b'_0 Z + \varepsilon_0$ . Compute:  $\lambda_{1g} = b_0 \lambda_{0g}$  and  $S_1 = b_0 S_0$  where  $b_0 = \hat{b}'_0 / \text{MSE}(\varepsilon_0)$ . Predictions from the 1-component model are given by  $\text{Logit.1}(Z) = \alpha + S_1$ . Since  $\beta_g^{(K)} = \sum_{k=1}^K b_k^{(K)} \lambda_{kg}$ , the standardization of  $S_0$  to  $S_1$  is such that  $b_1^{(1)} = 1$ , which allows the  $g$ th loading on component  $S_1$  to serve also as  $\beta_g^{(1)}$ , the weight for the  $g$ th predictor in the 1-component model.

**Step 2:** Determine component  $S_2$  such that it maximally improves prediction of  $Z$  over and above that provided by  $S_1$  alone. The loadings on  $S_2$ , denoted  $\lambda_{2g}$ , are estimated by a maximum likelihood method, where

$$S_2 = \sum_{g=1}^G \lambda_{2g} Y_g \quad (5)$$

The CCR algorithm proceeds as follows: Perform  $G$  separate linear regressions, the  $g$ th of which is the regression of  $Y_g$  on  $Z$  and  $S_1$ , providing an estimate for  $\lambda'_{2g}$  in (6):

$$Y_g = \alpha_{2g} + \lambda'_{2g} Z + \gamma_1 S_1 + \varepsilon_{2g} \quad (6)$$

Also get the associated p-value testing the null hypothesis  $H_0(1.g): \lambda'_{2g} = 0$ , which serves as the equivalent test for the loading  $\lambda_{2g} = 0$  where  $\lambda_{2g}$  in (5) is obtained by substituting the estimates for  $\lambda'_{2g}$  and MSE obtained in (6) into the equation  $\lambda_{2g} = \lambda'_{2g} / \text{MSE}(\varepsilon_{2g})$ . Obtain Logit.2 by estimating the b-weights in (1) corresponding to the logistic regression of  $Z$  on  $S_1$  and  $S_2$ :

$$\text{Logit.2}(Z|S_1, S_2) = \alpha + b_1^{(2)} S_1 + b_2^{(2)} S_2 \quad (7)$$

As in step 1 when we obtained Logit.1, we do not bother to estimate the intercept and the CCR algorithm obtains estimates for the b-coefficients as follows: Estimate the two linear regression models:

$$S_1 = a_1 + b'_{1.2} Z + d_1 S_2 + \varepsilon_1 \quad (8)$$

$$S_2 = a_2 + b'_{2.1} Z + d_2 S_1 + \varepsilon_2 \quad (9)$$

and compute:

$$b_1^{(2)} = b'_{1.2} / \text{MSE}(\varepsilon_1) \quad \text{and} \quad b_2^{(2)} = b'_{2.1} / \text{MSE}(\varepsilon_2) \quad (10)$$

From equation (2) the composite weight for the  $g$ th constituent in the  $K = 2$  component model is:

$$\beta_g^{(2)} = b_1^{(2)} \lambda_{1g} + b_2^{(2)} \lambda_{2g} \quad (11)$$

where  $\lambda_{1g}$  was obtained in Step 1, and  $\lambda_{2g}$  in Step 2. If the p-value associated with  $H_0: b'_{2,1} = 0$  is non-significant, the 2nd component does not provide a significant improvement over the 1-component model, and the algorithm terminates with  $K^*=1$ . Otherwise, return to Step 2 with  $K=K+1$ . For example, for  $K=3$  determine component  $S_3$  that improves prediction of  $Z$  over that provided by  $S_1$  and  $S_2$  alone. The algorithm terminates with the  $K^*$ -component model if the p-value associated with 1 or more  $b_k^{(K^*+1)}$  is not statistically significant, in which case we say that the  $K^*$ -component model has achieved ‘sequential independence’, the source of residual correlation between the predictors unable to improve prediction further.

Each predictor  $Y_g$  may have a different variance. From equation (2), reproduced here: ( $\text{Logit.K}(Z) = \alpha + \sum_{g=1}^G \beta_g^{(K)} Y_g$ ), it is clear that standardizing  $Y_g$  by dividing by its standard deviation results in the associated composite weight being multiplied by the standard deviation. Thus, we define standardized composite weights  $\beta_g^{*(K)} = \sigma_g \beta_g^{(K)}$ , the absolute value of which we use as a measure of importance of variable  $g$  in the  $K$ -component model. Standardized loadings can be obtained by multiplying the corresponding raw loadings by the standard deviations:  $\lambda_{kg}^* = \sigma_g \lambda_{kg}$ .

We propose the following strategy to reduce the number of predictors from  $G$  to  $G^*$ : Given a value for  $K^*$  (by default  $K^* = 4$ ), eliminate the predictor variable with the lowest measure of importance  $|\beta_g^{*(K^*)}|$ , and re-estimate the 1-component through  $K^*$ -component models with  $G-1$  predictors, determining again the lowest value for  $K$  for which sequential independence is achieved, and setting  $K^*$  to that value. Repeat this variable reduction process, eliminating 1 variable at a time until some stopping criteria is reached. For example, the stopping rule might be when a certain reduction in a validation performance measure occurs such as 1) AUC = the Area Under the ROC Curve, or 2) AMPS = Average Model Performance Statistic =  $E(\text{Logit.K}|Z = 1) - E(\text{Logit.K}|Z = 0)$ , as measured in validation data if available.

This measure of predictor importance can also be used as a step-down criteria with other component models such as PLS regression. It has advantages over other measures of importance. For a summary of weaknesses in other measures of importance, see Gromping (2009).

Generally,  $S_2$  is not predictive of  $Z$  directly, but is correlated with  $S_1$ , and improves prediction by suppressing irrelevant variation in  $S_1$ . In our analyses with gene expression data, we found that proxy variables are prevalent and contribute most among all predictors in the model. For example, with respect to a 6-gene model for early detection of prostate cancer, the mean gene expression difference between the cancer and normal subjects for the single most important predictor was found to be nil (Ross, et. al. 2010). For that model, it may be that the ‘proxy gene’ enhances the predictive effects of two ‘prime genes’ by (implicitly) predicting and subtracting out the expression

for each prime gene at an earlier time when the cancer subjects were normal, thus converting gene expression for the prime genes to the more predictive ‘change in expression’ on these genes. Formally, if component  $S_2$  in equation (7) is a pure proxy,  $S_2$  has no direct effect on  $Z$  and the linear regression of  $S_1$  on  $S_2$  has slope  $m = -b_2/b_1$  so that  $\text{Logit.2}(Z) = \alpha + b_1(S_1 - mS_2)$ ,  $S_2$  enhancing the predictive power of  $S_1$ .

Tables 1 and 2 show the results after application of the CCR variable selection algorithm to a training sample of  $N=41$  subjects to reduce the number of predictors from 2000 to 5 based on  $K^*=4$ . The goal is to discriminate between  $Z=1$  ( $N=40$  Colon cancer subjects) and  $Z=0$  ( $N=22$  Normal subjects). Table 2 shows that the predictor Hsa.25748 is a proxy variable, since it does not load significantly on  $S_1$  but has the sole significant loading on  $S_2$ , which itself acts as a proxy. Table 1 shows that Hsa.25748 is one of the most important variables in the  $K$ -component model, for  $K = 3$  or 4. More generally, the powerful enhancement effects of proxy (suppressor) variables is well documented (Friedman and Wall, 2005). Although common industry practice is to select from a large number of potential predictors only those predictive of the outcome variable(s), this strategy appears to be misguided, unnecessarily reducing the predictive power of a model by excluding proxy genes.

For the Colon Cancer Data, the 2-component model provides perfect prediction among the training data, and misclassifies only 3 in the validation data, 2 of which have been misclassified by many other models based on all 2000 genes. Results are similar for the 3-, 4-, and 5-component models. A larger number of misclassifications were reported from various PLS regression routines based on all 2000 genes (Fort, et. al., 2004).

Despite the fact that the 2-component model classifies all training subjects perfectly, the AMP statistic measured on the validation sample improves further when  $K$  increases from 2 to 3:  $\text{AMPS}(1) = 8.5$ ,  $\text{AMPS}(2) = 16.0$ ,  $\text{AMPS}(3) = 17.4$ ,  $\text{AMPS}(4) = 17.2$ , the large improvement from  $\text{AMPS}(1)$  being attributable largely to the enhancement effect of the suppressor variable. The  $p$ -values for the component weights (Table 2) show that inclusion of the 3rd component provides a marginally significant improvement ( $p=.04$ ) while the improvement for the 4th component is non-significant ( $p=.54$ ).

The model generalizes easily in many ways. When  $Z$  is ordinal with known category scores, say 0,  $Z^*[2], \dots, Z^*(J-1)$ , 1, the algorithm is unchanged, and for continuous  $Z$  only the division factor changes as described below <sup>1</sup>.

An equivalent form of the algorithm uses  $Z$  as the left-hand side variable and  $Y_g$  on the right and again uses least squares to estimate the parameter which is again divided by MSE. When  $Y_g$  is continuous, the division by MSE is omitted; when the equivalent form is used with  $Z$  on the left-hand side variable, rather than division by MSE, the division is by the factor  $W$  which provides the same estimate as obtained with the original form. For example,

<sup>1</sup> For  $Z$  continuous, an alternative is to use the latent class analysis extension described later, predicting the high versus low scoring classes.

**Table 1.** Standardized composite weights. Proxy gene results italicized. Output obtained from CORExpress program (Magidson, 2010)

K	beta*(K)			
	1	2	3	4
Hsa.8125	-1.5	-3.3	-4.3	-4.7
Hsa.8147	-2.6	-4.4	-5.4	-4.9
Hsa.6814	1.3	2.6	2.2	2.2
Hsa.9353	1.0	2.2	1.9	2.0
Hsa.25748	<i>-0.2</i>	<i>3.0</i>	<i>4.1</i>	<i>4.0</i>
AMPS (Training, N=41)	7.3	13.4	15.1	14.9
AMPS (Validation, N=21)	8.5	16.0	17.4	17.2

**Table 2.** p-values for loadings and component weights (b)

k	p-values for loadings (lambda)			
	1	2	3	4
Hsa.6814	0.002	0.90	0.05	0.94
Hsa.8125	0.0007	0.57	0.05	0.54
Hsa.8147	2.1E-06	0.54	0.41	0.54
Hsa.9353	0.013	0.60	0.06	0.82
Hsa.25748	0.65	4.2E-05	0.64	0.54
p-values for component weights (b)				
b(3,K)	2.18	1.06	0.36	
p	4.8E-14	1.8E-05	0.04	
b(4,K)	2.16	1.28	0.41	0.32
p	9.2E-14	0.01	0.03	0.54

when the variables are all standardized to have variance 1, when extracting  $S_2$ ,  $W = (1 - r_{ZS_1}^2)/(1 - r_{Y_1S_1}^2)$ . For a second outcome variable ZB, or for Z nominal <sup>2</sup> (say, with J = 3 categories where ZB = 1, 0 is a second dummy indicator variable), ZB is included as an additional variable on the right side of eqs. (3) and (6), separate B-components,  $S_{1B}$ ,  $S_{2B}$ , etc., are obtained along with an additional equation (7B) for LogitB.K corresponding to equation (7) with the additional components, and corresponding additional eqs. (8B), (9B), (10B) and (11B). Extension to an ordinal Z with unknown category scores can also be obtained, the scores being estimated for each component k using a baseline logit model extension (see Magidson, 1996).

Another important generalization occurs when Z is a latent variable, being defined in an earlier analysis (step 0). For example, Z may consist of two latent classes corresponding to subjects improving under a particular therapy (class 1) and those not improving (class 2), developed using multiple observed indicators of improvement. Example 2, described below, is based

<sup>2</sup> For J>3, there would be J-1 additional dummy variables, say ZB, ZC,...

on two latent classes of melanoma patients. In this extension, the algorithm is unchanged if  $Z$  is a dichotomy, each case being assigned to a class based on their modal category, or it may be extended to use posterior membership probabilities, obtained from the latent class model, as weights (see Magidson, 2005), adjusted based on modal or proportional assignment (Vermunt, 2010). The weights may be specified as case weights, replication weights, or sampling weights where the case ID is used as a primary sampling unit using the Latent GOLD program (Vermunt and Magidson, 2008), the latter approach being used below to get the appropriate p-values.

**Table 3.** Standardized composite weights. Proxy gene results italicized.

K	beta*(K)			
	1	2	3	4
Gene X	-0.20	-0.31	-0.20	-0.23
CTSD	-0.22	-0.62	-0.81	-0.82
<i>PLA2G7</i>	<i>0.06</i>	<i>0.41</i>	<i>0.32</i>	<i>0.32</i>
<i>TXNRD1</i>	<i>-0.01</i>	<i>0.49</i>	<i>0.63</i>	<i>0.64</i>

**Table 4.** p-values for loadings and component weights. Proxy gene results italicized.

k	p-values for loadings (lambda)			
	1	2	3	4
Gene X	2.0E-04	0.08	0.19	0.92
CTSD	4.4E-05	0.75	0.07	0.80
<i>PLA2G7</i>	<i>0.23</i>	<i>5.8E-04</i>	<i>0.07</i>	<i>0.92</i>
<i>TXNRD1</i>	<i>0.89</i>	<i>4.3E-08</i>	<i>0.36</i>	<i>0.92</i>
p-values for component weights (b)				
b(3,K)	3.15	0.76	0.43	
p	1.4E-16	9.0E-13	0.05	
b(4,K)	2.16	1.28	0.41	0.32
p	1.0E-13	7.1E-13	0.05	0.96

Previously, we obtained a 4-gene model that strongly predicted survival time among melanoma patients, which validated on a somewhat different melanoma population undergoing the same therapy (Gao, et. al, 2009). Here we re-analyzed that data using all  $G=169$  genes, and used the selection algorithm to obtain a 4-gene model. We first used a latent class proportional hazards model to identify 2 classes — long term survivors and others — solely based on survival time (see Vermunt, 2009), and obtained posterior membership probabilities for each class for each case. We then constructed a data file consisting of 2 records per case, one record for each class, along

with 169 gene expression variables. The resulting 4-gene model obtained from the unweighted as well as the weighted analyses consisted of 3 of the 4 genes obtained in the original analysis, the weakest (4th gene) in the original analysis being replaced by a different gene (identified as ‘gene X’ in Table 3 and Table 4 which show results from a weighted analysis). These 4-gene models performed even better than the original model, on the validation data, according to a log rank test where the risk score was used as a single covariate in a Cox model. Like the original model, the 4-genes again consisted of 2 prime and 2 ‘proxy genes’.

Acknowledgements: I am indebted to Karl Wassmann of Source MDx for his support and encouragement and for permission to use the melanoma data, to J. Alexander Ahlstrom for exceptional programming of CORExpress™, and Will Barker for his ongoing assistance. Multiple patent applications are pending regarding this technology.

## References

- FORT, G. and LAMBERT-LACROIX, S. (2004): Classification Using Partial Least Squares with Penalized Logistic Regression. *IAP-Statistics*.
- FRIEDMAN, L. and WALL, M. (2005): Graphical Views Of Suppression And Multicollinearity In Multiple Linear Regression. *American Statistician*, May 2005. Vol 59, No. 2, pp 127-136.
- GAO, F., BANKAITIS-DAVIS, D., MAGIDSON, J. and WASSMANN, K. (2010): Validation of a Cox Model based on peripheral blood gene expression measurements for melanoma patients receiving CTLA4-blockade. *forthcoming*.
- GROMPING, U. (2009): Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, November 2009. Vol 63, No. 4, pp 308-319.
- LYLES R., GUO, Y. and HILL, A. (2009): A Fresh Look at the Discrimination Function Approach for Estimating Crude or Adjusted Odds Ratios. *The American Statistician*, November 2009. Vol 63, No. 4, pp 320-327.
- MAGIDSON, J. (1996): Maximum Likelihood Assessment of Clinical Trials Based on an Ordered Categorical Response. *Drug Information Journal*, Maple Glen, PA: Drug Information Association, Vol. 30, No. 1, 143-170.
- MAGIDSON, J. (2005): An Extension of the CHAID Tree-based Segmentation Algorithm to Multiple Dependent Variables. in C. Weihs & W. Gaul, *Classification: The Ubiquitous Challenge*, 176-183. Heidelberg: Springer.
- MAGIDSON, J. (2010): *Correlated Extracted Predictors Software (CORExpress) User's Guide*. Belmont MA.: Statistical Innovations Inc.
- ROSS, R., et. al. (2010): A Whole-Blood RNA Transcript-Based Diagnostic Test Improves the Diagnosis of Prostate Cancer Compared with Prostate-Specific Antigen Alone. *forthcoming*.
- VERMUNT, J. (2010). Latent class modeling with covariates: Two improved three-step approaches. Under review.
- VERMUNT, J. (2009): Event history analysis. in R. Millsap (ed.) *Handbook of Quantitative Methods in Psychology*, 658-674. London: Sage.
- VERMUNT, J. and MAGIDSON, J. (2005): *Latent GOLD 4.0 Technical Guide*. Belmont MA.: Statistical Innovations Inc.

# A Bootstrap Method to Improve Brain Subcortical Network Segregation in Resting-State fMRI Data

Caroline Malherbe<sup>1,2,\*</sup>, Eric Bardinet<sup>3,4,5</sup>, Arnaud Messé<sup>1,2</sup>, Vincent Perlberg<sup>1,2</sup>, Guillaume Marrelec<sup>1,2</sup>, Mélanie Péligrini-Issac<sup>1,2</sup>, Jérôme Yelnik<sup>3,5</sup>, Stéphane Lehéricy<sup>2,3,4,5</sup>, Habib Benali<sup>1,2</sup>

<sup>1</sup> Inserm and UPMC Univ Paris 06, UMR\_S 678, Laboratoire d’Imagerie Fonctionnelle, 91 boulevard de l’Hôpital, Paris, France,

\* *caroline.malherbe@imed.jussieu.fr*

<sup>2</sup> Inserm, Université de Montréal, UPMC Univ Paris 06, LIneM, Laboratoire International de Neuroimagerie et Modélisation, Paris, France

<sup>3</sup> Inserm and UPMC Univ Paris 06, UMR\_S 975, CRICM, Paris, France

<sup>4</sup> UPMC Univ Paris 06, Centre for NeuroImaging Research – CENIR, Pitié-Salpêtrière Hospital, Paris, France

<sup>5</sup> CNRS, UMR 7225, CRICM, Paris, France

**Abstract.** Brain functional networks are sets of distant cortical, subcortical or cerebellar regions characterized by coherent dynamics. While spatial independent component analysis (sICA) reproducibly detects the cortical components of these networks from resting-state functional magnetic resonance imaging (fMRI) data, little is known about their subcortical (basal ganglia) components. We propose a method to detect cortico-subcortical networks across subjects. Cortical components are first detected using sICA. Subcortical components are then identified using a general linear model combined with bootstrap to ensure statistical robustness, and then compared with an atlas of the basal ganglia for validation.

**Keywords:** fMRI, functional networks, basal ganglia, sICA, bootstrap

## 1 Introduction

The basal ganglia (BG) are a set of grey matter nuclei that are located deep in the cerebral hemispheres. They have anatomical and functional connections with the cortex and are known to be involved in cortico-subcortical circuits that are critical not only for controlling motor function, but also for mediating cognition, emotions, and motivation (Purves et al. (2004)). Such loops have been extensively studied in the primate brain (Smith et al. (2004)), in which biological tracers can be used to reveal the links between the BG and the cortex.

In the human brain, using resting-state fMRI data, functional connectivity analyses attempt to study how the brain works by characterizing so-called large-scale functional networks (Bellec et al. (2006)). A large-scale functional

network is defined as a set of distant cortical, subcortical or cerebellar regions characterized by coherent dynamics (Varela et al. (2001), Beckmann et al. (2005)). At rest (i.e. for subjects lying still in the MRI scanner), such networks have been identified by various methods based on spatial independent component analysis (sICA) (Perlberg et al. (2008), Perlberg and Marrelec (2008)). However, while the cortical parts of these networks have already been described quite precisely in the literature, little has been said about the associated subcortical regions (Beckmann et al. (2005), Damoiseaux et al. (2006)). Indeed, sICA-based techniques do not explicitly account for possible signal heteroscedasticity between cortical and BG regions (Mériaux et al. (2006)). The consequence is that sICA alone can not segregate the BG in functional subregions and usually assigns large parts of the BG to a single ICA component (Damoiseaux et al. (2008)) instead of associating them with specific cortical areas.

The aim of this paper is to provide a robust method to detect precisely subcortical components in the large-scale functional networks observed in fMRI. To do so, we propose the following two-step approach. We first use sICA to extract the cortical components of the networks from resting-state fMRI data in which the BG are masked out. We then detect the subcortical components corresponding to these cortical regions using a general linear model. To find subcortical regions that are reproducible at the group level, robust statistical inference is needed across subjects. However, the sample size in fMRI acquisitions is usually quite low by statistical standards, i.e. typically less than 30 subjects are scanned, and the underlying distribution of the studied parameters is unknown and could be far from Gaussian. Therefore, group statistics conventionally based on Student  $t$  test are not expected to provide robust results. Instead, we resort to statistical inference using a bootstrap technique to determine significance levels and select BG regions that are robustly found across subjects. Each functional network is finally defined as the union of its cortical and subcortical components.

The method is carried out on real resting-state fMRI data from healthy volunteers. The identified subcortical components are validated by comparison with a functional atlas of the BG (Yelnik et al. (2007)).

## 2 Material and Methods

### 2.1 Data acquisition and preprocessing

Two runs of resting-state fMRI were acquired for 20 healthy volunteers who gave their informed consent. This protocol was approved by the local ethics committee. Subjects lied still in the scanner, eyes closed, refraining from any particular mental activity. Acquisition parameters were: FOV =  $224 \times 224$  mm; two functional runs of 160 volumes each with 41 contiguous slices; 3.5 mm isotropic voxels; (TR, TE) = (30, 2500) ms; flip angle:  $90^\circ$ . All



data were acquired using a 3 Teslas Siemens Trio TIM system at the Montreal Geriatric Institute Research Center, Montreal, QC, Canada. Individual data preprocessing using the SPM5 software<sup>1</sup> included slice-timing correction, spatial smoothing with an isotropic Gaussian kernel (full width at half maximum (FWHM)  $5 \times 5 \times 5$  mm) and coregistration of fMRI data on the corresponding subject's anatomical image. An affine and nonlinear transformation  $\mathbf{T}$  was calculated between each individual anatomical volume and the Montreal Neurological Institute (MNI) template (Evans et al. (1992)).

Besides, an immunohistochemical post-mortem atlas of the BG (Yelnik et al. (2007)) was used to provide a mask image of the BG in the MNI space, which was mapped back on each individual fMRI dataset using the inverse of the transformation  $\mathbf{T}$ , yielding one BG mask  $\text{BG}_{\text{ind}}$  per individual. This procedure finally yielded two fMRI datasets for each subject: one including the whole brain (wfMRI), and one comprising only cortical regions, the BG being masked out (mfMRI).

## 2.2 Identification of cortical networks

The first step of our method consisted of extracting cortical networks from the mfMRI data. More specifically, we used the NEDICA procedure (Perlberg et al. (2008)), which first computes sICA decomposition on each individual mfMRI data independently, and then uses hierarchical clustering on all individual ICA spatial components to define group maps. The resulting group maps characterize the functional networks.

Let  $\mathbf{X}$  be the  $T \times N_1$  matrix representing the mfMRI dataset for one subject, where  $T$  is the number of time samples and  $N_1$  the number of voxels per acquired volume. Spatial ICA solves the following decomposition problem

$$\mathbf{X} = \mathbf{A}\mathbf{F}, \quad (1)$$

where  $\mathbf{A}$  is the  $T \times T$  matrix of time courses, with  $T$  the number of time courses, and  $\mathbf{F}$  is the  $T \times N_1$  matrix of  $T$  spatial components. Following (Perlberg et al. (2008)),  $K = 40 \ll T$  components were used. Only  $P < K$  components corresponded to functional networks described in the literature and were considered for the following steps. The sICA model assumes statistical independence of the spatial components, which implies non-gaussianity for the resulting time course components.

To obtain information at the group level, individual spatial components were coregistered in the MNI space and clustered across subjects and runs based on their spatial similarity. The distance between two components was defined as their spatial correlation. A hierarchical clustering procedure was used that minimized the intra-class similarity, yielding a similarity tree. Thresholding this tree provided classes of similar components (one class per

<sup>1</sup> <http://www.fil.ion.ucl.ac.uk/spm/software/spm5/>

functional network), and a fixed-effect group map was finally computed for each class, representing a group-level cortical functional network. At the individual level, each spatial ICA component and corresponding time course can be associated to one group map (and, consequently, to one network); conversely, each group map (i.e. each network) is related to individual time courses that are characteristic of the network. In our case, since NEDICA was applied to mfMRI data, the group maps contained no voxels belonging to the BG. We have shown in a previous study that the cortical areas identified from mfMRI data were similar in terms of spatial distribution and power spectra to those identified from wfMRI data (Malherbe et al. (2010)).

### 2.3 Identification of subcortical components

The second step of our method consisted of complementing the cortical networks with their subcortical components. To this end, we assumed that the time courses of BG regions that were part of a given network were correlated to the time courses of the cortical regions belonging to the network. For each subject and each run, the individual time components of all networks selected from NEDICA were used as regressors in a general linear model (GLM)(Worsley et al. (2002)) applied to data in the  $BG_{ind}$  mask. This analysis, carried out with SPM5, yielded a parametric map per regressor, reflecting how the time course of each voxel in the mask was similar to that of the regressor. More specifically, let  $\mathbf{Y}$  be  $T \times N_2$  matrix of fMRI data in the  $BG_{ind}$  mask for each subject  $i$ , with  $N_2$  the number of voxels in the mask. The GLM reads:

$$\mathbf{Y}_i = \mathbf{R}_i \mathbf{B}_i + \boldsymbol{\varepsilon}_i, \quad (2)$$

where  $\boldsymbol{\varepsilon}_i$  is an i.i.d. Gaussian noise,  $\mathbf{R}_i$  is the  $T \times P$  time course matrix estimated from sICA, with  $P$  the number of networks of interest, and  $\mathbf{B}_i$  is the  $P \times N_2$  matrix of parameters to be estimated. The analysis yielded for each subject a map  $\hat{\mathbf{B}}_{ip}$  of BG regions for each network  $p$  or, in other words, a set  $\hat{\mathcal{B}}_p$  of 40 maps (20 subjects, 2 runs per subject) for each network.

To perform statistical inference at the group level, all parametric maps  $\hat{\mathcal{B}}_p$  were normalized to the MNI space using the transformation  $\mathbf{T}$  (see section 2.1). These normalized maps were then spatially smoothed with an isotropic Gaussian kernel (FWHM  $8 \times 8 \times 8$  mm). Then, we carried out a bootstrap analysis (Efron and Tibshirani, (1993)) per network as follows. For each network  $p$ , we first computed a conventional random effects parametric test, i.e. a Student  $t_0$  value for the  $\hat{\mathcal{B}}_p$  maps obtained for all subjects. A set  $\hat{\mathcal{B}}_p^*$  of  $S = 100$  surrogate data were then obtained by drawing randomly with replacement  $S$  times 40 maps from the initial  $\hat{\mathcal{B}}_p$  set. A Student  $t^*$  value was computed for each sample  $\hat{\mathcal{B}}_p^*$  and inference was made by calculating the achieved significance level (ASL) as follows:

$$ASL_{bootstrap} = \frac{\text{card}\{t^* \geq t_0\}}{S}. \quad (3)$$

Areas where  $ASL_{\text{bootstrap}} < 0.01$  were selected, yielding a group map of BG regions reproducibly found across subjects and associated with each cortical functional network. Finally, a cortical-subcortical functional network was obtained as the union of the group cortical map obtained from NEDICA analysis and the associated group subcortical areas obtained from bootstrap inference.

## 2.4 Validation using a functional atlas

The subcortical regions obtained in the previous step were finally validated by using an atlas of the BG (Yelnik et al. (2007)), which provides a segmentation of the BG in three functional domains: sensorimotor, limbic and associative. More precisely, we mapped the BG clusters selected as being reproducible at the group level onto the atlas, which allowed us to check whether the clusters and the atlas overlapped, and to validate that the functional segregation provided by our method was consistent with the atlas segmentation. For instance, we verified that BG clusters identified as being related to a motor cortical network correctly overlapped sensorimotor regions of the atlas.

## 3 Results

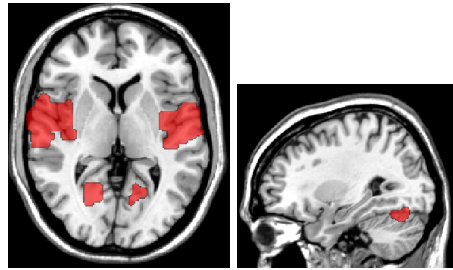
Using NEDICA,  $P = 14$  functional networks were identified from the mfMRI data and could be matched to 14 of the 16 networks identified from the wfMRI data. They were labelled following Beckmann et al. (2005) and Damoiseaux et al. (2006): five different attentional networks, one limbic network, one network related to executive control, one salient network, two default mode networks, two motor networks (see Figure 1 for an example), and two visual networks. The two networks extracted from wfMRI datasets that could not be matched with any of the mfMRI networks were: one pertaining to executive control and one including only subcortical structures such as the caudate nuclei, the putamen and part of the thalamus.

Figure 2 shows for the first motor network the group map obtained from the bootstrap analysis and mapped onto the functional atlas. The atlas segmentation for the motor part of the putamen and the pulvinar is outlined.

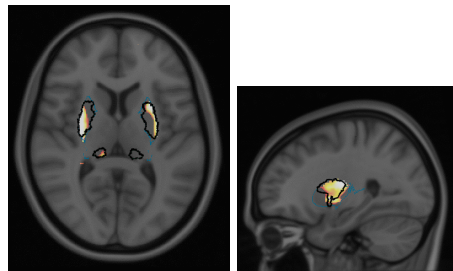
The first motor network obtained by the union of the first motor cortical network obtained by NEDICA and its subcortical components extracted by the individual GLM and the bootstrap analysis is shown in Figure 3.

## 4 DISCUSSION

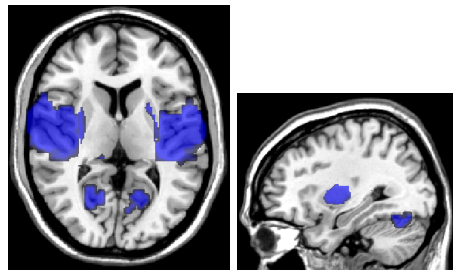
In this paper, we proposed a novel method to identify cortical-subcortical networks from fMRI data. First, cortical networks were detected with sICA and hierarchical clustering on data where the BG were masked out. Then,



**Fig. 1.** Axial (left) and sagittal (right) views of the cortical components of the first motor network obtained with NEDICA from mfMRI data.



**Fig. 2.** Axial (left) and sagittal (right) views of the subcortical components for the first motor network, mapped onto the functional atlas. The atlas segmentation for the motor part of the putamen and the pulvinar is outlined in black.



**Fig. 3.** Axial (left) and sagittal (right) views of the first motor network obtained by combining NEDICA (cortical parts) and our method (subcortical parts).

associated subcortical components were detected from fMRI data measured in the BG only, using an individual GLM and bootstrap inference at the group level. Bootstrap inference ensured statistical robustness at the group level, and with a high reproducibility in our population. In that respect, the proposed method appears to better characterize functional cortico-subcortical

loops than NEDICA alone. This segregation was qualitatively validated by using a functional atlas (Yelnik et al. (2007)), showing satisfactory match between the detected BG components and functional regions of the atlas.

NEDICA appeared to be particularly sensitive to the difference in BOLD signal that existed between BG and cortical regions on the whole fMRI data set, since it was not possible to segregate BG in subcomponents corresponding to the known cortex components. One possible reason why NEDICA failed could be that the fMRI signal within the BG may be composed of two components: one that is rather homogeneous within the BG (due to either similar localization, anatomy, or metabolic/hemodynamic features) and another one that is specific to the functional network the BG belong to.

An intuitive alternative could consist of applying NEDICA in the BG mask only to identify the subcortical components of functional networks. Unfortunately, this is not feasible, because sICA needs areas containing physiological noise to find reproducible spatial components. By contrast, using the individual GLM, we were able to separate the BG into areas that could be specifically associated with different large-scale networks.

In conclusion, the proposed method gives for the first time access to cortico-subcortical functional networks. A quantitative validation of the overlap between our results and the functional regions of the atlas is under investigation. Then, it would be interesting to quantify the functional interactions in terms of integration (Marrelec et al. (2008)) between the BG and the cortex in a given network or between networks. Another challenge would be to compare results obtained from healthy subjects with those obtained from patients with pathologies known to be associated with cortico-subcortical dysfunctions, such as the Tourette syndrome.

## Acknowledgements

The authors are grateful to the Montreal Geriatric Institute Research Center (Montreal, Canada), for providing us with data for this work. CM is funded by the French National Agency for Research (ANR 07 NEURO 023-01).

## References

- BECKMANN, C. F., DELUCA, M., DELVIN, J. T. and SMITH, S. M. (2005): Investigation into resting-state connectivity using independent component analysis, *Phil. Trans. R. Soc. B* 360:1001–1013.
- BELLEC, P., PERLBARG, V., JBABDI, S., PÉLÉGRINI-ISSAC, M., ANTON, J. L. and BENALI, H. (2006): Identification of large-scale networks in the brain using fMRI. *Neuroimage*, 29(4): 1231–1243.
- DAMOISEAUX, J. S., ROMBOUTS, S. A. R. B., BARKHOF, F., SCHELTENS, P., STAM, C. J., SMITH, S. M. and BECKMANN, C. F. (2006): Consistent resting-state networks across healthy subjects, *Proc. Natl. Acad. Sci. USA*, 103(37):13848–13853.

- DAMOISEAUX, J. S., BECKMANN, C. F., SANZ ARIGITA, E. J., BARKHOF, F., SCHELTENS, P., STAM, C. J., SMITH, S. M. and ROMBOUTS, S. A. R. B. (2008): Reduced resting-state brain activity in the "default network" in normal aging, *Cereb. Cortex*, 18:1856–1864.
- EFRON, B. and TIBSHIRANI, R. J. (1993): *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- EVANS, A. C., COLLINS, D. L. and MILNER, B. (1992): An MRI-based stereotactic atlas from 250 young normal subjects, *Journal Soc. Neurosci. Abstr.* 18:408.
- MALHERBE, C., PÉLÉGRINI-ISSAC, M., PERLBARG, V., LEHÉRICY, S., MARRELEC, G. and BENALI, H. (2010): Identification of functional cortico-subcortical networks in resting-state fMRI: a combined NEDICA and GLM analysis, *International Symposium on Biomedical Imaging (ISBI'10)*: in press.
- MARRELEC, G., BELLEC, P., DUFFAU, H., PÉLÉGRINI-ISSAC, M., KRAINIK, A., LEHÉRICY, S., DOYON, J. and BENALI, H. (2008): Regions, systems, and the brain: hierarchical measures of functional integration in fMRI, *Med. Image Anal.*, 12:484–496.
- MÉRIAUX, S., ROCHE, A., DEHAENE-LAMBERTZ, G., THIRION, B. and POLINE, J. B. (2006): Combined permutation test and mixed-effect model for group average analysis in fMRI, *Hum. Brain Mapp.*, 27(5):402–410.
- PERLBARG, V., MARRELEC, G., DOYON, J., PÉLÉGRINI-ISSAC, M., LEHÉRICY, S. and BENALI, H. (2008): NEDICA: Detection of group functional networks in fMRI using spatial independent component analysis, *International Symposium on Biomedical Imaging (ISBI'08)*:1247–1250.
- PERLBARG, V. and MARRELEC, G. (2008): Contribution of exploratory methods to the investigation of extended large-scale brain networks in functional MRI – methodologies, results and challenges, *Int. J. BioMed. Imaging, Article ID* 218519.
- PURVES, D., AUGUSTINE, G. J., FITZPATRICK, D., HALL, W. C., LAMANTIA, A. S. and McNAMARA, J. O. (2004): *Neuroscience*, Sinauer Associates Inc.
- SMITH, Y., RAJU, D. V., PARE, J-F. and SIDIBE, M. (2004): The thalamostriatal system: a highly specific network of the basal ganglia circuitry, *Trends Neurosci.*, 27:520–527.
- TOMASSINI, V., JBABDI, S., KLEIN, J. C., BEHRENS, T. E. J., POZZILLI, C., MATTHEWS, P. M., RUSHWORTH, M. F. S. and JOHANSEN-BERG, H. (2007): Diffusion-weighted imaging tractography-based parcellation of the human lateral premotor cortex identifies dorsal and ventral subregions with anatomical and functional specializations, *J. Neurosci.*, 27:10259–10269.
- VARELA, F. J., LACHAUX, J. P., RODRIGUEZ, E. and MARTINERIE, J. (2001): The brainweb: phase synchronization and large-scale integration, *Nature Reviews Neurosciences*, 2:229–239.
- WORSLEY, K. J., LIAO, C. H., ASTON, J., PETRE, V., DUNCAN, G. H. and MORALES, F. (2002): A general statistical analysis for fMRI data, *Neuroimage* 15:1–15.
- YELNIK, J., BARDINET, E., DORMONT, D., MALANDAIN, G., OURSELIN, S., TANDÉ, D., KARACHI, C., AYACHE, N., CORNU, P. and AGID, Y. (2007): A three-dimensional, histological and deformable atlas of the human basal ganglia. I. Atlas construction on immunohistochemical and MRI data, *Neuroimage*, 34:618–638.

# The Problem of Determining the Calibration Equations to Construct Model-calibration Estimators of the Distribution Function

Sergio Martínez<sup>1</sup>, Maria Rueda<sup>2</sup>, Antonio Arcos<sup>2</sup>, Helena Martínez<sup>1</sup> and Juan Francisco Muñoz<sup>3</sup>

- <sup>1</sup> Department of Statistics and Applied Mathematics  
04120 University of Almería, Spain *spuertas@ual.es*, *hmartinez@ual.es*  
<sup>2</sup> Department of Statistics and Operational Research  
18171 University of Granada, Spain *mrueda@ugr.es*, *arcos@ugr.es*  
<sup>3</sup> Department of Quantitative Methods in Economics  
18171 University of Granada, Spain *jfmunoz@ugr.es*.

**Abstract.** The calibration approach to estimating the finite population distribution function was proposed by Rueda et al. (2007). The proposed estimator is built by means of constraints that require the use of a set of fixed values  $t_1, \dots, t_p$ . Martínez et al. (2010), under the context of a linear regression working model, consider the case of only one point for the calibration and determine the optimum value  $t_1$  in the sense of minimum variance. In the present paper, assuming the use of more complex models, we study the problem of determining the optimal values  $t_i$  that gives the best estimation under simple random sampling without replacement for the case  $p = 2$

**Keywords:** distribution function, finite population, model-calibration approach

## 1 Introduction

Calibration is an important methodological instrument in the production of statistics. Calibration was introduced by Deville and Särndal (1992) to estimate the population total, but this approach adapts itself to the estimation of more complex parameters than a population total. Harms and Duchesne (2006), and Rueda et al. (2007) use different ways to implement the calibration approach in the estimation of the distribution function. Both methods give nearly design unbiased estimation and compare favorably with the earlier known estimation methods for the distribution function, not based on calibration thinking but on the same auxiliary information (see Särndal (2007)).

In the computationally simpler method of Rueda et al. (2007) one use the calibration with respect to the predicted  $y$ -values. The weights are obtained by minimizing the chi-square distance subject to calibration equations that require the use of  $p$  arbitrarily fixed values  $t_1, \dots, t_p$ . The precision of the

resulting calibration estimator change with the selected points  $t_i$ ,  $i = 1, \dots, p$ . Thus, the selection of the points  $t_i$  is a serious problem not analyzed in the above mentioned work. Martínez et al. (2010) under the context of a linear regression working model, consider the case of only one point for the calibration and determine the optimum value  $t_1$  in the sense of minimum variance.

In this paper we assume that the relationship between the study variable  $y$  and the auxiliary variable  $\mathbf{x}$  can be described by a superpopulation model  $y = m(x) + e$  where the regression function  $m$  can be linear or nonlinear. In this situation we consider a model-calibration estimator using two points and we study the problem of the optimal values  $t_1$  and  $t_2$  that gives the best estimation under simple random sampling without replacement. Finally, a simulation study compares the method proposed with other conventional methods.

## 2 Calibration estimator of the distribution function

Consider a finite population consisting of  $N$  identifiable units. Associated with the  $i$ th unit are, the study variable  $y_i$  and a vector of auxiliary variables  $\mathbf{x}_i$ . The values  $\mathbf{x}_i$  are known for the entire population but  $y_i$  is known only of the  $i$ th unit is selected on the sample,  $s$ . We assume that the inclusion probabilities  $\pi_k$  are strictly positive.

The finite population distribution function of the study variable  $y$ , is given by  $F_y(t) = \sum_{k \in U} \Delta(t - y_k)/N$  with  $\Delta(t - y_k) = 1$  if  $t \geq y_k$  and  $\Delta(t - y_k) = 0$  otherwise. A purely design based estimator of the distribution function is the Horvitz-Thompson estimator, defined by  $\hat{F}_{YH}(t) = \sum_{k \in s} d_k \Delta(t - y_k)/N$  with  $d_k = 1/\pi_k$ , the basic design weights.

Rueda et al. (2007) assume the relationship between  $y$  and  $\mathbf{x}$  can be described by a linear superpopulation model  $\xi$ :  $y_i = \beta' \mathbf{x}_i + \varepsilon_i$ ,  $i = 1, \dots, N$  where  $\varepsilon_i$ 's are independently and identically distributed random variables with  $E_\xi(\varepsilon_i) = 0$  and variance  $\sigma^2$ . The authors consider a calibration estimator by first defining a pseudo-variable  $g_k = \hat{\beta}' \mathbf{x}_k$  for  $k = 1, 2, \dots, N$ , where  $\hat{\beta}$  is a weighted estimator of the multiple regression coefficient  $\beta$  between  $y$  and  $\mathbf{x}$ . They then define the calibration estimator  $\hat{F}_{yc}(t) = \frac{1}{N} \sum_{k \in s} \omega_k \Delta(t - y_k)$  where the new weights  $\omega_k$  are modified from  $d_k = 1/\pi_k$  by minimizing the chi-square distance measure subject to the calibration equations  $\frac{1}{N} \sum_{k \in s} \omega_k \Delta(t_j - g_k) = F_g(t_j)$ ,  $j = 1, 2, \dots, P$ . The term  $F_g(t_j)$  denotes the finite distribution function of the pseudo-variable  $g$  evaluated at the point  $t_j$ , where  $t_j$  for  $j = 1, 2, \dots, P$  are points that we choose arbitrarily and assume that  $t_1 < t_2 < \dots < t_P$ .

Martínez et al. (2010) consider this estimator under simple random sampling and using a single point  $t_1$  in the calibration equations and they determine the best selection for this point in the sense of minimum variance.



Now, to incorporate auxiliary information  $x_i$  available for all  $i \in U$  we assume a superpopulation for  $y$  built on some mean function of  $\mathbf{x}$ :  $y_i = \mu(\mathbf{x}_i, \theta) + \varepsilon_i$ ,  $i = 1, \dots, N$ . The random vector  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_N)'$  is assumed to have zero mean and a positive definite covariance matrix which is diagonal ( $y_i$  are mutually independent). This model structure is quite general and includes the linear and the nonlinear regression model.

Of particular interest is the formulation of the predicted values  $y_i^0 = \hat{\mu}(\mathbf{x}_i, \theta)$ . If  $\mu(\mathbf{x}_i, \theta)$  is a known parametric function the superpopulation parameter  $\theta$  can be estimated using standard procedures and we can obtain the predicted values as  $y_i^0 = \hat{\mu}(\mathbf{x}_i, \theta) = \mu(\mathbf{x}_i, \hat{\theta})$ , being  $\hat{\theta}$  a design-based estimate from the sampled data (Wu and Sitter (2001))

Now, we define the new pseudo-variable  $h_k = y_k^0$  for each  $k = 1, \dots, N$ , the predicted values by the model, and then we can construct a model-calibrated estimator of the distribution function as  $\hat{F}_{ymc}(t) = \frac{1}{N} \sum_{k \in s} \omega_k \Delta(t - y_k)$  where the new weights  $\omega_k$  minimizing the chi-square distance measure  $\Phi_s = \sum_{k \in s} \frac{(\omega_k - d_k)^2}{d_k}$  subject to the calibration equations  $\frac{1}{N} \sum_{k \in s} \omega_k \Delta(t_j - h_k) = F_h(t_j)$ ,  $j = 1, 2, \dots, P$ . The term  $F_h(t_j)$  denotes the finite distribution function of the pseudo-variable  $h$  evaluated at the point  $t_j$ .

### 3 Determining optimal points under simple random sampling

In this section we assume that the sample  $s$  is selected by simple random sampling of size  $n$ . We consider the case of use only two points  $t_1$  and  $t_2$  for the calibration equations. In this situation the asymptotic variance of the model calibration estimator  $\hat{F}_{ymc}(t)$  can be written as:

$$V(\hat{F}_{ymc}(t)) = V(\hat{F}_{YH}(t)) + Z(t_1, t_2)$$

where

$$D_1 = \frac{F_h(t_2) \sum_{k \in U} \Delta(t_1 - h_k) \Delta(t - y_k) - F_h(t_1) \sum_{k \in U} \Delta(t_2 - h_k) \Delta(t - y_k)}{N F_h(t_1) (F_h(t_2) - F_h(t_1))} \quad (1)$$

$$D_2 = \frac{\sum_{k \in U} \Delta(t_2 - h_k) \Delta(t - y_k) - \sum_{k \in U} \Delta(t_1 - h_k) \Delta(t - y_k)}{N (F_h(t_2) - F_h(t_1))} \quad (2)$$

and

$$\begin{aligned} Z(t_1, t_2) = & D_1^2 V(\hat{F}_{hH}(t_1)) + D_2^2 V(\hat{F}_{hH}(t_2)) - 2D_1 \text{Cov}(\hat{F}_{YH}(t), \hat{F}_{hH}(t_1)) - \\ & - 2D_2 \text{Cov}(\hat{F}_{YH}(t), \hat{F}_{hH}(t_2)) + 2D_1 D_2 \text{Cov}(\hat{F}_{hH}(t_1), \hat{F}_{hH}(t_2)) \end{aligned} \quad (3)$$

We consider the expression (3), where under simple random sampling:

$$V(\widehat{F}_{hH}(t_i)) = \frac{N}{N-1} F_h(t_i)(1 - F_h(t_i)), i = 1, 2 \quad (4)$$

$$Cov(\widehat{F}_{YH}(t), \widehat{F}_{hH}(t_i)) = \frac{1}{N-1} \left[ \sum_{k \in U} \Delta(t_i - h_k) \Delta(t - y_k) - N F_y(t) F_h(t_i) \right] \quad (5)$$

$$\begin{aligned} Cov(\widehat{F}_{hH}(t_1), \widehat{F}_{hH}(t_2)) &= \frac{1}{N-1} \left[ \sum_{k \in U} \Delta(t_1 - g_k) \Delta(t_2 - y_k) - N F_h(t_1) F_h(t_2) \right] = \\ &= \frac{N}{N-1} F_h(t_1) [1 - F_h(t_2)] \end{aligned} \quad (6)$$

We consider again the population values of  $y$  in ascending order  $y_{[1]} \leq y_{[2]} \leq \dots \leq y_{[N]}$ . Next, we consider the population values of the variable  $h$ , arranged by the variable  $y$ , that is  $(y_{[1]}, h_{[1]}); (y_{[2]}, h_{[2]}); \dots; (y_{[N]}, h_{[N]})$

We define the set  $A_t$  and  $B_t$  given by  $A_t = \{h_{[k]} : k \in C_t\}$  with  $C_t$  is done by  $C_t = \{k \in \{1, 2, \dots, N\} : y_{[k]} \leq t\}$  and  $B_t = \{b_k : k = 1, 2, \dots, P\}$  with  $b_k = \max_{l \in U_k} \{h_l\}$  with  $U_k = \{l \in U : a_{k-1} \leq h_l < a_k\}$ ;  $k = 1, 2, \dots, P$  where we suppose that  $a_1 < a_2 < \dots < a_P$  are the  $P$  different elements of  $A_t$  in ascending order, with  $a_0 = -\infty$  and  $a_{P+1} = +\infty$ .

Then, if we want to estimate the distribution function  $F_y$  at the value  $t$ , we will proof that the function  $Z(t_1, t_2)$  has the global minimum at the point  $(t_1, t_2)$ , where  $t_1$  and  $t_2$  are points of the set  $A_t$  or  $B_t$ .

The function  $Z(t_1, t_2)$  can take different values by the election of  $t_1$  and  $t_2$ :

- a. If  $t_1 < a_1$  and  $t_2 < a_1$ ,  $Z(t_1, t_2) = 0$
- b. If  $t_1 < a_1$  and  $a_k \leq t_2 < a_{k+1}$  with  $k = 1, 2, \dots, P-1$

$$D_2 = \frac{\bar{k}}{N(F_h(t_2) - F_h(t_1))} \quad \text{with} \quad \bar{k} = \sum_{j \in U} \Delta(a_k - h_j) \Delta(t - y_j)$$

and  $D_1 = -D_2$ . Then

$$Z(t_1, t_2) = \frac{\bar{k}}{N(N-1)} \left[ 2N F_y(t) - \bar{k} - \frac{\bar{k}}{F_h(t_2) - F_h(t_1)} \right]$$

In this case, the function  $Z(t_1, t_2)$  has the minimum point at the minimum denominator  $F_h(t_2) - F_h(t_1)$ . In the set  $(-\infty, a_1) \times [a_k, a_{k+1})$  it's clear that the expression  $F_h(t_2) - F_h(t_1)$  has the minimum value at  $t_1 = b_1$  and  $t_2 = a_k$ .

- c. If  $t_1 < a_1$  and  $a_P \leq t_2$ , we have  $D_2 = \frac{F_y(t)}{(F_h(t_2) - F_h(t_1))}$  and  $D_1 = -D_2$ , and then

$$Z(t_1, t_2) = \frac{N(F_y(t))^2}{N-1} \left[ 1 - \frac{1}{(F_h(t_2) - F_h(t_1))} \right]$$

Thus, in the set  $(-\infty, a_1) \times [a_P, +\infty)$ , the function  $Z(t_1, t_2)$  has the minimum at the point  $(t_1, t_2) = (b_1, a_P)$ , because the denominator  $(F_h(t_2) - F_h(t_1))$  is minimum at this point.

- d. If  $a_k \leq t_1 < a_{k+1}$  and  $a_k \leq t_2 < a_{k+1}$  with  $t_1 < t_2$  and  $k = 1, 2, \dots, P-1$ ,  $D_1 = \frac{\bar{k}}{NF_h(t_1)}$  and  $D_2 = 0$  and then

$$Z(t_1, t_2) = \frac{\bar{k}}{N(N-1)} \left[ \frac{-\bar{k}}{F_h(t_1)} + 2NF_y(t) - \bar{k} \right]$$

Therefore the  $Z(t_1, t_2)$  in the set  $[a_k, a_{k+1}) \times [a_k, a_{k+1})$  has the local minimum at  $(t_1, t_2) = (a_k, t_2)$  where the point  $t_2$  is an arbitrary chosen point in the interval  $(a_k, a_{k+1})$ .

- e. If  $a_k \leq t_1 < a_{k+1}$  and  $a_l \leq t_2 < a_{l+1}$  where  $k = 1, 2, \dots, P-1$  and  $l = 1, 2, \dots, P$  with  $k < l$

$$D_1 = \frac{F_h(t_2)\bar{k} - F_h(t_1)\bar{l}}{NF_h(t_1)(F_h(t_2) - F_h(t_1))} \quad \text{and} \quad D_2 = \frac{\bar{l} - \bar{k}}{N(F_h(t_2) - F_h(t_1))}$$

and the function  $Z(t_1, t_2)$  take the form

$$Z(t_1, t_2) = \frac{-(\bar{k})^2}{N(N-1)F_h(t_1)} - \frac{(\bar{l} - \bar{k})^2}{N(N-1)(F_h(t_2) - F_h(t_1))} + \frac{\bar{l}(2NF_y(t) - \bar{l})}{N(N-1)}$$

Because the distribution function of the variable  $g$  is monotone nondecreasing, for all  $(t_1, t_2) \in [a_k, a_{k+1}) \times [a_l, a_{l+1})$  we have

$$\begin{aligned} Z(t_1, t_2) &= \frac{-(\bar{k})^2}{N(N-1)F_h(t_1)} - \frac{(\bar{l} - \bar{k})^2}{N(N-1)(F_h(t_2) - F_h(t_1))} \geq \\ &\geq \frac{-(\bar{k})^2}{N(N-1)F_h(t_1)} - \frac{(\bar{l} - \bar{k})^2}{N(N-1)(F_h(a_l) - F_h(t_1))} = G(t_1, a_l) \end{aligned}$$

and the election of  $t_2$  is  $t_2 = a_l$ .

Now, if we denote  $M = F_h(a_l)$ , we have to minimize the function  $h(t_1) = G(t_1, a_l)$  where

$$h(t_1) = G(t_1, a_l) = \frac{-(\bar{k})^2}{N(N-1)F_h(t_1)} - \frac{(\bar{l} - \bar{k})^2}{N(N-1)(M - F_h(t_1))}$$

For it, we consider the function

$$f(x) = \frac{-(\bar{k})^2}{N(N-1)x} - \frac{(\bar{l} - \bar{k})^2}{N(N-1)(M-x)}$$

The equation  $f'(x) = 0$  have two solutions when  $(2\bar{k} - \bar{l}) \neq 0$ :

$$x_1 = \frac{\bar{k}M}{2\bar{k} - \bar{l}} \quad ; \quad x_2 = \frac{\bar{k}M}{\bar{l}}$$

and  $f'(x) > 0$  for  $x \in (0, x_2)$  and  $f'(x) < 0$  for  $x \in (x_2, M)$ . Thus, we have

$f(x)$  is monotone nondecreasing for  $x \in (0, x_2)$

$f(x)$  is monotone noncreasing for  $x \in (x_2, M)$

For all  $t_1$  with  $a_k \leq t_1 < a_{k+1}$  we have that  $F_h(t_1) < M = F(a_l)$  because  $k < l$ . Then,  $0 < F_h(t_1) < M$  and the function  $F_h(t_1)$  is monotone nondecreasing and consequently the function  $h(t_1) = f(F_h(t_1))$  in the interval  $[a_k, a_{k+1})$  has the local minimum at the point  $t_1 = a_k$  or  $t_1 = b_{k+1}$ .

Anyway, in the set  $[a_k, a_{k+1}) \times [a_l, a_{l+1})$  the function  $Z(t_1, t_2)$  has the minimum at the point  $(t_1, t_2) = (a_k, a_l)$  or  $(t_1, t_2) = (b_{k+1}, a_l)$ .

- f. If  $(t_1, t_2) \in [a_P, +\infty) \times [a_P, +\infty)$ ,  $D_1 = \frac{F_y(t)}{F_h(t_1)}$  and  $D_2 = 0$  and the function  $Z(t_1, t_2)$  is given by

$$Z(t_1, t_2) = \frac{-N[F_y(t)]^2}{(N-1)F_h(t_1)} + \frac{N[F_y(t)]^2}{(N-1)}$$

It's clear that the function  $Z(t_1, t_2)$  has the local minimum at  $(t_1, t_2) = (a_P, t_2)$  where  $t_2$  is an arbitrary chosen point in the interval  $(a_P, +\infty)$ .

Whit these optimum points  $t_{1opt}, t_{2opt}$  we define the optimum model-calibrated estimator  $\hat{F}_{ymop}(t) = \hat{F}_{YH}(t) + (F_h(t_{1opt}) - \hat{F}_{hH}(t_{1opt}))D_1 + (F_h(t_{2opt}) - \hat{F}_{hH}(t_{2opt}))D_2$ . This optimum estimator can not be calculated because the optimal values  $t_{1opt}$  and  $t_{2opt}$  depend on some unknown values.

#### 4 Optimal estimator with optimal estimated points

We can obtain an estimation  $\hat{Z}(t_1, t_2)$  of  $Z(t_1, t_2)$  under simple random sampling with the estimation of the values (4); (5); (6); (1) and (2) by their corresponding sample expressions. Thus, if we consider again the sets  $A_{st}$  and  $B_{st}$ , given by  $A_{st} = \{h_k : k \in C_{st}\}$  with  $C_{st} = \{k \in s : y_k \leq t\}$  and assume that  $A_{st}$  has  $p$  points, that is

$$A_{st} = \{a_i : i = 1, 2, \dots, p\} \quad \text{and} \quad B_{st} = \{b_i : i = 1, 2, \dots, p\}$$

with  $b_k = \max_{l \in s_k} \{h_l\}$  with  $s_k = \{l \in s : a_{k-1} \leq h_l < a_k\}$ ;  $k = 1, \dots, p$ , the global minimum of the function  $\hat{Z}(t_1, t_2)$  is at one point of  $A_{st}$  or  $B_{st}$  (where  $a_0 = -\infty$ ).

Now we define the calibration estimator obtained with the values that minimize the function  $\hat{Z}(t_1, t_2)$ . This estimator is denoted by  $\hat{F}_{ymprop}(t)$ . Using the results of Randless (1982) one can prove that the proposed estimator  $\hat{F}_{ymprop}(t)$  and the obtained calibration estimator using the optimal values,  $\hat{F}_{ymop}(t)$  have the same asymptotic behavior.

## 5 Simulation study

We have performed an empirical study that illustrates the theoretical results and analyze the precision of the proposed method for small and moderate simple sizes. We compare the precision of the proposed calibration estimator  $\hat{F}_{ymprop}(t)$  with the following estimators:  $\hat{F}_{CD}(t)$  (Chambers, R.L. and Dunstan, R. (1986)),  $\hat{F}_{RKM}(t)$  (Rao et al. (1990)), the difference estimator  $\hat{F}_d(t)$ , the ratio estimator  $\hat{F}_r(t)$ , the usual calibration estimator  $\hat{F}_{yc1}(t)$  with  $t_1 = Q_h(0.5)$ , the usual calibration estimator  $\hat{F}_{yc2}(t)$  with  $t_1 = Q_h(0.5)$  and  $t_2 = \max_{k \in U} \{h_k\}$ , the calibration estimator  $\hat{F}_{yc3}(t)$  and  $\hat{F}_{yc4}(t)$  from Martínez et al (2010). The population considered is the "Cloud" population from UCI Machine Learning Repository. This population is constituted of 1024 units, and 10 variables, the study variable is the fourth variable and the auxiliary variable is the seventh variable (a more detailed description is available from Asuncion and Newman (2007)). We selected 1000 samples for three different sample sizes under simple random sampling without replacement (SRSWOR). The considered sample sizes were 50, 75 and 100. For each sample and for each estimator, estimates of the distribution function  $F(t)$  were calculated for 11 different values of  $t$ , namely the quantiles  $Q_y(\alpha)$  for  $\alpha = 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8$  and  $0.9$ . The performance of all the estimators is measured by means of the average relative bias (AVRB) and the average relative efficiency (AVRE), given respectively by

$$\text{AVRB}(t) = \frac{1}{11} \sum_{q=1}^{11} |\text{RB}(t_q)|, \quad \text{AVRE}(t) = \frac{1}{11} \sum_{q=1}^{11} \text{RE}(t_q)$$

where RB and RE are defined as

$$\text{RB}(t) = \frac{1}{B} \sum_{b=1}^B \frac{\hat{F}(t)_b - F_y(t)}{F_y(t)} \quad \text{and} \quad \text{RE}(t) = \frac{MSE[\hat{F}(t)]}{MSE[\hat{F}_{YH}(t)]},$$

where  $b$  indexes the  $b$ th simulation run,  $\hat{F}(t)$  is an estimator for the distribution function,  $MSE[\hat{F}(t)] = B^{-1} \sum_{b=1}^B [\hat{F}(t)_b - F_y(t)]^2$  is the empirical Mean Square Error for  $\hat{F}(t)$  and  $MSE[\hat{F}_{YH}(t)]$  is similarly defined for the

Horvitz-Thompson estimator. Table 1 shows our results and it can be seen that in terms of AVRE,  $\hat{F}_{ymprop}$  is the most efficient in all cases. Thus, the estimator  $\hat{F}_{ymprop}$  is more efficient than  $\hat{F}_{yc3}$  and  $\hat{F}_{yc4}$  proposed by Martínez et al. (2010), although both estimators provide very good results in the population, where they behave better than other estimators considered. We can also verify that the estimator  $\hat{F}_{ymprop}$  greatly improve the performance of usual calibrated estimators  $\hat{F}_{yc1}$  and  $\hat{F}_{yc2}$ . In all sample sizes the estimator  $\hat{F}_{ymprop}$  have low AVRB for all sample sizes, while most of estimators have low AVRB for the sample size  $n = 100$  but not for  $n = 50$  or  $n = 75$ . Moreover in all sizes the estimator  $\hat{F}_{ymprop}$  produces the lowest AVRB.

	CLOUD					
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 50$		$n = 75$		$n = 100$	
$\hat{F}_{CD}$	0.1609	1.0519	0.1272	1.0730	0.1096	1.0688
$\hat{F}_d$	0.1573	0.9721	0.1208	0.9634	0.1049	0.9673
$\hat{F}_r$	0.1869	1.4104	0.1456	1.4235	0.1258	1.4225
$\hat{F}_{RKM}$	0.1612	0.9999	0.1254	1.0030	0.1081	1.0003
$\hat{F}_{yc1}(t)$	0.1808	1.2617	0.1402	1.2646	0.1205	1.2558
$\hat{F}_{yc2}(t)$	0.1555	0.9237	0.1205	0.9257	0.1042	0.9274
$\hat{F}_{yc3}(t)$	0.1577	0.9771	0.1206	0.9736	0.1050	0.9755
$\hat{F}_{yc4}(t)$	0.1379	0.8751	0.1083	0.8716	0.0909	0.8698
$\hat{F}_{ymprop}(t)$	0.1220	0.7518	0.0939	0.7467	0.0790	0.7446

**Table 1.** Average relative bias (AVRB) and the average relative efficiency (AVRE).

## References

- ASUNCION, A. and NEWMAN, D.J. (2007): UCI Machine Learning Repository [<http://www.ics.uci.edu/mlearn/MLRepository.html>]. Irvine, CA: University of California, School of Information and Computer Science.
- CHAMBERS, R.L. and DUNSTAN, R. (1986): Estimating distribution functions from survey data. *Biometrika* 73, 597-604.
- DEVILLE, J. C. and SÄRNDAL, C. E. (1992): Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87, 376-382.
- ESTEVAO, V. M. and SÄRNDAL, C. E. (2006): Survey estimates by calibration on complex auxiliary information. *International Statistical Review* 42, 127-147.
- HARMS, T. and DUCHESNE, P. (2006): On calibration estimation for quantiles. *Survey Methodology* 32, 37-52.
- MARTÍNEZ, S., RUEDA, M., ARCOS, A. and MARTÍNEZ, H. (2010): Optimum calibration points estimating distribution functions. *Journal of Computational and Applied Mathematics* 233, 2265-2277.

- RANGLES, R. H. (1982): On the asymptotic normality of statistics with estimated parameters. *Annals of Statistics* 10, 462-474
- RAO, J.N.K., KOVAR, J.G. and MANTEL, H.J. (1990): On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* 77, 365-375.
- RUEDA, M., MARTÍNEZ, S., MARTÍNEZ, H. and ARCOS, A. (2007): Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference* 137 (2), 435-448.
- SÄRNDAL, C. E. (2007): The calibration approach in survey theory and practice. *Survey Methodology* 33 (2), 99-119.
- SINGH, S. (2001): Generalized calibration approach for estimating variance in survey sampling. *Annals of the Institute of Statistical Mathematics* 53 (2), 404-417.





# Dealing with Nonresponse in Survey Sampling: an Item Response Modeling Approach

Alina Matei<sup>1,2</sup>

<sup>1</sup> Institute of Statistics, University of Neuchâtel,  
Pierre à Mazel 7, 2000 Neuchâtel, Switzerland, [alina.matei@unine.ch](mailto:alina.matei@unine.ch)

<sup>2</sup> Institute for Pedagogical Research and Documentation (IRDP) Neuchâtel,  
Fbg. de l'Hôpital 43, cp 556, 2002 Neuchâtel, Switzerland

**Abstract.** Dealing with nonresponse is a very important topic, since nonresponse is present almost in all surveys, and can cause biased estimation. Nonresponse is defined as the failure to provide the required information by a unit selected in a sample. We distinguish between unit nonresponse and item nonresponse. Unit nonresponse implies that we have no information at all from the sampled unit. Item nonresponse means that the sampled unit does not fill some of the survey items. Each unit selected in the sample has associated a sampling weight and a response probability to answer the questionnaire. The response probability is unknown and should be estimated. The main method to deal with the unit nonresponse is to use rereighting. This method adjusts the initial sampling weights by the inverse of the estimated response probabilities, providing new weights. We focus on unit nonresponse adjustment in survey data and estimate the response probabilities using an item response model called the Rasch model. This model uses a latent parameter. We believe that this latent parameter can explain a part of the unknown behavior of a unit to respond in the survey. No information about the nonrespondents and no auxiliary information are required in the proposed method. Theoretical aspects and simulations are used to support our theory.

**Keywords:** survey sampling, nonresponse, item response modeling

## 1 Introduction

In survey sampling, a random sample  $s$  is drawn from a finite population in order to do inference. Usually, nonresponse occurs. We distinguish two types of nonresponse: *unit nonresponse*, when a sampled element refuses to answer, and *item nonresponse*, when a sampled element does not respond to all of the survey items. We focus on unit nonresponse. Dealing with nonresponse is a very important topic, since nonresponse is present in almost all surveys, and can highly bias estimation.

Randomness in survey sampling is generated by a sampling design. The sampling is made from a finite population  $U = \{1, 2, \dots, k, \dots, N\}$ . Let  $s \subset U$  denotes a random sample. Each unit  $k \in U$  has a given inclusion probability  $\pi_k$  to be included in the sample. Sampling design associates to each unit

$k \in s$  a sampling weight equal to  $1/\pi_k$ . In the presence of unit nonresponse, another source of randomness is added, which is generated by an unknown response mechanism. Due to unit nonresponse only a set  $r \subseteq s$  of units answer the questionnaire. Each unit  $k \in s$  has also associated a response probability  $p_k$  to respond to the questionnaire. The response probability  $p_k$  is unknown and is estimated using different methods like response propensity modeling or response homogeneity groups. A commonly used method to deal with unit nonresponse is to use reweighting. The sampling weights  $1/\pi_k$  are adjusted by the inverse of the estimated response probability  $\hat{p}_k$ , providing new weights equal to  $1/(\pi_k \hat{p}_k)$ . We propose to estimate  $p_k$  using an item response model fitted to the set of respondents  $r$ . In this method no information about the nonrespondents and no auxiliary information are required. The estimated response probabilities are used in a total estimator to reduce the nonresponse bias. Simulations suggest that the proposed method reduces the relative value of the nonresponse bias.

## 2 Estimating the response probabilities

The survey methodology literature defined the following missing data mechanisms. Data are *missing completely at random* if the probability of being a nonrespondent is independent of the observed or unobserved data. Data are *missing at random* if the missing data depends on the observed data, but not on the unobserved data. Finally, *non-ignorable nonresponse* allows dependence on both observed and unobserved data. Thus, unlike the assumptions of data missing completely at random or missing at random, non-ignorable nonresponse mechanism must be modeled. We consider a reweighting scheme to adjust for non-ignorable unit nonresponse based on item response models.

Item response models (IRM) (see Baker and Kim, 2004; De Boeck and Wilson, 2004) are predominantly used in measurement applications in psychology, education, and other social science areas. The basic principle in IRM is that item responses (typically questions to be answered, problems to be solved etc.) given by persons can be modeled as a function of some predictors. Such predictors may be: a) characteristics of items, of individuals, or a combinations of individuals and items; b) observed or latent (of either items or individuals) variables; c) latent continuous or latent categorical variables. The simplest IRM is the Rasch model (see Rasch, 1960, 1961). We consider a Rasch model to estimate the response probabilities  $p_k$ . The goal is to adjust for non-ignorable nonresponse. Weighting observations by the reciprocal of the estimated response probabilities reduces nonresponse bias given that the model is correct.

Suppose that each of  $n$  units in  $r$ , labelled  $k = 1, \dots, n$ , are exposed to  $m$  items, labelled  $j = 1, \dots, m$ . Let  $x_{kj}$  be the response indicator of unit  $k$  to item  $j$ , with  $j = 1, \dots, m$ . The indicator variable  $x_{kj}$  takes value 1 when the unit  $k$  asks the item  $j$  and 0, otherwise. A latent parameter, denoted  $\theta_k$ , is

computed for each respondent  $k \in r$ . It is often called the ‘ability’ of unit  $k$  to answer the items. In our context,  $\theta_k$  can be explained as a ‘will-to-respond to the questionnaire’ latent variable. Note that  $\theta_k \sim N(0, \sigma_\theta^2)$ . Each item  $j$  has associated a parameter  $\beta_j$ , called the difficulty of item  $j$ . For example, if the item  $j$  is a sensitive item, its difficulty will be high, since many people do not answer this item. Let  $Pr(x_{ij} = 1|\theta_k) = p_{kj}$  be the probability that unit  $k$  will answer the item  $j$  given the ability  $\theta_k$ . Rasch item analysis model assumes that

$$\log\left(\frac{p_{kj}}{1 - p_{kj}}\right) = \theta_k - \beta_j, \quad (1)$$

$k = 1, \dots, n, j = 1, \dots, m$ . An important feature of this model is the so-called *conditional independence* assumption, which postulates that the item responses are independent given the latent variables. The Rasch model exists in three variants named after the formulation of the method used to estimate it (the maximum likelihood): the joint maximum likelihood formulation, the conditional maximum likelihood formulation, and the marginal maximum likelihood formulation. We will follow here the marginal maximum likelihood formulation, where the unit parameters are sampled from a distribution, so that only the parameters of that distribution (and not the individual parameters) enter the likelihood that is maximized. Thus only the difficulties  $\beta_j$  and  $\sigma_\theta^2$  are estimated. The ability  $\theta_k$  requires a further step beyond the model estimation, based commonly on an empirical Bayes estimation (see De Boeck and Wilson, 2004 and the references therein).

For the unit  $k \in s$  we define a binary random variable  $R_k$  with value 1 if unit  $k \in r$  and 0 if  $k \in s \setminus r$ . Then  $E(R_k|s) = p_k$ , where the expectation operator is applied with respect to the response mechanism generating the unit nonresponse. We assume that  $p_k$  depends on  $k$  but not on the sample  $s$  of which  $k$  is a member. Our goal is to estimate the response probability  $p_k$  using a Rasch model. Let  $S_k = \sum_{j=1}^m x_{kj}$ . If the unit  $k \in s \setminus r$ , then  $k$  does not respond to any item and  $S_k = 0$ ; if  $k \in r$ , then  $S_k > 0$ . Then  $p_k$  is estimated by  $Pr(S_k > 0)$ . By the conditional independence assumption we have

$$\begin{aligned} Pr(S_k > 0) &= 1 - Pr(S_k = 0) \\ &= 1 - Pr(\cap_{j=1}^m (x_{kj} = 0)) \\ &= 1 - \prod_{j=1}^m (1 - p_{kj}), \end{aligned}$$

where  $x_{kj} \sim \text{Bernoulli}(p_{kj})$ . The response probability  $p_k = Pr(R_k = 1)$  is estimated by

$$\hat{p}_k = 1 - \prod_{j=1}^m (1 - \hat{p}_{kj}),$$

where  $\hat{p}_{kj}$  are computed using the model (1).

### 3 Simulated examples

A limited simulation study was carried out to study the behavior of our method in relatively small samples. The Rasch model was fitted using the R ‘ltm’ package (Rizopoulos, 2006), which uses the marginal maximum likelihood estimation method. We considered the selected sample  $s$  a simple random sample without replacement of size  $n_s$ . In this case  $\pi_k = n_s/N$ , where  $N$  is the population size.

*Example 1:* We drew 10 000 samples each of size  $n_s = 50$  by simple random simple without replacement from a population of size  $N = 300$ . For each sample  $s$ , a response set  $r$  was created by carrying out for each  $k \in s$  a Bernoulli experiment with parameter  $p_k$ . Thus, the unit  $k$  becomes a respondent with probability  $p_k$  and a non-respondent with the probability  $1 - p_k$ , where  $p_k$  is known for all  $k$ . The probabilities  $p_k$  were computed from  $p_{kj}$ , for all  $k$  and all  $j$ , using the formula  $1 - \prod_{j=1}^m (1 - p_{kj})$ ;  $p_{kj}$  were independent uniform randomly generated between 0 and 0.15. The mean of  $p_k$  was 0.69. The study variable  $y_k$  was randomly generated using the absolute value of a  $N(0, 1)$  random variable. The average response rate was about 0.7 in the simulation.

The goal was to estimate the total  $Y = \sum_{k \in U} y_k$ , where  $y_k$  is the value of the study variable for the  $k$ th unit. The following estimator of the total  $Y$  was used:

$$N \frac{\sum_{k \in r} w_k y_k}{\sum_{k \in r} w_k}. \quad (2)$$

Based on this general estimator and for every response set in simulations we computed three estimators:

- the proposed estimator, denoted by  $\hat{Y}$ ; in this case

$$w_k = \frac{1}{\pi_k \hat{p}_k};$$

- the ‘ideal’ estimator, denoted by  $\hat{Y}^{true}$  and calculated with the true response probabilities used to generate the data; in this case

$$w_k = \frac{1}{\pi_k p_k};$$

- the estimator based on the respondents, with equal estimated response probabilities  $n/n_s$ , denoted by  $\hat{Y}^r$ ; in this case

$$w_k = \frac{1}{\pi_k n/n_s} = \frac{N}{n}.$$

A Rasch model was used for each response set to compute the quantities  $\hat{p}_{kj}$  used in the proposed estimator. The model was assumed to be correctly specified.

A usual choice in practice is to use the estimator  $\sum_{k \in r} y_k / (\pi_k \hat{p}_k)$ . Here, the Hájek estimator applied to the nonresponse weighting adjustment estimation (2) was used because the previous estimator can be very unstable when  $\hat{p}_k$  is close to 0 (see Kim and Kim, 2007).

Table 1 presents the Monte Carlo relative bias and variance of the estimators obtained from the simulation study. The Monte Carlo relative bias is computed as the value of the Monte Carlo bias divided by the Monte Carlo standard error. Table 1 reveals that the relative bias of the three estimators are all small with values about 2%, and with the best result for the proposed estimator  $\hat{Y}$ . The variance results in Table 1 are similar for the three estimators.

Estimator	Relative bias	Variance
$\hat{Y}$	0.0189	27.8965
$\hat{Y}^{true}$	0.0212	27.8338
$\hat{Y}^r$	0.0264	27.8233

**Table 1.** Monte Carlo relative bias and variance of the estimators in Example 1

*Example 2:* A second simulation study has been carried out using the same set-up as before, but  $p_{kj}$  were independent uniform randomly generated between 0 and 0.1. The average response rate was about 0.58 in the simulation. Table 2 presents the Monte Carlo results. The proposed estimator has the best Monte Carlo relative bias equal to 0.0081. As in the previous case, the variance results in Table 2 are similar for the three estimators. The accuracy of the proposed estimator was almost equal to the accuracy of the ‘ideal’ estimator using the true response probabilities.

Estimator	Relative bias	Variance
$\hat{Y}$	0.0081	33.5262
$\hat{Y}^{true}$	0.0088	33.7678
$\hat{Y}^r$	0.0116	33.4684

**Table 2.** Monte Carlo relative bias and variance of the estimators in Example 2

## 4 Conclusion and discussion

A method of adjustment for nonresponse in sample surveys was considered. No information about the nonrespondents and no auxiliary information are required in this method. We have computed an estimator of a total based on estimated response probabilities using a Rasch model. We can consider that a pseudo-auxiliary information used in this estimator is the vector of  $x_{kj}$

and the latent variable  $\theta_k$ . Limited simulations show that adjustment using the Rasch model reduces bias, because it incorporates additional information contained in these pseudo-auxiliary variables. Further research should focus on the behavior of the Rasch model in the adjustment for nonresponse process.

Finally, two remarks should be made. The first remark concerns the auxiliary information, usually used in survey sampling to improve the estimators' performance. The proposed estimator does not require any auxiliary information about the nonrespondents or the respondents. The main advantage of this estimator is its possible application in surveys where no auxiliary information is available. For this reason, we did not include in the Monte Carlo simulation study any other estimator based on auxiliary information. It is well known that an available auxiliary information, highly correlated with the variable of interest, and used to estimate the response probabilities improves a lot the performance of the estimator (2); it is obvious that such an estimator will be superior to the proposed estimator.

The second remark concerns the conditional independence assumption in the Rasch model. In nonresponse literature it is a usual way to use Poisson sampling to model the response behavior by assuming that the sample units in the set  $r$  are selected with 'response probabilities' and this response is independent from unit to unit. The conditional independence assumption in the Rasch model is a similar condition applied to items. Both assumptions are strong, sometimes they are in doubt, yet they are necessary in the statistical inferential process.

## References

- BAKER, F., and KIM, S.H. (2004): *Item Response Theory*. Marcel Dekker, New York, 2nd edition.
- DE BOECK, P., and WILSON, M. (Eds.) (2004): *Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach*. New York: Springer.
- KIM, J.K. and KIM, J.J. (2007): Nonresponse weighting adjustment using estimated response probability, *Canadian Journal of Statistics*, 35, 501–514.
- RASCH, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.
- RASCH, G. (1961): On general laws and the meaning of measurement in psychology, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV*. Berkeley: University of Chicago Press, 1961, 321–333.
- RIZOPOULOS D. (2006): ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, *Journal of Statistical Software*, 17 (5), 1–25. URL <http://www.jstatsoft.org/v17/i05/>

# Estimation of the Bivariate Distribution Function for Censored Gap Times

Luís Meira-Machado<sup>1</sup> and Ana Moreira<sup>1</sup>

Department of Mathematics and Applications, University of Minho  
4800-058 Azurém, Guimarães, Portugal, *lmachado@math.uminho.pt*

**Abstract.** In many medical studies, patients may experience several events. The times between consecutive events (gap times) are often of interest and lead to problems that have received much attention recently. In this work we consider the estimation of the bivariate distribution function for censored gap times. Some related problems such as the estimation of the marginal distribution of the second gap time is also discussed. We introduce a nonparametric estimator of the bivariate distribution function based on Bayes' theorem and Kaplan-Meier survival function and compare its performance with related estimators in literature. In addition we explore the behavior of the estimators through simulations.

**Keywords:** bivariate censoring, Kaplan-Meier, nonparametric estimation

## 1 Introduction

In longitudinal studies of disease, patients can experience several events through a follow-up period. In these studies, the sequentially ordered events (gap times) are often of interest. The events of concern may be of the same nature (e.g. cancer patients may experience recurrent disease episodes) or represent different states in the disease process (e.g. alive and disease-free, alive with recurrence and dead). If the events are of the same nature this are usually referred as recurrent event, whereas if they represent different states (i.e. multi-state models) they are usually modeled through their intensity functions (Meira-Machado et al. 2009).

Let  $(T_1, T_2)$  be a pair of gap times of successive events, which are observed subjected to random right-censoring. Let  $C$  be the right-censoring variable, assumed to be independent of  $(T_1, T_2)$  and let  $Y = T_1 + T_2$  be the total time. Because of this, we only observe  $(\tilde{T}_{1i}, \tilde{T}_{2i}, \Delta_{1i}, \Delta_{2i})$ ,  $1 \leq i \leq n$ , which are  $n$  independent replications of  $(\tilde{T}_1, \tilde{T}_2, \Delta_1, \Delta_2)$ , where  $\tilde{T}_1 = T_1 \wedge C$ ,  $\Delta_1 = I(T_1 \leq C)$ , and  $\tilde{T}_2 = T_2 \wedge C_2$ ,  $\Delta_2 = I(T_2 \leq C_2)$  with  $C_2 = (C - T_1)I(T_1 \leq C)$  the censoring variable of the second gap time. Define  $\tilde{Y} = Y \wedge C$  and let  $F_1$  and  $G$  denote the distribution functions of  $T_1$  and  $C$ , respectively. Since  $T_1$  and  $C$  are independent, the Kaplan-Meier product-limit estimator based on the pairs  $(\tilde{T}_{1i}, \Delta_{1i})$ 's, consistently estimates the distribution  $F_1$ . Similarly, the distribution of the total time may be consistently estimated by the Kaplan-Meier estimator based on  $(\tilde{T}_{1i} + \tilde{T}_{2i}, \Delta_{2i})$ 's. Because  $T_2$  and  $C_2$  will be in

general dependent, the estimation of the marginal distribution for the second gap time is not a simple issue. The same applies to the bivariate distribution function  $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$ . This issue have received much attention recently. Among others it was investigated by Lin et al. (1999), van der Laan et al. (2002), van Keilegom (2004), de Uña-Álvarez and Meira-Machado (2008) or de Uña-Álvarez and Amorim (2009).

In this work we present a new estimator for the bivariate distribution function of the gap times. This estimator is based on Bayes' theorem and Kaplan-Meier survival function. This estimator is somehow related to that proposed in Lin et al. (1999) and with estimators proposed by de Uña since all use (in different ways) the Kaplan-Meier estimator (Kaplan and Meier, 1958). In 1999 Lin propose an estimator for the bivariate distribution function using inverse censoring weights based on the Kaplan-Meier estimator. On the other hand, the idea behind both estimators proposed by the Uña-Álvarez is using the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data. Difference between these two methods is that the more recent paper uses a presmoothed version of the Kaplan-Meier estimator (Dikta, 1998). Without smoothing, the estimator described in de Uña-Álvarez and Amorim (2009) reduces to that in de Uña-Álvarez and Meira-Machado (2008). We have conducted extensive simulation studies to compare all four methods regarding its bias and its variance. In these simulation studies we consider two simulated scenarios (using two different copulas) with different correlations between gap times. The performance of the estimators was also investigated for different sample sizes and different censoring percentages.

## 2 The estimators

In this section we will present four different approaches for estimating the bivariate distribution function  $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$ , all using the Kaplan-Meier estimator of survival. A simple estimator for the bivariate distribution function of the gap times is based on the on Bayes' theorem and Kaplan-Meier survival function (conditional Kaplan-Meier, CKM).

Since  $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y) = P(T_2 \leq y | T_1 \leq x)P(T_1 \leq x)$  one simple estimator for the bivariate distribution is given by

$$\hat{F}_{12}(x, y) = \hat{F}_1(x)\hat{F}_{KM}(y|T_1 \leq x, \Delta_1 = 1) \quad (1)$$

where  $\hat{F}_1(x)$  is the Kaplan-Meier product-limit estimator based on the pairs  $(\tilde{T}_{1i}, \Delta_{1i})$ 's and  $\hat{F}_{KM}(y)$  the Kaplan-Meier estimator based on the pairs  $(\tilde{T}_{2i}, \Delta_{2i})$ 's. The  $\hat{F}_{KM}(y|T_1 \leq x, \Delta_1 = 1)$  is the conditional distribution function for the subset of  $T_1 \leq x$  and  $\Delta_1 = 1$  (the Kaplan-Meier estimator based on the pairs  $(\tilde{T}_{2i}, \Delta_{2i})$ 's such that  $\tilde{T}_{1i} \leq x$  and  $\Delta_{1i} = 1$ ).

Another simple estimator was recently proposed by de Uña-Álvarez and Meira-Machado (2008). The idea behind the estimator is using the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the



bivariate data. The proposed estimator (Weighted Kaplan-Meier Estimator, WKME) is given by

$$\tilde{F}_{12}(x, y) = \sum_{i=1}^n W_i I(\tilde{T}_{1i} \leq x, \tilde{T}_{2i} \leq y) \quad (2)$$

where  $W_i = \frac{\Delta_{2i}}{n-R_i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{2j}}{n-R_j+1}\right]$  is the Kaplan-Meier weight attached to  $\tilde{Y}_i$  when estimating the marginal distribution of  $Y$  from  $(\tilde{Y}_i, \Delta_{2i})$ 's, and for which the ranks of the censored  $\tilde{Y}_i$ 's,  $R_i$ , are higher than those for uncensored values in the case of ties.

Recently, de Uña-Álvarez and Amorim (2009) propose an estimator related to (2), in which they assume a presmoothed version of the Kaplan-Meier estimator (Dikta, 1998). This estimator (Smooth Weighted Kaplan-Meier Estimator, SWKME) is expressed as

$$\tilde{\tilde{F}}_{12}(x, y) = \sum_{i=1}^n W_i^* I(\tilde{T}_{1i} \leq x, \tilde{T}_{2i} \leq y) \quad (3)$$

where  $W_i^* = \frac{m(\tilde{T}_{1i}, \tilde{Y}_i)}{n-R_i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m(\tilde{T}_{1j}, \tilde{Y}_j)}{n-R_j+1}\right]$  is the presmoothed Kaplan-Meier weight, where  $m$  stands for a parametric binary regression model. Throughout this paper we shall assume that  $m$  denotes a logistic regression model.

Note that, unlike (2), the SWKME may attach positive mass to pair of gap times with censored second gap time. However, both estimators (2) and (3) attach a zero weight to pairs of gap times with first gap time censored. Conditions under which both estimators are consistent is fully discussed in papers by de Uña-Álvarez and Meira-Machado (2008) and de Uña-Álvarez and Amorim (2009).

Another estimator for the bivariate distribution function was proposed in Lin et al. (1999). This estimator is based on inverse censoring (Kaplan-Meier) weights (ICKMW) and is expressed as

$$\tilde{F}_{12}(x, y) = \tilde{H}(x, 0) - \tilde{H}(x, y) \quad (4)$$

where

$$\tilde{H}(x, y) = \frac{1}{n} \sum_{i=1}^n \frac{I(\tilde{T}_{1i} \leq x, \tilde{T}_{2i} > y)}{1 - \hat{G}((\tilde{T}_{1i} + y)^-)}$$

and where  $\hat{G}$  stands for the Kaplan-Meier estimator of the censoring distribution based on the  $(\tilde{Y}_i, 1 - \Delta_{2i})$ 's.

From (1), (2) and (3) we can obtain an estimator for the marginal distribution of the second gap time,  $F_2(y) = P(T_2 \leq y)$ , namely

$$\hat{F}_2(y) = \hat{F}_{12}(+\infty, y) = \hat{F}_{KM}(y|\Delta_1 = 1) \quad (5)$$

$$\tilde{F}_2(y) = \tilde{F}_{12}(+\infty, y) = \sum_{i=1}^n W_i I(\tilde{T}_{2i} \leq y) \quad (6)$$

Note that estimator (5) is the Kaplan-Meier estimator based on  $(\tilde{T}_{2i}, \Delta_{2i})$ 's such that  $\Delta_1 = 1$  (i.e., for which the first gap time is uncensored). Estimator (6) is different because the Kaplan-Meier weights  $W_i$  in estimator are based on the  $\tilde{Y}_i$ -ranks rather than on the  $\tilde{T}_{2i}$ -ranks. Indeed, since  $T_2$  and  $C_2$  are expected to be dependent, the ordinary Kaplan-Meier estimator of  $F_2$  (estimator (5)) will be in general inconsistent. The corresponding estimator for (3) is obtained using the same ideas as for (6) by replacing the weights  $W_i$  by the presmoothed Kaplan-Meier weight  $W_i^*$  previously defined.

We note that, as mentioned in the introduction section other estimators were proposed to estimate the bivariate distribution function. A valid estimator of the bivariate distribution function, was provided by Van Keilegom (2004) which is based on Akritas (1994). However this approach has some limitations since some smoothing is required.

In the next section we report an extensive simulation study comparing the four approaches presented above.

### 3 Simulation study

In this section, we compare by simulations the four estimators (1)-(4), for the bivariate distribution function. We consider two simulated scenarios, the first scenario is the same as that described in Lin's paper (see their Section 3). In this scenario, the gap times were generated from Gumbel's bivariate distribution function, the so-called Fairlie-Gumbel-Morgenstern families of bivariate cdf's,  $F(x, y) = F_1(x)F_2(y)[1 + \delta(1 - F_1(x))(1 - F_2(y))]$  where  $|\delta| \leq 1$  for a bivariate density to exist. The marginal distributions,  $F_1$  and  $F_2$  are exponential with rate parameter 1. The case of independence is obtained for  $\delta = 0$  while the maximum of correlation (between  $T_1$  and  $T_2$ ) for the bivariate exponential distribution is obtained for  $\delta = 1$  with bound equal to 0.25. As in Lin's paper, for this scenario, the uniform censoring time  $C$  was generated according to models  $U[0, 4]$  and  $U[0, 3]$ . The first model ( $U[0, 4]$ ) resulted in 25% of censoring of the first gap time, and 46% of censoring in the second gap time. In the second model ( $U[0, 4]$ ) we have censoring levels of 32% and 60% for the corresponding gap times.

One limitation of the so-called Fairlie-Gumbel-Morgenstern families of bivariate cdf's, is that the correlation of  $T_1$  and  $T_2$  can never exceed 1/3 (0.25 in the bivariate exponential distribution). One potential category of bivariate distributions is the family of bivariate Weibull distributions. This distribution clearly allows for a larger correlation between the two gap times, making it superior than the bivariate exponential. For this reason, in our second scenario we consider the bivariate Weibull distribution with two-parameter marginal distributions. Its survival function is given by  $S(x, y) = P(T_1 >$

$$x, T_2 > y) = \exp \left[ - \left[ \left( \frac{x}{\theta_1} \right)^{\frac{\beta_1}{\delta}} + \left( \frac{y}{\theta_1} \right)^{\frac{\beta_2}{\delta}} \right]^{\delta} \right] \text{ where } 0 < \delta \leq 1, \text{ and each}$$

marginal distribution has shape parameter  $\beta_i$  and a scale parameter  $\theta_i$ . The correlation between the two gap times can be obtained though is a complicated function of the shape and scale parameters and of  $\delta$ . For our simulation we considered  $\delta = 0.6$ ,  $\theta_1 = \theta_2 = 7$  and shape parameters  $\beta_1 = \beta_2 = 2$ , for which we obtained about 54% of correlation.

For each scenario we have considered two sample sizes,  $n = 50$  and  $n = 100$  and for each simulation, 1000 samples were generated. For each setting we computed the mean and standard deviations for the bivariate estimators at pairs of time points  $(x, y)$ , where  $x$  and  $y$  takes values corresponding to: marginal survival probabilities of 0.8, 0.6, 0.4 and 0.2 for the bivariate exponential scenario; and to marginal survival probabilities of 0.8, 0.6, 0.4, 0.2 and 0.1 for the bivariate Weibull scenario. The true values of  $F_{12}(x, y)$  are reported in Tables 1 and 2.

$\delta = 0$					$\delta = 1$				
y 0.2231 0.5108 0.9163 1.6094					0.2231 0.5108 0.9163 1.6094				
x									
0.2231	0.0400	0.0800	0.1200	0.1600	0.0656	0.1184	0.1584	0.1856	
0.5108	0.0800	0.1600	0.2400	0.3200	0.1184	0.2176	0.2976	0.3584	
0.9163	0.1200	0.2400	0.3600	0.4800	0.1584	0.2976	0.4176	0.5184	
1.6094	0.1600	0.3200	0.4800	0.6400	0.1856	0.3584	0.5184	0.6656	

**Table 1.** True values for the bivariate Exponential distribution

y 3.3067 5.0030 6.7006 8.8805 10.622						
x						
3.3067	0.1130	0.1574	0.1800	0.1930	0.1972	
5.0030	0.1574	0.2610	0.3294	0.3741	0.3895	
6.7006	0.1800	0.3294	0.4494	0.5406	0.5751	
8.8805	0.1930	0.3741	0.5406	0.6872	0.7500	
10.622	0.1972	0.3895	0.5751	0.7500	0.8305	

**Table 2.** True values for the bivariate Weibull distribution

In this paper we only present the results for a sample size of  $n = 100$  and the highest censoring levels (for each setting). Remaining results (not shown) reveal that, in general, the bias increases for higher censoring levels and decreases with the increasing of the sample size. Results for the other sample sizes and other censoring percentages will be presented elsewhere.

Tables 3 and 4 report the mean estimate along with the corresponding standard deviation for estimators (1)-(4). As it can be seen, in all estimators

$\delta = 0$						$\delta = 1$			
y		0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
x									
CKM	0.2231	0.0652 (0.025)	0.1194 (0.036)	0.1565 (0.039)	0.1845 (0.043)	0.0652 (0.026)	0.1175 (0.035)	0.1585 (0.039)	0.1854 (0.042)
	0.5108	0.1203 (0.035)	0.2186 (0.044)	0.2993 (0.050)	0.3596 (0.052)	0.1194 (0.035)	0.2163 (0.042)	0.3002 (0.050)	0.3592 (0.056)
	0.9163	0.1635 (0.040)	0.2996 (0.052)	0.4259 (0.060)	0.5247 (0.061)	0.1639 (0.040)	0.3023 (0.051)	0.4217 (0.057)	0.5222 (0.063)
	1.6094	0.1969 (0.047)	0.3764 (0.058)	0.5369 (0.064)	0.6776 (0.070)	0.1985 (0.047)	0.3786 (0.061)	0.5405 (0.066)	0.6817 (0.068)
ICKMW	0.2231	0.0405 (0.023)	0.0783 (0.029)	0.1223 (0.037)	0.1603 (0.043)	0.0709 (0.027)	0.1174 (0.037)	0.1585 (0.042)	0.1830 (0.042)
	0.5108	0.0811 (0.032)	0.1576 (0.045)	0.2369 (0.050)	0.3222 (0.061)	0.1181 (0.034)	0.2178 (0.044)	0.3090 (0.058)	0.3575 (0.055)
	0.9163	0.1201 (0.040)	0.2404 (0.051)	0.3592 (0.062)	0.4816 (0.072)	0.1570 (0.042)	0.2981 (0.058)	0.4205 (0.066)	0.5237 (0.074)
	1.6094	0.1599 (0.050)	0.3165 (0.065)	0.4811 (0.080)	0.6570 (0.088)	0.1864 (0.056)	0.3681 (0.065)	0.5289 (0.078)	0.6938 (0.081)
WKME	0.2231	0.0411 (0.020)	0.0810 (0.028)	0.1187 (0.037)	0.1589 (0.043)	0.0652 (0.026)	0.1176 (0.035)	0.1589 (0.041)	0.1852 (0.045)
	0.5108	0.0807 (0.029)	0.1581 (0.040)	0.2388 (0.049)	0.3168 (0.058)	0.1174 (0.033)	0.2202 (0.046)	0.2949 (0.050)	0.3602 (0.059)
	0.9163	0.1206 (0.036)	0.2419 (0.050)	0.3608 (0.059)	0.4805 (0.069)	0.1563 (0.041)	0.2993 (0.053)	0.4210 (0.060)	0.5200 (0.069)
	1.6094	0.1627 (0.043)	0.3208 (0.058)	0.4816 (0.068)	0.6380 (0.082)	0.1900 (0.046)	0.3584 (0.057)	0.5162 (0.063)	0.6600 (0.082)
SWKME	0.2231	0.0427 (0.019)	0.0840 (0.026)	0.1218 (0.033)	0.1594 (0.037)	0.0659 (0.024)	0.1184 (0.033)	0.1593 (0.038)	0.1838 (0.042)
	0.5108	0.0835 (0.027)	0.1632 (0.037)	0.2425 (0.045)	0.3172 (0.054)	0.1193 (0.031)	0.2217 (0.044)	0.2953 (0.048)	0.3590 (0.055)
	0.9163	0.1232 (0.033)	0.2451 (0.046)	0.3609 (0.054)	0.4739 (0.062)	0.1590 (0.038)	0.3013 (0.049)	0.4203 (0.056)	0.5161 (0.063)
	1.6094	0.1626 (0.038)	0.3193 (0.053)	0.4745 (0.063)	0.6327 (0.072)	0.1926 (0.043)	0.3594 (0.051)	0.5093 (0.057)	0.6542 (0.073)

**Table 3.** Estimated values for the bivariate exponential distribution with standard deviation. Sample size of  $n = 100$ , uniform censoring  $C \sim U[0, 3]$ .

the bias of the bivariate distribution achieved reasonable levels. In all cases the variance increases at the right tail of the bivariate distribution, where the censoring effects are stronger. From these tables we can see that: (a) the CKM estimator has larger bias for higher values of  $T_1$ , the first gap time, but in general is one of the estimators with less variance; (b) the WKME estimator has less bias than its smooth version, SWKME; however as expected the later obtained less variance; (c) the WKME and ICKMW estimator are almost unbiased but the last one obtains higher levels of variance for small values

	y	3.3067	5.0030	6.7006	8.8805	10.622
x						
CKM	3.3067	0.1164 (0.036)	0.1608 (0.042)	0.1830 (0.046)	0.1944 (0.045)	0.1958 (0.045)
	5.0030	0.1658 (0.046)	0.2661 (0.053)	0.3333 (0.060)	0.3726 (0.059)	0.3899 (0.058)
	6.7006	0.1979 (0.051)	0.3496 (0.061)	0.4614 (0.066)	0.5431 (0.065)	0.5714 (0.065)
	8.8805	0.2222 (0.057)	0.4100 (0.070)	0.5606 (0.072)	0.7029 (0.065)	0.7558 (0.064)
	10.622	0.2347 (0.060)	0.4341 (0.072)	0.6139 (0.076)	0.7704 (0.065)	0.8390 (0.061)
	3.3067	0.1127 (0.038)	0.1565 (0.044)	0.1793 (0.045)	0.1933 (0.044)	0.1982 (0.045)
	5.0030	0.1555 (0.052)	0.2616 (0.059)	0.3315 (0.063)	0.3700 (0.059)	0.3922 (0.059)
	6.7006	0.1830 (0.058)	0.3306 (0.067)	0.4506 (0.072)	0.5377 (0.064)	0.5757 (0.066)
	8.8805	0.1931 (0.061)	0.3717 (0.077)	0.5407 (0.081)	0.6913 (0.073)	0.7521 (0.066)
	10.622	0.1988 (0.065)	0.3991 (0.078)	0.5763 (0.082)	0.7525 (0.080)	0.8304 (0.070)
ICKMW	3.3067	0.1107 (0.038)	0.1563 (0.045)	0.1811 (0.048)	0.1949 (0.050)	0.1999 (0.054)
	5.0030	0.1557 (0.046)	0.2608 (0.058)	0.3338 (0.064)	0.3803 (0.070)	0.3897 (0.070)
	6.7006	0.1806 (0.049)	0.3353 (0.064)	0.4475 (0.074)	0.5466 (0.078)	0.5796 (0.077)
	8.8805	0.1944 (0.052)	0.3746 (0.066)	0.5444 (0.075)	0.6876 (0.080)	0.7512 (0.078)
	10.622	0.1970 (0.052)	0.3910 (0.071)	0.5784 (0.079)	0.7502 (0.081)	0.8346 (0.073)
	3.3067	0.1036 (0.035)	0.1465 (0.045)	0.1704 (0.046)	0.1848 (0.049)	0.1894 (0.051)
	5.0030	0.1543 (0.041)	0.2578 (0.055)	0.3250 (0.062)	0.3744 (0.065)	0.3912 (0.070)
	6.7006	0.1813 (0.046)	0.3253 (0.060)	0.4520 (0.070)	0.5484 (0.076)	0.5859 (0.077)
	8.8805	0.2007 (0.048)	0.3689 (0.064)	0.5346 (0.075)	0.6860 (0.075)	0.7605 (0.073)
	10.622	0.2014 (0.047)	0.3841 (0.066)	0.5666 (0.076)	0.7440 (0.075)	0.8282 (0.067)
WKME	3.3067	0.1107 (0.038)	0.1563 (0.045)	0.1811 (0.048)	0.1949 (0.050)	0.1999 (0.054)
	5.0030	0.1557 (0.046)	0.2608 (0.058)	0.3338 (0.064)	0.3803 (0.070)	0.3897 (0.070)
	6.7006	0.1806 (0.049)	0.3353 (0.064)	0.4475 (0.074)	0.5466 (0.078)	0.5796 (0.077)
	8.8805	0.1944 (0.052)	0.3746 (0.066)	0.5444 (0.075)	0.6876 (0.080)	0.7512 (0.078)
	10.622	0.1970 (0.052)	0.3910 (0.071)	0.5784 (0.079)	0.7502 (0.081)	0.8346 (0.073)
	3.3067	0.1036 (0.035)	0.1465 (0.045)	0.1704 (0.046)	0.1848 (0.049)	0.1894 (0.051)
	5.0030	0.1543 (0.041)	0.2578 (0.055)	0.3250 (0.062)	0.3744 (0.065)	0.3912 (0.070)
	6.7006	0.1813 (0.046)	0.3253 (0.060)	0.4520 (0.070)	0.5484 (0.076)	0.5859 (0.077)
	8.8805	0.2007 (0.048)	0.3689 (0.064)	0.5346 (0.075)	0.6860 (0.075)	0.7605 (0.073)
	10.622	0.2014 (0.047)	0.3841 (0.066)	0.5666 (0.076)	0.7440 (0.075)	0.8282 (0.067)
SWKME	3.3067	0.1036 (0.035)	0.1465 (0.045)	0.1704 (0.046)	0.1848 (0.049)	0.1894 (0.051)
	5.0030	0.1543 (0.041)	0.2578 (0.055)	0.3250 (0.062)	0.3744 (0.065)	0.3912 (0.070)
	6.7006	0.1813 (0.046)	0.3253 (0.060)	0.4520 (0.070)	0.5484 (0.076)	0.5859 (0.077)
	8.8805	0.2007 (0.048)	0.3689 (0.064)	0.5346 (0.075)	0.6860 (0.075)	0.7605 (0.073)
	10.622	0.2014 (0.047)	0.3841 (0.066)	0.5666 (0.076)	0.7440 (0.075)	0.8282 (0.067)
	3.3067	0.1036 (0.035)	0.1465 (0.045)	0.1704 (0.046)	0.1848 (0.049)	0.1894 (0.051)
	5.0030	0.1543 (0.041)	0.2578 (0.055)	0.3250 (0.062)	0.3744 (0.065)	0.3912 (0.070)
	6.7006	0.1813 (0.046)	0.3253 (0.060)	0.4520 (0.070)	0.5484 (0.076)	0.5859 (0.077)
	8.8805	0.2007 (0.048)	0.3689 (0.064)	0.5346 (0.075)	0.6860 (0.075)	0.7605 (0.073)
	10.622	0.2014 (0.047)	0.3841 (0.066)	0.5666 (0.076)	0.7440 (0.075)	0.8282 (0.067)

**Table 4.** Estimated values for the bivariate Weibull distribution with standard deviation. Sample size of  $n = 100$ .

of the second gap time,  $T_2$ . (d) in the second setting, for larger values of  $T_2$ , the ICKMW obtains less variance than both WKME and SWKME.

## 4 Conclusion

In this paper we present several nonparametric estimators of the bivariate distribution function for censored gap times. We use these estimators to introduce also an estimator for the marginal distribution of the second gap time. Simulations showed that most of these estimators are virtually unbiased. We also study their efficiency with respect to their variance. To this point, we note that, though the smooth version of the estimator introduced by de Uña-Álvarez and Meira-Machado (2008) obtains slight higher bias, it may achieve efficiency levels above. We note that, in contrast to the other two methods, the estimators by Uña-Álvarez and Meira-Machado (2008) and Uña-Álvarez and Amorim (2009) are a proper distribution function, in the sense that it attaches positive mass to each observation.

## Acknowledgments

The authors acknowledge receiving financial support from the Portuguese Ministry of Science, Technology and Higher Education in the form of grants PTDC/MAT/104879/2008 and SFRH/BD/62284/2009. The research was also partially funded by CMAT and FCT under the POCI 2010 program.

## References

- UÑA-ÁLVAREZ, J. and MEIRA-MACHADO, L. (2008): A simple estimator of the bivariate distribution function for censored gap times. *Statistics and Probability Letters* 78, 2440-2445.
- UÑA-ÁLVAREZ, J. and AMORIM, A.P. (2009) A semiparametric estimator of the bivariate distribution function for censored gap times. Discussion Papers in Stats OR, Report 09/03. Dept. Estadística e IO, U. Vigo, Spain.
- DIKTA, G. (1998). On semiparametric random censorship models. *Journal of Statistical Planning and Inference* 66, 253-279.
- KAPLAN, E.L. and MEIER, P. (1958): Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457-481.
- LIN, D.Y., SUN, W. and YING, Z. (1999): Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* 86, 59-70.
- MEIRA-MACHADO, L., UÑA-ÁLVAREZ, J., CADARSO-SUÁREZ, C. and ANDERSEN, P.K. (2009): Multi-state models for the analysis of time to event data. *Statistical Methods in Medical Research* 18, 195-222.
- VAN DER LAAN, M.J., HUBBARD, A.E. and ROBINS, J.M. (2002): Locally efficient estimation of a multivariate survival function in longitudinal studies. *Journal of the American Statistical Association* 97, 494-507.
- VAN KEILEGOM, I. (2004): A note on the nonparametric estimation of the bivariate distribution under dependent censoring. *J. Nonpar. Statist.*, 16, 659-670.

# Two Measures of Dissimilarity for the Dendrogram Multi-Class SVM Model

Rafael Pino Mejías<sup>1</sup> and María Dolores Cubiles de la Vega<sup>1</sup>

<sup>1</sup> Departamento de Estadística e I.O. Avda. Reina Mercedes s/n, Sevilla, Spain,  
*rafaelp@us.es, cubiles@us.es*

**Abstract.** Several schemes for multi-class problems have been proposed. One of these approaches, the Dendrogram-based SVM model, builds a set of binary SVM models, arising from a hierarchical cluster analysis of the set of classes, where the matrix of dissimilarities between the classes is obtained by calculating the distances between the gravity centers. However, these vectors are not good representatives of their associated samples and other measures of dissimilarity could be more appropriate. We propose two measures, the first is based on the distance between matrices containing a set of sample quantiles, while the second one computes a distance between the empiric characteristic functions of the samples associated to the considered classes.

**Keywords:** SVM, classification, empiric characteristic function, R

## 1 Introduction

Support Vector Machines (SVM) are a powerful family of supervised machine learning techniques. They emerged from Statistical Learning Theory, or Vapnik-Chervonenkis theory, (Boser et al., 1992, Vapnik, 1998, Cristianini and Shawe-Taylor, 2002) and several extensions were successively proposed. When used for a two-class classification problem where the set of binary labeled training patterns is linearly separable, the SVM separates both classes with a hyper-plane that is maximally distant from them ("the maximal margin hyper-plane"). If linear separation is not possible, the feature space is enlarged using basis expansions such as polynomials or splines. Moreover, explicit specification of this transformation is not necessary, as a kernel function that computes inner products in the transformed space can be employed. However, the extension of this model for the multi-class scenario is still a research topic. A usual approach is based on the construction of a set of binary SVM models. For example, the one-against-all method, (Bottou et al., 1994) builds  $K$  binary SVM models for a problem with  $K$  classes, where the  $i$ -th model tries to separate the class  $i$  from the remaining categories. Thus, the classification rule for each model is based on the sign of a decision function  $m_i(x)$ . The final decision is based on the class which has the largest value of the decision functions  $m_1(x), \dots, m_K(x)$ .

Another approach is the one-against-one method, initially introduced by

Knerr et al. (1990) for neural networks, where  $K(K-1)/2$  models are obtained, one for each pair of classes, and a voting scheme provides the final decision. Although this second approach needs more SVM models, each one is computed with a subset of the training set, so it is usually more efficient than the first approach.

The directed acyclic graph SVM (DAGSVM) proposed by Platt et al. (2000) differs to the one-against-one method in the testing phase, where it uses a rooted binary directed acyclic graph which has internal nodes and leaves. Each node is a binary SVM of  $i$ -th and  $j$ -th classes. As it is explained in Hsu and Lin (2002), given an input vector to be classified, starting at the root node, the binary decision function is evaluated. Then it moves to either left or right depending on the output value. Therefore, we go through a path before reaching a leaf node which indicates the predicted class. These three methods are compared in Hsu and Lin (2002), suggesting that one-against-one and its variant DAGSVM are more suitable for practical use.

Binary tree of SVM uses multiple SVMs arranged in a binary tree structure (Fei and Liu, 2006). This approach selects two classes for training in every node, and employs probabilistic outputs to measure the similarity between remaining samples and the two classes used in training. Then, all the samples in the node are assigned to the two subnodes. The final tree arises from the repetition of the same steps on every node. Madzarov et al. (2009) proposes a clustering algorithm in the kernel space to design the binary tree of SVM. The Dendrogram-based SVM model (DSVM) (Benabdeslem and Bennani, 2006) is an alternative based on the previous realization of a hierarchical cluster analysis of the  $K$  classes. In each level of the dendrogram, a binary classification problem is formulated to separate two groups of classes. The final decision is computed by presenting the input vector to the set of SVM models in a tree decision form, until an assignation to a class is reached. However, the distance between classes is defined in Benabdeslem and Bennani (2006) as the distance between the  $K$  gravity centers. It is well known that the arithmetic mean can be a bad representative of a distribution, for example when there exist outliers in the sample, or when asymmetric distributions are obtained. The median could be more appropriate in such situations, and even a bigger number of quantiles could be computed, offering a more accurate description of the sample. This is the idea of the first measure of dissimilarity that we propose to employ for the cluster analysis inside the DSVM methodology. The second measure is based on a distance between the multivariate empiric characteristic functions of the  $K$  samples. The first measure and a practical application are presented in section 3, while the second measure and an application are presented in section 4. The main conclusions are explained in section 5. Previously, section 2 presents the SVM model as it is available in the R system.



## 2 Distance between quantiles

We consider a training data set formed by  $n$  training vectors  $x_i, y_i, i = 1, 2, \dots, n$ , where the  $p$ -dimensional vectors  $x_i$  contain the predictor features and the  $n$  labels  $y_i, i = 1, 2, \dots, K$  identify the class of each vector. A summary of each variable in each class can be obtained by computing a set of quantiles, say  $q_1, \dots, q_m$ . Thus, we define the quantile matrix  $Q_i$  of the class  $i$  as the  $m \times p$  matrix where the  $(r, l)$  position is the  $q_r$  quantile for the predictor variable  $l$  in the sample of that class, being  $i = 1, 2, \dots, K; r = 1, 2, \dots, m; l = 1, 2, \dots, p$ . Thus, a measure of dissimilarity between classes  $i$  and  $j$  can be defined through a distance  $d$  between their quantile matrices  $Q_i$  and  $Q_j$ . Let  $\mathbf{D}$  be the distance matrix so defined, whose  $(i, j)$  component is defined by  $D_{ij} = d(Q_i, Q_j), i, j = 1, 2, \dots, K$ .

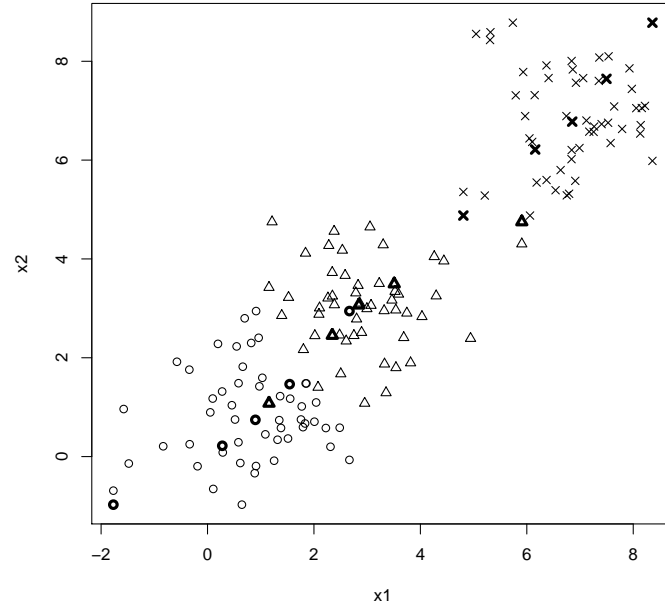
Having computed  $\mathbf{D}$ , an agglomerative cluster analysis of the  $K$  classes can be performed until only one cluster is defined. In each step,  $\mathbf{D}$  may be recomputed through the sample quantiles of the union of the previously joined classes. The obtained dendrogram is used to derive a set of binary SVM models. As a practical application, we consider a classification problem with 3 classes, where three samples of size 50 are generated from three bivariate normal populations, with zero correlation in the three populations, but with mean vectors equal to (1,1), (3,3) and (7,7). Figure 1 displays a training data set formed by 150 bi-dimensional points, where circles are points from class 1, triangles belong to class 2, and crosses are from class 3.

We have selected  $m=5$  and the following configuration of quantiles: 0, 0.25, 0.50, 0.75, 1, that is, the minimum, the quartiles and the maximum. The rows defining the three quantile matrices  $Q_i, i = 1, 2, 3$  have been superimposed over the sample points in figure 1, and they have been identified by bigger symbols. Table 1 shows the corresponding distance matrix  $\mathbf{D}$  for the training data set of figure 1, being  $d$  the euclidean distance.

Class	1	2	3
1	0	5.84	18.07
2	5.84	0	12.57
3	18.07	12.57	0

**Table 1.** Initial quantile distances for the data set of figure 1

From table 1 it is evident that the first step in the cluster process joins classes 1 and 2, and therefore, the second and last step will join the cluster 1,2 with class 3. Thus, two binary SVM models are necessary. The first model, SVM1, will try to separate class 3 from class 1,2. The second model, SVM2, is oriented to discriminate between classes 1 and 2. The final DSVM model is defined by the following process for an input vector  $x$ :

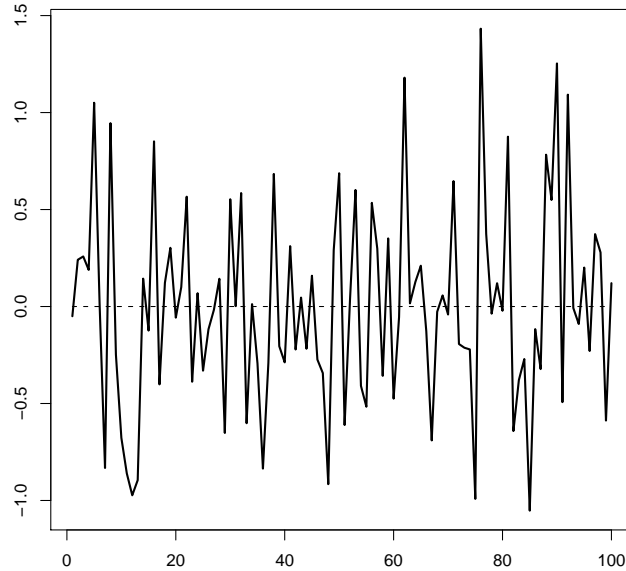


**Fig. 1.** A training data set from three bivariate normal populations.

- a. Input  $x$  to SVM1. If  $x$  is assigned to class 3, this is the decision, stop. Otherwise, go to step 2.
- b. Input  $x$  to SVM2. The decision of this model is the decision of the DSVM model (class 1 or 2).

We have compared the performance of the DSVM approach and the one-against-one method, by the random generation of 100 training data sets as in figure 1, and testing their ability over a test set formed by 1500 cases, 500 for each class. We have used the SVM implementation available in the library `e1071` of the R system. For each model, a tuning procedure has been followed to identify an appropriate pair of values for the cost  $C$  and the  $\gamma$  parameter of the radial basis kernel. Our grid search for  $C$  included small and big values:  $\{5, 50, 100, 250, 500, 750, 1000\}$ . The explored values for  $\gamma$  have been selected around  $1/p=0.5$ , namely  $\{0.1, 0.25, 0.5, 0.75, 0.9\}$ . Thus, and employing the function `tune.svm` available in the `e1071` library, for each binary SVM model a cross-validated search was performed over this grid.

Figure 2 exhibits the difference between the test error rates for the SVM one-against-one model and the DSVM method for each training data set. The 100 differences are randomly distributed around 0. Moreover, a paired t-test for the null hypothesis supporting the equality of both generalization



**Fig. 2.** Differences between the test error rates for the SVM one-against-one and DSVM, computed for 100 realizations of the data of figure 1.

errors was performed, obtaining  $p=0.923$ , and therefore accepting the similar performance of both methods.

We remark that DSVM method only needs two binary SVM models, one less than the usual one-against-one approach. This saving is more important for problems with more classes, for example when  $K=5$ , SVM only requires 4 models, while one-against-one requires 10 models. It can be argued that SVM based on one-against-all also needs  $K-1$  models, but all of them must be fitted over the whole training data set, while only one of the models included in the DSVM procedure is fitted over the complete training data set.

### 3 Distance between the empiric characteristic functions of the classes

Given the  $K$  samples from the  $K$  populations appearing in the multi-class problem, their  $K$  multivariate empirical characteristic functions (Feuerverger and Mureika, 1977, Epps, 1993), are defined as follows, for  $j = 1, 2, \dots, K$ :

$$\phi_j(t) = \int_{x \in R^p} e^{it'x} dG_j(x) = \frac{1}{n_j} \sum_{r=1}^{n_j} e^{it'x_{j,r}} \quad (1)$$

$G_1, \dots, G_K$  are the corresponding multivariate empirical distribution functions;  $x_{j,r}$  is the  $p$ -sized column vector corresponding to the  $r$ -th element of the  $n_j$  sized sample  $j$ , and  $t'$  denotes the transpose of the column vector  $t$ :

Fixed  $t$ , let  $d_{ij}(t)$  the euclidean distance between the values that  $i$ -th and  $j$ -th empiric functions take in  $t$ :

$$d_{ij}(t)^{1/2} = \|\phi_i(t) - \phi_j(t)\|. \quad (2)$$

We now consider the orthonormal basis  $T$  of the  $p$ -euclidean space:

$$T = \{t_\eta = (0, 0, \dots, \eta^{-1}, 0, 1, 0, \dots, 0)', \eta = 1, 2, \dots, p\}. \quad (3)$$

We define the following measure of dissimilarity between the  $i$ -th and  $j$ -th empiric characteristic functions, based on  $T$ , and therefore, between their corresponding classes:

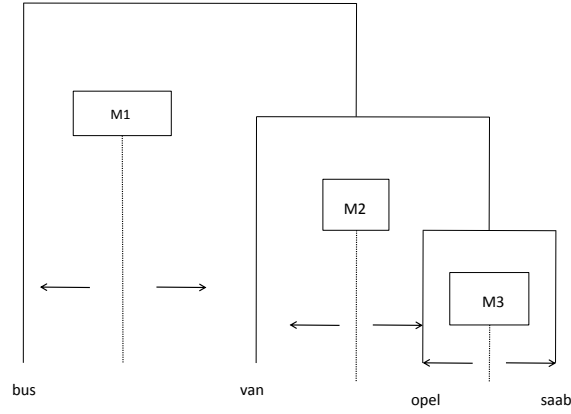
$$D(i, j) = \sum_{\eta=1}^p d_{ij}(t_\eta). \quad (4)$$

Some computational tips helping to obtain this distance matrix and to recompute the distances after each iteration of the clustering process were presented in Cubiles de la Vega et al. (1998). As an illustration, we have applied the DSVM methodology with this distance to the Vehicle data set in the library `mlbench` of R (Leisch and Dimitriadou, 2007). The purpose of this data set is to classify a given silhouette as one of four types of vehicle, using a set of 18 features extracted from the silhouette. The 846 cases were randomly split into training (75%) and test (25%) sets. This random split was independently repeated 100 times. For each training set, the distance matrix  $\mathbf{D}$  between the empiric characteristic function was computed. Table 2 contains the mean values of  $\mathbf{D}$  over the 100 iterations.

Class	bus	opel	saab	van
bus	0	1.17	1.22	1.17
opel	1.17	0	0.22	0.32
saab	1.22	0.22	0	0.29
van	1.17	0.32	0.29	0

**Table 2.** Mean distances between the empiric characteristic functions for the vehicle data set.

Table 2 reveals a clear separation between the bus class and the other three types of vehicle. Opel and saab are usually the nearest classes, although in 5



**Fig. 3.** Usual clustering of the four types of vehicle and the three binary SVM models.

splits the saab and van classes were joined in the first step. Figure 3 displays the clustering process observed in 95 of the 100 training data sets. In the other five cases, saab and van are joined in the first step, while opel forms with van and saab another cluster in the second step, being again bus the class that is finally aggregated.

Figure 3 also shows the three binary SVM models resulting from the hierarchical clustering process. Thus, M1 is fitted to separate bus from the aggregated class van, opel, saab. M2 is designed to discriminate between van and opel, saab. Finally, M3 is fitted to classify a vehicle as opel or saab. For each one of the fitted models, a grid search for an appropriate configuration of  $C$  and  $\gamma$  was also realized with the aid of the `tune.svm` function.  $C$  was studied in the set  $\{5, 50, 100, 250, 500, 750, 1000\}$ , while  $\{0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08\}$  was considered for  $\gamma$ . As in the previous example, a paired t-test for the null hypothesis supporting the equality of both generalization errors was also performed, obtaining  $p=0.521$ , and therefore the similar performance of both methods can be accepted.

## 4 Conclusion and future works

We have revised the idea of the Dendrogram-based SVM model, based on the realization of a hierarchical cluster analysis of the  $K$  classes, but considering a more general measure of dissimilarity between the classes, not restricted to the distance between the gravity centers. Two measures have been proposed in our work: a distance between quantile matrices, and a distance between the empiric characteristic functions. The practical cases presented in this paper

show an equivalent performance with respect to the usual one-against-one approach, but only  $K-1$  binary SVM models are required. We have already obtained similar conclusions with high dimensional bioinformatics data sets, but a wider empirical study is nowadays in process.

## References

- BENABDESLEM, K. and BENNANI, Y. (2006): Dendrogram-based SVM for Multi-Class Classification. *Journal of Computing and Information Technology*, 14(4), 283-286.
- BOSER, B., GUYON, I., and VAPNIK, V. (1992): A training algorithm for optimal margin classifiers. In: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* ACM Press, Pittsburgh, 144-152.
- BOTTOU, L., CORTES, C., DENKER, J., DRUCKER, H., GUYON, I., JACKEL, L., LECUN, Y., MULLER, U., SACKINGER, E., SIMARD, P., and VAPNIK, V. (1994): Comparison of classifier methods: A case study in handwriting digit recognition. In: *Proceedings of the International Conference on Pattern Recognition*, 77-87.
- CRISTIANINI, N. and SHAWE-TAYLOR, J. (2002): *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.
- CUBILES-DE-LA-VEGA, M., PINO-MEJÍAS, R. and MUÑOZ-GARCÍA, J. (1998): Cluster analysis of multivariate samples: a measure of dissimilarity between empiric characteristic functions. In: *Proceedings of the VI Conference International Federation of Classification Societies on Pattern Recognition*, 86-88.
- EPPS, T. (1993): Characteristic functions and their empiric counterparts: Geometrical interpretations and applications to statistical inference. *The American Statistician*, 47, 33-38.
- FEL, B. and LIU, J. (2006): Binary tree of SVM: A new fast multiclass training and classification algorithm. *IEEE Transactions on neural networks*, 17 (3), 696-704.
- FEUERVERGER, A. and MUREIKA, R. (1977): The empiric characteristic function and its applications. *The Annals of Statistics*, 5, 88-97.
- HSU, C., and LIN, C. (2002): A comparison of Methods for Multiclass Support Vector Machines. *IEEE Transactions on Neural Networks*, 13( 2), 415-425.
- KNERR, A., PERSONNAZ, L. and DREYFUS, G. (1990): Single-layer learning revisited: A stepwise procedure for building and training a neural network. In: J. Fogelman (Ed.): *Neurocomputing: Algorithms, Architectures and Applications*. New York, Springer-Verlag.
- LEISCH, F. and DIMITRIADOU, E. (2007): Mlbench: Machine Learning Benchmark Problems. R package version 1.1-3.
- MADZAROV, G., GJORGJEVIKJ, D. and CHORBEV, I. (2009): A multi-class SVM classifier utilizing binary decision tree. *Informatica*, 33, 233-241.
- PLATT, J., CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000): Large margin DAG's for multiclass classification. In: *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, vol. 12, 547-553.
- VAPNIK, V. (1998): *Statistical Learning Theory*. John Wiley, New York.

# Visualizing the Sampling Variability of Plots

Rajiv S. Menjoge<sup>1</sup> and Roy E. Welsch<sup>2</sup>

<sup>1</sup> Operations Research Center, M.I.T.

77 Massachussetts Avenue, Cambridge, MA, *menjoge@mit.edu*

<sup>2</sup> Sloan School of Management, M.I.T.

77 Massachussetts Avenue, Cambridge, MA, *rwelsch@mit.edu*

**Abstract.** A general method for providing a description of the sampling variability of a plot of data is proposed. The motivation behind this development is that a single plot of a sample of data without a description of its sampling variability can be uninformative and misleading in the same way that a sample mean without a confidence interval can be.

The method works by using bootstrap methods to generate several plots that could have arisen from different samples from the population, and then conveying the information given in the collection of plots by methodically selecting a few representative plots in the subset.

The method includes the capacity to incorporate prior knowledge and distributional assumptions and is useful in a broad range of situations. It is illustrated with a scatter plot example and a histogram example.

**Keywords:** uncertainty visualization, bootstrap, scatterplot

## 1 Introduction

In many cases, the data sets statisticians analyze are samples from larger populations. Therefore, when a plot is made of a given data set, it must be understood that the plot could have looked entirely different if the data had been a different sample from the population of interest. This caveat exists for any report that arises from a data set, and hence confidence intervals are typically reported for various summary statistics like the sample mean, in order to describe their sampling variability. The goal of this paper is extend the idea and implementation of the confidence interval to describe a plot's sampling variability.

We create and represent the confidence interval via the following three steps, which we describe in more detail in section 3. 1. Take  $k$  (a large number) bootstrap samples from the data set, and for each sample, create the plot of interest. 2. Compute the distance between each scatter plot pair using an appropriate distance metric, and let the envelope consist of the  $\gamma \times k$  plots which are most similar to a central plot, where  $\gamma$  is an approximate confidence level between 0 and 1. 3. Utilize the distance metric in the previous step to represent the collection of plots in this envelope. In section 3, a few useful methods for doing this are proposed.

Our method for representing a plot's sampling variability includes the capacity to incorporate prior knowledge and distributional assumptions and can be implemented for several types of plots. In this paper, we illustrate the method using a scatter plot and a histogram. However, we save a more in depth exploration of the method for a longer paper.

The rest of this paper is organized as follows: Section 2 reviews bootstrap procedures and mentions related work. Section 3 describes our method more thoroughly. Section 4 provides illustrations. Lastly, section 5 concludes with a discussion about the contributions of this paper and the directions for future research.

## 2 Literature Review

### 2.1 Bootstrap Procedures

Bootstrap procedures were introduced in Efron (1979) and have been used extensively since then to estimate standard errors, confidence intervals, and sampling distributions for statistics of interest, in cases where analytical methods would be too cumbersome.

A brief description of the procedure is as follows: 1. Assume a distributional form for the population. In the nonparametric case, the assumed distributional form would be that the population can only take the  $n$  values that the data set takes, and it takes those  $n$  values with certain probabilities. 2. Estimate the parameters for the assumed distributional form, usually by maximum likelihood. In the nonparametric case, one gets that the resulting estimated population distribution is the empirical distribution, where the population can only take the  $n$  values that the data set takes and it takes each of those with probability  $\frac{1}{n}$ . 3. Draw a large number of samples from the estimated population distribution, and for each of those samples, compute the statistic of interest. In the nonparametric case, drawing samples from the estimated population distribution corresponds to sampling the data with replacement.

The several values that the statistic took in the bootstrap samples are samples from an approximate sampling distribution of the statistic of interest and can be used to compute standard errors and confidence intervals, and to visualize the sampling distribution.

### 2.2 Related Literature

We are not aware of any methods that assess the sampling variability of plots of data in their raw form (such as scatter plots and histograms), other than the method developed in this paper. However, various attempts have been made to assess the variability of plots which evolve from functions of the data. In these cases, the general procedure is to resample from the data (usually



20-30 times), create a plot object for each bootstrap sample, and then find a way of representing the plot objects, such that they can be drawn on top of each other.

Examples of this include overlaying bootstrapped nonparametric regression curves formed by the data (see Härdle (1990) for more details), and overlaying bootstrapped principal components plots (Chateau and Lebart, 1996).

Our methodology is related to the literature above, but distinguishes itself in the types of plots, whose sampling variability it tries to assess, and in the method it uses to filter and represent the groups of bootstrapped plots. This method works well in cases where it is cumbersome or impossible for plot objects to be layered, thereby allowing easy generalizations to a variety of plots.

### 3 Methodology

In this section, we flesh out the details of the steps which were outlined in the introduction.

#### 3.1 Step 1

In step 1, we take  $k$  (a large number) bootstrap samples from our data set and form a plot with each of these samples, so that  $k$  plots are produced. Where  $n$  is the size of the data set, each bootstrap sample is a sample of size  $n$  from an estimated distribution.

This estimated distribution is created by assuming a distributional form for the data generating process and then estimating the parameters of the distributional form, usually by maximum likelihood. In the nonparametric bootstrap, the assumed distributional form is simply discrete with a probability  $p_i$  that the sample observation will take the value of observation  $i$  in the given data set, and the parameters are  $p_1, p_2, \dots, p_n$ . In the case of no prior knowledge, maximizing the likelihood for the nonparametric case yields  $p_i = \frac{1}{n} \forall i$ .

If prior knowledge exists, we merely modify our estimate of the distribution parameter and then proceed as before. The parameter estimate can be modified by replacing the estimate, which doesn't incorporate prior knowledge, with the posterior mean or the posterior mode.

As an example, if we were to make no parametric assumptions and impose the prior knowledge that the mean of the first variable,  $X_1$ , is less than or equal to 0.01, we could solve the following optimization problem in order to give us the posterior mode when the prior is uniform:  $\max \sum_{i=1}^n \log(p_i)$  (maximize log likelihood) subject to the constraints:  $\sum_{i=1}^n p_i = 1$ ,  $p_i \geq 0, \forall i$  (in other words, the values  $p_i$  are valid probabilities), and  $\sum_{i=1}^n p_i X_{1i} \leq 0.01$  (our prior knowledge is imposed), where  $x_{1i}$  represents the realization  $i$  of variable  $X_1$ .

### 3.2 Step 2

In this step, we choose an appropriate distance metric, which describes the distance between two plots. We use the chosen distance metric to create a  $k$  by  $k$  matrix of each plot's distance from each of the other plots. Following this, we define a central plot as the plot whose summed distances to other plots is minimized. Lastly, we collect the  $\gamma \times k$  plots which are closest to the central plot and let these be the plots in our envelope of interest. It is this envelope of plots which we seek to visualize in the next step.

The appropriate distance metric will depend on the type of plot being considered. In the examples given in this paper, we use Earth Mover's Distance (Peleg, Werman, and Rom, 1989), which can be used to compute plot distances for several types of plots. Earth Mover's Distance, a distance metric between multivariate histograms, has its name because if each distribution is viewed as a pile of dirt, the Earth Mover's Distance between the distributions is the minimal cost of turning one pile of dirt into the other, where the cost is assumed to be the amount of dirt moved times the distance by which it is moved. The Earth Mover's Distance satisfies the properties of a distance metric and has additional properties, such as an equivalence to Mallows's Distance (Levina and Bickel, 2001). Additionally, it has been shown to have very good empirical performance (in terms of agreeing with measures of distinction based on the human eye) (Rubner, Tomasi, and Guibas, 1998).

The optimization problem that needs to be solved in the computation of Earth Mover's Distance is a transportation problem: in particular, an uncapacitated minimum cost flow problem, whose computation time is  $O(n^3 \log(n))$ , where  $n$  is the number of bins in the histogram (Korte and Vygen, 2000).

Distance metrics between several types of plots can be derived using Earth Mover's Distance. Scatter plots, for instance, can be treated as bivariate histograms, where each point in a scatter plot represents a bar of height 1 in the corresponding bivariate histogram. Parallel coordinate plots can also be viewed as histograms, but in the parallel coordinate space. Meanwhile, the distance between two scatterplot matrices can be the sum of the distances between each corresponding pair of scatter plots in the matrices.

### 3.3 Step 3

In this step, we attempt to represent the multitude of plots in the envelope. Given that we have distances between each pair of plots, this can be done in several different ways. In the illustrations given in this chapter, we merely report the two plots which are furthest from one another for simplicity. This gives a sense of the "border" of the set of plots.

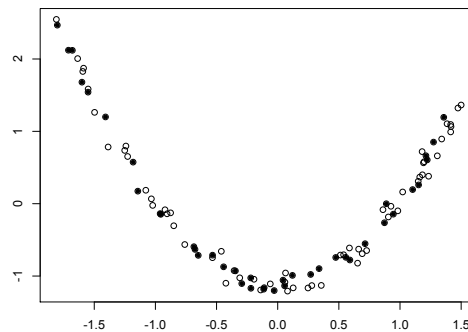
Nevertheless, several other summaries exist. As an example, the plots can be embedded as points in a higher dimensional space, such that the distances between the points in the higher dimensional space are close to the distances between the corresponding plots that the points represent. Border points and

cluster centers can then be extracted from the higher dimensional space and the corresponding plots could be represented.

## 4 Illustration

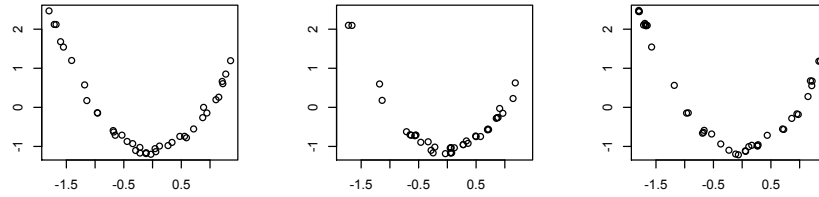
We apply the method on two contrasting examples in this section in order to demonstrate the output of the method. In our examples, we use  $k = 1000$  and  $\gamma = 95\%$ .

Our first illustration uses a sample of 40 points in two variables and exhibits a clear relationship. Figure 1 shows the sample embedded in the population and Figure 2 shows the sample plot, along with the two bootstrap plots in the envelope which are farthest apart. In the plots, a random jitter is added, so that overlapping points can be seen. In this case, the two scatter plots which are farthest apart still tell a similar story to each other, to the original plot, and to the population, in part because of the sample size, and in part because the noise around the relationship is quite small.



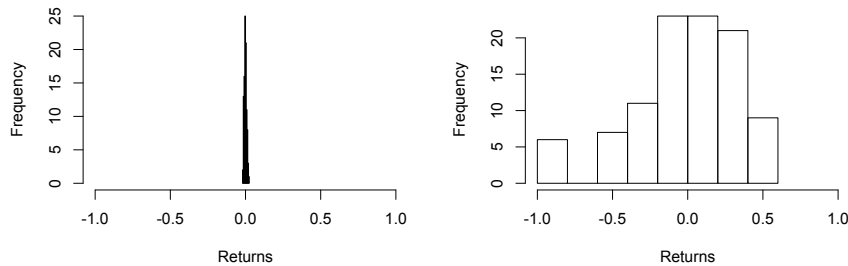
**Fig. 1.** A scatter plot of a population of 100 observations. A sample data set of 40 observations are filled in.

Our second illustration yields very different results. The data in the second example contain returns on 50 industries among the MSCI US Equity Indices. Returns are daily, beginning 01/03/1995 and ending 02/07/2005. In our example, we use the first 100 of these days to compute the portfolio weights that maximize the Sharpe ratio and then apply these portfolio weights for the next 100 days and plot the histogram to assess the distribution of future returns, given the investment strategy. The methodology used to find the portfolio weights which maximize the Sharpe ratio was originally proposed by Markowitz (1952), who initiated the mathematical framework for portfolio optimization.



**Fig. 2.** The sample of 40 points (left) and the two bootstrap sample scatter plots farthest from each other

Figure 3 demonstrates the sampling variability of such a histogram. The two histograms in the figure show the two farthest bootstrapped histograms based on our procedure. Not only is the variability of the distribution, as represented by the histogram, entirely different between the two plots, but the histogram in the right panel is skewed to the left. The analysis demonstrates that one must be careful when using the highly variable portfolio weights returned by portfolio optimization to invest in assets.



**Fig. 3.** The two extremes for return histograms with maximum Sharpe ratio portfolios

## 5 Contributions and Future Work

Our key contribution is to develop a general method for expressing the sampling uncertainty in plots of raw data, such as scatterplots and histograms. Other literature addressing the problem of the sampling variability of plots focuses on plots of certain types of projections of data, typically in cases

where the plot objects can be drawn on top of each other. Our method applies in these cases as well, though we believe that these cases aren't the only cases where sampling variability of plots should be presented. In fact, we believe that sampling variability should be reported even in cases as extreme as the first example of the previous section for validity.

We do not include comparisons with other types of plots in this paper because we are not aware of other methods which address the types of plots we considered in our illustrations. However, for the sake of discussion, one could argue that another valid approach to representing the sampling variability of plots is to simply output a sample of 20-30 bootstrapped plots. This may be possible in some cases, where the representation of the plot is relatively small. However, such an approach does nothing to filter plots which are outlying (for instance, as an extreme example, one of the 30 bootstrapped samples of Figure 1 could include 40 realizations of one point). Also, we believe that choosing a smaller representative sample out of 1000 plots is a statistically more valid approach and will likely make its way into statistical practice faster than outputting 20-30 plots for each plot given in an analysis. Furthermore, outputting 20-30 bootstrapped plots in the case of a scatterplot matrix or a set of regression diagnostic plots may be impossible, due to the amount of space it would use (20-30 pages) and the resulting difficulty of its interpretation. In fact, even in the case of simple plots, such as scatter plots and histograms, conducting a comparison in this paper would not be possible because it would cause the paper to exceed the page limit.

In this paper, we have only scratched the surface of the possible ways our method can be used. From a practical perspective, the method in this paper can be extended to several other types of plots and visualization problems. From a theoretical perspective, it would be interesting to analyze the extent to which incorporating prior knowledge is useful in our setting. In addition, there are many limitations and biases that bootstrap methods produce, which have not been addressed in this paper. One, for instance, is that bias corrections can produce better bootstrapped confidence intervals for many types of point estimates. We would expect this to be the case for plots as well.

## 6 Acknowledgements

This research was supported, in part, by a grant from the MIT-Singapore Alliance for Computational and Systems Biology and by the MIT Center for Computational Research in Economics and Management Science.

## References

- CHATEAU, R., LEBART, L. (1996): Assessing sample variability in visualization techniques related to principal component analysis: bootstrap and alternative simulation methods. *Computational Statistics* Prats, A. (ed.), 205-210.

- EFRON, B. (1979): Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7, 1-26.
- HAERDLE, W. (1990): Applied Non-parametric Regression. *Oxford University Press*.
- KORTE, B. and Vygen, J. (2000): *Combinatorial Optimization: Theory and Algorithms*. Springer, NY.
- LEVINA, E., BICKEL, P. (2001): The EarthMover's Distance is the Mallows Distance: Some Insights from Statistics. *Proceedings of ICCV*, 251-256. Vancouver, Canada.
- MARKOWITZ, H. (1952): Portfolio Selection. *Journal of Finance* 7, 77-91.
- PELEG, S., WERMAN, M., and ROM, H." (1989): A Unified Approach to the Change of Resolution: Space and Gray-level. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, 739-742.
- RUBNER, Y., TOMASI, C., and GUIBAS, L.J. (1998): A metric for distributions with applications to image databases. *Proceedings of IEEE International Conference on Computer Vision*, 59-66. Bombay, India

# Empirical Mode Decomposition for Trend Extraction: Application to Electrical Data

Farouk Mhamdi<sup>1</sup>, Mériem Jaïdane-Saïdane<sup>1</sup>, and Jean-Michel Poggi<sup>2,3</sup>

<sup>1</sup> Unité Signaux et Systèmes, ENIT,

*Farouk.Mhamdi@enit.rnu.tn, meriem.jaidane@enit.rnu.tn*

<sup>2</sup> Université Paris-Sud, Mathématiques Bât. 425, 91405 Orsay, France

*jean-michel.poggi@math.u-psud.fr*

<sup>3</sup> Université Paris Descartes, France

**Abstract.** This paper presents a method for trend extraction from seasonal time series through the Empirical Mode Decomposition (EMD). Experimental comparison of trend extraction based on EMD and Hodrick Prescott filter are conducted. First results proved the eligibility of EMD trend extraction. Tunisian real peak load<sup>1</sup> is finally used to illustrate the extraction of the intrinsic trend.

**Keywords:** Empirical Mode Decomposition, Trend extraction, Electrical

## 1 Introduction

Time series often contain many components such as seasonal and cyclical components, trends and irregularities, if we assume an additive decomposition model. However, even under this simplified form, trend extraction and seasonal adjustments are difficult tasks of time series analysis, due to the extreme variety of time series with their own time scales. Thereby, the trend has fuzzy general definition, despite its great practical importance. Nevertheless, it is considered to be a "smooth additive component that contain information about global change", see Alexandrov et al. (2008). This adopted definition make trend extraction an ambiguous task since we can found many several candidates from a time series which match this definition. Therefore, trend extraction should be related to time scales. For example, it is of great interest to extract very smoothed trend components from time series for long- or medium-term load forecasting.

Assuming additive decomposition model of time series, we will investigate in this paper the eligibility of trend extraction based on the Empirical Mode Decomposition (EMD), Huang et al. (2004). The EMD method is useful particularly to deal with possibly nonstationary and nonlinear which often characterize time series. The motivation to use the EMD is that it consider the signal as a superposition local sums of oscillatory components, extracted from upper and lower envelope, so-called Intrinsic Mode Functions (IMF).

---

<sup>1</sup> This work was supported by 2005/2007 VRR research project with Department of Studies and Planning of Tunisian Society of Electricity and Gas (STEG)

The IMFs are fully data-driven and local in time. This is important since it allows to identify various trends at different time scales. This method is easy to implement and does not use any predetermined transforms which depend on the choice of a particular theoretical structure. Furthermore, the EMD is an adaptive method which is entirely empirical and captures the characteristics in separate IMFs, explaining why it has been successfully applied in many engineering fields, e.g. Flandrin et al. (2004), Zhou et al. (2008).

The outline of the paper is as follows. Section 2 recalls some facts about Empirical Mode Decomposition and sketches how it is a good candidate for knowledge extraction. Section 3 experiments EMD-based trend extraction method on simulated seasonal time series, and compare it to the moving average (MA) filtering and the widely used trend extraction method based on Hodrick-Prescott filter (see e.g. Pollock (2003)). Then, Section 4 illustrates the method on real data by extracting trend component of the Tunisian peak load from 2000 to 2006.

## 2 Empirical Mode Decomposition and intrinsic trend

### 2.1 Time scales and trend extraction of time series

Let us consider observed additive time series  $y = (y_1, y_2 \dots y_{T_{obs}})$  supposed to be of the form:

$$y_t = T_t + S_t + C_t + I_t \quad (1)$$

where the different components are trend ( $T$ ), seasonal components ( $S$ ), cycles ( $C$ ) and irregular term ( $I$ ) for error modeling. With respect to this usual form, we will not make distinction between seasonal and cyclical components, in order to make these components identifiable. So, the signal decomposition reduces to :

$$y_t = T_t + SC_t + I_t \quad (2)$$

where  $SC$  represents seasonal and cyclical components. Several techniques have been traditionally used for time series components extraction and adjustment. As mentioned above, we will focus on trend extraction. Let us mention some the most frequently used approaches (see Alexandrov et al. (2009) for a recent review) for trend extraction such as local or global regressions, moving average filtering, X11, X12 and the Hodrick Prescott filter.

### 2.2 EMD residue as intrinsic trend

Empirical Mode Decomposition (EMD) has been introduced by Huang et al. (1998), as an important alternative to traditional methods for analyzing time series such as wavelets or Fourier methods. The key idea of EMD is



to locally decompose data  $y_t$  into oscillatory components so-called Intrinsic Mode Functions (IMF). The algorithm for the extraction of IMFs from a given time series  $y_t$  data is called sifting and it consists of the following steps:

- i Initialize the residue  $r_0(t) = y_t$ , set  $g_0(t) = r_{k-1}(t)$  and  $i = 1$ ; the index of IMF  $k = 1$
- ii Construct the lower minima  $Imin_{i-1}$  and the upper maxima  $Imax_{i-1}$  envelopes of the signal by the cubic spline method
- iii Calculate the mean values by averaging the upper envelope and the lower envelope. Set  $m_{i-1} = [Imax_{i-1} + Imin_{i-1}]/2$
- iv Subtract the mean from the original signal  
 $g_i = g_{i-1} - m_{i-1}$  and  $i = i + 1$ ,  
 and repeat steps (ii)-(iv) until  $g_i$  being an IMF (see below). If so, the  $k$ th IMF is given by  $IMF_k = g_i$
- v Update residue  $r_k(t) = r_{k-1}(t) - IMF_k$ . This residual component is treated as a new data and subjected to the process described above to calculate the next  $IMF_{k+1}$ .
- vi Repeat the steps above until the final residual component  $r(t)$  becomes a monotonic function.

It turns out that an IMF satisfy the two following properties. First: the upper and lower envelopes are symmetric and second: the number of zero-crossings and the number of extrema are equal or differ at most by one.

The advantage of this method is the fact that the oscillatory modes which are generated, are derived directly from the data without any reference to a predetermined dictionary of functions.

At the end of this process, the initial time series is decomposed into  $K$  IMF components and  $r$  is the final residue :

$$y_t = \sum_{k=1}^K IMF_k(t) + r(t) \quad (3)$$

Such a decomposition offers the opportunity to consider that  $r(t)$  as estimate of the trend of the data. Since, at the end of the algorithm, the number of extrema in the residue does not exceed 2. These results make EMD algorithm very suitable to extract trend. Note that end effect can affect the goodness fit of the trend extracted through EMD. e.g. Ren et al. (2006). In this case, no physical meaning IMFs can be obtained and the exact trend can be reconstructed by aggregation of the residue and the last or the last two IMFs.

### 2.3 Trend definitions and EMD

In this section we present a short overview of previous studies dealing with trend extraction through EMD. It is important to note that there are only a

few references, namely Zhaohua et al. 2007, Suling et al. 2009, Flandrin et al. 2004, and that there is no consensus about how to define trend, since trend definitions are related to the data peculiarities and fields of application.

Flandrin et al. 2004, has investigated the potentialities and limitations of EMD-based methods in detrending, relating the trend with the statistical properties of the IMFs. Indeed trend is defined as the sum of the IMFs having non-zero mean  $T_t = \sum_{k>D} IMF_k(t)$ . Application to heart-rate data illustrates its potential detrending usefulness.

Another definition is given in Suling et al. 2009, relating trend to time scales and  $T_t$  is supposed to be the trend of  $y_t$  on time scale  $T$  if  $\exists (t_1, t_2)$ ,  $(t_2 - t_1) > T$  such that  $(T_{t_2} - T_{t_1})(y_{t_2} - y_{t_1}) \geq 0$ . A short and partial comparison between EMD and a specific Moving Average method is made provided using Stock P&G time series.

Finally, let us mention that Zhou et al. 2008, have proposed an algorithm for removing trends from power-system oscillation data based on a slightly modified EMD. This ad-hoc adaptation is developed especially for highly oscillatory data.

In our case, trend definition and extraction are related to time scale. We investigate the performance of EMD-based approach for extraction classical long-term trend.

### 3 Simulated examples

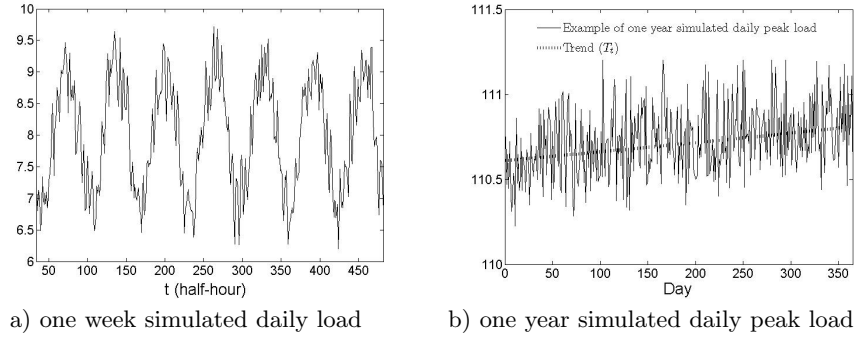
#### 3.1 Simulated seasonal series

The empirical EMD characteristic make difficult to quantify the EMD trend performance method analytically. For this, we will investigate its performance through experimental studies.

We consider elementary sinusoidal modeling for simulated daily power pattern ( $X_t$ ), supposed to be sampled every 22 minutes (see Figure 1.a). To examine trend extraction issue through EMD, a modified version ( $y_t$ ) of this simulated time series ( $X_t$ ) is obtained by adding classical trends (linear and exponential), even if they are unrealistic. The complete model is given by the following equations:

$$\begin{cases} y_t = X_t + T_t. \\ X_t = \beta_0 + \beta_1 m_1(t) + \beta_2 m_2(t) + \epsilon(t) \\ m_1(t) = \cos(\frac{2\pi t}{64}) + \sin(\frac{2\pi t}{64}) \\ m_2(t) = \cos(\frac{2\pi t}{6}) + \sin(\frac{2\pi t}{6}) \\ \epsilon(t) = \nu(t) + \theta \nu(t-1) \quad \nu(t) \sim \mathcal{N}(0, \sigma^2) iid \\ T_t = a + bt \quad \text{or} \quad T_t = a + e^{\alpha t} \end{cases} \quad (4)$$

where  $t = (1, 2, \dots, T_{obs})$ ,  $T_{obs} = 69120$ ,  $\beta_0 = 8$ ,  $\beta_1 = 0.8$ ,  $\beta_2 = 0.18$ ,  $\theta = 0.8$ ,  $\sigma^2 = 0.05$ ,  $a = 100$ ,  $b \in \{0.01, 0.02, \dots, 0.05\}$  and  $\alpha \in \{0.001, 0.0011, \dots, 0.005\}$ .



**Fig. 1.** Simulated daily power pattern

### 3.2 EMD trend extraction performance

To investigate the EMD trend extraction performance, a comparison with the nonparametric trend extraction method based on Hodrick-Prescott (HP) filter is performed. This last one is widely used by economists for trend estimation, see e.g. Pollock (2003).

For a time series  $y = (y_1, y_2 \dots y_{T_{obs}})$  supposed to contain a trend  $(T_t)$  and a cyclical component  $(C_t)$ , the best extracted trend  $(\hat{T}_t)$  is the solution of:

$$\min_{\{\hat{T}_t\}_{t=1}^{T-1}} \left\{ \sum_{t=1}^{T-1} (y_t - \hat{T}_t)^2 + \lambda \sum_{t=2}^{T-1} [(\hat{T}_{t+1} - \hat{T}_t) - (\hat{T}_t - \hat{T}_{t-1})]^2 \right\} \quad (5)$$

where the parameter  $\lambda$  is a positive number which penalizes variability in the growth rate of the trend component. The larger value of  $\lambda$ , the smoother the trend extracted and then a good extraction of a trend requires a suitably chosen value of  $\lambda$ , see Schlicht (2005) for theoretical investigation. Here we choose  $\lambda$  according to short empirical tuning based on simulated load curve for  $\lambda$  in the range  $10^2$  to  $10^{15}$ . In Table 1, are reported the Maximum of

$\alpha$		$Max\_MAE$	$Max\_AE$	Satisfactory HP $\lambda$ parameter range
$10^{-4}$	HP	0.0125	0.0365	$\lambda \in [10^9, 10^{11}]$
	EMD	0.009	0.0317	
$5 \cdot 10^{-4}$	HP	0.0146	0.0463	$\lambda \in [10^8, 3 \cdot 10^{11}]$
	EMD	0.02	0.1	
$10^{-3}$	HP	0.067	0.21	$\lambda \in [1.09 \cdot 10^5, 3.25 \cdot 10^8]$
	EMD	0.022	0.06	
$2 \cdot 10^{-3}$	HP	0.052	0.41	$\lambda \in [2.96 \cdot 10^5, 2.15 \cdot 10^8]$
	EMD	0.009	0.0317	

**Table 1.** Performances of the HP and EMD simulated daily peak trend extraction

Mean Absolute Error ( $Max\_MAE^2$ ) and the Maximum of Absolute Error ( $Max\_AE$ ) estimated for the HP and the EMD trends extracted for different values of  $\alpha$ :  $10^{-4}$ ,  $5 \cdot 10^{-4}$ ,  $10^{-3}$  and  $2 \cdot 10^{-3}$ . Note that, these values are chosen in order to allow simulated trends covering linear, quasi linear and exponential trend shapes.

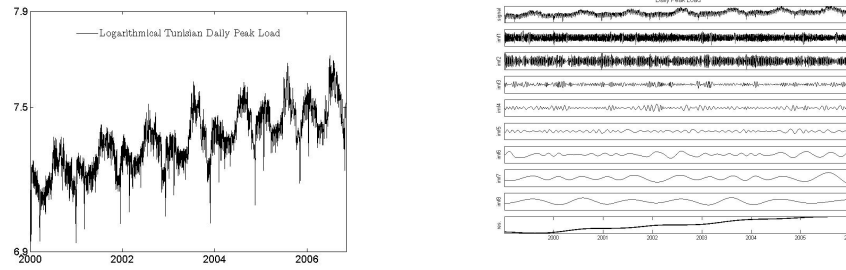
These first results show that the EMD-trend is very close to the optimal Hodrick-Prescott one and make the EMD as an effective alternative to trend extraction problem. Same results are also obtained through the EMD and a moving average filtering comparison. Indeed the EMD trends extracted are very close to those extracted through a conveniently chosen moving average filtering method. This expected finding is due to the adaptive nature of the principle of the EMD.

We notice in experiments not reported here, that high errors occur for high values of  $\alpha$  and that the end effects are so important for high value trends. In this case, the intrinsic trend can be obtained by aggregating the EMD residue and the last IMFs. It is also important to notice that there are various approaches to deal with the EMD end effect, for example by applying a window to the signal, see Ren et al. (2006). Another solution is to extrapolate end maxima and end minima to construct the lower minima  $Imin_{i-1}$  and the upper maxima  $Imax_{i-1}$  envelopes, see Zhaohua et al. (2009).

## 4 Real peak load time series

### 4.1 Peak load IMFs interpretation

We apply the EMD method to logarithmical Tunisian daily peak load from 2000 to 2006 (see Figure 2).



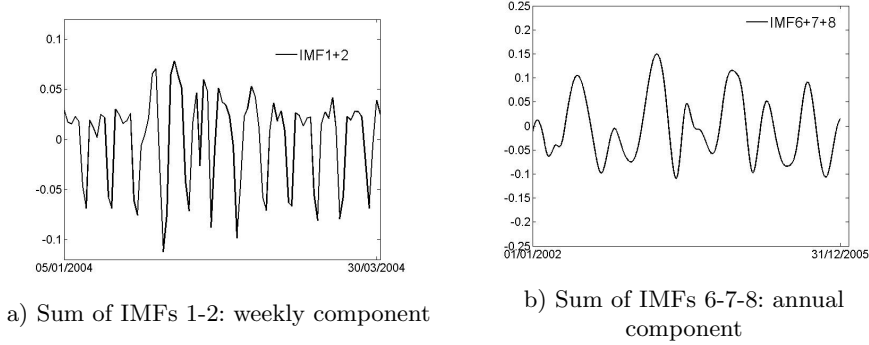
a) The logarithmical daily peak load 2000-2006      b) IMF components and the final residue or trend

**Fig. 2.** EMD of the logarithmical daily peak load 2000-2006 from STEG utility

As previously noticed by Ould Mohamed Mahmoud et al. (2009), we note that IMF 1 to 2 exhibit high frequency and can represent very short term

<sup>2</sup> For the EMD this statistic is reduced to the Mean Absolute Error  $MAE$

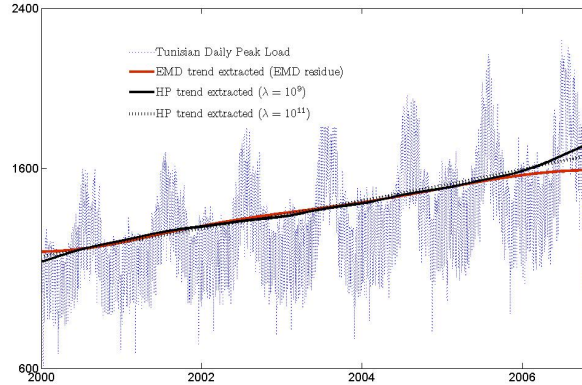
fluctuations (see Figure 3.a), IMF 3 to 5 capture small percentage of variance, indicating that such IMFs are not significant and finally IMF 6 to 8 capture mid-term effects described by seasonal variations (see Figure 3.b).



**Fig. 3.** IMFs connection to physical seasonal load components

#### 4.2 Peak load trend extraction

The trend estimate is given by the residue component of the EMD. It could represent the major trend of long term load demand which may be related to economic growth in Tunisia. The results obtained for the two methods are given in Figure 5. One can find the EMD trend and two HP trends obtained from the two bounds of the tuning parameter ( $\lambda$ ) interval evidenced for linear or quasi-linear in section 3.2.



**Fig. 4.** MA filtering, HP and EMD trends extracted of Tunisian daily peak load.

As previously mentioned, the results of trend extraction through the Hodrick-Prescott filter is very sensitive to the value of parameter  $\lambda$ . And

this fine tuning make trend extraction more difficult. On the contrary, the EMD does not require any parameter choice depending on the analyzed data.

## 5 Conclusion

Empirical Mode Decomposition appears to be an eligible method for trend extraction from seasonal time series. This finding has been illustrated through the comparison of EMD trend extracted with an improved and widely used method in economics, based on HP Filter. Since EMD-trend is very close to the optimal Hodrick Prescott trend obtained after approximation of the optimal parameter of the filter, it turns out that EMD trend extraction method does not require any optimal tuning parameter thanks to its adaptive nature.

## References

- ALEXANDROV, T., BIANCONCINI, S., BEE DAGUM, E., MAASS, P. and MCELROY, T. (2009): A Review of Some Modern Approaches to the Problem of Trend Extraction. *Research Report Series, Statistics 2008-3, U.S. Census Bureau, Washington, D.C.*
- FLANDRIN, P., GONCALVES, P. and RILLING, G. (2004): Detrending and Denoising with Empirical Mode Decomposition. *EUSIPCO 2004. September 6-10, Vienna, Austria.*
- HUANG, N.E., SHEN, Z., LONG, S.R., WU, M.C., SHIH, H.H., ZHENG, Q., YEN, N., TUNG, C.C., and LIU, H.H. (1998): The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis. *Royal Society London, 903-995.*
- OULD MOHAMED MAHMOUD, M., MHAMDI, F., JAIDANE-SAIDANE, M. (2009). Long Term Multi-Scale Analysis of the Daily Peak Load Based on the Empirical Mode Decomposition. *IEEE PowerTech, june 28-july 2, Romania.*
- POLLOCK, DSG. (2003). Sharp filters for short sequences. *Journal of Statistical Planning and Inference 113, 663-683.*
- REN, D., YANG, S., WU, Z. and YAN, G. (2006). Evaluation of the EMD end effect and a window based method to improve EMD. *International Technology and Innovation Conference, November 6-7, China.*
- SCHLICHT, E. (2005). Estimating the smoothing parameter in the so-called Hodrick-Prescott filter. *Journal of Japan Statistic Society 35, 99-119.*
- SULING, J., YANQIN, G., QIANG, W. and JIAN, Z. (2009): Trend Extraction and Similarity Matching of Financial Time Series Based on EMD Method. *World Congress on Engineering and Computer Science, San Francisco, 20-22 Oct 2009.*
- WU, Z., HUANG, N.E., LONG S.R. and PENG, C.K (2007): On the trend, detrending, and variability of nonlinear and nonstationary time series. *PNAS September 18, vol. 104, no. 38, 14889-14894.*
- ZHOU, N., TRUDNOWSKI, D., PIERRE, J.W., SARAWGI, S. and BHATT, N. (2008). An algorithm for removing trends from power-system oscillation data. *IEEE PES. July 20-24, Pittsburgh, PA.*

# The Evaluation of Non-centred Orthant Probabilities for Singular Multivariate Normal Distributions

Tetsuhisa Miwa

National Institute for Agro-Environmental Sciences  
3-1-3 Kannondai, Tsukuba 305-8604, Japan, [miwa@niaes.affrc.go.jp](mailto:miwa@niaes.affrc.go.jp)

**Abstract.** Miwa *et al.* (2003) proposed a procedure for evaluating non-centred orthant probabilities accurately for non-singular multivariate normal distributions. However, it was essential in their method that the covariance matrix should be non-singular. In this paper we consider an  $m$ -dimensional normal distribution with any singular covariance matrix of rank  $n$  ( $n < m$ ). It can be shown that the  $m$ -dimensional orthant probability is the probability volume of a polyhedron in the  $n$ -dimensional space. We show that a polyhedron can be expressed as differences between  $n$ -dimensional polyhedral cones, each of which can be evaluated by the procedure proposed by Miwa *et al.* (2003).

**Keywords:** multiple comparisons, normal distribution function, polyhedral cones, polyhedron

## 1 Introduction

Suppose that the  $m$ -dimensional random vector  $\mathbf{z} = (z_1, \dots, z_m)'$  is distributed as  $\mathbf{z} \sim N_m(\mathbf{0}, \mathbf{\Sigma})$  with density function  $\phi_m(\mathbf{z}; \mathbf{0}, \mathbf{\Sigma})$  where  $\mathbf{\Sigma}$  is any non-negative definite covariance matrix. We consider the non-centred orthant probability

$$\begin{aligned} P_m(\mathbf{c}, \mathbf{\Sigma}) &= \Pr\{z_i \geq c_i, 1 \leq i \leq m\} \\ &= \int_{c_1}^{\infty} \cdots \int_{c_m}^{\infty} \phi_m(\mathbf{z}; \mathbf{0}, \mathbf{\Sigma}) dz_1 \cdots dz_m \end{aligned} \quad (1)$$

for any constant vector  $\mathbf{c} = (c_1, \dots, c_m)'$ .

Let the rank of covariance matrix  $\mathbf{\Sigma}$  be  $n$  ( $n \leq m$ ). When the covariance matrix is non-singular ( $n = m$ ), probability (1) is called *non-singular* orthant probability. Miwa *et al.* (2003) provided an efficient and accurate procedure to evaluate a non-singular orthant probability by expressing it as differences between a finite number of orthant probabilities with tridiagonal covariance matrices. However it was essential in their method that the covariance matrix should be non-singular.

In this paper we consider any singular covariance matrix  $\mathbf{\Sigma}$ . Then the corresponding probability (1) is called *singular* orthant probability. We shall

show that any singular orthant probability can be expressed as differences of non-singular orthant probabilities, each of which can be evaluated by the procedure proposed by Miwa *et al.* (2003).

Note that the random vector  $\mathbf{z}$  would have any non-zero mean vector  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_m)'$ , which could be embedded into the constant vector  $\mathbf{c} = (c_1, \dots, c_m)'$ . Furthermore the covariance matrix  $\boldsymbol{\Sigma}$  is not changed if  $\mathbf{z}$  is replaced by  $-\mathbf{z} = (-z_1, \dots, -z_m)'$  and then the non-centred orthant probability (1) is equivalent to a multivariate normal distribution function

$$\begin{aligned} F_m(\mathbf{c}) &= \Pr\{z_i \leq c_i, 1 \leq i \leq m\} \\ &= \Pr\{-z_i \geq -c_i, 1 \leq i \leq m\} = P_m(-\mathbf{c}, \boldsymbol{\Sigma}). \end{aligned}$$

## 2 Orthant probabilities and polyhedra

Let the rank of covariance matrix  $\boldsymbol{\Sigma}$  be  $n$ . Then the  $m \times m$  non-negative definite matrix  $\boldsymbol{\Sigma}$  can be written in the form  $\boldsymbol{\Sigma} = \mathbf{A}'\mathbf{A}$  by an  $n \times m$  matrix  $\mathbf{A}$  of rank  $n$ . Denote  $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_m)$  with  $m$  column vectors  $\mathbf{a}_1, \dots, \mathbf{a}_m$  of dimension  $n$ . By expressing  $\mathbf{z} \sim N_m(\mathbf{0}, \boldsymbol{\Sigma})$  as  $\mathbf{z} = \mathbf{A}'\mathbf{x}$  with  $\mathbf{x} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ , it follows that

$$\begin{aligned} P_m(\mathbf{c}, \boldsymbol{\Sigma}) &= \Pr\{z_i \geq c_i, 1 \leq i \leq m\} \\ &= \Pr\{\mathbf{a}_i'\mathbf{x} \geq c_i, 1 \leq i \leq m\} \\ &= \Pr\{\mathbf{x} \in P\}, \end{aligned} \tag{2}$$

where  $P$  is a region in the  $n$ -dimensional space

$$P = \{\mathbf{x}: \mathbf{a}_i'\mathbf{x} \geq c_i, 1 \leq i \leq m\}. \tag{3}$$

When  $n = m$ , the region  $P$  is a polyhedral cone. If  $n < m$ , then the region  $P$  is generally a polyhedron. Therefore evaluating singular orthant probabilities reduces to the problem of evaluating the probability volumes of polyhedra in the  $n$ -dimensional space.

## 3 Expressing polyhedra as differences of polyhedral cones

### 3.1 Hyperplanes and half spaces

The polyhedron (3) is an intersection  $P = G_1 \cap \dots \cap G_m$  of half spaces  $G_i$  bounded by hyperplanes  $H_i$ , where

$$\begin{aligned} G_i &= \{\mathbf{x}: \mathbf{a}_i'\mathbf{x} \geq c_i\}, \quad 1 \leq i \leq m, \\ H_i &= \{\mathbf{x}: \mathbf{a}_i'\mathbf{x} = c_i\}, \quad 1 \leq i \leq m. \end{aligned}$$



We also define the opposite side of  $G_i$  as

$$G_i^c = \{\mathbf{x}: \mathbf{a}'_i \mathbf{x} < c_i\}, \quad 1 \leq i \leq m.$$

Since  $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_m)$  is of rank  $n$ , we assume without loss of generality that the first  $n$  vectors  $\mathbf{a}_1, \dots, \mathbf{a}_n$  are linearly independent. Then

$$Q = G_1 \cap \dots \cap G_n \quad (4)$$

is a polyhedral cone and its probability  $\Pr(Q)$  is easily calculated by the procedure given by Miwa *et al.* (2003).

### 3.2 Lemma

Consider another hyperplane and its corresponding half space

$$\begin{aligned} H_{n+1} &= \{\mathbf{x}: \mathbf{a}'_{n+1} \mathbf{x} = c_{n+1}\}, \\ G_{n+1} &= \{\mathbf{x}: \mathbf{a}'_{n+1} \mathbf{x} \geq c_{n+1}\} \end{aligned}$$

for any vector  $\mathbf{a}_{n+1}$  and any constant  $c_{n+1}$ . Then the region

$$Q \cap G_{n+1} = G_1 \cap \dots \cap G_n \cap G_{n+1}$$

forms a polyhedron, and its probability  $\Pr(Q \cap G_{n+1})$  is a singular orthant probability. If  $Q \cap G_{n+1}$  has  $s$  vertices ( $1 \leq s \leq n+1$ ), then its probability is expressed as a linear combination of cone probabilities:

$$\Pr(Q \cap G_{n+1}) = \sum_{j=1}^s \pm \Pr(Q_j), \quad (5)$$

where each  $Q_j$  is an  $n$ -dimensional polyhedral cone.

#### Proof of lemma

Let  $\mathbf{A}_* = (\mathbf{a}_1, \dots, \mathbf{a}_n)$ , and define  $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_n) = (\mathbf{A}'_*)^{-1}$  such that  $\mathbf{a}'_i \mathbf{b}_i = 1$  and  $\mathbf{a}'_i \mathbf{b}_j = 0$  for  $i \neq j$ . The vertex  $\mathbf{v}_0$  of the cone  $Q$  lies on all the hyperplanes  $H_i$  ( $1 \leq i \leq n$ ) and is found to be  $\mathbf{v}_0 = (\mathbf{A}'_*)^{-1} \mathbf{c}_* = \mathbf{B} \mathbf{c}_*$ , where  $\mathbf{c}_* = (c_1, \dots, c_n)'$ . The polyhedral cone  $Q$  can also be expressed in terms of the linearly independent vectors  $\mathbf{b}_1, \dots, \mathbf{b}_n$  as

$$Q = \{\mathbf{x}: \mathbf{x} = \mathbf{v}_0 + \lambda_1 \mathbf{b}_1 + \dots + \lambda_n \mathbf{b}_n, \lambda_i \geq 0\}. \quad (6)$$

We can consider three cases corresponding to the positions of vertex  $\mathbf{v}_0$  with respect to the hyperplane  $H_{n+1}$ .

Case 1:  $\mathbf{a}'_{n+1} \mathbf{v}_0 < c_{n+1}$  ( $\mathbf{v}_0 \in G_{n+1}^c$ ).

It follows from (6) that if  $\mathbf{a}'_{n+1} \mathbf{b}_j \leq 0$  for all  $1 \leq j \leq n$ , then  $Q \cap G_{n+1} = \emptyset$ . We suppose

$$\begin{aligned} \mathbf{a}'_{n+1} \mathbf{b}_j &> 0, \quad 1 \leq j \leq s, \\ \mathbf{a}'_{n+1} \mathbf{b}_j &\leq 0, \quad s < j \leq n. \end{aligned}$$

Then  $Q \cap G_{n+1}$  has  $s$  vertices

$$\mathbf{v}_j = \mathbf{v}_0 + \frac{c_{n+1} - \mathbf{a}'_{n+1} \mathbf{v}_0}{\mathbf{a}'_{n+1} \mathbf{b}_j} \mathbf{b}_j, \quad 1 \leq j \leq s.$$

It is easily seen that  $\mathbf{v}_j$  is contained in  $H_1, \dots, H_{j-1}, H_{j+1}, \dots, H_n, H_{n+1}$ , and we consider a polyhedral cone

$$Q_j = G_1^c \cap \dots \cap G_{j-1}^c \cap G_{j+1} \cap \dots \cap G_n \cap G_{n+1}.$$

with vertex  $\mathbf{v}_j$ . Then the target probability is expressed as

$$\Pr(Q \cap G_{n+1}) = \sum_{j=1}^s (-1)^{j-1} \Pr(Q_j). \quad (7)$$

This expression can be proved by starting with  $Q_s$  and proceeding backwards to  $Q_1$ . First note that  $Q_s \subset G_s$ , because if  $\mathbf{x} \in G_1^c \cap \dots \cap G_{s-1}^c \cap G_{s+1} \cap \dots \cap G_n$  and  $\mathbf{x} \in G_s^c$  then  $\mathbf{x} \notin G_{n+1}$ . Thus we have

$$Q_s = Q_s \cap G_s = G_1^c \cap \dots \cap G_{s-1}^c \cap G_s \cap G_{s+1} \cap \dots \cap G_n \cap G_{n+1} \subset Q_{s-1}$$

so that

$$Q_{s-1} \setminus Q_s = G_1^c \cap \dots \cap G_{s-2}^c \cap G_{s-1} \cap G_s \cap G_{s+1} \cap \dots \cap G_n \cap G_{n+1} \subset Q_{s-2}.$$

It follows that in general

$$\begin{aligned} & Q_j \setminus (Q_{j+1} \setminus \dots \setminus (Q_{s-1} \setminus Q_s)) \\ &= G_1^c \cap \dots \cap G_{j-1}^c \cap G_j \cap \dots \cap G_s \cap G_{s+1} \cap \dots \cap G_n \cap G_{n+1} \subset Q_{j-1} \end{aligned}$$

and finally we obtain

$$Q_1 \setminus (Q_2 \setminus \dots \setminus (Q_{s-1} \setminus Q_s)) = G_1 \cap \dots \cap G_n \cap G_{n+1} = Q \cap G_{n+1}$$

which gives

$$\Pr(Q \cap G_{n+1}) = \Pr(Q_1) - \{\Pr(Q_2) - \{\dots - \{\Pr(Q_{s-1}) - \Pr(Q_s)\}\}\}.$$

Case 2:  $\mathbf{a}'_{n+1} \mathbf{v}_0 = c_{n+1}$  ( $\mathbf{v}_0 \in H_{n+1}$ ).

If  $\mathbf{a}'_{n+1} \mathbf{b}_j \geq 0$  for all  $1 \leq j \leq n$ , then  $G_{n+1}$  is redundant, that is  $Q \subset G_{n+1}$  and  $Q \cap G_{n+1} = Q$ . On the other hand, if  $\mathbf{a}'_{n+1} \mathbf{b}_j \leq 0$  for all  $1 \leq j \leq n$ , then  $Q \cap G_{n+1} \subset H_{n+1}$  and  $\Pr(Q \cap G_{n+1}) = 0$ . Suppose  $\mathbf{a}'_{n+1} \mathbf{b}_j > 0$  for  $1 \leq j \leq s < n$ . Then we can show that the same decomposition (7) holds. In this case all the vertices in the above discussion are equal,

$$\mathbf{v}_1 = \dots = \mathbf{v}_s = \mathbf{v}_0$$

and the region  $Q \cap G_{n+1}$  is still a cone. We might define that the vertices are degenerated into  $s$ -th order.

Case 3:  $\mathbf{a}'_{n+1} \mathbf{v}_0 > c_{n+1}$  ( $\mathbf{v}_0 \in G_{n+1}$  and  $\mathbf{v}_0 \notin H_{n+1}$ ).

Since  $\Pr(Q \cap G_{n+1}) = \Pr(Q) - \Pr(Q \cap G_{n+1}^c)$ , we can apply the same discussion of case 1 to  $G_{n+1}^c = \{\mathbf{x}: \mathbf{a}'_{n+1} \mathbf{x} < c_{n+1}\} = \{\mathbf{x}: -\mathbf{a}'_{n+1} \mathbf{x} > -c_{n+1}\}$ .

This completes the proof of the lemma.

### 3.3 Corollary

Now that we can express  $\Pr(Q \cap G_{n+1})$  as differences of cone probabilities

$$\Pr(Q \cap G_{n+1}) = \sum_{j=1}^s \pm \Pr(Q_j),$$

we can add any number of half spaces. For example we have

$$\Pr(Q \cap G_{n+1} \cap G_{n+2}) = \sum_{j=1}^s \pm \Pr(Q_j \cap G_{n+2}), \quad (8)$$

and each  $\Pr(Q_j \cap G_{n+2})$  is expressed as differences of cone probabilities.

## 4 Applications to multiple comparison procedures

Consider the one-way analysis of variance model

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad 1 \leq i \leq k, \quad 1 \leq j \leq n_i,$$

where  $\epsilon_{ij}$  are independent  $N(0, \sigma^2)$  random variables. Let  $\bar{y}_i$ ,  $1 \leq i \leq k$ , be the  $i$ -th sample mean based on  $n_i$  observations, and let  $\hat{\sigma}^2$  be an unbiased estimator of  $\sigma^2$  distributed independently of  $\bar{y}_i$  as  $\hat{\sigma}^2 \sim \sigma^2 \chi^2(\nu)/\nu$  for some degrees of freedom  $\nu$ .

Suppose that we are interested in testing the null hypothesis

$$H_0: \mu_1 = \cdots = \mu_k$$

against the simply ordered alternative hypothesis

$$H_A: \mu_1 \leq \cdots \leq \mu_k$$

with at least one strict inequality.

One of the most frequently used multiple comparison procedures is the multiple contrast test, where the test statistic is the maximum value among

a set of contrasts of sample means. Hayter (1990) considered a set of ordered pairwise comparisons

$$z^{(ij)} = (-\bar{y}_i + \bar{y}_j) / \hat{\sigma} \sqrt{1/n_i + 1/n_j}, \quad 1 \leq i < j \leq k.$$

Marcus (1976) proposed the following set of contrasts

$$z^{(ij)} = \{-Y(1, i) + Y(j, k)\} / \hat{\sigma} \sqrt{2}, \quad 1 \leq i < j \leq k$$

$$Y(p, q) = \sum_{i=p}^q n_i \bar{y}_i / W(p, q), \quad W(p, q) = \sum_{i=p}^q n_i$$

in her modified Williams test. Miwa (1998) considered a set of standardised contrasts

$$z^{(ij)} = \frac{-Y(1, i) + Y(j, k)}{\hat{\sigma} \sqrt{1/W(1, i) + 1/W(j, k)}} \quad 1 \leq i < j \leq k.$$

Some simulation studies showed that these tests are powerful. However the contrasts involved in these multiple comparison tests are linearly dependent, and there has been no accurate method available for evaluating the distributions of test statistics in unbalanced cases. Since the algorithm presented in this paper copes with any singular covariance structure, we can apply it to these powerful multiple comparison procedures.

## 5 Discussions

Another possible approach is a Monte Carlo method (e.g. Genz and Bretz, 2002). Miwa *et al.* (2003) showed that the Monte Carlo method is faster for calculating non-singular orthant probabilities in low accuracy. However it is difficult for the Monte Carlo method to achieve high accuracy. The same results are expected to apply to singular orthant probabilities. The computer programs are being developed by the author.

## References

- GENZ, A. and BRETZ, F. (2002): Comparison of methods for the computation of multivariate  $t$  probabilities. *Journal of Computational and Graphical Statistics*, 11, 950-971.
- HAYTER, A. J. (1990): A one-sided studentized range test for testing against a simple ordered alternative. *Journal of American Statistical Association*, 85, 778-785.
- MARCUS, R. (1976): The powers of some tests of equality of normal means against an ordered alternative. *Biometrika*, 63, 177-183.
- MIWA, T. (1998): Bartholomew's test as a multiple contrast test and its applications. *Japanese Journal of Biometrics*, 19, 1-9.
- MIWA, T., HAYTER, A. J. and KURIKI, S. (2003): The evaluation of general non-centred orthant probabilities. *Journal of the Royal Statistical Society, Ser. B*, 65, 223-234.

# Variable Inclusion and Shrinkage Algorithm in High Dimension

Abdallah Mkhadri<sup>1</sup> and Mohamed Ouhourane<sup>2</sup>

<sup>1</sup> Department of Mathematics, Faculty of Sciences-Semlalia,  
B.P. 2390, Marrakech, Morocco. *mkhadri@ucam.ac.ma*

<sup>2</sup> Department of Mathematics, Faculty of Sciences-Semlalia,  
B.P. 2390, Marrakech, Morocco. *hourali@hotmail.com*

**Abstract.** We propose a new method to simultaneously select variables and encourage a grouping effect where strongly correlated predictors tend to be in or out of the model together. Moreover, our method is capable of selecting sparse models while avoiding over shrinkage of Lasso. It combines the idea of VISA algorithm which avoids over shrinkage of regression coefficients and those of the Elastic net which overcomes the limitation of Lasso in high dimension. Our method is based on a modified VISA algorithm, so is also computationally efficient. A detailed simulation study in small and high dimensional settings is performed, which illustrates the advantages of our approach in relation to several other possible methods.

**Keywords:** variable selection, VISA algorithm, LARS, linear regression

## 1 Introduction

Suppose that the data set has  $n$  observations with  $p$  predictors. We consider the standard linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where  $\mathbf{y} \in \mathbb{R}^n$  is the response and  $\mathbf{X}$  is the  $n \times p$  model matrix, with  $\mathbf{x}_j \in \mathbb{R}^n, j = 1, \dots, p$ , are the predictors,  $\boldsymbol{\beta}$  is a  $p$ -vector of unknown parameters which are to be estimated,  $t$  stands for the transpose and  $\boldsymbol{\varepsilon}$  is a  $n$ -vector of (i.i.d.) random errors with mean 0 and variance  $\sigma^2$ . It is assumed that the response is centered and the predictors are standardized. When  $p$  is large relative to  $n$ , there are many alternative procedures that outperform the ordinary least squares (OLS) which can be categorized into one of the two groups : regularization methods (like Ridge regression) and classical variable selection. But, the final fit of Ridge regression is difficult to interpret because all  $p$  predictors will remain in the model. While, the classical variable selection is computationally heavy to implement when  $p$  is large.

More recently interest has focused on an alternative class of methods which implement both the variable selection and coefficient shrinkage in a

single procedure. The Lasso (Tibshirani (1996)) is a popular one for regression that uses the  $\ell_1$  norm to achieve a sparse solution. The use of  $\ell_1$  penalty on the coefficients has the effect of automatically performing variable selection by setting certain coefficients to zero and shrinking the remainder. This method was made particularly appealing by the advent of the efficient LARS algorithm (Efron et al. (2004)) which computes the entire regularization path for Lasso.

However, one limitation of Lasso is that the same tuning parameter is used for both variable selection and shrinkage. It leads to selection of a final model with too many predictors to prevent over shrinkage of the regression components. An alternative is to use the hard thresholding approach enforced by the Relaxed Lasso (Meinshausen (2007)) which introduces a second tuning parameter on the  $\ell_1$  penalty. However, only variables included according to the first tuning parameter may enter the model. The second alternative is VISA (Radchenko and James (2008)) which instead allows for the potential inclusion of all variables. The first parameter in VISA divides the variables into two groups. The first one receives preference for model inclusion but variables from the second group may still be included if there is evidence that they are significant. The second limitation of Lasso is that when  $p \gg n$ , Lasso selects at most  $n$  variables. Moreover, the Lasso tends to select only one variable from the group of variables among which the pairwise correlations are very high. Using the combination of the  $\ell_1$  and  $\ell_2$  penalties, Zou and Hastie (2005) proposed the Elastic net which overcomes the latter two limitations.

Since VISA is based on the LARS algorithm, it inherits the latter two limitations of Lasso in high dimension. On the other hand, the Elastic net inherits the first limitation of Lasso of over shrinkage of regression coefficients. To overcome these problems, we propose an algorithm which is based on the combination of the ideas of VISA and Elastic net, we called VISA-Net. VISA-Net is based on the modification of the first step of VISA where some augmented data is used instead of initial data  $(\mathbf{y}, \mathbf{X})$ . Based on our empirical results, this modification produces considerable improvements over VISA and the latter thresholding approaches.

The paper is organized as follows. In Section 2, we review the variable selection problem in linear regression and present VISA and different shrinkage hard thresholding algorithms. In Section 3, we define the VISA-Net algorithm and discuss its computational aspects. A detailed simulation study is performed in Section 4, which illustrates the performance of VISA-Net in relation to its competitors. We end the paper with a brief conclusion.

## 2 Regularization methods

Let's review the principal regularization methods for linear models.

## 2.1 Lasso

The Lasso (least absolute shrinkage and selection operator) estimator is defined by solving the  $l_1$  penalized least squares problem

$$\hat{\boldsymbol{\beta}}_{Lasso} = \operatorname{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1. \quad (2)$$

The  $l_1$  penalty enables to reduce the variability of the estimates by shrinking the coefficients and at the same time produces interpretable models by shrinking some coefficients to exactly zero. This method was made particularly appealing by the advent of the LARS algorithm (Efron et al., (2004)) which provided a highly efficient means to simultaneously produce the set of Lasso fits for all values of the tuning parameter. The limitations of the Lasso are:

- In high dimension, the Lasso selects at most  $n$  variables.
- In high correlation, the Lasso tends to select only one variable from the group of variables among which the pairwise correlations are very high.
- The same tuning parameter is used for both variable selection and shrinkage, so it leads to selection of a final model with too many predictors to prevent over shrinkage of the regression coefficients.

## 2.2 Elastic net

The Elastic net is proposed to overcome the two first limitations of the Lasso in some situations. The Elastic net is based on a penalized least squares with a penalty which is a combination of the  $\ell_1$  and  $\ell_2$  norms. The  $\ell_1$  penalty is a Lasso-type thresholding that performs variable selection thus inducing a sparse model. The quadratic penalty, related to Ridge regression, encourages a grouping effect and places no limitation on the number of variables that may be selected for the model. However, the Elastic net has some limitations which are:

- Since it is based on LARS algorithm, it inherits the same third limitation of over shrinking of regression coefficients which leads to inclusion of a number of irrelevant variables.
- In some cases of correlations, some significant predictors not may enter in the final model.

## 2.3 VISA

VISA (Variables Inclusion Shrinkage Algorithms), proposed by Rodchenko and James (2008), is an alternative to Lasso which allows reducing the shrinkage of coefficients and finding other ignored variables in the first stage of Lasso. Thus, this algorithm is composed of three stages: based on a fixed

value of the tuning parameter  $\lambda$ , the first stage allows to select a sparse Lasso solution  $\hat{\beta}_\lambda(0)$  that satisfies

$$\|\mathbf{X}^t(\mathbf{y} - \mathbf{X}\hat{\beta})\| \leq \lambda \quad (3)$$

In the second stage, the variables with maximum importance, in term of correlation with the residual, are selected as the "primary variables". A path  $\beta_\lambda(s)$  is constructed that systematically drives their correlations to zero, while maintaining (3) for the remaining variables. Finally, once the primary covariances have reached zero, the third stage identifies a secondary set of variables whose correlations (with the residual) are now at the boundary  $\lambda$ . Since, they have large covariances, VISA path drives their correlations towards zero, while maintaining the primary covariances at zero. However, since VISA is based on LARS algorithm, it inherits the first two limitations of Lasso.

In what follows, we propose VISA-Net algorithm which combines the ideas of VISA and Elastic net methods. It allows us to select the groups of significant variables highly correlated and to find other significant variables which may be ignored by the Elastic net procedure.

### 3 VISA-Net algorithm

We propose VISA-Net algorithm for selection of variables in linear regression problem which combines the ideas of VISA and Elastic net algorithms. VISA uses the LARS algorithm with modifications of some of its steps. And yet, we know that Elastic net is but LARS using augmented data. So we can use the VISA algorithm with augmented data which will allow us to define VISA-Net algorithm which gathers the advantages of both algorithms. The VISA-Net algorithm consists of the following steps.

- Given data set  $(\mathbf{y}, \mathbf{X})$  and a grid of values of  $\lambda_2$ , define an augmented data set  $(\mathbf{y}^*, \mathbf{X}^*)$  by

$$\mathbf{X}_{(n+p) \times n}^* = (1 + \lambda_2) \left( \frac{\mathbf{X}}{\sqrt{\lambda_2} \mathbf{I}} \right), \mathbf{y}_{(n+p)}^* = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix}. \quad (4)$$

- Apply  $\text{VISA}_{lars}$  algorithm on augmented data  $(\mathbf{y}, \mathbf{X}^*)$ .
- The VISA-Net estimator is defined to be the  $\text{VISA}(\mathbf{y}^*, \mathbf{X}^*)_{lars}$  solution multiplied by the factor  $(1 + \lambda_2)^{1/2}$ .

The first step of this algorithm is exactly the reformulation of the Elastic net problem as Lasso problem based on augmented data  $(\mathbf{y}^*, \mathbf{X}^*)$ .

The estimation of the parameters of the model by VISA-Net needs tree regularization parameters. The first parameter  $\lambda_2$  gets its value in a grid of



values like the one used by Elastic net. The optimal value of  $\lambda_2$  is the one which produces the smallest prediction error by fold-cross validation. The second parameter allows selecting an initial sparse model  $\beta_\lambda(0)$  and a set of variables which should extend their correlations with the residual to zero. The third parameter allows eliminating the shrinkage in the last set of variables (cf. Radchenko and James 2008).

## 4 Numerical experiments

In this section we present a detailed simulation study comparing VISA-Net to tree competing methods (Lasso, Elastic net and VISA). We consider two dimensional setting: small to moderate ( $p \leq n$ ) and high dimensional ( $p > n$ ).

### 4.1 Simulation study in the case $p \leq n$

The simulation setting is similar to those used in the original paper of the Elastic net paper (Zou and Hastie (2005)). For each example, the data are simulated from the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n). \quad (5)$$

Exemple	Statistic	LASSO	ENET	VISA	VNET
1	MSE <sub>y</sub>	3.450	3.036	2.918	2.050
	<i>MSE</i> <sub><math>\beta</math></sub>	11.410	11.467	13.078	11.509
2	MSE <sub>y</sub>	4.196	3.257	4.112	2.812
	<i>MSE</i> <sub><math>\beta</math></sub>	63.452	59.722	63.789	55.802
3	MSE <sub>y</sub>	0.827	0.678	0.771	0.480
	<i>MSE</i> <sub><math>\beta</math></sub>	5.918	5.938	5.661	4.835
4	MSE <sub>y</sub>	48.525	34.599	47.833	21.086
	<i>MSE</i> <sub><math>\beta</math></sub>	81.598	56.108	85.739	21.573

**Table 1.** Mean squared errors for the simulated examples of four methods based on 100 replications.

For each example, 100 data sets were generated. Each data set consisted of a training set of size  $n$ , on which the model was fitted, an independent validation set of size  $n$  is used to select the tuning parameters and a test set is used for the evaluation of the performance. The four sample scenarios are:

- In Example one,  $n = 20$  and there are  $p = 8$  predictors. The true parameters are  $\boldsymbol{\beta} = (3; 1; 5; 0; 0; 2; 0; 0)^t$  and  $\sigma = 3$ . with the correlation matrix given by  $\rho(x_i; x_j) = 0.7^{|i-j|}$ . This example contains only positively correlated variables.
- The example 2 consists of  $p = 9$ ,  $\boldsymbol{\beta}$  is specified by  $\boldsymbol{\beta} = (1; 2; 3; 4; 0; 1; 2; 3; 4)^t$ ,  $\sigma = 3$  and  $\rho(x_i; x_j) = 1 - 0.25|i - j|$ . The same sample size is as in (1). In this example variables are positively and negatively correlated.

- Example 3 is the same as Example 1, except that  $\beta_j = 0 : 85$  for all  $j$ , creating a non-sparse underlying model.
- In Example 4,  $n = 100$  for each of the training and validation sets and there are 40 predictors. The true parameters are

$$\mathbf{beta} = (\underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10}, \underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10}) \quad (6)$$

Table 1 summarizes mean squared error of the estimation for the response  $y$  ( $\text{MSE}_y$ ) and the mean squared error for the estimation of  $\beta$  ( $\text{MSE}_\beta$ ). It can be seen that VISA-Net performs well than all other methods in terms of  $\text{MSE}_y$  and  $\text{MSE}_\beta$ . Except in Example 1 where Elastic net is slightly better than VISA-Net in term of  $\text{MSE}_\beta$ . Moreover, in the sparse example 4, the difference between VISA-Net and other methods is very high.

## 4.2 High-dimensional experiments

In this section, we give the differences between our VISA-Net, the Lasso, the Elastic net and VISA through simulations data in high dimensional setting. The simulation setting is similar to those used for VISA in Radchenko and James (2008). Our simulations contained tree parameters that we altered. Namely, the number of variables (50 or 70), the number of observations (50 or 100) and the correlations among the columns in the design matrix (0, 0.5 or 0.95). For each method and simulation we compute the average prediction error of on a test data set and the mean squared error of  $\beta$ . The true parameter vector is  $\beta = (\underbrace{1, \dots, 1}_{10}, \underbrace{0, \dots, 0}_{p-10})$ . The results are summarized in Table.2.

From Table 1 and Table 2 we can conclude that:

- VISA-Net is more efficient than the other methods given that the Mean prediction error and the mean squared error of  $\beta$  are smaller compared to other methods.
- In the absence of correlation, VISA is more efficient than Lasso and Elastic net which are equal. On the contrary, in the presence of correlation, VISA loses its efficiency compared to Elastic net which allows a higher efficiency than Lasso.
- VISA-Net is a refinement of VISA in case of high correlations and a refinement of Elastic net in case of low correlations.

Exemple	Statistic	LASSO	ENET	VISA	VNET
50 var 100 obs cor 0	MSE <sub>y</sub> $MSE\beta$	0.077 0.023	0.077 0.023	0.097 0.018	0.097 0.008
50 var 100 obs cor 0.5	MSE <sub>y</sub> $MSE\beta$	0.178 0.045	0.164 0.033	0.190 0.046	0.137 0.011
50 var 100 obs cor 0.95	MSE <sub>y</sub> $MSE\beta$	0.462 0.031	0.224 0.016	0.464 0.028	0.167 0.005
50 var 50 obs cor 0	MSE <sub>y</sub> $MSE\beta$	1.081 1.005	1.082 1.006	0.537 0.507	0.491 0.485
50 var 50 obs cor 0.5	MSE <sub>y</sub> $MSE\beta$	0.801 1.447	0.784 1.446	0.718 1.335	0.671 1.264
50 var 50 obs cor 0.95	MSE <sub>y</sub> $MSE\beta$	0.458 8.168	0.360 6.203	0.503 9.305	0.332 6.029
70 var 50 obs cor 0	MSE <sub>y</sub> $MSE\beta$	1.975 1.915	1.983 1.919	1.236 1.199	0.830 1.842
70 var 50 obs cor 0.5	MSE <sub>y</sub> $MSE\beta$	1.015 1.840	1.050 1.896	0.923 1.712	0.854 1.651
70 var 50 obs cor 0.95	MSE <sub>y</sub> $MSE\beta$	0.515 9.483	0.395 7.061	0.574 10.836	0.363 6.743

**Table 2.** Mean squared errors for the simulated examples of four methods based on 100 replications.

## 5 Conclusions

In this paper, we have proposed a method allowing simultaneously the selection and the estimation of a model under linear regression. It meets the advantages of both methods VISA and Elastic net. VISA is a method with two stages which allows eliminating the shrinkage of selected coefficients and finding others ignored by LARS. Elastic net favors the group effect and does not have any constraint of the size of the sample. Our method allows on one hand the encouragement of the group effect and on the other hand it allows eliminating the shrinkage and finding other groups which can be added to the model. Our empirical results confirm the efficiency of our method compared to other hard thresholding approaches. Finally, some theoretical justification of VISA-Net in terms of non-asymptotics bounds on the estimation error is not considered in this note and will be subject to further work.

## References

- EFRON, B. HASTIE, T. JOHNSTONE, I. and TIBSHIRANI, R. (2004): Least angle regression. *Annals of Statistics*, **32**, 407-499.  
 MEINSHAUSEN, N. (2007): Relaxed lasso. *Computational Statistics & Data Analysis*, **52**, 374-393.

- RADCHENKO, P. and JAMES, G. M.(2008): Variable inclusion and shrinkage algorithms. *Journal of the American Statistical Association*, vol 103, n 483, 1304-1315.
- TIBSHIRANI, R.(1996): Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, series B*, **58**, 267-288.
- ZOU, H. and HASTIE, T. (2005): Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, series B*, **67**, 301-320.

# Application of a Bayesian Approach for Analysing Disease Mapping Data: Modelling Spatially Correlated Small Area Counts

Mohammadreza Mohebbi<sup>1</sup> and Rory Wolfe<sup>1</sup>

1. Department of Epidemiology and Preventive Medicine, Faculty of Medicine,  
Nursing and Health Sciences, Monash University, Melbourne, Australia,  
*Mohammadreza.Mohebbi@med.monash.edu.au*

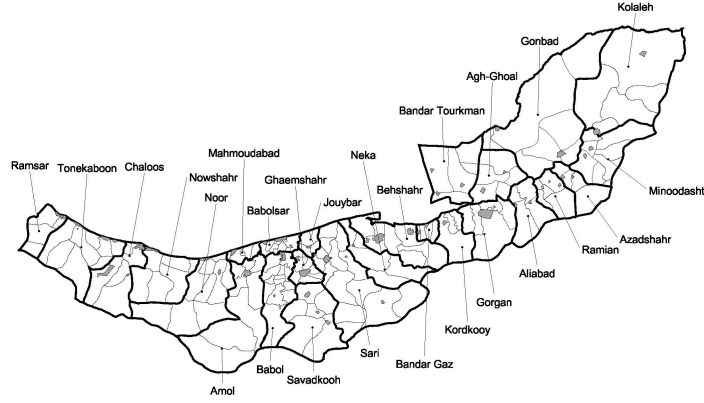
**Abstract.** We used a full Bayesian approach to estimate region specific geographically correlated disease rates. We performed three-stage hierarchical models in which disease counts were modelled as a function of area-specific relative risks at stage one; the collection of relative risks across the study region were modelled at stage two; and at stage three prior distributions were assigned to parameters of the stage two distribution. To illustrate the procedures, we present an analysis of esophageal cancer incidence in the Caspian region of Iran. The evidence suggested that models based on the use of spatial random effects work well and provide a robust basis for model inference.

**Keywords:** Bayesian inference, disease mapping, ecologic regression, Poisson regression, spatial correlation

## 1 Introduction

The analysis of disease rates from small areas often involves a trade-off between statistical stability of estimates and geographic precision. Detection of locally elevated risk or rates requires geographically small units to distinguish local risk/rates from area-wide values. On the other hand, smaller regions result in rate estimates based on smaller population sizes. For a rare disease, small population sizes result in particularly unstable rate estimates. The statistical literature contains various methods of combining information between regions to achieve local rate stabilization without losing geographic resolution. The most common approaches involve hierarchical models with random effects for each region.

Clayton and Kaldor (1987) introduce hierarchical models and associated empirical Bayesian inference for region-specific standardized mortality ratios which allow spatial correlation between neighbouring regions. Besag et al. (1991) extend these to a fully Bayesian setting using Markov chain Monte Carlo algorithms (MCMC) and Mollie (1996) provides a thorough introduction to the fully Bayesian approach. The plan of the paper is as follows. In the next section we describe model parameterisation for the Poisson model, prior distributions specification on the parameters



**Fig. 1.** Geographic boundaries of cities and rural agglomerations within wards in Mazandaran and Golestan provinces.

as well as give some details for monitoring the MCMC simulation. Section 3 summarizes the results from the analysis of the Caspian EC data. We conclude with a brief discussion in Section 4.

## 2 Method

### Data structure

The raw data are in the form of disease counts,  $Y_j$ , and population counts,  $N_j$ , where  $j=1, \dots, J$ , indexes geographical areas. For rare and non-infectious diseases we may then assume

$$Y_j | E_j, \psi_j \sim \text{Poisson}(E_j \psi_j) \quad (1)$$

Where  $E_j$  denote the expected number and  $\psi_j$  represents the relative risk of cases in area  $j$ .

### Hierarchical models for relative risks

In this section we describe a three-stage hierarchical model which may be used to analyse disease mapping data; Mollie (1996) contains further details. We begin by assuming that the first stage model given in equation (1) is appropriate, and return to the assessment of proportionality in the next section. The maximum likelihood estimator (MLE) of the relative risks from model (1) corresponds to a standardized mortality/incidence ratio (SMR/SIR):  $\frac{Y_j}{E_j}$ . The variance of this estimator is proportional to  $E_j^{-1}$  and so for areas with small populations there will be high sampling variability. Thus, for example, if the SMRs are mapped, large rural areas with low populations will often appear to display high risk due to the high variability of these estimates. The mapping of significance levels in order to overcome this problem produces its own difficulties since areas with large populations are more likely

to attain significance, even if the excess risk is small (Clayton DG, Kaldor J. 1987). In this research expected counts are thought of as fixed and proportional to the known population. Another difficulty with the use of SMRs for inference is that, for small areas in particular, SMRs in areas that are geographically close tend to display positive dependence, that is, positive spatial autocorrelation. If this aspect is ignored, incorrect inference will result; in particular, standard errors of ecological regression coefficients will be too small. To overcome these problems, Besag et al. (1991) suggested combining (1) with the following model for the relative risks:

$$\log \psi_j = X_j \beta_j^T + \theta_j + \phi_j \quad (2)$$

where  $X_j^T = (1, X_{j1}, \dots, X_{jk})^T$  is a  $(k+1) \times 1$  vector of area-level risk factors,  $\beta_j = (\beta_0, \beta_1, \dots, \beta_k)^T$  is a  $(k+1) \times 1$  vector of regression parameters with  $e^{(\beta_l)}$  representing the relative risk due to risk factor  $X_l$ ,  $l=(0, 1, \dots, k)$ ,  $\theta_j$ ,  $j=1, \dots, J$  represents a residual with no spatial structure (so that  $\theta_i$  and  $\theta_j$  are independent for  $i \neq j$ ), and  $\phi_j$ ,  $j=1, \dots, J$  represents a residual with spatial structure (so that  $\phi_i$  and  $\phi_j$  are modelled to have positive spatial dependence).

Model (2) forms the second stage of the hierarchical model. We let  $\theta = (\theta_1, \dots, \theta_J)^T$  and  $\phi = (\phi_1, \dots, \phi_J)^T$  denote the hyperparameters, that is, the parameters of the variance-covariance matrices of the distributions of  $\theta$  and  $\phi$ . As discussed above, unmeasured risk factors, artificial areal-stratum boundaries and data inaccuracies may lead to the Poisson model (1) being inadequate. In particular we typically find that  $\text{var}(Y) > E(Y)$ , that is, the area-specific disease counts exhibit overdispersion. Model (2) allows the spatial and non-spatial modelling of this overdispersion via the random effects  $\phi_j$  and  $\theta_j$ , respectively. In order to compare and select the most appropriate model we used the Deviance Information Criterion (DIC). If we believed that non-spatial overdispersion only needed to be considered, then it is natural to model  $e^{\theta_j}$  as arising from a gamma distribution as this leads to a tractable marginal distribution (the negative binomial).

### Prior specification

The gamma distribution cannot easily be extended to allow positive spatial dependence. However, the normal distribution does allow such an extension and is the common choice for both spatial and unstructured (non-spatial) random effects. The unstructured components are usually modeled independently as  $\theta_j \sim_{iid} N(0, \frac{1}{\tau_{\theta_j}})$  describing the non-spatial heterogeneity, we also assume normal distribution priors for regression parameters, e.g.,  $\beta_k \sim_{iid} N(0, \frac{1}{\tau_{\beta_k}})$ .

For the spatially-dependent random effects in (2), the problem is to model the  $J$ -dimensional random variable  $\phi$  allowing for dependence between  $\phi_i$  and  $\phi_j$ ,  $i \neq j$ . In the joint modelling approach a common model for  $\phi$  is the zero mean multivariate normal distribution  $N_I(0, \sigma_\phi^2 \Sigma(\nu))$ . The  $J \times J$  positive definite correlation matrix  $\Sigma(\nu)$  contains elements  $\Sigma_{ij}(\nu)$ ,  $i, j=1, \dots, J$  with diagonal elements equal to one and off-diagonal elements describing the correlation between  $\phi_i$  and  $\phi_j$ ,  $i \neq j$ . Various structured forms may be assumed for  $\Sigma(\nu)$ . Cressie (1993) and Wackernagel (1998) contain a discussion of more general forms for the correlation.

In the conditional approach, the intrinsic conditional autoregressive (CAR) Markov random field prior suggested by Besag et al. (1991) has commonly been used. In this

model the distribution of each  $\phi_j$  given all the other elements  $\{\phi_1, \dots, \phi_{j-1}, \phi_{j+1}, \dots, \phi_J\}$  depends only on its neighbourhood (Cressie, 1993). A commonly used form for this is the Gaussian whereby the conditional distribution of each  $\phi_j$  given by

$$\phi_j | \phi_{-j} \sim N\left(\frac{\sum_{-j} w_{ij} \phi_j}{\sum_{-j} w_{ij}}, \frac{1}{\tau_\phi \sum_{-j} w_{ij}}\right) \quad (3)$$

where  $\phi_{-j}$  represents the set of residuals for areas which neighbour area  $j$ . One important restriction in this specification is that the matrix of weights  $W$  must be symmetric. It should be noted that the specification of this CAR structure leads to a prior joint distribution for the relative risks given by (3).

As it is based on paired differences between the  $\phi_j$ 's this prior is improper. In practice, a sum to zero constraint is imposed on these random effects in order to guarantee identifiability. Although other possibilities exist, the simplest and most commonly used neighbourhood structure is defined by the existence of a common border of any length between the areas. In this case, the weights  $w_{ij}$  are constants and specified as  $w_{ij} = 1$  if  $i$  and  $j$  are adjacent and  $w_{ij} = 0$  otherwise. The conditional prior mean of  $\phi_j$  is given by the arithmetic average of the spatial effects from its neighbours and the conditional prior variance is proportional to the number of neighbours.

One great advantage of the conditional model is that it is very computationally efficient due to the conditional independencies that may be exploited in Markov chain Monte Carlo (MCMC) estimation approaches (Smith and Roberts (1993)).

### Fully Bayesian estimation

At the third stage of the model, the Bayesian approach that we follow requires specification of prior distributions for the second-stage parameters  $\phi$  and  $\theta$ . This prior distribution usually depends on hyperparameters  $\gamma$  so that the marginal posterior of  $\psi$  is given by

$$p(\psi | y) = \int p(\psi, \gamma | y) dy \quad (4)$$

Point estimates of the relative risks can be obtained via location measures of the distribution (4) while scale measures provide information on the uncertainty of these estimates. In general the integrals involved in the computation of these measures cannot be obtained analytically or even by numerical integration and approximation methods are necessary. In particular, Markov chain Monte Carlo methods (MCMC) will be employed to obtain a sample from the joint posterior distribution of  $(\psi, \gamma)$ , automatically generating samples from the marginal posteriors of  $\psi$  and hyperparameters  $\gamma$ . The joint posterior distribution of all parameters is expressed as

$$p(\theta, \phi, \beta, \tau_\theta, \tau_\phi, \tau_\beta) \sim p(y | \theta, \phi, \beta) p(\theta | \tau_\theta) p(\phi, \tau_\phi) p(\beta | \tau_\beta) p(\tau_\theta) p(\tau_\phi) p(\tau_\beta) \quad (5)$$

This joint posterior distribution takes into account a conditional independence structure. In the highest level of the hierarchy prior distributions are specified for the prior precisions  $\tau_\beta$ ,  $\tau_\theta$  and  $\tau_\phi$ . The Gamma family of prior distributions is conditionally conjugate, i.e. the full posterior conditional distribution is also Gamma. This conditional conjugacy allows that  $\tau_\theta$  and  $\tau_\phi$  be easily updated. A common choice in the literature is the non-informative (proper) prior Gamma  $(\epsilon, \epsilon)$  with small values for  $\epsilon$ . However, this specification attributes low prior probability to



small values of the standard deviation and consequently a spatial structure for example might be imposed a priori. Kelsall and Wakefield (2002) verified that the estimation of relative risks can be highly dependent of the choice of prior parameters and within a class of Gamma priors they suggest a Gamma (0.5, 0.0005) distribution as a sensible choice.

### Monitoring MCMC convergence

There are several strategies for monitoring convergence of MCMC sampling. Brooks and Gelman (1998), and Robert (1998, chap. 2) provide excellent details regarding the concept of convergence. The following methods have been considered to check convergence of MCMC in this work.

*i) Simple graphical methods (working on single/multiple chains)* *Time-series plot*: Provides a graphical check of the stability of the generated parameter values. *Running mean plot*: Checks graphically if the mean of a parameter has stabilized. *Running quantiles plot*: If the quantiles at the end of the sampling chain resemble those produced earlier in the chain, this suggests that the first portion of the simulated Markov chain has similar statistical properties to the entire chain and may indicate that stationarity has been achieved. *Autocorrelation function plots (ACF)*: although not convergence diagnostics, help indirectly to assess convergence of MCMC algorithms; higher ACFs will indicate the need of a longer run of the program.

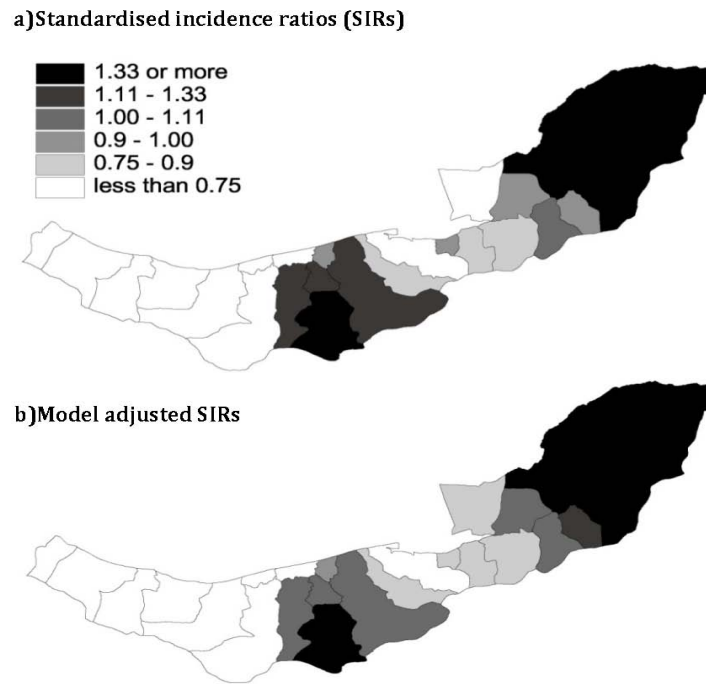
*ii) Methods using ratio of dispersions (multiple chains)* *Gelman-Rubin Potential Scale Reduction Factor*: Checks for each parameter if the scale/variance of its approximated posterior distribution will decrease significantly if simulations continue indefinitely. Brooks and Gelman (1998) interpret this as a check of any significant difference among the chains of inferences based only on sample mean and variance.

## 3 Application

Residents of Mazandaran and Golestan provinces of Iran constitute the study population. These provinces are located in the north of Iran (south of the Caspian Sea). There are 15 and 11 wards in Mazandaran and Golestan provinces respectively (see map in Figure 1). A total of 1701 new esophageal cancer cases were diagnosed in 2001-2005 in this region (see Mohebbi et al., 2008 a and b). Ward specific age standardised incidence ratios (SIR) for both sexes combined of esophageal cancer are modelled in this study. Figure 2.a plots the SIRs of esophageal cancer in the study area. Generally, the western areas of the region had lower cancer incidence compared to eastern and central wards.

Three selected models were considered in this study: non spatial structure, spatial structure, and a joint model with both spatial and non spatial structures. Table 1 summarizes the results for each of the three models. Map 'b' in Figure 2 was constructed with the posterior median of the  $\psi_j = e^{(\theta_j + \phi_j)}$  as point estimates of the SIRs.

Figure 2.a presents the map of the model adjusted (smoothed) incidence ratio of esophageal cancer in the study area. This index accounts for information on the unstructured heterogeneity random effect and the spatial dependence effect. The figure clearly showed characteristic Bayesian shrinkage of the crude rate toward the local average rate. In particular crude SIR ratio (see Table 1 for definition) of 4.83



**Fig. 2.** Maps of relative risks obtained via maximum likelihood estimates and model adjusted Bayesian hierarchical model.

had been reduced to 2.33 for the joint model. The high values in the eastern and central regions still remained high which indicated some tendency for local clustering of similar values.

Two parallel sampling chains were run with different initial values. WinBUGS was used to perform 900,000 simulations from the full conditional posterior distributions, from which the first 100,000 were discarded as burn-in. The three models described above had different 'burn-in' (pre-convergence) periods, with slower convergence for the more complex models. Several characteristics of the posterior distribution were estimated, and the main interest here was the relative risks density. The history graphs for selected posterior distributions in Figure 3 indicate that the model converged well after several thousand iterations.

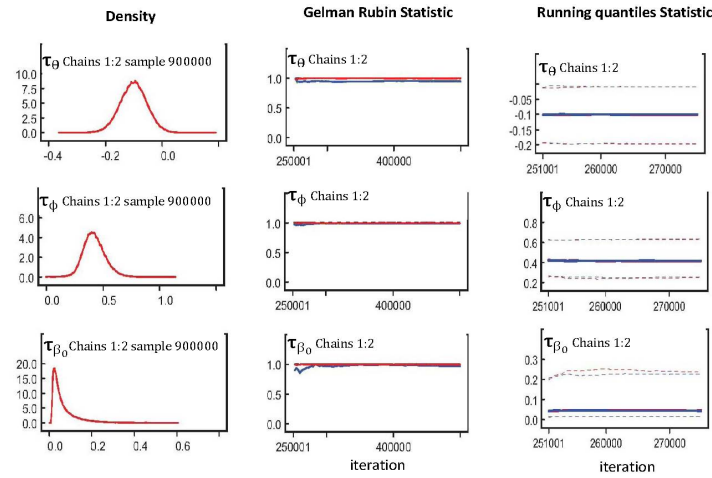
## 4 Discussion

In this study we adopted a Bayesian approach to estimate relative risks of a rare event occurrence in small areas. The problem of overdispersion found in the usual classical estimation was tackled via specification of suitable priors. The method was illustrated with a real world example. Estimates of the posterior distribution were obtained via MCMC methods where inference is based on an approximate sample from the posterior distribution.

**Table 1.** Posterior median summaries under three selected models: non spatial structure, spatial structure, and a joint model with both spatial and non spatial structures.

Model	Posterior median summaries				SIR ratio <sup>1</sup>	DIC
	$\tau_\theta$	$\tau_\phi$	$\tau_{\beta_0}$	Intercept		
Heterogeneity	0.18	-	0.12	-0.12	2.87	943.7
Spatial (CAR)	-	0.44	0.11	-0.14	2.60	926.5
Joint	0.06	0.42	0.12	-0.10	2.33	916.2

1. The SIR ratio refers to the ratio of the 95th and 5th percentiles of the distribution of the SIRs.



**Fig. 3.** Graphical monitoring of MCMC for selected parameters.

We have presented a three-level Bayesian hierarchical modelling approach to model the incidence ratio of esophageal cancer in the Caspian region of Iran. The first level was the likelihood of the Cancer SIR that follows a Poisson distribution. Level 2 modelled the log relative risk as a linear combination of three components which accounted for fixed effects of possible covariates, random effects of unstructured heterogeneity and spatial dependence. At level 3, non-informative hyperprior distributions were assigned to the precision parameters for the random effects. Three models that included/excluded the two random terms were compared based on Deviance Information Criterion. The full model which considered both unstructured and spatial dependence random effects was selected as the best fitting model.

**Conclusion:** Careful use of techniques based on hierarchical models is an improvement over non-hierarchical models. In particular, disease maps produced using Bayesian smoothing methods are likely to be less visually misleading than their predecessors based on non-hierarchical approaches. However, valid inference in spatial regression requires acknowledgment of residual spatial dependence.

## References

- BESAG, J., YORK, J. and MOLLIE, A. (1991): Bayesian image restoration with two applications in spatial statistics. *of the Institute of Statistics and Mathematics* 43, 1-59.
- BROOKS, S. and GELMAN, A. (1998): General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7(4), 434-456.
- CLATON, D. G. and KALDOR, J. (1987): Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics* 43, 671-682.
- CRESSIE, N. A. C. (1993): *Statistics for Spatial Data, revised edn.* Wiley, New York.
- KELSALL, J. and WAKEFIELD, J. (2002): Modelling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association* 97, 692-701.
- MOHEBBI, M., MAHMOODI, M., WOLFE, R., NOURIJELYANI, K., MOHAMMAD, K., ZERAATI, H. and FOTOUHI, A. (2008, a): Geographical spread of gastrointestinal tract cancer incidence in the Caspian Sea region of Iran: Spatial analysis of cancer registry data *BMC Cancer* 8, 137.
- MOHEBBI, M., NOURIJELYANI, K., MAHMOODI, M., MOHAMMAD, K., ZERAATI, H., FOTOUHI, A. and MOGHADASZADEH, B.(2008, b): Time of Occurrence and Age Distribution of Digestive Tract Cancers in Northern Iran. *Iranian Journal of Public Health*, 37(1), 8-19.
- MOLLIE, A. (1996): *Bayesian mapping of disease.* In GILKS, W. R., RICHARDSON, S., SPIEGELHALTER, D. J. (Eds): *Markov Chain Monte Carlo in Practice.* Chapman and Hall. London.
- ROBERT, C. P. (Ed.). (1998): *Discretization and MCMC convergence assessment.* Wiley, New York.

# Clusters of Gastrointestinal Tract Cancer in the Caspian Region of Iran: A Spatial Scan Analysis

Mohammadreza Mohebbi<sup>1</sup> and Rory Wolfe<sup>1</sup>

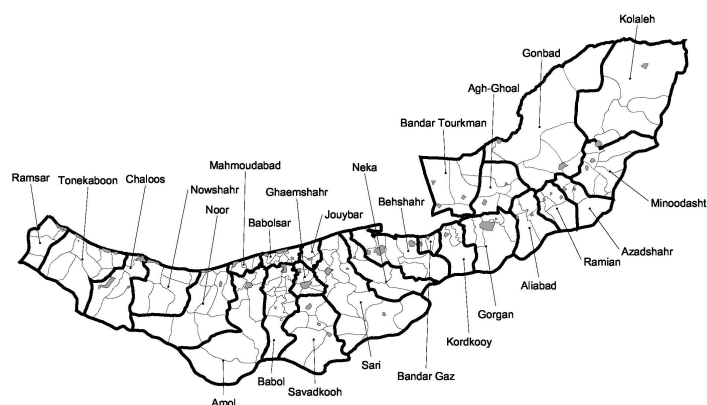
1. Department of Epidemiology and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia,  
*Mohammadreza.Mohebbi@med.monash.edu.au*

**Abstract.** Demographic data and age-specific gastrointestinal tract cancer incidence rates were obtained for all agglomerations of Caspian region for 2001-2006. A spatial scan statistic was used, which searched for clusters of disease without specifying their size or location ahead of time, and which tested for their statistical significance while adjusting for the multiple testing inherent in such a procedure. A primary cluster of high incidence was identified in the eastern agglomerations for esophageal and stomach cancer in male, female and both sexes combined. Surveillance findings such as these have the benefit of providing insight to the epidemiologist and might lead to monitoring geographical trends for cancer control activities.

**Keywords:** disease clustering, gastrointestinal cancer, Monte Carlo method, Poisson distribution, spatial autocorrelation

## 1 Introduction

Spatial cluster detection is an important tool in cancer surveillance to identify areas of elevated risk and to generate subsequent hypotheses about cancer etiology. A spatial disease cluster may be defined as an area with an unusually elevated disease incidence rate. There are several cluster detection methods used in spatial epidemiology to investigate apparently suspicious groupings of cancer occurrences in both regional count data and case-control data. Overviews of cluster detection methods can be found in Lawson et al. (1999) and Waller and Gotway (2004). An exploratory approach involving many overlapping circles was introduced in Openshaw et al. (1998). Turnbull et al. (1990) create overlapping circles with constant disease risk that partition the study region. While Turnbull et al. used a Monte Carlo simulation to assess statistical significance, Kulldorff and Nagarwalla (1995) provided a generalization that used a likelihood ratio test. Statistical cluster detection methods are generally classified into two main categories: Global and Local. Global tests of clustering identify areas with excess numbers of cases whereas local tests identify areas with excess numbers of cases in the presence of potential causes. In this research we focused on global tests and implemented methods that allow for diverse population sizes among the geographic areas. More specifically



**Fig. 1.** Geographic boundaries of cities (gray polygons) and rural agglomerations (outlined with light black limits) within wards (outlined with dark black limits) in Mazandaran (western) and Golestan (eastern) provinces.

we used Kulldorff's scan statistic developed by Kulldorff and Nagarwalla (1995) for the evaluation of the null hypothesis of no significant global spatial clustering. We began by describing the cluster detection method and used a testing algorithm analogous to those proposed in Kulldorff et al. (1997).

We illustrated the approach for a cancer registry data set from the Caspian region of Iran on gastrointestinal tract (GI) cancer. The study region contains two provinces: Mazandaran and Golestan. The provinces of Iran are subdivided into wards. There are usually a few cities and rural agglomerations in each ward. Rural agglomerations are a collection of a number of villages. Currently, Mazandaran province has 15 wards, 46 cities and 110 agglomerations and Golestan province has 11 wards, 24 cities and 50 agglomerations. The total population of these two provinces is approximately 4.5 million (1.6 million in Golestan province) constituting about 6.4 percent of the total Iranian population. Figure 1 shows geographic boundaries of cities and rural agglomerations within wards in Mazandaran and Golestan provinces.

The paper is organized as follows. In the methods section, we introduce the statistical methodology related to the spatial scan statistic; in the application section we first describe the geographic population under study as well as the demographic data and thereafter we describe the scan statistic setting to detect local clusters of EC and GC in the Caspian region. Finally we discuss the strengths of the methods and suggest analysis improve further on the results.

## 2 Method

### Data structure

The basic form of the data involves a set of observed counts (one count for each agglomeration) and a matching set of expected counts reporting the number of cases we expect in each agglomeration, under the null hypothesis. We assume that

the observed disease counts arose from a heterogeneous Poisson process (i.e., the data  $Y_1, Y_2, \dots, Y_S$  in agglomerations 1, 2, ..., S are mutually independent Poisson random variables). In addition, we assume population counts for each agglomeration, denoted by  $n_1, n_2, \dots, n_S$  are fixed (non-random). These population counts are used in determining the number of expected cases in each agglomeration, under a null hypothesis of no clusters, denoted by  $E_1, E_2, \dots, E_S$ .

### Null Hypothesis

In this regional count method observed incident disease counts are compared with census-based population counts for the same set of agglomerations. For this purpose the constant risk hypothesis was used. This means that people were equally likely to contract the disease regardless of location. The method compared observed counts to their corresponding expected counts based on a global incidence rate (proportion) applied to agglomeration specific population counts. The constant risk hypothesis assumes a constant disease risk,  $\theta$ , giving  $E_i = \theta n_i$   $i = 1, 2, \dots, S$ . The expected counts may also be standardized. For example, if we have age-specific rates  $\theta_j$  for age groups  $j = 1, 2, \dots, J$  and population sizes  $n_{ij}$  for the same age groups within each agglomeration, we can define an age-adjusted expected count via  $E_i = \sum_j \theta_j n_{ij}$  for  $i = 1, \dots, S$ .

We modelled the agglomeration specific regional counts as independent Poisson random variables based on one of the basic properties of a spatial Poisson process: event counts from non-overlapping agglomerations follow independent Poisson distributions where the underlying intensity function defines the expected values (and variances).

### The Spatial Scan Statistics

In an agglomeration specific count setting, Kulldorff (1997) considers distance-based circles with radii ranging from the smallest observed distance between a pair of agglomerations (e.g., intercentroid distance) to a user-defined upper bound (e.g., one-half the width of the study area). An agglomeration contributes all of its cases and individuals at risk to the circle if the agglomeration's centroid falls within the circle. At each possible radius in the user-defined interval (e.g., at each observed intercentroid distance) and for each circle having that radius, we calculate a likelihood ratio statistic testing the constant risk hypothesis versus the specific alternative that risk within agglomerations having their centroid within the circle, the zone of interest, differs from the risk in the rest of the study area.

Let  $n_+$  be the total population observed in the study region G, which is the sum of the population in each agglomeration,  $n_r$ ,  $r = 1, \dots, S$ . Similarly, we use  $Y_+$  and  $Y_r$  to denote the total number of cases in the whole study region and in the agglomerations ( $r$ ), respectively. For a given zone  $\mathcal{Z}$ , the likelihood is defined as

$$L(\mathcal{Z}) = \left[ \left( \frac{Y_{\mathcal{Z}}}{n_{\mathcal{Z}}} \right)^{Y_{\mathcal{Z}}} \left( \frac{n_{\mathcal{Z}} - Y_{\mathcal{Z}}}{n_{\mathcal{Z}}} \right)^{n_{\mathcal{Z}} - Y_{\mathcal{Z}}} \left( \frac{Y_{\mathcal{Z}'}}{n_{\mathcal{Z}'}} \right)^{Y_{\mathcal{Z}'}} \left( \frac{n_{\mathcal{Z}'} - Y_{\mathcal{Z}'}}{n_{\mathcal{Z}'}} \right)^{n_{\mathcal{Z}'} - Y_{\mathcal{Z}'}} \right] I \left( \frac{C_{\mathcal{Z}}}{n_{\mathcal{Z}}} > \frac{C_{\mathcal{Z}'}}{n_{\mathcal{Z}'}} \right) \quad (1)$$

where  $\mathcal{Z}$  is the collection of all the possible cells ( $r$ 's), in study region G, and  $\mathcal{Z}'$  is its complement, i.e.  $\mathcal{Z}' = G - \mathcal{Z}$ . ( $Y_{\mathcal{Z}}, Y_{\mathcal{Z}'}$ ) are the numbers of cases inside and

outside of  $Z$ , respectively, and  $(n_Z, n_{Z'})$  are the corresponding populations. Note that we are only interested in identifying clusters of cells with higher rates; hence, the indicator function is included in the likelihood. The statistic for our methods is  $\lambda = \max_{Z \in G}$ . The zone  $G$ , which maximizes  $\lambda$ , is called the most likely cluster. Under the constant risk hypothesis, the expected counts consist of age-standardized values or of regional population sizes multiplied by an estimate of the overall risk. Therefore, the spatial scan statistic is proportional to

$$\max \left( \frac{Y_Z}{E_Z} \right)^{Y_Z} \left( \frac{Y_{Z'}}{E_{Z'}} \right)^{Y_{Z'}} \quad (2)$$

For hypothesis testing a Monte Carlo method is used. The method generates independent data sets under the null hypothesis, calculates the likelihood ratio statistic for each circle, and stores the maximum statistic value. To find the distribution of the test statistic under the null hypothesis, Monte Carlo hypothesis testing is required. In this paper, p-value of the test is based upon the null distribution of likelihood ratio test statistic with a large number of Monte Carlo replications of the data set generated under the null hypothesis. This Monte Carlo based hypothesis test was first proposed by Dwass (1957). Statistics are correlated between circles within each simulation, but the maximum values are independent between simulations, providing a valid p-value for the most likely cluster, provided that one interprets the p-value as the probability of observing a more extreme maximal statistic anywhere in the study area (rather than the significance of observing the maximum at a particular location). Irrespective of the number of Monte Carlo replications chosen, the hypothesis test is unbiased, resulting in a correct significance level. The number of replications does affect the power of the test, with more replications giving slightly higher power.

We used Kulldorff's *SaTScan*<sup>TM</sup> (2006) software for implementation of the spatial scan statistic method. This test has been shown to have good power for detecting localized hot-spots of excess events (Kulldorff et al. 2003 and Song and Kulldorff, 2003). We imposed a circular window on the region's map and allowed its centre to move over the area so that at any given position, the window included different sets of neighbouring agglomerations (see Figure 1).

In addition to the most likely cluster, there may be secondary clusters that do not overlap the most likely cluster. We can report secondary clusters of this type if the likelihood ratio is larger than the likelihood ratio for the most likely cluster for at least one data set simulated under the null hypothesis. The *SaTScan*<sup>TM</sup> program can identify secondary clusters in the region and can order them according to their likelihood ratio.

### 3 Application

#### Study Population

All residents of Mazandaran and Golestan provinces of Iran constitute the study population. The cases of interest were all EC and GC patients registered between 2001 and 2006 among the study population. Data on incident cases of cancer was obtained from the Babol Cancer Registry; issues related to methods, quality and completeness of data collection for this cancer registry are described in Mohebbi



et al. (2008 a and b). In summary, the major sources of data collection related to cancer in the Babol cancer registry were reports from pathology laboratories, hospitals, and radiology clinics. Coding of cancer diagnosis samples was based on the international classification of disease for oncology (ICD-O) coding and was done under direct supervision of pathology specialists.

The estimated midyear population between 2001 until 2006, stratified for sex, age in five-year intervals, and place of residence (county/city) was obtained from the Statistical Centre of Iran (2003 a and b). We calculated the age standardized incidence ratio (SIR) of each ward for each sex separately and for both sexes combined. The population of the region was fairly stable between 2001 and 2006 so 2003 figures were used as the standard population, and indirect standardization was used to calculate SIR.

### Cluster detection method

We used the spatial scan statistic to detect local clusters in EC and GC SIR maps. For practical reasons, the centre of the window was positioned only at the 160 agglomeration centroids; and at each position, the radius of the circular window was varied continuously from zero up to a maximum radius so that the window never included more than 25 percent of the total population. In this way, the circular window was flexible both in location and size. In total, the method created a very large number of distinct circular windows, each with a different set of neighbouring agglomeration within it, and each a possible candidate for containing a cluster of high cancer incidence. In the present study the alternative hypothesis refers to elevated SIR inside the space as compared to outside, and a p-value less than 0.05 was used for statistical significance.

## 4 Results

The results from *SaTScan*<sup>TM</sup> identified clusters of agglomerations with significant high SIR in EC and GC for males, females and both sexes combined (see Table 1). As presented in Figure 2 all major cancer clusters were located in the eastern agglomerations. As shown in Figure 2, the primary cluster of EC for both sexes combined was located in the eastern agglomerations. A secondary cluster was also identified in the central part of the Mazandaran province at the location of the cities Sari and Ghaemshahr (see Figures 1 and 2). A similar pattern was seen for EC in males and females separately except that the primary cluster in Golestan province was divided into non-overlapping clusters. Similarly for GC the primary cluster of elevated SIR was located in the Eastern part of Golestan province and there were secondary clusters in Mazandaran province for males and both sexes combined.

## 5 Discussion

Compared with other cluster detection methods for spatial data (Lawson et al. (1999) and Waller and Gotway (2004)), the spatial scan statistic has some features

**Table 1.** Spatial scan statistics for detecting local clusters in EC and GC SIR's.

Cancer site	Cluster type <sup>1</sup>	p-value	No. Cases	No. Expected	Mean inside <sup>2</sup>	Mean outside <sup>3</sup>
Esophageal;	P	0.001	1045	628.3	1.86	0.90
Both Sexes	S	0.013	230	94.3	1.73	0.88
Esophageal;	P	0.0001	258	164.2	1.87	0.78
Male	S	0.0001	128	88.2	1.61	0.80
	S	0.0001	63	30.1	1.76	1.00
Esophageal;	P	0.0001	257	132.4	1.67	0.73
Female	S	0.0001	148	80.2	1.58	0.93
	S	0.011	44	23.1	1.47	0.71
	S	0.045	63	37.5	1.44	1.01
Gastric;	P	0.021	481	232.1	1.54	0.83
Both Sexes	S	0.008	442	238.3	1.21	0.98
Gastric;	P	0.031	282	181.6	1.17	0.93
Male	S	0.030	111	70.1	1.12	0.81
	S	0.002	18	10.5	1.08	0.75
Gastric;	P	0.015	179	114.9	1.39	1.09
Female	S	0.041	101	68.1	1.08	0.89

1. P = primary cluster; S = Secondary cluster.

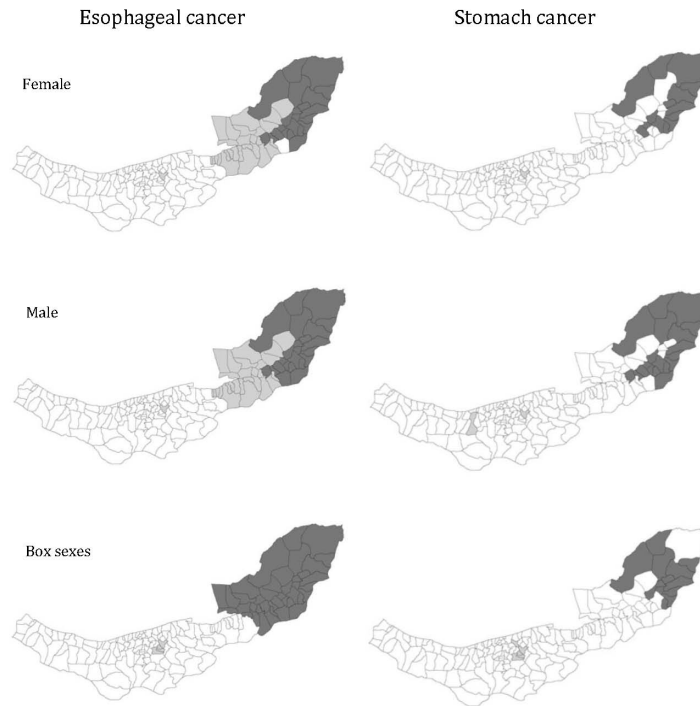
2, 3. SIR's mean in agglomerations inside and outside the circles generated by SatScan analysis

that make it preferable as a screening tool for evaluating reported disease clusters. We can use the method for the inhomogeneous population density and it can adjust for confounding variables. The problem of pre-selection bias in most other cluster detection methods is ameliorated in spatial scan test because it searches for clusters without specifying their size or location. The likelihood ratio-based test statistic takes multiple testing into account and calculates a single p-value for the test of the null hypothesis, and finally, if the null hypothesis is rejected, we can specify the approximate location of the cluster that caused the rejection.

Ecologic studies are perhaps best considered to be hypothesis generating, although small area analysis tends to reduce ecological fallacy, since the populations defined by agglomerations boundaries are more homogeneous. While this might well be true of villages and towns of average size, in large cities, however, the results reported here correspond to an overall mean, and important risk factors such as socio-economic and dietary patterns differences have been disregarded. In current work we are extending this work by assessing whether such relationships exist by implementing a multilevel spatial Poisson regression model.

## References

- DWASS, M. (1957): Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 28, 181-187.
- KULLDRFF, M. and NAGARWALLA, N. (1995): Spatial disease clusters: detection and inference. *Statistics in Medicine* 14, 799-810.



**Fig. 2.** Local clusters of esophageal (left) and gastric cancer's SIR in female, male and both sexes combined. Dark gray indicates primary clusters and light gray shows secondary clusters.

- KULLDRFF, M. (1997): A spatial scan statistic. *Communications in Statistics: Theory and Methods* 26, 1487-1496.
- KULLDRFF, M., TANGO, T. and PARK, P. J. (2003): Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis* 42, 665 - 684.
- KULLDRFF, M. (2006): Sat Scan V. 7.0. *Software for the spatial and space-time scan*.
- LAWSON, A., BIGGERI, A., BOHNING, D., LESAFFRE, E., VIEL, J. F. and BERTOLLINI, R. (1999): *Disease Mapping and Risk Assessment for Public Health*. John Wiley & Sons Ltd. West Sussex, UK.
- MOHEBBI, M., MAHMOODI, M., WOLFE, R., NOURIJELYANI, K., MOHAMMAD, K., ZERAATI, H. and FOTOUHI, A. (2008, a): Geographical spread of gastrointestinal tract cancer incidence in the Caspian Sea region of Iran: Spatial analysis of cancer registry data *BMC Cancer* 8, 137.
- MOHEBBI, M., NOURIJELYANI, K., MAHMOODI, M., MOHAMMAD, K., ZERAATI, H., FOTOUHI, A. and MOGHADASZADEH, B. (2008, b): Time of Occurrence and Age Distribution of Digestive Tract Cancers in Northern Iran. *Iranian Journal of Public Health*, 37(1), 8-19.

- OPENSHAW, S., CHARLTON, M., CRAFT, A. W. and BIRCH, J. M. (1988): Investigation of leukemia clusters by use of a geographical analysis machine. *Lancet* 331, 272-273.
- SONG, C. and KULLDRFF, M. (2003): Power evaluation of disease clustering tests. *International Journal of Health Geographic* 2(1)9.
- Statistical Center of Iran (2003, a) *Reconstruction and estimation of Golestan province population according to 2000 geographic boundaries*. Statistical Center of Iran, Tehran.
- Statistical Center of Iran (2003, b) *Reconstruction and estimation of Mazandaran province population according to 2000 geographic boundaries*. Statistical Center of Iran, Tehran.
- TURNBULL, B. W., IWANO, E. J., BURNETT, W. S., HOWE, H. L. and CLARK, L. C. (1990): Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology*, 132, S136-S143.
- WALLER, L. A. and GOTWAY, C. A. (2004): *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons Ltd. Hoboken, NJ.

# The Financial Crisis of 2008: Modelling the Transmission Mechanism Between the Markets

M. Pilar Muñoz<sup>1</sup> Maria Dolores Márquez<sup>2</sup> and Helena Chuliá<sup>3</sup>

<sup>1</sup> Departament of Statistics and Operations Research, Universitat Politècnica de Catalunya

C/ Jordi Girona 1-3, Campus Nord C5 204, 08034 Barcelona, Spain,  
*pilar.munoz@upc.edu*

<sup>2</sup> Departament of Business Economics, Universitat Autònoma de Barcelona  
C/ Emprius, 2 Sabadell, Barcelona, Spain, *mariadolores.marquez@uab.cat*

<sup>3</sup> Departament of Economics and Business, Open University of Catalunya  
C/ Jordi Girona 1-3, Campus Nord C5 204, 08034 Barcelona, Spain,  
*hhulia@uoc.edu*

**Abstract.** In this work we investigate how the current failures of the United States financial institutions have affected most of the stock markets in the world. First, we apply Time Series Factors Analysis (TSFA) in order to reduce the dimensionality of the number of indexes and obtain a lower number of new factors that can be related to regions. Then we use the dynamic conditional correlation (DCC) model to analyze the linkages between these regions. Our approach allows us to distinguish between contagion and interdependence. The results show evidence of a contagion effect between some regions.

**Keywords:** contagion, multivariate volatility, time series factor analysis and dynamical conditional correlation

## 1 Introduction

Over recent decades, we have seen how different financial crises, having originated in particular regions and countries, have extended geographically. As world capital markets have become increasingly integrated, information originating from one market is likely to become more important for other markets. In fact, understanding the nature of linkages between financial markets, whether intra- or international, is fundamental to establishing the limits of diversification, to security pricing, and to successfully allocating assets.

When analysing the interrelations across different financial markets it is necessary to distinguish between interdependence and contagion. Forbes and Rigobon (2002) define contagion as significant increases in cross-market comovements, while any continued market correlation at high levels is considered to be interdependence. Therefore, the existence of contagion must involve evidence of a dynamic increment in correlations.

The existing literature on contagion has several limitations such as the existence of a heteroskedasticity problem when measuring correlations caused by volatility

increases during the crisis (Forbes and Rigobon, 2002) and the choice of the window length to identify the crisis (Billio and Pelizzon, 2003)<sup>1</sup>.

This paper analyzes the linkages among the indexes of nineteen international stock markets during the last financial crisis, and models the transmission mechanism among them. Following Chiang et al. (2007) to overcome the limitations found in the existing literature, this paper employs a cross country, multivariate GARCH model, which is appropriate for measuring time-varying conditional correlations. Concretely, we use the multivariate GARCH model with dynamic conditional correlation (DCC-GARCH) introduced by Engle (2002). Although this model can be used to examine multiple asset returns without adding too many parameters, our analysis uses a large number of series (nineteen). For that reason, first we apply a Time Series Factors Analysis (TSFA) in order to reduce the dimensionality.

Results show that in most cases correlation coefficients are more volatile and increase during the different crises reducing the benefits of international portfolio diversification. The remainder of this paper is organized as follows. Section 2 introduces the data and formulates the econometric model. Section 3 discusses the empirical results. The paper concludes with a summary of the main results.

## 2 DATA and METHODOLOGY

In order to verify empirically if stock markets are globalized and, if there is a contagion effect imposed for the recent crisis, nineteen stock markets indexes are analyzed<sup>2</sup>. The data set consists of the daily stock-price indexes of the North American, European, Japanese and Asian markets. The considered indexes are Standard and Poor's 500 (SP), Dow Jones Industrial Index JONES (DJI), Nasdaq (NAS) and the Canadian Toronto Index SE300 (SE300) from the North-American market; from Europe, we take the indexes of the main markets: Germany (DAX), France (CAC40), Italy (MIB30), UK (FTSE) and Spain (IBEX35). Nikkei (NIK) and Topix (TOPX) are the indexes taken into account from Japan and finally we include the most representative stock indexes from Southeast Asia: Hong Kong (Hang Seng Index HSI), Philippines (Philippines SE Composite IPSE), Korean (Korea SE Composite, KS11), Singapore (Singapore Straits Time Index STI), Taiwan (Taiwan SE Weighted Index, TWII), Indonesia (Jakarta SE Composite Index JKSE), Malaysia (Kuala Lumpur SE Index KLCI) and Thailand (BANGKOK S.E.T., SET).

A data base<sup>3</sup> which begins with December 31, 1994 and ends with September 30, 2009 is considered. In this paper our interest is focused on the last period<sup>4</sup> but

<sup>1</sup> See Chiang et al. (2007) for a detailed explanation of the potential drawbacks and limitations of the existing contagion tests.

<sup>2</sup> Data are provided by Thomsom-Reuters and Datastream.

<sup>3</sup> In the case of national holidays in any country, the missing value is replaced by the last trading value. The number of observations used in each series is 3834, and the data are analyzed using Comprehensive R Archive Network (CRAN) and WinRats 6.0.

<sup>4</sup> The financial crisis began in 2007 with the mortgage and banking crisis. The collapse started in the middle of September 2008 with the bankruptcy of Lehman and the bailout of AIG.

analyzing the transmission of volatility in past crisis periods (Asian, Dot.com and Subprime crisis) will help us to understand the actual financial crisis.

The dynamics of most of the time series are similar, except in the group of Asian markets, where different patterns can be observed at the beginning, but from 2002 the dynamic is similar for all indexes. These facts show a clear interdependence between the markets, and perhaps the existence of a contagion effect which supposes that a dramatic movement in one stock market has a powerful impact on other markets. Contemporaneous correlations within the markets are analyzed allowing us to detect high correlations within the European indexes, the North American indexes and the Japanese Markets; there is a high correlation between NIKKEI and TOPIX, and also within Asian Markets. Finally, there is a high correlation among European and North-American Markets, Japanese Markets and the Philippine Market.

As usual, stock returns are calculated as the first difference of natural log of each stock-price index and the returns are expressed as percentages. Time series returns exhibit the usual features of financial time series: non-normality, skewness and high kurtosis. We also observe a high standard deviation for the Nasdaq index and the Korean SE Composite, KS11. Given the huge number of indexes to analyze it is necessary to introduce a useful factor model for studying the common patterns in the returns time series (Dungey and Martin, 2007, Bowman, Chang and Comer, 2007). At this point, our approach is based on the Time Series Factor Analysis (TSFA) method introduced by Gilbert and Meijer (2005). Next, a dynamic conditional correlation model with asymmetric GARCH (DCC-AGARCH) is fitted to the factors obtained in order to estimate the pair-wise correlations between factors. Finally, the effect of the several crises on the conditional correlations is analyzed.

Different methods are used to reduce the dimensionality and capture the underlying structure of the return time series. Factor Analysis is the most commonly used but this method assumes that the data have no serial correlations (this assumption is violated by financial data); Dynamic Factors Analysis allows that the observations to be dependent over time, but is necessary modelling the process dynamics of the underlying phenomena. TSFA obviates the need for explicitly modelling the dynamics of the process and estimates a model for a time series with as few assumptions as possible, the observations do not need to be independent and identically distributed (i.i.d.) and the data does not need to be covariance stationary. The factors identified by TSFA method are latent variables and, with a reduced number of factors we can explain the dynamic structure of the data.

The relationship between the observed time series  $y_t$  (M-vector of length T) and the unobserved factors  $\varepsilon_t$  (k-vector with  $k \ll M$ ) is explained by the model :

$$y_t = \alpha_t + B\varepsilon_t + e_t. \quad (1)$$

where  $\alpha_t$  is the M-vector of intercept parameters, B is a Mxk matrix parameter of loadings and  $e_t$  is a random M-vector of measurement errors.

The DCC-AGARCH model fitted to the factors obtained from the TSFA procedure is an autoregressive of order AR(1) for the mean equation of the factor plus a one-day lagged factor not included in the mean equation, following the suggestion of Chiang et al. (2007). The idea of including the lagged  $factor_j$  is for checking if this factor has a dynamic effect in the determination of the  $factor_i$ . Concretely, the mean equation for the factor is

$$factor_{i,t} = \gamma_0 + \gamma_1 factor_{i,t-1} + \gamma_2 factor_{j,t-1} + \epsilon_{i,t} \quad (2)$$

where  $t = 1, \dots, n$ ,  $i = 1, \dots, 4$   $j \neq i$  and  $\epsilon_t \mid F_{t-1} \sim N(0, H_t)$ .  $F_t = \{factor_{i,1}, \dots, factor_{i,t-1}\}$  is the set of the observations until time  $t-1$  and  $H_t$  is the conditional variance matrix, that is decomposed as  $H_t = D_t R_t D_t$ , where  $R_t$  is the (nxn) time varying correlation matrix and  $D_t$  is a (nxn) diagonal matrix of time-varying standard deviations  $\sqrt{h_{ii,t}}$  obtained from the asymmetric univariate GARCH model:

$$h_{ii,t} = c + a\epsilon_{ii,t-1}^2 + b h_{ii,t-1} + d\eta_{ii,t-1}^2 \quad (3)$$

where the variable  $\eta_{ii,t-1} = \max[0, -\epsilon_{ii,t}]$  picks up the asymmetric effect in the univariate GARCH model. The residuals  $\epsilon_{ii,t}$  have been standardized as  $z_{i,t} = \epsilon_{ii,t} / \sqrt{h_{ii,t}}$  and  $z_{i,t}$  is used for estimating the parameters of the dynamic conditional correlation  $R_t$  as:

$$R_t = (1 - \alpha^2 - \beta^2)\bar{R} + \alpha\epsilon_{t-1}'\epsilon_{t-1} + \beta R_{t-1} \quad (4)$$

where  $\bar{R}$  is the unconditional correlation

Finally, the effects of the several crises on the dynamic conditional correlations have been studied by means of introducing a set of dummy variables, one for each crisis, in the mean equation (2) and the same dummy variables in the conditional correlation(3)<sup>5</sup>. The equations system made is described by equations (5) and (6). A significant estimated coefficient for the dummy variable will be interpreted as a structural change in mean and/or variance that produces a shift in mean and/or variance of the conditional correlation. These dummy variables are indicators that take the value 1 in the crisis period and 0 otherwise. They are *Crisis<sub>1</sub>* for the Asian crisis (10/22/1997 - 11/21/1997), *Crisis<sub>2</sub>* for the Dot.com crisis (3/10/2000 - 4/7/2000), *Crisis<sub>3</sub>* for the Subprime crisis (8/17/2007 - 9/16/2007) and finally *Crisis<sub>4</sub>* for the Financial crisis (9/15/2008 - 10/14/2008). The order P has been chosen by means of the AIC statistics and the goodness of fit of those estimations has been checked by means of the Ljung-Box statistics for the residuals and squared residuals.

$$\rho_{ii,t} = \mu + \sum_{p=1}^P \phi_p \rho_{ij,t-p} + \sum_{k=1}^4 \alpha_k Crisis_{k,t} + e_{ij,t} \quad (5)$$

$$h_{ij,t} = \omega_0 + \omega_1 \epsilon_{ij,t-1}^2 + \beta_1 h_{ij,t-1} + \sum_{k=1}^4 \delta_k Crisis_{k,t} \quad (6)$$

### 3 EMPIRICAL RESULTS

First of all, Time Series Factor Analysis (TSFA) has been applied to the whole set of the return time series in order to reduce the dimensionality. Following the conventional rule that the number of factors should be equal to the number of eigenvalues that are larger than one, we consider four factors, moreover the comparative fit index (CFI), a pseudo- $R^2$ , is greater than 0,98 and the root mean square error of

<sup>5</sup> The estimation has been made by maximizing the likelihood of the model composed for both equations. We developed our own program in WinRats 6.0.



approximation (RMSA) compares a model with a saturated model, for the 4-factors model the RMSA is less than 0,05 and this implies that the model fits well. The first factor represents North-American indexes, the second represents only the two Japanese indexes; the next factor is built with the European market indexes; and finally the indexes of the Southeast Asia markets are loadings in the fourth factor.

The results of applying the DCC-AGARCH model to the four factors are reported in Table 1. The one-day lagged North-American factor has been introduced as an explicative variable of the evolution of the other factors because it is well known that this factor acts as a global factor (Chiang et al. 2007). In this case we can conclude that the North American effect is highly significant for the other factors and the estimated coefficient  $\gamma_2$  is positive in the three relationships. This means that a movement (positive or negative) in the North-American factor has a big influence on the other factors on day later. The coefficients associated with the variance equations are significant for all factors, showing that the DCC are affected by heteroskedasticity and the asymmetry coefficient ( $d$ ) is always negative. This denotes that periods with negative residuals will be followed by periods of high volatility. Finally, after the North-American effect is introduced, the tests for detecting changes in correlations are reported in Table 2. First of all, the goodness of fit is correct as we can deduce for the values of the Ljung-Box  $Q(20)$  and  $Q^2(20)$ . Observing the mean equation, we can assume that the Asian crisis has affected the correlations, on the one hand between North-America and Japan and on the other hand between Japan and Europe. The Dot.com crisis has an effect on the North-American and European relationships. Neither of the coefficients associated with the Subprime crisis are significant. To finish the mean equation we detect that the Financial Crisis has a positive effect in the pair-wise relationships between North-America/Europe, North-America/Southeast Asia and Europe/Southeast Asia. Things are very different when we examine the variance equation in this table. The coefficients associated with the GARCH(1,1) model are all significant, indicating that it is necessary to correct the dynamic correlations by heteroskedasticity. When we observe the coefficients related to the crisis variables, we can conclude that: the Asian crisis produces a transmission of volatility to all the pair-wise correlations; the Dot.com crisis is significant for all the pair-wise correlations with the exception of North-America/Europe; the Subprime crisis is significant for all the pair-wise correlations with the exception of North-America/Europe and Japan/Southeast Asian; and, finally, the shock in the volatilities due to the Financial Crisis is only significant for the pair-wise correlations of North-America/Japan and Europe/Southeast Asia.

## 4 CONCLUSIONS

In this paper we analyse how the current failures of the United States financial institutions have affected most of the stock markets in the world. The objective is to differentiate between interdependence (any continued market correlation at high levels) and contagion (significant increases in cross-market comovements). To do this we examine the time-series behavior of correlation coefficients and analyse the impact of the crisis on their movements and variability. The data set consists of nineteen international stock markets and the sample period covers December 1994 to September 2009. The analysis is carried out through a Time Series Factor Analysis

to reduce dimensionality. The multivariate GARCH model is used with a dynamic conditional correlation (DCC-GARCH) to estimate cross country correlations.

The most important results of the paper are the following: First, after applying Time Series Factor Analysis we find that the nineteen stock indices can be grouped into four regions: North-America, Japan, Europe and Southeast Asia. Second, we find that the response of the pair-wise correlation coefficients is heterogeneous within the different crises. The Asian, Dot.com and Financial crises had an effect on the correlation between some of the regions. However, the subprime crisis did not affect the correlation between any of the regions. Third, we observe that the variability of the pair-wise correlation coefficients increases in most cases after the occurrence of the different crises. As a result, as Chiang et al. (2007) point out, the estimates and statistical inference of risk from risk models based on constant correlation coefficients can be very misleading. The finding that in most cases pair-wise correlation coefficients are more volatile and increase during the different crises suggests that the gain from international diversification investment in multiple markets is likely to be lowest when it is most desirable.

	Factor equation			Variance equation			
	$\gamma_0$	$\gamma_2$	$\gamma_2$	c	a	b	d
Factor 1:	0.019**	0.006		0.010***	0.137***	0.924***	-0.149
North-America	(2.063)	(0.403)		(10.197)	(20.510)	(238.645)	(-21.174)
Factor 2:	-0.020*	-0.019	0.409***	0.018***	0.112***	0.913***	-0.087***
Japan	(-1.801)	(-1.360)	(30.878)	(7.485)	(13.552)	(162.626)	((-8.701)
Factor 3:	-0.019*	-0.239***	0.391***	0.010***	0.113***	0.921***	-0.096***
Europe	(1.918)	(-17.164)	(25.977)	(8.504)	(31.005)	(266.134)	((-14.847)
Factor 4:	-0.009	0.084***	0.433***	0.020***	0.132***	0.880***	-0.059***
Southeast Asia	(0.837)	(6.295)	(31.771)	(5.915)	(8.650)	(73.563)	(-4.431)
Conditional Correlation equation		$\alpha$	0.008***	(6.871)	$\beta$	0.991***	(666.317)

**Table 1.** Estimation results from the DCC with Asymmetric GARCH(1,1) model for the factors, introducing the North America effect (The t-statistics are in parenthesis. \*\*\*, \*\* and \* denote statistical significance at the 1%, 5% and 10% level)

## References

- BILLIO, M., PELIZZON, L., (2003): Contagion and interdependence in stock markets: have they been misdiagnosed?. *Journal of Economics and Business* 55, 405-426.
- BOWMAN, R. G., CHAN, K. F., COMER, M. R., (2007): Contagion in world equity markets and the Asian economic crisis. *Downloaded from SSRN: [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=965316](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=965316)*.

- CHIANG, T.C., JEON, B.N., LI H., (2007): Dynamic correlation analysis of financial contagion: Evidence from Asian markets. *Journal of International Money and Finance* 26, 1206-1228.
- DUNGEY, M., MARTIN, V.L., (2007): Unravelling financial market linkages during crisis. *Journal of Applied Econometrics*, 22, 89-119.
- ENGLE, R.E., (2002): Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics* 20, 339-350.
- FORBES, K., RIGOBON, R., (2002): No contagion, only interdependence: measuring stock market comovements. *Journal of Finance* 57 (5), 2223-2261.
- GILBERT, P. and MEIJER, E., (2005): Time Series Factor Analysis with an Application to Measuring Money *Research Report N 05F10. University of Groningen, SOM Research School. Available at <http://som.rug.nl> .*

	North- America/ Japan	North- America/ Europe	North- America/ Southeast Asia	Japan/ Europe	Japan/ Southeast Asia	Europe/ Southeast Asia
Mean Equation						
Constant	6.0e-04*** (2.96)	4.079e-04 (0.86)	5.99e-04*** (2.73)	0.001*** (3.91)	3.9e-04 (1.12)	3.61e-04 (1.30)
$\rho_{t-1}$	0.997*** (960.60)	0.999*** (1143.56)	0.996*** (756.12)	0.996*** (926.98)	0.99*** (1112.01)	0.998*** (92.47)
Crisis <sub>1,t</sub>	5.5e-03** (1.92)	-2.10e-03 (-0.90)	-3.35e-03 (-0.60)	0.007** (2.35)	0.001 (0.458)	0.009 (1.52)
Crisis <sub>2,t</sub>	1.3e-03 (0.28)	-5.2e-03** (-1.89)	-2.94e-03 (-0.14)	8.97e-05 (0.04)	0.005 (0.208)	0.002 (0.390)
Crisis <sub>3,t</sub>	-8.16e-03 (-0.39)	3.65e-03 (1.45)	-2.06e-04 (-0.14)	0.004 (1.22)	0.009 (0.49)	9.26e-05 (0.02)
Crisis <sub>4,t</sub>	-2.81e-04 (-0.05)	7.82e-03** (2.28)	0.012* (1.68)	0.005 (0.84)	0.002 (3.28)	0.015** (1.98)
Variance Equation						
Constant	3.7e-06*** (22.11)	1.5e-06*** (24.60)	6.61e-06*** (29.66)	5.09e-06 (23.99)	6.7e-06*** (29.73)	3.5e-06*** (26.60)
$\varepsilon_{t-1}^2$	0.105*** (29.56)	0.139*** (25.49)	0.121*** (24.62)	0.151*** (26.11)	0.177*** (32.10)	0.193*** (29.71)
$h_{t-1}$	0.831*** (154.50)	0.842*** (164.44)	0.771*** (105.49)	0.763*** (96.46)	0.720*** (97.70)	0.767*** (117.16)
Crisis <sub>1,t</sub>	2.96e-05* (1.72)	4.3e-05*** (3.13)	2.59e-04*** (4.54)	5.5e-05** (2.21)	4.7e-05*** (2.66)	1.2e-04*** (3.08)
Crisis <sub>2,t</sub>	2.6e-05*** (5.17)	1.16e-05 (1.51)	2.75e-04*** (6.14)	2.2e-05*** (2.67)	4.6e-04*** (3.49)	7.8e-05*** (6.16)
Crisis <sub>3,t</sub>	2.47e-04*** (2.44)	-5.07e-06 (-1.44)	-7.72e-06*** (-4.11)	2.1e-05** (2.30)	2.02e-04 (1.49)	2.72e-05** (2.38)
Crisis <sub>4,t</sub>	4.97e-05* (1.78)	4.07 (1.53)	2.013e-04 (3.05)	5.4e-05 (1.61)	3.39e-05 (1.56)	1.56e-04* (1.88)
Q(20)	19.284	29.855*	16.706	14.298	19.840	28.804*
Q <sup>2</sup> (20)	18.837	6.169	0.888	7.678	3.154	10.686

**Table 2.** Test for detecting changes in the dynamic correlations across the markets due to the financial crisis, adjusted by autocorrelation coefficient and conditional heteroscedasticity, introducing the *North-America effect*.

Notes: The t-statistics are in parenthesis. \*\*\*, \*\* and \* denote statistical significance at the 1%, 5% and 10% level.

$$\text{Mean equation: } \rho_{ij,t} = \mu + \phi \rho_{ij,t-1} + \sum_{k=1}^4 \alpha_k \text{Crisis}_{k,t} + e_{ij,t}$$

$$\text{Variance equation: } h_{ij,t} = \sigma_0 + \sigma_1 \varepsilon_{ij,t-1}^2 + \beta_1 h_{ij,t-1} + \sum_{k=1}^4 \delta_k \text{Crisis}_{k,t}$$

where  $\text{Crisis}_{k,t}$  are the dummy variables defined in eq. (4) and (5), indicators of the different crises. Crisis<sub>1</sub> is the dummy variable for the Asian crisis (11/22/1997 – 11/21/1997), Crisis<sub>2</sub> is the dummy variable for the Dot.com crisis (3/10/2000 – 7/04/2000), Crisis<sub>3</sub> is the dummy variable for the Subprime crisis (8/17/2007 – 9/16/2007) and Crisis<sub>4</sub> is the dummy variable for the Financial crisis (9/15/2008 – 10/14/2008). Q(20) is the Ljung-Box statistic up to 20 days for testing the independency of the residuals and Q<sup>2</sup>(20) is the Ljung-Box statistic up to 20 days for the squared residuals in order to test the heteroscedasticity of them.

# Determining the Direction of the Path Using a Bayesian Semiparametric Model

Kei Miyazaki<sup>1</sup>, Takahiro Hoshino<sup>1</sup>, and Kazuo Shigemasu<sup>2</sup>

<sup>1</sup> Graduate School of Economics, Nagoya University  
Furo-cho, Chikusa-Ku, Aichi 464-8601, Japan,  
*miyazaki.behaviormetrics@gmail.com*

<sup>2</sup> Department of Psychology, Teikyo University  
Otsuka 359, Hachioji-shi, Tokyo 192-0352, Japan, *kshige@main.teikyo-u.ac.jp*

**Abstract.** In Bayesian estimation, hierarchical models with Dirichlet process prior distributions have been applied to various kinds of models (Ansari & Iyengar, 2006). This method makes it possible to perform MCMC estimation for any assumed shape of distributions for the parameters.

In this study, we consider a simple single regression model, and set two Dirichlet process mixture models wherein the explanatory and dependent variables are alternated with each other under the assumption that the error variables do not follow normal distributions. Then, we decide which model is better by calculating the marginal likelihood (Basu & Chib, 2003) in simulation studies.

**Keywords:** Bayesian semiparametric models, Dirichlet process priors, marginal likelihood, equivalent models

## 1 Introduction

In behavioral sciences, data are sometimes obtained from heterogeneous populations. In such cases, we face a situation where the observed variables do not follow normal distributions. Thus, the methods that do not require normality for the error variables are required to resolve this situation. Preceding studies such as Browne (1984) suggested the asymptotic distribution free method, which is a form of the generalized least square method. Kano, Berkane, and Bentler (1993) presented an estimation method that employed elliptical distribution. Recently, as a solution-oriented approach for the above problem, an estimation method that uses a higher-order moment structure sparked interest among researchers in this field (Bentler, 1983; Shimizu & Kano, 2008). This method can be applied to data that are generated from very skewed distributions. Moreover, this method makes it possible to determine the direction of path among the models that have the same values of goodness of fit (that is, equivalent models).

On the other hand, in Bayesian estimation, we can compare several models accurately by using marginal likelihoods. While it is necessary to make a distributional assumption for random variables in existing general Bayesian estimation method, Bayesian hierarchical models with Dirichlet process prior distributions (in other words, Dirichlet process mixture modeling) have been applied to various kinds of

models (Ansari & Iyengar, 2006; Hoshino, accepted for publication). The Dirichlet process was proposed by Ferguson (1973), and Escobar (1994) suggested the MCMC algorithm with the Dirichlet process. This method makes it possible to perform MCMC estimation for any assumed shape of distributions for the parameters. That is, we can conclude that Dirichlet process mixture modeling is appropriate for analyzing nonnormal data. Moreover, since marginal likelihood can be used in Bayesian estimation, we can check the direction of the path among the equivalent models.

In this study, we consider a simple single regression model, and set two Dirichlet process mixture models wherein the explanatory and dependent variables are alternated with each other under the assumption that the error variables do not follow normal distribution. We then decide which model is better by calculating the marginal likelihood in simulation studies.

## 2 Model

We consider two simple single regression models wherein the explanatory and dependent variables are alternated with each other, as shown below:

$$y_i = \beta x_i + e_i \quad (1)$$

$$x_i = \beta' y_i + e'_i. \quad (2)$$

The error variables follow nonnormal distributions.

### Dirichlet process mixture models

According to Sethuraman (1994), when Dirichlet process priors  $F \sim DP(\gamma, G_0)$  are assumed,  $F$  is expressed as follows:

$$F(\cdot) = \sum_{l=1}^{\infty} \kappa_l \delta_{\theta_l}(\cdot), \quad \theta_l \sim G_0 \quad (3)$$

where  $\delta_{\theta_l}$  denotes a discrete measure concentrated at  $\theta_l$ ,  $\kappa_l = \prod_{k=1}^{l-1} (1 - V_k) V_l$  and  $V_1, V_2, \dots$  independently follow a beta distribution  $\text{Be}(1, \gamma)$ . In this,  $\gamma$  is the parameter that indicates ease of transition to other components while  $G_0$  indicates a reference distribution. Parameters are generated from this distribution when the new component is generated. When the above Dirichlet process priors are assumed, a random variable vector  $\mathbf{y}$  is expressed by the following Dirichlet process mixture models:

$$\mathbf{y} \sim \sum_{l=1}^{\infty} \kappa_l f(\cdot | \boldsymbol{\psi}_l) \quad (4)$$

where  $f$  is the sampling distribution of data  $\mathbf{y}$  and  $\boldsymbol{\psi}_l$  is a parameter vector. This equation indicates that any distribution can be expressed as a mixture distribution of conventional distributions such as normal distributions, and it is not necessary to set the number of mixed components for analysis. Ishwaran & Zarepour (2000) proposed the following finite-dimensional Dirichlet process priors:

$$F(\cdot) = \sum_{l=1}^L \kappa_l \delta_{\theta_l}(\cdot), \quad \theta_l \sim G_0 \quad (5)$$

Assuming the above finite-dimensional Dirichlet process priors, a random variable vector  $\mathbf{y}$  is expressed by the following finite-dimensional Dirichlet process mixture models:

$$\mathbf{y} \sim \sum_{l=1}^L \kappa_l f(\cdot | \psi_l) \quad (6)$$

where  $L$  is the maximum number of components. Ishwaran & James (2001) (Theorem 2) proved that Eq.(6) approximates infinite-dimensional Dirichlet process mixture models with satisfactory accuracy when the value of  $L$  is large enough and the value of the upper bound of errors that changes according to the sample size and the value of  $L$  is described with sketches of proofs. According to their paper, the truncation value of  $L$  has more influence than the sample size on reducing the error (Theorem 2).

### Hierarchical representation of the model

the entire structure of Bayesian semiparametric models with finite-dimensional Dirichlet process priors is represented hierarchically as follows (Ishwaran & James, 2001)B

$$\begin{aligned} (\mathbf{y}_i | \boldsymbol{\theta}, \mathbf{k}, \boldsymbol{\psi}) &\sim p(\mathbf{y}_i | \boldsymbol{\theta}_{k_i}, \boldsymbol{\psi}) \\ (\mathbf{k} | \mathbf{p}) &\sim \sum_{l=1}^L p_l \delta_l(\cdot) \\ (\mathbf{p}, \boldsymbol{\theta}) &\sim p(\mathbf{p}) p(\boldsymbol{\theta} | \boldsymbol{\tau}) \\ \boldsymbol{\psi} &\sim p(\boldsymbol{\psi}) \end{aligned} \quad (7)$$

$k_i$  is the component to which the  $i$ -th subject belongs,  $\boldsymbol{\psi}$  is the parameter vector that is common across components,  $\boldsymbol{\theta}$  is the parameter vector that differs across components,  $\boldsymbol{\theta}_{k_i}$  is the parameter value of the  $k_i$  component in  $\boldsymbol{\theta}$ ,  $\boldsymbol{\tau}$  be the hyperparameter value of  $\boldsymbol{\theta}$ , and  $\mathbf{p} = (p_1, \dots, p_L) \sim GD(\mathbf{s}, \mathbf{t})$ ,  $\mathbf{s} = (\gamma/L, \dots, \gamma/L)$ ,  $\mathbf{t} = (\gamma(L-1)/L, \dots, \gamma/L)$ , where GD is the generalized Dirichlet distribution (see for example, Ishwaran and James, 2001).

### Priors

In this study, we assume finite Dirichlet process prior distributions for the error variables:

$$\mathbf{e} \sim DP_L(\gamma, N(\boldsymbol{\mu}, \boldsymbol{\sigma})), \quad (8)$$

$$\beta \sim N(\beta_0, \sigma_\beta^2) \quad (9)$$

### Algorithm and the full conditional distributions

We apply the Blocked Gibbs Sampler proposed by Ishwaran and James (2001). Further, we describe the Blocked Gibbs Sampler for our model setup and introduce the full conditional distributions of each parameter's vector necessary for drawing samples in each iteration.

Let  $\{k_1^*, \dots, k_m^*\}$  be the set of the current  $m$  unique values of  $\mathbf{k}$ . To run the Blocked Gibbs Sampler, we draw parameter values in the following order:

- (1) Generate  $\beta$ : The full conditional-posterior distribution for  $\beta$  with the other given parameters is expressed as follows (We use the notation "... " which indicates that all the other parameters are given.):

$$\begin{aligned} p(\beta|\cdots) &\propto \prod_i^N p(y_i|k_i=l, \beta, \mu_l, \sigma_l) \times p(\beta) \\ &\propto \exp \left[ -\frac{1}{2\sigma_{k_i}^2} \sum_{i=1}^N (y_i - \beta x_i - \mu_{k_i}) - \frac{1}{2\sigma_\beta^2} (\beta - \beta_0)^2 \right] \\ &\propto \exp \left[ -\frac{1}{2B} (\beta - b)^2 \right], \end{aligned} \quad (10)$$

where

$$B = \left( \sum_i^N \sigma_{k_i}^{-2} x_i^2 + \sigma_\beta^{-2} \right)^{-1}, \quad b = B \left\{ \sigma_{k_i}^{-2} \sum_i^N x_i (y_i - \mu_{k_i}) + \sigma_\beta^{-2} \beta_0 \right\}$$

- (2) Generate the mean of the error variables  $\mu$ : The full conditional distribution is as follows. For simplicity, we let  $e_i = y_i - \beta x_i$ .

$$\begin{aligned} p(\mu_l|\cdots) &\propto \prod_i^N p(y_i|k_i=l, \mu_l, \sigma_l^2, \beta) p(\mu_l) \\ &\propto \exp \left[ -\sum_{i:k_i=l}^N \frac{1}{2\sigma_l^2} (e_i - \mu_l)^2 - \frac{1}{2V_{0,l}} (\mu_l - u_{0,l})^2 \right] \\ &\propto \exp \left[ -\frac{1}{2V_{n,l}} (\mu_l - u_{n,l})^2 \right] \end{aligned} \quad (11)$$

where

$$\begin{aligned} V_{n,l} &= (n_l \sigma_l^{-2} + V_{0,l}^{-1})^{-1}, \quad u_{n,l} = V_{n,l} (n_l \sigma_l^{-2} \bar{e}_l + V_{0,l}^{-1} u_{0,l}) \\ n_l &= \sum_{i:k_i=l}^N 1, \quad \bar{e}_l = n_l^{-1} \sum_{i:k_i=l}^N e_i \end{aligned}$$

- (3) Generate the variance of the error variables  $\sigma$ : The full conditional distribution is as follows

$$\begin{aligned} p(\sigma_l|\cdots) &\propto \prod_i^N p(y_i|k_i=l, \mu_l, \sigma_l^2, \beta) p(\sigma_l^2) \\ &\propto \sigma_l^{-(n_l+f_{0,l}+2)/2} \exp \left[ -\frac{1}{2\sigma_l^2} \sum_{i:k_i=l}^N (e_i - \mu_l)^2 \right] \times \exp \left[ -\frac{G_{0,l}}{2\sigma_l^2} \right] \\ &\propto \sigma_l^{-(f_{n,l}+2)/2} \exp \left[ -\frac{G_{n,l}}{2\sigma_l^2} \right], \end{aligned} \quad (12)$$

where

$$f_{n,l} = f_{0,l} + n_l, \quad G_{n,l} = G_{0,l} + \sum_{i:k_i=l}^N (e_i - \mu_l)^2$$



(4) Generate  $\mathbf{k}$  (the components that each sample belongs to):

$$p(k_i | \dots) \sim \sum_{l=1}^L \pi_{li} \delta_l(\cdot) \quad (13)$$

where

$$\pi_{li} = \frac{p_l \sigma_l^{-1/2} \exp \left[ -\frac{1}{2\sigma_l^2} (e_i - \mu_l)^2 \right]}{\sum_{l=1}^L p_l \sigma_l^{-1/2} \exp \left[ -\frac{1}{2\sigma_l^2} (e_i - \mu_l)^2 \right]} \quad (14)$$

(5) Generate  $\mathbf{p}$  (the probabilities of each component membership): The full conditional distribution of  $\mathbf{p}$  is the following generalized Dirichlet distribution.

$$p_l = \prod_{m=1}^{l-1} (1 - V_m) V_l \quad (15)$$

$$V_l \sim \text{Beta}(a_l + M_l, b_l + \sum_{m=l+1}^L M_m) \quad (16)$$

$M_l$  is the number of  $k_i$  that equals  $l$ .

### 3 Model Comparison

Chib (1995) suggested a calculation method for the marginal likelihood from the Gibbs output. Furthermore, Basu & Chib (2003) proposed a calculation method for the marginal likelihood when Dirichlet process mixture modeling was applied. Notice that in this study, it is necessary to perform a model comparison of the joint distribution of  $x$  and  $y$ . For simplicity, we denote  $\mathbf{y}^* = (\mathbf{x}^t, \mathbf{y}^t)^t$  and  $\mathbf{x} = (x_1, \dots, x_N)^t$ ,  $\mathbf{y} = (y_1, \dots, y_N)^t$ . Following Basu & Chib (2003), the logarithm of marginal likelihood is expressed as follows:

$$\log p(\mathbf{y}^*) = \log p(\mathbf{y}^* | \beta^*) + \log p(\beta^*) - \log p(\beta^* | \mathbf{y}^*) \quad (17)$$

While the parameters with the asterisk can be arbitrary values of the parameters, we use posterior means in this study.

According to Chib (1995),  $\log p(\beta^* | \mathbf{y}^*)$  can be calculated as follows:

$$\log \hat{p}(\beta^* | \mathbf{y}^*) = \frac{1}{G} \sum_{g=1}^G \log p(\beta^* | \mathbf{y}^*, \boldsymbol{\mu}^{(g)}, \boldsymbol{\sigma}^{(g)}) \quad (18)$$

where  $G$  is the number of iterations of the MCMC sampler used for parameter estimation. The estimation of  $p(\mathbf{y}^* | \beta^*)$  is rather difficult. Following Basu & Chib (2003), we used the *Sequential Important Sampling* mentioned below (Kong, Liu & Wong, 1994). For notational simplicity, denote  $\boldsymbol{\theta} = (\boldsymbol{\mu}^t, \boldsymbol{\sigma}^t)^t$ .

(a) Generate  $\boldsymbol{\theta}_1^{(g)}, \dots, \boldsymbol{\theta}_N^{(g)}$  from  $p(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N | \mathbf{y}^*, \beta)$

$$p(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N | \mathbf{y}^*, \beta) = \prod_{i=1}^N p(\boldsymbol{\theta}_i | \mathbf{y}_{(i)}^*, \boldsymbol{\theta}_{i-1}, \beta^*) \quad (19)$$

$$\begin{aligned} p(\boldsymbol{\theta}_i | \mathbf{y}_{(i)}^*, \boldsymbol{\theta}_{i-1}, \beta^*) &\propto \frac{\gamma}{\gamma + i - 1} p(y_i^* | \boldsymbol{\theta}_i, \beta^*) p(\boldsymbol{\theta}_i | \boldsymbol{\tau}) \\ &+ \sum_{j=1}^{m_i-1} \frac{n_{j,i-1}}{\gamma + i - 1} p(y_i^* | \boldsymbol{\theta}_i, \beta^*) \delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_i) \end{aligned} \quad (20)$$

where  $\mathbf{y}_{(i)}^* = (\mathbf{y}_1, \mathbf{x}_1, \dots, \mathbf{y}_i, \mathbf{x}_i)$ ,  $y_i^* = (y_i, x_i)$  and  $\boldsymbol{\theta}_{(i)} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_i)$ .  $m_{i-1}$  are unique values in  $\boldsymbol{\theta}_{(i-1)} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{i-1})$ , and  $n_{j,i-1}$  is the frequency of the  $j$ -th unique values.

(b) Calculate  $w$

$$w = w(\boldsymbol{\theta}_1^{(g)}, \dots, \boldsymbol{\theta}_N^{(g)}) = \prod_{i=1}^N p(y_i^* | \mathbf{y}_{(i-1)}^*, \boldsymbol{\theta}_{i-1}^{(g)}, \beta^*) \quad (21)$$

where

$$\begin{aligned} p(y_i^* | \mathbf{y}_{(i-1)}^*, \boldsymbol{\theta}_{i-1}, \beta^*) &= \frac{\gamma}{\gamma + i - 1} \int p(y_i^* | \boldsymbol{\theta}_i, \beta^*) p(\boldsymbol{\theta}_i | \boldsymbol{\tau}) d\boldsymbol{\theta}_i \\ &+ \sum_{j=1}^{m_{i-1}} \frac{n_{j,i-1}}{\gamma + i - 1} p(y_i^* | \boldsymbol{\theta}_i, \beta^*) \end{aligned} \quad (22)$$

The integral in the above equation can be estimated by the usual Monte Carlo method.

(c) Estimate the average of  $w$

$$\bar{w} = \frac{1}{G} \sum_{g=1}^G w(\boldsymbol{\theta}_1^{(g)}, \dots, \boldsymbol{\theta}_N^{(g)}) \quad (23)$$

The likelihood can be obtained via Monte Carlo estimation in the final step.

## 4 Simulation study

We let  $\mathbf{x}$  be the explanatory variable and  $\mathbf{y}$  be the dependent variable, and generated the data set. The error variables are generated from the normal mixture distribution with two components. The true values were  $\boldsymbol{\mu} = (-2.0, 2.0)^t$ ,  $\boldsymbol{\sigma} = (0.5, 2.0)^t$  and  $\beta = 0.5$ . We generated 100 observations and considered two cases. One was a case in which  $\mathbf{x}$  follow normal distributions, and in the other case,  $\mathbf{x}$  follow nonnormal distributions. In the former case, we generated  $\mathbf{x}$  from the normal distribution with mean 0 and variance 10.0. In the latter, we used the following mixture distribution:

$$x_i \sim 0.3 \cdot N(-2.0, 1.0) + 0.7 \cdot N(2.0, 2.0) \quad (24)$$

For each case, we changed the explanatory and dependent variables, and then analyzed these data sets. After employing 2,000 burn-in iterations, we employed 3,000 MCMC iterations and calculated the marginal likelihoods. The results are presented in Table 1.

**Table 1.** Results of the model comparison by marginal likelihood (The numbers of times that the models were selected are mentioned in parentheses)

data generating model	(1) $y_i = \beta x_i + e_i$ ( $x$ is normal)	(2) $y_i = \beta x_i + e_i$ ( $x$ is nonnormal)	(3) $y_i = \beta x_i^3 + e_i$
pair of models # 1	$y_i = \beta x_i + e_i$ ( <b>87</b> ) $y_i = \beta' x_i^3 + e'_i$ ( <b>13</b> )	$y_i = \beta x_i + e_i$ ( <b>78</b> ) $y_i = \beta' x_i^3 + e'_i$ ( <b>22</b> )	$y_i = \beta x_i^3 + e_i$ ( <b>85</b> ) $y_i = \beta' x_i + e'_i$ ( <b>15</b> )
pair of models # 2	$y_i = \beta x_i + e_i$ ( <b>92</b> ) $x_i = \beta'' y_i + e''_i$ ( <b>8</b> )	$y_i = \beta x_i + e_i$ ( <b>72</b> ) $x_i = \beta'' y_i + e''_i$ ( <b>28</b> )	$y_i = \beta' x_i + e'_i$ ( <b>42</b> ) $x_i = \beta''' y_i + e'''_i$ ( <b>58</b> )

## 5 Conclusion

In this study, we proposed a method of determining the direction of path with a Bayesian semiparametric model. In simulation study, we used a simple linear regression model, and set two Dirichlet process mixture models wherein the explanatory and dependent variables are alternated with each other. We then searched which model is better by calculating the marginal likelihood. In Table 1, the number of times of selecting the true direction of path is less than five out of ten for the data set generated from the nonlinear regression model ((3)). Since in real data analysis it is always possible that the true model is different from both of the models wherein the explanatory and dependent variables are alternated with each other, this result implies that it is meaningless to identify the causation by determining the direction of the path using nonnormal error variables.

## References

- ANSARI, A. and IYENGAR, R. (2006): Semiparametric thurstonian models for recurrent choices: A Bayesian analysis. *Psychometrika*, 71(4), 631-657.
- BASU, S. and CHIB, S. (2003): Marginal likelihood and Bayes factors for Dirichlet Process mixture models. *Journal of the American Statistical Association*, 98, 224-235.
- BENTLER, P. M. (1983): Some contributions to efficient statistics in structural models: specification and estimation of moment structures. *Psychometrika*, 48, 493-517.
- BROWNE, M. W. (1984): Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62-83.
- CHIB, S. (1995): Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90, 1313-1321.
- ESCOBAR, M. D. (1994): Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89, 268-277.
- FERGUSON, T. S. (1973): A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209-230.
- HOSHINO, T. (accepted for publication): Dirichlet process mixtures of structural equation modeling and direct calculation of posterior probabilities of the numbers of components. *Psychometrika*.

- ISHWARAN, H. and JAMES, L. F. (2001): Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96, 161-173.
- ISHWARAN, H. and ZAREPOUR, M. (2000): Markov Chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87, 371-390.
- KANO, Y., BERKANE, M. and BENTLER, P. M. (1993): Statistical inference based on pseudo-maximum likelihood estimators in elliptical populations. *Journal of the American Statistical Association*, 88, 135-143.
- KONG, A., LIU, J. S. and WONG, W. H. (1994): Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89, 278-288.
- SETHURAMAN, J. (1994): A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.
- SHIMIZU, S. and KANO, Y. (2008): Use of non-normality in structural equation modeling: application to direction of causation. *Journal of Statistical Planning and Inference*, 138, 3483-3491.

# Data Visualization and Aggregation

Junji Nakano<sup>1</sup> and Yoshikazu Yamamoto<sup>2</sup>

<sup>1</sup> The Institute of Statistical Mathematics

Tachikawa, Tokyo, Japan, *nakanoj@ism.ac.jp*

<sup>2</sup> Tokushima Bunri University

Takamatsu, Kagawa, Japan *yamamoto@is.bunri-u.ac.jp*

**Abstract.** Visualizing data using interactive and dynamic graphics is a useful first step of statistical data analysis, especially when the data are new to the analyst and the amount of them is very large. Recently, several data are collected by automatic data acquisition systems over networks, and become so huge that even high-speed computers require considerable time to draw interactive graphics that show all the observations. Therefore, we sometimes “aggregate” data by grouping them appropriately to reduce the amount of data without losing the information of the original data too much.

There exist several data aggregation techniques. Symbolic data analysis expresses a group of data as a “concept”, a second level data described by variables which take complicated values such as intervals and histograms. In relational database techniques, online analytical processing (OLAP) is extensively used to calculate the summation of variable values by interactively grouping the data.

Data usually contain both categorical and real valued variables. It is not easy to express the structure of such data clearly by traditional statistical graphics.

We propose an interactive and flexible aggregation of groups of data which are induced mainly by the values of categorical variables. The aggregation result is expressed by several graphics components, such as a dot, a boxplot and a histogram on usual graphics. We demonstrate an aggregation and visualization system which include extended parallel coordinate plot and scatter diagram. Simple example shows that an aggregation is a powerful visualization tool to reveal the structure of complex data.

**Keywords:** OLAP, parallel coordinate plot, symbolic data analysis

## 1 Introduction

When we have a new large data set, it is impossible to grasp the characteristics of the data set by looking at variable values expressed by numbers and characters. Instead of checking each observation in the huge data table, we usually calculate some basic statistics of variables such as means and variances, and drawing traditional statistical graphics such as scatter diagrams and barcharts. They are, however, not sufficient for expressing the structure of huge amount of complicated data clearly. Several new graphical methods have been developed with the help of powerful computers. They are sometimes called visual data mining techniques.

We note that statistical graphics are roughly divided into two categories. One category of graphics expresses the values of each observation by using an object.

This category includes a scatter diagram and a parallel coordinate plot (Inselberg, 1985). Another category expresses several summarized statistics as objects in graphics. Examples in this category are a boxplot and a histogram. When the number of data is large, graphics in the first category become very complicated and difficult to see, but the graphics in the second category does not change so much.

It often happens that we have more interest in groups of data than each observation. Consider sales data of convenience stores. Original observations are a huge amount of receipts, each of which records one individual purchase. We are often interested in the sales of a shop, shops in districts, week, month etc., not in each receipt itself. We sometimes hope to change the definitions of groups in which aggregation operations are performed. In the database system, it is realized by on-line analytical processing (OLAP) technique, which is as same as the pivot table operations in Microsoft Excel. For the convenience store example, we can get useful information of sales data by aggregating data based on specified shop, region, and time, etc., and showing the results of aggregation by tables and business graphs. OLAP provides interactive user interface to specify the unit of the aggregation dynamically and has several functions to perform flexible aggregation operations for calculating sum, average and median etc. Several new user interfaces has proposed by, for example, Stolte, Tang, and Hanrahan (2002) and Techapichetvanich and Datta (2005).

Symbolic data analysis (SDA) also handles a set of groups of original individual observations (Billard and Diday, 2006). In SDA, values of a variable can be more complex than the traditional data such as real numbers and categorical values. Typical symbolic data can take intervals, histograms and barcharts as variable values. We note that such values most often appear when we think a set of original observations as a new observation, which is called “concept” in SDA. As a concept expresses characteristics of a set of observations, the values of a variable in a concept should be expressed by one or more aggregated values. Consider the case where the original variable takes real values. One of the simplest summarizations is an average of all the values. A little complicated summarization is the interval which consist of the minimum and maximum values. If we use a histogram of the variable, it expresses more information of original data than the average and the interval. If we use symbolic data derived from original observations by aggregation, we may keep necessary information of original data with less data amount.

This paper considers aggregation in data visualization. In the next section, we illustrate some techniques of modern statistical graphics. Symbolic data is explained from the view point of aggregation in Section 3. Then we give a brief introduction to OLAP aggregation techniques in Section 4. In Section 5, we explain our aggregation techniques experimentally realized by an extension of a parallel coordinate plot and a scatter diagram. In the last section, we give a few concluding remarks.

## 2 Modern statistical graphics

We usually draw statistical graphics by a computer in these days. This gives several new means to statistical graphics compared with the past time when statistical graphics were drawn on paper by human hands. As a computer can redraw graphics easily, it is possible to redraw statistical graphics many times by changing the conditions. By using this ability, animation or dynamic graphics can be realized and

interactive operations such as selection, highlighting, zooming and linked graphics are implemented and extensively used. All these operations are known to be useful for revealing the structure of huge amount of data.

We note here that all these interactive operations have close relations with making a group of observations. Selection operation specifies a group of observations by surrounding objects which represent each observation (for example, a dot in a scatter diagram) or a group of observations (for example, a bar in a histogram). Zooming operation shows a graphics for a selected part of data, and highlighting operation draws the objects for selected data in emphasized way, for example, thickly or by using another color. Linked graphics is highlighting on several graphics for a selected group of data. Thus, modern interactive operations of statistical graphics are useful mainly because they focus on a specific group of data.

Traditional statistical variables are divided into some categories: continuous real valued, discrete real valued, ordered categorical and non-ordered categorical variables. We know that real valued variables and categorical variables sometimes should be considered differently in visualization and statistical analysis. When we consider one variable, the distribution of variable values is important. The distribution of a real valued variable is described by histogram sufficiently, a boxplot to some extent, and a mean with a variance awkwardly. The distribution of a categorical variable is described by barchart and piechart. When we consider two variables together, relation between two variables are of interest. If both variables are real valued, scatter diagram is the best graphics to show the relation. If both variables are categorical, a contingency table has all the information of the data, and a mosaic plot can directly show the information of the contingency table. If we consider three variables, 3 dimensional graphics may be helpful by using the 3 dimensional version of 2 dimensional graphics. However, more than 3 variables are difficult to be described by traditional statistical graphics. A parallel coordinate plot is useful if all the variables are real valued, and a mosaic plot can describe all the information of a multiple contingency table. These static graphics, however, are not easy to see if the number of variables and observations are large. To reduce the difficulty, we can use dynamic and interactive operations using several graphics together. Linked graphics is useful when some variables are real valued and others are categorical.

### 3 Symbolic data as aggregation results

Symbolic data analysis provides statistical methods to treat symbolic data (Billard and Diday, 2006). Symbolic data are described by variables which take real values and categorical values like traditional statistics, and can also take more complicated variables whose values are intervals, histograms, etc. This type of data often arises by aggregation operations.

We consider “aggregation” as an operation to summarize a group of observations by several statistics, including the case where the result is given by a single number such as a mean or median value. In symbolic data analysis, we sometimes consider new “concept”, which is a group of individual data. Consider that we have some measurements for many birds and are interested in the difference of species of birds. In this situation, one of species, for example, swallow is a concept which include many swallow observations. If a variable “weight” is measured for each observation, the value of the weight variable of a concept swallow can be expressed in

many ways. Traditional statistics typically use the mean of observations. In SDA, we may use the interval given by the range of values or histogram of the variable. It is clear that an interval has more information than a mean value, and the histogram has more information than the interval. On the other hand, a mean requires just one number to record, and an interval requires two numbers, and histogram requires more numbers. Note that the required number to record is considerably small compared with those of the original observations. Thus, aggregation is thought to be one of data reduction techniques. It is clear that one advantage of SDA is to emphasize the importance of expressing aggregation results by several numbers.

## 4 Aggregation in database by OLAP

Database is a computer technology to store large amount of data as compactly as possible in unified form, and to retrieve the required data as quickly as possible. The most famous database model is the relational database, which expresses data in the form of one or more tables. In the actual usage of database, the amount of data becomes very large. Therefore, it is impossible to capture the whole image of data by looking at these tables directly. Online analytical processing (OLAP) is a technology to capture the data structure as a whole by aggregating data from various ways of grouping data. In OLAP, groups are formed according to the values of the categorical variables, and simple aggregation functions such as summation and average are performed in each group of data. The relational database theory considers that data and results of operations are all expressed by tables. A resulted table can be also expressed by business graphs such as a barchart, a piechart and a linechart.

In the database theory, stochastic variations in the observations are not considered explicitly. On the other hand in Statistics, the mechanism of the data generation is explicitly described by using stochastic structure. Usual statistical models are written by the mixture of main mechanism which describes relations among variables and stochastic mechanism which describes inevitable error terms. Statistical analysis mainly focus on the main data generation mechanism and has less interest in each observation values. As OLAP technology performs aggregation for several groups and see the global structure of the data, it has similar role to statistical analysis.

## 5 Graphics for aggregation

Based on the considerations given in former sections, we conclude that the aggregation operations is important in the visualization of a large data set especially for defining appropriate groups. It is important to have groups which contain enough information of original individual data and reduce the amount of recording space for further analysis. When we have little prior information for a given data set, such groups can be obtained by trial and error. Our visualization system are designed to make this operation easy.

We propose to show aggregation results on the same framework as graphics for expressing the values of each observation, such as a scatter diagram and a parallel



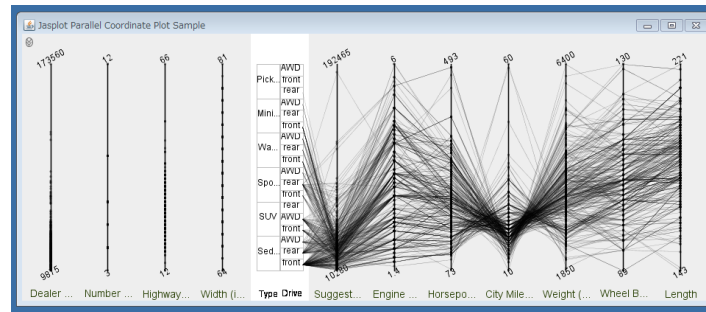


Fig. 1. Displaying all observations.

coordinate plot, with some modifications. Original data and grouped data are easily compared on the same framework.

As a parallel coordinate plot is more suitable to express a large number of variables than a scatter diagram, we mainly focus on a parallel coordinate plot in this section. A parallel coordinate plot (PCP) expresses the same information as the table of real data except the order of observations (Inselberg, 1985). Each observation is expressed by a polygonal line which connects all the variable axes placed in parallel. When a variable takes nominal values, they are sometimes transformed to discrete values such as integers to be plotted on one axis as same as the real valued variable. However, as nominal values have usually no definite order among them, we have difficulty how to transform a nominal value to a real value. The textile plot (Kumasaka and Shibata, 2008) is an extension of PCP to handle this transformation reasonably.

Transformed nominal values are useful to select observations which take a specific value of the nominal variable. In some software products, we use different colors for differently selected groups to clarify the difference among groups.

We have modified a parallel coordinate plot for realizing flexible aggregation operation. Our user interface enables us to perform OLAP like group specification. The results of aggregation are expressed by visual objects which illustrate SDA like complex values. We have used Jasplot software (Nakano, Yamamoto and Honda, 2008) to realize our experimental software. Fig.1 is an example to show the individual data of the 2004 Cars Data (Unwin, Theus and Hofmann, 2006) as a usual parallel coordinate plot.

For dividing raw data into groups, our system puts variable axes in three divided areas: an ignored variables area, an aggregation specification area and a data description area, from the left to right in Fig. 1. Variables can be moved to or from each area freely by drag-and-drop operation using a mouse.

Variable axes placed in the ignored variables area are ignored from drawing polygonal lines. Such variables are out of consideration in the data analysis.

The aggregation specification area is a place where variables are used to specify groups. Nominal variables are mainly put here as stacked boxes. One nominal variable can specify groups. Original observations that take the same categorical value consist one group. Groups can be also composed of two or more nominal variables like OLAP. Units of the aggregation can be specified by the Cartesian product of categorical variables. The way of expression in this area is similar to Double-decker

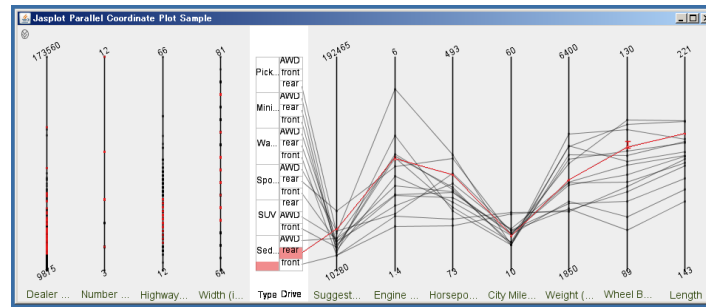


Fig. 2. Displaying means of groups.

plot (Unwin, Theus and Hofmann, 2006) from left to right. In Fig. 1, on each 6 values of variable **Type**, 3 values of variable **Drive** are stacked and the total number of groups is 18 (3 of them contain no observation). We can use real valued variable in this area if we divide them into several intervals and treat them as ordered categorical values.

In the data description area, observations is originally displayed just like a usual parallel coordinate plot (Fig. 1). When we perform aggregation based on the group specified in the aggregation specification area, just aggregated data are displayed and raw data disappeared. Default aggregation results are the means of groups. In Fig. 2, they are shown by each group, and one group, where **Type** is **Sedan** and **Drive** is **rear**, is selected. Other available one dimensional results are summations and medians. In addition, more detailed aggregation results are available as boxplot or histogram on each axes (Fig. 3). It is clear that the visibility becomes worse when we draw all the resulted graphics for all groups at the same time. Thus, we can select groups whose resulted graphics are shown on the axes. Our system has multiple selectors and can specify several groups at the same time. Example is shown in Fig. 4, where similar way of expression of groups are used on a scatter diagram. Left graphics of Fig. 4 is a usual scatter diagram of variables **Length** and **Horsepower**. We aggregate them as same groups as Fig. 2 and show means of groups in the right figure of Fig. 4. Superimposed histograms express more detailed distribution information for two groups: **Type = Sedan** and **Drive = rear** group is shown in red color and **Type = Sedan** and **Drive = AWD** group is shown in blue color. We express the location information (mean) of each group as it is, and scales of histograms are reduced (1/10 here) for ease of perception. Distributions of two variables of two groups are compared visually.

## 6 Concluding remarks

If the number of observations is huge, even a high-speed computer requires considerable time to perform interactive operations on graphics. The aggregation operation can be one remedy for this trouble. Aggregation means to summarize the information in one group of data by the less number of values than all the variable values for observations. Summarized information is expressed by explores and histograms in the case of real valued variables in our implementation.

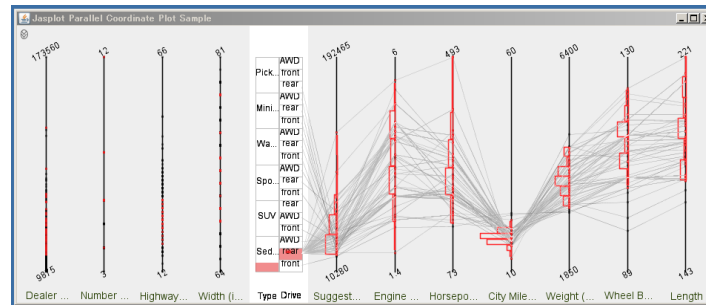


Fig. 3. Displaying histograms of a group.

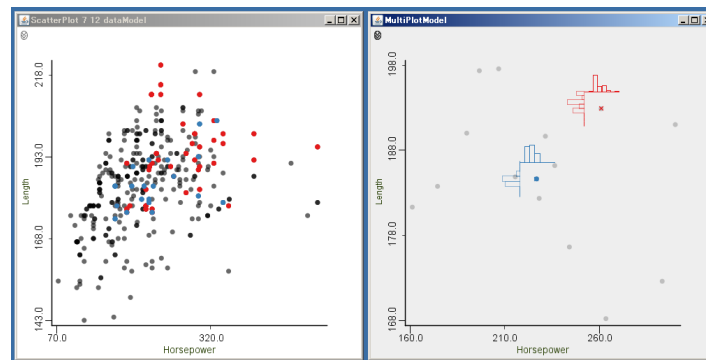


Fig. 4. Displaying a scatter diagram of data and a extended scatter diagram of groups.

Our experimental implementation of an extended parallel coordinate plot and a scatter diagram seems to be useful as an extended OLVA. Obtained groups can be analyzed by using symbolic data analysis techniques.

## References

- BILLARD, L. and DIDAY, E. (2006): *Symbolic Data Analysis: Conceptual statistics and data mining*. John Wiley, New York.
- INSELBERG, A. (1985): The plane with parallel coordinates. *The Visual Computer* 1, 69–91.
- KUMASAKA, N. and SHIBATA, R. (2008): High-dimensional data visualisation: The textile plot. *Computational Statistics & Data Analysis* 52 (7), 3616–3644.
- NAKANO, J., YAMAMOTO, Y. and HONDA, K. (2008): Promming statistical data visualization in the Java language. In: Chen, C-H., Hädle, W. and Unwin, A. (Eds.): *Handbook of Data Visualization*. Springer, Berlin, 725–756.
- STOLTE, C., TANG, D. and HANRAHAN, P. (2002): Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE transactions on visualization and computer graphics* 8 (1), 52–65.

- TECHAPICHETVANICH, K. and DATTA, A. (2005): Interactive Visualization for OLAP. In: Gervasi, O., Gavrilova, M. L., Kumar, V., Lagana, A., Lee, H. P., Mun, Y., Tanir, D. and Tan, C. J. K. (Eds.): *Computational Science and Its Applications - ICCSA 2005*. Springer, Berlin, 206–214.
- UNWIN, A., THEUS, M. and HOFMANN, H. (2006): *Graphics of Large Datasets: Visualizing a Million*. Springer, Berlin.

# Longitudinal Data Analysis Based on Ranks and Its Performance

Takashi Nagakubo<sup>1</sup> and Masashi Goto<sup>2</sup>

<sup>1</sup> Asubio Pharma Co., Ltd., Clinical Research & Development Department  
Orix Akasaka 2-Chome Building 3F, 2-9-11 Akasaka, Minato-ku, Tokyo  
107-8541, Japan  
*nagakubo.takashi.cw@asubio.co.jp*

<sup>2</sup> Biostatistical Research Association, NPO.  
2-22-10-A411 Kamishinden, Toyonaka-shi, Osaka 560-0085, Japan  
*info@bra.or.jp*

**Abstract.** In this study, we examine data measured repeatedly for a single subject over time, which is called longitudinal data. We propose the rank empirical distribution (RED) method, a method that is not reliant on distribution. We examine a case study in which the RED method produces outcomes different from those of repeated measures ANOVA. We then conduct simulations in which the underlying distribution is expected to be normal or skewed, and investigate differences in the power of the two methods. The results of the simulations were that the power of the RED method was higher than repeated measures ANOVA for skewed data. This suggests that the RED method is useful for longitudinal data analysis.

**Keywords:** repeated measures, cumulative distribution function, relative effect

## 1 Preface

Manipulation of quantitatively measurable responses is the primary concern when analyzing longitudinal data. In particular, when responses are obtained as quantitative values and those values can be assumed to describe a normal distribution and covariance structure is compound symmetry, repeated measures ANOVA is commonly used (Winer *et al.*, 1991). Benefits of this method include that it makes possible a quantitative evaluation of the main effects and interactions that are the factors of data variance, and furthermore allows for a simple interpretation of the effects. However, even if responses are obtained quantitatively, it does not necessarily follow that they will describe a normal distribution. It is therefore necessary to perform a follow-up investigation to determine the validity and appropriateness of using a parametric approach such as repeated measures ANOVA. As a way to weaken the requirements of such parametric approaches, we would like to consider the use of a method that utilizes rank and does not depend upon distribution.

Brunner & Puri (1996) defined relative effects to describe treatment effects in general nonparametric designs (for a review of such methods, see Brunner & Puri (2001)). Such effects are called nonparametric effects or relative effects because they are evaluated in relation to the average distribution of all measurements effects in the study. Previously, it has been difficult to display the results of nonparametric

methods using graphs and the like as an aid to interpretation, but the results of relative effects lend themselves well to graphic displays. Relative effects are drawn from empirical distributions and inferred by the rank of observations; accordingly, approximation using relative effects is called the rank empirical distribution (RED) method. The RED method is performed with regard to the rank of all observations, and thus is robust and invariable under strict monotonic transformations of the data. This method can also be applied to not only measured but also ordinal values.

This paper proposes the RED method of relative effects based on a ranking of longitudinal data with multiple groups, and evaluates the method's performance in case studies and simulations.

## 2 Relative effects

This paper considers the case where longitudinal data are composed of observations taken from multiple groups. Let  $I$  denote the number of groups, and  $T$  denotes the number of observations, with subject  $k$  ( $k = 1, \dots, n_i$ ) of group  $i$  ( $i = 1, \dots, I$ ) being observed at time  $t$  ( $t = 1, \dots, T$ ), resulting in observation  $X_{ikt}$ . Let  $n = \sum_{i=1}^I n_i$  denote the number of subjects from all groups, and  $N = nT$  denotes the number of all observations. This is the most frequently used model for longitudinal data. Marginal distributions are determined according to the group and time, and observation  $X_{ikt}$  will conform to the marginal distribution  $F_{it}$ .

In the two sample case, nonparametric effects are considered as measures of whether one observation is larger, smaller, or the same as another observation. If  $Y_1$  and  $Y_2$  are independent random variables with distributions  $F_1$  and  $F_2$ , respectively, then a nonparametric effect of  $Y_2$  with respect to  $Y_1$  (relative effect) may be defined as

$$p = \Pr(Y_1 < Y_2) + \frac{1}{2}\Pr(Y_1 = Y_2) = \int F_1 dF_2$$

(Mann & Whitney, 1947). This effect is the probability that a random variable under some distribution function is larger than a random variable under another distribution function. Applying this nonparametric effect to the longitudinal data format described in the previous section, we have the effect of group  $i$  at time  $t$  defined as

$$p_{it} = \int H dF_{it} \quad (1)$$

(Brunner & Puri, 1996). Where,  $F_{it}$  is the marginal distribution of group  $i$  at time  $t$ , and the weighted average of the marginal distribution  $F_{it}$  is

$$H = \frac{1}{N} \sum_{i=1}^I \sum_{t=1}^T n_i F_{it}.$$

This effect is evaluated as a comparison of the size of the effect of group  $i$  at time  $t$  with the distribution function  $H$ , which is the average of all the distribution functions, and is the so-called relative effect.

In the nonparametric setup introduced above, hypotheses are formulated by using distribution functions  $F_{it}$ . Let  $\mathbf{F} = (F_{11}, \dots, F_{IT})^T$  denote the vector of the

distribution functions and let  $\mathbf{C}$  denote a contrast matrix. Then, nonparametric hypotheses in their most general form are written as  $H_0^F : \mathbf{C}\mathbf{F} = \mathbf{0}$ . For mixed models, such hypotheses have been introduced by Akritas and Arnold (1994) and have been further developed and discussed by Akritas *et al.* (1997).

### 3 Statistics

#### 3.1 Estimation of relative effects

Using the relationship between the rank and the empirical distribution function shown in Eq. (1), the relative effect of group  $i$  at time  $t$  is inferred from

$$\begin{aligned}\hat{p}_{it} &= \int \hat{H} d\hat{F}_{it} = \frac{1}{n_i} \sum_{k=1}^{n_i} \hat{H}(X_{ikt}) \\ &= \frac{1}{N} (\bar{R}_{i \cdot t} - \frac{1}{2}).\end{aligned}\quad (2)$$

Here,  $\bar{R}_{i \cdot t}$  is the mean of the ranks in the  $i$ th group at time  $t$  and  $R_{ikt}$  is the rank of  $X_{ikt}$ . To test the hypothesis  $H_0 : \mathbf{C}\mathbf{F} = \mathbf{0}$ , it is necessary to consider the so-called rank version of the Wald type statistic. The distribution of  $Q_N(\mathbf{C}) = N\hat{\mathbf{p}}\mathbf{C}^T[\mathbf{C}\hat{\mathbf{V}}_N\mathbf{C}^T]^{-1}\mathbf{C}\hat{\mathbf{p}}$  converges extremely slowly to the  $\chi^2$ -distribution. Thus, we use an ANOVA type statistic. The idea is simply to leave out the estimated covariance matrix in  $Q_N(\mathbf{C})$ , and then to consider the asymptotic distribution of the statistic  $Q_N^*(\mathbf{C}) = N\hat{\mathbf{p}}\mathbf{T}\hat{\mathbf{p}}$  where  $\mathbf{T} = \mathbf{C}^T[\mathbf{C}\mathbf{C}^T]^{-1}\mathbf{C}$ . Under  $H_0$ , this statistic has, asymptotically, the distribution of a weighted sum  $\sum_{i=1}^I \lambda_i Z_i$  of independent  $\chi_1^2$ -random variables  $Z_i \sim \chi_1^2$  where  $\lambda_1, \dots, \lambda_I$  are the eigenvalues of  $\mathbf{T}\hat{\mathbf{V}}_N\mathbf{T}$ . According to Box (1954), this distribution is approximated by a scaled  $\chi^2$ -distribution  $g \cdot \chi_f^2$  such that the first two moments of  $\sum_{i=1}^I \lambda_i Z_i$  and  $g \cdot \chi_f^2$  coincide. If  $\text{trace}(\mathbf{T}\hat{\mathbf{V}}_N) > 0$ , then the statistic

$$F_N(\mathbf{C}) = \frac{Q_N^*(\mathbf{C})}{\hat{g}\hat{f}} = \frac{N}{\text{trace}(\mathbf{T}\hat{\mathbf{V}}_N)} \hat{\mathbf{p}}^T \mathbf{T} \hat{\mathbf{p}} \quad (3)$$

has, approximately, a central  $F(\hat{f}, \infty)$ -distribution under  $H_0$ , where  $\hat{f} = [\text{trace}(\mathbf{T}\hat{\mathbf{V}}_N)]^2 / \text{trace}(\mathbf{T}\hat{\mathbf{V}}_N\mathbf{T}\hat{\mathbf{V}}_N)$ . For the derivation of this result, we refer to Brunner *et al.* (1999).

### 4 Evaluation of RED method

In this section, we apply the RED method to an actual example to evaluate the features of the RED method. Repeated measures ANOVA is used for the target comparison.

#### 4.1 Application to case study

Let us now attempt to apply the RED method to an actual case study. The case involves data related to  $\gamma$ -GTP in patients undergoing cholecystectomy due to cholelithiasis (Brunner *et al.*, 2002). Significance levels of 0.05 are used.  $\gamma$ -GTP was

**Table 1.** Results for  $\gamma$ -GTP data

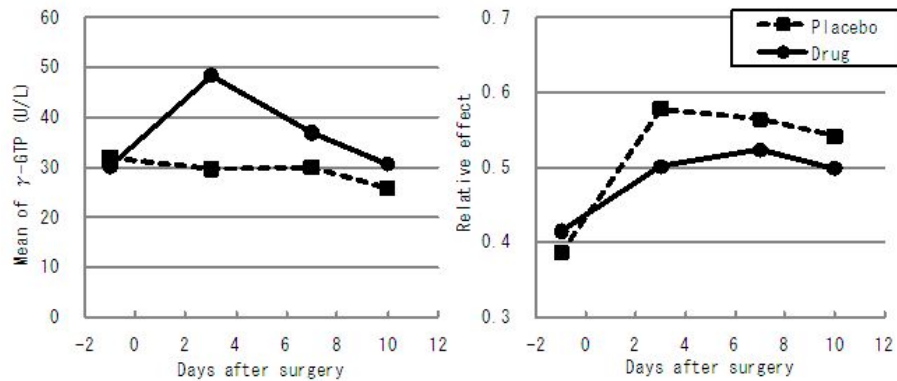
RED method				Repeated measures ANOVA			
Factor	df	F-value	p-value	Factor	df	F-value	p-value
Group	1.00	0.227	0.6338	Group	1	0.670	0.4173
Time	1.77	8.471	0.0004	Time	3	1.477	0.2234
Group $\times$ Time	1.77	0.930	0.3845	Group $\times$ Time	3	1.280	0.2836

measured prior to and after surgery in patients with cholelithiasis. Subjects were 50 cholecystectomy patients, of whom 26 were administered medication and 24 were administered a placebo.  $\gamma$ -GTP was measured at four times: the day before the operation, and 3, 7, and 10 days after the operation. The aim of the study was to see whether the medication worked faster than the placebo to lower  $\gamma$ -GTP levels, as well as whether there were pre- and post-operation differences in  $\gamma$ -GTP levels.

The means and the relative effects for each group are shown in Fig. 1. Means in the drug group show an initial post-operation increase, followed by a decrease. For relative effects, both groups show an increase in  $\gamma$ -GTP post-operation, followed by a gradual decline. The results for repeated measures ANOVA and the RED method are shown in Table 1. No significant group effect, time effect, or group  $\times$  time interaction was indicated by repeated measures ANOVA. Under the RED method, on the other hand, while group  $\times$  time interaction was not significant, indicating no effect in early lowering of  $\gamma$ -GTP levels due to the drug, a time effect was indicated for changes in the  $\gamma$ -GTP value due to the operation.

## 4.2 Simulations

**Motivation and aims** In the previous section, we applied the RED method and repeated measures ANOVA to a case study. Two methods show dissimilar results. This can likely be attributed to differences in potential distributions. We will next

**Fig. 1.** Transition of  $\gamma$ -GTP



conduct a simulation to evaluate the power related to data containing a latent distribution.

**Design** We take the following as a model for the simulation:

$$X_{ikt} = \mu_i + \tau_t + (\mu\tau)_{it} + e_{ikt}. \quad (4)$$

Where,  $\mu_i$  is the group effect,  $\tau_t$  is the time effect, and  $(\mu\tau)_{it}$  is the group  $\times$  time interaction.  $e_{ikt}$  is the error term of the  $T$  random distribution of the covariance matrix  $\mathbf{V}_{ik} = (v_{rc}), (r, c = 1, \dots, T)$ , here  $v_{rc} = \sigma_e^2 \rho^{|r-c|}$ . This covariance matrix has first-order autoregressive structure. The number of groups and the number of times are taken as  $I = 2, T = 4$ , as per the case examined in the previous section. When the sizes of the group effects, time effects, and group  $\times$  time interactions are fixed as  $\boldsymbol{\mu} = (0, 1)^T, \boldsymbol{\tau} = (0, 1, 2, 3)^T$ , and

$$(\boldsymbol{\mu}\boldsymbol{\tau}) = \begin{pmatrix} 0 & -1 & -2 & -3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

with an error variance  $\sigma_e^2 = 16$  and coefficient of correlation  $\rho = 0.6$ , the number of cases required for the power of repeated measures ANOVA to exceed 80% for all hypotheses is  $n_i = 30$ . For  $n_i = 20$ , retained power is 60%, and for  $n_i = 40$  power is 90%. Sample sizes per group are therefore set at  $n_i = 20, 30, 40$  with error variances of  $\sigma_e^2 = 16, 25, 36$  and coefficients of correlation  $\rho = 0.4, 0.6, 0.8$ . Four standards for distribution are set, namely, power normal distributions with transformation parameters  $\lambda = 0, 0.5, 1$  (Goto *et al.*, 1983) and a normal distribution. Here, the power normal distributions  $\lambda = 0, 0.5, 1$  are a logarithmic normal distribution, a square root normal distribution, and a truncated normal distribution, respectively.

**Comparison of power** We compare the power of the RED method and repeated measures ANOVA, performing 10,000 iterations to test to a 0.05 level of significance of the group effect, time effect, and group  $\times$  interaction. To estimate the power for each method, we divided the number of times the null hypothesis was rejected by the number of iterations. We also used ANOVA to determine whether the method, distribution, sample size, correlation coefficient, or error variance has an effect on the power of group effects, time effects, or group  $\times$  time interaction.

**Results and findings** The power from the simulation for  $n_i = 30, \sigma_e^2 = 16$  are shown in Table 2. The ANOVA results from the simulation of power related to group effect show a 0.05 level of significance for all factors. The error variance had a particularly high contribution of 27.6%. Among the interactions between method and other factors, method  $\times$  distribution made the highest contribution of 6.5%. This suggests that the power of group effects under the RED method and repeated measures ANOVA differs according to the latent distribution of the data. As correlation coefficients increased, the power of group effects decreased. In the case where the latent distribution is a normal distribution and  $\lambda = 1$ , the power of repeated measures ANOVA for group effects was slightly higher than that of RED method. For a latent distribution with  $\lambda = 0, 0.5$ , the power of the RED method for

**Table 2.** The power of RED method and repeated measures ANOVA ( $n_i = 30, \sigma_e^2 = 16$ ).

Distribution	$\rho$	Group		Time		Group $\times$ Time	
		RM-ANOVA	RED	RM-ANOVA	RED	RM-ANOVA	RED
Normal	0.4	0.947	0.937	0.494	0.466	0.489	0.462
	0.6	0.878	0.864	0.586	0.553	0.587	0.547
	0.8	0.775	0.758	0.807	0.770	0.803	0.758
Power normal	$\lambda = 0$	0.4	0.950	1.000	0.501	0.908	0.508
		0.6	0.906	1.000	0.579	0.945	0.580
		0.8	0.834	0.999	0.728	0.992	0.733
	$\lambda = 0.5$	0.4	0.948	0.987	0.485	0.579	0.484
		0.6	0.881	0.953	0.576	0.666	0.566
		0.8	0.782	0.882	0.787	0.855	0.778
	$\lambda = 1$	0.4	0.986	0.981	0.607	0.561	0.612
		0.6	0.952	0.943	0.703	0.657	0.700
		0.8	0.882	0.874	0.896	0.850	0.890

group effects was greater than that under repeated measures ANOVA, and showed an even larger trend where  $\lambda = 0$ .

For power related to time effects, the ANOVA results were significant to a 0.05 level for all factors except for method  $\times$  sample size interaction, method  $\times$  correlation coefficient interaction, sample size  $\times$  error variance interaction, and correlation coefficient  $\times$  error distribution interaction. As with power related to group effects, the contribution of the error variance was high, at 23.5%. Among interactions between method and other factors, the contribution of method  $\times$  distribution was high of 16.2%. Power related to time effects increased as the correlation coefficient became larger, indicating that its effect is greater than the effect of the correlation coefficient with regard to the power of group effects. In the case where the latent distribution is a normal distribution and  $\lambda = 1$ , the power of repeated measures ANOVA for time effects was slightly higher than that of the RED method. In the case where  $\lambda = 0, 0.5$ , the power of the RED method for time effects was considerably greater than that of repeated measures ANOVA.

The ANOVA results for the simulation of group  $\times$  time interaction were significant to a level of 0.05 for all factors except method  $\times$  sample size interaction, method  $\times$  correlation coefficient interaction, sample size  $\times$  correlation coefficient interaction, and correlation coefficient  $\times$  error variance interaction. The contribution of the error variance was highest, at 23.5%. The power of group  $\times$  time interaction was the same as power for time effects.

The RED method also showed almost identical power as repeated measures ANOVA for power of group effects, time effects, and group  $\times$  time interactions in cases where data followed a normal distribution, as well as in cases where data followed a distorted distribution power was clearly higher than repeated measures ANOVA. Therefore, use of the RED method is appropriate in cases where the distribution of observations is unknown and data cannot be assumed to follow a normal distribution.

## 5 Conclusion

This paper has raised for discussion the RED method, a relative effects-based method for analyzing longitudinal data. We presented rank-based inferences and asymptotic distributions for relative effects, which are the basis of the RED method. We applied the RED method to an actual case with longitudinal data, and compared the analysis results with those of repeated measures ANOVA. Using relative effects, we illustrated results using a nonparametric method framework. The results showed a case of differing results between repeated measures ANOVA and the RED method for testing time effects. Some differences between the test results of group effects and group  $\times$  time interactions were also demonstrated. These results are assumed to be due to differences in the latent distribution of the data. To verify these results, power was evaluated through simulation of several data sets following a latent distribution. The results indicated that when the latent distribution of data is skewed, the RED method retained a higher power than did repeated measures ANOVA for all effects. Furthermore, in situations where data followed a normal distribution, the power of the RED method was almost the same as that of repeated measures ANOVA, indicating the robustness of the RED method.

## References

- AKRITAS, M. G. and ARNOLD, S. F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. *J. Amer. Statist. Assoc.* 89, 336-343.
- AKRITAS, M. G., ARNOLD, S. F. and BRUNNER, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *J. Amer. Statist. Assoc.* 92, 258-265.
- ARNOLD, S. F. (1981). *The Theory of Linear Models and Multivariate Analysis*. John Wiley and Sons.
- BOX, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann. Math. Statist.* 25, 290-302.
- BRUNNER, E., DOMHOF, S. and LANGER, F. (2002). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. John Wiley and Sons.
- BRUNNER, E., MUNZEL, U. and PURI, M. L. (1999). Rank-score tests in factorial designs with repeated measures. *J. Multivariate Anal.* 70, 286-317.
- BRUNNER, E. and PURI, M. L. (1996). Nonparametric methods in design and analysis of experiments. *Handbook of Statistics* 13, 631-703.
- BRUNNER, E. and PURI, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers* 42, 1-52.
- GOTO, M., MATSUBARA, Y. and TSUCHIYA, Y. (1983). Power-normal distribution and its applications. *Rep. Stat. Appl. Res., JUSE* 30, 8-28.
- KOCH, G. G., ELASHOFF, J. D. and AMARA, I. A. (1988). Repeated measurements: Design and analysis. In: KOTZ, S. and JOHNSON, N. L. (Eds.): *Encyclopedia of Statistical Science*. John Wiley and Sons, Vol.8, 46-73.
- MANN, H. B. and WHITNEY, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 18, 50-60.

- MUNZEL, U. (1999). Linear rank score statistics when ties are present. *Statist. Probab. Lett.* 41, 389-395.
- WINER, B., BROWN, D. and MICHELS, K. (1991). *Statistical Principles in Experimental Design*, 3rd edition. McGraw-Hill.

# Multiple Change Point Detection by Sparse Parameter Estimation

Jiří Neubauer<sup>1</sup> and Vítězslav Veselý<sup>2</sup>

<sup>1</sup> Department of Econometrics, University of Defence  
Kounicova 65, Brno, Czech Republic, *Jiri.Neubauer@unob.cz*

<sup>2</sup> Department of Applied Mathematics and Computer Science, Masaryk  
University, Lipová 41a, Brno, Czech Republic, *vesely@econ.muni.cz*

**Abstract.** The contribution is focused on multiple change point detection in a one-dimensional stochastic process by sparse parameter estimation from an overparametrized model. Stochastic process with changes in the mean is estimated using dictionary consisting of Heaviside functions. The basis pursuit algorithm is used to get sparse parameter estimates. Some properties of mentioned method are studied by simulations.

**Keywords:** multiple change point detection, overparametrized model, sparse parameter estimation, basis pursuit algorithm

## 1 Introduction

Chen, S. S. et al. (1998) proposed a new methodology based on basis pursuit for spectral representation of signals (vectors). Instead of just representing signals as superpositions of sinusoids (the traditional Fourier representation) they suggested alternate dictionaries – collections of parametrized waveforms – of which the wavelet dictionary is only the best known. A recent review paper by Bruckstein et al. (2009) demonstrates a remarkable progress in the field of sparse modeling since that time. Theoretical background for such systems (also called frames) can be found for example in Christensen, O. (2003). In traditional Fourier expansion a presence of jumps in the signal slows down the convergence rate preventing sparsity. The Heaviside dictionary (see Chen et al. (1998)) merged with the Fourier or wavelet dictionary can solve the problem quite satisfactorily.

A lot of other useful applications in a variety of problems can be found in Veselý and Tonner (2005), Veselý et al. (2009) and Zelinka et al. (2004).

In Zelinka et al. (2004) kernel dictionaries showed to be an effective alternative to traditional kernel smoothing techniques. In this paper we are using Heaviside dictionary in the same manner to denoise signal (the univariate time series sample path) exhibiting jumps (discontinuities) in the mean (unlike kernel smoothing where the mean is supposed to be sufficiently smooth). Consequently, the basis pursuit approach can be proposed as an alternative to conventional statistical techniques of change point detection (see Neubauer and Veselý (2009)). The mentioned paper is focused on using the basis pursuit algorithm with the Heaviside dictionary for one change point detection.

This paper presents results of an introductory empirical study for the simplest case of detecting two change points buried in additive gaussian white noise.

## 2 Heaviside Dictionary for Change Point Detection

In this section we propose the method based on basis pursuit algorithm (BPA) for the detection of the change point in the sample path  $\{y_t\}$  in one dimensional stochastic process  $\{Y_t\}$ . We assume a deterministic functional model on a bounded interval  $\mathcal{I}$  described by the dictionary  $G = \{G_j\}_{j \in J}$  with atoms  $G_j \in L^2(\mathcal{I})$  and with additive white noise  $e$  on a suitable finite discrete mesh  $\mathcal{T} \subset \mathcal{I}$ :

$$Y_t = x_t + e_t, \quad t \in \mathcal{T},$$

where  $x \in \text{sp}(\{G_j\}_{j \in J})$ ,  $\{e_t\}_{t \in \mathcal{T}} \sim WN(0, \sigma^2)$ ,  $\sigma > 0$ , and  $J$  is a big finite indexing set. Smoothed function  $\hat{x} = \sum_{j \in J} \hat{\xi}_j G_j =: \mathbf{G}\hat{\xi}$  minimizes on  $\mathcal{T}$   $\ell^1$ -penalized optimality measure  $\frac{1}{2}\|\mathbf{y} - \mathbf{G}\xi\|^2$  as follows:

$$\hat{\xi} = \underset{\xi \in \ell^2(J)}{\text{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{G}\xi\|^2 + \lambda\|\xi\|_1, \quad \|\xi\|_1 := \sum_{j \in J} \|G_j\|_2 \xi_j,$$

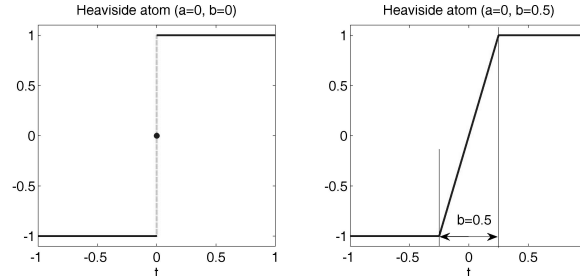
where  $\lambda = \sigma\sqrt{2 \ln(\text{card } J)}$  is a smoothing parameter chosen according to the soft-thresholding rule commonly used in wavelet theory. This choice is natural because one can prove that with any orthonormal basis  $G = \{G_j\}_{j \in J}$  the shrinkage via soft-thresholding produces the same smoothing result  $\hat{x}$ . (see Bruckstein et al. (2009)). Such approaches are also known as basis pursuit denoising (BPDN).

Solution of this minimization problem with  $\lambda$  close to zero may not be sparse enough: we are searching small  $F \subset J$  such that  $\hat{x} \approx \sum_{j \in F} \hat{\xi}_j G_j$  is a good approximation. That is why we apply the following four-step procedure described in Zelinka et al. (2004) in more detail and implemented in Veselý (2001–2008).

- (A0) Choice of a raw initial estimate  $\xi^{(0)}$ , typically  $\xi^{(0)} = \mathbf{G}^+ \mathbf{y}$ .
- (A1) We improve  $\xi^{(0)}$  iteratively by stopping at  $\xi^{(1)}$  which satisfies optimality criterion BPDN. The solution  $\xi^{(1)}$  is optimal but not sufficiently sparse in general (for small values of  $\lambda$ ).
- (A2) Starting with  $\xi^{(1)}$  we are looking for  $\xi^{(2)}$  by BPA which tends to be nearly sparse and is optimal.
- (A3) We construct a sparse and optimal solution  $\xi^*$  by removing negligible parameters and corresponding atoms from the model, namely those satisfying  $|\xi_j^{(2)}| < \alpha \|\xi^{(2)}\|_1$  where  $0 < \alpha \ll 1$  is a suitable sparsity level, a typical choice being  $\alpha = 0.05$  following an analogy with the statistical significance level.
- (A4) We repeat the step (A1) with the dictionary reduced according to the step (A3) and with a new initial estimate  $\xi^{(0)} = \xi^*$ . We expect to obtain a possibly improved sparse estimate  $\xi^*$ .

Hereafter we refer to this four-step algorithm as to BPA4. The steps (A1), (A2) and (A4) use Primal-Dual Barrier Method designed by M. Saunders (see Saunders (1997–2001)). This up-to-date sophisticated algorithm allows one to solve fairly general optimization problems minimizing convex objective subject to linear constraints. A lot of controls provide a flexible tool for adjusting the iteration process.

We build our dictionary from heaviside-shaped atoms on  $L^2(\mathbb{R})$  derived from a fixed 'mother function' via shifting and scaling following the analogy with the construction of wavelet bases.



**Fig. 1.** Heaviside atoms with parameters  $a = 0, b = 0$  and  $a = 0, b = 0.5$

We construct an oversized shift-scale dictionary  $G = \{G_{a,b}\}_{a \in \mathcal{A}, b \in \mathcal{B}}$  derived from the 'mother function' by varying the shift parameter  $a$  and the scale (width) parameter  $b$  between values from big finite sets  $\mathcal{A} \subset \mathbb{R}$  and  $\mathcal{B} \subset \mathbb{R}^+$ , respectively ( $J = \mathcal{A} \times \mathcal{B}$ ), on a bounded interval  $\mathcal{I} \subset \mathbb{R}$  spanning the space  $H = \text{sp}(\{G_{a,b}\}_{a \in \mathcal{A}, b \in \mathcal{B}})$ , where

$$G_{a,b}(t) = \begin{cases} 1 & \text{for } t - a > b/2, \\ 2(t - a)/b & |t - a| \leq b/2, b > 0, \\ 0 & t = a, b = 0, \\ -1 & \text{otherwise.} \end{cases}$$

In the simulations below  $\mathcal{I} = [0, 1]$ ,  $\mathcal{T} = \{t/T\}$  (typically with mesh size  $T = 100$ ),  $\mathcal{A} = \{t/T\}_{t=t_0}^{T-t_0}$  ( $t_0$  is a boundary trimming,  $t_0 = 4$  was used in the simulations) and scale  $b$  fixed to zero ( $\mathcal{B} = \{0\}$ ). Clearly the atoms of such Heaviside dictionary are normalized on  $\mathcal{I}$ , i.e.  $\|G_{a,0}\|_2 = 1$ . Some examples of Heaviside functions are displayed in the figure 1.

### 3 Change Point Detection by Basis Pursuit

Neubauer and Veselý (2009) proposed the method of change point detection if there is just one change point in a one-dimensional stochastic process (or in its sample path). We briefly describe a given method. We would like to find a change point in a stochastic process

$$Y_t = \begin{cases} \mu + \epsilon_t & t = 1, 2, \dots, c, \\ \mu + \delta + \epsilon_t & t = c + 1, \dots, T, \end{cases} \quad (1)$$

where  $\mu, \delta \neq 0, t_0 \leq c < T - t_0$  are unknown parameters and  $\epsilon_t$  are independent identically distributed random variables with zero mean and variance  $\sigma^2$ . The parameter  $c$  indicates the change point in the process. Using the basis pursuit algorithm we obtain some significant atoms, we calculate correlation between significant atoms and analyzed process. The shift parameter of the atom with the highest correlation is taken as an estimator of the change point  $c$ .

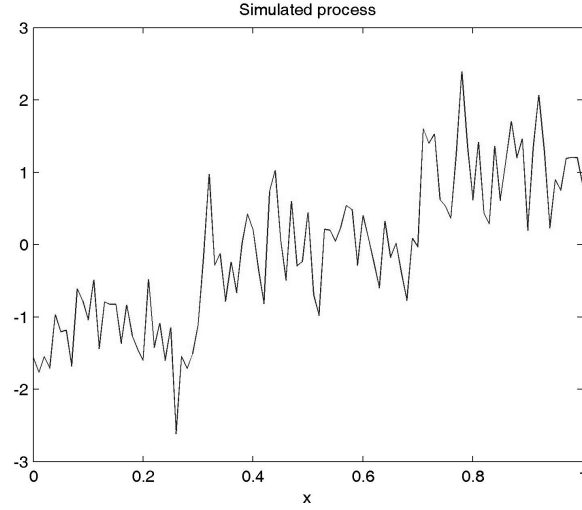


Fig. 2. Simulated process

#### 4 Multiple Change Point Detection by Basis Pursuit

Now let us assume the model with two change points

$$Y_t = \begin{cases} \mu + \epsilon_t & t = 1, 2, \dots, c_1 \\ \mu + \delta_1 + \epsilon_t & t = c_1 + 1, \dots, c_2, \\ \mu + \delta_2 + \epsilon_t & t = c_2 + 1, \dots, T, \end{cases} \quad (2)$$

where  $\mu, \delta_1, \delta_2 \neq 0, t_0 \leq c_1 < c_2 < T - t_0$  are unknown parameters and  $\epsilon_t$  are independent identically distributed random variables with zero mean and variance  $\sigma^2$ .

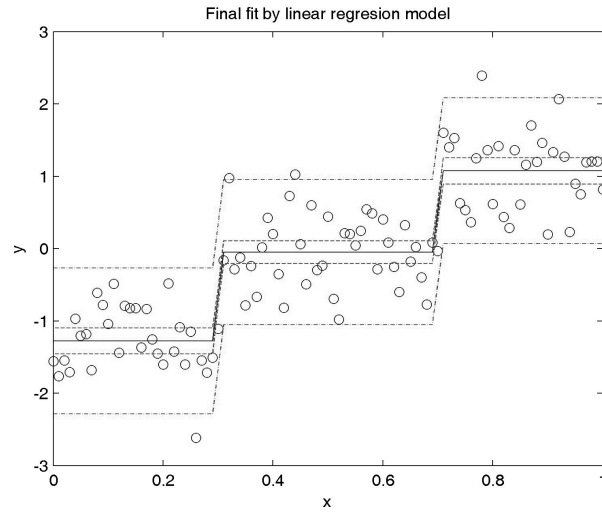
We use the method of change point estimation described above for detection two change points  $c_1$  and  $c_2$  in the model (2). Instead of finding only one significant atom with the highest correlation with the process  $Y_t$  we can identify two significant atoms with the highest correlation. The shift parameters of these atoms determine estimators for the change points  $c_1$  and  $c_2$ . Another possibility is to apply the procedure of one change point detection two times in sequence. In the first step we identify one change point in the process  $Y_t$ , then we subtract given significant atom from the process (by linear regression)

$$\begin{aligned} Y_t &= \beta G_{0, \hat{c}_1} + e_t \\ Y'_t &= Y_t - \hat{\beta} G_{0, \hat{c}_1} \end{aligned}$$

and finally we apply the method to the new process  $Y'_t$ . The shift parameters of selected atoms are again identifiers of the change points  $c_1$  and  $c_2$ . Observe that this can be seen as two steps of orthogonal matching pursuit (OMP) combined with BPA.

We demonstrate the method of multiple change point detection by BPA4 on simulations of the process (2) with the change points  $c_1 = 30$  and  $c_2 = 70, T = 100$ ,





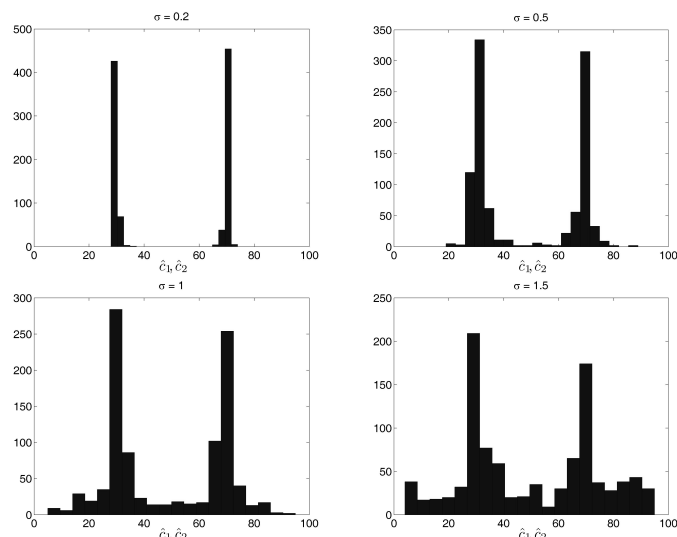
**Fig. 3.** Final fit by linear regression model

$\mu = -1, \delta_1 = 1, \delta_2 = 2$  and  $\sigma = 0.5$  (for BPA4 we transform it to the interval  $[0, 1]$ , see figure 2). After first applying the method of one change point detection, we get estimate  $\hat{c} = 30$ . Using linear regression we can subtract this identified atom from the process and repeat the procedure. We obtain estimation of the second change point  $\hat{c}_2 = 70$ . We use linear regression model

$$y_t = \beta_1 G_{0.3,0} + \beta_2 G_{0.7,0} + u_t,$$

to get final fit on the simulated process. We obtain  $\hat{\beta}_1 = 0.651$  with the confidence interval  $[0.459, 0.664]$  and  $\hat{\beta}_2 = 0.613$  with the confidence interval  $[0.511, 0.716]$ , see figure 3. Dashed lines denote 95% confidence and prediction intervals.

For the purpose of introductory performance study of the proposed method of multiple change point detection we use simulations of the process (2). We put, analogously to the example,  $\mu = -1, \delta_1 = 1, \delta_2 = 2$  and  $T = 100$  where the error terms are independent normally distributed with zero mean and the standard deviations  $\sigma = 0.2, 0.5, 1$  and  $1.5$ , respectively. We calculate simulations of this model with change points  $c_1 = 30$  and  $c_2 = 70$  (500 simulations for each choice of standard deviation). We preferred the second method of multiple change point detection (an application the procedure of one change point detection two times in succession) which proved to be more suitable. The reason being that first method based on selecting two significant atoms exhibiting highest correlation with the analyzed process becomes sensitive (as  $\sigma$  increases) to the number of selected significant atoms. This fact affects the simulations negatively. All results are summarized in the figure 4 where histograms display frequencies of change point estimates for given values of the standard deviation  $\sigma$ . For the small values of the standard deviation of the white noise  $\epsilon_t$  the proposed method detects the change points satisfactorily. In accordance with our expectation, the uncertainty of the change point positions grows with the increasing noise variance  $\sigma^2$  but is still acceptable.



**Fig. 4.** Histograms of estimated change points for  $\sigma = 0.2, 0.5, 1$  and  $1.5$

The model (2) can be easily extended to more than two change points. The number of the change points in real situation unknown. Using the BP approach we assume that if there are any change points, we can detect significant atoms in BPA4 algorithm. In case there is not a significant atom, change point cannot be detected. In the first step we identify one change point, then we subtract given significant atom from the process by linear regression (according to the procedure of two change points detection mentioned above). If there are some significant atoms in the new process, we find the atom with highest correlation and subtract it from the new process etc. We stop when it is not possible to detect any significant atom.

## 5 Conclusion

According to the introductory simulation results the basis pursuit approach proposes a reasonable detection method of two change points in one-dimensional process. The outlined method can be used for detection of two or more change points, or another sort of change point with a dictionary  $G$  of different kind.

The change point detection techniques may be useful for instance in modeling of economical or environmental time series where jumps can occur.

## References

- BRUCKSTEIN, A. M. et al. (2009): From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. *SIAM Review* 51 (1), 34–81.  
 CHEN, S. S. et al. (1998): Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* 20 (1), 33–61 (2001 reprinted in *SIAM Review* 43 (1), 129–159).

- CHRISTENSEN, O. (2003): *An introduction to frames and Riesz bases*. Birkhuser, Boston-Basel-Berlin.
- NEUBAUER, J. and VESELÝ, V. (2009): Change Point Detection by Sparse Parameter Estimation. In: *The XIIIth International conference: Applied Stochastic Models and Data Analysis*. Vilnius, 158–162.
- SAUNDERS, M. A. (1997–2001): *pdsco.m: MATLAB code for minimizing convex separable objective functions subject to  $Ax = b, x \geq 0$* .
- VESELÝ, V. (2001–2008): *framebox: MATLAB toolbox for overcomplete modeling and sparse parameter estimation*.
- VESELÝ, V. and TONNER, J. (2005): Sparse parameter estimation in overcomplete time series models. *Austrian Journal of Statistics*, 35 (2&3), 371–378.
- VESELÝ, V. et al. (2009): Analysis of PM10 air pollution in Brno based on generalized linear model with strongly rank-deficient design matrix. *Environmetrics*, 20 (6), 676–698.
- ZELINKA et al. (2004): Comparative study of two kernel smoothing techniques. In: Horová, I. (ed.) *Proceedings of the summer school DATASTAT'2003, Svatka*. Folia Fac. Sci. Nat. Univ. Masaryk. Brunensis, Mathematica 15: Masaryk University, Brno, Czech Rep., 419–436.

**Acknowledgment:** supported by grants GAČR P402/10/P209 and MSM0021622418



# Quasi-Maximum Likelihood Estimators for Threshold ARMA Models: Theoretical Results and Computational Issues

Marcella Niglio<sup>1</sup> and Cosimo Damiano Vitale<sup>2</sup>

<sup>1</sup> Department of Economics and Statistics  
Via Ponte Don Melillo, Fisciano (SA), Italy *mniglio@unisa.it*

<sup>2</sup> Department of Economics and Statistics  
Via Ponte Don Melillo, Fisciano (SA), Italy *vitale@unina.it*

**Abstract.** In this paper we derive quasi-maximum likelihood estimators for the parameters of the threshold autoregressive moving average process (TARMA). After the presentation of the model, we discuss some property that makes this model of interest in most empirical domains. The derivation of the estimators is proposed in details and computational issues are examined in a simulation study.

**Keywords:** threshold model, Q-ML estimators, parameters initialization

## 1 Introduction

Let  $\{X_t, t \geq 0\}$  a stochastic process defined on  $(\Omega, \mathcal{F}_t, P_\beta)$  and values in  $\mathcal{R}$ , with  $\beta$  the  $(v \times 1)$ -vector of unknown parameters, such that  $\beta \in B$  open, and  $\mathcal{F}_t$  is a  $\sigma$ -field generated by  $\{X_t, t \leq T\}$ .

The process  $X_t$  is called Self Exciting TARMA( $\ell; p, q$ ) (shortly called SETARMA( $\ell; p, q$ ), if his form is given by:

$$X_t = \left( \phi^{(k)\top} \mathbf{X}_{t-1} + e_t - \theta^{(k)\top} \mathbf{e}_{t-1} \right) I(X_{t-d} \in \mathcal{R}_k), \quad k = 1, \dots, \ell, \quad (1)$$

where  $^\top$  denotes transposition,  $\phi^{(k)\top} = (\phi_1^{(k)} \dots \phi_p^{(k)})$ ,  $I(\cdot)$  is an indicator function,  $\theta^{(k)\top} = (\theta_1^{(k)} \dots \theta_q^{(k)})$ , for  $k = 1, \dots, \ell$ ,  $\mathbf{X}_{t-1}^\top = (X_{t-1}, \dots, X_{t-p})$ ,  $\mathbf{e}_{t-1}^\top = (e_{t-1}, \dots, e_{t-q})$ ,  $\{e_t\}$  is a sequence of independent continuous random variables, with  $E(e_t) = 0$  and finite variance  $V(e_t) = \sigma_e^2 > 0$ ,  $d \in \mathbb{N}^*$  is the threshold delay,  $\mathcal{R}_j \cap \mathcal{R}_i = \emptyset$ , for  $i \neq j$ ,  $\cup_{k=1}^\ell \mathcal{R}_k = \mathcal{R}$ . Model (1), originally proposed in Tong (1983), is characterized by a switching among regimes regulated by the process  $\{X_i\}$ , with  $i = t - d$ .

In more general settings the threshold variable  $X_{t-d}$  of model (1) can be assumed exogenous and is denoted with  $Y_{t-d}$ . This last specification is of wide interest in many empirical domains where the data generating process of  $X_t$  is related to the dynamic of another phenomenon observed at time  $t - d$ . A further generalization of the TARMA model can be obtained assuming that the error component  $e_t$  may change in each regime such that model (1) becomes:

$$X_t = \left( \phi^{(k)\top} \mathbf{X}_{t-1} + e_t^{(k)} - \theta^{(k)\top} \mathbf{e}_{t-1}^{(k)} \right) I(X_{t-d} \in \mathcal{R}_k), \quad k = 1, \dots, \ell.$$

As remarked in Tong (1990), the dynamic structure of threshold models makes them able to catch asymmetric effects often recognized in the data sets. Starting from this comment, in Section 2 we discuss the asymmetry of model (1) and we briefly present the stationarity of this model. In Section 3 the quasi-maximum likelihood (Q-ML) estimators of the model parameters are presented and some computational problem is discussed and illustrated by means of a Monte Carlo simulation study.

## 2 Properties of the Threshold ARMA process

One of the main features of the threshold models is related to their ability to catch asymmetric effects in the data generating process.

In particular, if the distribution of a SETARMA(2;  $p, q = 0$ ) (commonly called SETAR(2;  $p$ )) and of a SETARMA(2;  $p, q$ ) model are compared, it can be noted that the presence of a moving average component could make the distribution more skewed and the intercepts are relevant for its shape. In more detail, when the asymmetry and the kurtosis are measured in terms of third and fourth moment of  $Z_t = [X_t - E(X_t)]/[V(X_t)]^{1/2}$  respectively, it can be shown that the shape of the distribution of  $X_t$  is strictly related to the mean of both regimes and that the process  $X_t$  is asymmetric and/or leptokurtic even when the two regimes have  $\gamma_1^{(k)} = E[(Z_t^{(k)})^3] = 0$  and  $\gamma_2^{(k)} = E[(Z_t^{(k)})^4] - 3 = 0$ , where  $Z_t^{(k)} = [X_t^{(k)} - E(X_t^{(k)})]/[V(X_t^{(k)})]^{1/2}$ , for  $k = 1, 2$ .

In this regard it can be shown that in presence of a SETARMA(2;  $p, q$ ) model,  $\gamma_1 = E[Z_t^3]$  becomes:

$$\gamma_1 = p\gamma_1^{(1)} + (1-p)\gamma_1^{(2)} + 3 \frac{p\mu_2^{(1)}(\mu_1^{(1)} - \mu) + (1-p)\mu_2^{(2)}(\mu_1^{(2)} - \mu)}{\sigma^3} - 2 \frac{p(\mu_1^{(1)})^3 + (1-p)(\mu_1^{(2)})^3 - \mu^3}{\sigma^3},$$

and in a similar way the kurtosis can be measured in terms of  $\gamma_2 = E[Z_t^4] - 3$  that becomes:

$$\gamma_2 = p\gamma_2^{(1)} + (1-p)\gamma_2^{(2)} + 3 \frac{p(\mu_1^{(1)})^4 + (1-p)(\mu_1^{(2)})^4 - \mu^4}{\sigma^4} + 4 \frac{p\mu_3^{(1)}(\mu_1^{(1)} - \mu) + (1-p)\mu_3^{(2)}(\mu_1^{(2)} - \mu)}{\sigma^4} - 6 \frac{p\mu_2^{(1)}((\mu_1^{(1)})^2 - \mu^2) + (1-p)\mu_2^{(2)}((\mu_1^{(2)})^2 - \mu^2)}{\sigma^4},$$

with  $\mu_j^{(k)} = E[(X_t^{(k)})^j]$ , for  $k = 1, 2$  and  $j = 1, 2, 3$ ,  $\mu = E(X_t)$ ,  $\sigma = [V(X_t)]^{1/2}$  and  $p = E[I_{t-d}]$ .

The results given on  $\gamma_1$  and  $\gamma_2$  are based on the assumption that the TARMA model is stationary. This statistical property has preminent importance even to face inferential problems and it has been differently examined for model (1): Liu and Susko (1992), Brockwell *et al.* (1992), Ling (1999) discuss the strong stationarity whereas more recently Amendola *et al.* (2009) provide sufficient conditions

under which model (1) is weakly stationary. This last property has been further discussed in Niglio and Vitale (2010) in the context of Threshold ARMA process with exogenous threshold variable  $Y_{t-d}$  which is assumed to be independent from the errors  $\{e_t\}$ . More precisely they distinguish between the so called “global” and “local” stationarity of the model: the authors give the conditions under which a TARMA model, with  $\ell \geq 2$  regimes and a stationary exogenous threshold variable  $Y_{t-d}$ , can be globally stationary even in presence of local unit roots in one or more regimes. They provide these conditions showing that they are based not only on the autoregressive coefficients of the model but even on the proportion of observations belonging to each regime. For example, if the generating process is  $X_t \sim \text{TARMA}(\ell; 1, q)$ , a sufficient condition for the weak stationarity is given by  $\prod_{k=1}^{\ell} |\phi_1^{(k)}|^{p_k} < 1$ , where  $p_k$  is the proportion of observations in regime  $k$ , for  $k = 1, 2, \dots, \ell$ . In other words, they show that the process  $X_t$  is weakly stationary even in presence of regimes having unit roots in modulus, if at least one regime is locally stationary. More general results are given to the TARMA( $\ell; p, q$ ) model. The study of the statistical properties of model (1) is an essential requirement to introduce the estimation of its parameters. More precisely in presence of stationary processes consistent estimators can be obtained for the model under analysis using some results widely developed in nonlinear domain (among the others see Tong (1990)). In the following Section, Q-ML estimators are presented for the stationary model (1) and some computational issues are given in this context.

### 3 Parameters estimation and initialization

As expected a greater computational effort is requested to estimate a TARMA model with respect to other threshold models without moving average component. To introduce some detail on this problem, in the following we consider model (1) with  $\ell = 2$  regimes. It simplifies the notation but does not constrain the generality of the proposed results that can be easily extended to cases with  $\ell > 2$ .

Following the approach described in Tong (1990) in a quite different setting, we start the presentation of the estimation procedure assuming that the orders  $p$  and  $q$ , the threshold delay  $d$  and the threshold value  $r$  are all known.

Let  $X_t$ , a stationary and ergodic process with unknown vector of parameters  $\beta \in B \subset \mathcal{R}^v$  (with  $v \geq 1$  and  $B$  an open set), the conditional Q-ML function is defined as:

$$L(\beta, \mathcal{F}_T) = \prod_{i=1}^T f_{i|i-1}(x_i, \beta | \mathcal{F}_{i-1}),$$

where  $\mathcal{F}_{i-1}$  is a  $\sigma$ -field generated by  $(X_1, \dots, X_{i-1})$ , with  $\mathcal{F}_i \subset \mathcal{F}_{i-1}$ , for  $i = 1, 2, \dots, T$ ,  $\mathcal{F}_0$  is the trivial  $\sigma$ -field and  $f_{i|i-1}$  is the conditional Gaussian density of  $X_i$ . Starting from these results, the conditional Q-ML function becomes:

$$\ell(\beta, \mathcal{F}_T) = \sum_{i=1}^T [\log(L_i(\beta, \mathcal{F}_{i-1})) - \log(L_{i-1}(\beta, \mathcal{F}_{i-2}))], \quad (2)$$

with  $L_0 = 1$  for convenience and  $L_i(\beta, \mathcal{F}_{i-1}) = \prod_{j=1}^i f_{j|j-1}(x_j, \beta | \mathcal{F}_{j-1})$ .

When  $X_t \sim \text{SETARMA}(2; p, q)$ , then  $\beta = (\phi^{(1)}, \phi^{(2)}, \theta^{(1)}, \theta^{(2)}, \sigma_e^2)_{[v \times 1]}$ , with  $v = 2p + 2q + 1$ , the function (2) has form:

$$\ell_t(\beta, \mathcal{F}_T) = -\frac{T-s}{2} [\log(2\pi) + \log(\sigma_e^2)] - \frac{1}{2} \sum_{i=s}^T \frac{e_i^2(\beta)}{\sigma_e^2}, \quad (3)$$

where:

- $s = \max(p, q, d) + 1$ ;
- $e_i^2(\beta) = e_{i,1}^2(\beta)$ , if  $I(X_{t-d} \in \mathcal{R}_1) = 1$ , and  $e_i^2(\beta) = e_{i,2}^2(\beta)$  otherwise, with  $e_{i,k}(\beta) = X_i - E(X_i | \mathcal{F}_{i-1}) = X_t - (\phi^{(k)})^\top \mathbf{X}_{t-1} - \theta^{(k)} \mathbf{e}_{t-1}$ , for  $k = 1, 2$ ;
- $\sigma_e^2 = V(e_i)$ .

Noting that  $\hat{\sigma}_e^2 = (T-s)^{-1} \sum_{i=s}^T e_i^2(\beta)$ , the conditional Q-ML function (3) can be shortly given as:

$$\ell_t(\beta, \mathcal{F}_T) = C - \frac{T-s}{2} \log \sum_{i=s}^T e_i^2(\beta), \quad (4)$$

with  $C = -\frac{T-s}{2} [\log(2\pi) - \log(T)] - \frac{T}{2}$  and such that the Q-ML estimator for  $\beta$  is obtained from:

$$\hat{\beta} = \arg \min_{\beta \in B} Q_T(\beta), \quad \text{with } Q_T(\beta) = \sum_{i=s}^T e_i^2(\beta), \quad (5)$$

where the minimization of  $Q_T(\beta)$  can be performed using nonlinear optimization algorithms based on the score vector whose elements are computed from:

$$\frac{\partial e_t}{\partial \phi_0^{(k)}} = -1 + \sum_{i=1}^q \left( \theta_i^{(1)} I(X_{t-d-i} \leq r) + \theta_i^{(2)} I(X_{t-d-i} > r) \right) \frac{\partial e_{t-i}}{\partial \phi_0^{(k)}} \quad (6a)$$

$$\frac{\partial e_t}{\partial \phi_j^{(k)}} = -X_{t-j} + \sum_{i=1}^q \left( \theta_i^{(1)} I(X_{t-d-i} \leq r) + \theta_i^{(2)} I(X_{t-d-i} > r) \right) \frac{\partial e_{t-i}}{\partial \phi_j^{(k)}} \quad (6b)$$

$$\frac{\partial e_t}{\partial \theta_u^{(k)}} = e_{t-u} + \sum_{i=1}^q \left( \theta_i^{(1)} I(X_{t-d-i} \leq r) + \theta_i^{(2)} I(X_{t-d-i} > r) \right) \frac{\partial e_{t-i}}{\partial \theta_u^{(k)}}, \quad (6c)$$

for  $j = 1, \dots, p$ ,  $u = 1, \dots, q$ ,  $t = s, \dots, T$  and  $k = 1, 2$ .

To select the remaining  $(d, p, q, r)$  parameters of the model, a grid search can be performed defining a set of candidate values for  $(d, p, q)$  and where the threshold value  $r$  is chosen in the range of the threshold variable  $X_{t-d}$  delimited from the quantiles  $[q_\alpha, q_{1-\alpha}]$ , with a value selected for  $\alpha \in [0.2, 0.8]^1$  that should guarantee an adequate number of observations in both regimes. For each combination of parameters  $(d, p, q, r)$  the Akaike Information Criterion (AIC) can be computed and the model having minimum AIC is selected.

This iterative procedure is quite common in non linear domain, but from the computational point of view, it is of interest to evaluate how the estimates of  $\beta$  can be initialized and how to carry out the grid search. As expected the initializations can

<sup>1</sup> Note that when  $\alpha = 0$  or  $\alpha = 1$  the process degenerates to a linear ARMA model.



be of help to try to avoid local maxima and to increase the speed of convergence of the estimates. Our proposal is to adapt an estimation procedure, originally presented in Tong (1983), to the model under analysis noting that it can be seen as the “natural” starting point to minimize  $Q_T(\beta)$  in (5).

In more detail, let  $d, p$  and  $q$  three fixed parameters and let  $s = \max(p, q, d) + 1$ , the data set  $(X_s, X_{s+1}, \dots, X_T)$  may be splitted in two non-overlapping subsets,  $\mathcal{S}_1$  and  $\mathcal{S}_2$ , according to the value selected for the threshold  $r$ , such that:

$$X_t \in \mathcal{S}_1 \quad \text{iff} \quad X_{t-d} \in \mathcal{R}_1 \quad \text{and} \quad X_t \in \mathcal{S}_2 \quad \text{iff} \quad X_{t-d} \in \mathcal{R}_2$$

for  $t = s, \dots, T$ , where  $\mathcal{R}_1 = (-\infty, r]$  and  $\mathcal{R}_2 = (r, \infty)$  form a partition of the real line  $\mathcal{R}$ .

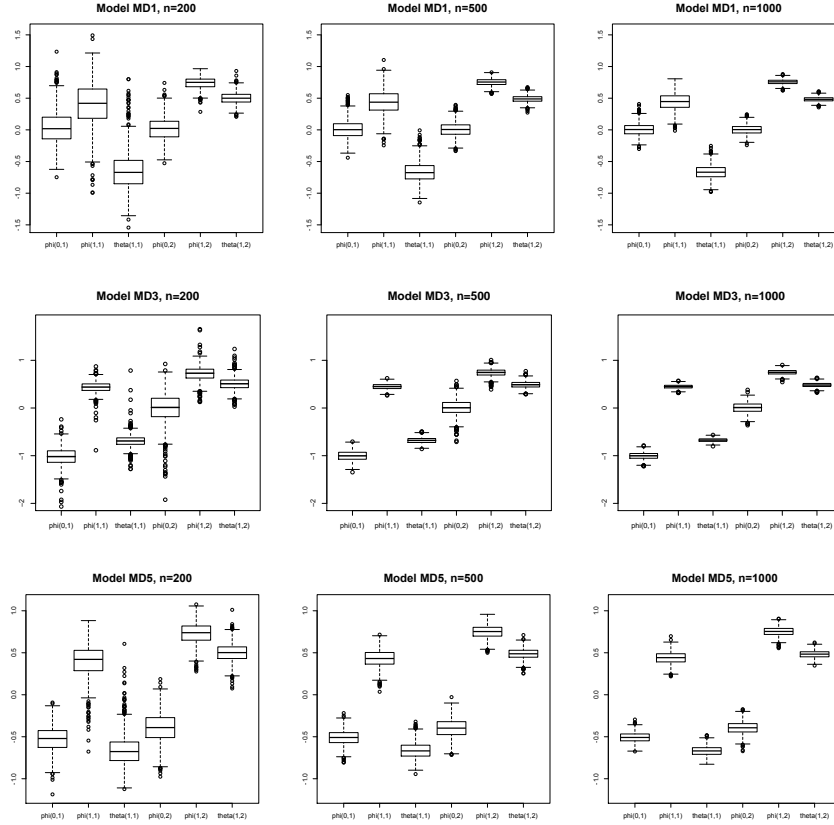
If the cardinality of  $\mathcal{S}_k$  (denoted  $\#(\mathcal{S}_k)$ ) is adequate, with  $k = 1, 2$  and  $\#(\mathcal{S}_1) + \#(\mathcal{S}_2) = T - s$ , or, in other words, if each regime has a moderate or a large number of observations, the parameters of each ARMA regime  $(\phi^{(k)}, \theta^{(k)})$  are estimated independently (neglecting the dependence between regimes) using conditional least squares estimates (Box and Jenkins, 1976) whereas the preliminary estimate of the white noise variance is obtained from  $\hat{\sigma}_e^2 = (T - s)^{-1} \sum_{i=s}^T \hat{e}_i^2$ .

Starting from the previous theoretical results, the procedure to estimate the parameters of the SETARMA model can be summarized in five main steps:

- E1. select a set of non negative integer values for the  $p$  and  $q$  orders, define a set  $\mathcal{D} \in N^*$  of possible threshold delays  $d$  and a subset  $[R_L, R_U]$  of  $\alpha$ -quantiles of  $X_t$  that represent the candidate values of the threshold parameter  $r$ ; for all combinations of  $(p, q, d, r)$  define a grid;
- E2. select the cells of the grid defined in step E1. such that  $\#\mathcal{S}_k$  is at least moderate, for  $k = 1, 2$ ;
- E3. for each cell selected in E2., estimate the parameters of the ARMA models fitted to each regime in order to obtain preliminary estimates for  $\beta$ ;
- E4. estimate the  $\phi_j^{(k)}$  and  $\theta_u^{(k)}$  parameters of model (1) by means of (5) using the estimates obtained in step E3. as starting values;
- E5. use the residuals  $\hat{e}_t = X_t - \hat{X}_t$ , for  $t = s, \dots, T$ , obtained from step E4., and estimate the variance  $\hat{\sigma}_e^2 = (T - s)^{-1} \sum_{t=s}^T \hat{e}_t^2$ .

This five steps procedure has been used in a simulation experiment where the parameters of three SETARMA models are estimated and compared. The properties of Q-ML estimators in presence of stationary processes are widely known in the literature. This is the reason why the interest risen by this study is mainly related to the speed of convergence and to the distributions of the estimated parameters. To investigate these properties we have simulated 1000 time series from a SETARMA(2;1,1) model, denoted MD1, with parameters  $\phi_0^{(1)} = \phi_0^{(2)} = 0, \phi_1^{(1)} = 0.45, \theta_1^{(1)} = -0.67, \phi_1^{(2)} = 0.76$  and  $\theta_1^{(2)} = 0.48$ , and model MD3 that differs from model MD1 only for the intercept  $\phi_0^{(1)} = -1$ , using series of different length  $n = 200, 500, 1000$  (all series are obtained dropping the first 100 data points to avoid the effect of initial seed). For each model the vector  $\beta$  has been estimated using Q-ML estimators. This simulation scheme has been replicated for another SETARMA(2;1,1) model, denoted MD5, that differs from model MD1 and MD3 only for the intercepts of both regimes having values  $\phi_0^{(1)} = -0.5$  and  $\phi_0^{(2)} = -0.4$ . It is widely known that the intercepts have remarkable impact on the moments of threshold models, on their identifiability and even on their existence. It is due to

the fact that small changes of the intercepts values can cause significant changes on the generating process. To better understand this last remark, in Figure 1 the box plots of the estimated parameters are shown for model MD1, MD3 and MD5 respectively. As expected, in all cases, the variability of the empirical distributions



**Fig. 1.** Box-plots of the 1000 estimates obtained from model MD1, MD3 and MD5 using three time series length:  $n = 200, 500, 1000$ .

decreases as the series length grows. In more detail, when the variability is measured in terms of interquartile range,  $IQR = q_{0.75} - q_{0.25}$ , in Table 1 it can be noted that the changes in variability are due not only to the number of observations that belong to each regime (evaluated through  $\hat{\lambda}$  that represents the mean value, over the 1000 replicates, of the proportion of observations belonging to the first regime) but even to the presence of the intercept values that are of help to discriminate between the two regimes. In fact, if the attention is focused on model MD3, it can be noted that, despite the small number of observations generated from the second regime, the IQR computed for the empirical distribution of  $\phi_1^{(2)}$  and  $\theta_1^{(2)}$  has

not remarkable differences with respect to the other two models. On the contrary, when the attention is focused on the parameters  $\phi_1^{(1)}$  and  $\theta_1^{(1)}$  of model MD1 where  $\hat{\lambda} \approx 0.21$  and  $\phi_0^{(1)} = \phi_0^{(2)} = 0$ , the variability increases with respect to MD3 and MD5. The estimate  $\hat{\lambda}$  in the last column of Table 1 is of primary importance for the model under analysis for different reasons: it changes remarkably even in presence of small changes of  $\phi_0^{(1)}$  and  $\phi_0^{(2)}$ ; it allows to define the stationarity region over the parametric space (as hinted in Section 2 and extensively discussed in Niglio and Vitale (2010)); it impacts the identifiability of the model and the subsequent parameters estimation.

	$\phi_0^{(1)}$	$\phi_1^{(1)}$	$\theta_1^{(1)}$	$\phi_0^{(2)}$	$\phi_1^{(2)}$	$\theta_1^{(2)}$	$\hat{\lambda}$
MD1 ( $n = 200$ )	0.3415 (0.0184)	0.4611 (0.4194)	0.3688 (-0.6714)	0.2455 (0.0229)	0.1202 (0.7502)	0.1208 (0.4999)	0.2112
MD1 ( $n = 500$ )	0.1866 (0.0030)	0.2574 (0.4372)	0.2098 (-0.6754)	0.1486 (0.0057)	0.0755 (0.7552)	0.0708 (0.4884)	0.2119
MD1 ( $n = 1000$ )	0.1297 (0.0052)	0.1811 (0.4470)	0.1445 (-0.6685)	0.1011 (0.0042)	0.0524 (0.7581)	0.0489 (0.4821)	0.2111
MD3 ( $n = 200$ )	0.2425 (-1.0189)	0.1369 (0.4389)	0.1356 (-0.6913)	0.3872 (0.0122)	0.1864 (0.7288)	0.1592 (0.5078)	0.7185
MD3 ( $n = 500$ )	0.1461 (-1.0029)	0.0826 (0.4498)	0.0837 (-0.6790)	0.2079 (0.0049)	0.1017 (0.7462)	0.0950 (0.4877)	0.7185
MD3 ( $n = 1000$ )	0.0996 (-1.0022)	0.0546 (0.4480)	0.0526 (-0.6729)	0.1519 (0.0089)	0.0713 (0.7488)	0.0636 (0.4817)	0.7165
MD5 ( $n = 200$ )	0.2006 (-0.5209)	0.2417 (0.4214)	0.2208 (-0.6764)	0.2371 (-0.3907)	0.1689 (0.7391)	0.1380 (0.5021)	0.4836
MD5 ( $n = 500$ )	0.1186 (-0.5087)	0.1385 (0.4316)	0.1294 (-0.6683)	0.1536 (-0.3972)	0.1047 (0.7527)	0.0824 (0.4880)	0.4836
MD5 ( $n = 1000$ )	0.0816 (-0.5084)	0.0971 (0.4419)	0.0793 (-0.6686)	0.0992 (-0.3930)	0.0714 (0.7540)	0.0599 (0.4838)	0.4836

**Table 1.** Interquartile range (IQR) of the empirical distribution obtained for the Autoregressive and Moving Average parameters of three SETARMA models (in parenthesis is given the median).  $\hat{\lambda}$  is the mean value, over the 1000 simulated time series, of the proportion of observations generated from the first regime.

## References

- AMENDOLA, A., NIGLIO, M. and VITALE, C.D. (2009): Statistical properties of SETARMA models, *Communications in Statistics: Theory and Methods*, 38, 2479-2497.
- BOX, J.E.P. and JENKINS, G.M. (1976): *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco.
- BROCKWELL, P.J., LIU, J. and TWEEDIE, R.L. (1992): On the existence of stationary threshold AR-MA process, *Journal of the Time Series Analysis*, 13, 95-107.
- LIU, J. and SUSKO, E. (1992): On strict stationarity and ergodicity of a non-linear ARMA model, *Journal of Applied Probability*, 29, 363-373.

- LING, S. (1999): On the probabilistic properties of a double threshold ARMA conditional heteroskedasticity model, *Journal of Applied Probability*, 36, 688-705.
- NIGLIO, M. and VITALE, C.D. (2010): Local unit roots and global stationarity of TARMA models, *Methodology and Computing in Applied Probability*, DOI:10.1007/s11009-010-9166-y.
- TONG, H. (1983): *Threshold Models in Nonlinear Time Series Analysis*, Springer-Verlag, London.
- TONG, H. (1990): *Non-Linear Time Series. A Dynamical System Approach*, Clarendon Press Oxford.
- TONG, H. and LIM, K.S. (1980): Threshold autoregression, limit cycles and cyclical data, *Journal of Royal Statistical Society (B)*, 42, 245-292.

# A Case Study of Bank Branch Performance Using Linear Mixed Models

Peggy Ng<sup>1</sup>, Claudia Czado<sup>2</sup>, Eike Christian Brechmann<sup>2</sup>, and Jon Kerr<sup>1</sup>

<sup>1</sup> School of Administrative Studies, York University  
4700 Keele Street, Toronto, Canada, *peggyng@yorku.ca*, *jonkerr@yorku.ca*

<sup>2</sup> Center for Mathematical Sciences, Technische Universität München  
Boltzmannstr. 3, D-85747 Garching, Germany, *cczado@ma.tum.de*,  
*eike.brechmann@mytum.de*

**Abstract.** The assessment of performance and potential is central to decisions pertaining to the location of bank branches. A common method for evaluating branch performance is data envelope analysis in which in-branch variables are typically considered. This paper adopts an alternate methodology that quantifies the influence of local socio-economic variables on bank deposits (a common measure of performance) using linear mixed models (LMM). It also illustrates the potential of using LMM to build a predictive model to support branch location decisions.

**Keywords:** bank performance, branching, linear mixed models

## 1 Introduction

Commercial banks can operate as a single unit bank or develop a network of bank branches that act as the key contact points between customers and the central bank. As such, branches occupy key positions in banking organizations and their locations reflect important strategic and operating decisions. Several empirical studies have dealt with the issues concerning banks' optimal branch networks (as examples see: Boufounou (1995), DeYoung et al. (2004), Gart (1994), Jayaratne and Strahan (1996), Shiers (2002)).

The optimum number of branches and their optimum locations are interrelated issues that have to be addressed by bank managers. Chelst et al. (1988) provide a general procedure to facilitate this task in which the assessment of performance and potential of the branch network is central. Yet, such assessments are complex multidimensional processes. In fact, Doyle et al. (1979) found 38 independent variables needed to fully describe branch performance. Boufounou (1995) analyzed a similar number of variables for a commercial bank in Greece while Avkiran (1997) tested 91 potential variables and six performance variables for evaluating branch performance.

Approaches to evaluating performance vary, but since the 1990's data envelope analysis (DEA) has been used extensively to evaluate banking institutions (Berger and Humphrey (1997), Mostafa (2007)). DEA is a non-parametric linear programming technique used to compute a comparative ratio of inputs to outputs for each unit. Variables under consideration are typically in-branch variables (see Berger

and Humphrey (1997)), such as number of employees, physical space and branch operating expenses.

The focus of this study is to quantify the influence of local socio-economic variables on branch performance [in this paper we consider the performance measure of total deposits as in Boufounou (1995) because the data is easily collected and amenable to statistical analysis. For a discussion of the drawbacks of this simple measure see Avkiran (1997)]. We are especially interested in investigating the effect of local wealth (as measured by county unemployment rates and income per capita) and local bank competition. The later, measured as the sum of distances of a branch to other branches of other banks, is shown to influence total deposits. The statistical model utilized is chosen from the class of linear mixed models. In particular, a non-hierarchical model specification with interaction effects is shown to be preferred over a standard linear model as well as a hierarchical specification. Finally, the predictive capabilities of the models are investigated.

## 2 Linear mixed models

Introductions to the theory and the use of linear mixed models can be found e.g. in West et al. (2006) and in Pinheiro and Bates (2000). The latter also describes the *R*-library *nlme* (Pinheiro et al. (2009)) which is designed for statistical analyses with mixed models and used in this case study.

The well-known standard linear model can be written as

$$Y = X\beta + \varepsilon,$$

where  $Y \in \mathbb{R}^n$  denotes the response vector,  $X \in \mathbb{R}^{n \times p}$  is the design matrix,  $\beta \in \mathbb{R}^p$  are the regression coefficients, and  $\varepsilon \in \mathbb{R}^n$  is the vector of random errors. Usually one assumes  $\varepsilon \sim N_n(0, \sigma^2 I_n)$ , which implies response variables are independent. Often, this is not the case in observational studies. Data could be clustered due to hierarchical structure (nested within levels/groups) or longitudinal (measured at multiple occasions or times). Linear mixed models, an extension of the standard linear model that contains both fixed and random effects, are useful for modeling such data and could be written as

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_i,$$

where the first part is simply a standard linear model for the  $n_i$  observations in group  $i$ . Additionally,  $Z_i \in \mathbb{R}^{n_i \times q}$  is the design matrix for the random effects with  $q \leq p$  and  $b_i \in \mathbb{R}^q$  being the random-effect coefficients. Since random effects and errors are random, the following distributions are specified:

$$\varepsilon_i \sim N_{n_i}(0, R_i), \quad b_i \sim N_q(0, G),$$

where  $\varepsilon_i$  and  $b_i$  are independent.  $R_i \in \mathbb{R}^{n_i \times n_i}$  and  $G \in \mathbb{R}^{q \times q}$  are the covariance matrices for the errors and the random effects respectively. Usually it is assumed that  $R_i = \sigma^2 I_{n_i}$  with  $\sigma^2 > 0$ , and that  $G$  is the same for all groups  $i$ . However, heteroscedasticity or correlation among within-group errors can be captured in more general specifications of  $R_i$ .

Level	Name	Description
State	<i>no.fail</i>	number of branches that closed in NY during the year
	<i>mshare</i>	market share in NY
	<i>branch.total</i>	share of the number of branches in NY compared to the USA
	<i>dep.total</i>	share of the total deposits of the bank in NY compared to the USA
	<i>av.dep</i>	average deposit per bank in NY
County	<i>pop</i>	population in the county (in 1000)
	<i>inc.pc</i>	per capita income (in 1000)
	<i>unemp</i>	unemployment rate in the county
Branch	<i>branch</i>	branch identity number (constant over the years)
	<i>log.dep</i>	total deposits (in USD) in the branch in log form
	<i>comp</i> <sup>1</sup>	measure of geographical competition of the branch (different for each year; values between 0 and 100, where a value of 100 is an indication of a high geographical competition)

**Table 1.** Variables on the state, county and branch level.

### 3 Geographical and macroeconomic determinants for total branch deposits

#### 3.1 Data

The data considered in this study consists of 2,988 branch-year records from 506 unique branches in the State of New York of a major US bank and covers the period from 1994 to 2002. The data is clustered: a branch resides within a county, which is located within a state. The data is also longitudinal: observations are made over a period of nine years. However, bank branches enter/ exit the market freely. So, not all branches provide nine years of data. Therefore, a mixed model is appropriate for modeling the dependencies in the data that were measured on the same statistical unit (branch) at various levels (county, state) over time.

The total deposits in logarithmic form, *log.dep*, is used as the dependent variable. Please find the definition of variable names in Table 1.

The relationship between *log.dep* and all three levels of covariates were examined using exploratory data analysis (EDA) tools. Scatter plots show a weak positive overall influence of *comp*, *pop*, and *inc.pc* on *log.dep*, but the variation is high. *unemp* shows no clear association with *log.dep*. In general, *log.dep* vary across branches and counties. State variables *mshare* and *av.dep* appear to be related to *log.dep*. EDA also suggests possible interaction effects that may mask the marginal effects of the independent variables. To reduce the complexity of the model, only second order interactions are considered.

<sup>1</sup> Scaled and standardized sum of the distances between a branch and all branches of other banks which have only one single branch or multiple branches respectively.

### 3.2 Analysis

In contrast to Boufounou (1995) and Avkiran (1997), we fit a regression model with mixed effects as described in section 2. The initial mixed model includes fixed effects for all branch, county and state variables and their interactions; random intercepts and slopes on the branch level  $b_{ij0}, b_{ij1}$ ; and random intercepts on the county level  $b_j$ .

The following distributions for the errors and the random effects are assumed (where  $n_i$  denotes the number of observations of branch  $i$ ):

$$\begin{aligned}\varepsilon_{ij} &= (\varepsilon_{ij1}, \dots, \varepsilon_{ijn_i})^T \sim N_{n_i}(0, \sigma^2 I_{n_i}), \\ b_j &\sim N(0, g_{00}^2),\end{aligned}\quad (1)$$

$$b_{ij} = (b_{ij0}, b_{ij1})^T \sim N_2(0, G) \text{ with } G = \begin{pmatrix} g_0^2 & g_{01} \\ g_{01} & g_1^2 \end{pmatrix}. \quad (2)$$

This model is not hierarchical because the random effects  $b_{ij0}$  and  $b_{ij1}$  are crossed with the fixed effects of the county variables (e.g.  $pop_{jt}$ ). The structure of the random effects is examined by testing whether the random effects specified in this model should be included. While the random effects for the branch level intercept and slope of *comp* are significant ( $p$ -value  $< .0001$ ), the random effects for the intercept on the county level  $b_j$  are not. They are subsequently removed from the model. Indeed, according tests show that both branch level random effects stay significant.

Since the number of observations and the values of *log.dep* vary over the years, the within-group errors might be different over time. Therefore, heterogeneous residual variances  $\sigma_t^2$  for each year  $t, t = 1994, \dots, 2002$  are considered:

$$\varepsilon_{ijt} \sim N(0, \sigma_t^2). \quad (3)$$

The test of this variance structure ( $H_0 : \sigma_t^2 = \sigma^2$  for each year  $t, t = 1994, \dots, 2002$  at the 5% level) shows a significant improvement in the fit ( $p$ -value  $< .0001$ ).

The variance structure can be extended further. Since the observations are taken longitudinally on the same statistical unit (branch), the within-group (i.e. within-branch) errors are probably autocorrelated. Further, because data is collected from at most 9 time points (1994-2002), only the first three or four lags should be considered. Since the empirical autocorrelation function from the residuals of the previous model shows that the autocorrelation of the first lag is significantly not equal to zero, an  $AR(1)$  model is chosen as correlation structure. An additional moving average term is also included, giving an  $ARMA(1, 1)$  model. Note that the parameters  $\phi_1$  and  $\theta_1$  are estimated using all available observations, not only with the at most 9 observations of a branch, and thus the same for all groups (cp. Pinheiro and Bates (2000)).

$$\begin{aligned}\varepsilon_{ijt} &= \phi_1 \varepsilon_{ijt-1} + \theta_1 a_{t-1} + a_t \\ \{a_t\} &= \text{zero mean white noise process with constant variance } \sigma_a^2\end{aligned}\quad (4)$$

Testing  $H_0 : \phi_1 = \theta_1 = 0$  at the 5% level confirms the significance of this extended variance structure ( $p$ -value  $< .0001$ ). As a result, this heterogeneous autoregressive variance structure of the errors is included in the model.



Variable	Estimate	Std. Error	p-value	Interact.	Estimate	Std. Error	p-value
<i>Intercept</i>	1.12 E+1	4.41 E-1	0.0000	<i>unemp</i> ×	-1.04 E-4	4.39 E-5	0.0184
<i>pop</i>	5.77 E-4	8.83 E-5	0.0000	<i>no.fail</i>			
<i>inc.pc</i>	-7.32 E-4	1.44 E-3	0.6121	<i>unemp</i> ×	1.37 E+0	3.67 E-1	0.0002
<i>unemp</i>	-2.69 E-1	6.91 E-2	0.0001	<i>mshare</i>			
<i>no.fail</i>	5.50 E-4	2.95 E-4	0.0624	<i>unemp</i> ×	1.05 E+0	2.82 E-1	0.0002
<i>mshare</i>	-6.23 E+0	2.23 E+0	0.0054	<i>branch.t</i>			
<i>branch.t</i>	-3.90 E+0	1.80 E+0	0.0303	<i>unemp</i> ×	-9.43 E-1	2.66 E-1	0.0004
<i>dep.total</i>	3.44 E+0	1.68 E+0	0.0410	<i>dep.total</i>			
<i>av.dep</i>	1.70 E-6	2.87 E-7	0.0000	<i>inc.pc</i> ×	1.05 E-8	3.94 E-9	0.0078
				<i>av.dep</i>			

**Table 2.** Significant effects with their estimates, standard errors and p-values in the final model ( $branch.total = branch.t$ ).

Finally, the model is reduced by a stepwise approach based on  $t$ -tests of the fixed effects at the 5% level ( $H_0 : \beta_i = 0$ ). In the resulting model all fixed effects are significant at the 5% level or left in the model in order to maintain the hierarchical structure of the fixed effects. All significant effects are displayed with their estimated regression coefficients in Table 2.

The assumption of the error distributions specified in this final model is then examined. The within-group errors are assumed to have a heterogeneous autoregressive variance structure (compare (1), (3) and (4)):  $\varepsilon_{ijt} \sim N(0, \sigma_t^2)$  and  $\varepsilon_{ijt} = \phi_1 \varepsilon_{ijt-1} + \theta_1 a_{t-1} + a_t$ , where  $\{a_t\}$  is a zero mean, white noise process with constant variance  $\sigma_a^2$ . Therefore, the errors depend on those of the years prior to and including  $t$ . As  $\phi_1 = 0.73$ , one expects approximately similarly distributed standardized within-group residuals per year. In fact, 93.1% of the residuals lie in the  $[-2, 2]$ -band, i.e. the approximate 95% confidence band. Checking the assumption of normality, QQ-plots for each year showed that the model fit is quite good for some years, but there are clear deviations from normality in other years.

The final model includes random effects for the intercept and for the slope of *comp* on the branch level with distribution as given in (2). Empirical Best Linear Unbiased Predictors of the random effects for each branch confirm the zero-mean assumption, while the marginal normality of the random effects was examined through QQ-plots which show that the assumption is plausible.

The final fitted model contains five interactions of county variables with state variables. Among those, four are interactions with *unemp* and one with *inc.pc*. The interaction effects complicate the interpretation of main effects. In order to facilitate this, we considered the expected deposits at different levels of the covariates: first, each county variable is considered at its 25%- and its 75%-quantile (denoted by 'low' and 'high'), the state variables are taken at their respective medians, and, second, we did it the other way around with the county variables at their medians. The results are shown in Table 3 (only the best four and the worst four combinations of the variables are displayed for the second comparison).

Next, we investigate if the fitted non-hierarchical linear mixed model (NHLMM) is an improvement over a standard linear model in modeling *log.dep*. The comparison of NHLMM to a linear model with the same fixed effects shows that the AIC of NHLMM is much smaller: 531 vs. 7937 of the linear model. Furthermore, a com-

<i>pop</i>	<i>inc.pc</i>	<i>unemp</i>	Deposits	<i>no.f</i>	<i>msh</i>	<i>branch.t</i>	<i>dept</i>	<i>av.dep</i>	Deposits
high	high	high	77,368	low	high	high	low	high	139,487
high	high	low	75,756	high	high	high	low	high	139,043
high	low	high	73,061	low	low	high	low	high	118,323
high	low	low	71,539	high	low	high	low	high	117,947
low	high	high	66,975	low	high	low	high	low	58,184
low	high	low	65,580	high	high	low	high	low	57,999
low	low	high	63,248	low	low	low	high	low	49,356
low	low	low	61,930	high	low	low	high	low	49,199

**Table 3.** Expected deposits at different levels of the county and state variables, respectively, using the abbreviations  $no.fail = no.f$ ,  $mshare = msh$ ,  $branch.total = branch.t$  and  $dep.total = dept$ .

parison was made between NHLMM and the generalized least squares (GLS) model with heteroscedastic and correlated within-group errors (but no random effects). The latter model also shows a larger AIC than NHLMM: 531 vs. 3675 of the GLS model. Thus, the additional inclusion of random effects in a mixed model, modeling the variability and dependency, lead to a considerable improvement in the fit.

A hierarchical linear mixed model was also considered. Following the same process described in model selection, the following main effects and interactions are kept in the model: an intercept and all branch, county and state variables except for  $no.fail$  as well as the interactions  $av.comp \times mshare$ ,  $av.pop \times mshare$ ,  $av.unemp \times mshare$ ,  $av.unemp \times branch.total$  and  $av.unemp \times dep.total$ . This model is denoted as HLMM.

QQ-plots on residuals showed clear deviations from normality, suggesting the distributional assumptions on the errors for the hierarchical mixed model are probably not accurate. However, the distributional assumptions on the random effects are acceptable.

### 3.3 Prediction

To evaluate the predictive capability of the NHLMM model and the HLMM model, the following intuitive approach is taken: Both NHLMM and HLMM are estimated using the data from 1994 to 2001, with the same fixed effects, same random effects, and same variance structures. The fitted models are then used to predict the response for Year 2002.

A comparison of predicted and observed values of 2002 shows that the predictive capability of NHLMM is quite good and better than that of HLMM, which underestimates the observed values. The sum of squared deviations of HLMM is much larger: 126 vs. 18 of NHLMM.

### 3.4 Results

Analyses showed that the predictive capability of the NHLMM model is superior to that of the HLMM model. So, interpretation regarding the influences on the total deposits of a bank branch is based on the results of the NHLMM model as developed in section 3.2.

The examination of the data showed that the total deposits of a bank branch depend significantly on local geographic effects such as wealth and competition. Deposits positively depend on the county's population and on the per capita income, but the overall effect of the unemployment rate is unclear because it interacts with other effects. Likewise, there is no uniform influence of the local competition on bank deposits. When competition increases, it stimulates business but, when there is more competition each branch enjoys a smaller market share.

To investigate certain influences, the following approach is helpful: all branches are classified as being in a 'rural'/'urban', 'poor'/'rich' area with a low/high unemployment rate (the classification is done using the medians of the respective variables). For all branches in a specific area the branch-specific effects of the competition are averaged and then compared.

Besides these dependencies, there is an additional branch-specific effect: some branches have more deposits than others if all other influences are disregarded. This can be explained by specific characteristics of a branch such as a long-term customer loyalty or a particular good location in an area.

Detailed results and plausible explanations will be presented at the conference.

## 4 Summary and discussion

Our approach illustrates the potential of linear mixed models in the context of measuring branch performance as an aid to optimizing branch location decisions. Compared to DEA, regression analysis summarizes performance information of *all* branches in the sample in a model that could be used to forecast deposits of new branches (Boufounou (1995)). As such, it allows an easy evaluation of a single existing branch and of the potential of a new location. In contrast to Boufounou (1995) and Avkiran (1997), we include interactions and random effects in our models in order to take into account different local market environments and thus make the model more reliable.

While the specific performance measure of total deposits is easily available for a statistical analysis, other performance variables (e.g. fee income) could easily be included in a mixed-effects regression analysis. Similarly, other in-branch variables or competitive factors could also be included in the modeling to improve the fit further, but would increase the computational complexity. However we used all the covariates available to us. A detailed explorative data analysis is crucial to identify potential random effects and interactions before fitting an initial model in order to ensure a model which can be appropriately analyzed and interpreted.

## References

- AVKIRAN, N. K. (1997): Models of retail performance for bank branches: predicting the level of key business drivers. *International Journal of Bank Marketing* 15 (6), 224-237.
- BERGER, A. N. and HUMPHREY, D. B. (1997): Efficiency of financial institutions: International survey and directions for future research. *European Journal of Operational Research* 98 (2), 175-212.

- BOUFOUNOU, P. V. (1995): Evaluating bank branch location and performance: A case study. *European Journal of Operational Research* 87, 389-402.
- CHELST, K. R., SCHULTZ, J. P. and SANGHVI, N. (1988): Issues and decision aids for designing branch networks. *Journal of Retail Banking* 10 (2), 5-17.
- DEYOUNG, R., HUNTER, W. C. and UDELL, G. F. (2004): The past, present, and probable future for community banks. *Journal of Financial Services Research* 25, 85-133.
- DOYLE, P., FENWICK, L. and SAVAGE, G. P. (1979): Management planning and control in multi-branch banking. *Journal of Operational Research Society* 30 (2), 105-111.
- GART, A. (1994): *Regulation, Deregulation, Reregulation*. Wiley, New York.
- JAYARATNE, J. and STRAHAN, P. E. (1996): The finance-growth nexus: evidence from bank branch deregulation. *Quarterly Journal of Economics* 111, 639-670.
- PINHEIRO, J. C. and BATES, D. M. (2000): *Mixed-effects models in S and S-PLUS*. Springer, New York.
- PINHEIRO, J. C., BATES, D. M., DEBROY, S., SARKAR, D., and the R Core team (2009): nlme: Linear and Nonlinear Mixed Effects Models. *R package version 3.1-96*.
- SHIERS, A. F. (2002): Bank branching, economic diversity and bank risk. *The Quarterly Review of Economics and Finance* 42, 587-598.
- WEST, B. T., WELCH, K. B. and GALECKI, A. T. (2006): *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall/CRC, Boca Raton.

# Numerical Methods for some Classes of Matrices with Applications to Statistics and Optimization

Juan M. Peña<sup>1</sup>

Departamento de Matemática Aplicada  
Universidad de Zaragoza, 50009 Zaragoza, Spain, *jmpena@unizar.es*

**Abstract.** Recent advances on numerical methods for matrices with applications to Statistics and Optimization are presented. We first consider sign-regular matrices as well as some subclasses of these matrices. The interest of these matrices comes from their characterization as variation diminishing linear maps, which leads to a unified presentation of hypothesis tests. We also consider the class of  $H$ -matrices, which plays an important role in linear complementarity problems, and some classes of  $P$ -matrices.

**Keywords:** numerical algorithms, statistical computing, sign-regular matrices,  $H$ -matrices, error bounds, conditioning

## 1 Introduction

This paper surveys some recent developments on advances on numerical methods for matrices with applications to Statistics and Optimization. An important advance performed during the last years consists of finding some classes of matrices for which relevant computations (such as computing singular values or solving linear systems) can be carried out with high relative accuracy. These classes of matrices are formed either by sign-regular matrices or by  $H$ -matrices diagonally dominant. As we shall recall in Sections 2 and 3, respectively, both classes of matrices, sign-regular matrices and  $H$ -matrices, present important applications to Statistics and Optimization. We also present other recent developments for numerical methods related to these matrices, including results on the localization of eigenvalues, optimally conditioned matrices and error bounds for the linear complementarity problem.

Section 2 is devoted to the advances for the sign-regular matrices. As recalled in this section, the interest of these matrices comes from their characterization as variation diminishing linear maps, which leads to a unified presentation of hypothesis tests (see Brown et al. (1981)). We also focus on its important subclass of totally positive matrices (arising with many important density functions occurring in statistical theory, such as the Normal, Gamma, Binomial or Poisson distributions) and, in particular, on Bernstein-Vandermonde matrices. These matrices present interesting probabilistic interpretations (see Goldman (1985), Goldman (1988, 1) and Goldman (1988, 2)), and they satisfy a nice result on optimal conditioning.

Section 3 is devoted to the advances for the  $H$ -matrices and to its related class of  $P$ -matrices (an  $H$ -matrix whose diagonal entries are positive is a  $P$ -matrix).

$P$ -matrices and  $H$ -matrices plays an important role in the linear complementarity problem (see Chen and Xiang (2006) and Chen and Xiang (2007)). New results on error bounds for this problem are presented. The paper finishes with some concluding remarks in Section 4.

## 2 Sign-regular matrices: recent numerical advances

An  $n \times m$  matrix  $A$  is *sign-regular* if, for each  $k$  ( $1 \leq k \leq \min\{n, m\}$ ), all  $k \times k$  submatrices of  $A$  have determinant with the same sign. The interest of nonsingular sign-regular matrices comes from their characterization as variation diminishing linear maps: the number of sign changes in the consecutive components of the image of a vector is bounded above by the number of sign changes in the consecutive components of the vector (see Ando (1987)). Many applications of these matrices come from this property. In particular, the variation diminishing property allows a unified presentation of hypothesis tests, as shown in Brown et al. (1981). Other statistical applications of this property and of sign-regular matrices can be seen in Karlin (1968). Algebraic properties of nonsingular sign-regular matrices can be seen in Peña (2003, 1). An advantageous pivoting strategy for these matrices was introduced in Peña (1997). Later, this pivoting strategy was used in García-Esnaola and Peña (2009, 1) to derive a stable dual simplex method for linear programming problems whose associated matrices are sign-regular.

If all minors of a sign-regular matrix are nonzero, then the matrix is called *strictly sign-regular*. Factorizations of these matrices with applications in numerical methods were analyzed in Cortés and Peña (2008, 1), where  $LDU$ ,  $QR$  and symmetric-triangular factorization were considered. A stable test to check if a given matrix is strictly sign-regular has been obtained in Cortés and Peña (2008, 2). A crucial tool to derive this test has been the use of an elimination procedure called *Neville elimination*. Roughly speaking, Neville elimination is an elimination procedure to produce zeros in each column of a matrix alternative to Gauss elimination, in the sense that we subtract to each row an adequate multiple of the previous one. This elimination procedure has also been used to provide a factorization of sign-regular matrices in terms of bidiagonal factors, which has been recently very useful to obtain numerical algorithms with high relative accuracy for the singular values of matrices belonging to some families of sign-regular matrices. Below we shall mention one of these families. Finding other classes of sign-regular matrices with accurate numerical algorithms is a very important open problem.

Let us now focus on a very important subclass of sign-regular matrices. A matrix with all its minors nonnegative is called *totally positive* (TP). Many applications of TP matrices are mentioned in Karlin (1968) and in Ando (1987). In particular, in Section 2 of Chapter 1 of Karlin (1968) it is shown that many density functions occurring in statistical theory are totally positive, in the sense that associated collocation matrices are totally positive. For instance, the Normal, Gamma, Binomial or Poisson distributions are totally positive.

It is well known that the eigenvalues of a nonsingular TP matrix are positive, as can be seen in Ando (1987). Lower bounds of the minimal eigenvalue of a TP matrix can be used in many practical issues. The following result of Peña (2009) provides such a bound, improving the corresponding one of Gerschgorin circles theorem. We

shall use the following notation for index subsets: given  $i \in \{1, \dots, n\}$  let

$$J_i := \{j \mid |j - i| \text{ is odd}\}, \quad K_i := \{j \neq i \mid |j - i| \text{ is even}\}. \quad (1)$$

**Theorem 1.** *Let  $A$  be a nonsingular totally nonnegative matrix, and let  $\lambda_{\min}(> 0)$  be its minimal eigenvalue. For each  $i \in \{1, \dots, n\}$ , let  $J_i$  be the index subset defined by (1) Then:*

$$\lambda_{\min} \geq \min_i \{a_{ii} - \sum_{j \in J_i} a_{ij}\}. \quad (2)$$

An important source of TP matrices comes from the collocation matrices of systems of functions. This provides many applications of TP matrices to Computer Aided Geometric Design and to Approximation Theory. Let  $\mathcal{U}$  be a vector space of real functions defined on a real interval  $I$  and  $(u_0(t), \dots, u_n(t))$  ( $t \in I$ ) be a basis of  $\mathcal{U}$ . The *collocation matrix* of  $(u_0(t), \dots, u_n(t))$  at  $t_0 < \dots < t_m$  in  $I$  is given by the matrix

$$M \begin{pmatrix} u_0(t), \dots, u_n(t) \\ t_0, \dots, t_m \end{pmatrix} := (u_j(t_i))_{i=0, \dots, m; j=0, \dots, n}. \quad (3)$$

A system of functions is TP when all its collocation matrices (3) are TP. In CAGD the functions  $u_0, \dots, u_n$  also satisfy that  $\sum_{i=0}^n u_i(t) = 1 \forall t \in I$  (i.e., the system  $(u_0, \dots, u_n)$  is *normalized*), and a normalized TP system is denoted by NTP. In fact, in Computer Aided Geometric Design shape preserving representations are associated with NTP bases.

An important NTP basis is the Bernstein basis  $(b_0^n, \dots, b_n^n)$  of the space  $P_n$  of polynomials of degree less than or equal to  $n$  on  $[0, 1]$ , given by

$$b_i^n(t) = \binom{n}{i} t^i (1-t)^{n-i}, \quad i = 0, 1, \dots, n.$$

This basis is widely used in Computer Aided Geometric Design due to its good properties. This basis and other bases used frequently in Computer Aided Geometric Design (as B-spline basis) have a nice probabilistic interpretation, as shown in Goldman (1985), Goldman (1988, 1) and Goldman (1988, 2). Collocation matrices of Bernstein basis are called Bernstein-Vandermonde matrices. In Marco and Martínez (2008), numerical algorithms with high relative accuracy have been derived for these matrices.

Now we present a recent result on the Bernstein-Vandermonde matrices published in Delgado and Peña (2009). First, we need some notation. Given a matrix  $A = (a_{ij})_{1 \leq i, j \leq n}$ , we denote by  $|A|$  the matrix whose  $(i, j)$ -entry is  $|a_{ij}|$ . Given a nonsingular matrix  $A$ , let us consider the condition number

$$\kappa_{\infty}(A) := \|A\|_{\infty} \|A^{-1}\|_{\infty}.$$

Let us also recall the condition number

$$\text{Cond}(A) := \| |A^{-1}| |A| \|_{\infty},$$

introduced in Skeel (1979), which measure effects of perturbations of the data in linear systems. The following two properties hold:  $\text{Cond}(A) \leq \kappa_{\infty}(A)$  and it can be

much smaller, and, in contrast to  $\kappa_\infty(A)$ ,  $\text{Cond}(A)$  is invariant to row scaling: if  $D$  is a nonsingular diagonal matrix, then

$$\text{Cond}(DA) = \text{Cond}(A).$$

These properties provide some of the reasons that explain why the Skeel condition number  $\text{Cond}(A)$  is more satisfying than the traditional condition number  $\kappa_\infty(A)$ .

The first inequality of the following result (obtained in Delgado and Peña (2009)) shows that the collocation matrices of the Bernstein basis are the best conditioned among all the corresponding collocation matrices of NTP bases of the space  $P_n$  on  $[0, 1]$ . The second inequality of the following result shows that the transposes of the collocation matrices of the Bernstein basis are the best conditioned for the Skeel condition number among all the transposes of the corresponding collocation matrices of TP bases of the space  $P_n$  on  $[0, 1]$ . Let us recall that The transposes of collocation matrices arise in the construction of formulas of numerical integration and numerical differentiation of interpolatory type.

**Theorem 2.** *Let  $(b_0, \dots, b_n)$  be the Bernstein basis, let  $(v_0, \dots, v_n)$  be another TP basis of  $P_n$  on  $[0, 1]$ , and let  $0 \leq t_0 < t_1 < \dots < t_n \leq 1$ ,  $V := M_{t_0, \dots, t_n}^{(v_0, \dots, v_n)}$ , and  $B := M_{t_0, \dots, t_n}^{(b_0, \dots, b_n)}$ . Then*

- (i)  $\kappa_\infty(B) \leq \kappa_\infty(V)$  if  $(v_0, \dots, v_n)$  is normalized.
- (ii)  $\text{Cond}(B^T) \leq \text{Cond}(V^T)$ .

### 3 H-matrices and P-matrices: recent numerical advances

In this section we comment recent numerical results for  $P$ -matrices and related classes of matrices. We also illustrate some applications of these classes of matrices to optimization. In particular, a recent result for the error bound of the linear complementarity problem is presented.

Let us recall that an  $n \times n$  real matrix  $M$  is a  $P$ -matrix if all its principal minors are positive. Recent results on eigenvalue bounds for some subclasses of  $P$ -matrices are provided in Peña (2009).

Let us now introduce some subclasses of  $P$ -matrices. As recalled in Ando (1987), a nonsingular TP matrix is a  $P$ -matrix. A nonsingular matrix  $M$  is called an  $M$ -matrix if its inverse is nonnegative and all its off-diagonal entries are nonpositive. It is also known that nonsingular  $M$ -matrices are  $P$ -matrices. Given a matrix  $M = (m_{ij})_{1 \leq i, j \leq n}$ , its comparison matrix  $\tilde{M} = (\tilde{m}_{ij})_{1 \leq i, j \leq n}$  has entries  $\tilde{m}_{ii} := |m_{ii}|$  and  $\tilde{m}_{ij} := -|m_{ij}|$  for all  $j \neq i$  and  $i, j = 1, \dots, n$ . We say that a matrix is an  $H$ -matrix if its comparison matrix is a nonsingular  $M$ -matrix. Finally, we say that a matrix  $M = (m_{ij})_{1 \leq i, j \leq n}$  is *strictly diagonally dominant* (by rows) if  $|m_{ii}| > \sum_{j \neq i} |m_{ij}|$  for all  $i = 1, \dots, n$ . If  $|m_{ii}| \geq \sum_{j \neq i} |m_{ij}|$  for all  $i = 1, \dots, n$ , then we say that the matrix  $M$  is *diagonally dominant* (by rows). An  $H$ -matrix with positive diagonals is a  $P$ -matrix and a strictly diagonally dominant matrix is an  $H$ -matrix.

Algorithms with high relative accuracy that can be used for the computation of  $M$ -matrices diagonally dominant can be found in Demmel and Koev (2004) and Peña (2004). They use complete pivoting and a diagonal dominance pivoting



(analyzed in Peña (1998)), respectively. More recently, both pivoting strategies have been used in Ye (2008) to obtain algorithms with high relative accuracy for the general class of diagonally dominant matrices.

Subclasses of  $P$ -matrices have also provided localization results for the eigenvalues of a square real matrix, which can be considered alternative to Gerschgorin circles. This was proved in Peña (2001) and Peña (2003, 2) the subclass of  $B$ -matrices, which will be introduced later.

The linear complementarity problem consists of finding vectors  $x \in \mathbf{R}^n$  satisfying

$$Mx + q \geq 0, \quad x \geq 0, \quad x^T(Mx + q) = 0, \quad (4)$$

where  $M$  is an  $n \times n$  real matrix and  $q \in \mathbf{R}^n$ . We denote this problem by  $\text{LCP}(M, q)$  and its solutions by  $x^*$ . Many problems can be posed in the form (4). For instance, problems in linear and quadratic programming, the problem of finding a Nash equilibrium point of a bimatrix game or some free boundary problems of fluid mechanics. As recalled in Chen and Xiang (2007),  $M$  is a  $P$ -matrix if and only if the  $\text{LCP}(M, q)$  has a unique solution  $x^*$  for any  $q \in \mathbf{R}^n$ . In Chen and Xiang (2007), error bounds for  $\|x - x^*\|$  were derived when  $M$  in (4) is a  $P$ -matrix.

If  $M$  is a  $P$ -matrix, then we can apply the third inequality of Theorem 2.3 of Chen and Xiang (2006) and obtain for any  $x \in \mathbf{R}^n$  the inequality:

$$\|x - x^*\|_\infty \leq \max_{d \in [0, 1]^n} \|(I - D + DM)^{-1}\|_\infty \|r(x)\|_\infty, \quad (5)$$

where  $I$  is the  $n \times n$  identity matrix. The diagonal matrix  $D = \text{diag}(d_i)$  with  $0 \leq d_i \leq 1$  for all  $i = 1, \dots, n$ ,  $x^*$  is the solution of the  $\text{LCP}(M, q)$  and  $r(x) := \min(x, Mx + q)$ , where the min operator denotes the componentwise minimum of two vectors.

If the  $P$ -matrix also satisfies additional properties, then additional consequences can be derived on the corresponding linear complementarity problem. A typical example is provided by the subclass of  $P$ -matrices given by the  $H$ -matrices with positive diagonals. There are many bounds for the linear complementarity problem when the involved matrix is an  $H$ -matrix with positive diagonals (see also Chen and Xiang (2007)). By (2.4) of Chen and Xiang (2006), given in Theorem 2.1 of Chen and Xiang (2006), when  $M = (m_{ij})_{1 \leq i, j \leq n}$  is an  $H$ -matrix with positive diagonals, then

$$\max_{d \in [0, 1]^n} \|(I - D + DM)^{-1}\|_\infty \leq \|\tilde{M}^{-1} \max(\Lambda, I)\|_\infty, \quad (6)$$

where  $\tilde{M}$  is the comparison matrix of  $M$ ,  $\Lambda$  is the diagonal part of  $M$  ( $\Lambda := \text{diag}(m_{ii})$ ) and  $\max(\Lambda, I) := \text{diag}(\max\{m_{11}, 1\}, \dots, \max\{m_{nn}, 1\})$ .

We now mention a recent bound derived for another class of  $P$ -matrices. A square real matrix  $A = (a_{ik})_{1 \leq i, k \leq n}$  with positive row sums is a  $B$ -matrix if all its off-diagonal elements are bounded above by the corresponding row means, i.e., for all  $i = 1, \dots, n$

$$\sum_{k=1}^n a_{ik} > 0, \quad \text{and} \quad \frac{1}{n} \left( \sum_{k=1}^n a_{ik} \right) > a_{ij} \quad \forall j \neq i.$$

$B$ -matrices have been used to derive localization results for the real eigenvalues of a matrix alternative to Gerschgorin circles (see Peña (2001) and Peña (2003, 2)).

By Corollary 2.6 of Peña (2001), a  $B$ -matrix is a  $P$ -matrix. However, there exist  $B$ -matrices that are not  $H$ -matrices. For instance, for any  $0 < \varepsilon < 1$ , the matrix

$$A = \begin{pmatrix} 1 + \varepsilon & 1 & 1 \\ 1 & 1 + \varepsilon & 1 \\ 1 & 1 & 1 + \varepsilon \end{pmatrix}$$

is a  $B$ -matrix and is not an  $H$ -matrix.

Since a  $B$ -matrix  $M$  is a  $P$ -matrix, then we can apply (5) to  $M$ . The following result corresponds to Theorem 2.2 of García-Esnaola and Peña (2009, 2) and provides a bound for  $\max_{d \in [0,1]^n} \|(I - D + DM)^{-1}\|_\infty$  when  $M$  is an  $n \times n$   $B$ -matrix. This bound can be calculated with  $O(n^2)$  elementary operations. We first need an auxiliary notation. Given a real matrix  $M = (m_{ij})_{1 \leq i, j \leq n}$ , for each  $i = 1, \dots, n$  let us define

$$r_i^+ := \max\{0, m_{ij} | j \neq i\}.$$

Then we can write  $M = B^+ + C$ , where

$$B^+ = \begin{pmatrix} m_{11} - r_1^+ & \dots & m_{1n} - r_1^+ \\ \vdots & & \vdots \\ m_{n1} - r_n^+ & \dots & m_{nn} - r_n^+ \end{pmatrix} \text{ and } C = \begin{pmatrix} r_1^+ & \dots & r_1^+ \\ \vdots & & \vdots \\ r_n^+ & \dots & r_n^+ \end{pmatrix}. \quad (7)$$

By Proposition 2.3 of García-Esnaola and Peña (2009, 2),  $M$  is a  $B$ -matrix if and only if  $B^+$  is strictly diagonal dominant by rows with positive diagonal entries.

**Theorem 3.** *Suppose that  $M = (m_{ij})_{1 \leq i, j \leq n}$  is an  $n \times n$   $B$ -matrix and let  $B^+ = (b_{ij})_{1 \leq i, j \leq n}$  be the matrix of (7). Let  $\beta_i := b_{ii} - \sum_{j \neq i} |b_{ij}|$  and  $\beta := \min_{i \in \{1, \dots, n\}} \{\beta_i\}$ . Then*

$$\max_{d \in [0,1]^n} \|(I - D + DM)^{-1}\|_\infty \leq \frac{n-1}{\min\{\beta, 1\}}.$$

Sharpness of the previous bound is illustrated in García-Esnaola and Peña (2009, 2).

## 4 Concluding remarks

Sign-regular matrices and  $P$ -matrices have arisen in many applications of Statistics and Optimization. For some subclasses of these matrices, accurate computations can be performed, due to some recent advances commented in the paper. Relationship with eigenvalue localization has been considered and matrices with optimal conditioning (Bernstein-Vandermonde matrices) are also found. Recent results on new classes of  $P$ -matrices (such as  $B$ -matrices) with error bounds for the linear complementarity problem are also presented.

## Acknowledgements

This work has been partially supported by the Spanish Research Grant MTM2009-07315 and by Gobierno de Aragón.

## References

- ANDO, T.(1987): Totally positive matrices. *Linear Algebra Appl.* 90 , 165-219.
- BROWN, L. D., JOHNSTONE, I. M. and MacGIBBON, K. B. (1981): Variation Diminishing transformations: a direct approach to total positivity and its statistical applications. *J. Amer. Statist. Assoc.* 76, 824-832.
- CHEN, X. and XIANG, S. (2006): Computation of error bounds for P-matrix linear complementarity problems. *Math. Program., Ser. A*, 106, 513-525.
- CHEN, X. and XIANG, S. (2007): Perturbation bounds of P-matrix linear complementarity problems. *SIAM J. Opt.* 18, 1250-1265.
- CORTES, V. and PEÑA J. M. (2008, 1): A stable test for strict sign regularity. *Math. Comp.* 77, 2155-2171.
- CORTES, V. and PEÑA J. M. (2008, 2): Decompositions of strictly sign regular matrices. *Linear Algebra Appl.* 429, pp. 1071-1081.
- DELGADO, J. and PEÑA, J. M. (2009): Optimal conditioning of Bernstein collocation matrices. *SIAM J. Matrix Anal. Appl.* 31, 990-996.
- DEMMEL, J. and KOEV, P. (2004): Accurate SVDs of weakly diagonally dominant  $M$ -matrices. *Numer. Math.* 98, 99-104.
- GARCIA-ESNAOLA, M. and PEÑA, J. M. (2009, 1): Sign consistent linear programming problems. *Optimization* 58, 935-946.
- GARCIA-ESNAOLA, M. and PEÑA, J. M. (2009, 2): Error bounds for linear complementarity problems for  $B$ -matrices. *Appl. Math. Lett.* 22, 1071-1075.
- GOLDMAN, R. N. (1985): Pólya's urn model and computer aided geometric design. *SIAM J. Algebraic Discrete Methods* 6, 1-28.
- GOLDMAN, R. N. (1988, 1): Urn models and  $B$ -splines. *Constr. Approx.* 4, 265-288.
- GOLDMAN, R. N. (1988, 2): Urn models, approximations, and splines. *J. Approx. Theory* 54, 1-66.
- KARLIN, S. (1968): *Total Positivity*. Stanford University Press, Stanford.
- MARCO, A. and MARTINEZ, J. J. (2007): A fast and accurate algorithm for solving Bernstein-Vandermonde linear systems. *Linear Algebra Appl.* 422 , 616-628.
- PEÑA, J. M. (1997): Backward stability of a pivoting strategy for sign-regular linear systems. *BIT* 37, 910-924.
- PEÑA, J. M. (1998): Pivoting strategies leading to diagonal dominance by rows. *Numer. Math.* 81, 293-304.
- PEÑA, J. M. (2001): A class of P-matrices with applications to the localization of the eigenvalues of a real matrix. *SIAM J. Matrix Anal. Appl.* 22, 1027-1037.
- PEÑA, J. M. (2003, 1): On nonsingular sign regular matrices. *Linear Algebra Appl.* 359, 91-100.
- PEÑA, J. M. (2003, 2): On an alternative to Gerschgorin circles and ovals of Cassini. *Numer. Math.* 95, 337-345.
- PEÑA, J. M. (2004): LDU decompositions with L and U well conditioned. *Electronic Transactions of Numerical Analysis* 18, 198-208.
- PEÑA, J. M. (2009): Eigenvalue bounds for some classes of P-matrices. *Numerical Linear Algebra with Applications* 16, 871-882.
- PRÉKOPA, A. (1990): Sharp Bounds on Probabilities using Linear Programming. *Operation Research* 38, 227-239.

- SKEEL, R. D. (1979): Scaling for numerical stability in Gaussian elimination. *J. Assoc. Comput. Mach.* 26, 494-526.
- YE, Q. (2008): Computing singular values of diagonally dominant matrices to high relative accuracy. *Math. Comp.* 77, 2195-2230.

# Maximum Margin Learning of Gaussian Mixture Models with Application to Multipitch Tracking

Franz Pernkopf and Michael Wohlmayr\*

Signal Processing and Speech Communication Laboratory (SPSC)  
Graz University of Technology, Austria.  
*pernkopf@tugraz.at*  
*michael.wohlmayr@tugraz.at*

**Abstract.** We present a maximum-margin based learning algorithm for Gaussian mixture models. In contrast to existing methods, our approach includes the sum-to-one constraint of probabilistic models. Model parameters are optimized by maximizing the margin between training samples of distinct classes. Optimization is based on the extended Baum-Welch procedure, which attains a local maximum of the proposed optimization criterion. We apply the proposed algorithm to the task of multipitch tracking given single-channel recordings of two simultaneously speaking subjects. Using the mixture-maximization interaction model, we are able to combine classifiers trained on single speakers to classify the mixture of both speakers. We demonstrate the superior performance over generative training based on the expectation maximization algorithm under low-noise conditions.

**Keywords:** Gaussian mixture model, discriminative classifiers, extended Baum Welch, maximum margin learning

## 1 Introduction

One of the most successful discriminative classifiers, namely the support vector machine (SVM), maximizes the margin of confidence between samples of distinct class label, which leads to a good generalization performance (Vapnik, 1998). Taskar et al. (2003) observed that undirected graphical models can be efficiently trained to maximize the margin. More recently, Guo et al. (2005) introduced the maximization of the margin to Bayesian networks. Unlike in undirected graphical models, the main difficulty for Bayesian networks is the normalization constraint of the local conditional probabilities. Guo et al. (2005) relaxed this constraint to obtain a convex optimization problem. Sha and Saul (2006) applied margin optimization to Gaussian mixture models (GMMs), but similar as above, the normalization constraint has been neglected leading to a convex optimization problem.

In this paper, we aim to follow a quite different approach to maximize the margin in GMMs. We keep the sum-to-one constraint, which maintains the probabilistic interpretation of the network, e.g. marginalization over missing variables is

---

\* This work was supported by the Austrian science fund (project number P19737-N15 and S10604-N13)

still possible. This also enables the application of the mixture-maximization (MIX-MAX) interaction model (Nadas et al. (1989)) for multipitch tracking, where we combine discriminatively learned single-speaker GMMs. We no longer have a convex optimization problem, however we can optimize the parameters using the extended Baum-Welch (EBW) algorithm (Gopalakrishnan et al. (1991)).

The outline of the paper is as follows: Section 2 describes the proposed max-margin training algorithm for Gaussian mixture classifiers. In section 3, we apply the proposed algorithm to the task of multipitch tracking. Section 4 presents experimental results, and section 5 concludes the paper.

## 2 Max-Margin Learning of GMMs

Given a feature vector  $\mathbf{x}$ , classification within a probabilistic framework is the task of finding the class  $c$  with maximal posterior probability, i.e.  $c^* = \arg \max_c p(c|\mathbf{x})$ . In a generative setting, we model the joint probability of class and observations,  $p(c, \mathbf{x}) = p(\mathbf{x}|c)p(c)$ , and obtain the posterior using Bayes rule:  $p(c|\mathbf{x}) = p(\mathbf{x}|c)p(c)/p(\mathbf{x})$ .

In this paper,  $p(\mathbf{x}|c)$  is modelled by a GMM,  $p(\mathbf{x}|\boldsymbol{\Theta}_c) = \sum_{m=1}^M \alpha_c^m \mathcal{N}(\mathbf{x}|\boldsymbol{\theta}_c^m)$ , where  $\boldsymbol{\Theta}_c = \{\alpha_c^m, \boldsymbol{\theta}_c^m\}_{m=1}^M$  are class-specific parameters, and  $M$  is the number of components. The component weights  $\alpha_c^m$  are constrained to be positive and sum to one, i.e.  $\sum_{m=1}^M \alpha_c^m = 1$ . Each Gaussian component is specified by the mean vector  $\boldsymbol{\mu}_c^m$  and covariance matrix  $\boldsymbol{\Sigma}_c^m$ , i.e.  $\boldsymbol{\theta}_c^m = \{\boldsymbol{\mu}_c^m, \boldsymbol{\Sigma}_c^m\}$ . Denoting the class prior as  $\rho_c = p(c)$ , the full parameter set of the model is given by  $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_c, \rho_c\}_{c=1}^C$ . Usually, each class conditional GMM is trained generatively, i.e. GMMs are optimized using the expectation-maximization (EM) algorithm, and  $\rho_c$  is the normalized frequency count of class  $c$  in the training data. In contrast, we use the multi-class margin (Crammer and Singer (2001)) of training sample  $n$ ,

$$d_{\boldsymbol{\Theta}}^n = \min_{c \neq c^n} \frac{p(c^n|\mathbf{x}^n, \boldsymbol{\Theta})}{p(c|\mathbf{x}^n, \boldsymbol{\Theta})} = \min_{c \neq c^n} \frac{p(c^n, \mathbf{x}^n|\boldsymbol{\Theta})}{p(c, \mathbf{x}^n|\boldsymbol{\Theta})} = \frac{p(c^n, \mathbf{x}^n|\boldsymbol{\Theta})}{\max_{c \neq c^n} p(c, \mathbf{x}^n|\boldsymbol{\Theta})}, \quad (1)$$

to optimize  $\boldsymbol{\Theta}$ . If  $d_{\boldsymbol{\Theta}}^n > 1$ , then sample  $n$  is correctly classified and vice versa. We replace the max operator by the differentiable approximation  $\max_x f(x) \approx [\sum_x (f(x))^\eta]^\frac{1}{\eta}$ , where  $\eta \geq 1$  and  $f(x)$  is non-negative. In the limit of  $\eta \rightarrow \infty$  the approximation converges to the max operation. Replacing the max with its approximation, we obtain

$$d_{\boldsymbol{\Theta}}^n = \frac{p(c^n, \mathbf{x}^n|\boldsymbol{\Theta})}{\left[ \sum_{c \neq c^n} (p(c, \mathbf{x}^n|\boldsymbol{\Theta}))^\eta \right]^\frac{1}{\eta}}.$$

Usually, the max margin approach maximizes the margin of the sample with the smallest margin, i.e.  $\min_{n=1, \dots, N} d_{\boldsymbol{\Theta}}^n$  for a separable classification problem (Schölkopf and Smola (2001)). We aim to relax this by introducing a soft margin, i.e. we focus on samples with a margin  $d_{\boldsymbol{\Theta}}^n$  close to one. Therefore, the *hinge* error function usually used in SVMs can be considered according to

$$C(\mathcal{X}|\boldsymbol{\Theta}) = \prod_{n=1}^N \min \left[ 2, (d_{\boldsymbol{\Theta}}^n)^\lambda \right],$$

where  $\mathcal{X}$  is the set of training samples. Maximizing this function with respect to the parameters  $\Theta$  implicitly means to increase the margin  $d_{\Theta}^n$  whereas the emphasis is on samples with a margin  $(d_{\Theta}^n)^{\lambda} < 2$ , i.e. samples with a large positive margin have no impact on the optimization. The parameter  $\lambda > 0$  scales the margin and is set by cross-validation. Maximizing  $C(\mathcal{X}|\Theta)$  with gradient based methods is not straightforward due to the discontinuity in the derivative at  $(d_{\Theta}^n)^{\lambda} = 2$ . Therefore, we propose to use instead a *smooth hinge* function  $h(y)$  which enables a smooth transition of the derivative and has similar shape as  $\min[2, y]$ :

$$h(y) = \begin{cases} y + \frac{1}{2}, & \text{if } y \leq 1 \\ 2 - \frac{1}{2}(y - 2)^2, & \text{if } 1 < y < 2 \\ 2, & \text{if } y \geq 2 \end{cases}.$$

This requires to divide the data  $\mathcal{X}$  into three partitions depending on  $y$ , i.e.  $\mathcal{X}^1$  contains samples where  $y \leq 1$ ,  $\mathcal{X}^2$  consists of samples with a margin in the range  $1 < y < 2$ , and  $\mathcal{X}^3 = \mathcal{X} \setminus \{\mathcal{X}^1 \cup \mathcal{X}^2\}$ . Now, our objective function for maximization is

$$C(\mathcal{X}|\Theta) = \prod_{n=1}^N h((d_{\Theta}^n)^{\lambda}) = \prod_{n_1 \in \mathcal{X}^1} \left[ (d_{\Theta}^{n_1})^{\lambda} + \frac{1}{2} \right] \prod_{n_2 \in \mathcal{X}^2} \left[ 2 - \frac{1}{2} \left( (d_{\Theta}^{n_2})^{\lambda} - 2 \right)^2 \right] 2^{|\mathcal{X}^3|}.$$

## 2.1 The Extended Baum-Welch algorithm

Let  $\phi = \{\phi_i^j\}$  denote a set of discrete probability distributions, with  $\phi_i^j > 0$  and  $\sum_i \phi_i^j = 1$ . Given a rational function  $R$  over  $\phi$ , the extended Baum-Welch (EBW) algorithm (Gopalakrishnan et al. (1991)) is an iterative procedure that attains a local maximum of  $R$ , and is given according to

$$\phi_i^j \leftarrow \frac{\phi_i^j \left( \frac{\partial \log R}{\partial \phi_i^j} + D \right)}{\sum_l \phi_l^j \left( \frac{\partial \log R}{\partial \phi_l^j} + D \right)}, \quad (2)$$

where  $D$  is a constant. Note that  $C(\mathcal{X}|\Theta)$  is a rational function over parameters  $\alpha_c^m$  and  $\rho_c$ . We will employ a discrete approximation to the Gaussian distribution, such that we can apply EBW to update as well the continuous parameters  $\mu_c^m$  and  $\Sigma_c^m$ .

The required partial derivative with respect to  $\Theta$  is

$$\frac{\partial \log C(\mathcal{X}|\Theta)}{\partial \Theta} = \sum_{n=1}^N s^n \frac{\partial \log d_{\Theta}^n}{\partial \Theta}$$

where  $s^n$  denotes a sample dependent weight given as follows:

$$s^n = \begin{cases} \frac{\lambda (d_{\Theta}^n)^{\lambda}}{(d_{\Theta}^n)^{\lambda} + \frac{1}{2}}, & \text{if } n \in \mathcal{X}^1 \\ 2\lambda, & \text{if } n \in \mathcal{X}^2 \\ 0, & \text{if } n \in \mathcal{X}^3 \end{cases}.$$

Further, the partial derivative of  $\log d_{\Theta}^n$  with respect to  $\rho_c$  and  $\alpha_c^m$  is

$$\frac{\partial \log d_{\Theta}^n}{\partial \rho_c} = \frac{1}{\rho_c} (\mathbf{1}_{\{c=c^n\}} - \mathbf{1}_{\{c \neq c^n\}} r_c^n), \quad (3)$$

$$\frac{\partial \log d_{\Theta}^n}{\partial \alpha_c^m} = \frac{\gamma_c^{n,m}}{\alpha_c^m} (\mathbf{1}_{\{c=c^n\}} - \mathbf{1}_{\{c \neq c^n\}} r_c^n), \quad (4)$$

where  $\mathbf{1}_{\{i=j\}}$  is the indicator function (i.e. equals 1 if  $i = j$  and 0 if  $i \neq j$ ), and we introduced

$$r_c^n = \frac{[p(\mathbf{x}^n | \Theta_c) \rho_c]^\eta}{\left[ \sum_{c' \neq c^n} p(\mathbf{x}^n | \Theta_{c'}) \rho_{c'} \right]^\eta},$$

$$\gamma_c^{n,m} = \frac{\alpha_c^m \mathcal{N}(\mathbf{x}^n | \theta_c^m)}{\sum_{m'=1}^M \alpha_c^{m'} \mathcal{N}(\mathbf{x}^n | \theta_c^{m'})}.$$

The derivatives in Eq. (3) and (4) are sensitive to small parameter values. Therefore, we approximate these derivatives using the approximation suggested by Meriäldo (1988):  $\frac{\partial \log d_{\Theta}^n}{\partial x_c} = \frac{1}{x_c} (a_c^n - b_c^n) \approx \frac{a_c^n}{\sum_{c'} a_{c'}^n} - \frac{b_c^n}{\sum_{c'} b_{c'}^n}$ .

In order to optimize the continuous parameters  $\mu_c^m$  and  $\Sigma_c^m$  using EBW, we use the discrete approximation proposed in Normandin and Morgera (1991) assuming diagonal covariance matrices. This leads to the re-estimation equation for  $\bar{\mu}_c^m$  and  $\bar{\Sigma}_c^m$  given as

$$\bar{\mu}_c^m \leftarrow \frac{\sum_{n=1}^N [s^n \gamma_c^{n,m} (\mathbf{1}_{\{c^n=c\}} - \mathbf{1}_{\{c \neq c^n\}} r_c^n) \mathbf{x}^n] + D_1 \mu_c^m}{\sum_{n=1}^N [s^n \gamma_c^{n,m} (\mathbf{1}_{\{c^n=c\}} - \mathbf{1}_{\{c \neq c^n\}} r_c^n)] + D_1} \quad (5)$$

$$\bar{\Sigma}_c^m \leftarrow \frac{\sum_{n=1}^N [s^n \gamma_c^{n,m} (\mathbf{1}_{\{c^n=c\}} - \mathbf{1}_{\{c \neq c^n\}} r_c^n) (\mathbf{x}^n)^2] + D_1 (\Sigma_c^m + (\mu_c^m)^2)}{\sum_{n=1}^N [s^n \gamma_c^{n,m} (\mathbf{1}_{\{c^n=c\}} - \mathbf{1}_{\{c \neq c^n\}} r_c^n)] + D_1} - (\bar{\mu}_c^m)^2,$$

where  $(\mathbf{x})^2$  denotes the element-wise square operation on vector  $\mathbf{x}$ .

The EBW algorithm converges to a local optimum of  $C(\mathcal{X} | \Theta)$  providing a sufficiently large value for  $D$ . Indeed, setting the constant  $D$  is not trivial. If it is chosen too large then training is slow and if it is too small the update may fail to increase the objective function. In practical implementations heuristics have been suggested (Woodland and Povey (2002), Klautau et al. (2003)).

Setting  $D = 1 + \left| \min_{c,m} \left[ \frac{\partial C(\mathcal{X} | \Theta)}{\partial \alpha_c^m} \alpha_c^m, \frac{\partial C(\mathcal{X} | \Theta)}{\partial \rho_c} \rho_c \right] \right|$  shows good performance in our experiments. Furthermore, the value for  $D_1$  is initialized to  $D$ . Then we double  $D_1$  until all variances in  $\bar{\Sigma}_c^m$  are positive in the re-estimation step. The values  $D$  and  $D_1$  are adapted in each iteration of the algorithm.

### 3 Multipitch Tracking

We apply discriminative GMMs to the task of multipitch tracking. The term pitch refers to the fundamental frequency of voiced utterances in speech, and is an important feature in speech signal analysis. The term multipitch tracking refers to the task of estimating the pitch of multiple speakers speaking



simultaneously, where only a single-channel recording is available ('mixture signal'). For a more detailed description on the multipitch tracking method used herein, we refer the interested reader to Wohlmayr and Pernkopf (2010). Tracking is based on factorial hidden Markov models (FHMMs) (Ghahramani and Jordan (1997)), see Figure 1(a). The hidden state random variables are denoted by  $c_k^{(t)}$ , where  $k$  indicates the Markov chain and  $t$  the time index from 1 to  $T$ . Similarly, realizations of the observed random variables at  $t$  are collected in a vector  $\mathbf{y}^{(t)} \in \mathbb{R}^D$ . Each  $c_k^{(t)}$  represents a discrete random variable with cardinality  $|C|$ . The dependency of hidden variables between two consecutive time instances is represented for each Markov chain by the transition probability  $p(c_k^{(t)} | c_k^{(t-1)})$ . The dependency of the observed variables  $\mathbf{y}^{(t)}$  on the hidden variables of the same time frame are specified by the observation probability  $p(\mathbf{y}^{(t)} | c_1^{(t)}, c_2^{(t)})$ . Finally, the prior distribution of the hidden variables in every chain is denoted by  $p(c_k^{(1)})$ . Denoting the whole sequence of variables, i.e.  $\{c^{(t)}\} = \bigcup_{t=1}^T \{c_1^{(t)}, c_2^{(t)}\}$  and  $\{\mathbf{y}^{(t)}\} = \bigcup_{t=1}^T \mathbf{y}^{(t)}$ , the joint distribution of all variables is given by

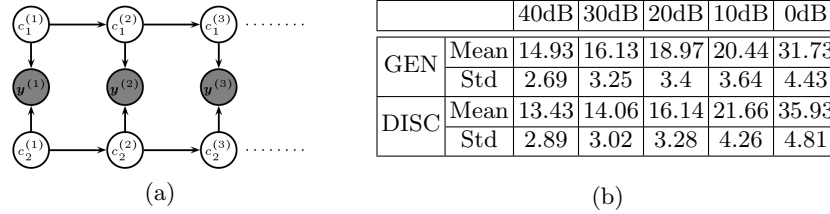
$$p(\{c^{(t)}\}, \{\mathbf{y}^{(t)}\}) = \prod_{k=1}^2 \left[ p(c_k^{(1)}) \prod_{t=2}^T p(c_k^{(t)} | c_k^{(t-1)}) \right] \prod_{t=1}^T p(\mathbf{y}^{(t)} | c_1^{(t)}, c_2^{(t)}).$$

In this work, we perform multipitch tracking of two simultaneously speaking subjects, i.e. the FHMM has two Markov chains. Each Markov chain models the pitch trajectory of one speaker, hence the hidden variable  $c_k^{(t)}$  denotes the pitch state (i.e. class) of speaker  $k$  at time  $t$ . Each hidden variable has  $|C| = 170$  states, where state value '1' refers to 'no pitch' (i.e. unvoiced or silent), and state values '2'-'170' encode different pitch frequencies ranging from 80 to 500Hz. Specifically, the pitch value  $f_0$  corresponding to state  $c > 1$  is obtained as  $f_0 = \frac{f_s}{30+c}$ , where the sampling rate  $f_s = 16\text{kHz}$ . This results in a nonuniform quantization of the pitch interval (Wu et al. (2003)). The feature vector  $\mathbf{y}^{(t)}$  contains the log-spectrum of the mixture signal observed at time  $t$ . This allows the usage of the MIXMAX interaction model, as outlined in the next subsection. For the sake of brevity, we omit from now on the explicit dependence of random variables on  $t$ , where appropriate.

### 3.1 MIXMAX Interaction Model

Given a set of  $N_i$  single speaker log-spectra for speaker  $i$ ,  $\mathcal{X}_i = \{\mathbf{x}_i^{(1)}, \dots, \mathbf{x}_i^{(N_i)}\}$ , together with corresponding reference pitch labels,  $\{c_i^{(1)}, \dots, c_i^{(N_i)}\}$  we can easily learn a speaker dependent classifier  $p(c_i | \mathbf{x}_i)$  using the proposed training algorithm for discriminative GMMs, where  $\mathbf{x}_i$  is the log-spectrum of the *single* speaker  $i$ .<sup>1</sup> To track a mixture of two specific speakers, we then combine both speaker-specific classifiers via the MIXMAX interaction model to obtain the joint posterior  $p(c_1, c_2 | \mathbf{y})$ . The MIXMAX interaction model is

<sup>1</sup> Note that in this case we do not have to learn the class priors  $\rho_c$  since they are not used in the interaction model.



**Fig. 1.** (a) Factorial hidden Markov model (FHMM) for multipitch tracking. Shaded nodes indicate observed random variables. (b) Error measure  $E_{Total}$  (in percent) of multipitch tracking algorithm using generative (GEN) or discriminative (DISC) GMMs in conjunction with an FHMM on Mocha-TIMIT database. For each noise condition, tracking was evaluated on 60 test utterances.

guided by the insight that the log-spectrogram of two speakers can be approximated by their elementwise maximum (Nadas et al. (1989)). Specifically,  $\mathbf{y} \approx \max(\mathbf{x}_1, \mathbf{x}_2)$ , where  $\mathbf{x}_i$  is the log-spectrum of speaker  $i$ . Thus,  $p(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2) = \delta(\mathbf{y} - \max(\mathbf{x}_1, \mathbf{x}_2))$ , where  $\delta(\cdot)$  denotes the Dirac delta. In general, we obtain the joint posterior by marginalizing over the unknown single speaker log-spectra:

$$p(c_1, c_2|\mathbf{y}) = \frac{p(c_1)p(c_2)}{Z} \int \int p(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2)p(\mathbf{x}_1|c_1)p(\mathbf{x}_2|c_2)d\mathbf{x}_1d\mathbf{x}_2. \quad (6)$$

where we obtain  $p(\mathbf{x}_i|c_i)$  from the trained classifier  $p(c_i|\mathbf{x}_i)$  using Bayes rule, and  $Z$  is a normalization constant. Note that for tracking, we use  $p(\mathbf{y}|c_1, c_2) = p(c_1, c_2|\mathbf{y})Z / (p(c_1)p(c_2))$ . We denote the set of GMM parameters trained for speaker  $i$  as  $\{\boldsymbol{\Theta}_{i,c}\}_{c=1}^{|C|}$ . It can be shown that

$$p(\mathbf{y}|c_1, c_2) = \sum_{m=1}^M \sum_{n=1}^M \alpha_{1,c_1}^m \alpha_{2,c_2}^n \prod_{d=1}^D \left[ \mathcal{N}(y_d|\theta_{1,c_1}^{m,d}) \Phi(y_d|\theta_{2,c_2}^{n,d}) + \Phi(y_d|\theta_{1,c_1}^{m,d}) \mathcal{N}(y_d|\theta_{2,c_2}^{n,d}) \right],$$

where  $y_d$  is the  $d^{\text{th}}$  element of  $\mathbf{y}$ ,  $\theta_{i,c_i}^{m,d}$  is the  $d^{\text{th}}$  element of the corresponding mean and variance, and  $\Phi(y|\theta) = \int_{-\infty}^y \mathcal{N}(x|\theta)dx$  denotes the univariate cumulative normal distribution. For a more detailed discussion, we refer the interested reader to Wohlmayr and Pernkopf (2010).

### 3.2 Tracking

Given the set of observations  $\{\mathbf{y}^{(t)}\}$ , the task of tracking involves searching the sequence of hidden states  $\{c^{(t)}\}^*$  that maximizes the conditional distribution  $p(\{c^{(t)}\}|\{\mathbf{y}^{(t)}\})$ . We use a variant of the junction tree algorithm (Ghahramani and Jordan (1997)) to obtain an exact solution for tracking.

## 4 Experimental Results

We compare the performance of discriminative and generative GMMs used in conjunction with FHMMs for multipitch tracking. Tests are performed using

the Mocha-TIMIT database (Wrench (2000)), which consists of 460 English utterances from both a male and a female speaker, sampled at 16kHz. In addition, laryngograph signals are available for all recordings, from which the reference pitch  $f_0[t]$  was acquired using the RAPT method (Talkin (1995)) together with manual removal of erroneous pitch estimates in nonaudible regions. The speaker dependent classifiers were trained on 400 sentences each, while 60 test instances were obtained by mixing the remaining male and female utterances at 0dB. Each test utterance was mixed with white Gaussian noise at SNR conditions ranging from 40dB down to 0dB in 10dB steps.

The input features  $\mathbf{y}^{(t)}$  are based on the log-spectrogram of the speech mixture. Given an input signal at sampling rate  $f_s = 16\text{kHz}$ , we compute the spectrogram via the 1024 point FFT, using a Hamming window of length 32ms and step size of 10ms. Next, we obtain each observation vector  $\mathbf{y}^{(t)} \in \mathbb{R}^{64}$  by taking the log of spectral bins 2-65, which corresponds to a frequency range up to 1000Hz. This covers the most relevant frequency range, while keeping the model complexity low.

For discriminative training of the single speaker GMMs, parameters were initialized to the ML solution. Parameters  $\lambda$  and  $\eta$  were set to 0.1 and 5, respectively. Both transition matrices of the FHMM,  $p(c_k^{(t)}|c_k^{(t-1)})$ , are obtained by counting and normalizing the transitions of the reference pitch values from single speaker recordings in the training set. Additionally, we apply Laplace smoothing on both transition matrices. Prior distributions  $p(c_k^{(1)})$  are obtained likewise.

To evaluate the results, we used the error measure  $E_{Total}$  as used in Wohlmayr and Pernkopf (2010). The performance of both methods on Mocha-TIMIT is shown in Figure 1(b) for various noise conditions, where the parameters of the MIXMAX-FHMM remained optimized for clean speech. The FHMM with discriminatively optimized GMMs slightly outperforms the purely generative model.

## 5 Conclusions

We have proposed an algorithm for maximum margin training of GMM based classifiers. The resulting classifiers remain within the probabilistic framework, which facilitates their application in the probabilistic interaction models. For the task of multipitch tracking, we combined single-speaker classifiers into one double-speaker classifier, and have shown an increased performance over the generative model under low-noise conditions.

## References

- CRAMMER, K. and SINGER, Y. (2001): On the algorithmic interpretation of multiclass kernel-based vector machines, *Journal of Machine Learning Research* 2, 265-292.

- GHAHRAMANI, Z. and JORDAN, M. I. (1997): Factorial Hidden Markov Models, *Machine Learning* 29(2-3), 245-273.
- GOPALAKRISHNAN, O., KANEVSKY, D. and NADAS, A. and NAHAMOO, D. (1991): An inequality for rational functions with applications to some statistical estimation problems. *IEEE Transactions on Information Theory* 37(1), 107-113.
- GUO, Y., WILKINSON, D. and SCHUURMANS, D. (2005): Maximum margin Bayesian networks, *International Conference on Uncertainty in Artificial Intelligence (UAI)*.
- KLAUTAU, A., JEV TIC, N. and ORLITSKY, A. (2003): Discriminative Gaussian mixture models: A comparison with kernel classifiers, *International Conference on Machine Learning (ICML)*, 353-360.
- MERIALDO, B. (1988): Phonetic recognition using hidden Markov models and maximum mutual information training. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 111-114.
- NADAS, A., NAHAMOO, D. and PICHENY, M. A. (1989): Speech recognition using noise-adaptive prototypes. *IEEE Transactions on Acoustics, Speech and Signal Processing* 37(10), 1495-1503.
- NORMANDIN, Y. and MORGERA, S. D. (1991): An improved MMIE training algorithm for speaker-independent small vocabulary, continuous speech recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 537-540.
- SCHÖLKOPF, M. and SMOLA, A. J. (2001): *Learning with Kernels: Support Vector Machines, regularization, optimization, and beyond*. MIT Press..
- TALKIN, D. (1995): A robust algorithm for pitch tracking (RAPT). In: W.B. Kleijn and K.K. Paliwal (Eds.): *Speech Coding and Synthesis*. Elsevier Science, 495-518.
- TASKAR, B., GUESTRIN, C. and KOLLER, D. (2003): Max-margin Markov networks, *Advances in Neural Information Processing Systems (NIPS)*.
- SHA, F. and SAUL, L. (2006): Large margin Gaussian mixture modeling for phonetic classification and recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.
- VAPNIK, V. (1998): *Statistical learning theory*, Wiley & Sons.
- WOHLMAYR, M. and PERNKOPF, F. (2010): A Mixture Maximization Approach to Multipitch Tracking With Factorial Hidden Markov Models, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.
- WOODLAND, P. and Povey, D. (2002): Large scale discriminative training of hidden Markov models, *Computer Speech and Language* 16, 25-47.
- WRENCH, A. (2000): A multichannel/multispeaker articulatory database for continuous speech recognition research. *Phonus* 5, 3-17.
- WU, M., WANG, D. and BROWN, G. J. (2003): A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing* 11(3), 229-241.

# Low-Pass Filter Design and Discrete Prolate Spheroidal Sequences

Tommaso Proietti<sup>1</sup> and Alessandra Luati<sup>2</sup>

<sup>1</sup> University of Rome “Tor Vergata”, S.E.F. e ME. Q.,  
via Columbia 2, 00133 Roma, Italy, *tommaso.proietti@uniroma2.it*

<sup>2</sup> University of Bologna, Department of Statistics,  
via Belle Arti 41, 40126 Bologna, Italy, *alessandra.luati@unibo.it*

**Abstract.** The paper concerns the design of nonparametric low-pass filters that have the property of reproducing a polynomial of a given degree. Two approaches are considered. The first is locally weighted polynomial regression (LWPR), which leads to linear filters depending on three parameters: the bandwidth, the order of the fitting polynomial, and the kernel. We find a remarkable linear (hyperbolic) relationship between the cutoff period (frequency) and the bandwidth, conditional on the choices of the order and the kernel.

The second hinges on a generalization of the maximum concentration approach, leading to filters related to discrete prolate spheroidal sequences (DPSS). In particular, we propose a new class of low-pass filters that maximize the concentration over a specified frequency range, subject to polynomial reproducing constraints.

**Keywords:** low-pass filters, kernels, concentration, filter design

## 1 Introduction

Trend filters that arise from fitting a locally weighted polynomial to a time series have a well established tradition in time series analysis and signal extraction, see, e.g., Loader (1999). Locally weighted polynomial regression generates finite impulse response filters (LWPR filters, henceforth), that are widely applied in practice. LWPR are trend extraction filters, and as such they somehow pass low frequency components and reduce the amplitude of high frequency ones.

The first aim of the paper is to enforce the interpretation of LWPR filters as low-pass filters and to illustrate how the cutoff period (or frequency) is related to the three key ingredients: bandwidth, polynomial order and kernel. We define the cutoff period  $P_c$  as the smallest periodicity of the fluctuations that are passed i.e. that are preserved to a great extent, by the filter, whereas the fluctuations with smaller period are compressed. Conversely, the cutoff frequency in radians  $\omega_c \in (0, \pi)$  is the maximum frequency in radians that is preserved by the filter, whereas higher frequencies are suppressed. The two quantities are inversely related, as  $P_c = 2\pi/\omega_c$ . An ideal low-pass filter has unit transfer function (TF) in the frequency range  $(0, \omega_c)$ , and zero TF

elsewhere (see Percival and Walden, 1993). For any LWPR filter, we derive the underlying cutoff frequency by least squares filter design principles, as the frequency at which a given LWPR filter provides the best approximation to an ideal low pass filter. Interestingly, we find out that given the kernel and polynomial order, the cutoff period is linearly related to the bandwidth; the intercept and slope of the linear relationship depend on the the order of the polynomial and the kernel.

The second contribution of the paper is to propose a generalization of a well-known class of filters that are designed by using the discrete prolate spheroidal sequences (DPSS). The DPSS approach to filter design is based on the maximization of the concentration of a filter at a reference frequency, denoted  $\varpi$  (see Slepian, 1978; see also Lii and Rosenblatt, 2008, for recent applications). A DPSS filter reproduces a constant or linear function of time. The filter that solves the traditional concentration problem is given by a zeroth order discrete prolate spheroidal sequence. We generalize the DPSS approach by imposing the constraint that the filter reproduces a polynomial trend of any degree; we thus obtain a class of DPSS filters that depend on three parameters, the bandwidth, the order of the polynomial and the concentration frequency,  $\varpi$ . We then discuss the interpretation of the latter. Our interpretation differs from the mainstream one: in fact, we show that the choice of the concentration frequency does not really differ from the choice of the kernel.

## 2 LWPR as low-pass filters

Let  $y_t$  denote a time series measured at discrete and equally spaced time points. The series can be decomposed as  $y_t = \mu_t + \varepsilon_t$ , where  $\mu_t$  is the underlying trend, and  $\varepsilon_t$  is the noise. We further assume that  $\mu_t$  is a smooth but unknown deterministic function of time, which can be approximated in a neighborhood of time  $t$  by a polynomial of degree  $p$  of the time distance,  $j$ , between  $y_t$  and the neighboring observations  $y_{t+j}$ ,  $j = 0, \pm 1, 2, \dots, h$ . This enables to write  $\mu_{t+j} \approx m_{t+j}$ , with  $m_{t+j} = \beta_0 + \beta_1 j + \dots + \beta_p j^p$ ,  $j = 0, \pm 1, \dots, \pm h$ . We assume throughout that we are interested in the estimation of the trend at an interior point, and we will not be concerned with the treatment of end points. Provided that  $p \leq 2h$ , the  $p+1$  unknown coefficients  $\beta_k$ ,  $k = 0, \dots, p$ , can be estimated by the method of weighted least squares (WLS), which consists of minimising with respect to the  $\beta_k$ 's the objective function:  $S(\hat{\beta}_0, \dots, \hat{\beta}_p) = \sum_{j=-h}^h \kappa_j \left( y_{t+j} - \hat{\beta}_0 - \hat{\beta}_1 j - \dots - \hat{\beta}_p j^p \right)^2$ . In matrix notation, the local polynomial approximation can be written as  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ,  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ , where  $\mathbf{y} = [y_{t-h}, \dots, y_t, \dots, y_{t+h}]'$ ,  $\boldsymbol{\varepsilon} = [\varepsilon_{t-h}, \dots, \varepsilon_t, \dots, \varepsilon_{t+h}]'$ , the  $k+1$ -th column of  $\mathbf{X}$  is the  $[(-h)^k, \dots, (h-1)^k, (h)^k]'$ ,  $k = 0, \dots, p$ , and  $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_p]'$ . Defining  $\mathbf{K} = \text{diag}(\kappa_h, \dots, \kappa_1, \kappa_0, \kappa_1, \dots, \kappa_h)$ , the WLS estimate of the coefficients is  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}\mathbf{y}$ . In order to obtain

$\hat{m}_t = \hat{\beta}_0$ , we need to select the first element of the vector  $\hat{\beta}$ . Hence, denoting by  $\mathbf{e}_1$  the  $p+1$  vector  $\mathbf{e}_1' = [1, 0, \dots, 0]$ ,  $\hat{m}_t = \mathbf{e}_1' \hat{\beta} = \mathbf{e}_1' (\mathbf{X}' \mathbf{K} \mathbf{X})^{-1} \mathbf{X}' \mathbf{K} \mathbf{y} = \mathbf{w}' \mathbf{y} = \sum_{j=-h}^h w_j y_{t-j}$ , which expresses the estimate of the trend as a linear combination of the observations with coefficients

$$\mathbf{w}' = \mathbf{e}_1' (\mathbf{X}' \mathbf{K} \mathbf{X})^{-1} \mathbf{X}' \mathbf{K}. \quad (1)$$

The linear combination yielding our trend estimate is often termed a (linear) *filter*, and the weights  $w_j$  constitute its impulse responses. The latter are time invariant and carry essential information on the nature of the estimated signal; they enjoy two important properties, symmetry,  $w_{-j} = w_j$ , and reproduction of  $p$ -th degree polynomials. As far as the second is concerned, from (1) we have that  $\mathbf{X}' \mathbf{w} = \mathbf{e}_1$ . As a consequence, the filter  $\mathbf{w}$  is said to preserve a deterministic polynomial of order  $p$ , which means that if the series is  $\mathbf{y} = \mathbf{X} \beta$  then the filter will reproduce it exactly, i.e.  $\hat{m}_t = \beta_0 = y_t$ .

The LWPR filter are low-pass filters, i.e., they leave almost unchanged low frequency components, such as the trend, and attenuate high frequency fluctuations associated with the noise. This is evidenced by the transfer function of the filter

$$G_{h,p,\kappa}(\omega) = \sum_{j=-h}^h w_j e^{-i\omega j}$$

where  $i$  is the imaginary unit and  $\omega \in (-\pi, \pi)$  is the frequency measured in radians. Due to the symmetry of the weights  $w_j$ ,  $G_{h,p,\kappa}(\omega)$  is real;  $|G_{h,p,\kappa}(\omega)|$  is the gain of the filter.

It should be recalled that an ideal low-pass filter retains only the low frequency fluctuations in the series and reduces the amplitude of fluctuations with frequencies higher than a cutoff frequency  $\omega_c$ . Its transfer function takes the following form: for  $\omega \in (0, \pi)$ ,

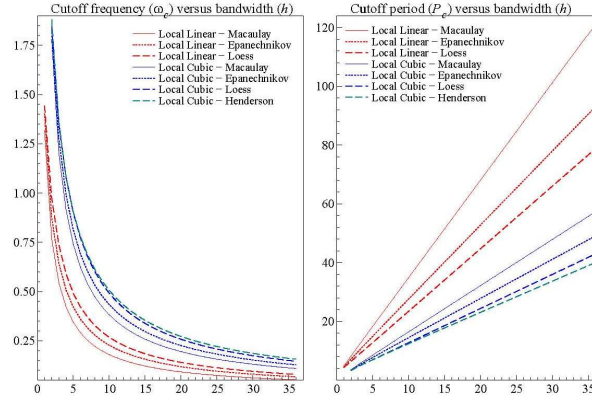
$$G_{lp}(\omega) = \begin{cases} 1 & \text{if } \omega \leq \omega_c \\ 0 & \text{otherwise.} \end{cases}$$

As it is well known the ideal filter is available analytically, but unfeasible. Taking the inverse Fourier transform, yields the weights of the ideal filter:

$$w_j = \begin{cases} \frac{\sin(\omega_c j)}{\pi j}, & j = \pm 1, 2, \dots, \\ \frac{\omega_c}{\pi} & j = 0. \end{cases}$$

The filter can be approximated by truncating the weights at lead and lag  $h$ , and rescaling them so that their sum is one.

We now derive the cutoff frequency associated with any LWPR filter by a least squares filter design approach. The latter is usually applied for determining the filter which minimizes the squared modulus of the discrepancy between its transfer function,  $G_{h,p,\kappa}(\omega)$  and that of the ideal low pass filter,  $G_{lp}(\omega)$ , for a given cutoff frequency  $\omega_c$ . See Percival and Walden (1993, sec.



**Fig. 1.** Relationship between bandwidth (horizontal axis) and cutoff frequency  $\omega_c$  (left panel), and cutoff period,  $P_c = 2\pi/\omega_c$ , (right panel) for different LWPR filters.

5.8). In our problem, instead, the filter weights are given and the cutoff is unknown. The latter will be determined as the frequency

$$\omega_c = \operatorname{argmin} \{D(\omega_c; h, p, \kappa)\}, \quad D(\omega_c; h, p, \kappa) = \int_0^\pi |G_{h,p,\kappa}(\omega) - G_{lp}(\omega)|^2 d\omega.$$

for given  $h, p$ , and kernel  $\kappa$ . The solution is straightforward: rewriting

$$D(\omega_c; h, p, \kappa) = \int_0^{\omega_c} |G_{h,p,\kappa}(\omega) - 1|^2 d\omega + \int_{\omega_c}^\pi |G_{h,p,\kappa}(\omega)|^2 d\omega,$$

and differentiating with respect to  $\omega_c$ , the first order condition for the above problem is

$$\sum_{j=-h}^h w_j e^{-i\omega_c j} = \frac{1}{2}, \quad (2)$$

i.e. we need to locate the frequency at which the transfer function is  $1/2$ . The solution to equation (2) is obtained as the phase of the conjugate pair of roots with unit modulus of the polynomial  $\sum_j w_j x^j$ .

Figure 1 displays the result for the LWPR filters most commonly in use. The left panel shows the cutoff frequency (vertical axis) corresponding to a particular bandwidth  $h$ : as  $h$  increases, for given  $p$  and kernel function, the cutoff frequency is pushed towards zero, i.e. we obtain a filter that produces a smoother trend. It is interesting to present the same results in terms of the cutoff period,  $P_c = 2\pi/\omega_c$ , since this enables to discover a remarkable linear relationship between cutoff period and bandwidth. For instance, for the Henderson filter,  $P_c = 1.4 + 1.1h$ , so that when  $h = 12$ ,  $P_c = 14.6$ . It should be noticed that the intercept and slope depend on the order of the polynomial and the kernel. If we move from a linear to a cubic fit, *ceteris paribus*,



i.e. using the same  $h$  and kernel, the cutoff period decreases. Secondly, higher order kernels yield lower cutoff periods for fixed  $h$  and  $p$ . Figure 1 provides an effective tool for designing the LWPR filter that is appropriate for a particular cutoff frequency or period: drawing an horizontal straight line at a particular value and looking at the intersection with the hyperboles (left panel) or lines (right panel) would make immediately available the bandwidth, order and kernel of the relevant LWPR filter.

### 3 Generalized local polynomial DPSS filter design

We turn our attention to a different approach to the design of a low-pass filter which can be adapted to the problem of fitting a local polynomial. This adaptation configures the main contribution of the paper to this strand of literature.

Let  $\mathbf{w}$  denote a filter characterized by the transfer function  $G(\omega)$ ,  $\omega \in [0, \pi]$ ; then its concentration at frequency  $\varpi \in (0, \pi)$  is defined as

$$\beta^2(\varpi) = \frac{\int_0^\varpi |G(\omega)|^2 d\omega}{\int_0^\pi |G(\omega)|^2 d\omega}. \quad (3)$$

The concentration can be expressed as a ratio of quadratic forms,

$$\beta^2(\varpi) = \frac{\mathbf{w}' \mathbf{A}(\varpi) \mathbf{w}}{\mathbf{w}' \mathbf{w}}, \quad (4)$$

where  $\mathbf{A}(\varpi)$  is the symmetric and positive definite matrix whose generic  $ij$ -th element,  $i, j = 1, \dots, 2h+1$ , is

$$a_{ij}(\varpi) = \begin{cases} \frac{\sin(\varpi(i-j))}{\pi(i-j)} & \text{for } i \neq j \\ \varpi/\pi & \text{for } i = j \end{cases}$$

as follows by replacing  $|G(\omega)|^2 = G(\omega)G^*(\omega) = \sum_{i=-h}^h w_i^2 + 2 \sum_{i=-h}^h \sum_{j=-h}^{i-1} w_i w_j \cos((i-j)\omega)$  in (3) and then performing the integrations;  $G^*(\omega)$  denotes the complex conjugate of  $G(\omega)$ .

A strategy for designing filters is, for a given bandwidth  $h$ , to choose the filter that maximizes the concentration at a given concentration frequency, under the constraint that the weights sum to one. Under this condition the filter is capable of reproducing a polynomial of order 0, 1. Hence, given  $\varpi$ , the filter weights result from the solution of the following constrained maximization problem:

$$\max_{\mathbf{w}_j, j=-h, \dots, h} \beta^2(\varpi) \text{ subject to } \mathbf{w}' \mathbf{i} = 1,$$

where  $\mathbf{i} = [1, \dots, 1]'$ . The quantity (4) is a Rayleigh quotient that reaches its maximum when it is equal to the largest eigenvalue of  $\mathbf{A}(\varpi)$ , let us call

it  $\lambda_0(\varpi)$ . The corresponding eigenvector, denoted  $\mathbf{v}_0$ , is the filter that maximizes the concentration (3) or (4) at the concentration frequency  $\varpi$ , when no restrictions are imposed to the weights except, possibly, to be of unitary norm. The filter weights arise from the normalization:  $\mathbf{w} = \mathbf{v}_0/(\mathbf{v}_0'\mathbf{i})$ . The (unit norm) vector  $\mathbf{v}_0$  is a subsequence of length  $2h+1$  of a zeroth-order discrete prolate spheroidal sequence (DPSS). Its direct computation from the spectral decomposition  $\mathbf{A}(\varpi)$  is computationally unattractive and unstable, but it has been shown that it can be derived as the eigenvector associated with the largest eigenvalue of a symmetric tridiagonal matrix, see Slepian (1978).

The DPSS approach described above generates a filter that passes a local constant/linear polynomial,  $m_{t+j}$ , for  $p = 0, 1$ ; this is a simple consequence of the constraint  $\mathbf{w}'\mathbf{i} = 1$ . It seems therefore natural to propose a generalization by introducing further linear constraints on the weights that aim at the reproduction of higher order polynomials. We show below that there is a solution to this problem. The price to be paid is that the solution is no closed form solution to this problem and the solution has to be found iteratively.

For any bandwidth  $h$ , we aim at designing the filter that maximizes the concentration at a given frequency, under the constraint of reproducing a polynomial of degree  $p$ . This produces the following constrained maximization problem:

$$\max_{\mathbf{w}_j, j=-h, \dots, h} \beta^2(\varpi) \text{ subject to } \mathbf{w}'\mathbf{X} = \mathbf{e}'_1, \quad (5)$$

Obviously, for  $\mathbf{X} = \mathbf{i}$ , we are back to the standard problem discussed above.

The objective function to be maximized is

$$\phi(\mathbf{w}, \mu) = \frac{\mathbf{w}'\mathbf{A}(\varpi)\mathbf{w}}{\mathbf{w}'\mathbf{w}} - 2(\mathbf{w}'\mathbf{X} - \mathbf{e}'_1)\mu,$$

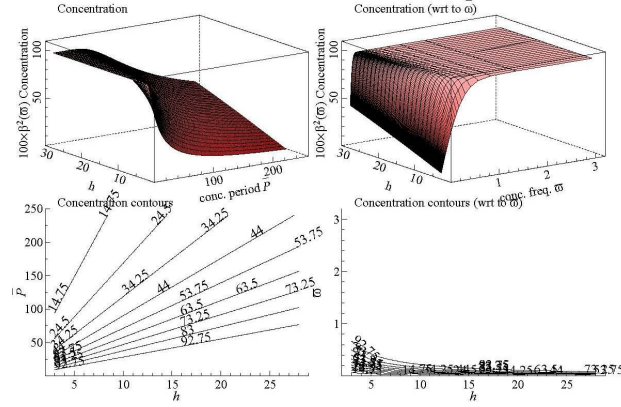
where  $\mathbf{e}'_1 = [1, 0, \dots, 0]$  and  $\mu$  is a vector of Lagrange multipliers. The first order conditions give  $\mathbf{B}(\varpi)\mathbf{w}(\mathbf{w}'\mathbf{w})^{-1} = \mathbf{X}\mu$  and  $\mathbf{w}'\mathbf{X} = \mathbf{e}'_1$ , where  $\mathbf{B}(\varpi) = \mathbf{A}(\varpi) - \beta^2(\varpi)\mathbf{I}$ . If  $\mathbf{w}$  is an eigenvector of  $\mathbf{A}(\varpi)$ , associated with the eigenvalue  $\beta^2(\varpi)$ , then  $\mathbf{B}(\varpi)$  is singular and the first order conditions imply that  $\mu$  is the null vector, i.e. the problem collapses to the unconstrained maximization discussed above. On the other hand, if  $\mathbf{w}$  is not an eigenvector of  $\mathbf{A}(\varpi)$ , then  $\mathbf{B}(\varpi)$  is nonsingular and its inverse  $\mathbf{B}(\varpi)^{-1}$  exists. Rearranging,  $\mathbf{w}(\mathbf{w}'\mathbf{w})^{-1} = \mathbf{B}(\varpi)^{-1}\mathbf{X}\mu$  and  $\mu'\mathbf{X}'\mathbf{B}(\varpi)^{-1}\mathbf{X} = \mathbf{e}'_1(\mathbf{w}'\mathbf{w})^{-1}$ , the solution of (5) is the filter

$$\mathbf{w}' = \mathbf{e}'_1(\mathbf{X}'\mathbf{B}(\varpi)^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}(\varpi)^{-1}.$$

It is evident that this is an implicit solution, as  $\mathbf{B}(\varpi)$  depends on  $\mathbf{w}$  through  $\beta^2(\varpi)$ .

Hence, the solution can be obtained by the following iterative algorithm.

- i. Start from a symmetric LWPR filter  $\mathbf{w}_1$  that satisfies the polynomial reproducing constraints.



**Fig. 2.** Concentration for standard DPSS filters, as a function of the concentration period and frequency.

- ii. For  $r = 2, 3, \dots$ ,
  - a. compute the concentration:

$$\beta_r^2(\varpi) = \frac{\mathbf{w}_r' \mathbf{A}(\varpi) \mathbf{w}_1}{\mathbf{w}_r' \mathbf{w}_r};$$

- b. construct the matrix  $\mathbf{B}_r(\varpi) = \mathbf{A}(\varpi) - \beta_r(\varpi)^2 \mathbf{I}$ ;
  - c. update the solution

$$\mathbf{w}_r' = \mathbf{e}_1' (\mathbf{X}' \mathbf{B}_r(\varpi)^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{B}_r(\varpi)^{-1}. \quad (6)$$

- iii. Repeat until convergence, i.e. for some  $r$  the discrepancy  $\|\mathbf{w}_r - \mathbf{w}_{r-1}\|^2$  is negligible.

We would like to point out the correspondence of (6) with the expression (1) for the LWPR filters. The weights now arise from generalized least square regression using  $\mathbf{B}_r(\varpi)^{-1}$  en lieu of the kernel. We shall refer to  $\beta_r^2(\varpi)$  as the *restricted concentration* index. It is interpreted as the maximum concentration at the frequency  $\varpi$  that a filter reproducing a polynomial of a degree  $p$  can achieve.

The concentration frequency has often been interpreted as a genuine cutoff frequency. We will argue that this interpretation, which has been responsible for the dismissal of this approach for low-pass filter design, is unwarranted. We are going to illustrate that perhaps  $\varpi$  is better interpreted as a kernel. Indeed, this interpretation is in line with the view of DPSS filters as *convergence factors*, supported for instance by Percival and Walden (1993, p. 182). Moreover, the comparison of expression (6) with the general expression of an LWPR filter (1) reinforces such interpretation: the only way  $\varpi$  affects the shape of the filter is via the matrix  $\mathbf{B}(\varpi)$ .

Figure 2 displays the value of percentage concentration,  $100 \times \beta^2(\varpi)$ , as a function of  $h$  and either  $\bar{P}$  (top left panel) or  $\varpi \in (0, \pi)$  (top right panel), for the standard, i.e. locally constant/linear, DPSS filters. The value of  $\beta^2(\varpi)$  is obtained from the first eigenvalue of the matrix  $\mathbf{A}(\varpi)$ . To enhance the interpretability of the underlying patterns, we present also the contour plots of  $100 \times \beta^2(\varpi)$  on the  $(h, \bar{P})$  plane (bottom left panel) and  $(h, \varpi)$  plane (bottom right panel). For given  $h$ , the concentration is a decreasing function of  $\bar{P}$  and an increasing function of  $\varpi$ . For small periods (high concentration frequencies), and for  $\bar{P} < h$ , the DPSS filter is 100% concentrated in the desired frequency range. The concentration is poor, instead, when the concentration period (frequency) is high (small) and the bandwidth is small; the latter provides a limiting factor.

The concentration contours enable to assess the trade-off between  $h$  and the concentration frequency: to achieve a given concentration level we can decrease  $h$ , providing we increase  $\bar{P}$  along a straight line, whose intercept and slope depend on the level of  $\beta^2(\varpi)$ , or equivalently decrease  $\varpi$  along an hyperbole. A remarkable feature is indeed that the combinations of  $(h, \bar{P})$  giving the same concentration lie along straight lines, whereas the equi-concentration points describe hyperboles on the  $(h, \omega_c)$  plane. The plot also informs us that if  $h = 12$  (e.g. three years of quarterly data, so that 25 consecutive observations are used to estimate the trend), we can estimate with unit concentration fluctuations with period up to 22 time units. To get a concentration equal to 1 for this concentration period a bandwidth  $h \geq 18$  would be needed.

## 4 Concluding remarks

The paper has characterized local polynomial filters as low-pass filters. We found a remarkable linear (hyperbolic) relationship between the cutoff period (frequency) and the bandwidth, conditional on the choices of the order and the kernel.

Secondly, the paper has proposed a generalized DPSS filter design approach that depends on three parameters: the bandwidth, the order of the polynomial that is reproduced by the filter, and the concentration frequency. The latter has been shown to play the role of a kernel, rather than of a cutoff frequency, in a low pass interpretation of DPSS filters.

## References

- LII, K.S. and ROSENBLATT, M. (2008): Prolate Spheroidal Spectral Estimates. *Statistics and Probability Letters*.
- LOADER, C. (1999): *SLocal Regression and Likelihood*. Springer-Verlag, New York.
- PERCIVAL, D. and WALDEN, A. (1993): *Spectral Analysis for Physical Applications*. Cambridge University Press, Cambridge.
- SLEPIAN, D. (1973): Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty, V. *The Bell System Technology Journal* 57, 1371–1430.

# A Statistical Survival Model Based on Counting Processes

Jose-Manuel Quesada-Rubio, Julia Garcia-Leal,  
Maria-Jose Del-Moral-Avila, Esteban Navarrete-Alvarez  
and Maria-Jesus Rosales-Moreno

Dpto. Estadística e I.O. - Facultad de Ciencias  
Campus de Fuentenueva s/n, Granada, Spain,  
*quesada@ugr.es, juliagl@ugr.es, delmoral@ugr.es,*  
*estebang@ugr.es, mrosales@ugr.es*

**Abstract.** We discuss some survival models with the intensity process of the counting process having a multiplicative structure. The most commonly used model is the Cox multiplicative hazard model. This model can be extended in different ways. We propose an additive-multiplicative model, where some of the covariates act multiplicatively on the risk function and others do so additively.

**Keywords:** survival analysis, counting process, martingales

## 1 Introduction

Survival analysis is a field of statistics that focuses on the study of time until the occurrence of a given event, generally called a "failure".

A large part of the modern theory of survival analysis uses counting processes. For a survival time we can define a counting process from which, by using the theorem of decomposition of Doob-Meyer, it is possible to connect directly to the theory of martingales. In this way, the theory of counting processes provides the theoretical tools necessary for the development of the theory of survival models, some related to the theory of martingales and others to stochastic integration, which help in the study of asymptotic properties of the estimators of the parameters and which have permitted a big advance in survival models in recent years.

Let  $T_1, T_2, \dots, T_n$  be i.i.d. continuous and non negative random variables on the probability space  $(\Omega, \mathcal{F}, P)$ , where  $T_i$  represents the survival time of the  $i$ -th individual for  $i = 1, 2, \dots, n$ , with *distribution function*  $F$ , *density function*  $f$  and *survival function*  $S$ , represents the probability of failure not occurring before time  $t$

$$S(t) = P(T_i > t) = 1 - F(t) \ .$$

The *hazard function*,  $\alpha(t)$ , measures the instantaneous risk of failure for individuals who have survived to time  $t$  (Cox (1972) or Cox and Oakes (1984))

$$\alpha(t) = \lim_{\Delta t \rightarrow 0+} \frac{P(T_i \in [t, t + \Delta t) / T_i \geq t)}{\Delta t} = \frac{f(t)}{S(t)} .$$

This hazard function completely determines the distribution of survival time from the relation

$$S(t) = P(T_i > t) = \exp \left\{ - \int_0^t \alpha(s) ds \right\} .$$

The cumulative hazard function is

$$A(t) = \int_0^t \alpha(s) ds .$$

Individuals that are not observed throughout their survival time are *censored observations*. The most common form of censoring is *right censoring*, in which the observed time is less than the actual survival time. Thus we observe the variable  $X_i$  defined as

$$X_i = \min\{T_i, U_i\} = T_i \wedge U_i$$

where  $U_i$  is the censoring time.

If an observation is censored, the *censoring indicator* is

$$D_i = I\{T_i = X_i\} = \begin{cases} 1 & \text{si } T_i \leq U_i \\ 0 & \text{si } T_i > U_i \end{cases}$$

where  $I\{\cdot\}$  is the indicator variable.  $D_i$  takes the value 1 if we observe the failure and 0 if the individual ceases to be observed before failure.

If the observation of some individuals is not possible until some time after the start of the study, left truncated data is obtained.

Furthermore, survival time can be affected by a number of Features, called *covariates*, these covariates are usually grouped into a vector, which for the  $i$ -th individual is usually represented as  $\mathbf{Z}_i$  or  $\mathbf{Z}_i(t)$ . One of the main objectives of studying survival time is to assess the influence that covariates have on the this.

A counting process is a stochastic process that counts/reckons the occurrences of a particular event over a period of time. A typical example of a counting process is the Poisson process. Fleming and Harrington (1991) define a counting process as “a stochastic process  $\{N(t) : t \in \mathcal{T}\}$  adapted to a filtration  $\{\mathcal{F}_t : t \in \mathcal{T}\}$  with  $N(0) = 0$  and  $N(t) < \infty$  a.s., and whose paths have a probability of one right-continuous, piecewise constant, and have only jump discontinuities, with jumps of size +1 ”.

A counting process is defined by its intensity process

$$\lambda(t) = \lim_{h \rightarrow 0} \frac{E(N(t+h) - N(t) | \mathcal{F}_t)}{h}.$$

The counting processes theory provides the theoretical tools needed for a rigorous development of models in several environments such as survival analysis. For a survival time we can define a counting process from which, by using the theorem of decomposition of Doob-Meyer, it is possible to connect directly to the theory of martingales. In right-censored survival data, the counting process is given by  $N(t) = I(X \leq t, D = 1)$ , where  $X = \min\{T, U\}$  and  $D = I(T \leq U) = I(X = T)$ , being  $T$  failure time and  $U$  censored time.

The Doob-Meyer theorem for counting processes can be expressed: Let  $N$  be arbitrary counting process. Then there exists a unique predictable process, right continuous and not decreasing  $A$ , such that  $A(0) = 0$  c.s.,  $A(t) < \infty$  c.s. for all  $t \in \mathcal{T}$  and  $M = N - A$  is a local martingale. If  $A$  is bounded locally, then  $M$  is a locally square integrable martingale.

In the events defined by counting processes, the intensity process is modeled by means of a multiplicative structure  $\lambda(t) = Y(t)\alpha(t)$ , where  $Y(t)$  is a left-continuous adapted process, and  $\alpha(t)$  is the hazard function which usually specifies the covariables dependence.

This treatment of survival analysis by means of counting processes has its origins in the work of Aalen (1978). Afterwards, Andersen and Gill (1982) studied the multiplicative hazard model of Cox incorporated the framework of the counting processes into, generalising the usual treatment of survival models. Thus they consider that the intensity process associated to the counting process for an individual  $i$  has the following form

$$\lambda_i(t) = Y_i(t)\alpha_0(t) \exp(\beta^T \mathbf{Z}_i(t)) ,$$

where  $\alpha_0(t)$ , the baseline hazard function, is an unknown nonnegative function;  $\beta = (\beta_1, \dots, \beta_p)^T$  is a  $p$ -vector of regression coefficients;  $\mathbf{Z}_i(t)$  is a  $p$ -vector of covariates processes, predictable and locally bounded,  $\mathbf{Z}_i(t) = (Z_{i1}(t), \dots, Z_{ip}(t))^T$ ; and  $Y_i(t)$  is a predictable processes which indicates when the individual  $i$ -th is under observation. Under certain stability and regularity conditions Andersen and Gill study the asymptotic properties of the estimators, using the theory of martingales and to the stochastic integration.

Another alternative to the models of multiplicative hazard are the models of additive hazard, in this sense Lin and Ying (1994) considered semiparametric models of the form

$$\lambda_i(t) = Y_i(t)\{\alpha_0(t) + \beta^T \mathbf{Z}_i(t)\}.$$

Some efforts have been made to combine the additive and multiplicative models in a mixed or combined hazard model. Lin and Ying (1995) consider the following to be an additive-multiplicative hazard model

$$\lambda_i(t) = Y_i(t) \{g(\delta^T \mathbf{W}_i(t)) + \alpha_0(t)h(\beta^T \mathbf{Z}_i(t))\}$$

where  $(\mathbf{Z}_i(t)^T, \mathbf{W}_i(t)^T)$  is a  $p + r$ -vector of covariates, which is predictable, and they propose the estimation of the parameters  $(\beta, \delta)$  by means of a semiparametric study of the model.

Parallel to the semiparametric models various non-parametric models have also been considered. Thus we can highlight the additive intensity model of Aalen (Aalen (1980, 1989), Huffer and McKeage (1991)) which has the form

$$\lambda_i(t) = Y_i(t)\beta(t)^T \mathbf{Z}_i(t) \ .$$

Scheike and Zhang (2002) suggest a non parametric model which combines the multiplicative model of Cox and the additive of Aalen, where the intensity is given as

$$\lambda_i(t) = Y_i(t)(\delta(t)^T \mathbf{W}_i(t)) \exp(\beta^T \mathbf{Z}_i(t)) \ ,$$

in this model the effects of some covariates act additively on the hazard and other multiplicatively.

One of the principal differences between nonparametric, semiparametric and parametric models lies in the baseline hazard function, the nonparametric models do not have baseline hazard function, the semiparametric models have an unknown baseline hazard function and the parametric models have a known baseline hazard function.

## 2 A parametric model of additive-multiplicative hazard

In what follows we propose a parametric additive-multiplicative survival model, with the starting point being the non-parametric multiplicative-additive model proposed by Scheike and Zhang (2002). The aim is to introduce a survival model that allows us to take advantage of the underlying information in the probability distributions involved in the study, after assessing the goodness of fit of the model to the considered data. The use of the information concerning the probability distribution that follows from the survival data allows more accurate estimates than those obtained by the nonparametric model. Let us consider

$$\lambda_i(t) = Y_i(t)\alpha_0(t, \gamma)(\delta^T \mathbf{W}_i(t)) \exp(\beta^T \mathbf{Z}_i(t))$$

where  $(\mathbf{Z}_i(t)^T, \mathbf{W}_i(t)^T)$  is a row vector with  $p + r$ -dimensional covariables processes, that we will suppose predictable;  $\beta$ ,  $\gamma$  and  $\delta$  are column vectors with  $p \times 1$ ,  $q \times 1$  and  $r \times 1$  dimensions respectively,  $\theta = (\gamma^T, \beta^T, \delta^T)^T$ ;  $Y_i(t)$  is a predictable process that indicates when the  $i$ -th individual is under observation and  $\alpha_0(t, \gamma)$  is the fixed baseline hazard function of which we know the form.

The local square integrable martingale associated to the counting process  $N_i$  is

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\alpha_0(s, \gamma)(\delta^T \mathbf{W}_i(s)) \exp(\beta^T \mathbf{Z}_i(s))ds \ .$$



The log-likelihood function for the estimation of the parameters is given by

$$\begin{aligned} C_\tau(\gamma, \beta, \delta) &= \\ &= \log L_\tau(\gamma, \beta, \delta) = \\ &= \sum_{i=1}^n \left( \int_0^\tau \log \{ Y_i(t) \alpha_0(t, \gamma) (\delta^T \mathbf{W}_i(t)) \exp(\beta^T \mathbf{Z}_i(t)) \} dN_i(t) \right. \\ &\quad \left. - \int_0^\tau Y_i(t) \alpha_0(t, \gamma) (\delta^T \mathbf{W}_i(t)) \exp(\beta^T \mathbf{Z}_i(t)) dt \right). \end{aligned}$$

The likelihood equations for  $(\gamma, \beta, \delta)$ ,  $\mathbf{U}_\tau(\gamma, \beta, \delta) = \mathbf{0}$ , are given by

$$\begin{aligned} \frac{\partial}{\partial \gamma} C_\tau(\gamma, \beta, \delta) &= \\ &= \sum_{i=1}^n \int_0^\tau \psi(t, \gamma) \{ dN_i(t) - Y_i(t) \alpha_0(t, \gamma) (\delta^T \mathbf{W}_i(t)) \exp(\beta^T \mathbf{Z}_i(t)) dt \} \\ &= \sum_{i=1}^n \int_0^\tau \psi(t, \gamma) dN_i(t) - \int_0^\tau \alpha_0(t, \gamma) S^{(0)(0)}(\beta, \delta, t) dt, \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \beta} C_\tau(\gamma, \beta, \delta) &= \\ &= \sum_{i=1}^n \int_0^\tau \mathbf{Z}_i(t) \{ dN_i(t) - Y_i(t) \alpha_0(t, \gamma) (\delta^T \mathbf{W}_i(t)) \exp(\beta^T \mathbf{Z}_i(t)) dt \} = \\ &= \sum_{i=1}^n \int_0^\tau \mathbf{Z}_i(t) dN_i(t) - \int_0^\tau \alpha_0(t, \gamma) \mathbf{S}^{(1)(0)}(\beta, \delta, t) dt, \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \delta} C_\tau(\gamma, \beta, \delta) &= \\ &= \sum_{i=1}^n \int_0^\tau \frac{\mathbf{W}_i(t)}{\delta^T \mathbf{W}_i(t)} dN_i(t) - \int_0^\tau \alpha_0(t, \gamma) \mathbf{S}^{(0)(1)}(\beta, \delta, t)^T dt = \\ &= \sum_{i=1}^n \int_0^\tau \frac{\mathbf{W}_i(t)}{\delta^T \mathbf{W}_i(t)} \{ dN_i(t) - Y_i(t) \alpha_0(t, \gamma) (\delta^T \mathbf{W}_i(t)) \exp(\beta^T \mathbf{Z}_i(t)) dt \}, \end{aligned}$$

where  $\psi(t, \gamma) = \frac{\partial}{\partial \gamma} \log \alpha_0(t, \gamma)$ , and

$$\mathbf{S}^{(k)(h)}(\beta, \delta, t) = \sum_{i=1}^n \mathbf{Z}_i(t)^{\otimes k} \{ \mathbf{W}_i(t)^{\otimes h} \}^T (\delta^T \mathbf{W}_i(t))^{1-h} Y_i(t) \exp(\beta^T \mathbf{Z}_i(t))$$

for  $k, h = 0, 1, 2$  and  $k + h \leq 2$ ,  $\mathbf{Z}^{\otimes 0} = 1$ ,  $\mathbf{Z}^{\otimes 1} = \mathbf{Z}$  and  $\mathbf{Z}^{\otimes 2} = \mathbf{Z}\mathbf{Z}^T$ .

The observed information matrix is then

$$\mathcal{I}_\tau(\theta) = -\frac{\partial}{\partial \theta} \mathbf{U}_\tau(\theta) = \begin{pmatrix} \mathcal{I}_{\gamma\gamma}(\theta)_{q \times q} & \mathcal{I}_{\gamma\beta}(\theta)_{q \times p} & \mathcal{I}_{\gamma\delta}(\theta)_{q \times r} \\ \mathcal{I}_{\gamma\beta}^T(\theta)_{p \times q} & \mathcal{I}_{\beta\beta}(\theta)_{p \times p} & \mathcal{I}_{\beta\delta}(\theta)_{p \times r} \\ \mathcal{I}_{\gamma\delta}^T(\theta)_{r \times q} & \mathcal{I}_{\beta\delta}^T(\theta)_{r \times p} & \mathcal{I}_{\delta\delta}(\theta)_{r \times r} \end{pmatrix}$$

with

$$\begin{aligned} \mathcal{I}_{\gamma\gamma}(\theta) &= \\ &= -\frac{\partial^2}{\partial \gamma \partial \gamma^T} C_\tau(\gamma, \beta, \delta) = \\ &= \sum_{i=1}^n \left[ \int_0^\tau Y_i(t) \alpha_0^{(2)}(t, \gamma) (\delta^T \mathbf{W}_i(t)) \exp(\beta^T \mathbf{Z}_i(t)) dt - \right. \\ &\quad \left. - \int_0^\tau \frac{\partial^2}{\partial \gamma \partial \gamma^T} \log \alpha_0(t, \gamma) dN_i(t) \right] = \\ &= \int_0^\tau \psi(t, \gamma)^{\otimes 2} S^{(0)(0)}(\beta, \delta, t) dt - \sum_{i=1}^n \int_0^\tau \frac{\partial^2}{\partial \gamma \partial \gamma^T} \log \alpha_0(t, \gamma) dM_i(t), \end{aligned}$$

where  $\alpha_0^{(k)}(t, \gamma) = \frac{\partial^k}{\partial \gamma^{\otimes k}} \alpha_0(t, \gamma) \quad k = 0, 1, 2$  ,

$$\mathcal{I}_{\gamma\beta}(\theta) = -\frac{\partial^2}{\partial \gamma \partial \beta^T} C_\tau(\gamma, \beta, \delta) = \int_0^\tau \alpha_0^{(1)}(t, \gamma) \mathbf{S}^{(1)(0)}(\beta, \delta, t)^T dt ,$$

$$\mathcal{I}_{\gamma\delta}(\theta) = -\frac{\partial^2}{\partial \gamma \partial \delta^T} C_\tau(\gamma, \beta, \delta) = \int_0^\tau \alpha_0^{(1)}(t, \gamma) \mathbf{S}^{(0)(1)}(\beta, \delta, t) dt ,$$

$$\mathcal{I}_{\beta\beta}(\theta) = -\frac{\partial^2}{\partial \beta \partial \beta^T} C_\tau(\gamma, \beta, \delta) = \int_0^\tau \alpha_0(t, \gamma) \mathbf{S}^{(2)(0)}(\beta, \delta, t) dt ,$$

$$\mathcal{I}_{\beta\delta}(\theta) = -\frac{\partial^2}{\partial \beta \partial \delta^T} C_\tau(\gamma, \beta, \delta) = \int_0^\tau \alpha_0(t, \gamma) \mathbf{S}^{(1)(1)}(\beta, \delta, t) dt ,$$

$$\begin{aligned} \mathcal{I}_{\delta\delta}(\theta) &= -\frac{\partial^2}{\partial \delta \partial \delta^T} C_\tau(\gamma, \beta, \delta) = \sum_{i=1}^n \left[ \int_0^\tau \frac{\mathbf{W}_i(t)^{\otimes 2}}{(\delta^T \mathbf{W}_i(t))^2} dN_i(t) \right] = \\ &= \int_0^\tau \alpha_0(t, \gamma) \mathbf{S}^{(0)(2)}(\beta, \delta, t) dt + \sum_{i=1}^n \int_0^\tau \frac{\mathbf{W}_i(t)^{\otimes 2}}{(\delta^T \mathbf{W}_i(t))^2} dM_i(t) . \end{aligned}$$

The estimations are obtained from the likelihood equations through numeric methods, due to the difficulty that these would otherwise present for their resolution.

By the Rebolledo's theorem for local martingale and on the Lengart's inequality we have derived the asymptotic properties of the estimators.

### 3 Discussion

When survival analysis uses non-parametric models, a generic model that fits the available data without taking into account the information provided by the nature of such data is formulated. In this way the information that would have been obtained if it were assumed that this data followed a specific probability distribution is unused or wasted, hence leading to models which depend very closely on the specific data considered in their study. This can be overcome using parametric survival models because a great deal of information about the underlying distributions can be derived from such models (Weibull, exponential, etc). It is only necessary to evaluate that the proposed model fits the data available. In situations where one knows the basic risk function, the use of parametric models allows for more accurate estimates using information concerning the probability distribution that follows from the survival data. In these cases the use of parametric survival models is therefore more appropriate. In this article, and with reference to the article of Scheike and Zhang (2002) in which they propose a non-parametric additive-multiplicative survival model, we propose a parametric survival additive-multiplicative model. The additive-multiplicative models allow the description of what is a common occurrence in practice: some covariates act additively on the hazard and others do so in a multiplicative manner. Added to this is the fact that these models allow the effects of some covariates that may vary over time, while others may be constant over time. One of the main advantages of considering the additive component in survival models as previously mentioned is that, in this way, the calculations on the effects of covariates that vary over time can be simplified.

### References

- AALEN, O.O. (1978): Nonparametric inference for a family of counting processes. *Ann. Statist.* 6, 701-726.
- AALEN, O.O. (1980): A model for non-parametric regression analysis of counting processes. In Klonecki, W., Kozek, A. & Rosinski, J., editors, *Lecture Notes in Statistics-2: Mathematical Statistics and Probability Theory*. Springer-Verlag New York, 1-25.
- AALEN, O.O. (1989): A linear regression model for the analysis of life times. *Statistics in Medicine* 8, 907-925.
- ANDERSEN, P.K., BORGAN, Ø, GILL, R.D. and KEIDING, N. (1993): *Statistical Models Based on Counting Processes*. Springer-Verlag New York.

- ANDERSEN, P.K. & GILL, R.D. (1982): Cox's regression model for counting processes: A large sample study. *Ann. Statist.* 10, 1100-1120.
- BHATTACHARJEE, A. (2004): Estimation in hazard regression models under ordered departures from proportionality. *Comput. Stat. Data Anal.* 47, 517-536.
- CORTESE, G. and SCHEIKE, T.H. (2008): Dynamic regression hazards models for relative survival. *Statistics in medicine* 27, 3563-3584.
- COX, D.R. (1972): Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* 34, 187-220.
- COX, D.R. and OAKES, D. (1984): *Analysis of Survival Data*. Chapman and Hall, London.
- FLEMING, T.R. and HARRINGTON D.P. (1991): *Counting Processes and Survival Analysis*. Wiley & Sons, New York.
- HUFFER, F.W. and MCKEAGE, I.W. (1991): Weighted least squares estimation for Aalen's additive risk model. *J. Amer. Statist. Assoc.* 86, 114-129.
- KRAUS, D. (2004): Goodness-of-fit inference for the Cox-Aalen additive-multiplicative regression model. *Statistics & Probability Letters* 70, 285-298.
- KULICH, M. and LIN, D.Y. (2000): Additive hazards regression for case-cohort studies. *Biometrika* 87, 73-87.
- LIN, D.Y. and YING, Z. (1994): Semiparametric analysis of the additive risk model. *Biometrika* 81, 61-71.
- LIN, D.Y. and YING, Z. (1995): Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *Ann. Statist.* 23, 1712-1734.
- MARTINUSSEN, T. and SCHEIKE, T.H. (2006): *Dynamic Regression Models for Survival Data*. Springer. New York.
- QUESADA-RUBIO, J.M., GARCIA-LEAL, J., LARA-PORRAS, A.M. and NAVARRETE-ALVAREZ, E. (2001): An Additive Intensity Model in a Multivariate Process Counting. *Revista de Estatística. Portugal. Volume II, 2 Quadrimestre*, 329-330.
- QUESADA-RUBIO, J.M. (2002): *Aportaciones en Análisis de Supervivencia*. PhD thesis, Univ. of Granada.
- SCHEIKE, T.H. and ZHANG, M. (2002): An additive-multiplicative Cox-Aalen regression model. *Scand. J. Statist.* 29, 75-88.
- SCHEIKE, T.H. and ZHANG, M. (2003): Extensions and Applications of the Cox-Aalen Survival Model. *Biometrics* 59, 1036-1045.
- SUNDARAM, R. (2006): Semiparametric inference for the proportional odds model with time-dependent covariates. *Journal of Statistical Planning and Inference* 136, 320-334.

# Bootstrapping Additive Models in Presence of Missing Data

Rocío Raya-Miranda, M. Dolores Martínez-Miranda and Andrés  
González-Carmona

Department of Statistics and O.R.  
c/ Severo Ochoa, s/n 18071 Granada, Spain, *rraya@ugr.es*

**Abstract.** The problem of estimating nonparametric additive models when missing data appear in the response variable is considered. Two estimators are constructed from the Backfitting Local Polynomial estimators by Opsomer (2000). The simplest (SB) is defined by using the available data and another imputed version (IB) is based on a complete sample from imputation techniques. The problem of selecting the smoothing parameter is solved by using a local selector, which is based on a Wild Bootstrap approximation of the Mean Squared Error. Several simulation experiments illustrate the performance of the proposed methods.

**Keywords:** imputation, backfitting, wild bootstrap, smoothing parameter

## 1 Introduction

Statistical inference with missing data carries a long historical background. Missing values arise in empirical studies for many reasons, in sample surveys, in clinical essays, in studies with longitudinal and ecological data, etc. If any of the items that form the index are missing, some procedure is needed to deal with the missing data.

The standard nonparametric regression estimation methods usually consider complete samples for independent observations. There exist many ways to deal with missing data problems, ranging from the most naive one of focusing on the complete cases only to other more complex method consists of imputing the missing values with an estimate. However dealing with missing data via parametric imputation methods usually implies stating several strong assumptions. If such parametric assumptions do not hold, the imputed data are not appropriate and might produce inconsistent estimators and thus misleading results. Other more refined imputation methods are those of Dempster et al. (1977), which develop the EM algorithm, or those based on multiple imputation (Rubin (1987)). Recently, multiple imputation methods have been proposed based on nonparametric and semiparametric estimation (Aerts et al. (2002)).

Several inferential problems have been considered in the presence of missing data, however, the nonparametric additive models with missing data have not been paid special attention yet. In multivariate regression, the missing

data can arise on the response and/or the covariates in the model. In this paper we will focus our attention only when missing data occur only in the responses and the covariates are totally observed. In this paper we look at two nonparametric estimators based on the Backfitting algorithm with local polynomial smoothers (Opsomer and Ruppert (1997)). The first is the Simplified Backfitting (SB), which uses only complete observations for the estimation and therefore discards incomplete vectors. The second denoted as the Imputed Backfitting (IB) consists of initially employing the SB in order to impute the missing responses and then estimates the regression function with the completed sample.

An important issue in any nonparametric estimation is the problem of choosing the smoothing or bandwidth parameter from the available data. Martínez-Miranda et al. (2008) proposed a local bandwidth selector for Backfitting additive estimates based on the Bootstrap method. They proved the consistency of the bootstrap approximation for the bivariate case (when local linear smoothers are involved) and also the optimality of the bootstrap bandwidth selector. They also obtained an expression that can be evaluated in practice by computing exactly the expectation according to the resampling distribution. Thus it is possible to achieve the practical implementation of the proposed local bootstrap bandwidth selector without getting involved in Monte Carlo approximations. In this paper the estimators are based in the ideal situation of whole datasets. In this paper we propose an extension of the method involving missing-data adjustments. The selector also extends the previous work in multivariate local linear estimation by González-Manteiga et al. (2004).

## 2 Simple imputation

For  $i = 1, \dots, n$ , it is assumed for one-dimensional response variables  $Y_1, \dots, Y_n$  that

$$Y_i = m_0 + m_1(X_{1i}) + \dots + m_d(X_{di}) + \varepsilon_i, \quad (1)$$

where  $m_1, \dots, m_d$  are unknown functions satisfying  $E[m_j(X_{ji})] = 0$ ,  $m_0$  is an unknown constant,  $\varepsilon_i$  are error variables and  $\mathbf{X}_i = (X_{1i}, \dots, X_{di})$  are random variables  $\mathbb{R}^d$  ( $i = 1, \dots, n$ ,  $j = 1, \dots, d$ ). Throughout the paper we make the assumption that  $\mathbf{X}_1, \dots, \mathbf{X}_n$  are independent identically distributed (i.i.d.) and that  $X_{ji}$  takes its values in a bounded interval  $I_j$ . Furthermore, the error variables,  $\varepsilon_1, \dots, \varepsilon_n$ , are assumed to be i.i.d. with mean zero and being independent of  $\mathbf{X}_1, \dots, \mathbf{X}_n$ .

In the case where no observations are missing, a sample of i.i.d. vectors with respect to the random vector  $\{(\mathbf{X}_i^t, Y_i)\}_{i=1}^n$  is available. In our case it may be possible that  $Y_i$  is not observed for any index  $i$ . In order to check whether an observation is complete or not, a new variable  $\delta$  is introduced into the model as an indicator of the missing observations. Thus  $\delta_i = 1$  if  $Y_i$  is observed and zero if  $Y_i$  is missing, for  $i = 1, \dots, n$ .

Following the literature (see Little and Rubin (2002)), we need to establish whether the loss of an item of data is independent or not of the value of the observed data and/or the missing data. In this paper we suppose that the data are missing at random (MAR), i.e.  $P[\delta = 1|Y, \mathbf{X}] = P[\delta = 1|\mathbf{X}] = p(\mathbf{X})$ .

If there are not missing data, the nonparametric estimation of the component additive functions,  $\mathbf{m}_j$ , can be given by solving the system of normal equations:

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_d & \mathbf{S}_d & \cdots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_d \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_d \end{bmatrix} \mathbf{Y}. \quad (2)$$

where  $\mathbf{S}_j$  represents the  $n \times n$  smoother matrix with respect to the  $j$ th covariate vector. The smoother matrices for local polynomial regression are  $\mathbf{S}_j = (\mathbf{s}_{j,X_{d1}}, \dots, \mathbf{s}_{j,X_{dn}})^T$ , where  $\mathbf{s}_{j,x_j}$  represents the equivalent kernel for the  $j$ th covariate at the point  $x_j$ :

$$\mathbf{s}_{j,x_j} = \mathbf{e}_1^T \left( \mathbf{X}_{j,x_j}^T \mathbf{K}_{x_j} \mathbf{X}_{j,x_j} \right)^{-1} \mathbf{X}_{j,x_j} \mathbf{K}_{x_j},$$

with  $\mathbf{e}_i$  a vector with a one in the  $i$ th position and zeros elsewhere, the matrix  $\mathbf{K}_{x_j} = \text{diag}\{K_{h_j}(X_{j1} - x_j), \dots, K_{h_j}(X_{jn} - x_j)\}$  for some kernel function  $K$  and bandwidth  $h_j$ ,

$$\mathbf{X}_{j,x_j} = \begin{bmatrix} 1 & (X_{j1} - x_j) & \cdots & (X_{j1} - x_j)^{p_j} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_{jn} - x_j) & \cdots & (X_{jn} - x_j)^{p_j} \end{bmatrix},$$

and  $p_j$  the degree of the local polynomial for fitting  $\mathbf{m}_j$ .

The Backfitting algorithm, Buja et al. (1999) provides an iterative solution of (2). Opsomer (2000) wrote the estimators directly as

$$\begin{bmatrix} \hat{\mathbf{m}}_1 \\ \hat{\mathbf{m}}_2 \\ \vdots \\ \hat{\mathbf{m}}_d \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_d & \mathbf{S}_d & \cdots & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_d \end{bmatrix} \mathbf{Y} \equiv \mathbf{M}^{-1} \mathbf{C} \mathbf{Y},$$

provided the inverse of  $\mathbf{M}$  exists. This expression shows that the additive model estimators are also linear smoothers, and it is possible to define the additive smoother matrix  $\mathbf{W}_j$  as  $\mathbf{W}_j = \mathbf{E}_j \mathbf{M}^{-1} \mathbf{C}$ , where  $\mathbf{E}_j$  is a partitioned matrix of dimension  $n \times nd$  with an  $n \times n$  identity matrix as the  $j$ th block and zeros elsewhere, so that  $\hat{\mathbf{m}}_j = \mathbf{W}_j \mathbf{Y}$ , for  $j = 1, \dots, d$ .

If there are missing observations in the response variable, a very simple way to estimate the regression function is the Simplified Backfitting (SB),

which consists of using only complete observations, in other words, those where  $\delta_i = 1$ . Thus the SB can be obtained as

$$\hat{\mathbf{m}}_{SB,j} = \mathbf{W}_j^\delta \mathbf{Y}, \quad j = 1, \dots, d, \quad (3)$$

where  $\mathbf{W}_j^\delta = \mathbf{E}_j(\mathbf{M}^\delta)^{-1} \mathbf{C}^\delta$ .

And the smoother matrices  $\mathbf{S}_j^\delta = (\mathbf{s}_{j,X_{d1}}^\delta, \dots, \mathbf{s}_{j,X_{dn}}^\delta)^T$ , where  $\mathbf{s}_{j,x_j}^\delta$  :

$$\mathbf{s}_{j,x_j}^\delta = \mathbf{e}_1^T \left( \mathbf{X}_{j,x_j}^T \mathbf{K}_{x_j}^\delta \mathbf{X}_{j,x_j} \right)^{-1} \mathbf{X}_{j,x_j} \mathbf{K}_{x_j}^\delta,$$

the matrix  $\mathbf{K}_{x_j}^\delta = \text{diag}\{K_{h_j}(X_{j1} - x_j)\delta_1, \dots, K_{h_j}(X_{jn} - x_j)\delta_n\}$ .

Another option is the Imputed Backfitting (IB), which is constructed in two stages. In the first stage, the SB is used to estimate the missing observations so as to complete the sample. In this way a completed sample,  $\{(\mathbf{X}_i, \hat{Y}_i), i = 1, \dots, n\}$ , is obtained where  $\hat{Y}_i = \delta_i Y_i + (1 - \delta_i) \hat{\mathbf{m}}_{SB}(\mathbf{X}_i)$ , with  $\hat{\mathbf{m}}_{SB}(\mathbf{X}_i) = \hat{\mathbf{m}}_{SB,0} + \sum_{j=1}^d \hat{\mathbf{m}}_{SB,j}(X_{ji})$  being the SB estimation of the additive function  $m$ , evaluated on  $\mathbf{X}_i$ . Once the sample is completed, Backfitting is applied to the data  $\{(\mathbf{X}_i, \hat{Y}_i), i = 1, \dots, n\}$ , where  $(\hat{Y}_1, \dots, \hat{Y}_n)^t$  is the imputed response vector.

### 3 Local bootstrap bandwidth selection

The first step to propose a proper data-driven bandwidth selector is to define the optimality criterion. The easiest and most common criterion in nonparametric regression is the conditional Mean Integrated Squared Error (MISE). Bandwidth parameters chosen from such a global criterion are constant along the estimation interval. This feature can be inefficient when the target regression surface exhibits irregularities and also when the sample sizes are smaller. On the other hand local bandwidth selectors based on local error criteria such as the Mean Squared Error (MSE) allow a nice adaptation to the available data and capture the real features in the underlying surface.

The method consists of obtaining the local optimal bandwidth parameter by minimizing the Bootstrap estimator of the Mean Squared Error. For the SB and IB estimators we have, respectively:

$$\min_{\mathbf{h}} MSE_{SB}^*(\mathbf{x}; \mathbf{h}) = \min_{\mathbf{h}} E^* [\hat{m}_{SB;\mathbf{h}}^*(\mathbf{x}) - \hat{m}_{SB;\mathbf{h}_0}(\mathbf{x})]^2, \quad (4)$$

$$\min_{\mathbf{h}, \mathbf{g}} MSE_{IB}^*(\mathbf{x}; \mathbf{h}, \mathbf{g}) = \min_{\mathbf{h}, \mathbf{g}} E^* [\hat{m}_{IB;\mathbf{h}, \mathbf{g}}^*(\mathbf{x}) - \hat{m}_{IB;\mathbf{h}_0, \mathbf{g}_0}(\mathbf{x})]^2, \quad (5)$$

where  $\mathbf{h}_0, \mathbf{g}_0$  are pilot bandwidth vectors,  $E^*$  is the expectation operator for the Bootstrap resampling, and  $\hat{m}^*$  is the regression function estimation with the Bootstrap sample.

The minimization of the above expressions, (4 and 5), has to be carefully made. Indeed the results can be dramatically different depending on how you



minimize. Here we follow the discussion of Nielsen and Sperlich (2005) who consider an algorithm which look for the minimum of each component by separately. The problem is that we cannot decompose the MSE in each component but we can focus first in the minimization in a particular component and fix the other ones. Therefore we avoid the problem of minimizing in a multidimensional grid with high size. More specifically we have considered the following strategy to derive the bootstrap smoothing parameter:

First, for some initial bandwidth  $\mathbf{g}^0 \in \mathbb{R}^d$ , calculate the  $\hat{\mathbf{m}}_{SB, \mathbf{g}^0}$  and  $\hat{\mathbf{m}}_{IB, \mathbf{g}^0}$ . Then, direction for direction look for the  $h_j(x_j)$  minimizing the *MSE* value. Having arrived at  $j = d$  and found the appropriate  $h_j(x_j)$  we must update all estimates, with the new bandwidths. The same procedure has been carried out for the SB and IB estimators. The algorithm scheme, in general, is as follows.

- Step 1: take starting bandwidths  $g_1^{(0)}, \dots, g_d^{(0)}$ , set  $r = 0$ . A natural choice of starting values  $g_j^{(0)}$ ,  $j = 1, \dots, d$ , would be minimize a global cross-validation criterion for an additive model.
- Step 2: calculate  $\hat{m}_j^{(r)}(x_j)$ ,  $j = 1, \dots, d$ .
- Step 3: for  $j = 1, \dots, d$ ,
  - (a) find the  $h_j^{(r)}(x_j)$  that minimizes the bootstrap estimator of mean squared error value, by a simple one-dimensional grid search;
  - (b) when the optimal  $h_j^{(r)}(x_j)$  has been found, update the components  $\hat{m}_j^{(r)}(x_j)$ ; then set  $j$  to  $j + 1$ .
- Step 4: if the  $MSE^*(\mathbf{x})$  has improved, set  $r$  to  $r + 1$  and go to step 3; otherwise stop.

Define the vector  $\hat{\mathbf{h}}_{opt}(\mathbf{x}) = \mathbf{h}^{(k)}(\mathbf{x})$  with  $k$  the index of the last iteration. Then,  $\hat{\mathbf{h}}_{opt}(\mathbf{x})$  is the bandwidth that minimizes the Bootstrap estimator of Mean Squared Error at each estimation point  $\mathbf{x}$ .

## 4 Simulation study

We next present a simulation study in which both estimators (SB and IB) are compared for Mean-Squared Error (*MSE*) over 100 samples,  $(\mathbf{X}_i, Y_i)$ ,  $i = 1, \dots, n$ , of size 200 from two additive models, with around 40% of missing data. We also investigate the performance of the proposed bootstrap approximation of the conditional mean squared error and the local bootstrap bandwidth selector based on it.

To this end we estimate the additive model by SB and IB. As far as estimates are concerned, we have compared the local bootstrap selector proposed above and a global selector based on the cross-validation method. The local bootstrap bandwidth selector requires the choice of a pilot bandwidth parameter,  $\mathbf{g}$ ; we have relied on the pilot bandwidths obtained by a global cross-validation selector.

The component functions, design distributions and missing data mechanism specified as follows:

The model 1 with  $m_j(x_j) = 2 \sin(\pi x_j)$ ,  $j = 1, 2$ , with two different missing data mechanisms,  $p_1(\mathbf{x}) = 1/(2.2 + \exp\{-0.6x_1^2\})$  and  $p_2(\mathbf{x}) = 1/(2.2 + \exp\{-0.6x_1x_2\})$ , and the model 2 with  $m_1(x_1) = x_1^2$ ,  $m_2 = 2 \sin(0.5\pi x_2)$  and  $p(\mathbf{x}) = 1/(2 + 0.5 \exp\{x_1^2\})$ .

The explicative variables were generated as follows: In model 1, for the design we first draw variables  $z_j$ ,  $j = 1, 2$  from  $N(0, 1)$  with correlation  $\rho_{12} = \rho$ . To facilitate the grid search, we projected the variables into the interval  $(-1.25, 1.25)$  by the transformation  $x_j := 2.5 \tan^{-1}(z_j)/\pi$ . Therefore  $\rho$  is not the correlation between the  $x_j$  but is related to it. We concentrate on the estimation inside  $[-1, 1]^2$ . In model 2, the covariates were generated from truncated bi-dimensional normal distributions  $(X_1, X_2)^T \sim N(0, \Sigma_\gamma)$ ,  $\gamma = 1, 2$ , where the covariance matrices are given by

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}.$$

The truncation was done for the covariates to have the compact support  $[-1, 1]^2$ . To be specific, a random variable generated from one of the bi-dimensional normal distributions was discarded if one of the covariates fell outside the interval  $[-1, 1]$ . The residuals were generated from a normal distribution with mean zero and constant variance of 1. The local linear smoother involved in the Backfitting algorithm was calculated with a gaussian kernel,  $K(x) = (2\pi)^{(-1/2)} \exp(-x^2/2)$ .

Here we evaluate the empirical performance of an estimate  $\hat{m}$  of  $m$ , by calculating its Integrated Squared Error (*ISE*). This measure is computed over the performed replications of the model and it is given by

$$ISE(\hat{m}) = \sum_{j=1}^r (m(\mathbf{v}_j) - \hat{m}(\mathbf{v}_j))^2 / r$$

where  $\mathbf{v}_j$  is a grid of  $21 \times 21$  bidimensional estimation points. An approximation to the *MISE* was obtained as the mean over the 100 replications of *ISE*.

Tables 1 and 2 summarize the results of the simulations. Table 1 gives the values of the *MISE*, the efficiency for the SB and IB estimators and also the percentage of times that the *ISE* for the IB is smaller than the *ISE* for the SB when using the local bootstrap bandwidth.

The reported efficiency of the IB estimator with respect to the SB estimator was computed as

$$EF_{SB, IB} = \frac{MISE_{SB} - MISE_{IB}}{MISE_{SB}} \times 100,$$

where  $MISE_{SB}$  denoted the *MISE* of the SB estimator and  $MISE_{IB}$  that of the IB estimator.

	Model 1			Model 2	
	$\rho = 0, p_1$	$\rho = 0, p_2$	$\rho = 0.3$	$\rho = 0$	$\rho = 0.4$
<i>SB</i>	0.353895 (0.092856)	0.880437 (0.263632)	0.360574 (0.115311)	0.174554 (0.070258)	0.168112 (0.077478)
<i>IB</i>	0.278069 (0.067206)	0.796633 (1.256642)	0.300156 (0.075724)	0.126046 (0.035916)	0.105898 (0.038871)
$EF_{SB,IB}$	21.426	9.518	16.794	27.789	37.007
$\%ISE_{IB} < ISE_{SB}$	87	78	72	87	95

**Table 1.** Approximated Mean Integrated Squared Error (*MISE*) for the SB and IB estimators. The standard deviation is showed into brackets.

The results reported show that the IB estimator provides smaller *ISE* values than those associated to the SB ones and also very high percentages of times that the *ISE* for the IB is smaller than the *ISE* for the SB.

In order to compare the results obtained between the bootstrap local bandwidth and a global one the Table 2 gives the values of the *MISE* for the SB and IB estimators when considering the local bandwidth obtained applying the bootstrap method and the global one obtained from a cross validation method, for the model 1 (the results for the other model have been omitted because they provide similar information). It also included the efficiency and the percentage of times that the *ISE* for the Bootstrap is smaller than the *ISE* for the cross-validation method.

	Model 1: SB estimator		
	$\rho = 0, p_1$	$\rho = 0, p_2$	$\rho = 0.3$
$SB_{CV}$	0.535402 (0.334541)	0.930065 (0.213821)	0.492153 (0.298474)
$SB_{Boot}$	0.353895 (0.092856)	0.880437 (0.263632)	0.360574 (0.115311)
$EF_{CV,Boot}$	33.901	5.336	26.7353
$\%ISE_{Boot} < ISE_{CV}$	62	55	56

	Model 1: IB estimator		
	$\rho = 0, p_1$	$\rho = 0, p_2$	$\rho = 0.3$
$IB_{CV}$	0.487348 (0.337810)	0.859504 (0.206315)	0.443691 (0.297271)
$IB_{Boot}$	0.278069 (0.067206)	0.796634 (0.256641)	0.300156 (0.075724)
$EF_{CV,Boot}$	42.942	7.315	32.350
$\%ISE_{Boot} < ISE_{CV}$	61	50	54

**Table 2.** Approximated Mean Integrated Squared Error (*MISE*) for local (Bootstrap) and global (CV) bandwidth for the SB and IB estimators. The standard deviation is showed into brackets.

The reported efficiency of the local bandwidth with respect to the global one was computed as

$$EF_{CV,Boot} = \frac{MISE_{CV} - MISE_{Boot}}{MISE_{CV}} \times 100,$$

where  $MISE_{CV}$  denotes the  $MISE$  of the estimator using a global bandwidth calculated by cross validation and  $MISE_{Boot}$  that of the estimator using a local bootstrap bandwidth. From these results we see that the local bandwidths outperform over the global ones, achieving smaller  $ISE$  values.

#### 4.1 Concluding remarks

The simulations reveal that the IB estimator performs better than the SB estimator, which is reflected in very high percentages of times that the  $ISE$  for the IB is smaller than the  $ISE$  for the SB. On the other hand, the performance of the IB and SB estimators is sensitive (the  $MISE$  of the estimators increase) to the presence of correlation among the variables. The proposed local bootstrap bandwidth selector provides satisfactory bandwidth estimates which outperform over classical global cross-validation.

#### Acknowledgments

The authors would like to thank the reviewers for their suggestions about the simulation study which has led to improve the presentation of the paper. This research was financially supported by Project MTM2008-03010/MTM.

#### References

- AERTS, M., CLAESKENS, G., HENS, N. and MOLENBERGS, G. (2002): Local multiple imputation. *Biometrika* 89 (2), 375-388.
- BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989): Linear smoothers and additive models (with discussion). *The Annals of Statistics* 17, 453-555.
- DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977): Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society* 39, 1-38.
- GONZÁLEZ-MANTEIGA, W. and PÉREZ-GONZÁLEZ, A. (2004): Nonparametric mean estimation with missing data. *Communications in Statistics, Theory and Methods* 33 (2), 277-303.
- LITTLE, R.J.A. and RUBIN, D.B. (2002): *Statistical Analysis with Missing Data*. Wiley-Interscience.
- MARTÍNEZ-MIRANDA, M.D., RAYA-MIRANDA, R., GONZÁLEZ-MANTEIGA, W. and GONZÁLEZ-CARMONA, A. (2008): A bootstrap local bandwidth selector for additive models. *Journal of Computational and Graphical Statistics* 17, 38-55.
- NIELSEN, J.P. and SPERLICH, S. (2005): Smooth backfitting in practice. *Journal of the Royal Statistical Society, Series B* 67 (1), 43-61.
- OPSOMER, J.D. (2000): Asymptotic properties of backfitting estimators. *Journal of the Multivariate Analysis* 73, 166-179.
- OPSOMER, J.D., and RUPPERT, D. (1997): Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, 25, 186-293.
- RUBIN, D.B. (1987): *Multiple Imputation for Nonresponse in Surveys*. Wiley & Sons.

# On Aspects of Quality Indexes for Scoring Models

Martin Řezáč<sup>1</sup> and Jan Kolářček<sup>1</sup>

Department of Mathematics and Statistics, Masaryk University  
Kotlářská 2, 611 37 Brno, Czech Republic, *mrezac@math.muni.cz*

**Abstract.** Credit scoring models are widely used to predict a probability of an event like client's default. To measure the quality of the scoring models it is possible to use quantitative indexes such as Gini index, K-S statistics, C-statistics and Lift. They are used for comparison of several developed models at the moment of development as well as for monitoring of quality of those models after deployment into real business. The paper deals with mentioned quality indexes, their properties and relationships. The main contribution of the paper is proposition and discussion of indexes and curves based on Lift. Curve of ideal Lift is defined, Lift ratio is proposed as analogy to Gini index. Integrated Relative Lift is defined and discussed.

**Keywords:** credit scoring, quality indexes, gini index, lift, integrated relative lift

## 1 Introduction

Banks and other financial institutions receive thousands of credit applications every day (in case of consumer credits it can be tens or hundreds of thousands every day). Since it is impossible to process them manually, automatic systems are widely used by these institutions for evaluating credit reliability of individuals who ask for credit. The assessment of the risk associated with granting of credits has been underpinned by one of the most successful applications of statistics and operations research: credit scoring.

Credit scoring is the set of predictive models and their underlying techniques that aid financial institutions in the granting of credits. These techniques decide who will get credit, how much credit they should get, and what further strategies will enhance the profitability of the borrowers to the lenders. Credit scoring techniques assess the risk in lending to a particular client. They do not identify "good" or "bad" (negative behaviour is expected, e.g. default) applications on an individual basis, but they forecast probability, that an applicant with any given score will be "good" or "bad". These probabilities or scores, along with other business considerations such as expected approval rates, profit, churn, and losses, are then used as a basis for decision making.

Several methods connected to credit scoring were introduced during last six decades. The most known and widely used are logistic regression, classification trees, linear programming approach and neural networks.

Methodology of credit scoring models and some measures of their quality were discussed in surveys like Hand and Henley (1997), Thomas (2000) or Crook et al. (2007). Even if ten years ago the list of books devoted to the issue of credit scoring was not some large, the situation has improved in the last decade. Particularly it concerns Anderson (2007), Crook et al. (2007), Siddiqi (2006), Thomas et al. (2002) and Thomas (2009).

The aim of this paper is to give an overview of widely used techniques of assessment of credit scoring models quality, to discuss their properties and to extend some known results. We review widely used quality indexes, their properties and relationships. Main part of the paper is devoted to Lift. Curve of ideal Lift is defined, Lift ratio is proposed as analogy to Gini index. Integrated Relative Lift is defined and discussed.

## 2 Measuring the quality

We can consider two basic types of quality indexes. First, indexes based on cumulative distribution function like Kolmogorov-Smirnov statistics, Gini index and Lift. The second, indexes based on likelihood density function like Mean difference (Mahalanobis distance) and Informational statistics. For further available measures and appropriate remarks see Wilkie (2004), Giudici (2003) or Siddiqi (2006).

Assume that realization  $s \in \mathbb{R}$  of random variable  $S$  (score) is available for each client and put the following markings.

$$D = \begin{cases} 1, & \text{client is good} \\ 0, & \text{otherwise} \end{cases}$$

Distribution functions, respectively their empirical forms, of scores of good (bad) clients are given by

$$F_{n.GOOD}(a) = \frac{1}{n} \sum_{i=1}^N I(s_i \leq a \wedge D = 1)$$

$$F_{m.BAD}(a) = \frac{1}{m} \sum_{i=1}^N I(s_i \leq a \wedge D = 0), \quad a \in [L, H],$$

where  $s_i$  is score of  $i$ -th client,  $n$  is number of good,  $m$  is number of bad clients.  $L$  is the minimum value of given score,  $H$  is the maximum value. Empirical distribution function of scores of all clients is given by

$$F_{N.ALL}(a) = \frac{1}{N} \sum_{i=1}^N I(s_i \leq a), \quad a \in [L, H],$$

where  $N = n + m$  is number of all clients. The proportion of bad (good) clients we denote by

$$p_B = \frac{m}{n + m}, \quad p_G = \frac{n}{n + m}.$$

An often-used characteristic in describing the quality of the model (scoring function) is Kolmogorov-Smirnov statistics (K-S or KS). It is defined as

$$KS = \max_{a \in [L, H]} |F_{m.BAD}(a) - F_{n.GOOD}(a)|.$$

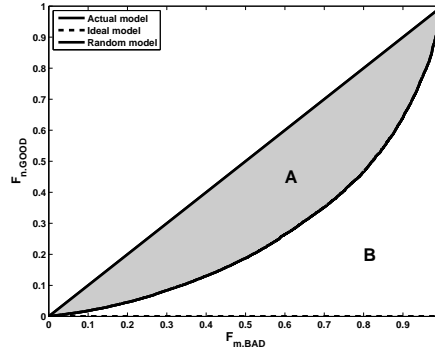
It takes values from 0 to 1. Value 0 corresponds to random model, value 1 corresponds to ideal model. The higher KS, the better scoring model.

The Lorenz curve (LC), sometimes called ROC curve (Receiver Operating Characteristic curve), can also be successfully used to show the discriminatory power of scoring function, i.e. the ability to identify good and bad clients. The curve is given parametrically by

$$\begin{aligned} x &= F_{m.BAD}(a) \\ y &= F_{n.GOOD}(a), \quad a \in [L, H]. \end{aligned}$$

In connection to LC we consider the next quality measure, Gini index. This index describes a global quality of scoring model. It takes values from 0 to 1 (it can take negative values for contrariwise models). The ideal model, i.e. scoring function that perfectly separate good and bad clients, has the Gini index equal to 1. On the other hand, model that assigns a random score to the client has this index equal to 0. It can be shown that Gini index is greater than or equal to KS for any scoring model. Using Figure 1 it can be defined as following

$$Gini = \frac{A}{A + B} = 2A.$$



**Fig. 1.** Lorenz curve, Gini index. Source: own work.

Using previous markings, computational formula of Gini index is given by

$$Gini = 1 - \sum_{k=2}^{n+m} [(F_{m.BAD_k} - F_{m.BAD_{k-1}}) (F_{n.GOOD_k} - F_{n.GOOD_{k-1}})],$$

where  $F_{m.BAD_k}$  ( $F_{n.GOOD_k}$ ) is  $k$ -th vector value of empirical distribution function of bad (good) clients. For further details see Anderson (2007) or Xu (2003). The Gini index is a special case of Somers' D (Somers (1962)), which is an ordinal association measure. It can be found in Thomas (2009), that Somers' D assessing performance of given credit scoring model, denoted as  $D_S$ , one can calculate as

$$D_S = \frac{\sum_i g_i \sum_{j < i} b_j - \sum_i g_i \sum_{j > i} b_j}{n m}$$

where  $g_i$  ( $b_j$ ) is number of goods (bads) in  $i$ -th interval of scores.

In connection to the Gini index, c-statistics (Siddiqi 2006) is defined as

$$c - stat = \frac{1 + Gini}{2}.$$

It represents the likelihood that randomly selected good client has higher score than randomly selected bad client, i.e.

$$c - stat = P(s_1 \geq s_2 | D_1 = 1 \wedge D_2 = 0).$$

It takes values from 0.5, for random model, to 1, for ideal model. Another name for c-statistics can be found in literature. It is Harrell's c, which is a reparameterization of Somers' D (Newson 2006).

## 2.1 Lift.

Another possible indicator of the quality of scoring model is Lift, which says, how many times, at a given level of rejection, is the scoring model better than random selection (random model). More precisely, the ratio indicates the proportion of bad clients with less than a score  $a$ ,  $a \in [L, H]$ , to the proportion of bad clients in the general population. Formally, it can be expressed by

$$Lift(a) = \frac{BadRate(a)}{BadRate} = \frac{\frac{\sum_{i=1}^N I(s_i \leq a \wedge D=0)}{\sum_{i=1}^N I(s_i \leq a)}}{\frac{\sum_{i=1}^N I(D=0)}} = \frac{\frac{\sum_{i=1}^N I(s_i \leq a \wedge D=0)}{\sum_{i=1}^N I(s_i \leq a)}}{\frac{\sum_{i=1}^N I(D=0 \vee D=1)}} = \frac{\frac{\sum_{i=1}^N I(s_i \leq a \wedge D=0)}{\sum_{i=1}^N I(s_i \leq a)}}{\frac{m}{N}}$$

It can be easily verified that the Lift can be equivalently expressed as

$$Lift(a) = \frac{F_{m.BAD}(a)}{F_{N.ALL}(a)}, \quad a \in [L, H].$$

Now we would like to discuss the form of the Lift function for the case of ideal model. It means the model for which sets of output scores of bad and



good clients are disjoint. So there exists a cut-off point  $c$ , for which

$$P(S \leq a) = \begin{cases} P(S \leq a \wedge D = 0), & a \leq c \\ P(D = 0) + P(S \leq a \wedge D = 1), & a > c. \end{cases}$$

Thus we can derive the form for the Lift function

$$Lift_{ideal}(a) = \begin{cases} \frac{1}{p_B}, & a \leq c \\ \frac{1}{F_{N.ALL}(a)}, & a > c. \end{cases}$$

In practice, Lift is computed corresponding to 10%, 20%, ..., 100% of clients with the worst score (see Coppock (2002)). In context with this approach we define

$$QLift(q) = \frac{F_{m.BAD}(F_{N.ALL}^{-1}(q))}{F_{N.ALL}(F_{N.ALL}^{-1}(q))} = \frac{1}{q} F_{m.BAD}(F_{N.ALL}^{-1}(q)), \quad q \in (0, 1]$$

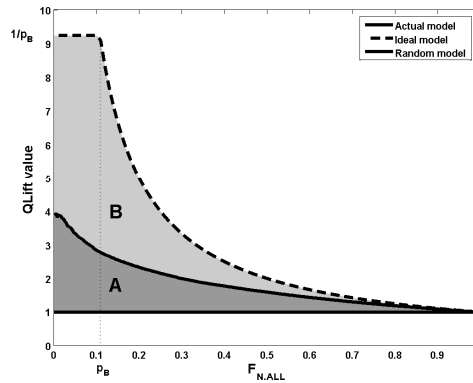
where  $q$  represents the score level of 100 $q$ % of the worst scores and  $F_{N.ALL}^{-1}(q)$  can be computed as

$$F_{N.ALL}^{-1}(q) = \min\{a \in [L, H], F_{N.ALL}(a) \geq q\}.$$

It can be easily shown that Lift function for ideal model is now

$$QLift_{ideal}(q) = \begin{cases} \frac{1}{p_B}, & q \in (0, p_B] \\ \frac{1}{q}, & q \in (p_B, 1]. \end{cases}$$

The following Figure 2 gives an example of Lift function for ideal, random and actual models.



**Fig. 2.** QLift function, Lift Ratio. Source: own work.

Using previous Figure 2, we define Lift Ratio as analogy to Gini index.

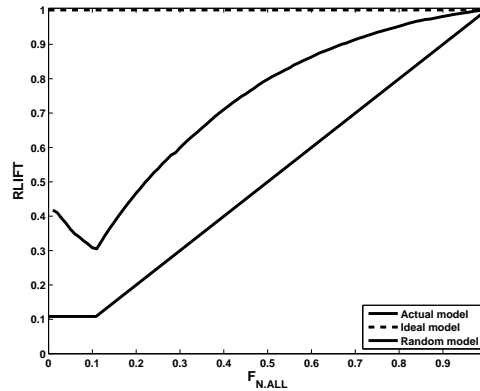
$$LR = \frac{A}{A+B} = \frac{\int_0^1 QLift(q) dq - 1}{\int_0^1 QLift_{ideal}(q) dq - 1}$$

It is obvious that it is global measure of model's quality and that it takes values from 0 to 1. Value 0 corresponds to random model, value 1 match to ideal model. Meaning of this index is quite simple. The higher, the better. Important feature is that Lift Ratio allows us to fairly compare two models developed on different data samples, which is not possible with Lift.

Since Lift Ratio compares areas under Lift function for actual and ideal models, next concept is focused on comparison of Lift functions themselves. We define Relative Lift function by

$$RLift(q) = \frac{QLift(q)}{QLift_{ideal}(q)}, \quad q \in (0, 1].$$

An example of this function is presented in Figure 3. Definition domain of the function is  $(0, 1]$ , range is a subinterval of  $[0, 1]$ . The graph starts in point  $[q_{min}, p_B \cdot Lift(q_{min})]$ , where  $q_{min}$  is a positive number near to zero. Then it falls down to a local minima in point  $[p_B, p_B \cdot Lift(p_B)]$  and then rises up to point  $[1, 1]$ . It is obvious that graph of Relative Lift function for better model is closer to top line, which represents the function for ideal model.

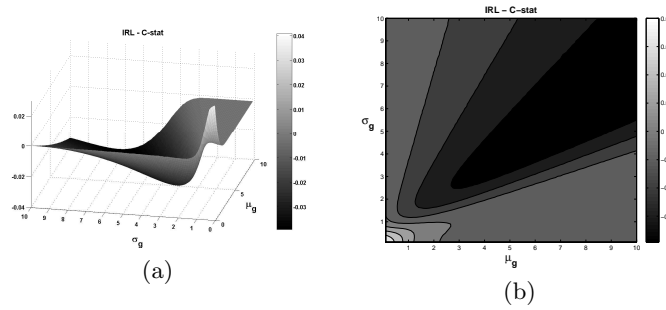


**Fig. 3.** Relative Lift function. Source: own work.

Now it is natural to ask what we obtain when integrate Relative Lift function. We define Integrated Relative Lift (IRL) by

$$IRL = \int_0^1 RLift(q) dq.$$

It takes values from  $0.5 + \frac{p_B^2}{2}$ , for random model, to 1, for ideal model. Again it holds: the higher, the better. This global measure of scoring model's quality has interesting connection to c-statistics. We made a simulation with scores generated from normal distribution. Scores of bad clients had mean equal to 0 and variance equal to 1. Scores of good clients had mean and variance from 0.1 to 10 with step equal 0.1. Number of samples and sample size was 1000,  $p_B$  was equal to 0.1. IRL and c-statistics were computed for each sample and each value of mean and variance of good client's scores. Finally, means of IRL and c-statistics were computed. Results are presented in Figure 4. Part (b) represents countour plot of figure in part (a).



**Fig. 4.** Difference of IRL and c-stat (a) and its contour plot (b). Source: own work.

The simulation shows that IRL and c-statistics are approximately equal in case that variances of good and bad clients are equal. Furthermore it shows that they significantly differ when variances are different and ratio of mean and variance of good clients is near to 1.

### 3 Conclusions

All referred quantitative indexes can be successfully used to measure the quality of different credit scoring models or their individual parts, i.e. predictors entering the models. They can be used as benchmarks for comparison of several proposed models at the time of model development as well as monitoring tools after deployment into rel business.

Main part of the paper was devoted to Lift. Formulas for Lift in basic and quantile form were stated as well as their forms for ideal models. It is obvious that we need to have the best performance of given scoring model near by expected cutoff value. From this point of view, QLift seems to be the best choice for our purpose. According to its definition it is easy to read how many times is actual model better than random model at given level of rejection.

Lift ratio was stated. Important feature is that Lift Ratio allows us to fairly compare two models developed on different data samples, which is not possible with Lift. Furthermore it was proposed Relative Lift function, which shows ratio of QLift of actual and ideal model. Finally, Integrated Relative Lift was stated. Connection to c-statistics was presented via simulation on normally distributed scores. This simulation shows that IRL and c-statistics are approximately equal in case that variances of good and bad clients are equal.

## 4 Acknowledgments

The research was supported by our department and by The Jaroslav Hájek center for theoretical and applied statistics (grant No. LC 06024).

## References

- ANDERSON, R. (2007): *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, Oxford.
- CROOK, J.N., EDELMAN, D.B., THOMAS, L.C. (2007): Recent developments in consumer credit risk assessment. *European Journal of Operational Research* 183 (3), 1447-1465.
- COPPOCK, D.S. (2002): Why Lift?. *DM Review Online*, [www.dmreview.com/news/5329-1.html](http://www.dmreview.com/news/5329-1.html). Accessed on 1 December 2009.
- GIUDICI, P. (2003): *Applied Data Mining: statistical methods for business and industry*. Wiley, Chichester.
- HAND, D.J. and HENLEY, W.E. (1997): Statistical Classification Methods in Consumer Credit Scoring: a review. *Journal. of the Royal Statistical Society, Series A*. 160 (3), 523-541.
- NEWSON R. (2006): Confidence intervals for rank statistics: Somers' D and extensions. *The Stata Journal* 6 (3), 309-334.
- SIDDIQI, N. (2006): *Credit Risk Scorecards: developing and implementing intelligent credit scoring*. Wiley, New Jersey.
- SOMERS R. H. (1962): A new asymmetric measure of association for ordinal variables. *American Sociological Review* 27, 799-811.
- THOMAS, L.C. (2000): A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting* 16 (2), 149-172.
- THOMAS, L.C. (2009): *Consumer Credit Models: Pricing, Profit, and Portfolio*. Oxford University Press, Oxford.
- THOMAS, L.C., EDELMAN, D.B., CROOK, J.N. (2002): *Credit Scoring and Its Applications*. SIAM Monographs on Mathematical Modeling and Computation, Philadelphia.
- WILKIE, A.D. (2004): Measures for comparing scoring systems, In: Thomas, L.C., Edelman, D.B., Crook, J.N. (Eds.): *Readings in Credit Scoring*. Oxford University Press, Oxford, 51-62.
- XU, K. (2003): How has the literature on Gini's index evolved in past 80 years?. *economics.dal.ca/RePEc/dal/wparch/howgini.pdf*. Accessed on 1 December 2009.

# Clustering with Mixed Type Variables and Determination of Cluster Numbers

Hana Řezanková<sup>1</sup>, Dušan Húsek<sup>2</sup>, and Tomáš Löster<sup>1</sup>

<sup>1</sup> University of Economics, Prague, nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic, {*hana.rezankova|tomas.loster*}@vse.cz

<sup>2</sup> ICS, AS CR, Pod vodárenskou věží 2, 182 07 Praha 8, Czech Republic, *dusan.husek@cs.cas.cz*

**Abstract.** The main applications of cluster analysis methods concern clustering objects characterized by quantitative variables. In the paper, the evaluation criteria for the case of mixed type variables are dealt with. The criteria based on variability measures for nominal and quantitative variables are described. A combination of variance and entropy or Gini's coefficient of mutability is used. The coefficients for determination of cluster numbers based on variability measures for mixed type variables are applied to a real-data to show viability of our approach.

**Keywords:** cluster analysis, entropy, Gini's coefficient of mutability, cluster number determination, Schwarz's Bayesian information criterion

## 1 Introduction

Cluster analysis is an important tool in many areas. Different cluster analysis methods have been proposed, including special techniques for mixed data types (Huang, 1997; He et al., 2005).

Further, many coefficients for evaluation of clustering and determination of cluster numbers have been proposed. However, these coefficients are proposed mainly for cases when objects are characterized by quantitative variables (Gordon, 1999; Gan et al., 2007; Kogan, 2007).

In this paper, the principles of determination of cluster numbers for the case of quantitative data are extended to the data files with mixed type variables (nominal and quantitative). Two-step cluster analysis is applied to questionnaire survey data several times with numbers of clusters as a parameter. Then proposed coefficients are used for optimal assignment of respondents to clusters.

## 2 Basic methods for clustering with mixed type variables and its evaluation

The simplest way to cluster objects characterized by mixed type variables is to create a proximity matrix based on Gower's similarity coefficient (Gower,

1971), and apply hierarchical cluster analysis. However, this coefficient has not been implemented in the commercial statistical packages yet.

Another measure of the relationship between two objects (and also between two clusters) is the *log-likelihood distance measure*. Its implementation in the software products is linked with *two-step cluster analysis* in the SPSS system (now IBM SPSS Statistics). This method has been designed for clustering of a large number of objects and it is based on the BIRCH algorithm (Zhang et al., 1996). The log-likelihood distance measure is determined for data files with combinations of quantitative and nominal variables. Two objects are the most similar, if the cluster composed of them has the smallest variability. The entropy is applied to nominal variables as a variability measure.

Furthermore, separate clustering of the objects characterized by categorical variables, see Řezanková (2009) for the survey, and clustering of the objects characterized by quantitative variables can be applied to be followed by a cluster ensemble technique (Punera and Ghosh, 2007). Moreover, it is possible to transform nominal or ordinal variables to the sets of binary variables and use the techniques of cluster analysis for binary or quantitative data. This variant is not considered in this paper.

For clustering results evaluation, the variability of the all clusters can be calculated. For  $k$  clusters we can write

$$H(k) = \sum_{g=1}^k n_g \left( \sum_{l=1}^{m_1} \frac{1}{2} \ln(s_l^2 + s_{gl}^2) - \sum_{l=1}^{m_2} \sum_{u=1}^{K_l} \left( \frac{n_{glu}}{n_g} \ln \frac{n_{glu}}{n_g} \right) \right), \quad (1)$$

where  $k$  is the number of cluster,  $m_1$  is the number of quantitative (continuous) variables,  $m_2$  is the number of nominal variables,  $s_l^2$  is a sample variance of the  $l$ th variable, and  $s_{gl}^2$  is a sample variance of the  $l$ th variable in the  $g$ th cluster (quantitative variables are standardized first),  $K_l$  is the number of categories of the  $l$ th variable,  $n_{glu}$  represents the frequency of the  $u$ th category of the  $l$ th variable in the  $g$ th cluster, and  $n_g$  is the number of objects in the  $g$ th cluster.

For determination of cluster numbers, the information criteria can be used. In the SPSS, both *Schwarz's Bayesian Information Criterion* (BIC) and *Akaike's Information Criterion* (AIC) are implemented. In the case when variables are of a mixed type, the former is calculated as

$$I_{BIC} = 2H(k) + k \left( 2m_1 + \sum_{l=1}^{m_2} (K_l - 1) \right) \ln(n), \quad (2)$$

where  $n$  is the number of objects.

The latter is calculated as

$$I_{AIC} = 2H(k) + 2k \left( 2m_1 + \sum_{l=1}^{m_2} (K_l - 1) \right). \quad (3)$$

First, the values of a criterion are calculated for each number of clusters within a specified range. The initial estimate for the number of clusters is determined as the minimum of these values. Then, this initial estimate is modified by finding the largest increase in the distance between the two closest clusters in each hierarchical clustering stage. According to our experience, the BIC criterion often provides the local minimum but the AIC criterion does not (usually the latest value is minimum). For this reason, we will consider only the BIC criterion in the following text.

### 3 Evaluation criteria of clustering based on variability measures

Quality of partitioning of objects to clusters can be evaluated in different ways. The aim of the disjunctive clustering is to create clusters with a small variability of variables within them. In the case of quantitative variables, the variance is applied as a variability measure. The variability within clusters, between clusters and the total variability of all variables are investigated. It is an analogy of one factor multivariate analysis of variance (MANOVA) where a new variable containing a label of clusters to which the object was assigned is a factor.

Similarly, for nominal variables, the specific coefficients of variability (mutability) can be applied. Besides entropy, it is *Gini's coefficient* (measure of mutability, see Gini, 1912).

For mixed type variables, the combination of entropy and variance can be used, see (1). Using Gini's coefficient of mutability, we can write

$$G(k) = \sum_{g=1}^k n_g \left( \sum_{l=1}^{m_1} \frac{1}{2} \ln(s_l^2 + s_{gl}^2) + \sum_{l=1}^{m_2} \left( 1 - \sum_{u=1}^{K_l} \left( \frac{n_{glu}}{n_g} \right)^2 \right) \right). \quad (4)$$

If an analogy of MANOVA is applied, several coefficients have been proposed for the quantitative data. All of these coefficients can be modified for mixed type variables. In the following section, we will present the applications of formulas (1) and (4).

We can distinguish a comparison of results of different clustering methods and a determination of a suitable number of clusters. In the latter, the local optimum from the specified range is usually searched for.

For the method comparison, we suppose that the partitioning results are compared to the same number of clusters. In this case, the total variability

(mutability or entropy) in the data set (the sum of the variability of all the variables) is the same, and only the comparison of within-cluster variability is a sufficient tool for evaluation.

If we consider an analogy with the R-squared coefficient from ANOVA, the ratio of the between-cluster variability and the total variability is calculated. We obtain the values from the interval from 0 to 1 when the greater value indicates the better clustering.

The modification of R-square with the use of the entropy as a variability measure for one dependent nominal variable is called the *uncertainty coefficient* (Press and Flannery, 1988). Using  $H$  defined in (1), we can define *uncertainty index*

$$I_U(k) = \frac{H(1) - H(k)}{H(1)}, \quad (5)$$

where  $H(1)$  is the variability of the whole data set ( $k = 1$ ).

The modification of R-square with the use of *Gini's coefficient* as a variability measure for one dependent nominal variable is called *Goodman and Kruskal's tau* (Goodman and Kruskal, 1954). Using  $G$  defined in (4), we can define *tau index*

$$I_\tau(k) = \frac{G(1) - G(k)}{G(1)}. \quad (6)$$

It is obvious that partitioning to a greater number of clusters has a lower within-cluster variability and the values of coefficients defined in (5) and (6) are higher. To avoid this, some special indices were proposed for determination of cluster numbers. Let us mention semipartial R-squared index (*SPRSQ*), root-mean-square standard deviation (*RMSSTD*) index and  $I_{CHF_U}$ , respective  $I_{CHF_\tau}$  indices. The *SPRSQ* index for  $k$  clusters is expressed as

$$I_{SPRSQ}(k) = I_{RSQ}(k+1) - I_{RSQ}(k). \quad (7)$$

For variables of mixed types, we can write  $I_U$  or  $I_\tau$ , instead of  $I_{RSQ}$ . The smallest value indicates the best partitioning of the objects into clusters.

The *RMSSTD* index (Sharma, 1995), is the root mean square standard deviation of the variables within each cluster. First, the within-group sum of squares of each cluster is computed. Then, it is normalized by the product of the number of objects in the cluster and the number of variables. However, in the case of mixed type variables, there is a problem with their number (in the presented example with the increasing number of clusters the values of this index decrease).

The pseudo  $F$  or *CHF* index (Calinski and Harabasz, 1974), is the ratio of between-cluster variance to within cluster variance. Two analogous variants of this index for mixed type variables are



$$I_{CHFU}(k) = \frac{(n-k)(H(1) - H(k))}{(k-1)H(k)}, \quad (8)$$

or

$$I_{CHF\tau}(k) = \frac{(n-k)(G(1) - G(k))}{(k-1)G(k)}. \quad (9)$$

Peaks in the plot of index values for different numbers of clusters are indicators of greater cluster separation.

Some other coefficients for determination of cluster numbers based on comparison of compactness of clusters and their separation could be used. We can consider the modification of the *Davies and Bouldin (DB) index* (Davies and Bouldin, 1979), in the form

$$I'_{DB}(k) = \frac{\sum_{h=1}^k \max_{h', h' \neq h} \left\{ \frac{H_h + H_{h'}}{H_{hh'} - (H_h + H_{h'})} \right\}}{k}, \quad (10)$$

where  $H_h$  is defined as

$$H_h = n_h \left( \sum_{l=1}^{m_1} \frac{1}{2} \ln(s_l^2 + s_{hl}^2) - \sum_{l=1}^{m_2} \sum_{u=1}^{K_l} \left( \frac{n_{hlu}}{n_h} \ln \frac{n_{hlu}}{n_h} \right) \right). \quad (11)$$

Similarly  $H_{h'}$  and  $H_{h,h'}$  represent variability of the cluster created by joining the  $h$ th and  $h'$ th clusters. The value of the expression  $(H_{hh'} - H_h - H_{h'})$  represents the distance of the  $h$ th and  $h'$ th clusters in the two-step cluster analysis in SPSS. The best partitioning of objects to cluster is in the case when the value of this index is minimal.

## 4 Example

We analyzed the data file created on the basis of the answers provided by 50 participants in a chemistry seminar (<http://statistika.vse.cz/literatura/KSICHT.sav>). We chose the nominal variables with the frequencies of at least 10 for each category. For this reason, some new variables were created both by merging variables and by joining categories. The analyzed data set contains seven categorical (binary, nominal and ordinal) variables and one quantitative variable (for all pairs, the variables are independent at 5% significance level when chi-square test for independence for categorical variables, and analysis of variance for testing independence of a quantitative variable on a categorical variable are used).

There are four binary variables (interest in research news, participant in either mathematics or physics Olympiad, participant in biology Olympiad, and participant in some other Olympiad), two variables with three categories (where the respondent got information about the seminar with categories

of school, educational camp, and other place, and class of the secondary school with categories of first and second, third, and fourth), the dichotomous variable of sex (there were 25 males and 25 females), and the quantitative variable of the number of solved tasks.

We used the SPSS system for analyses. We applied two-step cluster analysis with the log-likelihood distance measure. In our experiments, we clustered participants in the chemistry seminar to two, three and four clusters. The sizes of the individual clusters are in Table 1.

**Table 1.** Sizes of clusters (numbers of objects) obtained by two-step cluster analysis

Order of cluster	Number of clusters		
	2	3	4
1	38	20	10
2	12	18	11
3		12	17
4			12

We calculated the measures of within-cluster variability (based both on the entropy and on Gini's coefficient), differences of these values (the difference expresses the distance of two closest clusters), uncertainty index  $I_U$ , tau index  $I_\tau$ , semipartial  $I_U$ , semipartial  $I_\tau$ ,  $I_{CHF\tau}$ , respective  $I_{CHF_U}$  index, BIC criterion based on the entropy, and analogous  $BGC$  criterion based on Gini's coefficient. The obtained results are shown in Table 2 (measures based on the entropy) and 3 (measures based on Gini's coefficient). According to the maxima of variability differences and  $I_{CHF_U}$ , respective  $I_{CHF\tau}$  indices, and the minima of semipartial  $I_U$ , semipartial  $I_\tau$ , and modified BIC criterion, respective analogous  $BGC$  criterion (the values marked by the asterisk), 3 clusters were found as optimal in all these cases.

In comparison with the case when the respondents were characterized only by seven categorical variables, 3 clusters were found as optimal when semipartial  $I_\tau$  and modified  $F$  indices  $I_{CHF_U}$  based on the entropy and  $I_{CHF\tau}$  based on Gini's coefficient were applied. By means of variability differences, the maxima were found for 2 clusters (from 2 to 6 clusters). According to the semipartial  $I_U$ , 4 clusters were found as optimal (from 2 to 5 clusters), as well as BIC criterion based on the entropy.

## 5 Conclusion

In the paper, we proposed the evaluation measures for clustering with mixed type variables based on the combination of variance and entropy or Gini's coefficient. To determine the cluster numbers, the following measures can be applied: variability difference (based both on the entropy and on Gini's

**Table 2.** Measures based on the entropy

Measure	Number of clusters			
	1	2	3	4
Within-cluster variability	273.92	241.17	206.39	186.51
Variability difference	—	32.75	34.78*	19.88
Coefficient $I_U$	0	0.12	0.25	0.32
Semipartial $I_U$	0.12	0.13	0.07*	—
$I_{CHF_U}$	0	6.52	7.69*	7.19
$I_{BIC}$	590.85	568.41	541.88*	545.15

**Table 3.** Measures based on Gini's coefficient

Measure	Number of clusters			
	1	2	3	4
Within-cluster variability	185.41	162.57	137.83	127.86
Variability difference	—	22.84	24.74*	9.97
Coefficient $I_\tau$	0	0.12	0.26	0.31
Semipartial $I_\tau$	0.12	0.13	0.05*	—
$I_{CHF_\tau}$	0	6.74	8.11*	6.90
$I_{BGC}$	413.85	411.20	404.75*	427.84

coefficient), semipartial uncertainty index, semipartial tau index,  $I_{CHF_U}$  index, respective  $I_{CHF_\tau}$  index, and  $BIC$  criterion based on the entropy and analogous  $BGC$  criterion based on Gini's coefficient. One can find either the local maximum (in the case of variability difference or both indices  $I_{CHF_U}$  and  $I_{CHF_\tau}$ ) but local minimum (in the other cases). As the results using the mentioned measures can differ, it is preferable to make calculation using several of these measures and choose such a number of clusters for which the values of most of the measures are favorable.

Another possibility for clustering evaluation is the investigation of dissimilarity of the objects in the pairs both from the same clusters and from the different clusters, dissimilarity of the object and "its" centroid (centroid of the cluster to which the object belongs) and dissimilarity of centroids. Several coefficients based on these dissimilarities have been proposed for quantitative data (Gan, 2007). In the case of mixed type variables, the centroid can contain the modal values for nominal variables and the medians for ordinal and quantitative variables. Problem can arise if a variable has two or more modal categories. Moreover, these approaches are more time-consuming both for programming and for calculations.

**Acknowledgement.** This work was supported by projects AV0Z10300504, GACR P202/10/0262, 205/09/1079, and IGA VSE F4/3/2010.

## References

- CALINSKI, T., HABARASZ, J. (1974): A dendrite method for cluster analysis. *Communications in Statistics* 3, 1-27.
- DAVIES, D. L. and BOULDIN, D. W. (1979): A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1 (4), 224-227.
- GAN, G., MA, C. and WU, J. (2007): *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM, Philadelphia.
- GINI, C. W. (1912): *Variability and Mutability*. Contribution to the study of statistical distributions and relations. Studi Economico-Giuridici della R. Università de Cagliari. Reviewed in: LIGHT R. J. and MARGOLIN B. H.: An Analysis of Variance for Categorical Data. J. American Statistical Association 66, 534-544, 1971.
- GOODMAN, L. A. and KRUSKAL, W. H. (1954): Measures of association for crossclassification. *Journal of the American Statistical Association* 49, 732-764.
- GORDON, A. D. (1999): *Classification, 2nd Edition*. Chapman & Hall/CRC, Boca Raton.
- GOWER, J. C. (1971): A general coefficient of similarity and some of its properties. *Biometrics* 27, 857-874.
- HE, Z., XU, X. and DENG, S. (2005): Clustering mixed numeric and categorical data: a cluster ensemble approach. <http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0509011>, 14p.
- HUANG, Z. (1997): Clustering large data sets with mixed numeric and categorical values. *Proc. of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*. World Scientific, Singapore.
- KOGAN, J. (2007): *Introduction to Clustering Large and High-Dimensional Data*. Cambridge University Press, New York.
- PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. and VETTERLING, W. T. (1988): *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge, p. 634.
- PUNERA, K. and GHOSH, J. (2007): Soft cluster ensembles. In: J.V. Oliveira and W. Pedrycz (Eds.): *Advances in Fuzzy Clustering and its Applications*. John Wiley & Sons, Chichester, 69-91.
- ŘEZANKOVÁ, H. (2009): Cluster analysis and categorical data. *Statistika* 89, 216-232.
- SHARMA, S. (1995): *Applied multivariate techniques*. John Wiley & Sons, Inc., New York.
- ZHANG, T., RAMAKRISHNAN, R. and LIVNY, M. (1996): BIRCH: An efficient data clustering method for very large databases. *ACM SIGMOD Record* 25, 103-114.

# A General Strategy for Determining First-Passage-Time Densities Based on the First-Passage-Time Location Function

Patricia Román-Román, Juan José Serrano-Pérez and Francisco Torres-Ruiz

Departamento de Estadística e Investigación Operativa (Universidad de Granada)  
Avda Fuentenueva s/n, 18071 Granada, Spain, {proman,jjserra,fdeasis}@ugr.es

**Abstract.** This paper presents a general strategy for the efficient application of numerical schemes for solving Volterra integral equations which have as solution first-passage-time density functions associated to certain stochastic processes. Such strategy is based on the information provided by the First-Passage-Time Location function about the location of the variation range of the first-passage-time variable, and it is valid for general type situations that expand on the particular cases considered in Román et al. (2008). In addition, numerical applications are shown to prove the validity of the strategy as well as its computational advantages.

**Keywords:** diffusion processes, first-passage-times, volterra integral equations, first-passage-time location function

## 1 Introduction

Let be  $\{X(t); t_0 \leq t \leq T\}$  a stochastic process defined on a real interval  $I$ . For this process we consider the first-passage-time (f.p.t.) problem of the process through the boundary  $S(t)$ , provided  $X(t_0) = x_0$ . This problem is defined by the variable

$$T_{S(t),x_0} = \begin{cases} \inf_{t \geq t_0} \{t : X(t) > S(t) \mid X(t_0) = x_0\}, & x_0 < S(t_0) \\ \inf_{t \geq t_0} \{t : X(t) < S(t) \mid X(t_0) = x_0\}, & x_0 > S(t_0), \end{cases} \quad (1)$$

and its solution requires determining the distribution of this variable. Nevertheless, there are no general procedures for this purpose, but instead a wide variety of studies have been carried out for specific types of processes. In particular, if the process considered is a diffusion process with infinitesimal moments  $A_1(x, t)$  and  $A_2(x, t)$ , and if  $S(t)$  is a continuous function, the density function of  $T_{S(t),x_0}$ ,  $g(S(t), t \mid x_0, t_0)$ , is the solution to the Volterra integral equation of the second kind (see Ricciardi et al. (1999) and references therein).

$$g(S(t), t|x_0, t_0) = \rho \left\{ -2\Psi(S(t), t|x_0, t_0) + 2 \int_{t_0}^t g(S(\tau), \tau|x_0, t_0) \Psi(S(t), t|S(\tau), \tau) d\tau \right\}, \quad (2)$$

where  $\rho = \text{Sgn}(S(t_0) - x_0)$  and

$$\begin{aligned} \Psi(S(t), t|y, \tau) = & \frac{1}{2} f(S(t), t|y, \tau) \left[ S'(t) - A_1(S(t), t) \right. \\ & \left. + \frac{3}{4} \frac{\partial A_2(x, t)}{\partial x} \Big|_{x=S(t)} \right] + \frac{1}{2} A_2(S(t), t) \frac{\partial f(x, t|y, \tau)}{\partial x} \Big|_{x=S(t)}, \end{aligned}$$

being  $f(x, t|y, s)$  the transition probability density function of the process.

Nevertheless, and apart from some particular processes and boundaries, closed-form solutions for the integral equation are not available. For this reason, in the cases without explicit solutions, numerical procedures are needed. The most usual methods are based on numerical quadrature procedures, as proposed by Buonocore et al. (1987) on the basis of the composite trapezoid method. Specifically,

$$\begin{aligned} g(S(t_0 + h), t_0 + h|x_0, t_0) &= -2\rho\Psi(S(t_0 + h), t_0 + h|x_0, t_0), \\ g(S(t_0 + kh), t_0 + kh|x_0, t_0) &= -2\rho\Psi(S(t_0 + kh), t_0 + kh|x_0, t_0) \\ &\quad + 2\rho h \sum_{j=1}^{k-1} g(S(t_0 + jh), t_0 + jh|x_0, t_0) \\ &\quad \times \Psi(S(t_0 + kh), t_0 + kh|S(t_0 + jh), t_0 + jh) \\ k &= 2, 3, \dots \end{aligned} \quad (3)$$

In the application of numerical procedures of type (3) for the resolution of the Volterra integral equation (2), several problems may arise (see Román et al. (2008) for a more detailed analysis). Thus, considering an inadequate integration step for the variation range of the f.p.t. variable may lead to bad approximations to its density function or, in the case of good approximations, to an unnecessarily high computational cost. This latter situation can also be associated to a poor choice of the initial instant for the application of the algorithm, or to the use of stopping rules unable to detect tails in the distribution of the variable.

Regarding this sort of problems, Román et al. (2008) introduced the use of the First-Passage-Time Location (FPTL) function, which allows to locate the variation range for the variable under study. Furthermore, for particular behaviors of the FPTL function, a strategy for reaching optimal values for the initial and final instants of application of the algorithm, as well as an

optimal integration step for the resolution of the problems aforementioned, was proposed. Specifically, the following cases were considered

- the FPTL function increases in  $[t_0, T]$ ,
- the FPTL increases up to a certain time and then decreases subsequently,

situations that occur naturally in several contexts, as the study of f.p.t. through constant boundaries in diffusion processes with a strictly increasing trend, in which case its density function shows a single mode.

However, this situation is not usual in most real applications. In fact, considering certain time-dependent boundaries can lead to f.p.t. densities showing several modes, even if the process has a strictly increasing trend. In such cases the FPTL function does not show a behavior as the one considered in Román et al. (2008) (see applications in the present work). This situation can also appear in the study of f.p.t. through constant boundaries when the trend of the process considered swings around the boundary. For example, for modeling economic variables as the G.N.P., considering external variables or exogenous factors can affect the trend of the process under consideration, leading to this kind of behavior (see Gutiérrez et al., 1999).

The present work deals with a general strategy to determine the necessary parameters for an efficient application of numerical procedures of the type (3), associated to any behavior exhibited by the FPTL function. The most general situation is the FPTL function being piecewise monotonous in  $[t_0, T]$ , with a global maximum and possible local maxima. The following section will propose an adequate strategy for such situation. Finally, we will propose some examples of application, showing the validity of that strategy and its computational advantages.

## 2 The FPTL function. Proposed strategy

The FPTL function (see Román et al. 2008) is defined as

$$FPTL(t) = \begin{cases} P[X(t) > S(t) | X(t_0) = x_0] = 1 - F(S(t), t | x_0, t_0) & \text{if } x_0 < S(t_0) \\ P[X(t) < S(t) | X(t_0) = x_0] = F(S(t), t | x_0, t_0) & \text{if } x_0 > S(t_0) \end{cases}$$

for  $t_0 \leq t \leq T$ , being  $F(x, t | y, s)$  the transition probability distribution function of the process.

The strategy proposed in this paper, which extends that considered by Román et al. (2008), is determined by the following steps:

Step 1: Splitting the interval  $[t_0, T]$  in subintervals which are each associated to a maximum of the FPTL function. In order to do this, we must determine the  $t_1 < t_2 < \dots < t_m \in [t_0, T]$  instants, from which the function starts growing. Taking  $t_{m+1} = T$ , subintervals are determined by  $I_i = [t_i, t_{i+1}]$ ,  $i = 0, 1, 2, \dots, m-1$  and  $I_m = [t_m, T]$ . The first subinterval can be ignored since the FPTL function almost equals zero, taking into account that  $FPTL(t_0) =$

0. For each of the following subintervals the FPTL function grows up to a maximum value and then remains constant and/or decreases.

Step 2: For each of the subintervals  $I_i, i = 1, \dots, m$ , found in the first step, find:

- The maximum value of the FPTL function,  $p_{max,i}$ , and the first time instant  $t_{max,i} \in I_i$  in which the FPTL function reaches that maximum, that is,  $FPTL(t_{max,i}) = p_{max,i}$ .
- The first time instant  $t_i^* \in [t_i, t_{max,i}]$  in which the FPTL function is bigger than or equals  $FPTL(t_i) + 10^{-k}(p_{max,i} - FPTL(t_i))$  with a sufficiently large  $k$ , that is, with a significantly bigger value relative to that of the lower end of the subinterval.
- The first instant  $t_{max,i}^- \in [t_i, t_{max,i}]$ , in which the FPTL function is bigger than or equal to

$$p_{max,i}^- = p_{max,i} (1 - 0.05(p_{max,i} - FPTL(t_i))),$$

so that  $p_{max,i}^-$  approaches  $p_{max,i}$  as the magnitudes of  $p_{max,i}$  and  $p_{max,i} - FPTL(t_i)$  decrease.

- The first instant  $t_{max,i}^+ \in [t_{max,i}, t_{i+1}]$ , in which the FPTL function is smaller than or equal to

$$p_{max,i}^+ = \max \{1 - (1 - p_{max,i}^2)^{(1+q)/2}, FPTL(t_{i+1})\},$$

where

$$q = \frac{p_{max,i} - FPTL(t_i)}{p_{max,i}},$$

so that  $p_{max,i}^+$  diverges from  $p_{max,i}$  as the magnitudes of  $p_{max,i}$  and  $p_{max,i} - FPTL(t_i)$  decrease.

However, if in the subinterval  $[t_m, T]$  the FPTL function is strictly increasing, then  $t_{max,m}^+ = t_{max,m}$  coincides with  $T$ .

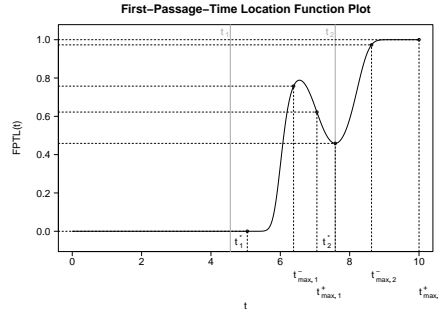
Figure 1 shows every time instant considered for a FPTL function with a general behavior, and Figure 2 shows them for two FPTL functions of the types studied in Román et al (2008). It can be observed that the first case is an extension of the following.

Step 3: With the information provided by the FPTL function, we propose the application of the numerical algorithm (3) in order to determine an approximation,  $g_1$ , of the f.p.t. density  $g$ , by taking

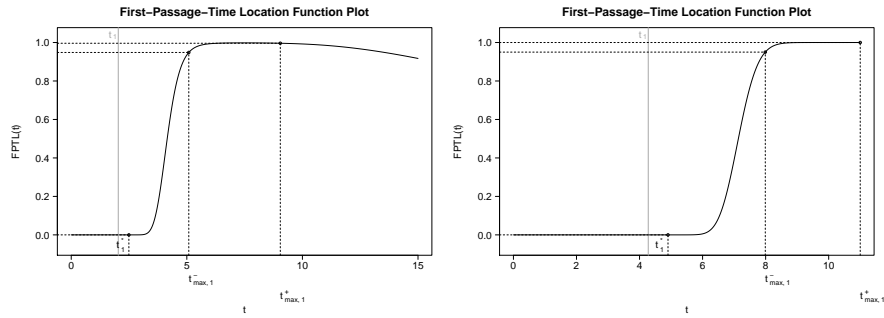
- Initial instant:  $t_1^*$
- Integration step. Two choices could be considered:
  - Fixed integration step

$$h = \frac{\min_{i=1, \dots, m} (t_{max,i}^- - t_i^*)}{n}.$$





**Fig. 1.** FPTL function and the information it provides for a general case.



**Fig. 2.** FPTL function and the information it provides for the particular cases analyzed in Román et al. (2008).

– Variable integration step

$$h_i = \frac{t_{max,i}^- - t_i^*}{n} \text{ in subintervals } [t_i^*, t_{max,i}^+] \subset I_i, \quad i = 1, \dots, m,$$

$$h_i^* = \frac{t_{i+1}^* - t_{max,i}^+}{[np + 0.5]} \text{ in subintervals } [t_{max,i}^+, t_{i+1}^*], \quad i = 1, \dots, m-1.$$

where  $[x]$  is the integer part of  $x$ . In each case,  $n \in \mathbb{N}$  and  $0 < p < 1$  depend on the desired accuracy. For both options, it could be checked, for each  $i = 1, \dots, m$ , if  $g_1(t_{max,i}^+) \simeq 0$ . If so, the application of the numerical algorithm in the  $[t_{max,i}^+, t_{i+1}^*]$  subinterval could be avoided, and then continue from  $t_{i+1}^*$ , considering  $g_1(t) = 0, \forall t \in [t_{max,i}^+, t_{i+1}^*]$ .

- Stopping rule. Check for each  $t_{max,i}^+, i = 1, \dots, m$ , the accumulated value of the integral and stop the application of the algorithm if it is close to one. In any case, stop the application of the algorithm in  $t_{max,m}^+$ .

### 3 Applications

This section shows several applications for which the FPTL function presents a general situation as the one described earlier on this paper.

First, and in order to prove the usefulness of the strategy proposed, the density of the f.p.t. is approximated for the example shown in Figure 4.

Secondly, an application is shown in which, from a computational standpoint, a variable integration step ostensibly improves over a fixed integration step, due to the difference in the amplitude of the corresponding growth subintervals.

### 3.1 Application 1

Let us consider the problem of f.p.t. through the boundary  $S(t) = 7 + 3.2t + 1.4t \sin(1.75t)$  for the lognormal homogeneous process, defined as a diffusion process  $\{X(t); t_0 \leq t \leq T\}$ , taking values in  $\mathbb{R}^+$  and with infinitesimal moments  $A_1(x) = mx$  and  $A_2(x) = \sigma^2 x^2$ , where  $m \in \mathbb{R}$  and  $\sigma > 0$ , in the particular case  $m = 0.48$ ,  $\sigma^2 = 0.0049$ ,  $t_0 = 0$ ,  $T = 10$  and  $P[X(t_0) = 1] = 1$ . For this case, the FPTL function is the one shown in Figure 1, and from the information that it provides the f.p.t. density has been approximated, starting at  $t_1^* = 5.0475$  and using the following variable integration steps (for  $n = 250$  and  $p = 0.2$ ):

$$\begin{aligned} h_1 &= (6.3826 - 5.0475)/n \text{ in } [5.0475, 7.0567), \\ h_1^* &= (7.5848 - 7.0567)/[np + 0.5] \text{ in } [7.0567, 7.5848), \text{ and} \\ h_2 &= (8.6309 - 7.5849)/n \text{ in } [7.5849, 10]. \end{aligned}$$

Figure 3 shows the mean function of the process together with the boundary considered, the FPTL function and the approximation obtained for the f.p.t. density, which verifies  $P(5.0475 < T < 10) > 1 - 10^{-4}$ .

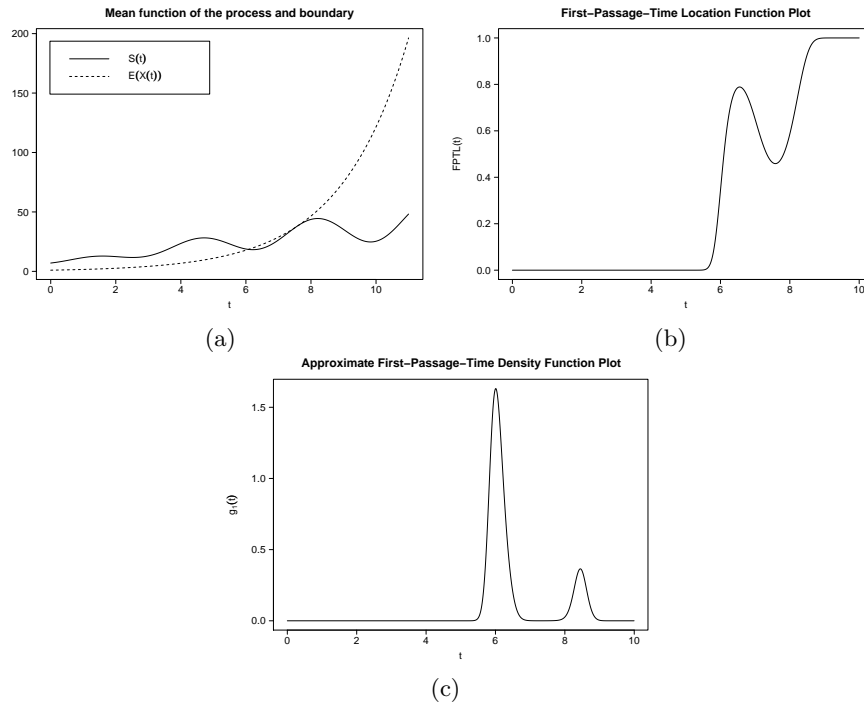
### 3.2 Application 2

Let us now consider the lognormal homogeneous process with  $m = 0.48$ ,  $\sigma^2 = 0.0049$ ,  $t_0 = 0$  and  $T = 18$ . By taking  $P[X(t_0) = 1] = 1$ , we consider the f.p.t. variable through the boundary  $S(t) = 4.5 + 4t^2 + 7t\sqrt{t} \sin(6\sqrt{t})$ .

For this case, Figure 4a shows the mean function of the process together with the boundary in the interval  $[0, 18]$ , and Figure 4b shows the FPTL function for the f.p.t. problem at hand. From the information provided by the FPTL function the f.p.t. density has been approximated from  $t_1^* = 3.0531$  and, initially, until  $t_{max,2}^+ = 18$ , first considering the fixed integration step  $h = \min\{(3.3195 - 3.0531)/n, (14.2556 - 11.0675)/n\}$  for  $n = 250$ , and, secondly, the following variable integration steps (for  $n = 250$  and  $p = 0.2$ ):

$$\begin{aligned} h_1 &= (3.3195 - 3.0531)/n \text{ in } [3.0531, 3.4689), \\ h_1^* &= (11.0675 - 3.4689)/[np + 0.5] \text{ in } [3.4689, 11.0675), \text{ and} \\ h_2 &= (14.2556 - 11.0675)/n \text{ in } [11.0675, 18]. \end{aligned}$$

Furthermore, for each of these cases there is a possibility to avoid subinterval  $[3.46895, 11.06751)$  in the application of the numerical algorithm by giving value zero to the f.p.t. density function in every point of the subinterval, since  $g_1(3.46895) \simeq 0$ .



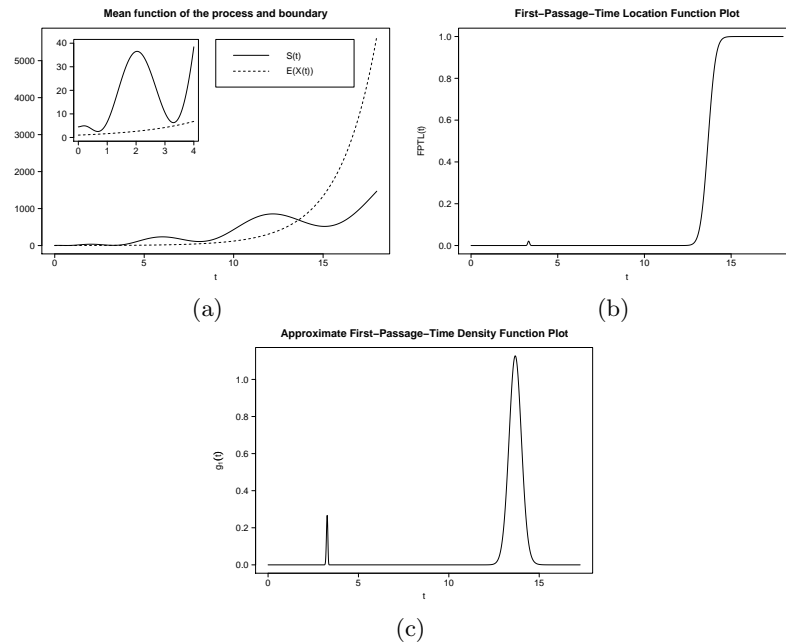
**Fig. 3.** Mean function of the process and boundary  $S(t)$  (a), FPTL function (b) and approximate f.p.t. density function (c) for application 1.

Table 1 shows the computational cost for each case, and Figure 4c shows the approximation obtained for the case with the smaller computational cost. In this case the algorithm was stopped at  $t = 15.35234$ , since the value of the accumulated probability by the approximation is over  $1 - 10^{-5}$  (alternative stopping rule). In all cases, the approximation is practically coincident, and  $P(3.05311 < T_{S(t),x_0} < 15.73197) > 1 - 10^{-5}$  is verified.

**Table 1.** Computational costs in application 2, avoiding subintervals in which the FPTL function is near zero (a), and applying the algorithm in all subintervals (b).

	Step			
	Fixed		Variable	
	Iterations	CPU time in seconds	Iterations	CPU time in seconds
(a)	4754	68.203	728	6.578
(b)	11883	277.906	776	7.265

The main conclusion is that, for both cases, considering a variable integration step considerably reduces computational cost. Also, not applying the algorithm to a, sometimes big, certain subintervals of values can have in this scenario an almost negligible effect.



**Fig. 4.** Mean function of the process and boundary  $S(t)$  (a), FPTL function (b) and approximate f.p.t. density function (c) for application 2.

### Acknowledgments

This work was supported in part by the Ministerio de Educación y Ciencia, Spain, under Grants MTM2008-05785 and HI2007- 0034, and by the Consejería de Innovación, Ciencia y Empresa (Junta de Andalucía), Spain, under Grant P06-FQM-02271.

### References

- BUONOCORE, A., NOBILE, A.G. and RICCIARDI, L.M. (1987): A new integral equation for the evaluation of first-passage-time probability densities. *Adv. Appl. Probab.* 19, 784-800.
- GUTIÉRREZ, R., ROMÁN, P., TORRES, F. (1999): Inference and first-passage-times for the lognormal diffusion process with exogenous factors: application to modelling in economics. *Appl. Stoch. Model. Bus. Ind.* 15, 325-332.
- RICCIARDI, L., DI CRESCENZO, A., GIORNO, V., NOBILE, A.G. (1999): An outline of theoretical and algorithmic approaches to first passage time problems with applications to biological modeling. *Math. Jpn.* 50, 247-322.
- ROMÁN, P., SERRANO, J. J., TORRES, F. (2008): First-passage-time location function: Application to determine first-passage-time densities in diffusion processes. *Comput. Stat. Data Anal.* 52, 4132-4146.

# Rplugin.Econometrics: R-GUI for Teaching Time Series Analysis

Dedi Rosadi<sup>1</sup>

Department of Mathematics, Research Group: Statistics  
Gadjah Mada University, Indonesia, *dedirosadi@ugm.ac.id*

**Abstract.** In this paper, we introduce `RcmdrPlugin.Econometrics` (Rosadi, Marhadi and Rahmatullah, 2009), a R-GUI for time series analysis, as a plug-in for R-Commander. `RcmdrPlugin.Econometrics` has nearly all of the typical models introduced in undergraduate time series courses, such as exponential smoothing, ARIMA/SARIMA, ARIMAX, ARCH/GARCH, etc. We discuss the philosophy of the plug-in design, compare and show its difference with a similar purpose R-Commander plug-in, called as `RcmdrPlugin.epack` (Hodgess and Vobach, 2008) and a commercial econometrics software EViews.

**Keywords:** R commander plug-ins, open source, time series analysis

## 1 Introduction

R (R Development Core Team, 2009), an open source programming environment for data analysis and graphics, has becoming the 'lingua franca' of data analysis and statistical computing. It is available for a number of platforms (Windows, Mac OS and various Unix and Linux), frequently updated (latest version per January 2010 is 2.10.0), a 'reliable' open-source software (its kernel developed by leading statistician and programmers, called as R-Core Team) and has an extensive online information and user-help (<http://www.R-project.org> and [r-help@r-project.org](mailto:r-help@r-project.org)).

The functionality of R is based on the add-on packages/library (similar to Toolbox in MATLAB). Default installation of R will automatically install and load several basic packages i.e., `base`, `datasets`, `utils`, `grDevices`, `graphics`, `stats`, `methods`. Beyond these packages, there are thousand of contributed packages, available in CRAN and related websites.

For time series analysis, there are various packages of R, available under the taskviews `Econometrics`, `Finance` and `Time Series` in CRAN), with the main user interaction via Command Line Interface (CLI). See, for instance, Cryer and Chan (2008), Kleiber and Zeilis (2008), Pfaff (2008), Racine and Hyndman (2002) and Cribari-Neto and Zarkos (1999) for a comprehensive discussion of R application in Time Series and Econometrics modeling. Unfortunately, for teaching purpose, R-CLI seems to be less user friendly and relatively difficult to use, especially if we compare it with the commercial softwares which has an extensive GUI capabilities, such as Eviews. For solving

this problems, Hodgess and Vobach (2008) introduced `RcmdrPlugin.epack`, a R-GUI package for doing time series analysis. In this paper, we introduce a new GUI package for time series analysis, called as `RcmdrPlugin.Econometrics` (Rosadi et al., 2009).

The rest of this paper is organized as follows. In the second section, we review the R-GUI, and especially the R-Commander which is the most popular R-GUI currently available for basic statistical analysis. In the third section, we provide a detail of design and implementation of the plugin. In the last section, we provide empirical examples of application of plugin for analysis of time series data, showing the unique features of the plugin `RcmdrPlugin.Econometrics`.

## 2 R-GUI

The main user interaction of R is using CLI (Command Line Interface), therefore, for some (beginners) user, it is less interactive, less user-friendly or relatively difficult to learn, especially compared to the statistical softwares with extensive GUI. For this purpose, some statistican and programmers have been developed the R-GUI version, such as R-Commander (Fox, 2005), R Sciviews (<http://www.sciviews.org/SciViews-R>), JGR (R-GUI Interface using Java, see <http://www.rosuda.org>), etc. (see <http://www.sciviews.org> for more information on R-GUI).

One of the most popular R-GUI package is R Commander (`Rcmdr`). It is developed using language tcl/tk (Welch, Jones and Hoobs, 2003; Dalgaard, 2001a,b and 2002) and provide the point and click GUI for doing some basic statistical analysis, such as:

- Data Management, such as for inputting, exporting (from external formats) and manipulating data.
- Basic statistics : such as for obtainig descriptive statistics, mean, proportion and variance test, etc.
- Advanced statistics tools: such as doing multivariate analysis, non parametric analysis, regression analysis, etc.
- Various graphics for visualizing data.
- Various functions for doing statistical analysis related to the distribution function.

In general, R commander is a useful GUI for doing the most common use basic statistical analysis. However, it does not contain menus for econometric analysis (except for regression analysis). However, it can be easily extended using suitable plug-in (Fox, 2009).

## 3 Design and Development of `RcmdrPlugin.Econometrics`

`Rcmdr Plug-in` is a plug-in that allow the package developers to develop GUI to their R package, with the R Commander providing most of the necessary

intrastructure (in a similar way as the add-ins menu in Microsoft excel). `RcmdrPlugin` introduced by John Fox and illustrated using his package called as `RcmdrPlugin.TeachingDemos` (Fox, 2009). Currently, in CRAN server, there are several `Rcmdr` plug-ins, one of them is `Rcmdrplugin.epack`. This plug-in contains GUI menu for analysing several standard time series models, studied in undergraduate forecasting and time series courses. However, it suffers from several problems, such as:

- The package contain a lot of errors
- The menu are not well organized, it scattered on several sub menus of `Rcmdr`
- The menus are mainly available for doing the univariate time series analysis
- The input and output dialog for doing a particular analysis is less interactive and too simple, especially if we compare it with GUI menu available in the commercial econometrics softwares, e.g. *Eviews*

Motivated by these reasons, in Rosadi et al. (2009), we develop a new but similar purpose plugins for R Commander which we called `RcmdrPlugin.Econometrics`. There are several goals design and developments for this plugins, as follows:

- **Open Source and Multiplatforms**

`RcmdrPlugin.Econometrics` is designed to be multi platforms, available for Windows and Linux. The plug in is an open-source and will be released under GPL, and it is available in CRAN server.

- **The ease of use**

The menu sctructure of `RcmdrPlugin.Econometrics` is grouped appropriately according to their purpose of the analysis, self-defined and easy to be accessed. Furthermore, when design the GUI layout, we only used the standard version of package `tcltk` and `tcl/tk`, and avoid to use the function that requires to install the extended version of `tcltk` and `tcl/tk`.

- **Input/Output dialog**

When we design the input-output dialog and layout for the GUI, we try to be more comprehensive, and to be compatible with GUI inputs and outputs dialog that are available in the commercial software, especially *Eviews*. Furthermore, since R is used as the computation engine in the background, then a particular menu can be programmed to give only the necessary output, following the standard procedure in statistical analysis. For instance, the menu for doing ARIMA analysis can be used to identify, to estimate the model, to do diagnostic check and to forecast based on the best model, following the standard Box-Jenkins ARIMA modeling procedure.

- **Menu coverage**

In the current version of the plug-in, the plug-in menu contains almost all the models that have been introduced in the course of Introduction to Forecasting and Introduction to Time Series Analysis, offered in our Department. `RcmdrPlugin.Econometrics` is consisting of five main menus:

**Simulation** (of ARIMA and ARCH/GARCH models), **Statistics** (Numerical summary, Jarque Berra Statistics for normality test), **Transformation** (Box-Cox Transformation, Difference, Log Difference, Time Series conversion, ADF Test for stationarity), **Plot** (Time Series Plot and ACF/PACF plot) and **Univariate time series analysis** (Smoothing, Decomposition, ARIMA, Automatic ARIMA, ARIMAX, ARCH/GARCH). As the 'to do list' for the future development of the plug-in, we will add some additional menus for doing spectral analysis, multivariate time series analysis such as VAR and Cointegration analysis, Granger Causality, ECM and VECM, etc.; dynamic linear model (ADL), panel model and several other linear and non linear models that are popular for Econometric society and studied in various econometrics related courses in our Department.

In the next section, we provide empirical examples showing some unique features of `RcmdrPlugin.Econometrics`.

## 4 Empirical example

The menus layout of the plug-in can be seen in the Figure 1. To illustrate the design philosophy of the plug-in, we will show its application for doing exponential smoothing and ARCH/GARCH analysis. We start by considering the data Spain from Enders (1995), containing the monthly number of tourism visiting Spain during January 1970 until March 1989.

The exponential smoothing GUIs in Eviews, `RcmdrPlugin.epack` and `RcmdrPlugin.Econometrics`, are given in Figure 2, 3 and 4, respectively. We see that the GUI dialog of `RcmdrPlugin.Econometrics` is more comprehensive than `RcmdrPlugin.epack`, it contains not only the dialog for estimation of model, but user also can enter the time lag for forecasting using a particular forecasting method that have been chosen by the user. Furthermore, as the output of smoothing dialog in `RcmdrPlugin.Econometrics`, it will print the plots of the confidence interval of the prediction, the fitted values and the original data (see Figure 5). In R Commander output window, it will also print the smoothing parameter, the data filtered using the chosen model, and the interval confidence for  $n$ -step ahead forecasting. The filtered data will be added to the active data sheet.

In the second example, we consider the data of daily exchange rates from DeutschMark to Poundsterling (time periods is unknown). We will analyze the log returns series of this data using ARCH/GARCH model. The GUI dialog for ARCH/GARCH analysis using Eviews, `RcmdrPlugin.epack` and `RcmdrPlugin.Econometrics` are depicted in Figure 6, Figure 7 and Figure 8, respectively. These figures shows again that the GUI dialog `RcmdrPlugin.Econometrics` is more comprehensive than `RcmdrPlugin.epack`, it contains not only the dialog for model estimation, but it also provide dialogs to enter the time lag for mean and volatility forecasting (this important feature is not available in



EViews). Furthermore, users can also choose the distribution assumption for the error in the ARCH/GARCH model.

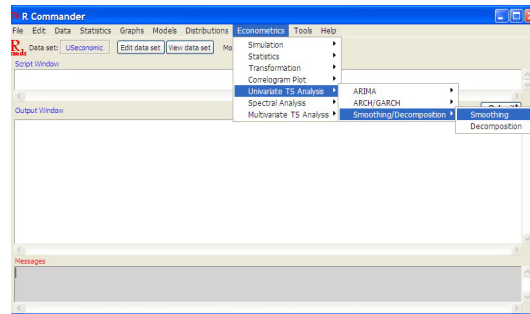
## 5 Concluding remarks

`RcmdrPlugin.Econometrics` is the plug-in of R Commander, which is designed for doing econometrics/time series analysis. From the previous discussion, we can see that its GUI layout is more comprehensive than the similar purpose R-package, called as `RcmdrPlugin.epack` and more compatible with EViews. For teaching and research purposes, we expect that in the future `RcmdrPlugin.Econometrics` can be used in addition to or in place of commercial econometrics software available in the market.

**Acknowledgement.** The financial support from the Ministry of Science and Technology of Indonesia via "Insentif Riset Terapan" in 2009, see Rosadi et al. (2009), is gratefully acknowledged.

## References

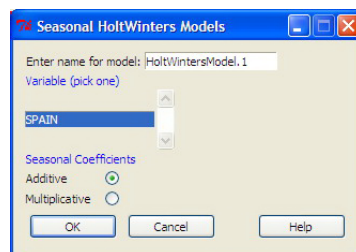
- CRIBARI-NETO, F. and ZARKOS, S.G. (1999): R: Yet another econometric programming environment. *Journal of Applied Econometrics*, 14, 319 – 329.
- CRYER, J.D. and CHAN, K-S. (2008): *Time Series Analysis, With Applications in R. Second Edition*, Springer.
- DALGAARD, P. (2001a): The R-Tcl/Tk interface. In: K. Hornik and F. Leisch (Eds.), *Proceedings of the 2nd International Workshop on Distributed Statistical Computing*. Technische Universität Wien, Vienna, Austria, 2001. URL <http://www.ci.tuwien.ac.at/Conferences/DSC-2001/Proceedings/>. ISSN 1609-395X.
- DALGAARD, P. (2001b): A Primer on the R-Tcl/Tk Package, *R News*. vol. 1/3, September 2001.
- DALGAARD, P. (2002): Changes to the R-Tcl/Tk package, *R News*. vol. 2/3, December 2002.
- ENDERS, W. (1995): *Applied Econometric Time Series*. Wiley.
- FOX, J. (2005): The R Commander : A Basic Statistics Graphical User Interface to R,. *Journal of Statistics Software*, 14 (9).
- FOX, J. (2009): *RcmdrPlugin.TeachingDemos*. [Online] Available at [www.cran.r-project.org](http://www.cran.r-project.org) .
- HODGESS, E. and VOBACH, C. (2008), RcmdrPlugin.epack: A Menu Driven Package for Time Series in R. Paper presented at *The annual meeting of the The Mathematical Association of America MathFest*, TBA, Madison, Wisconsin, Jul 28, 2008.
- KLEIBER, C. and ZEILIS, A. (2008): *Applied Econometrics with R*, Springer.
- PFAFF, B. (2008): *Analysis of Integrated and Cointegrated Time Series with R*, Springer
- R Development Core Team (2009): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.



**Fig. 1.** Rcmdr with RcmdrPlugin.Econometrics menu



**Fig. 2.** Eviews dialog for Exponential Smoothing



**Fig. 3.** Exponential smoothing dialog in RcmdrPlugin.epack

- RACINE and HYNDMAN, R. (2002): Using R to teach econometrics. *Journal of Applied Econometrics*, 17 (2), 175 – 189.
- ROSADI, D., MARHADI, A. and RAHMATULLAH, F. (2009): *Developing R GUI for econometrics analysis using open source Tcl/Tk and its application for oil demand modeling*, Research Report. Unpublished. In Bahasa Indonesia.
- WELCH, B., JONES, K. and HOBBS, J., (2003): *Practical Programming in Tcl and Tk*, 4th eds. Prentice Hall.

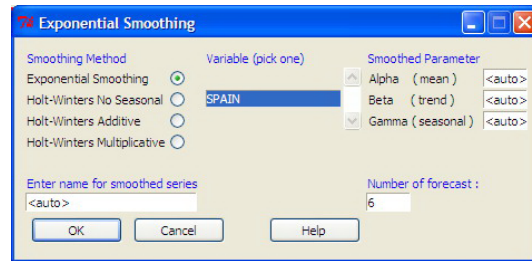


Fig. 4. Exponential smoothing dialog in RcmdrPlugin.Econometrics

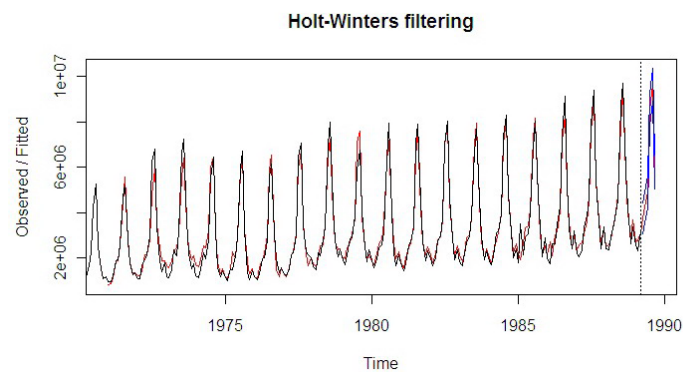


Fig. 5. Exponential smoothing prediction plot

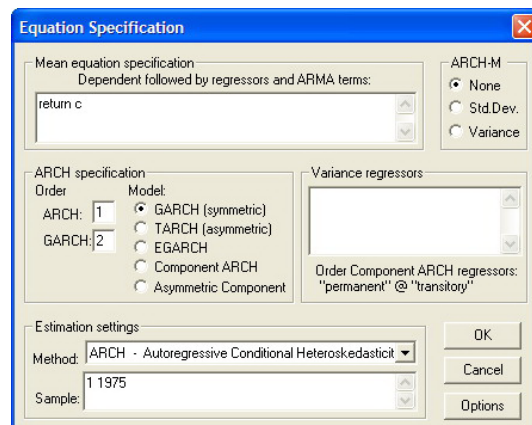
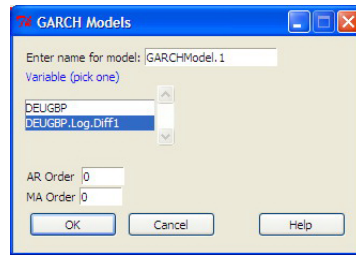
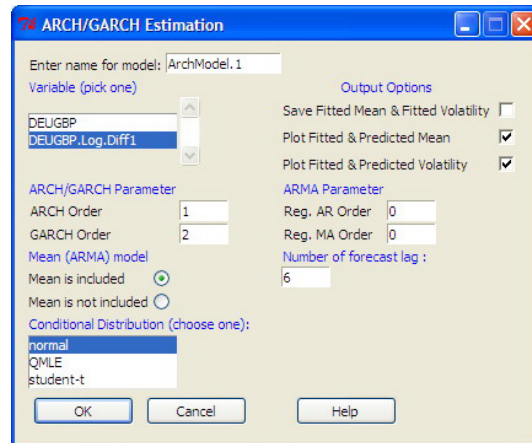


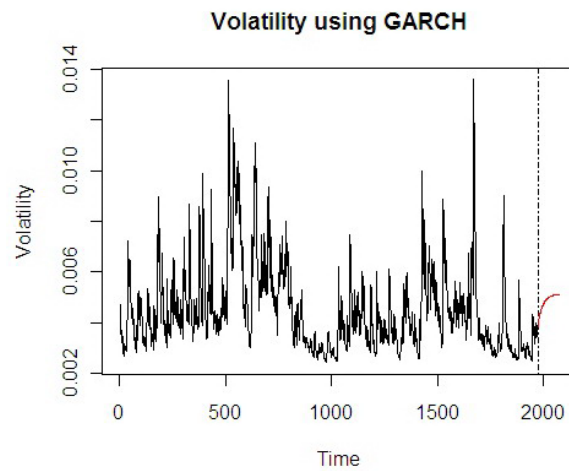
Fig. 6. ARCH/GARCH Estimation dialog in Eviews



**Fig. 7.** ARCH/GARCH Estimation dialog in RcmdrPlugin.epack



**Fig. 8.** ARCH/GARCH Estimation dialog in RcmdrPlugin.Econometrics



**Fig. 9.** Volatility forecasting using GARCH(1,1) models

# EOFs for Gap Filling in Multivariate Air Quality data: a FDA Approach

Mariantonietta Ruggieri, Francesca Di Salvo, Antonella Plaia, and  
Gianna Agró

Department of Statistical and Mathematical Sciences, University of Palermo  
viale delle Scienze - building 13, 90128 Palermo, Italy.

*ruggieri@dssm.unipa.it, disalvo@dssm.unipa.it, plaia@unipa.it, agro@unipa.it*

**Abstract.** Missing values are a common concern in spatiotemporal data sets. During recent years a great number of methods have been developed for gap filling. One of the emerging approaches is based on the Empirical Orthogonal Function (EOF) methodology, applied mainly on raw and univariate data sets presenting irregular missing patterns. In this paper EOF is carried out on a multivariate space-time data set, related to concentrations of pollutants recorded at different sites, after denoising raw data by FDA approach. Some performance indicators are computed on simulated incomplete data sets with also long gaps in order to show that the EOF reconstruction appears to be an improved procedure especially when long gap sequences occur.

**Keywords:** FDA, EOF, missing data, gap filling

## 1 Introduction

In air pollution data sets missing values are usually due to measurement errors or misfunctions of instruments in the monitoring network, representing a crucial problem to face in order to perform subsequent analyses. An inadequate preprocessing may lead to low-quality data and hamper the extraction of meaningful information about their temporal and spatial characteristics. The simplest method for imputing missing data, or also for replacing potential unreliable data, is using a sample average, having the merit of not needing any a priori assumption or complex calculations. Actually, the sample average presents the drawback of assuming all of the observations as equally important and does not consider they are collected over time and space. A lot of more sophisticated methodologies aiming at filling in the gaps have been developed in literature; polynomial functions and splines, iterative approaches based on model parameter estimation, such as Expectation-Maximization (EM) and kriging, are only some examples. A brief review on more recent imputation methods is presented in Section 2. The contribution of our work lies in using EOF based interpolation jointly with Functional Data Analysis (FDA) approach (Ramsay and Silverman (2002, 2005)) and in applying them on a multivariate space-time data set. Actually, the FDA alone could solve

the problem of missing values, when only short gaps occur; here, time series exhibiting long gaps are also considered in order to show that the EOF reconstruction is a more valid method especially for long gap sequences. The FDA is used as a powerful tool for denoising observed data as well as for its well-known computational advantages; in fact, converting discrete time series into functional data, it allows to deal with a low number of spline coefficients rather than a large number of observations. In Section 3 the starting air pollution data set is presented. In Section 4 the FDA approach and the methodology used for gap filling are described. In Section 5 the method to generate simulated incomplete data sets and some performance indicators, computed in order to validate the performed imputation procedure, are reported. Finally, Section 6 is devoted to test imputation method and comment obtained results.

## 2 A gap filling review

The approaches for processing missing values can be classified in parameter-dependent model-based and nonparametric methods, relying only on the data. Classical parametric methods are mainly based on optimal interpolation (OI), such as objective analysis; more advanced approaches may be considered as a variant of OI. All these implemented methods require the knowledge of the spatio-temporal covariance structure for both the analyzed data set and related errors. Improvements over traditional OI are represented by the EM and by the geostatistical filling-in procedure, but data are required to be gaussian and missing values distributed at random in time. A recent parameter-free interpolation method is proposed by Beckers and Rixen (2003); it is based on EOFs. The EOF is a deterministic methodology, enabling a linear projection to a high-dimensional space and a continuous interpolation even if a high percentage of gaps occurs. With respect to classical interpolation method as OI, the EOFs take the advantage of not requiring any information about a correlation function of the data themselves, since EOFs and missing values are iteratively estimated. In order to find the best number of leading EOFs to be retained, cross-validation algorithms are adopted. The EOF analysis is widely used in climate research to fill in missing data as well as a denoising tool; actually, it is also applied to other scientific areas, although with different names, such as Proper Orthogonal Decomposition, in a functional framework, or Proper Orthogonal Modes and Principal Component Analysis, for discrete data sets. A new method to fill in gaps is proposed by Kondrashov and Ghil (2006). They use a novel, iterative form of Singular Spectrum Analysis (SSA), considering only temporal correlations in the data, and multi-channel SSA (M-SSA), considering both spatial and temporal correlations. The method may be regarded as a generalization of the Beckers and Rixen's spatial-EOF based reconstruction and is particularly useful for data sets exhibiting a relatively large and continuous fraction of gaps in space or

in time. An improved version of the standard EOF methodology, called EOF pruning, is presented by Sorjamaa et al. (2009). It enhances the accuracy of the EOF method and decreases the calculation time needed to approximate the missing values, taking into account the information possible contained in some smaller singular values, when removing the noise is not the main purpose of the analysis.

### 3 The air pollution data set

In this paper a spatiotemporal data set of concentrations for four main pollutants ( $CO$ ,  $NO_2$ ,  $PM_{10}$  and  $SO_2$ ) is considered. Data, provided by AMIA (Azienda Municipalizzata Igiene Ambientale, <http://www.amianet.it/>), are hourly (bi-hourly for  $PM_{10}$ ) recorded by the 9 monitoring stations distributed in Palermo (Italy) during 2005. In order to obtain daily syntheses, we aggregate by time hourly (or bi-hourly) data at each site for each pollutant, using the functions suggested by EC guidelines. Daily average is not computed with more than 25% of missing values on a day, giving rise to a gap to be imputed for that day. Data are standardized by linear interpolation (Di Salvo et al. (2009)) according to EC directives thresholds, to allow a comparison among pollutants different for measurement unit or order of magnitude. We convert the standardized observed data set into a functional data set, as described in Section 4. Such a conversion, by assuming the existence of a continuous function giving rise to the observed data, enables to work with a few coefficients rather than a large number of data and denoises time series by fluctuations due to contingent factors by preserving their temporal pattern.

### 4 FDA and EOF approaches

According to FDA approach, the generic realization  $x_i^{p_j}(t)$ , recorded at time  $t$  ( $t = 1, \dots, T$ ) for the pollutant  $p_j$  ( $j = 1, \dots, P$ ) at the station  $i$  ( $i = 1, \dots, N$ ), is considered as a signal,  $\tilde{x}_i^{p_j}(t)$ , plus a noise  $\varepsilon(i, t)$ :

$$x_i^{p_j}(t) = \tilde{x}_i^{p_j}(t) + \varepsilon(i, t). \quad (1)$$

The functional datum  $\tilde{x}_i^{p_j}(t)$  is represented by a linear expansion in terms of a basis functions system  $\phi_k$ , that we assume to be unique for all the considered pollutants:

$$\tilde{x}_i^{p_j}(t) = \sum_k^K c_{i,k,p_j} \phi_k(t), \quad (2)$$

with  $K$  the number of basis function coefficients.

In particular, we choose the cubic B-spline basis system with equally spaced interior knots. Once defined the simultaneous expansion of the  $N$

curves (2) for the  $P$  pollutants, their curvature is penalized by a parameter  $\lambda \geq 0$ ; so, the smoothing strategy depends on the values of  $\lambda$  and  $K$ . More technical details on the roughness penalty approaches for functional data are exhaustively treated in Ramsay & Silverman (2005).

Before transforming the 3D ( $N \times T \times P = 9 \times 365 \times 4$ ) array  $\mathbf{X}$  of observed data into the functional  $\tilde{\mathbf{X}}$ , the  $m$  missing values are replaced by the annual mean, given the station and the pollutant. If the functional datum corresponding to the missing value may be a good representation of the real datum for short gap sequences, a better reconstruction for long gaps might be obtained by means of the EOFs, computed by performing the standard Functional Principal Components Analysis (FPCA), treated by Ramsay and Silverman (2005), or the Singular Value Decomposition (SVD) on  $\tilde{\mathbf{X}}$ . As it is known, the first selected  $\nu < \text{rank}(\tilde{\mathbf{X}})$  singular values, sorted to decreasing order, are chosen on the basis of the explained variation and used, with the corresponding singular vectors, to reconstruct  $\tilde{\mathbf{X}}$ . In fact, it is assumed that the vectors corresponding to the largest singular values hold more signal than noise with respect to the ones corresponding to the smallest values. Denoted by  $\hat{\mathbf{X}}$  the reconstruction of  $\tilde{\mathbf{X}}$ , the problem is to test if the generic element  $\hat{x}_l$  ( $l = 1, 2, \dots, m$ ) of  $\hat{\mathbf{X}}$ , corresponding to a missing value, may be considered a good imputation of such a missing value. Obviously, this must be verified for the whole set of the  $m$  missing values and above all at long gap sequences. In order to evaluate the performance of the imputation method, simulated incomplete data sets are generated, then the method is applied and some performance indicators are computed (cfr. Section 5).

It should be underlined that to extract the EOFs by FPCA or SVD, the latter chosen in this paper, the three way data array  $\tilde{\mathbf{X}}$  has to be matricized to obtain a two way data matrix; this can be done in different ways, according to what the researcher wants to focus on. We choose to focus on the monitoring sites, so the data array is matricized by concatenating horizontally the four matrices related to each pollutant:

$$\tilde{\mathbf{X}}_{9 \times (365 \times 4)} = [\tilde{\mathbf{X}}^{p_1}, \tilde{\mathbf{X}}^{p_2}, \tilde{\mathbf{X}}^{p_3}, \tilde{\mathbf{X}}^{p_4}].$$

Actually, from a computational point of view, both the FPCA and the SVD are not performed on  $\tilde{\mathbf{X}}$ , but on a matrix  $\mathbf{Z}$ , of size  $9 \times (K \times 4)$ , containing the coefficients of the basis functions. In our case, after doing the SVD, for  $q < \text{rank}(\mathbf{Z})$  singular values  $\rho_i$ , the matrix  $\mathbf{Z}$  can be approximated as:

$$\mathbf{Z}_{9 \times (K \times 4)} \approx \mathbf{U}_{9 \times q} \mathbf{\Gamma}_{q \times q} \mathbf{A}_{q \times (K \times 4)},$$

where  $\mathbf{U}$  is the matrix whose columns are the standardized principal scores, allowing more interpretative results in some situations, and the diagonal matrix  $\mathbf{\Gamma}$  has elements  $\sqrt{\rho_i}$  (for details see Di Salvo et al. (2009)).



## 5 Missing data simulation and imputation method validation

It is usually difficult to determine and compare the accuracy of different imputation methods because, being unable to retrieve the missing data, a method must be designed to create a data set that mimics real life data and missing data patterns; as a matter of fact, the imputation performance does not depend only on the amount of missing data but also on the characteristics of the missing data mechanism. The standard classification of missing data mechanism considers data: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Little and Rubin (1987)). Usually, environmental data are at least MAR.

Many studies on imputation methods use real data sets and simulated missing data patterns by deleting values. In this work, we create some "artificial" gaps in the observed array  $\mathbf{X}$ , by reproducing the pattern of missing values in the observed data set. Table 1 summarizes the observed missing data patterns, conditionally to the monitoring site and the pollutant, together with the total amount of missing daily values (percentage). As it reveals, most of the sequences of missing values are very short, from 1 to 3 values long, only few sequences from 4 to 9 values long and very few sequences (only 8) greater than 10 values long. In order to reproduce the actual pattern of missing data, 50 missing data indicator matrices  $\mathbf{M}$ , with dimensions as our data set ( $9 \times 365 \times 4$ ), are randomly generated from a Bernoulli distribution with parameter  $\pi$  equal to the actual percentage of missing in each monitoring site, since in MAR, conditionally to the observed data, remaining missingness is completely random. Sometime (but not so rarely) it can happen that very long gap sequences are observed in a data set: this can be due to long time failures not easily solvable or to data coming from a mobile monitoring station. To reproduce this kind of patterns, very long gap sequences (the length is randomly generated in the 45-90 days range) are randomly generated and randomly positioned in the matrices  $\mathbf{M}$  above described, in order to test if a different performance of EOFs reconstruction is obtained. Then, each matrix  $\mathbf{M}$  is applied to the observed data set  $\mathbf{X}$  creating "artificially" missing data (actually real values are known) and this allows to compute the value of some performance indicators to assess the goodness of the imputation method.

Two performance indicators are here considered (Plaia and Bondí (2006)):

- the coefficient of correlation  $\rho$  between functional and imputed data

$$\rho = \left[ \frac{1}{m} \frac{\sum_{l=1}^m [(\tilde{x}_l - M_{\tilde{x}_l})(\hat{x}_l - M_{\hat{x}_l})]}{\sigma_{\tilde{x}_l} \sigma_{\hat{x}_l}} \right] \quad (3)$$

- the root mean square deviation  $RMSD$

$$RMSD = \left( \frac{1}{m} \sum_{l=1}^m [\tilde{x}_l - \hat{x}_l]^2 \right)^{1/2} \quad (4)$$

where:

Gap length	Pollutant	St1	St2	St3	St4	St5	St6	St7	St8	St9
$1 \leq l \leq 3$	<i>CO</i>	2	12	8	8	3	2	6	5	3
	<i>NO<sub>2</sub></i>	24	8	17	4	5	8	10	17	9
	<i>PM<sub>10</sub></i>	4	16	12	8	4	7	6	9	4
	<i>SO<sub>2</sub></i>	13	13	23	10	12	7	13	13	7
$4 \leq l \leq 9$	<i>CO</i>	0	2	4	3	1	0	3	0	0
	<i>NO<sub>2</sub></i>	3	3	6	2	2	2	4	2	3
	<i>PM<sub>10</sub></i>	0	0	4	3	2	0	1	0	0
	<i>SO<sub>2</sub></i>	1	0	6	2	1	0	3	1	0
$l \geq 10$	<i>CO</i>	0	0	1	0	0	0	0	0	0
	<i>NO<sub>2</sub></i>	0	0	2	0	1	0	0	0	1
	<i>PM<sub>10</sub></i>	0	0	0	0	0	0	0	0	0
	<i>SO<sub>2</sub></i>	0	0	0	0	0	1	1	1	0
% of missing	<i>CO</i>	0.82	8.77	14.25	7.12	1.92	0.55	7.40	1.92	2.19
	<i>NO<sub>2</sub></i>	12.33	7.95	17.53	3.84	3.01	4.93	10.14	11.78	13.70
	<i>PM<sub>10</sub></i>	1.64	6.58	21.10	6.85	7.95	2.19	4.11	2.47	1.64
	<i>SO<sub>2</sub></i>	6.30	6.03	18.08	5.48	9.59	5.75	13.70	9.86	3.29
	Total	5.41	7.51	18.19	5.97	5.76	3.44	9.06	6.67	5.34

**Table 1.** Missing data patterns.

- $m$  is the number of imputations (that is the number of missing data);
- $\tilde{x}_l$  ( $\hat{x}_l$ ) is the  $l^{th}$  functional (imputed) data point,  $l = 1, 2, \dots, m$ ;
- $M_{\tilde{x}_l}$  ( $M_{\hat{x}_l}$ ) and  $\sigma_{\tilde{x}_l}$  ( $\sigma_{\hat{x}_l}$ ) are the average and the standard deviation of all the  $\tilde{x}_l$  ( $\hat{x}_l$ ), respectively.

The index (4), lower the better, with respect to the coefficient of correlation (3), higher the better, is related to the sizes of the discrepancies between predicted and functional values.

## 6 Testing imputation procedure and results

As a first step, real missing data present in the observed matrix  $\mathbf{X}$  are imputed with the annual mean, given the station and the pollutant, then the matrix  $\mathbf{X}$  is transformed into the functional data matrix  $\tilde{\mathbf{X}}$ . Subsequently, each of the matrices  $\mathbf{M}$  (cfr. Section 5) is overlapped to  $\mathbf{X}$  creating "artificial" gaps. On each of these 50 matrices, missing values are imputed with the annual mean, given the station and the pollutant, and on the resulting matrices FDA is applied, getting matrices  $\tilde{\mathbf{X}}^{\mathbf{M}}$ . By extracting the EOFs from  $\tilde{\mathbf{X}}^{\mathbf{M}}$ , the reconstructed matrices  $\hat{\mathbf{X}}^{\mathbf{M}}$  are obtained. For each matrix  $\tilde{\mathbf{X}}^{\mathbf{M}}$ , the subset containing the imputed values only is named  $I_{FD}$ ; similarly, the subsets containing the imputed values of each matrix  $\tilde{\mathbf{X}}^{\mathbf{M}}$  is named  $I_{EOF}$ .

The number of EOFs ( $\nu = 5$ ), extracted by SVD, is chosen on the basis of the explained total variability (about 95% for each matrix), while two different values of  $K$  (number of coefficients of the basis functions) and  $\lambda$  (smoothing parameter) are considered (cfr. Table 4.2).

The entire procedure is implemented in R (<http://cran.rproject.org>.)

The imputed values  $I_{EOF}$  and  $I_{FD}$  are compared with the corresponding values in  $\tilde{\mathbf{X}}$  by means of the performance indicators (3) and (4). The distributions of the performance indicators, over the 50 matrices  $\mathbf{M}$ , are summarized by their means  $\hat{\mu}$  and standard deviations  $\hat{\sigma}$  in Table 4.2, considering the whole set of  $m$  missing values or only long gap sequences. As it can be observed,  $I_{EOF}$  and  $I_{FD}$  results are very similar, for both the two performance indicators and the two sets of smoothing parameters, when all missing values are considered.

Smoothing	Imputed	all missing values				long gap sequences			
		$\hat{\mu}_\rho$	$\hat{\sigma}_\rho$	$\hat{\mu}_{RMSD}$	$\hat{\sigma}_{RMSD}$	$\hat{\mu}_\rho$	$\hat{\sigma}_\rho$	$\hat{\mu}_{RMSD}$	$\hat{\sigma}_{RMSD}$
$K = 183$	$I_{FD}$	0.973	0.009	3.202	0.543	0.120	0.205	4.373	2.276
$\lambda = 2$	$I_{EOF}$	0.970	0.007	3.475	0.339	0.718	0.215	3.681	1.912
$K = 53$	$I_{FD}$	0.992	0.004	1.622	0.456	0.485	0.520	2.745	1.820
$\lambda = 20$	$I_{EOF}$	0.993	0.003	3.473	0.566	0.562	0.570	5.230	3.060

**Table 2.**  $\hat{\mu}$  and  $\hat{\sigma}$  of performance indicators.

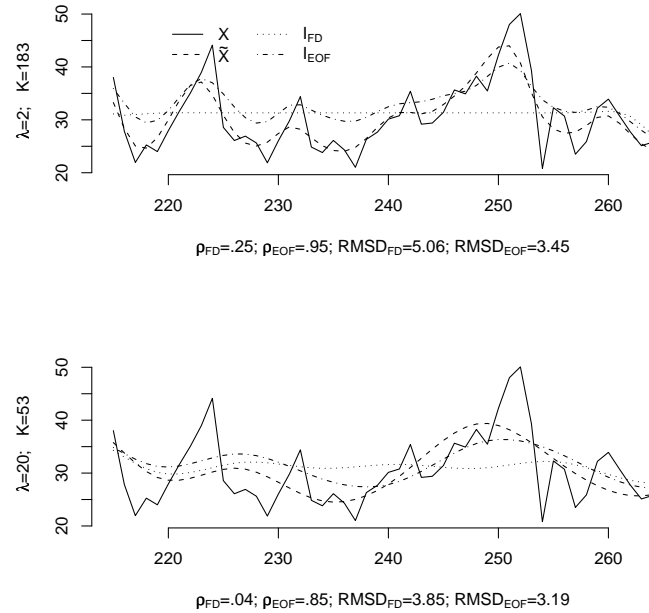
If only long gap sequences are taken into account, the less is the smoothing ( $K = 183$ ,  $\lambda = 2$ ) the more improvements are evident for  $I_{EOF}$  results with respect to  $I_{FD}$ . This behaviour is shown in Figure 1 for one long gap sequence, where the reconstructions  $I_{FD}$  and  $I_{EOF}$ , together with the corresponding sequences of  $\mathbf{X}$  and  $\tilde{\mathbf{X}}$ , are reported.

On the basis of the obtained results, the EOF procedure that, unlike what is found in literature, is here applied on a multivariate space-time data set, seems to provide interesting results especially if very long gap sequences occur. Such a feature is particularly attractive when data from mobile and fixed monitoring stations need to be integrated. Furthermore, it is worth to underline that it might also be considered in order to extend series into the future, that is for solving forecast issues.

**ACKNOWLEDGEMENTS:** This research is funded by the University of Palermo (Plaia and Ruggieri). The authors thank the anonymous referees for their helpful comments.

## References

- BECKERS, J.M. and RIXEN, M. (2003): EOF Calculations and Data Filling from Incomplete Oceanographic Datasets. *Journal of Atmospheric and Oceanic Technology* 20 (12), 1839-1856.
- DI SALVO, F., AGRÓ, G., PLAIA, A. and RUGGIERI, M. (2009): Exploring spatio-temporal patterns in air pollution data by FPCA. *Submitted to Environmetrics*.



**Fig. 1.**  $I_{FD}$  and  $I_{EOF}$  for one example of long gap sequence and two sets of smoothing parameters.

- KONDRASHOV, D. and GHIL, M. (2006): Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics* 13, 151-159.
- LITTLE, R.J.A. and RUBIN, D.B. (1987): *Statistical Analysis with Missing Data*. Wiley, New York.
- MURENA, F. (2004): Measuring air quality over large urban areas: development and application of an air pollution index at the urban area of Naples. *Atmospheric Environment* 38, 6195-6202.
- OTT, W.R. and HUNT, W.F. (1976): A quantitative evaluation of the pollutant standards index. *Journal of the Air Pollution Control Association* 26, 1050-1054.
- PLAIA, A. and BONDÍ, M. (2006): Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment* 40, 7316-7330.
- RAMSAY, J.O. and SILVERMAN, B.W. (2002): *Applied Functional Data Analysis*. Springer-Verlag.
- RAMSAY, J.O. and SILVERMAN, B.W. (2005): *Functional Data Analysis. Second Edition*. Springer-Verlag.
- SORJAMAA, A., LENDASSE, A., CORNET, Y. and DELEERSNIJDER, E. (2009): An improved methodology for filling missing values in spatiotemporal climate data set. *Computational Geosciences* 14 (1), 55-64.

# A Transient Analysis of a Complex Discrete $k$ -out-of- $n$ : $G$ System with Multi-State Components

Juan Eloy Ruiz-Castro<sup>1</sup> and Paula R. Bouzas<sup>2</sup>

<sup>1</sup> Department of Statistic and Operations Research.  
University of Granada, 18071-Granada, Spain. *jeloy@ugr.es*

<sup>2</sup> Department of Statistics and Operations Research.  
University of Granada, 18071-Granada, Spain. *paula@ugr.es*

**Abstract.** A system with  $n$  components that works if and only if at least  $k$  of them does it is called a  $k$ -out-of- $n$ : $G$  system. A discrete  $k$ -out-of- $n$ : $G$  system is modelled by considering multi-state components. Phase-type distributions for the lifetime of the units are considered. The units can undergo repairable and non-repairable failures from any state. We assume a general number of repairpersons. The repair time for each repairperson is general distributed and the phase type representation is considered. The system is modelled and some performance measures of interest are built. All results have been implemented computationally with *Matlab*. A numerical application shows the versatility of the model.

**Keywords:** discrete  $k$ -out-of- $n$ : $G$  system, phase-type distribution, reliability

## 1 Introduction

The  $k$ -out-of-system:  $G$  system is a popular type of redundancy that finds applications in several fields, such as electronic, industrial and military systems. The literature about this class of systems is very extensive but they are often considered with components that work in continuous time. An analysis of advanced reliability systems is shown in Pham (2008). When a complex system is considered for modelling, it is usual to assume the exponential distribution for the times involved in this one. Moustafa (2001) used the imbedded Markov chain to obtain the availability of  $k$ -out-of- $n$ :  $G$  systems subject to exponential failures and general repairs. There are several aspects of interest to study when a system is modelled. Among others, the modeling, the analysis of the performance measures and optimal values if a cost function associated to the system is defined. Krishnamoorthy and Ushakumari (2001) consider a  $k$ -out-of- $n$ :  $G$  system with repair under  $D$ -policy. They model the system assuming that lifetimes of components are exponentially distributed, work out some performance measures by considering phase type distributions and compute optimal values for the repair policy.

As it has been mentioned above, the most of the studies of redundant systems in the reliability literature have been performed by considering that the units work on continuous time. Nevertheless, due to several causes, such as the inner structure of the system, all the systems cannot be continuously monitored and they must be observed at certain epochs. Thus, some reliability systems, such as digital computer systems, have a discrete behaviour by time. In this case, the state of the system is known in discrete times. Redundant systems in discrete times have been modelled. Ruiz-Castro et al. (2009a) model a discrete warm-standby system and the stationary distribution, performance reliability measures in transient and stationary regime; the up period, and the involved costs are worked out in a matrix and algorithmic form. Also, Ruiz-Castro et al. (2009b) study a cold-standby system subject to different types of failures with loss of units. In both cases, phase type distributions are considered.

Phase type distributions (PH) are described in detail in Neuts (1981). This class of distributions enables us to model systems where each component has several performance stages in a well-structured matrix form. Furthermore, in discrete case, any distribution with finite support is a phase type distribution. Thus, any discrete distribution has a phase type representation.

In the present paper a discrete  $k$ -out-of- $n$ :  $G$  system is modeled. Each unit can occupy several performance stages for its lifetime. The lifetime of each unit is phase type distributed. The units are subject to two types of failures: repairable and non-repairable. When a repairable failure occurs, the unit goes to the repair facility. This repair facility is composed by a general  $R$  number of repairpersons and the repair time is general distributed. The model is built and performance measures of interest are computed in transient regime. The results have been implemented computationally and a numerical example is shown.

## 2 The system and its modelling

A discrete  $k$ -out-of- $n$ :  $G$  system is considered with  $R$  repairpersons. The system is performing when at least  $k$  units are operational. Each unit can occupy several performing stages and two types of failures can undergo: repairable and non-repairable. When a failure occurs, it is repairable with probability  $p$  or non-repairable with probability  $1 - p$ . If the failure is repairable then the unit goes to the repair facility. In other case, the unit is replaced instantaneously by a new and identical one. There are  $R$  repairpersons in the repair facility. We assume that the number of repairpersons is less or equal than the number of units in the system. If the failure of the system occurs, only the repair is operating. The discrete case is more complex than the continuous one given that several transitions can occur at same time.

## 2.1 Assumptions and state space

*Assumption 1:* Each unit can occupy  $m$  states before the failure. The lifetime of each unit is PH distributed with representation  $(\alpha, T)$ . The order of this matrix is equal to  $m$ .

*Assumption 2:* If a failure occur this one is repairable with probability  $p$  and non-repairable with probability  $1 - p$ .

*Assumption 3:* If a non-repairable failure occurs, the unit is replaced instantaneously by a new and identical one.

*Assumption 4:* The repair time of each repairperson is PH distributed with representation  $(\beta, S)$ . The order of this matrix is equal to  $q$ .

*Assumption 5:* The times involved in the model are independents.

### State Space

Each unit can occupy several states as it was mentioned above. The situation of the system is determined through the following macro-states. These ones are defined as the number of non-operational units. The macro-states can be partitioned as follows

$$\begin{aligned} E_0 &= \{(i_1, \dots, i_n); 0 \leq i_\nu \leq m, 1 \leq \nu \leq n\} \\ E_h &= \{(i_1, \dots, i_{n-h}; j_1, \dots, j_{\min\{h, R\}}); 0 \leq i_\nu \leq m, 0 \leq j_s \leq q, \\ &\quad 1 \leq \nu \leq n - h, 1 \leq s \leq \min\{h, R\}; 1 \leq h < n \\ E_n &= \{(j_1, \dots, j_R); 0 \leq j_s \leq q, 1 \leq s \leq R\}, \end{aligned}$$

where  $i_\nu$  is the state of the  $\nu$ -th operational unit and  $j_s$  the stage of the  $s$ -th repairing.

The up macro-state (the system is working on) is given by  $U = \bigcup_{h=0}^{n-k} E_h$  and the down macro-state is  $D = \bigcup_{h=n-k+1}^n E_h$ .

## 2.2 The transition probability matrix

The behaviour and evolution of the system is analyzed through a Discrete Markov Process with state space described above. The transition probability matrix by blocks is given by

$$P = (B_{ij})_{i,j=0,\dots,n}. \quad (1)$$

The matrix  $B_{ij}$  is equal to zero in several situations. If the system is broken ( $i \geq n - k + 1$ ) then the transition until a new failure is not possible ( $j > i$ ). On the other hand, if there are  $i$  units in repair, then at most there will be  $i - R$  units in repair at next time. Then, the next transition is also not possible,  $E_i \rightarrow E_j$  with  $j < \max\{0, i - R\}$ . In this case  $B_{ij} = \mathbf{0}$ .

We build the transition probability matrix by blocks by considering the macro-states described above. We introduce some functions for doing more

understandable the methodology. These ones have been developed and they are available from the authors. We define

- $b(i; r; n_r)$ : matrix probability, according to the state space, that there are  $i$  units in the repair facility and that  $r$  repairable and  $n_r$  non-repairable failures occur at the next time. This matrix has order  $m^{n-i} \times m^{n-i-r}$ . The element  $(a, b)$  of this matrix is the probability that the operational units are in the phases given by the row  $a$  before the failures and, then the units that do not fail are in the phases given by the column  $b$  after the failures.
- $rep(i; j)$ : matrix probability, according to the state space, that there are  $i$  units under repair and that  $j$  units are repaired at the next time. This matrix has order  $q^{\min\{i, R\}} \times q^{\min\{i, R\}-j}$ . The element  $(a, b)$  of this matrix is the probability that the units in repairing are in the phases given by the row  $a$  before the failures, and then, the units that are not repaired ( $\min\{i, R\} - j$  units) are in the phases given by the column  $b$  after the repairs.

*Transition  $E_i \rightarrow E_i$*

We describe the transition  $E_i \rightarrow E_i$  in detail. This transition occurs when repairable failures and repairs occur at same number at one unit of time. The transition  $E_0 \rightarrow E_0$  occurs when all units are operational and at next time the system is in the same macro-state. It happens because 0 repairable and  $i$  non-repairable failures occur for  $i = 0, \dots, n$ . It is

$$B_{00} = \sum_{i=0}^n b(n; 0; i).$$

A similar reasoning can be performed for the transitions between the macro-states  $i$  units are broken. We assume that the system is working on with  $n - i$  operational units,  $i = 1, \dots, n - k$ . At next time, the system is in the same macro-state if there are  $r$  repairable failures and repairs and any number of non-repairable failures, for  $r = 0, \dots, \min\{i, R, n - i\}$ . The maximum number of repairable failures that can occur is the minimum among the operational units, the number of repairpersons and the broken units. If the stages are considered the probability is equal to  $b(n - i; r; n_r) \otimes rep(i; r)$  summing up over  $r$  and  $n_r$ . When a unit is repaired it begins working with initial distribution  $\alpha$ , and analogously, when a unit entries in the repair it occurs in phase  $j$  with probability  $\beta_j$ . It is introduced in the structure with the corresponding Kronecker product. In this case  $B_{ii}$  is equal to

$$B_{ii} = \sum_{r=0}^{\min\{i, R, n-i\}} \sum_{n_r=0}^{n-i-r} \alpha^{(r)} \otimes b(n - i; r; n_r) \otimes rep(i; r) \otimes \beta^{(r)},$$



for  $i = 1, \dots, n - k$  and being  $\beta^{(0)} = \alpha^{(0)} = 1$ ,  $\alpha^{(l)} = \alpha \otimes \dots \otimes \alpha$  and  $\beta^{(l)} = \beta \otimes \dots \otimes \beta$ .

If the system is broken, more or equal than  $n - k + 1$  units in repair, the rest of the operational units are stopped and transitions between repair stages occur without final repairing. Thus, in this case it is equal to

$$B_{ii} = I \otimes \overset{n-i}{\dots} \otimes I \otimes S \otimes \overset{\min\{R,i\}}{\dots} \otimes S,$$

for  $i = n - k + 1, \dots, n$  and being  $I$  the identity matrix with appropriate order.

The other transitions can be built in an analogous form. The blocks are shown.

*Transition  $E_i \rightarrow E_j$  with  $j > i$*

- For  $i = 1, \dots, n - 2$ ;  $j = i + 1, \dots, n - 1$  with  $i < n - k + 1$

$$B_{ij} = \sum_{r=0}^{\min\{i,R,n-j\}} \sum_{n_r=0}^{n-j-r} \alpha^{(r)} \otimes b(n-i; j-i+r; n_r) \otimes rep(i; r) \otimes \beta^{(\max\{r, \min\{j, k-i+r\}\})},$$

being  $B_{ij} = \mathbf{0}$  if  $i > n - k + 1$ .

- For  $i = 1, \dots, n - 1$  with  $i < n - k + 1$

$$B_{in} = b(n-i; n-i; 0) \otimes S \overset{\min\{i,R\}}{\dots} \otimes S \otimes \beta^{(\max\{0, R-i\})}$$

being  $B_{in} = \mathbf{0}$  if  $i \geq n - k + 1$ .

- For  $1 \leq j \leq n - 1$

$$B_{0j} = \sum_{n_r=0}^{n-j} b(n; j; n_r) \otimes \beta^{(\min\{j,R\})}$$

$$B_{0,n} = p^n T^0 \otimes \overset{n}{\dots} T^0 \otimes \beta^{(R)}$$

being  $T^0 = \mathbf{e} - T\mathbf{e}$  and  $\mathbf{e}$  a column vector of ones with appropriate order.

*Transition  $E_i \rightarrow E_j$  with  $j < i$*

- For  $i = 1, \dots, n-1$ ;  $j = \max\{i-R, 0\}, \dots, i-1$  with  $i \leq n-k$

$$B_{ij} = \sum_{r=0}^{\min\{n-i, R-i+j, j\}} \sum_{n_r=0}^{n-i-r} \alpha^{(i-j+r)} \otimes b(n-i; r; n_r) \otimes rep(i; i-j+r) \\ \otimes \beta^{(\min\{i-j+r, \max\{i-r, 0\}+r\})},$$

On the other hand if  $i > n-k$  then

$$B_{ij} = \alpha^{(i-j)} \otimes I \otimes \overset{n-i}{\dots} \otimes rep(i; i-j) \otimes \beta^{(\min\{i-j, \max\{i-R, 0\}\})}$$

- For  $j = n-R+1, \dots, n-1$

$$B_{nj} = \alpha^{(n-j)} \otimes rep(n; n-j) \otimes \beta^{(\min\{n-R, n-j\})} \\ B_{n, n-R} = \alpha^{(R)} \otimes S^0 \otimes \overset{R}{\dots} \otimes S^0 \otimes \beta^{(\min\{R, n-R\})}$$

being  $S^0 = \mathbf{e} - S\mathbf{e}$ .

### 3 Performance measures

Several measures associated to the system described above, such as the reliability and the availability, are calculated. We focus in the analysis of the conditional probability of failure of the system.

#### 3.1 Transition probabilities

Let  $\epsilon = (\alpha \otimes \overset{n}{\dots} \otimes \alpha, \mathbf{0})$  be the initial distribution of the system, considering that the system is new initially. The transition probabilities can be worked out given the initial distribution and the transition probability matrix given in (1). The probability that there are  $j$  units in repair at time  $\kappa$ , given the structure of the state space, can be computed. We denote this probability vector through  $p_j^\kappa$  and it is equal to the vector  $\epsilon P^\kappa$  limited to the elements of the phases of the macro-state  $E_j$  for  $j = 0, \dots, n$ .

#### 3.2 Availability and Reliability

The availability is the probability that the system is working at a certain time. This measure is denoted through  $A(\kappa)$  and it is equal to

$$A(\kappa) = \sum_{j=0}^{n-k} p_j^\kappa \mathbf{e}.$$

The reliability function is the probability that the system continues working at a certain time without a previous break. The distribution of this random time is PH with representation  $\left(\left(\alpha \otimes \cdots \otimes \alpha, \mathbf{0}\right), P^*\right)$ , being  $P^*$  the transition probability matrix given in (1) limited to the macro states  $E_0, \dots, E_{n-k}$ .

### 3.3 Conditional probability of failure of the system

For analysing the behaviour of a repairable system, it is interesting to know the probability of failure of this one at certain time if the system is working on. Thus, if  $j \leq n - k$ , we define

$$sf_j^\kappa = Pr \left( \begin{array}{l} \text{system fails at time } \kappa \text{ given that at time } \kappa - 1 \\ \text{the system is working on with } j \text{ units in repair} \end{array} \right)$$

This probability is equal to

$$sf_j^\kappa = p_j^{\kappa-1} \cdot w_j,$$

where  $w_j$  is a column vector whose entries are the probabilities of failure from each state of the macro-state  $E_j$  in lexicographical order.

The probability that the system fails at time  $\kappa$  is achieved summing up over  $j$  and it is equal to

$$sf^\kappa = \sum_{j=0}^{n-k} sf_j^\kappa.$$

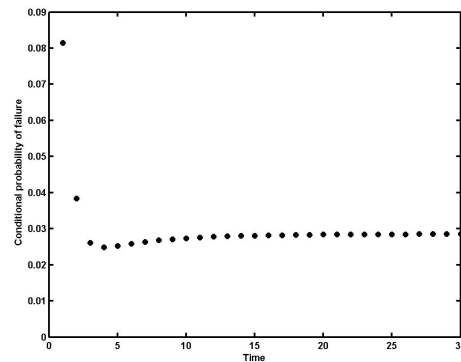
## 4 Numerical Application

We assume a system with five units that is working on while at least four units are operational: 4-out-of-5:  $G$  system. Each unit can occupy three operational states. The lifetime of each unit is phase type distributed with representation  $(\alpha, T)$  being

$$\alpha = (1, 0, 0) ; T = \begin{pmatrix} 0.2 & 0.5 & 0.2 \\ 0 & 0.6 & 0.39 \\ 0 & 0 & 0.98 \end{pmatrix}.$$

The repair time is also PH distributed. We assume two possible stages of repairing and the PH representation is given by  $(\beta, S)$  being

$$\beta = (1, 0) ; S = \begin{pmatrix} 0.6 & 0.2 \\ 0.2 & 0.7 \end{pmatrix}.$$



**Fig. 1.** Conditional probability of failure of the system

We assume that all failures are repairable and if a unit undergoes a failure, this one goes to repair. There are two repairpersons. Some performance measures of interest for this system have been built.

The conditional probability of failure has been computed and plotted. This one for the system is shown in Figure 1.

### Acknowledgements

This paper is supported by the Junta de Andalucía, Spain, under the Grant FQM-307.

### References

- KRISHNAMOORTHY, A. and USHAKUMARI, P.V. (2001): k-out-of-n: G system with repair: the D-policy. *Computers & Operations Research* 28, 973-981.
- MOUSTAFA, M.S. (2001): Availability of K-out-of-N:G Systems with Exponential Failures and General Repairs. *Economic Quality Control* 16 (1), 75-82.
- NEUTS, M. F. (1981): *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Johns Hopkins University Press, Baltimore.
- PHAM, HOANG (2008): *Recent Advances in Reliability and Quality in Design*. Springer.
- RUIZ-CASTRO, J.E., FERNÁNDEZ-VILLODRE, G. and PÉREZ-OCÓN, R. (2009a): A level-dependent general discrete system involving phase-type distributions. *IIE Transactions* 41 (1), 45-56.
- RUIZ-CASTRO, J.E., FERNÁNDEZ-VILLODRE, G. and PÉREZ-OCÓN, R. (2009b): A Multi-Component General Discrete System Subject to Different Types of Failures with Loss of Units. *Discrete Event and Dynamic Systems* 19 (1), 31-65.

# Using Logitboost for Stationary Signals Classification

Pedro Saavedra, Angelo Santana, Carmen Nieves Hernández, Juan Artiles,  
and Juan-José González

Departamento de Matemáticas. Universidad de Las Palmas de Gran Canaria  
35017 Las Palmas de Gran Canaria. Spain *saavedra@dma.ulpgc.es*

**Abstract.** The use of Boosting in conjunction with decision trees has been shown to be an effective method for classification problems characterized by high dimensionality. We propose using Boosting for classification of signals generated by stationary processes with mixed spectra, taking as features vector the ordinates of the components of the spectral distribution obtained at Fourier frequencies. The proposed method is evaluated by means of two studies with simulated data and a third study with real data from electro-encephalogram (EEG) signals measured on healthy and epileptic subjects in seizure-free intervals. The error rates are compared with those obtained from another proposed classification method in the literature.

**Keywords:** logitboost, stationary signals

## 1 The problem of classifying stationary signals

Signal classification problems are common in the field of biomedical sciences. In many cases the observed signals can be considered stationary, but the assumption of its spectral distribution being absolutely continuous is often unrealistic. A better modelling is achieved by considering mixed spectral distributions. Pardey et al (1996), in a review of modelling techniques for EEG analysis, consider mixed spectral distributions for this class of signals. Bhansali (1979) use mixed spectra to analyse the annual record of the number of Canadian lynx trapped in the Mackenzie River district of North-West Canada for the period 1821-1934 (Canadian lynx data set).

It is also unrealistic to assume that signals corresponding to subjects of the same population are all generated by the same stationary process. Diggle and Al-Wasel (1997) showed that the time series corresponding to levels of LH hormone in blood samples from subjects of a given population can not be considered realizations of a same stationary process. Instead, they suggested a random effects model based on the asymptotic representation of the periodogram of linear processes.

In this paper we deal with the problem of discriminating between two groups of stationary time series, assuming that within each group, the time series are generated by a random effects model with a possibly mixed spectrum (Koopman et al. (1995), Saavedra et al. (2008)). This description is

included in Section 2. With this aim in Section 3 we propose a LogitBoost classifier based on a features vector formed by the ordinates of an estimate of the spectral distribution of each time series (Friedman et al. (2000)). A feature preselection based on the ideas of Park et al. (2001) is carried out in order to reduce the dimensionality of the feature vector, thus alleviating computational burden and obviating the overfitting problem. To test the efficiency of our classification method, in Section 4 we conduct four studies with simulated data and one study with real EEG data available on line (Andrzejak et al. (2001)), corresponding to healthy subjects and epileptic ones, being the signals from epileptic recorded during seizure free intervals in the epileptogenic zone of the brain. In the literature, several classification methods can be found based on obtaining spectral disparity measures in the frequency domain using Kulback-Leibler discrepancy (Shumway (1974)). The results of our tests show the superiority of the proposed method in a comparison with one of the latest methods in this category, the one developed by Vilar and P rtega (2004).

## 2 Random effects model for the set of time series

We consider a population of objects  $A$ , such that on each  $a \in A$  a stationary process  $X_t(a)$  can be observed. We assume that this process is of the form:

$$X_t(a) = \sum_{j=1}^p R_{a,j} \cos(t\lambda_{a,j} + \phi_{a,j}) + Y_t(a) \quad (1)$$

where  $\lambda_{a,j}$  and  $R_{a,j}$  are random variables such that  $-\pi < \lambda_{a,j} \leq \pi$ ,  $R_{a,j} > 0$  and conditionally to  $a \in A$ ,  $\phi_{a,j}$  are independent and uniformly distributed random variables on the interval  $[-\pi, \pi]$ . Moreover,  $Y_t(a)$  is a second order stationary process with absolutely continuous spectral distribution, being  $\{f_a(\omega) : |\omega| \leq \pi\}$  the set of spectral density functions. In general, the doubly stochastic process  $\{Y_t(a) : t \in \mathbb{Z}\}$  can be represented as a linear process with random coefficients. Saavedra et al. (2008) studied this class of processes. Thus, each spectral distribution function  $F_a(\omega) : |\omega| \leq \pi$  can be considered as a realization of a stochastic process on the space  $A$ . In addition, it can be expressed by:

$$F_a(\lambda) = \sum_{\omega \leq \lambda} d_a(\omega) + \int_{-\pi}^{\lambda} f_a(\omega) d\omega \quad (2)$$

Here,  $d_a(\omega)$  are the so called spectrum lines which takes on the form  $d_a(\lambda) = R_{a,j}^2/2$  if  $\lambda = \pm\lambda_{a,j}$  and  $d_a(\lambda) = 0$  otherwise.

### 3 LogitBoost classifier

In this section we propose a classifier for a partition:

$$A = A_0 \cup A_1 \quad (A_0 \cap A_1 = \emptyset)$$

of the population under study, based on a data set of the form:

$$\{(X_t(a_i), y_i) : i = 1, \dots, n ; t = 1, \dots, T\} \quad (3)$$

being  $a_1, \dots, a_n$  a random sample of objects of  $A$ ,  $X_t(a_i)$  a time series consisting of the realization of a stationary process of the form (1) on the object  $a_i$  and finally,  $y_i \in \{1, -1\}$  the binary variable indicating the group membership of the object ( $y_i = 1$  or  $-1$  depending on  $a_i \in A_1$  or  $a_i \in A_0$ ). In epidemiological studies,  $A_1$  and  $A_0$  often represent cases and controls for a specific pathology.

#### 3.1 Spectral estimation.

From the data of the form (3), we calculate the periodogram of each object  $a \in A$ , which is defined by:

$$I_a^{(T)}(\lambda) = \frac{1}{2\pi T} \left| \sum_{t=1}^T X_t(a) \exp(-i\lambda t) \right|^2 : -\pi \leq \lambda \leq \pi$$

From the set of periodograms and using the procedure proposed by Kooperberg et al. (1995), we estimate the discrete and absolutely continuous components of the spectral distribution. In what follows, we represent the estimate of  $\sum_{\omega \leq \lambda} d_a(\omega)$  by  $\hat{D}_a(\lambda)$  and that of the  $f_a(\lambda)$  by  $\hat{f}_a(\lambda)$ .

#### 3.2 Feature vector.

In order to construct the classifier, we consider the following feature vector:

$$\mathbf{v}_i = \left( \hat{D}_{a_i}(\omega_1), \dots, \hat{D}_{a_i}(\omega_{[T/2]}) ; \hat{f}_{a_i}(\omega_1), \dots, \hat{f}_{a_i}(\omega_{[T/2]}) \right)$$

namely, the ordinates of the estimations of the components of the spectral distributions evaluated at the Fourier frequencies.

#### 3.3 Feature pre-selection.

Note that the dimension of the feature vector is  $2 \cdot [T/2]$ , which means that for a signal observed at a high sampling rate (for example,  $T = 1000$ ), the number of data  $n$  is usually lesser than the dimension of the discriminant vector. However, many components of the features have no discriminant power. A pre-selection of variables can then be carried out in order to reduce the dimensionality of the features vector, thus alleviating computational burden and obviating the overfitting problem. To this aim we use the method based on the nonparametric Wilcoxon test for two samples by Park et al. (2001).

### 3.4 LogitBoost with decision trees.

Boosting procedures introduced by Freund and Schapire (1997) are a powerful classification technique, especially in high dimension. Its aim is to produce an accurate combined classifier from a sequence of weak classifiers. We use here the LogitBoost procedure introduced by Friedman et al. (2000): it relies on the binomial log-likelihood as loss function. We use in addition the stumps as weak classifiers, which are decision trees with only two terminal nodes (see Breiman et al. (1984)).

## 4 Numerical study

We illustrate the procedure proposed in the previous section by means of four simulation studies and a fifth study with real data, using electroencephalogram (EEG) records from healthy and epileptic subjects, as described in Andrzejak et al. (2001). In all cases, the dataset was split into a training data set and a validation set. The proposed classification method was compared with the one by Vilar and P ertega (2004) which uses discrepancy measures based on the Kullback-Leibler discrimination information rate. They assume that all series are generated by stationary Gaussian processes with an absolutely continuous spectral distribution, and thus, groups  $A_1$  and  $A_0$  are characterized by its spectral density functions  $f_1(\lambda)$  and  $f_0(\lambda)$ . They then consider the following classifier: an object with spectral density associated  $f_a(\lambda)$  is assigned to the class whose spectral density has minimal discrepancy with  $f_a(\lambda)$ . The authors propose nonparametric estimates for spectral densities.

### 4.1 Simulation studies.

All series in simulation studies from this section were generated by stationary processes of the form (1) being  $Y_t(\mathbf{a}) = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t + a_1 \varepsilon_{t-1} + a_2 \varepsilon_{t-2}$ , with  $\varepsilon_t$  being a standard Gaussian white noise and  $\mathbf{a} = (a_1, a_2)'$  the vector of random coefficients which has a bivariate probability distribution  $N_2(\mu, \mathbf{C})$ . In each study we consider a group of cases and a group of controls. In the first study we consider that in both groups the spectral distributions contain only the absolutely continuous component. In the second and third studies, the processes are characterized by properly mixed spectra, with exactly the same continuous components in the third study, and similar but different continuous components in the second. In the fourth study we have considered an absolutely continuous spectral distribution for series corresponding to cases and a mixed spectral distribution for controls (both with the same continuous part). The parameters used for these simulations are summarized in Table 1.

In all cases the covariance matrix

$$\mathbf{C} = \begin{pmatrix} 1/3 & 1/6 \\ 1/6 & 1/3 \end{pmatrix}$$



Study	Spectral Component	Group	
		$A_1$ (Cases)	$A_0$ (Controls)
I	Absolutely Continuous	$\beta_1 = 0.9; \beta_2 = -0.5$ $\mu_1 = (2; 1)$	$\beta_1 = 0.5; \beta_2 = -0.5$ $\mu_0 = (2; 2)$
	Singular	$R \cong N_3((16; 20; 12), \mathbf{I}_3)$ $\lambda = (0.38; 0.18; 0.025)$	$R \cong N_4((5; 6; 7; 10), \mathbf{I}_4)$ $\lambda = (0.38; 0.12; 0.6; 0.025)$
II	Absolutely continuous	$\beta_1 = 0.9; \beta_2 = -0.5$ $\mu_1 = (2; 1)$	$\beta_1 = 0.5; \beta_2 = -0.5$ $\mu_0 = (2; 2)$
	Singular	$R \cong N_3((16; 20; 12), \mathbf{I}_3)$ $\lambda = (0.38; 0.18; 0.025)$	$R \cong N_4((5; 6; 7; 10), \mathbf{I}_4)$ $\lambda = (0.38; 0.12; 0.6; 0.025)$
III	Absolutely continuous	$\beta_1 = 0.9; \beta_2 = -0.5$ $\mu_1 = (2; 1)$	$\beta_1 = 0.9; \beta_2 = -0.5$ $\mu_0 = (2; 1)$
	Singular		$R \cong N_4((5; 6; 7; 10), \mathbf{I}_4)$ $\lambda = (0.38; 0.12; 0.6; 0.025)$
IV	Absolutely continuous	$\beta_1 = 0.9; \beta_2 = -0.5$ $\mu_1 = (2; 1)$	$\beta_1 = 0.9; \beta_2 = -0.5$ $\mu_0 = (2; 1)$

**Table 1.** Parameters of the simulated signals

was considered (which guarantees heterogeneity within classes for the absolutely continuous component). In each simulation, two hundred and fifty signals were simulated per group. Of these 500 objects, 333 were randomly selected to construct the LogitBoost classifier (training) and the remaining 167 were reserved for validation (estimation of classification error rates by the method of Vilar & Pertega and the proposed logitboost method). In each study this procedure was repeated 200 times. The simulation study was performed entirely using the R package, version 2.10. The results of the simulation studies are summarized in Table 2, where we show the observed mean values and standard deviations of classification error rates.

## 4.2 Diagnosis of epilepsy.

Epilepsy is one of the most common neurological disorders. The epilepsy is characterized by a sudden and recurrent malfunction of the brain, which is termed "seizure." Epileptic seizures reflect the clinical signs of an excessive and hypersynchronous activity of neurons in the brain. Traditionally, suspected seizures are evaluated using a routine electroencephalogram (EEG). An automated classification system can thus be of great interest for discriminating EEG activity in healthy subjects from epileptic ones, and in these to distinguish between seizure and seizure-free periods. We used the dataset described in Andrzejak et al. (2001), which is publicly available. This dataset includes five subsets (denoted as Z, O, N, F and S) each containing 100 single-channel EEG segments, each one having 23.6 second duration. For our study we have used subsets Z (acquired using surface EEG recording of five healthy volunteers with eyes open) and N (corresponding to signals measured

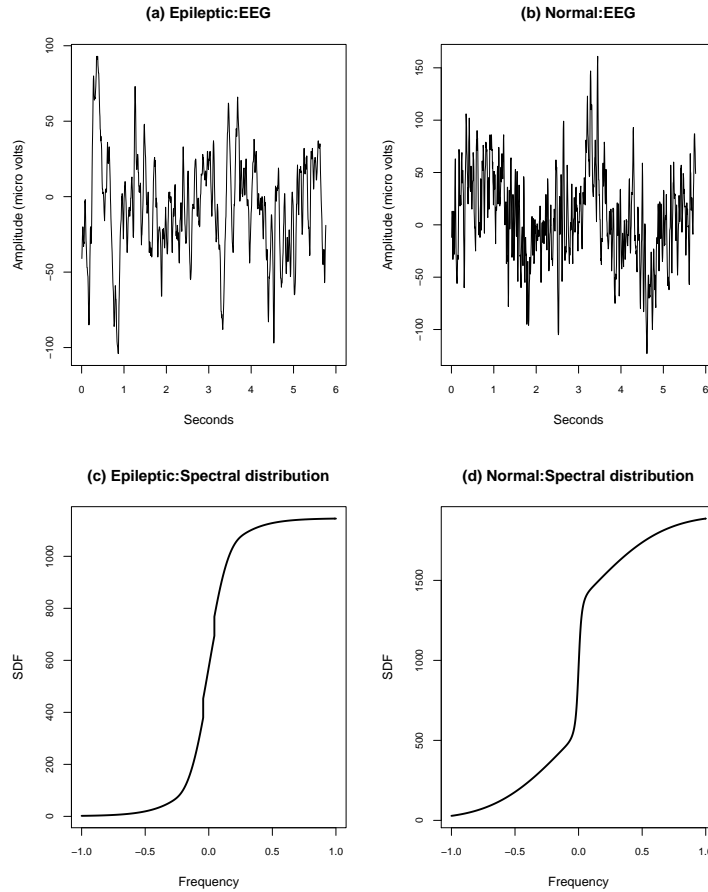
during seizure-free intervals from five epileptic patients in the hippocampal formation of the brain). The segments were selected and cut from multichannel records that were collected after a visual inspection of artifacts such as muscle activity or eye movement. Figure 1 shows records from a signal corresponding to an epileptic subject measured during seizure free intervals (1.a) and from a normal subject (1.b). Figures 1.c and 1.d show the corresponding estimations of the spectral distribution functions. It can be easily seen that both spectral distributions contain atoms with positive mass. In fact, 107 of the 200 signals analyzed, have an estimated spectral distribution containing atoms. For evaluating the classification error rate we repeated 200 times the procedure of randomly splitting the set of 200 EEG records into 133 records for training and the remaining 67 for testing. On each test set the classification error rate was calculated using the method of Vilar & Pertega and the proposed logitboost method. The averaged error rate and its standard deviation are also summarized in Table 2.

Study	Classification Method	
	Vilar & Pertega	LogitBoost
I	0.0667 (0.0220)	0.0287 (0.0125)
II	0.1113 (0.0252)	0.0162 (0.0099)
III	0.4864 (0.0120)	0.0608 (0.0197)
IV	0.1878 (0.0285)	0.0124 (0.0099)
EEG	0.1999 (0.0757)	0.0438 (0.0507)

**Table 2.** Evaluation of boosting classification method compared with that of Vilar and P ertega (V & P) for the simulation studies and real EEG signals. Mean classification error rates (sd) are shown.

## 5 Conclusions

Classical works on signals discrimination assume that these signals are generated by stationary Gaussian processes with an absolutely continuous spectral distribution, and the construction of the classifier relies heavily on this assumption. A drawback of this assumption is a dramatic reduction in the fields for potential application. In many cases this assumption fails since the spectrum distribution contains atoms, as we have seen in EEG records. By using the boosting classifiers, however, due to its ability for handling high dimensional vectors we are able to include more information about the signals,



**Fig. 1.** EEG signals corresponding to an epileptic subject (a) and to normal subject (b) and its corresponding spectral distribution functions (c) and (d).

namely the discrete and absolutely continuous components of the spectral distribution. Also, the decision trees which form the basis of the logitboost classifiers do not require additional hypothesis on the form of these distribution to accomplish their classification task.

In the first simulation study, signals were generated by Gaussian processes with an absolutely continuous spectral distribution, thus satisfying the conditions in which Vilar and Pértiga (V&P) propose for a classifier based on the Kullback-Leibler discrepancy. It can be observed (Table 1) that the two sets of series (cases and controls) are generated by very close spectral patterns. In this case, the mean classification rate of the V&P method is only 6.67%. However, the LogitBoost based method reduces this rate to 2.87%. When any of the spectral distributions has atoms (as is the case in the other simulation studies), the error rate of the V&P method is significantly worse (up to a

48.6% in the third study in which the continuous component is the same in cases and controls). This performance occurs because the classifier was designed for a different scenario in which the spectral distribution of signals is absolutely continuous. However the proposed method can detect the atoms in the signals and use them effectively in the classification, thereby obtaining a much lesser error rating. Note that the records of the EEG may be such that their spectral distributions have atoms. In this way, the proposed classifier provides predictions that are clearly better than the V&P method.

## References

- ANDRZEJAK R. G., LEHNERTZ K., MORMANN F., RIEKE C., DAVID P., ELGER C. E. (2001): Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. *Physical review. E, Statistical, nonlinear, and soft matter physics*; 64 (6 Pt 1):061907.
- BHANSALI, R.J. (1979): A Mixed spectrum analysis of the lynx data. *J. R. Statist. Soc. A*, 142, 199-209.
- BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., AND STONE, C. J., (1984): *Classification and regression trees*. Monterey, Calif., U.S.A.: Wadsworth, Inc.
- DIGGLE, P. J. AL-WASEL, I. (1997): Spectral Analysis of Replicated Biomedical Time Series. *Appl. Statist.*, 46, 31-71.
- FREUND, Y. AND SCHAPIRE, R. (1997): A decision-theoretic generalization of online learning and applications to boosting. *Journal of Computer and System Sciences*, 55, 249-266.
- FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2000): Additive logistic regression: a statistical view of boosting. *Annals of Statistical* 28, 337-407.
- KOOPERBERG, C., CHARLES J. STONE, C.J. AND TRUONG, Y. K. (1995): Log spline estimation of a possibly mixed spectral distribution. *Journal of Time Series Analysis*, 16, 359-388.
- PARDEY, J., S. ROBERTS, S. and L. TARASSENKO, L. (1996): A review of parametric modelling techniques for EEG analysis. *Med. Eng, Phys.* 18, 2-11.
- PARK, P., PAGANO, M. AND BONETTI, M. (2001): A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pac. Symp. Biocomput.*, 6, 52-63.
- R DEVELOPMENT CORE TEAM (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- SAAVEDRA, P., HERNÁNDEZ, C.N., LUENGO, I., ARTILES, J. AND SANTANA, A. (2008): Estimation of Population Spectrum for Linear Proceses with Random Coefficients. *Computational Statistics*, 23, 79-98.
- SHUMWAY, R. H., AND UNGER, A. N. (1974): Linear Discriminant Functions for Stationary Time Series. *Journal of the American Statistical Association*, 69, 948-956.
- VILAR, J. A. AND PÉRTEGA, S. (2004): Discriminant and cluster analysis for Gaussian stationary processes: local linear fitting approach. *Journal of Nonparametric Statistics*, 16, 443-462.

# Test of Mean Difference for Longitudinal Data Using Circular Block Bootstrap

Hirohito Sakurai and Masaaki Taguri

National Center for University Entrance Examinations  
2-19-23 Komaba, Meguro-ku, Tokyo 153-8501, Japan  
{sakurai, taguri}@rd.dnc.ac.jp

**Abstract.** This paper proposes a testing method for detecting the difference of two means or mean curves in longitudinal data using the circular block bootstrap. For the detection of mean difference, we consider four types of test statistics. Monte Carlo simulations are carried out in order to examine the sizes and powers of the proposed test.

**Keywords:** circular block bootstrap, resampling, longitudinal data, comparison of mean curves

## 1 Introduction

Suppose that there are two samples given by  $\{Y_i(t)\}_{i=1}^{q_1}$  and  $\{X_j(t)\}_{j=1}^{q_2}$  for  $t = 1, \dots, n$ , and assume that they are mutually independent, where  $q_1$  and  $q_2$  are numbers of subjects, and  $n$  is the number of observed points. We also assume that, for fixed  $t$ ,  $Y_i(t)$  and  $X_j(t)$  are independent over  $q_1$  and  $q_2$  subjects, respectively. Then we consider the model

$$\begin{cases} Y_i(t) = p_1(t) + \varepsilon_i(t), & i = 1, \dots, q_1, \\ X_j(t) = p_2(t) + \eta_j(t), & j = 1, \dots, q_2, \end{cases} \quad (1)$$

and define

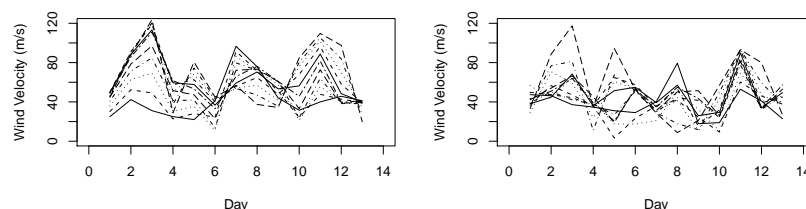
$$D_t = Y_t - X_t, \quad t = 1, \dots, n,$$

where  $p_1(t)$  and  $p_2(t)$  are unknown regression functions,  $Y_t = \sum_{i=1}^{q_1} Y_i(t)/q_1$ ,  $X_t = \sum_{j=1}^{q_2} X_j(t)/q_2$ , and  $\varepsilon_i(t)$  and  $\eta_j(t)$  are the error terms having means 0 and finite variances, respectively. More general formulations for (1) can be found, for example, in Hall and Hart (1990), Ferreira and Stute (2004), and references therein.

Let us now consider an example: Fig. 1 shows the wind velocity data measured by an artificial satellite (left panel) and a radar on the earth (right panel), and is obtained at altitudes from 80 km to 90 km every 1 km ( $(q_1 = q_2 =)11$  subjects) during 13 days ( $t = 1, \dots, 13$ ). For these data, we want to know whether the mean behavior of the two devices in measuring wind velocity is equal or not. The problem of our interest is then formulated as

$$H_0 : p_1(t) = p_2(t) \text{ for all } t \quad \text{vs.} \quad H_1 : p_1(t) \neq p_2(t) \text{ for some } t, \quad (2)$$

where  $H_0$  and  $H_1$  are the null and alternative hypotheses, respectively. The significant difference between  $p_1(t)$  and  $p_2(t)$  is detected by some methods, which is briefly summarized in Section 4.



**Fig. 1.** Wind velocity data (left: satellite, right: radar)

Several methods for the comparison of two means or regression curves have been proposed, and most of them assume that the error terms are independent and identically distributed (i.i.d.), and that their distributions are normal. However, if we cannot assume the normality, nonparametric methods are available for implementing the comparison. Graphical approach by Bowman and Young (1996) would be a possible choice.

Another approach could be an application of resampling such as nonparametric bootstrap. The bootstrap proposed by Efron (1979) has become the most popular resampling technique and is used to tackle a wide range of statistical problems. However, some modifications are needed for dependent data because the naive bootstrap ignores the order of observations when resampling is done. In order to overcome this problem, two kinds of bootstrap methods are proposed. One is a model-based approach which resamples from approximately i.i.d. residuals, and the other is a nonparametric, purely model-free bootstrap scheme, which resamples from blocks of observations. Among the latter, many papers on block bootstraps are proposed after Hall's (1985) pioneering work; see, for example, Bühlmann (2002), Härdle et al. (2003), Lahiri (2003), and references therein.

This paper concerns with the case where longitudinal data from two groups are given, and proposes a testing method for (2) using circular block bootstrap proposed by Politis and Romano (1992). As seen in the numerical examinations given below, our approach is superior to Bowman and Young's (1996) test.

The rest of this paper is constructed as follows. Section 2 proposes a testing procedure which generates the resamples corresponding to two samples by circular block bootstrap and calculates a  $p$ -value (achieved significance level). In order to investigate the properties of sizes and powers of the proposed testing method, Monte Carlo simulations are carried out in Section

3, and some concluding remarks and results of the above data analysis are summarized in Section 4.

## 2 Testing method

In this paper, our interest concentrates on the behavior of four types of test statistics given below, although there are some approaches to detecting the difference between two mean curves,  $p_1(t)$  and  $p_2(t)$ , in (1). The following statistic is proposed by Hall and Hart (1990):

$$S_n = S_n(D_1, \dots, D_n) = \left[ \sum_{j=0}^{n-1} \left( \sum_{t=j+1}^{j+g} D_t \right)^2 \right] \left[ n \sum_{t=1}^{n-1} \frac{(D_{t+1} - D_t)^2}{2} \right]^{-1}, \quad (3)$$

where  $D_t = Y_t - X_t$  for  $t = 1, \dots, n$  or  $D_t = Y_{t-n} - X_{t-n}$  for  $t = n + 1, \dots, n + g$ ,  $Y_t = \sum_{i=1}^q Y_i(t)/q$ ,  $X_t = \sum_{i=1}^q X_i(t)/q$ ,  $g = [np]$  is the integer part of  $np$ , and  $p$  is a tuning constant satisfying  $0 < p < 1$  which is determined by the fully data-driven approach; the second approach described in Hall and Hart (1990, pp.1043–1044). The statistic (3) is essentially based on kernel estimators of  $p_1(t)$  and  $p_2(t)$ . As another type of test statistics, we can consider

$$T_{1n} = T_{1n}(D_1, \dots, D_n) = \sum_{t=1}^n |D_t|, \quad (4)$$

$$T_{2n} = T_{2n}(D_1, \dots, D_n) = \sum_{t=1}^n D_t^2. \quad (5)$$

In addition to (3), (4) and (5), our attention also focuses on the area-difference expressed by  $A = \int |p_1(t) - p_2(t)| dt$ . Note that the quantity  $A$  is 0 under  $H_0$  and positive under  $H_1$ . Thus, the hypothesis of our interest reduces to testing

$$H_0 : A = 0 \quad \text{vs.} \quad H_1 : A > 0. \quad (6)$$

Then, we consider the following test statistic:

$$T_{3n} = T_{3n}(D_1, \dots, D_n) = \frac{1}{2} \sum_{t=1}^{n-1} (|D_t| + |D_{t+1}|) I_1 + \frac{1}{2} \sum_{t=1}^{n-1} \frac{|D_t|^2 + |D_{t+1}|^2}{|D_t| + |D_{t+1}|} I_2, \quad (7)$$

where  $I_1 = I\{D_t D_{t+1} \geq 0\}$ ,  $I_2 = I\{D_t D_{t+1} < 0\}$  and  $I\{\cdot\}$  is the indicator function, respectively. The statistic (7) is an estimator of  $A$  constructed by the trapezoidal rule with linear interpolations of adjacent observations.

The values of  $S_n$  and  $T_{rn}$  ( $r = 1, 2, 3$ ) will be small when  $H_0$  is true, and large when  $H_0$  is false. Therefore, the above four statistics enable us to measure the discrepancy between  $p_1(t)$  and  $p_2(t)$ .

In this section, we propose a nonparametric testing method for the problem (2) or (6) using (3), (4), (5) and (7). We call it “Mixed Circular Block Bootstrap (MCBB) Test.” The main ideas of MCBB test are to make blocks of observations in each sample similar to the circular block bootstrap, and to generate resamples corresponding to two samples by drawing blocks with replacement from the mixed circular blocks. This is motivated from the technique that can reflect the null hypothesis by resampling from a combined sample. The idea of combining observations of two samples and drawing resamples with replacement from the combined sample is previously considered by Boos et al. (1989) and Wang and Taguri (1996). The former is the test of homogeneity of scale, and the latter is that of equality of two means.

For simplicity, let  $T$  be a generic notation for statistics (3), (4), (5) or (7). For a given significance level  $\alpha$ , the proposed testing algorithm together with Monte Carlo method is described as follows.

- a. Calculate  $t_{obs} = T(Y, X) = T(D_1, \dots, D_n)$ .
- b. Put  $C_{y,t} = Y_t - \bar{Y}$  and  $C_{x,t} = X_t - \bar{X}$ , where  $\bar{Y} = \sum_{t=1}^n Y_t/n$  and  $\bar{X} = \sum_{t=1}^n X_t/n$ .
- c. Divide  $\{C_{y,1}, \dots, C_{y,n}\}$  and  $\{C_{x,1}, \dots, C_{x,n}\}$  into  $n$  collections of blocks,  $\xi_y = \{\xi_{y,1}, \dots, \xi_{y,n}\}$  and  $\xi_x = \{\xi_{x,1}, \dots, \xi_{x,n}\}$ , respectively, where blocks  $\xi_{y,t}$  and  $\xi_{x,t}$  are of length  $l$  and obtained in the manner of Politis and Romano (1992);

$$\xi_{y,t} = \begin{cases} \{C_{y,t}, \dots, C_{y,t+l-1}\}, & t = 1, \dots, n-l+1, \\ \{C_{y,t}, \dots, C_{y,n}, C_{y,1}, \dots, C_{y,t+l-n-1}\}, & t = n-l+2, \dots, n, \end{cases}$$

and

$$\xi_{x,t} = \begin{cases} \{C_{x,t}, \dots, C_{x,t+l-1}\}, & t = 1, \dots, n-l+1, \\ \{C_{x,t}, \dots, C_{x,n}, C_{x,1}, \dots, C_{x,t+l-n-1}\}, & t = n-l+2, \dots, n. \end{cases}$$

- d. Combine  $\xi_y$  and  $\xi_x$ , and put  $\xi_{pooled} = \{\xi_{y,1}, \dots, \xi_{y,n}, \xi_{x,1}, \dots, \xi_{x,n}\}$ .
- e. Draw  $\xi_y^{*b} = \{\xi_{y,1}^{*b}, \dots, \xi_{y,m}^{*b}\}$  and  $\xi_x^{*b} = \{\xi_{x,1}^{*b}, \dots, \xi_{x,m}^{*b}\}$  with replacement from  $\xi_{pooled}$  to obtain resamples  $Y^{*b} = \{Y_1^{*b}, \dots, Y_n^{*b}\}$  and  $X^{*b} = \{X_1^{*b}, \dots, X_n^{*b}\}$  ( $b = 1, \dots, B$ ), where  $m = n/l$  (if  $[n/l]$  is an integer) or  $m = [n/l] + 1$  (otherwise), and  $[n/l]$  is the integer part of a real  $n/l$ .
- f. Calculate  $t^{*b} = T(Y^{*b}, X^{*b}) = T(D_1^{*b}, \dots, D_n^{*b})$ .
- g. Repeating steps 5 and 6 an appropriate number of times  $B$ , calculate  $t^{*1}, \dots, t^{*B}$ .
- h. From steps 1 and 7, approximate the achieved significance level by  $\widehat{ASL} = \sum_{b=1}^B I\{t^{*b} \geq t_{obs}\}/B$ , and reject  $H_0$  when  $\widehat{ASL} \leq \alpha$ .

Note that, if  $l = 1$ , a resample of size  $n$  is generated with replacement from a centered sample. This reduces to the test proposed by Wang and Taguri (1996).



### 3 Numerical examination

In order to investigate the size and power properties of our testing procedure proposed in Section 2, we carry out Monte Carlo simulations including the comparison with Bowman and Young's (1996) test for unpaired data (hereafter termed "BY" for short). In the level and power studies, the nominal level is  $\alpha = 0.05$  and  $0.10$ . All our results are based on independent 2000 pairs of two samples,  $\{Y_i(t)\}$  and  $\{X_j(t)\}$ , where  $B = 2000$  replications of resampling are applied to every two samples in our test. Naturally, the same initial samples are used for the comparisons.

We generate initial two samples according to (1) whose means are specified by  $p_1(t) = 0$  and  $p_2(t) = c$ , where  $c = 0, 0.2, 0.4, 0.6, 0.8, 1.0$ ;  $c = 0$  or  $c \neq 0$  corresponds to the null hypothesis or the alternative hypothesis being true. The values,  $q_1, q_2$  and  $n$ , are  $(q_1, q_2) = (10, 10), (10, 20), (10, 30), (20, 20), (20, 30), (30, 30)$  and  $n = 10$ . The size of  $n$  may be small, however it is useful in practice to examine such a behavior of similar settings for the real data described in Section 1. As for the error terms,  $\varepsilon_i(t)$  and  $\eta_j(t)$ , we choose the following Gaussian AR(1) errors:  $\varepsilon_i(t) = \phi\varepsilon_i(t-1) + z_i(t)$  and  $\eta_j(t) = \phi\eta_j(t-1) + z_j(t)$ , where  $z_i(t) \stackrel{i.i.d.}{\sim} N(0, \tau_1^2)$ ,  $z_j(t) \stackrel{i.i.d.}{\sim} N(0, \tau_2^2)$ ,  $\phi = 0, \pm 0.1, \pm 0.2$ ,  $\tau_1^2 = \tau_2^2 = (1 - \phi^2)V(\varepsilon_i(t))$ , and  $V(\varepsilon_i(t)) = 1, 2, 3, 4, 5$ . The computation has been carried out for all combinations of the parameters described above, however, to save space, we restrict ourselves to discussing the case of  $\alpha = 0.05$  and  $V(\varepsilon_i(t)) = 3$ .

In MCBB test, it contains a block length parameter  $l$  to be estimated. Since it is preferable that the empirical level is nearly equal to the nominal level  $\alpha$ , our choice of  $l$  is then the length where the empirical level is close to  $\alpha$ . If there are some candidates which have the same level errors, we make the conservative choice, viz., we choose the block length such that the empirical level is less than the nominal level. Further if there are some candidates whose empirical levels are equal, we select the length where the empirical power is the highest among them. The resulting block lengths are summarized in Table 1.

Now, let us first summarize the results of the level studies. The empirical level of BY and our tests are given in Table 2. BY test shows a tendency to have large level errors. It underestimates the nominal level except for the case of  $(q_1, q_2) = (10, 10)$ . It is also observed that the empirical levels of MCBB test with  $T_{rn}$  ( $r = 1, 2, 3$ ) and  $S_n$  tend to keep the nominal level  $\alpha$ , however it is not true for most cases of  $\phi = 0.2$ .  $S_n$  does not need longer block length than  $T_{rn}$  ( $r = 1, 2, 3$ ) to keep the nominal level as is shown in Table 1. Most of the resulting block length for  $S_n$  is 1 or 2, while those for  $T_{1n}, T_{2n}$  and  $T_{3n}$  are greater than or equal to 3. On the other hand, MCBB for  $\phi \leq 0$  has a tendency to keep the nominal level, however the level error of MCBB for  $\phi > 0$  seems to be slightly larger.

Next, we discuss the power studies. Since we found similar tendencies among the six cases of  $(q_1, q_2)$ , we show the results for  $(q_1, q_2) = (10, 10)$ ,

(10, 20), (20, 20) with  $\phi = 0, 0.1, -0.2$ . The empirical power seems to be affected by the variance of noise, and the increase of noise variance causes

**Table 1.** Optimum block length in MCBB test for  $\alpha = 0.05$  and  $V(\varepsilon_i(t)) = 3$

$\phi$	$(q_1, q_2) = (10, 10)$					$(q_1, q_2) = (10, 20)$					$(q_1, q_2) = (10, 30)$				
	-0.2	-0.1	0	0.1	0.2	-0.2	-0.1	0	0.1	0.2	-0.2	-0.1	0	0.1	0.2
$T_{1n}$	1	1	9	3	3	1	1	1	3	2	1	1	1	1	8
$T_{2n}$	9	9	7	2	2	1	9	1	5	2	2	2	3	4	9
$T_{3n}$	8	4	2	2	2	9	7	2	2	7	3	4	1	9	9
$S_n$	2	2	1	2	2	3	2	2	2	2	2	2	1	1	2

$\phi$	$(q_1, q_2) = (20, 20)$					$(q_1, q_2) = (20, 30)$					$(q_1, q_2) = (30, 30)$				
	-0.2	-0.1	0	0.1	0.2	-0.2	-0.1	0	0.1	0.2	-0.2	-0.1	0	0.1	0.2
$T_{1n}$	9	9	9	3	3	1	9	9	2	3	2	1	1	1	4
$T_{2n}$	9	9	8	1	3	9	9	9	1	1	9	9	9	4	4
$T_{3n}$	9	7	1	3	2	9	4	1	1	7	9	3	1	4	4
$S_n$	2	2	1	1	2	2	2	1	2	2	2	2	1	1	2

**Table 2.** Empirical level for  $\alpha = 0.05$  and  $V(\varepsilon_i(t)) = 3$

	$\phi$	$(q_1, q_2) = (10, 10)$					$(q_1, q_2) = (10, 20)$				
		-0.2	-0.1	0	0.1	0.2	-0.2	-0.1	0	0.1	0.2
MCBB	$T_{1n}$	0.037	0.044	0.052	0.050	0.067	0.033	0.041	0.051	0.050	0.054
	$T_{2n}$	0.036	0.046	0.050	0.054	0.075	0.030	0.037	0.049	0.050	0.065
	$T_{3n}$	0.049	0.053	0.061	0.085	0.116	0.044	0.051	0.054	0.076	0.101
	$S_n$	0.048	0.051	0.053	0.072	0.095	0.039	0.046	0.054	0.072	0.091
BY		0.065	0.066	0.064	0.068	0.068	0.034	0.034	0.030	0.033	0.034

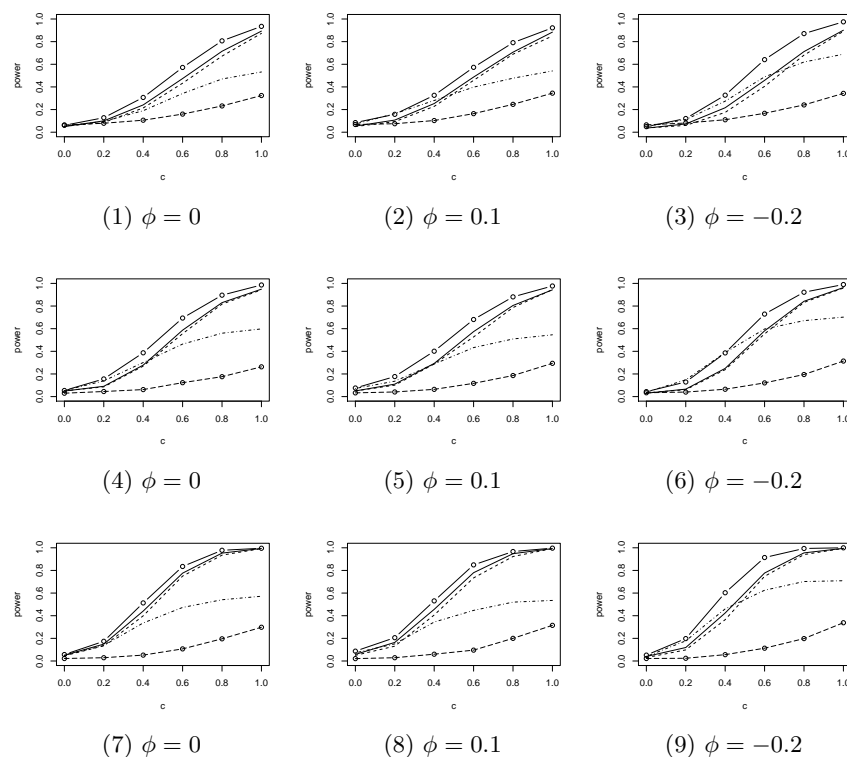
	$\phi$	$(q_1, q_2) = (10, 30)$					$(q_1, q_2) = (20, 20)$				
		-0.2	-0.1	0	0.1	0.2	-0.2	-0.1	0	0.1	0.2
MCBB	$T_{1n}$	0.019	0.025	0.037	0.051	0.051	0.030	0.039	0.046	0.050	0.068
	$T_{2n}$	0.023	0.029	0.041	0.050	0.053	0.040	0.047	0.051	0.058	0.077
	$T_{3n}$	0.030	0.049	0.051	0.072	0.092	0.052	0.051	0.057	0.087	0.118
	$S_n$	0.052	0.059	0.058	0.078	0.102	0.040	0.050	0.051	0.072	0.088
BY		0.041	0.041	0.041	0.040	0.038	0.022	0.022	0.022	0.021	0.022

	$\phi$	$(q_1, q_2) = (20, 30)$					$(q_1, q_2) = (30, 30)$				
		-0.2	-0.1	0	0.1	0.2	-0.2	-0.1	0	0.1	0.2
MCBB	$T_{1n}$	0.027	0.034	0.042	0.049	0.062	0.027	0.035	0.043	0.050	0.064
	$T_{2n}$	0.032	0.040	0.050	0.055	0.076	0.028	0.037	0.049	0.052	0.072
	$T_{3n}$	0.048	0.051	0.056	0.087	0.115	0.051	0.053	0.055	0.087	0.117
	$S_n$	0.041	0.053	0.059	0.075	0.088	0.050	0.058	0.051	0.075	0.092
BY		0.030	0.030	0.030	0.021	0.032	0.020	0.022	0.020	0.018	0.016

the decrease of empirical power as a whole. However, the power properties among the five variances considered in this section are nearly the same each other. Thus, we choose and discuss the case of  $V(\varepsilon_i(t)) = 3$ .

Fig. 2 shows the empirical powers corresponding to MCBB test with the four test statistics and BY test. We can observe that the empirical power of  $T_{3n}$  is most powerful among them, and that the overall relationship among powers corresponding to our and BY tests is given by  $T_{3n} \geq T_{2n} \geq T_{1n} \geq S_n \geq \text{BY}$ . This indicates the numerical superiority of our test using the four statistics in power. Especially, the superiority of  $T_{3n}$  in power can be confirmed from Fig. 2. As the number of subjects increases, the empirical power is improved. Within the values of  $0 \leq c \leq 1$ , the empirical power of  $T_{1n}$  is nearly equal to that of  $T_{2n}$ , however  $T_{2n}$  is slightly higher than  $T_{1n}$ .



**Fig. 2.** Empirical power for  $(q_1, q_2) = (10, 10), (10, 20), (20, 20)$ . The panels (1)–(3), (4)–(6) and (7)–(9) are the cases of  $(q_1, q_2) = (10, 10), (10, 20)$  and  $(20, 20)$ , respectively. MCBB with  $T_{1n}$ ,  $T_{2n}$ ,  $T_{3n}$  and  $S_n$ , and BY tests are corresponding to dashed, solid, solid with circles, dashed-dotted and dashed with circles lines, respectively.

## 4 Concluding remarks

In this paper we have proposed a testing method for detecting the difference of two means in longitudinal data based on the circular block bootstrap. Our numerical studies indicate the applicability of MCBB test for weakly dependent data even when the observed points are quite few. In some cases the effectiveness of application of block resampling could be confirmed.

Applying our MCBB with every possible block length and BY tests to the data given in Fig. 1, we obtain the results that MCBB with  $S_n$  and BY tests reject the null hypothesis. Therefore there is a possibility of the significant difference between the satellite and radar in measuring wind velocity. However, the problem on block length selection in the block resampling is very important, and the development of a fully data-driven approach to selecting block length in MCBB test will be needed for practical data analyses. It is also a future task to examine the accuracy of MCBB test.

## References

- BOOS, D., JANSSEN, P. and VERAVERBEKE, N. (1989): Resampling from centered data in the two sample problem. *Journal of Statistical Planning and Inference* 21 (3), 327–345.
- BOWMAN, A. and YOUNG, S. (1996): Graphical comparison of nonparametric curves. *Applied Statistics* 45 (1), 83–98.
- BÜHLMANN, P. (2002): Bootstraps for time series. *Statistical Science* 17 (1), 52–72.
- EFRON, B. (1979): Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7 (1), 1–26.
- FERREIRA, E. and STUTE, W. (2004): Testing for differences between conditional means in a time series context. *Journal of the American Statistical Association* 99, 169–174.
- HALL, P. (1985): Resampling a coverage pattern. *Stochastic Processes and their Applications* 20 (2), 231–246.
- HALL, P. and HART, J. D. (1990): Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association* 85, 1039–1049.
- HÄRDLE, W., HOROWITZ, J. and KREISS, J.-P. (2003): Bootstrap methods for time series. *International Statistical Review* 71 (2), 435–459.
- LAHIRI, S. N. (2003): *Resampling Methods for Dependent Data*. Springer, New York.
- POLITIS, D. N. and ROMANO, J. P. (1992): A circular block-resampling procedure for stationary data. In: R. LePage and L. Billard (Eds.): *Exploring the Limit of Bootstrap*. Wiley, New York, 263–270.
- WANG, J. and TAGURI, M. (1996): Bootstrap method — An introduction from a two sample problem (in Japanese). *Proceedings of the Institute of Statistical Mathematics* 44 (1), 3–18.

# An Empirical Study of the Use of Nonparametric Regression Methods for Imputation

Ismael R. Sánchez-Borrego, Maria Rueda and Encarnación Álvarez-Verdejo

Department of Statistics and Operational Research  
18071 University of Granada, Spain *ismasb@ugr.es*, *mrueda@ugr.es*,  
*encarniav@ugr.es*

**Abstract.** We address the problem of data incompleteness. A new algorithm based on Multivariate Adaptive Regression Splines is proposed to impute missing observations. A comparison with several imputations methods is addressed by considering missing at random (MAR) and missing completely at random (MCAR) missing data mechanisms. Two different ways of adding a disturbance to the imputation estimators are also addressed. A simulation study has been performed and a real-life data have been considered to illustrate the precision of the proposed method.

## 1 Introduction

Incomplete data sets can be encountered in a wide range of fields, including social and behavioral sciences, biological systems and computer vision. More examples can be found in clinical studies and in industrial and research databases, among others.

Although there exist several different methods to manage data with missing items, incomplete-data problems are often addressed with imputation: replacing missing data by filling in specific values. By treating these imputed values as true observations, traditional analysis may be carried out using the standard procedures developed for data without any missing observations.

Traditional single imputation methods have been widely used, such as the ratio imputation, multiple regression imputation, nearest neighbor imputation, respondent mean imputation or hot deck imputation (see e.g. Särndal and Lundström (2005), Schafer (1997), Little and Rubin (2002)) A comparison of several imputation methods is given e.g. in Montaquila and Ponikowski (1995) and in Nitter (2004). Iacus and Porro (2007) describe applications of regression tree to hot-deck missing data imputation. Ding and Simonoff (2010) consider popular missing data methods for classification tree algorithms applied to binary response. Other important related works in this area are those given by D'Ambrosio et al. (2007) and by Conversano and Siciliano (2009).

The first basic step begins by considering the random sample

$$(y_j, x_j, \delta_j), \quad j = 1, \dots, N, \quad (1)$$

where all the  $x_j$ 's are observed and  $\delta_j = 0$  if  $y_j$  is missing, otherwise  $\delta_j = 1$ .

By a purely nonparametric approach to (1), Chu and Cheng (1995) and Nitter (2004) among others, assume that the data are missing completely at random (MCAR). A relaxed version of MCAR is missing at random (MAR) (Cheng (1994) and Little and Rubin (1987), Ch. 1 among others). MAR assumes that there is a chance mechanism  $p(x)$ , such that

$$P(\delta = 1 \mid X = x, Y = y) = P(\delta = 1 \mid X = x) = p(x). \quad (2)$$

MCAR considers  $\delta$  independent of both  $X$  and  $Y$ , for example  $p(x)$  being a constant between 0 and 1.

Multiple regression estimation is one of the most commonly used imputation methods. Nevertheless, it states strong assumptions over the functional form of the underlying regression function. When those assumptions do not hold, nonparametric methods are more appropriate, as only smoothing assumptions over the regression function are made.

Multivariate Adaptive Regression Splines (MARS) were introduced by Friedman (1991) in the general context of multivariate nonparametric regression. MARS can automate variable selection and can handle a large number of independent variables (until  $k = 20$ ). We consider both MAR and MCAR missing data mechanisms to analyze the performance of the proposed imputation method.

The paper is organized as follows: Section 2 contains a description of the proposed method for estimating the population mean in the presence of missing data and Section 3 and Section 4 includes the simulation study and the numerical example to test the practical performance of the proposed method.

## 2 Proposed method

Let  $U = \{1, \dots, N\}$  be the population of  $N$  units from which a random sample  $s$  of fixed size  $n$  is drawn according to a specified sampling design  $d$ . Let  $y_j$  be the value of the response variable  $y$ , and  $x_{1j}, \dots, x_{kj}$  the corresponding values of the auxiliary variables  $x_1, \dots, x_k$ .

Let  $n_2$  be the number of missing values in a given sample  $s$  of size  $n$ . We denote  $s_r$  the set of observed data in  $s$ , of size  $n_1 = n - n_2$ . If  $j \in s_r$ ,  $y_j$  is an observed value of the study variable  $y$ . If  $j \in s - s_r$ , the observation  $y_j$  is unknown and will be estimated and denoted by  $\hat{y}_j$ .

We assume the following model:

$$y_j = m(x_{1j}, \dots, x_{kj}) + e_j, \quad j = 1, \dots, N, \quad (3)$$

where the  $e_j$ ,  $j = 1, \dots, N$  are independent and identically distributed with  $E(e_j) = 0$  and  $Var(e_j) = \sigma^2$  for  $j = 1, \dots, N$ . The unknown regression

function  $m$  is defined over  $D \subseteq \mathbb{R}^k$ . We use MARS technique to estimate this unknown regression function.

Based on the regression tree methodology (Morgan and Sunquist (1963) and Breiman et al. (1984)),  $\hat{m}(\mathbf{x})$  can be expanded

$$\hat{m}(\mathbf{x}) = \sum_{l=1}^L a_l B_l(\mathbf{x}), \quad (4)$$

where  $L$  is the number of basis functions,  $B_l$  take the form  $B_l(\mathbf{x}) = I[\mathbf{x} \in R_l]$ , where  $\{R_l\}_1^L$  are disjoint regions obtained by regression tree and  $I$  is the indicator function having the value one if its argument is true and zero otherwise.

The coefficients  $\{a_l\}$   $l = 1, \dots, L$  are fitted to the data by the MARS first algorithm. It provides new basis functions and improves the accuracy of the approximation by relying on truncated power splines functions instead of piecewise (continuous) functions as recursive partitioning regression does.

The MARS second algorithm removes basis functions that no longer contribute sufficiently to the accuracy of the fit.

The resulting MARS estimator after these two algorithms is a model of the form

$$\hat{m}(\mathbf{x}) = a_0 + \sum_{l=1}^L a_l \prod_{k=1}^{K_l} [s_{kl}(x_{v(k,l)} - t_{kl})]_+, \quad (5)$$

where  $t_{kl}$  are the knot locations,  $K_l$  the number of factors,  $s_{kl} = \pm 1$ ,  $v(k, l)$  label the predictor variables and  $a_l$  are the fitted coefficients of the basis function.

Now, we can calculate the "predicted" value  $\hat{y}_j = \hat{m}(x_j)$  for each missing value  $y_j$ .

The missing values can be replaced by the predicted values. If we use  $\hat{y}_j$  as the imputed value, this method artificially reduces the variance of the variable in question. To overcome the underestimated variance, we may add a small disturbance  $d_j$  drawn from a normal distribution with a zero mean and a variance obtained from the observed data (in the empirical study performed, we have used two methods to estimate the error variance  $\sigma^2$ ).

Finally, a complete set of observations  $s_c$  of size  $n$  is obtained, compounded of observed and estimated data and given by

$$y_j^* = \begin{cases} y_j & j \in s_r \\ \hat{y}_j + d_j & j \in s - s_r, \end{cases}$$

being  $\hat{y}_j = \hat{m}(\mathbf{x}_j)$  the MARS estimated value of  $y_j$ .

Once the missing observations have been estimated, we have a complete set of values  $y_j^*$ , thus standard complete-data methods of analysis can be used. In addition, parameters such as mean, variance and the distribution function can be estimated by using the values of the variable  $y$ .

We describe the algorithm including each required operation.

- Step 1.** Determine if there are missing values of the variable of interest.
- Step 2.** If Step 1 is true, then a selection of the auxiliary variables with completed observations is performed. Then, go to Step 4.
- Step 3.** If Step 1 is false, standard complete data methods of analysis can be used. Then, Halt.
- Step 4.** MARS estimation procedure: Based on the recursive partitioning regression methodology (Morgan and Sunquist (1963) and Breiman et al. (1984)), the estimated regression function can be expanded as a linear combination of the basis functions.
- Step 4a. MARS Algorithm 1:** Provide basis functions based on a piecewise continuous function.
- Step 4b. MARS Algorithm 2:** The coefficients of the expansion are fitted to the data providing new basis functions by relying on truncated power splines functions instead of piecewise continuous functions as MARS Algorithm 1 does.
- Step 4c. MARS Algorithm 3:** Removal of basis functions that no longer contribute sufficiently to the accuracy of the fit.
- Step 5.** Estimate the variance of errors ( $\sigma^2$ ) from the observed data:  $\hat{\sigma}^2 = \sum_{s_r} \frac{(y_j - \hat{m}(\mathbf{x}_j))^2}{n_1}$ .
- Step 6.** Impute the missing value  $y_j$   $j \in s - s_r$  for the predicted value plus the disturbance:  $\hat{y}_j + d_j$  where  $d_j$  is generated from a normal distribution with a zero mean and variance  $\hat{\sigma}^2$ .
- Step 7.** Build a complete matrix of observations  $y_j^*$ ,  $j \in s_c$ , then Halt.

### 3 Simulation study

In this section we perform a brief simulation to test the performance of the proposed methods under a practical point of view. We consider  $X$  a standard normal random variable, we consider model (3) with  $m(x) = x$  and piecewise linear missing function  $p(x) = 0.9 - 0.2|x|$  if  $|x| \leq 4.5$  and 0.1 elsewhere (taken from Cheng (1994)).  $R = 1000$  samples of size  $n = 100$  are drawn under simple random sampling without replacement (SRSWOR) and 10 %, 30 % and 50 % of missing values are generated missing at random (MAR) according to probabilities given by  $p(x)$ .

We study the practical behaviour of the proposed estimator for the population mean  $\bar{Y}$ . We note  $\hat{\theta}_{prop1}$  the above-mentioned proposed estimator. Another imputation estimator noted by  $\hat{\theta}_{prop2}$  can be derived by changing the way the disturbance  $d_j$  is calculated. We consider the normalized residuals  $d_j = \frac{(y_j - \hat{m}(\mathbf{x}_j))^2}{\hat{\sigma}^2} (1 - k/n_1)^{-1/2}$ ,  $j = 1, \dots, n_1$ . A random sample with replacement of size  $n_2$  from the residuals is drawn. Finally the missing value  $y_j$ ,  $j \in s - s_r$  is imputed by using the predicted value plus the selected residual  $d_j$ :  $\hat{y}_j + d_j$ .



We compare the performance of the proposed estimators with the local linear kernel regression estimator ( $\hat{\theta}_{LL}$ ) (Fan and Gijbels (1996), among others), with the addition of the same disturbance as  $\hat{\theta}_{prop1}$ . We also compare it with the regression estimator ( $\hat{\theta}_{reg}$ ), which assumes the functional form of the underlying regression function to be linear, with the hot-deck estimator ( $\hat{\theta}_{hd}$ ) and the nearest-neighbor estimator ( $\hat{\theta}_{NN}$ ).

To analyze the practical performance of the proposed estimator we consider the relative bias  $RB$  and the relative efficiency  $RE$  respect to  $\hat{\theta}_r$ , the estimator using only the observed units  $s_r$  in  $s$

$$RE(\hat{\theta}) = \frac{\sum_{i=1}^R (\hat{\theta}(s_i) - \bar{Y})^2}{\sum_{i=1}^R (\hat{\theta}_r(s_i) - \bar{Y})^2}, \quad RB(\hat{\theta}) = \frac{1}{R} \sum_{i=1}^R \frac{\hat{\theta}(s_i) - \bar{Y}}{\bar{Y}}. \quad (6)$$

being  $R$  the number of replications and  $\hat{\theta}$  is the population mean estimator considered.

(7)

The calculations and every estimator were obtained using the  $R$  program and the mda package. Programming details are available from the authors. Tables 1 and 2 shows the good performance of the proposed estimators when compared to other approaches. Regression estimator ( $\hat{\theta}_{reg}$ ) performs best when the regression model is well-specified. As  $m(x) = x$ , the regression estimator is a correct specification of the corresponding model and performs even better than the local linear kernel regression smoother. Nevertheless, when the model is misspecified, a superior efficiency can be gained by the nonparametric regression estimators. As the regression function is typically unknown, these estimators are likely to be a desirable choice for mean population estimation.

**Table 1.** Relative efficiency ( $RE$ ) for for the simulated population and sample size  $n = 100$ .

	$\hat{\theta}_r$	$\hat{\theta}_{prop1}$	$\hat{\theta}_{prop2}$	$\hat{\theta}_{LL}$	$\hat{\theta}_{reg}$	$\hat{\theta}_{hd}$	$\hat{\theta}_{nn}$
10%	1.00	0.75	0.66	0.81	0.69	0.78	0.70
30%	1.00	0.77	0.61	0.84	0.69	0.71	1.14
50%	1.00	0.46	0.36	0.50	0.44	0.42	1.57

## 4 Empirical study

In this section a numerical example illustrate the practical performance of the proposed estimator. We consider the real-life population of the Sugar

**Table 2.** Relative bias ( $RB$ ) for the simulated population and sample size  $n = 100$ .

	$\hat{\theta}_r$	$\hat{\theta}_{prop1}$	$\hat{\theta}_{prop2}$	$\hat{\theta}_{LL}$	$\hat{\theta}_{reg}$	$\hat{\theta}_{hd}$	$\hat{\theta}_{nn}$
10%	-0.277	-0.067	-0.050	-0.271	0.092	-0.479	-0.059
30%	-0.704	-0.015	-0.086	-0.247	0.053	-0.801	-0.334
50%	0.556	0.041	-0.030	-0.292	0.051	-0.222	-0.368

cane. The variable of interest is income from cane taken from 338 Sugar cane farms. The auxiliary variables are the costs and the area assigned for growing cane. It was used by Chambers and Dunstan (1986) and by Rao, Kovar and Mantel (1990).

We study the practical behaviour of the proposed estimator for the population mean  $\bar{Y}$ . We consider the above-mentioned estimators. Missing data are selected for each sample by simple random sampling without replacement (SRSWOR) under a missing completely at random (MCAR) mechanism and three proportions of missing values (constant values of  $p(x)$  in MCAR:  $p = 0.1$ ,  $p = 0.3$  and  $p = 0.5$ ) are considered.  $R = 1000$  samples of sizes  $n = 100$ ,  $n = 75$  and  $n = 50$  are drawn under simple random sampling without replacement.

Numerical results supports the easy applicability of the method and its advantageous properties. The proposed method is based on MARS, which is well-known for its easy applicability and advantageous properties. We may conclude that the proposed method may be a good alternative to other classical imputation estimators.

## References

- BREIDT, F.J., CLAESKENS, G. and OPSOMER, J.D. (2005) Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831–846.
- BREIDT, F.J. and OPSOMER, J.D. (2000) Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4), 1026–1053.
- BREIMAN, L., FRIEDMAN, J.H., OLHSEN, R.A. and STONE, C.J. (1984) *Classification and Regression Trees* Wadsworth, Belmont, California.
- CONVERSANO, C. and SICILIANO, R. (2009) Incremental Tree-Based Missing Data Imputation with Lexicographic Ordering *Journal of classification*, 26, 3, 361–379.
- D'AMBROSIO, A., ARIA, M. and SICILIANO, R. (2007) *Robust Tree-Based Incremental Imputation Method for Data Fusion*. Advances in Intelligent Data Analysis VII. Ed. Springer, Berlin/Heidelberg.
- DING, Y. and SIMONOFF, J.S. (2010) An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research*, 11, 131–170.

**Table 3.** Relative efficiency ( $RE$ ) for the Sugar cane population

$n$	$p$	$\hat{\theta}_r$	$\hat{\theta}_{prop1}$	$\hat{\theta}_{prop2}$	$\hat{\theta}_{reg}$	$\hat{\theta}_{hd}$	$\hat{\theta}_{nn}$
$n = 100$	$p = 0.1$	1.00	0.89	0.88	0.89	1.08	0.91
	$p = 0.3$	1.00	0.68	0.70	0.72	1.16	0.77
	$p = 0.5$	1.00	0.43	0.46	0.50	1.00	0.60
$n$	$p$	$\hat{\theta}_r$	$\hat{\theta}_{prop1}$	$\hat{\theta}_{prop2}$	$\hat{\theta}_{reg}$	$\hat{\theta}_{hd}$	$\hat{\theta}_{nn}$
$n = 75$	$p = 0.1$	1.00	0.91	0.90	0.91	1.08	0.93
	$p = 0.3$	1.00	0.69	0.70	0.70	1.00	0.78
	$p = 0.5$	1.00	0.45	0.49	0.52	1.01	0.62
$n$	$p$	$\hat{\theta}_r$	$\hat{\theta}_{prop1}$	$\hat{\theta}_{prop2}$	$\hat{\theta}_{reg}$	$\hat{\theta}_{hd}$	$\hat{\theta}_{nn}$
$n = 50$	$p = 0.1$	1.00	0.90	0.91	0.91	1.10	0.94
	$p = 0.3$	1.00	0.70	0.73	0.73	1.10	0.76
	$p = 0.5$	1.00	0.47	0.49	0.55	1.00	0.61

**Table 4.** Relative bias ( $RB$ ) for Sugar cane population

$n$	$p$	$\hat{\theta}_r$	$\hat{\theta}_{prop1}$	$\hat{\theta}_{prop2}$	$\hat{\theta}_{reg}$	$\hat{\theta}_{hd}$	$\hat{\theta}_{nn}$
$n = 100$	$p = 0.1$	-0.003	-0.001	-0.002	-0.002	-0.004	-0.003
	$p = 0.3$	-0.003	-0.001	-0.002	-0.003	-0.002	-0.007
	$p = 0.5$	-0.006	-0.002	-0.002	-0.003	-0.006	-0.012
$n$	$p$	$\hat{\theta}_r$	$\hat{\theta}_{prop1}$	$\hat{\theta}_{prop2}$	$\hat{\theta}_{reg}$	$\hat{\theta}_{hd}$	$\hat{\theta}_{nn}$
$n = 75$	$p = 0.1$	0.002	0.001	0.001	0.001	0.001	0.001
	$p = 0.3$	-0.002	-0.001	-0.001	-0.001	-0.002	-0.001
	$p = 0.5$	-0.004	0.000	-0.001	-0.001	0.005	-0.012
$n$	$p$	$\hat{\theta}_r$	$\hat{\theta}_{prop1}$	$\hat{\theta}_{prop2}$	$\hat{\theta}_{reg}$	$\hat{\theta}_{hd}$	$\hat{\theta}_{nn}$
$n = 50$	$p = 0.3$	0.001	0.001	0.001	0.000	0.004	-0.002
	$p = 0.3$	-0.001	-0.002	-0.002	-0.004	-0.003	-0.010
	$p = 0.5$	-0.001	0.000	-0.001	-0.002	-0.001	-0.013

- CHAMBERS, R.L., DORFMAN, A.H. and WHERLY, T.E. (1993) Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268–277.
- CHENG, P.E. (1994) Nonparametric Estimation of Mean Functionals with Data Missing at Random. *Journal of the American Statistical Association*, 89, 425, 81–87.
- CHU, C.K. and CHENG, P.E. (1995) Nonparametric regression estimation with missing data. *Journal of Statistical Planning and Inference*, 48, 85–99.
- EILERS, P.H.C. and MARX, B.D. (1996) Flexible smoothing with B-splines and penalties. *Stat. Science*, 11(2), 86–121.

- FAN, J. and GIJBELS, I. (1996) *Local Polynomial Modelling and Its Applications*. Ed. Chapman and Hall.
- FRIEDMAN, J.H. (1991) Multivariate Adaptive Regression Splines *The Annals of Statistics*, Vol. 19, No. 1 (Mar. 1991), pp. 1–67.
- HASTIE, T. (1996) Pseudosplines. *Journal of the Royal Statistical Society, Series B*, 58, 376–396.
- IACUS, S.M. and PORRO, G. (2007) Missing data imputation, matching and other applications of random recursive partitioning. *Computational Statistics and Data Analysis*, 52, 2, 773–789.
- LITTLE, R.J.A. and RUBIN, D. (2002) *Statistical Analysis with missing data*. Wiley.
- MARX, B.D. and EILERS, P.H.C. (1998) Direct generalized additive modelling with penalized likelihood. *Computational Statistical and Data Analysis*, 28, 193–209.
- MONTAQUILA, J.M. and PONIKOWSKI, C.H. (1995) An Evaluation Of Alternative Imputation Methods, *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- MORGAN, J.N. and SONQUIST, J.A. (1963) Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58, 415–434.
- NITTNER, T. (2004) The additive model affected by missing completely at random in the covariate. *Computational Statistics*, 19, 2, 261–282.
- RAO, J.N.K., KOVAR, J.G. and MANTEL, H.J. (1990) On estimating distribution functions and quantiles from survey data using auxiliary information, *Biometrika*, 77, 365–375.
- RUPPERT, D. and WAND, M. P. (1994) Multivariate locally weighted least squares regression *The Annals of Statistics*, 22(3), 1346–1370.
- SCHAFER, J. (1997) *Analysis of incomplete multivariate data*. Chapman Hall.
- SÄRNDAL, C.E. and LUNSTRÖM, S. (2005) *Estimation in Surveys with Nonresponse*. Wiley Series in Survey Methodology.
- SINGH, S. (2003), *Advanced sampling theory with applications: How Michael "selected" Amy*, pp. 1–1247. Kluwer Academic Publisher. The Netherlands.
- WAND, M. (2003) Smoothing and mixed models. *Computational Statistics*, 18, 223–249.

# A Simulation Study of the Bayes Estimator of Parameters in an Extension of the Exponential Distribution

Samira Sadeghi

Department of Mathematics, Statistics and Computer Sciences, University of Tehran, Iran, *samira.sadeghi1@gmail.com*

**Abstract.** Recently a generalization of the exponential distribution has been introduced by Nadarajah & Haghighi (2009). In this paper, we consider the Bayes estimators of the scale and shape parameters of this family under the assumptions of gamma priors and squared error loss function. We used the idea of Lindley for obtaining the approximate Bayes estimators. Under assumptions of non-informative priors, the approximate Bayes estimators are computed and compared with the corresponding maximum likelihood estimators. One real data set has been analyzed to demonstrate how the proposed method can be used in practice. For this data set, we also computed the approximate Bayes estimators using MCMC technique and compared the results.

**Keywords:** Bayes estimator, exponential distribution, Lindley approximation, MCMC, Monte Carlo simulation

## 1 Introduction

In survival analysis, times with monotone hazard rate are commonly modelled using Weibull and gamma distributions. Weibull is often preferred, for having a closed form survival function, contrary to gamma, which requires numerical integration. Gupta and Kundu's exponentiated exponential (EE) distribution (2001) has many properties in common with gamma, but has closed form survival and hazard rate functions. Nadarajah and Haghighi (2009) proposed a new distribution, as a competitor to the exponentiated exponential and Weibull distributions. The two-parameter Extension of the Exponential distribution has a decreasing probability density function (pdf), yet allowing for increasing, decreasing and constant hazard rates like EE and Weibull, and having closed form survival and hazard rate functions. Real data applications showed that it could perform better than EE and Weibull both in an industrial context (failure times) and in a natural one (interval between earthquakes in Iran over 20 years). The family has the following density function:

$$f(t) = \alpha\lambda(1 + \lambda t)^{\alpha-1} \exp\{1 - (1 + \lambda t)^\alpha\}, \quad t > 0, \quad \alpha > 0, \quad \lambda > 0. \quad (1)$$

The aim of this paper is to consider the problem of estimating the shape and scale parameters,  $\alpha$  and  $\lambda$ , of a family defined by (1). The maximum likelihood

estimators often provide satisfactory estimates of these parameters. However, maximum likelihood estimates turn out to be imprecise for small samples. In survival analysis and reliability studies, available samples are often small. Also, in many circumstances, there exists some knowledge on the underlying failure mechanism. Hence Bayesian methods can be expected to be useful in such cases. In this paper, we consider the Bayes estimators and compare their performance with the maximum likelihood estimators. In Bayesian inference, a prior distribution  $\pi(\theta)$  is specified and leads to the posterior distribution  $\pi(\theta|t) \propto L(t; \theta)\pi(\theta)$ . The Bayesian estimation of any function of interest  $g(\theta)$ , under the squared error loss function is the posterior mean,

$$E(g(\theta)) = \int g(\theta)\pi(\theta|t)d\theta.$$

It is observed that this integral cannot be expressed in an explicit form and numerical integration is required. We use the idea of Lindley to compute the approximate Bayes estimates under the assumption of non-informative prior and compare them with the MLE's by Monte Carlo simulations. For a real data set, we also used Markov Chain Monte Carlo (MCMC) method, the Gibbs Sampler, to generate samples from the posterior distribution and in turn compute the Bayes estimates.

## 2 Estimation of parameters

Suppose  $\{t_1, t_2, \dots, t_n\}$  is a random sample from (1). The likelihood function of  $\alpha$  and  $\lambda$  based on the observed data is

$$l(\alpha, \lambda|data) = \alpha^n \lambda^n \exp\left[\sum_{i=1}^n (1 - (1 + \lambda t_i)^\alpha)\right] \prod_{i=1}^n (1 + \lambda t_i)^{\alpha-1}. \quad (2)$$

### 2.1 Bayes estimators

When both the shape and scale parameters are unknown, the proposed independent priors for parameters may be taken as

$$\pi_1(\lambda) \propto \lambda^{b-1} e^{-a\lambda}, \quad \lambda > 0, \quad (3)$$

$$\pi_2(\alpha) \propto \alpha^{d-1} e^{-c\alpha}, \quad \alpha > 0. \quad (4)$$

respectively, to give the joint prior distribution for  $\lambda$  and  $\alpha$  as:

$$\pi(\alpha, \lambda) = \frac{a^b c^d}{\Gamma(d)\Gamma(b)} \alpha^{d-1} \lambda^{b-1} \exp\{-(a\lambda + c\alpha)\}, \quad a, b, c, d > 0. \quad (5)$$

The joint posterior density function of  $\alpha$  and  $\lambda$  can be written as

$$l(\alpha, \lambda|data) = \frac{l(data|\alpha, \lambda)\pi_1(\lambda)\pi_2(\alpha)}{\int_0^\infty \int_0^\infty l(data|\alpha, \lambda)\pi_1(\lambda)\pi_2(\alpha)d\alpha d\lambda}. \quad (6)$$

Under the SEL function the Bayes estimator of any function of  $\alpha$  and  $\lambda$ , say  $g(\alpha, \lambda)$ , is

$$E_{\alpha, \lambda | data}(g(\alpha, \lambda)) = \frac{\int_0^\infty \int_0^\infty g(\alpha, \lambda) l(data | \alpha, \lambda) \pi_1(\lambda) \pi_2(\alpha) d\alpha d\lambda}{\int_0^\infty \int_0^\infty l(data | \alpha, \lambda) \pi_1(\lambda) \pi_2(\alpha) d\alpha d\lambda}. \quad (7)$$

It may be noted that (7) cannot be reduced to a closed form and numerical approximations are needed. There exist many procedures producing such approximations. In this paper, we study the use of Lindley's approximation on simulated data and compare this method to a MCMC technique on a real data set.

**Lindley's approximation** Lindley (1980) considered an approximation for the ratio of integrals of the form

$$I = \frac{\int \omega(\theta) \exp[L(\theta)] d\theta}{\int v(\theta) \exp[L(\theta)] d\theta} \quad (8)$$

where  $\theta = (\theta_1, \theta_2, \dots, \theta_n)$  is the parameter,  $L(\theta)$  is the logarithm of the likelihood function,  $\omega(\theta)$  and  $v(\theta)$  are arbitrary functions of  $\theta$ . Let  $v(\theta)$  be the prior distribution of  $\theta$  and  $\omega(\theta) = g(\theta)v(\theta)$ . From (8) we have the posterior expectation

$$I = E[g(\theta) | x] = \frac{\int g(\theta) \exp[L(\theta) + \rho(\theta)] d\theta}{\int \exp[L(\theta) + \rho(\theta)] d\theta}. \quad (9)$$

where  $\rho(\theta) = Ln[(v\theta)]$ .

Lindley's expansion of (9) leads to

$$I = g(\hat{\theta}) + \frac{1}{2} \sum \left[ g_{ij}(\hat{\theta}) + 2g_i(\hat{\theta})\rho_j(\hat{\theta}) \right] \sigma_{ij} + \frac{1}{2} \sum L_{ijk}(\hat{\theta}) g_l(\hat{\theta}) \sigma_{ij} \sigma_{kl}$$

where

$$g_i = \frac{\partial g}{\partial \theta_i}, \quad g_{ij} = \frac{\partial^2 g}{\partial \theta_i \partial \theta_j}, \quad L_{ijk} = \frac{\partial^3 L}{\partial \theta_i \partial \theta_j \partial \theta_k}, \quad \rho_j = \frac{\partial \rho}{\partial \theta_j}, \quad \sigma_{ij} = E(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j).$$

Here, Lindley's approximation of (7) can be written as follows:

$$\begin{aligned} I = & g(\hat{\lambda}, \hat{\alpha}) + \frac{1}{2} [(\hat{g}_{\lambda\lambda} + 2\hat{g}_{\lambda}\hat{\rho}_{\lambda})\hat{\sigma}_{\lambda\lambda} + (\hat{g}_{\alpha\lambda} + 2\hat{g}_{\alpha}\hat{\rho}_{\lambda})\hat{\sigma}_{\alpha\lambda} + (\hat{g}_{\lambda\alpha} + 2\hat{g}_{\lambda}\hat{\rho}_{\alpha})\hat{\sigma}_{\lambda\alpha} + (\hat{g}_{\alpha\alpha} + 2\hat{g}_{\alpha}\hat{\rho}_{\alpha})\hat{\sigma}_{\alpha\alpha}] \\ & + \frac{1}{2} [(\hat{g}_{\lambda\lambda}\hat{\sigma}_{\lambda\lambda} + \hat{g}_{\alpha}\hat{\sigma}_{\lambda\alpha})(\hat{L}_{\lambda\lambda\lambda}\hat{\sigma}_{\lambda\lambda} + \hat{L}_{\lambda\alpha\lambda}\hat{\sigma}_{\lambda\alpha} + \hat{L}_{\alpha\lambda\lambda}\hat{\sigma}_{\alpha\lambda} + \hat{L}_{\alpha\alpha\lambda}\hat{\sigma}_{\alpha\alpha}) \\ & + (\hat{g}_{\lambda}\hat{\sigma}_{\alpha\lambda} + \hat{g}_{\alpha}\hat{\sigma}_{\alpha\alpha})(\hat{L}_{\alpha\lambda\lambda}\hat{\sigma}_{\lambda\lambda} + \hat{L}_{\lambda\alpha\alpha}\hat{\sigma}_{\lambda\alpha} + \hat{L}_{\alpha\lambda\alpha}\hat{\sigma}_{\alpha\lambda} + \hat{L}_{\alpha\alpha\alpha}\hat{\sigma}_{\alpha\alpha})] \end{aligned} \quad (10)$$

where  $(\theta_1, \theta_2) = (\alpha, \lambda)$ ,

$$\hat{g}_i = \frac{\partial g(\hat{\theta}_1, \hat{\theta}_2)}{\partial \hat{\theta}_i}, \quad \hat{g}_{ij} = \frac{\partial^2 g(\hat{\theta}_1, \hat{\theta}_2)}{\partial \hat{\theta}_i \partial \hat{\theta}_j}, \quad \hat{L}_{ijk} = \frac{\partial^3 L(\hat{\theta}_1, \hat{\theta}_2)}{\partial \hat{\theta}_i \partial \hat{\theta}_j \partial \hat{\theta}_k}, \quad \hat{\rho}_j = \frac{\partial \rho(\hat{\theta}_1, \hat{\theta}_2)}{\partial \hat{\theta}_j}, \quad \hat{\sigma}_{ij} = \frac{-1}{\hat{L}_{ij}}.$$

and  $\hat{\theta}$  is MLE of  $\theta$ .

Following Lindley's approximation, the approximate Bayes estimators of the parameters, when both are unknown, are obtained as follows:

$$\begin{aligned}\lambda_B = \hat{\lambda} &+ \frac{\left(\frac{b-1}{\hat{\lambda}}\right) - a}{\frac{n}{\hat{\lambda}^2} + \hat{\alpha}(\hat{\alpha}-1) \sum_{i=1}^n t_i^2 (1 + \hat{\lambda}t_i)^{\hat{\alpha}-2} + (\hat{\alpha}-1) \sum_{i=1}^n \frac{t_i^2}{(1+\hat{\lambda}t_i)^2}} \\ &+ \frac{\frac{2n}{\hat{\lambda}^3} - \hat{\alpha}(\hat{\alpha}-1)(\hat{\alpha}-2) \sum_{i=1}^n t_i^3 (1 + \hat{\lambda}t_i)^{\hat{\alpha}-3} + (\hat{\alpha}-1) \sum_{i=1}^n \frac{2t_i^3}{(1+\hat{\lambda}t_i)^3}}{2 \left[ \frac{n}{\hat{\lambda}^2} + \hat{\alpha}(\hat{\alpha}-1) \sum_{i=1}^n t_i^2 (1 + \hat{\lambda}t_i)^{\hat{\alpha}-2} + (\hat{\alpha}-1) \sum_{i=1}^n \frac{t_i^2}{(1+\hat{\lambda}t_i)^2} \right]^2} \\ &+ \frac{-\hat{\alpha} \sum_{i=1}^n t_i (1 + \hat{\lambda}t_i)^{\hat{\alpha}-1} \ln^2(1 + \hat{\lambda}t_i) - 2 \sum_{i=1}^n t_i (1 + \hat{\lambda}t_i)^{\hat{\alpha}-1} \ln(1 + \hat{\lambda}t_i)}{2 \left[ \frac{n}{\hat{\lambda}^2} + \hat{\alpha}(\hat{\alpha}-1) \sum_{i=1}^n t_i^2 (1 + \hat{\lambda}t_i)^{\hat{\alpha}-2} + (\hat{\alpha}-1) \sum_{i=1}^n \frac{t_i^2}{(1+\hat{\lambda}t_i)^2} \right] \left[ \frac{n}{\hat{\alpha}^2} + \sum_{i=1}^n (1 + \hat{\lambda}t_i)^{\hat{\alpha}} \ln^2(1 + \hat{\lambda}t_i) \right]}\end{aligned}$$

and

$$\begin{aligned}\alpha_B = \hat{\alpha} &+ \frac{\frac{d-1}{\hat{\alpha}} - c}{\frac{n}{\hat{\alpha}^2} + \sum_{i=1}^n (1 + \hat{\lambda}t_i)^{\hat{\alpha}} \ln^2(1 + \hat{\lambda}t_i)} + \frac{\frac{2n}{\hat{\alpha}^3} - \sum_{i=1}^n (1 + \hat{\lambda}t_i)^{\hat{\alpha}} \ln^3(1 + \hat{\lambda}t_i)}{2 \left[ \frac{n}{\hat{\alpha}^2} + \sum_{i=1}^n (1 + \hat{\lambda}t_i)^{\hat{\alpha}} \ln^2(1 + \hat{\lambda}t_i) \right]^2} \\ &+ \frac{-\sum_{i=1}^n t_i^2 (1 + \hat{\lambda}t_i)^{\hat{\alpha}-2} [2\hat{\alpha} - 1 + \hat{\alpha}(\hat{\alpha}-1) \ln(1 + \hat{\lambda}t_i)] - \sum_{i=1}^n \frac{t_i^2}{(1+\hat{\lambda}t_i)^2}}{2 \left[ \frac{n}{\hat{\lambda}^2} + \hat{\alpha}(\hat{\alpha}-1) \sum_{i=1}^n t_i^2 (1 + \hat{\lambda}t_i)^{\hat{\alpha}-2} + (\hat{\alpha}-1) \sum_{i=1}^n \frac{t_i^2}{(1+\hat{\lambda}t_i)^2} \right] \left[ \frac{n}{\hat{\alpha}^2} + \sum_{i=1}^n (1 + \hat{\lambda}t_i)^{\hat{\alpha}} \ln^2(1 + \hat{\lambda}t_i) \right]}\end{aligned}$$

where  $\hat{\alpha}, \hat{\lambda}$  are the MLE's.

If  $\alpha$  is known, under the proposed prior, the approximate Bayes estimator of  $\lambda$  under SEL function is

$$\begin{aligned}\lambda_B = \hat{\lambda} &+ \frac{\left(\frac{b-1}{\hat{\lambda}}\right) - a}{\frac{n}{\hat{\lambda}^2} + \alpha(\alpha-1) \sum_{i=1}^n t_i^2 (1 + \hat{\lambda}t_i)^{\alpha-2} + (\alpha-1) \sum_{i=1}^n \frac{t_i^2}{(1+\hat{\lambda}t_i)^2}} \\ &+ \frac{\frac{2n}{\hat{\lambda}^3} - \alpha(\alpha-1)(\alpha-2) \sum_{i=1}^n t_i^3 (1 + \hat{\lambda}t_i)^{\alpha-3} + (\alpha-1) \sum_{i=1}^n \frac{2t_i^3}{(1+\hat{\lambda}t_i)^3}}{2 \left[ \frac{n}{\hat{\lambda}^2} + \alpha(\alpha-1) \sum_{i=1}^n t_i^2 (1 + \hat{\lambda}t_i)^{\alpha-2} + (\alpha-1) \sum_{i=1}^n \frac{t_i^2}{(1+\hat{\lambda}t_i)^2} \right]^2}\end{aligned}$$

If  $\lambda$  is known, under the proposed prior, the approximate Bayes estimator of  $\alpha$  under SEL function is

$$\alpha_B = \hat{\alpha} + \frac{\frac{d-1}{\hat{\alpha}} - c}{\frac{n}{\hat{\alpha}^2} + \sum_{i=1}^n (1 + \lambda t_i)^{\hat{\alpha}} \ln^2(1 + \lambda t_i)} + \frac{\frac{2n}{\hat{\alpha}^3} - \sum_{i=1}^n (1 + \lambda t_i)^{\hat{\alpha}} \ln^3(1 + \lambda t_i)}{2 \left[ \frac{n}{\hat{\alpha}^2} + \sum_{i=1}^n (1 + \lambda t_i)^{\hat{\alpha}} \ln^2(1 + \lambda t_i) \right]^2}$$

### 3 Numerical Comparisons

In this section we compare the Bayes estimators with the maximum likelihood estimators. We assume the non-informative priors on both the shape and



scale parameters and compute approximated Bayes estimators using Lindley's approximation. The average estimates (AE) and square root of the mean squared error (RMS) are empirically evaluated based on a Monte-Carlo simulation study of 1000 replications for different sample sizes ( $n = 20, 50, 100$ ). The results are summarized in Table 1. It is noted that as sample size increases, the RMS and the bias for all estimators decrease, as we expected. When  $\lambda$  is known the Bayes estimates of  $\alpha$  perform better than the MLE's in terms of both bias and RMS for all cases. When both the shape and the scale parameters are unknown and the non-informative priors are considered, it is observed that when  $\alpha < 0.5$ , the MLE's of  $\alpha$  are better than the Bayes estimates in terms of bias but not of RMS and the MLE's of  $\lambda$  are better than the Bayes estimates in terms of both bias and RMS. For  $\alpha \geq 0.5$ , the Bayes estimates of  $\alpha$  perform better than the MLE's in terms of both the bias and the RMS and for a small sample size the Bayes estimates of  $\lambda$  perform better than the MLE's in terms of both the bias and the RMS when  $\alpha > 0.5$ . The results are summarized in Table 2.

#### 4 Data Analysis

To provide the real-life example for motivating our study, we apply our method to a real data set. Linhart and Zucchini (1986) gave the failure times of the air conditioning system of an airplane. Data set: 23, 261, 87, 7, 120, 14, 62, 47, 225, 71, 246, 21, 42, 20, 5, 12, 120, 11, 3, 14, 71, 11, 14, 11, 16, 90, 1, 16, 52, 95. We analyzed the data using (1) and computed the MLE's and the Bayesian estimators of  $\alpha$  and  $\lambda$  under non-informative priors. For computing the approximate Bayes estimates, we used Lindley's idea and the MCMC technique. Also it was assumed that we do not have any prior information, i.e.  $a = b = c = d = 0$ . We also computed the exact Bayes estimates of  $\alpha$  and  $\lambda$  using the numerical integration and compared it with the approximate Bayes estimates resulting from MCMC and Lindley's approximation. To check the validity of the model, we computed the Kolmogorov-Smirnov distance between the empirical distribution function and the fitted distribution function when the parameters are obtained by different methods. We also computed chi-squared statistics. The results are summarized in Table 3. It was observed that the MCMC method provided the approximate Bayes estimates which are closer to exact Bayes estimates and the Lindley's approximation gave estimates which are closer to the maximum likelihood estimates (Figure 1). We plotted the empirical distribution function and the estimated cumulative distribution function obtained using the MLE's, exact Bayes and approximate Bayes in Figure 2. From Figure 2, it is clear that the estimated cumulative distribution function obtained using MLE's and Lindley's approximation are closer to the empirical distribution function and provide a good fit to the given data.

## 5 Conclusion

In this paper, we consider the problem of estimating the parameters of an extension of exponential family. This family always has a decreasing probability density function and yet allows for monotone hazard rates. We are interested in estimating the parameters of this family by Bayesian methods. It is observed that the Bayes estimators cannot be obtained under explicit form and the numerical integration is required. We focused on the Lindley's approximation and compared the approximate Bayesian estimators under non-informative priors to the maximum likelihood estimators through a simulation study. The results showed that the Lindley's approximation and the maximum likelihood method produce the same estimates. So the approximation works very well. For a real data set, we used the Monte Carlo Markov Chain method to generate the samples from the posterior distribution and computed the approximate Bayes estimates. Comparison of results shows that for this data set, Lindley's approximation performs better than the MCMC method. As mentioned, the above results turned up with non-informative priors.

## References

- KUNDU, D., and GUPTA, R. D. (2008): Generalized exponential distribution: Bayesian estimations. *Computational Statistics & Data Analysis* 52, 1873–1883.
- LINDLEY, D. V. (1980): Approximate Bayesian method. *Trabajos Estadist* 31, 223–237.
- LINHART, H., and ZUCCHINI, W. (1986): Model Selection. Wiley, New York.
- NADARAJAH, S., and Haghighi, F. (2009): An extension of the exponential distribution. accepted in Statistics .

n	$\alpha$	0.2	0.5	1	1.5	2	2.5	3
20	MLE	1.0387 (0.1563)	1.0453 (0.1575)	1.0406 (0.1588)	1.0410 (0.1588)	1.0335 (0.1561)	1.0374 (0.1574)	1.0361 (0.1550)
	ABAYES	1.0193 (0.1498)	1.0285 (0.1503)	1.0216 (0.1544)	1.0216 (0.1520)	1.0143 (0.1503)	1.0180 (0.1511)	1.0167 (0.1488)
50	MLE	1.0150 (0.0933)	1.0176 (0.0911)	1.0115 (0.0905)	1.0165 (0.0917)	1.0090 (0.0887)	1.0151 (0.0922)	1.0114 (0.0888)
	ABAYES	1.0074 (0.0917)	1.0099 (0.0893)	1.0038 (0.0892)	1.0089 (0.0900)	1.0014 (0.0876)	1.0075 (0.0906)	1.0038 (0.0875)
100	MLE	1.0059 (0.0616)	1.0095 (0.0659)	1.0055 (0.0608)	1.0096 (0.0633)	1.0092 (0.0640)	1.0085 (0.0643)	1.0052 (0.0638)
	ABAYES	1.0021 (0.0611)	1.0057 (0.0652)	1.0017 (0.0604)	1.0057 (0.0626)	1.0054 (0.0633)	1.0047 (0.0637)	1.0014 (0.0633)

Table 1: The (AE) and (RMS) for the MLE's and the approximate Bayes estimate of  $\alpha$  when  $\lambda$  is known.

n		$\alpha = 0.2, \lambda = 0.2$		$\alpha = 0.2, \lambda = 0.5$		$\alpha = 0.2, \lambda = 1$	
		MLE	ABAYES	MLE	ABAYES	MLE	ABAYES
20	$\alpha$	0.2228	0.2238	0.2247	0.2259	0.2246	0.2257
		(0.0629)	(0.0613)	(0.0687)	(0.0667)	(0.0762)	(0.0740)
	$\lambda$	0.2448	0.2760	0.6317	0.7275	1.2345	1.4285
		(0.2828)	(0.3425)	(0.7409)	(0.9342)	(1.4233)	(1.7872)
50	$\alpha$	0.2104	0.2112	0.2084	0.2091	0.2085	0.2093
		(0.0319)	(0.0318)	(0.0297)	(0.0296)	(0.0318)	(0.0317)
	$\lambda$	0.2078	0.2170	0.5385	0.5633	1.0826	1.1327
		(0.1167)	(0.1249)	(0.3164)	(0.3408)	(0.6455)	(0.6951)
100	$\alpha$	0.2038	0.2042	0.2036	0.2040	0.2036	0.2041
		(0.0193)	(0.0192)	(0.0188)	(0.0188)	(0.0189)	(0.0189)
	$\lambda$	0.2062	0.2107	0.5167	0.5281	1.0332	1.0561
		(0.0767)	(0.0795)	(0.1906)	(0.1977)	(0.3817)	(0.3959)
n		$\alpha = 0.5, \lambda = 0.2$		$\alpha = 1, \lambda = 0.2$		$\alpha = 1, \lambda = 1$	
		MLE	ABAYES	MLE	ABAYES	MLE	ABAYES
20	$\alpha$	0.6855	0.6643	1.8029	1.7212	1.7970	1.7160
		(0.4743)	(0.4443)	(1.4501)	(1.3504)	(1.5022)	(1.4013)
	$\lambda$	0.2209	0.2211	0.2106	0.2061	1.0775	1.0548
		(0.2011)	(0.2079)	(0.2473)	(0.2469)	(1.2175)	(1.2118)
50	$\alpha$	0.5483	0.5429	1.2951	1.2730	1.3050	1.2827
		(0.1506)	(0.1460)	(0.7394)	(0.7167)	(0.7409)	(0.7181)
	$\lambda$	0.2054	0.2050	0.2010	0.1988	0.9899	0.9792
		(0.0967)	(0.0973)	(0.1246)	(0.1238)	(0.6013)	(0.5975)
100	$\alpha$	0.5248	0.5223	1.13434	1.1249	1.1354	1.1261
		(0.0893)	(0.0878)	(0.3895)	(0.3823)	(0.3919)	(0.3846)
	$\lambda$	0.2010	0.2008	0.1978	0.1967	0.9891	0.9835
		(0.0714)	(0.0716)	(0.0837)	(0.0834)	(0.4188)	(0.4174)
n		$\alpha = 1, \lambda = 1.5$		$\alpha = 1.5, \lambda = 1$		$\alpha = 0.5, \lambda = 1$	
		MLE	ABAYES	MLE	ABAYES	MLE	ABAYES
20	$\alpha$	1.6993	1.6233	2.9071	2.7643	0.6697	0.6496
		(1.4678)	(1.3697)	(2.4205)	(2.2501)	(0.4267)	(0.3995)
	$\lambda$	1.7080	1.6732	1.0634	1.0338	1.1119	1.1147
		(1.9380)	(1.9362)	(1.2540)	(1.2320)	(1.0622)	(1.1017)
50	$\alpha$	1.3159	1.2933	2.1864	2.1438	0.5502	0.5448
		(0.7422)	(0.7191)	(1.4808)	(1.4347)	(0.1477)	(0.1430)
	$\lambda$	1.4553	1.4396	0.9935	0.9809	1.0144	1.0123
		(0.8964)	(0.8919)	(0.6812)	(0.6745)	(0.4865)	(0.4898)
100	$\alpha$	1.1246	1.1154	1.8368	1.8191	0.5211	0.5188
		(0.3725)	(0.3657)	(0.8554)	(0.8399)	(0.0873)	(0.0859)
	$\lambda$	1.4823	1.4739	0.9778	0.9714	1.0156	1.0143
		(0.5923)	(0.5904)	(0.4704)	(0.4681)	(0.3489)	(0.3500)

Table 2: The (AE) and (RMS) for the MLE's and the approximate Bayes estimates of  $\alpha$  and  $\lambda$  when both parameters are unknown.

Methods	$(\hat{\alpha}, \hat{\lambda})$	K-S	$\chi^2$
MLE	(0.5985, 0.0434)	0.5301	10.44232
Lindley	(0.5895, 0.0429)	0.5373	10.67144
MCMC	(0.6571, 0.0534)	0.4718	11.82650
Exact Bayes	(0.6477, 0.0519)	0.4710	11.10849

Table 3: Parameter estimations, Kolmogorov-Smirnov and Chi-squared statistics for the data set.

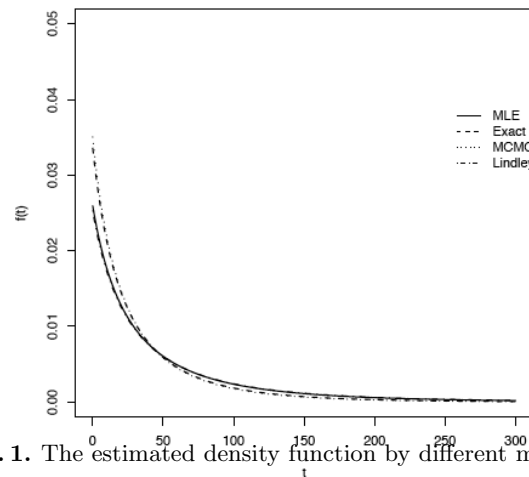


Fig. 1. The estimated density function by different methods.

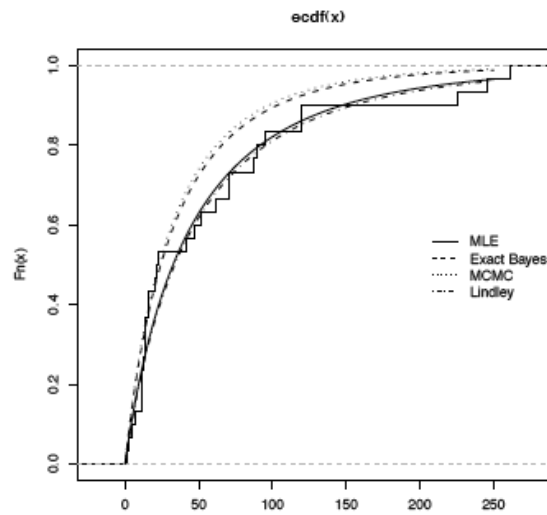


Fig. 2. The estimated cumulative distribution functions and empirical distribution function.

# A Cluster-Target Similarity Based Principal Component Analysis for Interval-Valued Data

Mika Sato-Ilic

Faculty of Systems and Information Engineering, University of Tsukuba  
Tennodai 1-1-1, Tsukuba, Ibaraki 305-8573, Japan, *mika@sk.tsukuba.ac.jp*

**Abstract.** This paper proposes a new principal component analysis for interval-valued data. The merit of this analysis is the consideration of dissimilarity of objects in a higher dimensional space when we obtain the projected space by using a covariance matrix involving the contribution degree for the fuzzy classification structure of objects, based on dissimilarity of objects in the higher dimensional space. In order to obtain the adaptable classification structure which is closely related with a selection of an appropriate number of clusters, we propose an alignment criterion which measures similarity between original similarity data and the restored similarity consisting of a fuzzy clustering result under a given number of clusters which we call cluster-target similarity. In addition, we prove the concentration of the alignment criterion which shows that empirical alignment is close to its expectation.

**Keywords:** fuzzy clustering, symbolic data, alignment, metric projection

## 1 Introduction

The aim of principal component analysis (PCA) is to summarize the latent similarity structure of data observed in high dimensional space by projecting the data into a much smaller dimensional space. However, classical PCA has the following problem. Since the methodology of classical PCA is based on orthogonal projection, the metric projection defined in convex vector space, which is the data space, is non-expansive. Therefore, there is a possibility that a norm between two projected vectors (objects) in a smaller dimensional space is "inevitably" smaller than the norm between the corresponding pre-projected two vectors (objects) in a high dimensional space. The root cause of this problem is that PCA only focuses on minimizing the sum of square of distances from objects in a high dimensional space to a hyper plane in a lower dimensional space, and does not consider similarities among objects in a high dimensional space.

Therefore, in this study, we extract similarity structure of objects in a high dimensional space by using a fuzzy clustering method. By tacking the result of the fuzzy clustering method to the PCA, we propose a new PCA considering the similarity structure of objects in a high dimensional space in

order to obtain more accurate result of the PCA. In order to quantify the validity of the fuzzy clustering, we define a criterion of alignment between original similarity data and the restored similarity created by the obtained result of fuzzy clustering when we assume several numbers of clusters. Once we assume a number of clusters, then we can obtain a fuzzy clustering result and based on the result, we create a restored similarity. So, if we assume several numbers of clusters, then we can obtain the corresponding several numbers of restored similarities. Then we evaluate which restored similarity based on a clustering result has the closest relation with the original similarity data by using an alignment criterion. That is, we evaluate which clustering result is the most adaptable for the original similarity data.

## 2 Clustering interval-valued data

Suppose the observed interval-valued data  $y_{ia}$  which are values of  $n$  objects with respect to  $p$  variables are denoted by the following:

$$Y = (y_{ia}) = ([\underline{y}_{ia}, \bar{y}_{ia}]), \quad i = 1, \dots, n, \quad a = 1, \dots, p,$$

where  $y_{ia} = [\underline{y}_{ia}, \bar{y}_{ia}]$  shows the interval-valued data of the  $i$ -th object with respect to a variable  $a$  which has the minimum value  $\underline{y}_{ia}$  and the maximum value  $\bar{y}_{ia}$ . The dissimilarity between  $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})$  and  $\mathbf{y}_j = (y_{j1}, \dots, y_{jp})$  is defined as follows:

$$d_{ij} = \sum_{a=1}^p \sup\{d(x, y_{ja}) | x \in y_{ia}\}, \quad d(x, y_{ja}) = \inf\{d(x, y) | y \in y_{ja}\}, \quad (1)$$

$$d_{ji} = \sum_{a=1}^p \sup\{d(y_{ia}, y) | y \in y_{ja}\}, \quad d(y_{ia}, y) = \inf\{d(x, y) | x \in y_{ia}\}. \quad (2)$$

Where,  $d(x, y)$  shows distance between  $x$  and  $y$ ,  $\forall x \in y_{ia}, \forall y \in y_{ja}$ . Therefore,  $d_{ij} \neq d_{ji}$ , ( $i \neq j$ ). We transform this dissimilarity to similarity as follows:

$$S = (s_{ij}), \quad s_{ij} = 1 - d_{ij} / \max_{i,j}\{d_{ij}\}, \quad i, j = 1, \dots, n. \quad (3)$$

Since  $S = (s_{ij})$  is asymmetric similarity data, the asymmetric fuzzy clustering model (Sato et al. (1995)) will be used. The model is defined as follows:

$$s_{ij} = \sum_{k=1}^K \sum_{l=1}^K w_{kl} u_{ik} u_{jl} + \varepsilon_{ij}, \quad (4)$$

where  $s_{ij} \neq s_{ji}$ , ( $i \neq j$ ),  $w_{kl} \neq w_{lk}$ , ( $k \neq l$ ). In this model, the state of fuzzy clustering is represented by a partition matrix  $U = (u_{ik})$  whose elements show the degree of belongingness of the objects to the clusters,

$u_{ik}$ ,  $i = 1, \dots, n$ ,  $k = 1, \dots, K$ , where  $n$  is number of objects and  $K$  is number of clusters. In general,  $u_{ik}$  satisfies the following conditions:

$$u_{ik} \in [0, 1], \quad \sum_{k=1}^K u_{ik} = 1. \quad (5)$$

The weight  $w_{kl}$  is considered to be a quantity which shows the asymmetric similarity between a pair of clusters. In this paper, we define the  $w_{kl}$  as derived from an assumption of normal distribution of objects in each cluster as follows:

$$w_{kl}^{(K)} = 1 - 1/(1 + e^{-\tilde{w}_{kl}^{(K)}}), \quad (6)$$

where

$$\tilde{w}_{kl}^{(K)} = \frac{1}{2} \left( \|\boldsymbol{\mu}_{(k,K)} - \boldsymbol{\mu}_{(l,K)}\|_{\Sigma_{(k,K)}^{-1}} + \text{tr}(\Sigma_{(k,K)}^{-1} \Sigma_{(l,K)} - I) + \log \frac{|\Sigma_{(k,K)}|}{|\Sigma_{(l,K)}|} \right), \quad (7)$$

$$\|\boldsymbol{\mu}_{(k,K)} - \boldsymbol{\mu}_{(l,K)}\|_{\Sigma_{(k,K)}^{-1}} = (\boldsymbol{\mu}_{(k,K)} - \boldsymbol{\mu}_{(l,K)})' \Sigma_{(k,K)}^{-1} (\boldsymbol{\mu}_{(k,K)} - \boldsymbol{\mu}_{(l,K)}), \quad \forall k, l.$$

Where  $\tilde{w}_{kl}^{(K)}$  is derived from Kullback-Leibler's divergence (Bock, et al. (2000)) and  $w_{kl}^{(K)}$  is obtained by using monotone transformation to similarity having a range of  $[0, 1]$  shown in equation (6).  $w_{kl}^{(K)}$  shows the similarity from a cluster  $k$  to a cluster  $l$  when we assume the number of clusters is  $K$ .  $I$  is a unit matrix.  $\boldsymbol{\mu}_{(k,K)}$  and  $\Sigma_{(k,K)}$  are an expected value and a variance-covariance matrix of  $S_{(k,K)}$  which is shown as follows:

$$S_{(k,K)} = \{\tilde{\mathbf{y}}_i \mid p_{ik}^{(K)} = 1\}, \quad \tilde{\mathbf{y}}_i = (\tilde{y}_{i1}, \dots, \tilde{y}_{ip}), \quad \tilde{y}_{ia} = (\underline{y}_{ia} + \bar{y}_{ia})/2, \quad \forall k,$$

where  $p_{ik}^{(K)}$  satisfy the following:

$$u_{ik}^{(K)} = \max_{1 \leq k \leq K} u_{ik}^{(K)} \rightarrow p_{ik}^{(K)} = 1, \quad i = 1, \dots, n,$$

under the condition of  $\sum_{k=1}^K p_{ik}^{(K)} = 1$ . In the case that  $\max_{1 \leq k \leq K} u_{ik}^{(K)}$  is not unique, we select the first cluster which appears having the maximum degree of belongingness over the clusters.

$u_{ik}^{(K)}$  shows degree of belongingness of an object  $i$  to a cluster  $k$  when we assume the number of clusters is  $K$ , and satisfy the condition (5). From equations (6) and (7),  $w_{kl}^{(K)} \neq w_{lk}^{(K)}$ ,  $(k \neq l)$ ,  $w_{kl}^{(K)} \in [0, 1]$  are satisfied.

### 3 Selection of number of clusters

The criterion of selection of an appropriate number of clusters is defined as follows:

$$C(K) = \sum_{i \neq j=1}^n s_{ij} \tilde{s}_{ij}^{(K)} / \left( \sqrt{\sum_{i \neq j=1}^n s_{ij}^2} \sqrt{\sum_{i \neq j=1}^n \tilde{s}_{ij}^{(K)^2}} \right), \quad (8)$$

where  $\tilde{s}_{ij}^{(K)}$  shows the restored similarity obtained as follows:

$$\tilde{s}_{ij}^{(K)} = \sum_{k=1}^K \sum_{l=1}^K w_{kl}^{(K)} u_{ik}^{(K)} u_{jl}^{(K)}. \quad (9)$$

$C(K)$  shows the degree of alignment between  $s_{ij}$  and  $\tilde{s}_{ij}^{(K)}$ . Therefore, the larger value of  $C(K)$  is better when compared with several cases in which we assume several numbers of clusters shown as  $K$ . In other words, selecting the best  $K$  when we obtain the largest value of  $C(K)$  means selecting the best matched latent classification structure of original similarity data,  $S = (s_{ij})$ , since  $\tilde{s}_{ij}^{(K)}$  shown in equation (9) involves the latent classification structure of  $s_{ij}$  when the number of clusters is fixed as  $K$ .

Next, we discuss the concentration around the expected value of the criterion for the different  $w_{kl}$ . Equation (8) is shown as the following for different  $\hat{w}_{kl}$  and  $\tilde{w}_{kl}$ :

$$\hat{C}(s_{ij}, \hat{s}_{ij}, \hat{w}_{kl}) = \hat{C}(s, \hat{s}, \hat{w}), \quad \hat{C}(s_{ij}, \tilde{s}_{ij}, \tilde{w}_{kl}) = \hat{C}(s, \tilde{s}, \tilde{w}).$$

Then we obtain as follows:

$$|\hat{C}(s, \hat{s}, \hat{w}) - \hat{C}(s, \tilde{s}, \tilde{w})| \leq (6/(ma^2)), \quad m = n(n-1). \quad (10)$$

Let us rewrite

$$\hat{C}(s, \tilde{s}, \tilde{w}) = \hat{C}(s, \tilde{s} + d\tilde{s}, \tilde{w}).$$

Now we evaluate as follows:

$$\begin{aligned} \hat{C}(s, \tilde{s} + d\tilde{s}, \tilde{w}) - \hat{C}(s, \tilde{s}, \tilde{w}) &\leq \frac{2 \langle \tilde{s}, d\tilde{s} \rangle}{\|\tilde{s}\|(\|\tilde{s}\| + \|\tilde{s} + d\tilde{s}\|)} + \frac{\|d\tilde{s}\|^2}{\|\tilde{s}\|(\|\tilde{s}\| + \|\tilde{s} + d\tilde{s}\|)} \\ &\quad + \langle \frac{d\tilde{s}}{\|\tilde{s}\|}, \frac{s}{\|s\|} \rangle. \end{aligned} \quad (11)$$

Since  $\langle d\tilde{s}, \tilde{s} \rangle \leq 2/m$ , and from equation (11) we obtain equation (10). From equation (10) and the following theorem, we obtain the following which is concerned with the concentration around the expected value:

$$P(|\hat{C}(s, \hat{s}, w) - E[\hat{C}(s, \hat{s}, w)]| \geq \varepsilon) \leq 2 \exp((-m^2 a^4 \varepsilon^2)/c), \quad (12)$$

where  $c$  is a constant.

**Theorem 1** (McDiarmid, (1989)): Let  $X_1, \dots, X_n$  be independent random variables taking values in a set  $A$ , and assume that  $f : A^n \rightarrow R$  satisfies for  $1 \leq i \leq n$

$$\sup_{x_1, \dots, x_n, \hat{x}_i} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, \hat{x}_i, x_{i+1}, \dots, x_n)| \leq c_i,$$

then for all  $\varepsilon > 0$ ,

$$P(|f(X_1, \dots, X_n) - Ef(X_1, \dots, X_n)| \geq \varepsilon) \leq 2 \exp\left(-2\varepsilon^2 / \sum_{i=1}^n c_i^2\right).$$



That is, from theorem 1 and equation (10), we calculate as follows and obtain equation (12):

$$2 \exp \left( -2\varepsilon^2 / \sum_{l=1}^{K(K-1)} c_l^2 \right) = 2 \exp ((-m^2 a^4 \varepsilon^2)/c),$$

where  $c$  is a constant and

$$\varepsilon = \sqrt{(c_1 \log(\frac{2}{\delta}))/m^2}, \quad \delta = 2 \exp ((-m^2 a^4 \varepsilon^2)/c).$$

#### 4 PCA based on fuzzy clustering

First, we discuss single-valued PCA which is interpreted geometrically as finding a projected space spanned by vectors that show direction of the principal components. Let  $L$  be a nonempty subset of the inner product space  $X$ . Then we define a mapping  $P_L$  from  $X$  into the subsets of  $L$  called the metric projection onto  $L$ . Then  $P_L(\mathbf{o}_1)$  is defined as follows:

$$P_L(\mathbf{o}_1) = \{\mathbf{o}_2 \in L \mid \|\mathbf{o}_1 - \mathbf{o}_2\| = d(\mathbf{o}_1, L)\},$$

where  $\mathbf{o}_1 \in X$  and  $d(\mathbf{o}_1, L) = \inf_{\mathbf{o}_2 \in L} \|\mathbf{o}_1 - \mathbf{o}_2\|$ . Let  $L$  be a convex Chebyshev set in which for each  $\mathbf{o}_1 \in X$ , there exists at least one nearest point in  $L$ . Then  $P_L$  is nonexpansive, that is,

$$\|P_L(\mathbf{o}_1) - P_L(\mathbf{o}_2)\| \leq \|\mathbf{o}_1 - \mathbf{o}_2\|, \quad \forall \mathbf{o}_1, \mathbf{o}_2 \in X. \quad (13)$$

The problem of the PCA is that the metric projections only satisfies equation (13) and PCA does not consider the size of values shown as follows:

$$C(\mathbf{o}_1, \mathbf{o}_2) = \|\mathbf{o}_1 - \mathbf{o}_2\| - \|P_L(\mathbf{o}_1) - P_L(\mathbf{o}_2)\|.$$

Our obtained data is interval-valued data. The empirical joint density function for bivariate  $a$  and  $b$  for interval-valued data has been defined (Billard et al. (2000)) as follows:

$$f(y_a, y_b) = \frac{1}{n} \sum_{i=1}^n I_i(y_a, y_b) / \|Z(i)\|, \quad (14)$$

where  $I_i(y_a, y_b)$  is the indicator function where each element of  $(\mathbf{y}_a, \mathbf{y}_b)$  is or is not in the rectangle  $Z(i) = y_{ia} \times y_{ib}$  consisted of two sides which are intervals  $[y_{ia}, \bar{y}_{ia}]$  and  $[y_{ib}, \bar{y}_{ib}]$ .  $y_a$  and  $y_b$  are random variables.  $\|Z(i)\|$  is the area of this rectangle.  $\mathbf{y}_a$  is  $a$ -th column vector of  $Y$  and is shown as follows:  $\mathbf{y}_a = (y_{1a}, \dots, y_{na})^t = ([y_{1a}, \bar{y}_{1a}], \dots, [y_{na}, \bar{y}_{na}])^t$ . We extend the empirical joint density function shown in equation (14) as follows:

$$\tilde{f}(y_a, y_b) = \frac{1}{n} \sum_{i=1}^n (w_i I_i(y_a, y_b) / \|Z(i)\|), \quad (15)$$

$$w_i = \sum_{k=1}^K u_{ik}^m / \sum_{i=1}^n \sum_{k=1}^K u_{ik}^m, \quad i = 1, \dots, n, \quad m \in (1, \infty), \quad (16)$$

where  $u_{ik}$ ,  $i = 1, \dots, n$ ,  $k = 1, \dots, K$  show the obtained degree of belongingness of the objects to the clusters when  $K$  is the selected appropriate number of clusters. Then fuzzy covariance for interval-valued data between variables  $a$  and  $b$  is derived as follows:

$$\hat{c}_{ab} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y_a - \bar{y}_a)(y_b - \bar{y}_b) \tilde{f}(y_a, y_b) dy_a dy_b, \quad \bar{y}_a = \frac{1}{2n} \sum_{i=1}^n (\underline{y}_{ia} + \bar{y}_{ia}). \quad (17)$$

Substituting equation (16) into equation (17), and from equation (5), we have obtained the following:

$$\begin{aligned} \hat{c}_{ab} = & (1/(4n)) \sum_{i=1}^n w_i (\bar{y}_{ia} + \underline{y}_{ia})(\bar{y}_{ib} + \underline{y}_{ib}) - (1/n) \bar{y}_b \sum_{i=1}^n (w_i (\bar{y}_{ia} + \underline{y}_{ia}))/2 \\ & - (1/n) \bar{y}_a \sum_{i=1}^n (w_i (\bar{y}_{ib} + \underline{y}_{ib}))/2 + (1/n) \bar{y}_a \bar{y}_b. \end{aligned} \quad (18)$$

From equations (5) and (16),  $w_i$  satisfy the following condition:

$$w_i > 0, \quad \sum_{i=1}^n w_i = 1. \quad (19)$$

In a hard clustering when  $u_{ik} \in \{0, 1\}$ ,  $\sum_{k=1}^K u_{ik} = 1$  is satisfied, equation (16) is

$$w_i = 1/n, \quad \forall i. \quad (20)$$

Since  $u_{ik}$  satisfies conditions shown in equation (5), the weight  $w_i$  in equation (16) shows how an object is clearly classified for the obtained classification structure. If an object  $i$  is clearly classified to a cluster, then the weight  $w_i$  becomes larger, and if the classification situation with respect to an object  $i$  is an uncertainty situation, then the value of  $w_i$  becomes smaller. Therefore, it can be seen that the weights shown in equation (16) show a degree of fuzziness of the clustering with respect to each object and the proposed fuzzy covariance matrix for interval-valued data,  $\hat{C} = (\hat{c}_{ab})$ ,  $a, b = 1, \dots, p$  shown in equation (18) involve a classification structure over the variables which is obtained by reflecting the dissimilarity structure of objects in a higher dimensional space shown as  $\|\mathbf{o}_1 - \mathbf{o}_2\|$  in equation (13). Then based on the covariance matrix, we obtain principal components.

## 5 Numerical example

We use overall performance rankings data for possible energy options shown in table 1 (Yoshizawa et al., (2008)). The data shows value of evaluation obtained by nine energy specialists or policy actors over two countries, United

Kingdom and Japan. Each value is obtained as interval-valued data by considering subjective uncertainty involvement in such evaluation. First, we calculate an asymmetric dissimilarity matrix obtained by using equations (1) and (2). Next, we symmetrize this data as  $\hat{d}_{ij} = (d_{ij} + d_{ji})/2$  and apply  $\hat{d}_{ij}$  to a fuzzy clustering method called FANNY (Kaufman, et al. (1990)). The result of the fuzzy clustering is used for obtaining several restored similarity shown in equation (9) by using equation (6).

Table 1. Performance Ranking Data for Energy

Energy	JP1	JP2	...	UK1	UK2	...
1. Oil	[60,90]	[60,70]	...	[81,91]	[51,71]	...
2. Coal	[90,120]	[80,95]	...	[80,91]	[80,91]	...
3. Coal with CCS	[60,100]	[20,40]	...	[50,60]	[65,75]	...
4. Nuclear	[70,120]	[50,85]	...	[45,65]	[60,80]	...
5. Geothermal	[60,80]	[30,45]	...	[0,20]	[0,20]	...
6. Solar PV	[30,70]	[30,40]	...	[0,10]	[10,40]	...
7. Biomass	[40,100]	[20,35]	...	[60,70]	[60,70]	...
8. On. Wind, large	[70,100]	[50,60]	...	[60,72]	[50,70]	...
9. Mun/Ind Waste	[83,111]	[50,65]	...	[60,80]	[80,90]	...
10. Hydro	[70,100]	[40,60]	...	[65,75]	[60,70]	...
11. Gas	[80,120]	[65,85]	...	[87,97]	[87,97]	...

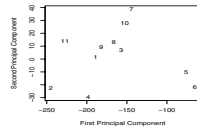


Fig. 1 Result for Proposed PCA

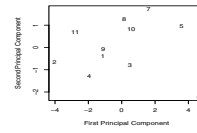


Fig. 2 Result for Centers Method

When compared with the original similarity obtained in equation (3) and the restored similarity obtained in equation (9), the result of alignment between the two similarities,  $C(K)$ , is obtained as follows:  $C(2) = 0.93$ ,  $C(3) = 0.90$ ,  $C(4) = 0.91$ . From this result, it can be seen that an appropriate number of clusters is selected as 2, since the value of  $C(2)$  is the largest. Figure 1 shows the results of the first and the second principal components by using the proposed PCA. In this figure, each number shows the number of energy shown in table 1. From this result, we can see the similarity structure of eleven energies through nine specialists evaluation. Figure 2 shows the results of the

first and the second principal components by applying the data consisting of centers of intervals to the conventional PCA. This method used for figure 2 is referred as the centers method in the symbolic data analysis (Bock, et al. (2000)). This method is also identical with a method in which we use the conventional empirical joint density function shown in equation (14) and derive the covariance and then apply the obtained covariance into the conventional PCA. From equation (20), this method is the same as a case in which we use a hard clustering in a high dimensional space in our proposed PCA. Table 2 shows a comparison of values of cumulative proportion which is the sum of the first and the second proportions corresponding to the first and the second principal components shown in the results of figures 1 and 2. From this result, we can see that the proposed PCA could obtain a better result.

Table 2. Comparison of Cumulative Proportion

Proposed PCA	Ordinal PCA (Centers Method)
0.86	0.82

## 6 Conclusion

A new principal component analysis is proposed involving the dissimilarity structure in high dimensional space through a result of fuzzy clustering. In order to select an appropriate number of clusters in the fuzzy clustering, we define a criterion of alignment and prove the concentration. A numerical example shows a better performance of the proposed PCA when compared with an ordinal PCA.

## References

- BILLARD, L. and DIDAY, E. (2000): Regression analysis for interval-valued data. In: H.A.L. Kiers, et al. (Eds.): *Data Analysis, Classification, and Related Methods*. Springer, 369–374.
- BOCK, H.H. and DIDAY, E. (Eds.)(2000): *Analysis of Symbolic Data*. Springer.
- CRISTIANINI, N., KANDOLA, J., ELISSEEFF, A. and SHQWE-TAYLOR, J. (2006): On kernel target alignment. In: D.E. Holmes and L.C. Jain (Eds.): *Innovations in Machine Learning*. Springer.
- KAUFMAN, L. and ROUSSEEUW, P.J. (1990): *Finding Groups in Data*. John Wiley & Sons.
- MCDIARMID, C. (1989): *On the Method of Bounded Differences, Surveys in Combinatorics*. Cambridge University Press.
- SATO, M. and SATO, Y. (1995): Extended fuzzy clustering models for asymmetric similarity. In: B. Bouchon-Meunier, R.R. Yager, L.A. Zadeh (Eds.): *Fuzzy Logic and Soft Computing*. World Scientific, 228–237.
- YOSHIZAWA, G., STIRLING, A. and SUZUKI, T. (2008): Electricity system diversity in the UK and Japan: a multicriteria diversity analysis. *GraSPP Working Paper Series, Graduate School of Public Policy, The University of Tokyo*.

# Bayesian Flexible Modelling of Mixed Logit Models

Luisa Scaccia<sup>1</sup> and Edoardo Marcucci<sup>2</sup>

<sup>1</sup> Dip. di Istituzioni Economiche e Finanziarie, Università di Macerata,  
via Crescimbeni 20, 62100 Macerata, Italy, *scaccia@unimc.it*

<sup>2</sup> Dip. di Istituzioni Pubbliche, Economia e Società, Università di Roma Tre,  
via G. Chiabrera 199, 00145 Roma, Italy, *edoardo.marcucci@uniroma3.it*

**Abstract.** The widespread use of the Mixed Multinomial Logit model, in the context of discrete choice data, has made the issue of choosing a mixing distribution very important. The choice of a specific distribution may seriously bias results if that distribution is not suitable for the data. We propose a flexible hierarchical Bayesian approach in which the mixing distribution is approximated through a mixture of normal distributions. Numerical results on a real data set are provided to demonstrate the usefulness of the proposed method.

**Keywords:** hierarchical Bayes, mixed logit, mixture of distributions, random taste heterogeneity, semi-parametric estimation

## 1 Introduction

The multinomial logit (MNL) model has provided for a long time the foundation for the analysis of discrete choice modelling, due to its advantages in terms of closed-form solution and simplicity of interpretation and use (McFadden (1974)). However, some restrictive assumptions underlying the model have motivated researchers to consider alternative specifications, the most popular of which is probably the mixed logit (MMNL) model (McFadden and Train (2000), Train (1998)). In its simplest specification, the utility of each individual is a function of the alternative attributes, with attribute coefficients that are random and reflect individual preferences.

In MMNL models, however, a crucial issue is that of specifying an appropriate mixing distribution of the random coefficients that may be interpreted as representing random taste heterogeneity. Most popular specifications have been the normal, triangular, uniform and lognormal distributions. However, in practical applications, any of them has shown its deficiencies (Hess et al. (2005)). An inappropriate choice of the mixing distribution can lead to problems in interpretation and potentially misguided policy-decisions (Cirillo and Axhausen (2006), Fosgerau (2006)).

To deal with this issue, Fosgerau and Hess (2009) proposed two approaches: the first one improves on the flexibility of a base distribution by adding in a series approximation using Legendre polynomials; the second one

makes use of a semi-parametric mixing distribution consisting of a discrete mixture of normal distributions (MOD). Both approaches can approximate any continuous distribution, allowing also for multiple modes, a significant advantage compared to typically used distributions, restricted to a single mode. Allowing for multiple modes means that the population may be composed of distinct groups with different behaviour. In a Monte Carlo study, Fosgerau and Hess (2009) show that the two approaches do about equally well in outperforming commonly used distributions, over a range of situations. The MOD approach has a particular ability in approximating point masses. A heightened mass at zero is useful in representing taste heterogeneity for attributes that some individuals are indifferent to, as discussed by Cirillo and Axhausen (2006) with regard to valuation of travel time savings.

In this paper, we consider the MOD approach and we illustrate how to estimate this model in a Bayesian framework. Moreover, we extend the approach to the case in which multiple random coefficients, potentially correlated, are present in the model. As Fosgerau and Hess (2009), we will fix the number of components in the mixture. The extension to mixtures allowing the number of components to vary is a topic for further research.

We rely on Bayesian procedures since these avoid two of the most prominent difficulties associated with classical procedures. Firstly, the Bayesian procedures do not require maximization of any function, thus avoiding the numerical difficulties that often arise in maximizing the simulated likelihood function of some MMNL models. Secondly, desirable estimation properties, such as consistency and efficiency, can be attained under more relaxed conditions with Bayesian procedures (Train (2001)). Moreover, Bayesian procedures usually avoid the need to simulate choice probabilities, which is quite cumbersome with MMNL models. Finally, individual-level parameters can be easily obtained. The Bayesian perspective has been adopted in the context of discrete choice models by, for example, Train (2001) for mixed logits with normal, lognormal, uniform and triangular distributed coefficients. Allenby et al. (1998) used a mixture of normal distribution for random parameters in mixed logits, in the context of marketing research. We extend their approach, including fixed parameters in the model and allowing for a more flexible hierarchical structure. Our approach also relates to the one proposed by Ho and Hu (2008) for linear mixed models with random effects.

The paper is organized as follows: the model and prior assumptions are illustrated in Section 2; Section 3 deals with computational implementation; Section 4 discusses an application to the analysis of stated preference data on public transport demand. Conclusions are given in Section 5.

## 2 The Mixed logit model

In this Section, we illustrate the MOD approach in a hierarchical Bayesian fashion. We, then, specify priors for the parameters in the model.

## 2.1 The MOD approach

Person  $n$  faces a choice among  $J$  alternatives in each of  $T$  time periods. According to the specification of a MMNL model that we will use, the person's utility from alternative  $i$  in period  $t$  is:

$$U_{nit} = \alpha' w_{nit} + \beta'_n x_{nit} + \epsilon_{nit} \quad (n = 1, \dots, N; i = 1, \dots, J; t = 1, \dots, T)$$

where  $\epsilon_{nit} \sim$  i.i.d. extreme value,  $w_{nit}$  is a vector of  $R$  attributes (characterizing the alternative and/or the subject) whose coefficients  $\alpha$  are fixed and  $x_{nit}$  is a vector of  $K$  attributes whose coefficients  $\beta_n$  are supposed to be random and to vary in the population, according to the density  $g(\beta_n|\mu, \Sigma)$ , for  $n = 1, \dots, N$ , with  $\mu$  and  $\Sigma$  being hyperparameters. Person  $n$  chooses alternative  $i$  in period  $t$  if  $U_{nit} > U_{njt}, \forall j \neq i$ . Let  $y_n = (y_{n1}, \dots, y_{nT})$  be the person's sequence of choices over the  $T$  time periods. The probability of observing this sequence, conditional on the person-specific parameters  $\beta_n$  and the common fixed parameters  $\alpha$ , is the product of standard logit formulas:

$$L(y_n|\alpha, \beta_n) = \prod_{t=1}^T \frac{e^{\alpha' w_{ny_{nt}t} + \beta'_n x_{ny_{nt}t}}}{\sum_{j=1}^J e^{\alpha' w_{njt} + \beta'_n x_{njt}}}. \quad (1)$$

The MOD approach, adopting a semi-parametric perspective, assumes that  $g(\beta_n|\mu, \Sigma)$  is a mixture of  $C$  multivariate normal distributions:

$$\beta_n|\mu, \Sigma \sim \sum_{c=1}^C s_c \phi(\cdot|\mu_c, \Sigma_c) \quad (n = 1 \dots, N), \quad (2)$$

where  $\phi(\cdot|\mu_c, \Sigma_c)$  is a multivariate normal density with mean vector  $\mu_c = (\mu_{c1}, \dots, \mu_{cK})'$  and  $K$  by  $K$  covariance matrix  $\Sigma_c$ , and  $s_c$  are weights satisfying  $0 \leq s_c \leq 1$ , for  $c = 1, \dots, C$ , and  $\sum_{c=1}^C s_c = 1$ .

Notice that this model provides both the flexibility of the latent class model and the parsimony of the traditional MMNL model. Indeed, both models are special cases of the proposed model: the latent class model is obtained by letting the within-class variances go to zero, and the traditional MMNL model corresponds to using only one class or component.

## 2.2 Latent allocation variables

An alternative perspective, leading to the same mixture model in (2), involves the introduction of latent allocation variables  $z = (z_1, \dots, z_N)$  and the assumption that the vector  $\beta_n$ , relative to individual  $n$ , arose from an unknown component  $z_n$  of the mixture of multivariate normal distributions. The allocation variables are given probability mass function

$$p(z_n = c) = s_c \quad \text{independently for } n = 1, \dots, N, \quad (3)$$

and conditional on them, the random taste parameters  $\beta_n$  are independently drawn, for each subject  $n$ , from the density:

$$\beta_n|z, \mu, \Sigma \sim \phi(\cdot|\mu_{z_n}, \Sigma_{z_n}). \quad (4)$$

Integrating out  $z_n$  in (4), using the distribution in (3), leads back to (2).

### 2.3 Prior settings

We assume the number of components  $C$  to be fixed to a reasonable small number, as in Fosgerau and Hess (2009). In a forthcoming paper, we will consider  $C$  to be unknown and subject to inference, as well as the other parameters of the model. From past experience, we would not expect inference about the model proposed to be highly sensitive to prior specification. We use a weakly informative priors approach, according to which we use some information from the sample to set the values of the hyperparameters. In particular, we fit a standard mixed logit model to the data, with normal distribution of the taste parameters to get some idea from the estimated mean and standard errors of the random parameters (Ho and Hu (2008)).

In particular, we assume a priori:

- a)  $(s_1, \dots, s_C) \sim \mathcal{D}(\delta, \dots, \delta)$ , where  $\mathcal{D}$  denotes the Dirichlet distribution. We choose  $\delta = 1$ .
- b)  $z_n \sim p(z_n = c) = s_c$ , independently for  $n = 1, \dots, N$ .
- c)  $\Sigma_c \sim \mathcal{IW}(r, \Theta^{-1})$ , independently for  $c = 1, \dots, C$ , where  $\mathcal{IW}$  denotes the Inverse Wishart distribution.
- d)  $\mu_c \sim \mathcal{N}_K(\xi, D)$ , independently for  $c = 1, \dots, C$ , where  $\mathcal{N}_K$  denotes the  $K$ -dimensional multivariate normal distribution.
- e)  $\Theta \sim \mathcal{IW}(a, S^{-1})$ .
- f)  $\alpha \sim \mathcal{N}_R(\nu, \Omega)$ .

### 2.4 Complete hierarchical model

Let  $y = (y'_1, \dots, y'_N)'$ ,  $\mu = (\mu'_1, \dots, \mu'_C)'$  and  $\Sigma$  be the matrix obtained by stacking the covariance matrices  $\Sigma_c$  on top of each other. We exploit the natural conditional independence structure so that the joint distribution of all variables, conditional to the fixed values of the hyperparameters, is

$$\begin{aligned} p(y, s, z, \Sigma, \Theta, \mu, \alpha, \beta|C, \nu, \gamma, \xi, D, a, S, r, \delta) \\ = p(s|C, \delta)p(z|s, C)p(\Theta|a, S)p(\Sigma|r, \Theta, C) \\ \cdot p(\mu|\xi, D, C)p(\alpha|\nu, \Omega)p(\beta|z, \mu, \Sigma)p(y|\alpha, \beta), \end{aligned}$$

where  $p(z|s, C) = \prod_{n=1}^N s_{z_n}$ ,  $p(y|\alpha, \beta) = \prod_{n=1}^N L(y_n|\alpha, \beta_n)$ , with  $L(y_n|\alpha, \beta_n)$  defined in (1) and  $p(\beta|z, \mu, \Sigma)$  given in (4). The prior distributions  $p(s|C, \delta)$ ,  $p(\Theta|a, S)$ ,  $p(\Sigma|r, \Theta, C)$ ,  $p(\mu|\xi, D, C)$ ,  $p(\alpha|\nu, \Omega)$  are all given in Section 2.3.



### 3 Computational implementation

The complexity of the model presented requires Markov chain Monte Carlo (MCMC) methods to approximate the posterior distribution. Our sampler uses seven fixed-dimension moves. Gibbs samplers are used to update all model parameters, except  $\alpha$  and  $\beta$ , which are updated by means of Metropolis algorithm. The notation ‘ $\dots$ ’ will be used to denote ‘all other variables’.

**Updating  $s$ .** Before considering the updating of  $s$ , we comment briefly on the issue of labeling the components. The whole model is, in fact, invariant to permutation of the labels  $c = 1, \dots, C$ . For identifiability, we adopt a unique labeling in which the component weights are in increasing numerical order. As a consequence, the joint prior distribution of  $s$  is a Dirichlet density, restricted to the set  $s_1 < s_2 < \dots < s_C$ . The weights are updated by drawing them from their full conditional distribution

$$(s_1, \dots, s_C) | \dots \sim \mathcal{D}(\delta + m_1, \dots, \delta + m_C)$$

where  $m_c = \#\{n : z_n = c\}$  is the number of subjects currently allocated to the  $c$  component of the mixture. To preserve the ordering constraints on  $s$ , the move is accepted provided the ordering is unchanged.

**Updating  $z$ .** The allocation variable  $z_n$  has conditional probability

$$p(z_n = c | s, C, \beta_n) = \frac{s_c \phi(\beta_n | \mu_c, \Sigma_c)}{\sum_{c=1}^C s_c \phi(\beta_n | \mu_c, \Sigma_c)}.$$

We can update the  $z_n$  independently, sampling from this distribution.

**Updating  $\mu$ .** The  $\mu_c$  can be updated independently, drawing them from

$$\mu_c | \dots \sim \mathcal{N}_K \left( \frac{D^{-1} \xi + m_c \Sigma_c^{-1} \bar{\beta}_c}{D^{-1} + m_c \Sigma_c^{-1}}, \frac{1}{D^{-1} + m_c \Sigma_c^{-1}} \right)$$

where  $\bar{\beta}_c = m_c^{-1} \sum_{n: z_n = c} \beta_n$ .

**Updating  $\Theta$ .** We update  $\Theta$  sampling from its full conditional:

$$\Theta | \dots \sim \mathcal{IW} \left( a + Cr, S^{-1} + \sum_{c=1}^C \Sigma_c^{-1} \right)$$

**Updating  $\Sigma$ .** We update  $\Sigma_c$  independently, sampling from

$$\Sigma_c | \dots \sim \mathcal{IW} \left( m_c + r, \Theta^{-1} + \sum_{n: z_n = c} (\beta_n - \mu_c)(\beta_n - \mu_c)' \right)$$

**Updating  $\alpha$ .** The Metropolis algorithm to update  $\alpha$  proposes, at step  $h+1$ , a new value  $\alpha^*$  drawn from a symmetric proposal density  $\mathcal{N}_R(\alpha^{(h)}, \tau_1 \Omega)$ , where  $\tau_1$  is a tuning parameter. This proposal is accepted with probability

$$\min \left\{ 1, \frac{\prod_{n=1}^N L(y_n | \alpha^*, \beta_n^{(h)}) \phi(\alpha^* | \nu, \Omega)}{\prod_{n=1}^N L(y_n | \alpha^{(h)}, \beta_n^{(h)}) \phi(\alpha^{(h)} | \nu, \Omega)} \right\}.$$

If the proposal is accepted,  $\alpha^{(h+1)} = \alpha^*$ , otherwise  $\alpha^{(h+1)} = \alpha^{(h)}$ .

**Updating  $\beta$ .** We update  $\beta_n$  independently, by means of Metropolis algorithm. At the  $h + 1$  step of the algorithm, we use a  $\mathcal{N}_K(\beta_n^{(h)}, \tau_2 \Sigma_{z_n})$  as symmetric proposal density, with  $\tau_2$  being a tuning parameter, and we accept the new value  $\beta_n^*$ , drawn from it, with probability

$$\min \left\{ 1, \frac{L(y_n | \alpha^{(h)}, \beta_n^*) \phi(\beta_n^* | \mu_{z_n}, \Sigma_{z_n})}{L(y_n | \alpha^{(h)}, \beta_n^{(h)}) \phi(\beta_n^{(h)} | \mu_{z_n}, \Sigma_{z_n})} \right\}.$$

## 4 An application to public transport demand

The data set refers to a study carried out in Urbino (Italy) to analyse the attributes of the local public transport and to investigate possible interventions to improve the service (Marcucci and Scaccia (2005), Scaccia (2009)). Five attributes of the service were considered: cost of monthly ticket, headway, first and last run, real time information displays, bus shelters. Each attribute was further described by five levels. Questionnaires contained 15 choice exercises, 11 of which were random, 2 aimed at testing the quality of the answers, and 2 aimed at testing preference stability. Each choice exercise contained four hypothetical alternatives. A total number of 50 respondents took part in the study, providing a data set of 750 observations.

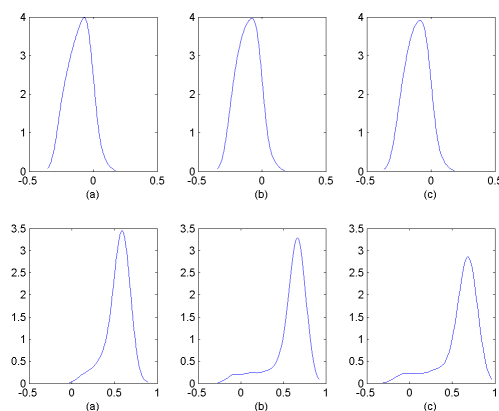
To specify the models, the Lagrange multiplier test (McFadden and Train (2000)) was used to decide which parameters are to be random. The null hypothesis of no mixing was rejected for the parameters of the attributes headway and daily operating time. The cost parameter was treated as non random to simplify the estimation of marginal willingness to pay for an improvement in a certain attribute (see Scaccia (2009)).

To estimate the proposed model, we performed 100,000 sweeps of the MCMC algorithm, allowing for a burn-in of 50,000 sweeps. Posterior means of relevant parameters are given in Table 1. The posterior estimates for the fixed parameters are very close to those obtained by Scaccia (2009). The signs of the cost and bus shelters parameters are as expected, while the information displays attribute seems to have a non significant influence on utility.

The marginal posterior densities for the random parameters are shown in Figure 1. The estimated posterior distribution of  $\beta_{\text{headway}}$  does not change when moving from the MMNL model with normal mixing density specification to the MOD models. In this case, a simple normal mixing density would have been appropriate to approximate taste heterogeneity with respect to the headway attribute. This shows how the MOD approach can also be seen as a diagnostic tool to get an idea of the shape of the true distribution and to help in the choice of an appropriate model. The estimated posterior distribution of  $\beta_{\text{run time}}$  is, instead, different under the three models. Under both the MOD models, a mass point at zero can be noticed, revealing the presence of

	MMNL		MOD (2 components)		MOD (3 components)	
Parameter	Est.	Std. dev.	Est.	Std. dev.	Est.	Std. dev.
$\alpha_{\text{cost}}$	-0.2671	0.0336	-0.2751	0.0346	-0.2764	0.0349
$\alpha_{\text{displays}}$	-0.0452	0.1753	-0.0190	0.1678	-0.0242	0.1828
$\alpha_{\text{shelters}}$	0.3258	0.1771	0.3167	0.1835	0.3320	0.1747
$\mu_{\text{headway}}$	-0.1039	0.0247	-0.0673	0.0806	-0.0485	0.1525
			-0.1116	0.0254	-0.0750	0.0864
					-0.1127	0.0265
$\mu_{\text{run time}}$	0.5322	0.0537	0.0403	0.1454	0.0180	0.1926
			0.6561	0.0798	0.0863	0.1697
					0.6840	0.0867
$s$	1.0000	0.0000	0.1599	0.0891	0.0503	0.0398
			0.8401	0.0891	0.1623	0.0818
					0.7873	0.0957

**Table 1.** Posterior mean estimates of relevant parameters.



**Fig. 1.** Estimated marginal posterior densities for the random parameters  $\beta_{\text{headway}}$  (upper panel) and  $\beta_{\text{run time}}$  (lower panel) under a) the MMNL model, b) the MOD model with 2 components, c) the MOD model with 3 components.

individuals that are indifferent to the availability of bus departures early in the morning or late at night. This interesting feature is not revealed by the MMNL model, which is not flexible enough to represent taste heterogeneity with respect to run time.

## 5 Conclusions and further development

We proposed modelling mixed logit models under a Bayesian hierarchical framework, making use of a mixture of normal distribution to approximate the density of the random parameters. This approach is conceptually simple and is flexible enough to approximate well a variety of distributions, allowing for multiple modes, as well as for saddle points in a distribution. Furthermore, it is possible to have point-mass at a specific value. Lastly, the mixture components can be used to classify the individuals into homogeneous groups, facilitating cluster analysis and classification. The approach proposed is not restricted to being based on the normal distribution but can use any continuous distribution. This could be an interesting avenue for further research.

## References

- ALLENBY, G.M., ARORA, N. and GINTER, J.L. (1998): On the heterogeneity of demand. *Journal of Marketing Research* 35 (3), 384-389.
- CIRILLO, C. and AXHAUSEN, K.W. (2006): Evidence on the distribution of values of travel-time savings from a six-week travel diary. *Transportation Research Part A: Policy and Practice* 40 (5), 444-457.
- FOSGERAU, M. (2006): Investigating the distribution of the value of travel time savings. *Transportation Research Part B: Methodological* 40 (8), 688-707.
- FOSGERAU, M. and HESS, S. (2009): A comparison of methods for representing random taste heterogeneity in discrete choice models. *European Transport Trasporti Europei* 42, 1-25.
- HESS, S., BIERLAIRE, M. and POLAK, J.W. (2005): Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice* 39 (2-3), 221-236.
- HO, R.K.W. and HU, I. (2008): Flexible modelling of random effects in linear mixed models-A Bayesian approach. *Computational Statistics and Data Analysis* 52 (3), 1347-1361.
- MARCUCCI, E. and SCACCIA, L. (2005): Alcune applicazioni dei modelli a scelta discreta al settore dei trasporti. In: E. Marcucci (Ed.): *I Modelli a Scelta Discreta nel Settore dei Trasporti. Teoria, Metodologia e Applicazioni*, Carocci editore, Roma.
- McFADDEN, D. (1974): Conditional logit analysis of qualitative choice behaviour. In: P.C. Zarembka (Ed.): *Frontiers in Econometrics*. Academic Press, New York, 105-142.
- McFADDEN, D. and TRAIN, K. (2000): Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15 (5), 447-470.
- SCACCIA, L. (2009): Random parameters logit models applied to public transport demand. *Global & Local Economic Review* 13 (2), 147-166.
- TRAIN, K. (1998): Recreation demand models with taste differences over people. *Land Economics*, 74 (2), 230-239.
- TRAIN, K. (2001): A Comparison of hierarchical bayes and maximum simulated likelihood for mixed logit. *Working Paper*, Department of Economics, University of California, Berkeley.

# A Decision Tree for Interval-valued Data with Modal Dependent Variable

Djamal Seck<sup>1</sup>, Lynne Billard<sup>2</sup>, Edwin Diday<sup>3</sup> and Filipe Afonso<sup>4</sup>

<sup>1</sup> Departement de Mathematiques et Informatique,  
Université Cheikh Anta Diop de Dakar, Senegal *djamal.seck@ucad.edu.sn*,

<sup>2</sup> Department of Statistics, University of Georgia  
Athens GA 30602 USA *lynne@stat.uga.edu*

<sup>3</sup> CEREMADE, University of Paris Dauphine  
75775 Paris Cedex 16 France *edwin.diday@ceremade.dauphine.fr*

<sup>4</sup> Syrokko, Aéroport de Roissy, Bat. Aéronef, 5 rue de Copenhague, 95731 Roissy  
Charles de Gaulle Cedex France, *afonso@syrokko.com*

**Abstract.** The CART (Breiman et al., 1984) methodology for classical data is extended to symbolic data in Seck (2010). This new methodology, called STREE, is capable of building a pure CART tree, a pure divisive hierarchy, or a weighted combination of both. This paper presents an application of STREE and compares its results with the traditional CART analysis.

**Keywords:** classification regression tree, divisive hierarchy tree, Fisher's iris data

## 1 Introduction

Breiman et al. (1984) introduced the classification and regression tree methodology (CART) for classical data. This now well-known methodology produces a divisive top-down hierarchical tree with regression methods informing the binary partition of each node in the tree's construction process.

This CART method has been adapted so as to construct a classification and regression tree (STREE) for so-called symbolic data in Seck(2010). Symbolic data typically are lists, intervals, histograms, and the like (see, e.g., Bock and Diday, 2000, and Billard and Diday, 2006). Thus, symbolic data take values as hypercubes or cartesian products of distributions in  $p$ -dimensional space  $\mathcal{R}^p$ , in contrast to classical data which are points in  $\mathcal{R}^p$ . However, classical values are special cases; e.g., the classical point  $x = a$  is equivalent to the symbolic interval  $x = [a, a]$ .

Both the classical CART method and the symbolic STREE method are divisive top-down methods, restricted to recursive binary partitions. The binary partition is induced by the variable which implies the best binary splitting of the variables for a given criterion. When there is a single categorical dependent variable (as in the classical case), the criterion used by both methods

is a discrimination criterion  $D(N)$  which measures the impurity of a node  $N$  with respect to the prior  $r^{th}$  partition  $P_r = \{C_1, \dots, C_r\}$ . The discrimination criterion is the Gini measure with the node impurity reaching a value of zero when only one class is present at a node. With priors estimated from class sizes, the Gini measure is computed as the sum of products of all pairs of class proportions for classes present at the node; it reaches its maximum value when class sizes at the node are equal. Therefore, from Breiman, et al. (1984),

$$D(N) = \sum_{i \neq f} p_i p_f = 1 - \sum_{i=1, \dots, r} p_i^2 \quad (1)$$

with  $p_i = n_i/n$ ,  $n_i = \text{card}(N \cap C_i)$  and  $n = \text{card}(P)$  in the classical case. For symbolic data,  $n_i$  is the number of individuals belonging to  $N$  which verify the current description of  $N$  and at the same time belong to  $C_i$ , and  $n$  is the total number of the individuals belonging to  $N$ . To normalize  $D(N)$ , we multiply by  $r/(r-1)$ ; where  $r$  is the number of prior classes; the normalized  $D(N)$  takes values in  $[0, 1]$ .

The STREE algorithm allows for the dependent variable to take modal-categorical values; in this case the discrimination criterion is the inertia associated with that variable (such as, e.g., the  $L_2$  distance; see Seck, 2010). This extends Périnel (1996, 1999), Limam (2005) and Winsberg et al. (2006) who developed decision trees for symbolic data with (non-modal) categorical dependent variables. When all the input variables have classical point values, the STREE methodology is the same as the CART methodology.

In addition to the pure decision tree (or, pure CART for classical data), the STREE methodology also allows for a pure hierarchical divisive tree (DIV) to be constructed, as well as a weighted average of both the pure decision tree and DIV trees to be built. The construction criteria are as follows.

Let  $D(N)$  be the criterion used to explain a partition of node  $N$  as in a CART analysis, and let  $H(N)$  be the inertia associated with the explanatory variables as in a pure hierarchy tree analysis. Then, as in Limam (2005), Limam et al. (2004) and Winsberg et al. (2006), we define the mixture as, for  $\alpha > 0$ ,  $\beta > 0$ ,

$$I = \alpha D(N) + \beta H(N) \quad \text{with} \quad \alpha + \beta = 1. \quad (2)$$

A pure decision tree corresponds to the case that  $\alpha = 0$ ; whereas a pure divisive tree corresponds to  $\alpha = 1$ . The criterion for the CART component ( $D(N)$ ) is as given in eqn (1) when there is a single categorical value for the dependent variable, or is the inertia when the dependent variable takes modal-categorical values. The inertia for the DIV component is the homogeneity criterion  $H(N)$  which is a squared distance measure between observations. Thus if the data consist of observations  $\{\omega_1, \dots, \omega_n\} \in \Omega$ , then

$$H(N) = \sum_{\omega_{i_1} \in \Omega} \sum_{\omega_{i_2} \in \Omega} \frac{p_{i_1} p_{i_2}}{2\mu} d^2(\omega_{i_1}, \omega_{i_2}) \quad (3)$$

where  $d(\omega_{i_1}, \omega_{i_2})$  is a distance measure between  $\omega_{i_1}$  and  $\omega_{i_2}$ ,  $p_i$  is the weight associated with  $\omega_i$  and  $\mu = \sum_{i=1}^N p_i$ . For example, observations can be equally weighted. The STREE algorithm uses an  $L2$  distance for a modal categorical variable, the Hausdorff distance for an interval variable, the Euclidean distance for a classical continuous variable and the  $(0, 1)$  distance for a classical categorical variable. Chavent (1998) has a divisive algorithm for intervals using the Hausdorff distance.

Concept	Species <sup>a</sup>	Sepal length	Sepal width	Petal length	Petal width
$\omega_1$	{1, 1.0}	[4.8, 5.4]	[3.3, 3.8]	[1.5, 1.9]	[0.2, 0.6]
$\omega_2$	{3,1.0}	[6.4, 6.9]	[3.0, 3.2]	[5.1, 5.5]	[2.0, 2.3]
$\omega_3$	{3,1.0}	[7.6, 7.7]	[2.6, 3.0]	[6.1, 6.9]	[2.0, 2.3]
$\omega_4$	{1,1.0}	[4.5, 4.5]	[2.3, 2.3]	[1.3, 1.3]	[0.3, 0.3]
$\omega_5$	{1,1.0}	[4.6, 4.7]	[3.2, 3.6]	[1.0, 1.4]	[0.2, 0.3]
$\omega_6$	{2,1.0}	[6.1, 6.4]	[2.8, 2.9]	[4.0, 4.3]	[1.3, 1.3]
$\omega_7$	{1,1.0}	[5.1, 5.4]	[3.7, 4.1]	[1.3, 1.7]	[0.1, 0.4]
$\omega_8$	{1,1.0}	[4.3, 4.4]	[2.9, 3.2]	[1.1, 1.4]	[0.1, 0.2]
$\omega_9$	{2,1.0}	[5.2, 5.6]	[2.3, 2.7]	[3.7, 4.0]	[1.0, 1.4]
$\omega_{10}$	{2, 1.0}	[5.9, 6.1]	[2.8, 3.4]	[4.2, 4.7]	[1.2, 1.6]
$\omega_{11}$	{2, 1.0}	[5.8, 6.0]	[2.2, 2.7]	[3.9, 4.1]	[1.0, 1.2]
$\omega_{12}$	{2,.9; 3,.1}	[4.9, 5.7]	[2.5, 3.0]	[4.1, 4.5]	[1.2, 1.7]
$\omega_{13}$	{1, 1.0}	[4.9, 5.5]	[3.2, 3.6]	[1.2, 1.5]	[0.1, 0.3]
$\omega_{14}$	{3, 1.0}	[7.7, 7.9]	[3.8, 3.8]	[6.4, 6.7]	[2.0, 2.2]
$\omega_{15}$	{1, 1.0}	[5.0, 5.0]	[3.0, 3.0]	[1.6, 1.6]	[0.2, 0.2]
$\omega_{16}$	{3, 1.0}	[7.1, 7.4]	[2.8, 3.2]	[5.8, 6.3]	[1.6, 2.1]
$\omega_{17}$	{2, 1.0}	[5.6, 5.7]	[2.6, 2.9]	[3.5, 3.6]	[1.0, 1.3]
$\omega_{18}$	{2,.2; 3,.8}	[6.0, 6.3]	[2.5, 2.8]	[4.8, 5.1]	[1.5, 1.9]
$\omega_{19}$	{2,.5; 3,.5}	[6.0, 6.3]	[2.2, 2.5]	[4.9, 5.0]	[1.5, 1.5]
$\omega_{20}$	{2,1.0}	[6.7, 7.0]	[2.8, 3.2]	[4.7, 5.0]	[1.4, 1.7]
$\omega_{21}$	{2,1.0}	[4.9, 5.1]	[2.0, 2.5]	[3.0, 3.5]	[1.0, 1.1]
$\omega_{22}$	{1,1.0}	[5.5, 5.7]	[4.2, 4.4]	[1.4, 1.5]	[0.2, 0.4]
$\omega_{23}$	{3,1.0}	[6.1, 6.7]	[2.5, 3.1]	[5.3, 5.8]	[1.4, 2.2]
$\omega_{24}$	{2,.11; 3,.89}	[5.6, 6.1]	[2.5, 3.2]	[4.8, 5.1]	[1.8, 2.4]
$\omega_{25}$	{3,1.0}	[6.2, 6.9]	[3.0, 3.4]	[5.4, 6.0]	[2.1, 2.5]
$\omega_{26}$	{1,1.0}	[5.7, 5.8]	[3.8, 4.0]	[1.2, 1.7]	[0.2, 0.3]
$\omega_{27}$	{3,1.0}	[7.2, 7.2]	[3.6, 3.6]	[6.1, 6.1]	[2.5, 2.5]
$\omega_{28}$	{2,1.0}	[6.3, 6.7]	[2.8, 3.3]	[4.4, 4.7]	[1.3, 1.6]
$\omega_{29}$	{1,1.0}	[4.6, 4.9]	[3.0, 3.4]	[1.4, 1.6]	[0.1, 0.3]
$\omega_{30}$	{2,1.0}	[6.2, 6.3]	[2.2, 2.3]	[4.4, 4.5]	[1.3, 1.5]

<sup>a</sup>Species identified by 1,2,3 for *setosa*, *versicolor*, *virginica*, respectively.

**Table 1: Fisher's Iris Data as Intervals**

More complete details of the STREE algorithm can be found in Seck(2010) and Seck et al. (2010). The focus of this paper is to apply the STREE method-

ology to random variables which take interval values and dependent variable which takes modal categorical values.

## 2 The data

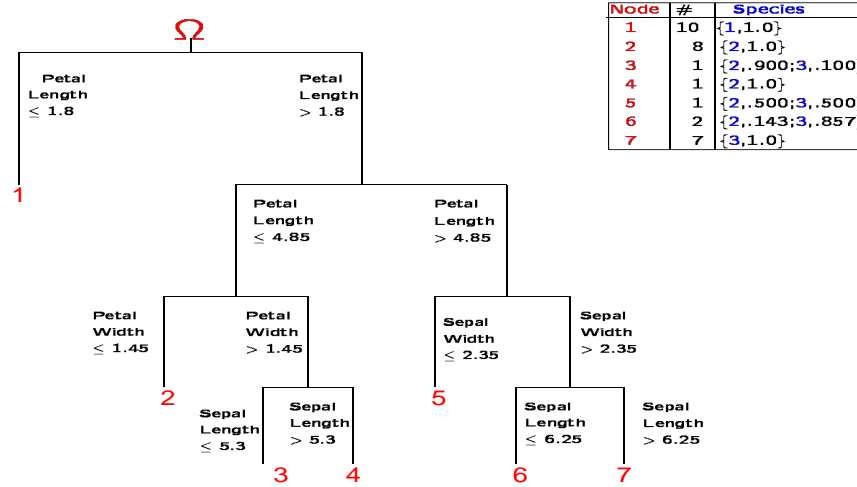
The new methodology will be illustrated on the familiar Fisher (1936) Iris data. There are 150 observations, 50 each from the species *setosa*, *versicolor*, *virginica*. There are  $p = 4$  variables,  $Y_1 =$  Sepal Length,  $Y_2 =$  Sepal Width,  $Y_3 =$  Petal Length,  $Y_4 =$  Petal Width. These 150 observations were clustered into 30 sets of observations using the  $k$ -means clustering methodology based on the 4 random variables. The observations within each set formed a new symbolic-valued observation taking now interval values for the  $Y_j, j = 1, \dots, 4$ ; each set had a distribution of the variable species in the form  $\{setosa, p_1; versicolor, p_2; virginica, p_3\}$  where  $p_k$  is the relative frequency for species  $k = 1, 2, 3$ . The data now assume the form shown in Table 1. Thus, in the observation  $\omega_{12}$ , 90% are from the species *versicolor* and 10% are from the species *virginica*; the aggregated species now assume interval values, here e.g.,  $Y_1 = [4.9, 5.7]$  implies that species aggregated to form  $\omega_1$  have a sepal length from 4.9 to 5.7 units of measurement. In contrast, all species in the aggregation of  $\omega_{10}$  come from the species *versicolor*. Note that observation  $\omega_4$  corresponds to a classical value, in this case from the species *setosa*.

## 3 STREE analyses

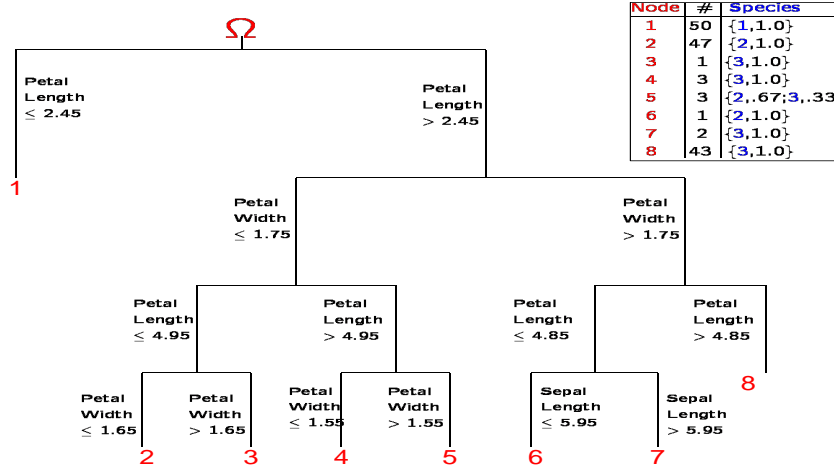
First, the STREE algorithm was applied to the set of 30 interval-valued iris data, with  $\alpha = 0$  and equal weights to obtain a pure decision tree, shown in Figure 1. Since classical data are a special case of interval data, the algorithm was also applied to the original 150 classical observations, with  $\alpha = 0$  to obtain the pure CART tree of Figure 2. These figures also show the proportions of each species at each of the respective terminal nodes.

It is immediately observed that node 1 in each case consists entirely and exhaustively of the species *setosa*. Further, each emerges from the first partitioning of the entire dataset  $\Omega$  and are each based on the Petal Length. Nodes 2, 3 and 4 in Figure 1 are almost all from the species *versicolor* and match node 2 in Figure 2 which node is entirely *versicolor*. Both sets of nodes were reached through cut points that are not dissimilar, running in order through Petal Length, Petal Width, Petal Length, and Petal Width, with the final cut variable being Petal. Likewise, the consistencies between the two trees occur also for the species *virginica*. In this case, the node 7 in Figure 1 and nodes 7 and 8 of Figure 2 account for most of this species. As for the other two species, not surprisingly, the first cutting variables are not dissimilar, in both cases being Petal Length and Petal Width with slightly varying cut values. Interestingly though for these nodes, the third cutting variable differs, with





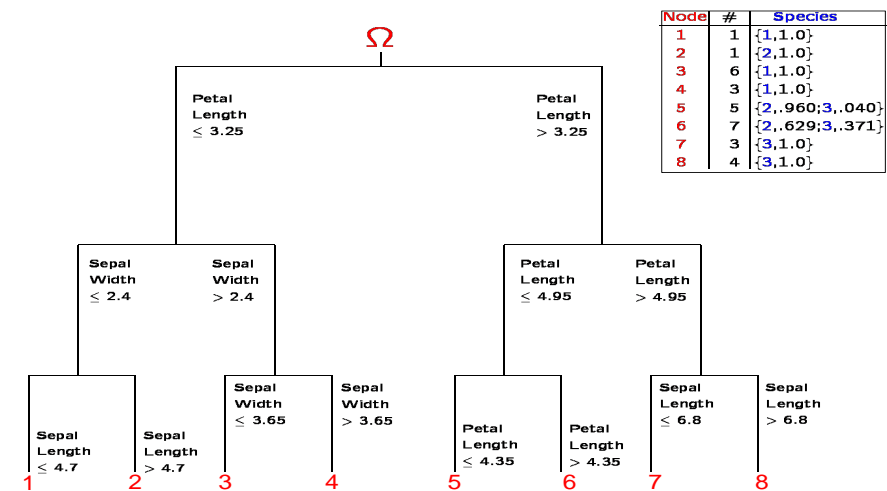
**Fig. 1.** Pure Decision tree on 30 Iris intervals:  $\alpha = 0$  (Species 1,2,3  $\equiv$  *setosa*, *versicolor*, *virginica*)



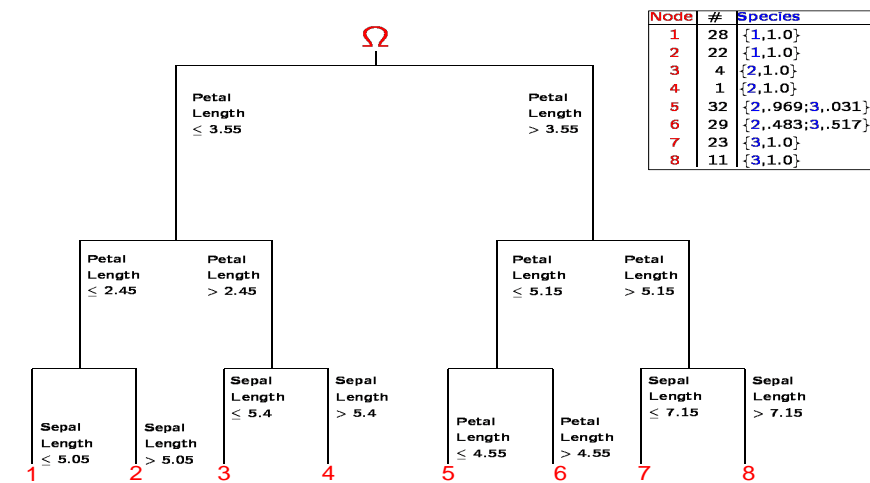
**Fig. 2.** Pure CART tree on original 150 Iris observations:  $\alpha = 0$  (Species 1,2,3  $\equiv$  *setosa*, *versicolor*, *virginica*)

Sepal Width and Petal Length being the cutting variable for the symbolic and classical analyses, respectively.

A pure divisive hierarchy was also constructed where now  $\alpha = 1$ ; see Figure 3. The corresponding DIV tree constructed on the original 150 observations is displayed in Figure 4. As for the comparison of the CART analyses of Figures 1 and 2, a comparison of the two DIV analyses, Figures 3 and 4, shows similar consistencies across the two trees. However, while the CART



**Fig. 3.** Pure DIV tree on 30 Iris intervals:  $\alpha = 1$  (Species 1,2,3  $\equiv$  *setosa*, *versicolor*, *virginica*)



**Fig. 4.** Pure DIV tree on original 150 Iris observations:  $\alpha = 1$  (Species 1,2,3  $\equiv$  *setosa*, *versicolor*, *virginica*)

analyses tends to find clusters which are predominately of the same species, the DIV analyses tend to obtain clusters that are genuine mixtures of the three species (as at nodes 5 and 6), reflecting the different criteria used in the construction.

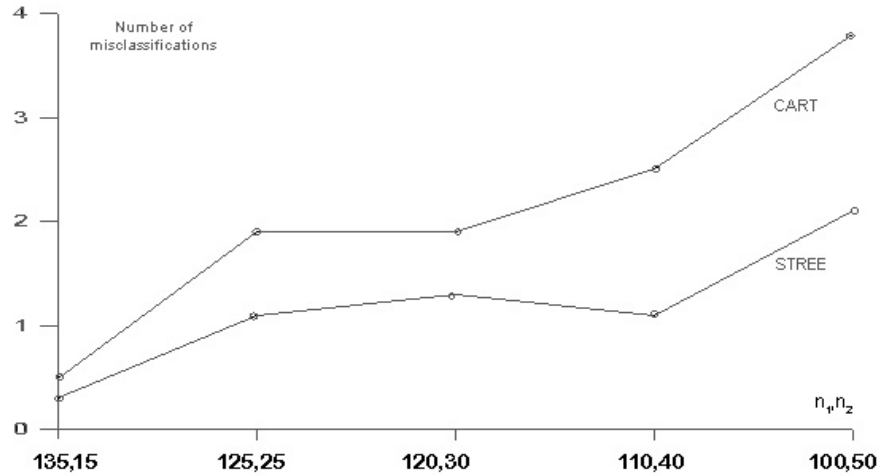


Fig. 5. Comparison of STREE and CART Misclassifications for Test Subsets

#### 4 Comparison of STREE and CART Algorithms

A comparison of the STREE and CART algorithms is performed by randomly dividing the full set of the original 150 classical observations into a training subset and a test subset of sizes  $n_1$  and  $n_2$ , respectively, with  $n_1 + n_2 = 150$ . For the training and test subsets, the  $n_1$  and  $n_2$  observations are drawn randomly from the original set. The CART algorithm is then run on this training subset. The procedure is repeated ten times for each set of  $(n_1, n_2)$ .

For the STREE analysis, we first find 30 interval clusters in the  $n_1$  classical observations by running STREE as a divisive algorithm (i.e., DIV, with  $\alpha = 1$ ,  $\beta = 0$ ) on this training subset. We then conduct a decision tree analysis on the resulting (training subset of) clusters with the STREE algorithm where now  $\alpha = 0$ ,  $\beta = 1$ . After that, the tree is tested with the test subset (i.e., the remaining  $n_2$  observations) to obtain the number of misclassifications. This was repeated ten times.

Figure 5 displays the plot of the average number of misclassifications (over the ten repetitions) for each of the  $n_1$  and  $n_2$  subsets considered. From these results, it is clear that the STREE algorithm works well, and indeed has a lower rate of misclassifications than does the pure CART algorithm. It is also observed that as the relative size of the test subset increases, the rate of misclassifications increases, not surprisingly.

## 5 Conclusion

This paper has applied the new STREE algorithm for constructing decision trees to interval-valued data. Further details along with numerous other applications and to other types of symbolic data are detailed in Seck (2010). This includes examination of validation procedures. Also, Seck (2010) has included bagging and boosting components to the STREE algorithm. How these are incorporated for symbolic data will be presented elsewhere.

## References

- BILLARD, L. and DIDAY, E. (2006): *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley, London.
- BOCK, H. H. and DIDAY, E. (Eds.) (2000): *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin.
- BREIMAN, L., FREIDMAN J. H., OLSHEN, R. A. and STONE C. J. (1984): *Classification and Regression Trees*. Wadsworth.
- CHAVENT, M. (1998): A monothetic clustering algorithm. *Pattern Recognition Letters* 19, 989-996.
- FISHER, R. A. (1936): The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188.
- LIMAM, M. M. (2005): *Méthodes de Description de Classes Combinant Classification et Discrimination en Analyse des Données Symboliques*. Université Paris.
- LIMAM, M. M., DIDAY, E. and WINSBERG, S. (2004): Probabilist allocation of aggregated statistical units in classification trees for symbolic class description. In: D. Banks, L. House, F. R. McMorris, P. Arabie and W. Gaul (Eds.): *Classification, Clustering and Data Mining Applications*. Springer, Heidelberg, 371-379.
- PÉRINEL, E. (1996): *Segmentation et Analyse des Données Symbolique: Applications à des Données Probabilités Imprécises*. Doctoral Dissertation, Université Paris Dauphine.
- PÉRINEL, E. (1999): Binary discrimination tree construction from imprecise data. *Revue de Statistique Appliquée* 47, 5-30.
- SECK, D. (2010): *Thèse de Doctorat*, Université Paris Dauphine.
- SECK, D., DIDAY, E. and BILLARD, L. (2010): STREE: A classification and regression tree for symbolic data. *Technical Report* (in preparation).
- WINSBERG, S., DIDAY, E. and LIMAM, M. M. (2006): A tree structured classifier for symbolic class description. In: A. Rizzi and M. Vichi (Eds.) *COMPSTAT: Proceedings in Computational Statistics 17th Symposium*. Springer, 927-936.

# The Set of $3 \times 4 \times 4$ Contingency Tables has 3-Neighborhood Property

Toshio Sumi<sup>1</sup> and Toshio Sakata<sup>2</sup>

<sup>1</sup> Faculty of Design, Kyushu University  
4-9-1 Shiobaru, Minami-ku, Fukuoka, 815-8540, Japan,  
*sumi@design.kyushu-u.ac.jp*

<sup>2</sup> Faculty of Design, Kyushu University  
4-9-1 Shiobaru, Minami-ku, Fukuoka, 815-8540, Japan,  
*sakata@design.kyushu-u.ac.jp*

**Abstract.** We consider the sequential conditional test for three-way contingency tables. Conditional tests of no interaction for three-way contingency tables use as the frame of conditional inference the set of all contingency tables with three fixed two-way marginal tables. Lifting between three-way contingency tables means a method of calculating the frame  $\Omega_t$  of the  $t$ -stage from  $\Omega_{t-1}$  of the  $(t-1)$ -stage, which makes it easy to perform the sequential conditional test efficiently. In the previous paper, Sakata and Sumi (COMPSTAT'2008), we treated  $3 \times 3 \times 3$  tables and 2-neighborhood property. As a continuation, in this paper, we treat  $3 \times 4 \times 4$  tables and show that the conditional inference frame  $\Omega_t$  is obtained from  $\Omega_{t-1}$  by transformations made by at most three elements of Markov basis for  $3 \times 4 \times 4$  contingency tables, that is, 3-neighborhood property.

**Keywords:**  $3 \times 4 \times 4$  contingency tables, sequential conditional test, Markov basis, 3-neighborhood property

## 1 Introduction

For three-way contingency tables conditional tests are commonly used (for example see Agresti (2007)). For the problem of testing no three factor interaction of a three-way table the conditional test treats a distribution on the set of all contingency tables with three fixed two-way tables. These set of contingency tables are called the frame of conditional inference, or in short, the frame. In this paper we consider a sequential conditional test and address the problem how to construct the frames of the conditional inference in the sequential conditional test. Let  $\Omega_t$  be the frame at the  $t$ -stage of the sequential conditional test. Then we have a sequence of frames in an experiment,

$$\Omega_1 \rightarrow \Omega_2 \rightarrow \cdots \rightarrow \Omega_{t-1} \rightarrow \Omega_t \rightarrow \cdots,$$

and it is not efficient to calculate each  $\Omega_t$  from scratch at each time. Thus we address the following problem: how can we calculate  $\Omega_t$  from  $\Omega_{t-1}$  successively, where  $\Omega_t$  is the set of marginal fixed contingency tables at the

$t$ -stage. For this purpose we proposed a method by using a Markov basis in the previous paper (Sakata and Sumi (2008)). Note that for each of the set of  $I \times J \times 2$  contingency tables, the set of  $3 \times 3 \times K$  contingency tables and the set of  $3 \times 4 \times K$  contingency tables there is the unique minimal Markov basis respectively, (Sturmfels (1996), Aoki and Takemura (2003), and Aoki (2004)). Using these bases, we showed that the set of  $I \times J \times 2$  contingency tables has 1-neighborhood property, the set of  $3 \times 3 \times K$  contingency tables has 2-neighborhood property for  $K = 3$  (Sakata and Sumi (2008), Sumi and Sakata (2009a)) and 3-neighborhood property for  $K \geq 4$  (Sumi and Sakata (2009b)). In this paper, we show that the set of  $3 \times 4 \times 4$  contingency tables has 3-neighborhood property. Here we define that the set of contingency tables has  $r$ -neighborhood property if any element  $H$  with  $H_{ijk} = 0$  in the set has an element  $H'$  with  $H'_{ijk} = 1$  within its  $r$ -neighborhood and no element  $H'$  with  $H'_{ijk} = 1$  within its  $(r - 1)$ -neighborhood. Fix  $i_1, j_1$ , and  $k_1$ . We denote by  $\Phi(I, J, K)$  the set of  $I \times J \times K$  contingency tables  $H$  such that  $H_{i_1 j_1 k_1} = 0$  and  $H$  can be transformed to  $H'$  with  $H'_{i_1 j_1 k_1} > 0$ .

We get the following theorems.

**Theorem 1.** *There is a family of 105 tables satisfying that for arbitrary  $H \in \Phi(3, 4, 4)$ , there are a table  $H'$  of the family and permutations  $\sigma_I, \sigma_J$  and  $\sigma_K$  on  $\{1, \dots, 3\}, \{1, \dots, 4\}, \{1, \dots, 4\}$  preserving  $i_1, j_1, k_1$ , respectively such that  $H_{ijk} \geq H'_{\sigma_I(i)\sigma_J(j)\sigma_K(k)}$  for each  $i, j, k$ .*

**Theorem 2.** *Let  $H \in \Omega(3, 4, 4)$ . Let  $H'$  be a contingency table made by  $H'_{ijk} = 2$  if  $H_{ijk} \geq 2$  and  $H'_{ijk} = H_{ijk}$  otherwise. Then  $H \in \Phi(3, 4, 4)$  if and only if  $H' \in \Phi(3, 4, 4)$ .*

This paper is organized as follows. We define subordination in section 2 and in section 3, we characterize that a set of all contingency tables with fixed marginals has no table  $H$  with  $H_{i_1 j_1 k_1} > 0$  by using subordination (see Theorem 5). In section 4, we give a minimal set of moves obtaining the frame of  $t$ -stage from the frame of  $(t - 1)$ -stage (see Theorem 6). In the last section we show the computational algorithms to obtain results.

## 2 Subordination and non-movableness

Let  $\sigma_I, \sigma_J$  and  $\sigma_K$  be permutations on  $\{1, \dots, I\}, \{1, \dots, J\}, \{1, \dots, K\}$ , respectively and let  $H$  be an  $I \times J \times K$  table. For the triad  $\sigma = (\sigma_I, \sigma_J, \sigma_K)$  of permutations, we denote by  $\sigma(H)$  a table whose  $(i, j, k)$ -entry is  $H_{\sigma_I(i), \sigma_J(j), \sigma_K(k)}$  for each  $i, j, k$ . Recall that  $\Phi(I, J, K)$  depends on  $i_1, j_1$  and  $k_1$ . Suppose that  $\sigma_I(i_1) = \sigma_J(j_1) = \sigma_K(k_1) = 1$ . Then the set  $\sigma(\Phi(I, J, K))$  is a set consisting of tables  $H$  with  $H_{111} = 0$  which is reachable to  $H'$  with  $H'_{111} > 0$ . Thus we may assume that  $i_1 = j_1 = k_1 = 1$  without loss of the generality.

**Definition 11.** Let  $H = (H_{ijk})$  and  $H' = (H'_{ijk})$  be an  $I \times J \times K$  table and an  $I \times J \times K'$  table, respectively, with  $H_{i_1 j_1 k_1} = H'_{i_1 j_1 k_1} = 0$ . We call

that  $H$  is  $K$ -subordinated to  $H'$  if there is a partition  $\mathbb{P} = P_1, \dots, P_{K'}$  on  $\{1, 2, \dots, K\}$  such that

- a.  $P_1 \sqcup \dots \sqcup P_{K'} = \{1, 2, \dots, K\}$ ,
- b.  $P_k \neq \emptyset$  for any  $k$ , and
- c.  $\sum_{\ell \in P_k} H_{ij\ell} \leq H'_{ijk}$  for any  $i, j, k$ .

We define ‘ $I$ -subordinated’ and ‘ $J$ -subordinated’ similarly.

Note that  $K' \leq K$  and that  $H$  is  $K$ -subordinated to  $H$  itself. The following proposition is important for detecting non-movableness by smaller tables.

**Proposition 18.** *If  $H$  is  $K$ -subordinated to  $H' \notin \Phi(I, J, K')$  then it holds that  $H \notin \Phi(I, J, K)$ . If  $H$  is  $I$ -subordinated to  $H' \notin \Phi(I', J, K)$  then  $H \notin \Phi(I, J, K)$ . If  $H$  is  $J$ -subordinated to  $H' \notin \Phi(I, J', K)$  then  $H \notin \Phi(I, J, K)$ .*

*Proof.* We show only the first assertion since the other assertion can be similarly shown. Suppose that  $H$  is  $K$ -subordinated to  $H'$  by some partition  $\mathbb{P} = P_1, \dots, P_{K'}$ . Let  $\phi_{\mathbb{P}}$  be a map from the set of  $I \times J \times K$  tables to the set of  $I \times J \times K'$  tables, sending  $G = (G_{ijk})$  to  $G' = (G'_{ijk})$  where

$$G'_{ijk} = \sum_{\ell \in P_k} G_{ij\ell}$$

for each  $i, j, k$ . Let suppose that  $H \in \Phi(I, J, K)$ . The set  $\Omega$  of contingency tables including the table  $H$  has a table  $G$  with  $G_{i_1 j_1 k_1} > 0$ . Then the image  $\phi_{\mathbb{P}}(\Omega)$  is a set of contingency tables including the table  $\phi_{\mathbb{P}}(H)$  and a table  $\phi_{\mathbb{P}}(s)$  with  $\phi_{\mathbb{P}}(G)_{i_1 j_1 k_1} > 0$ . Therefore,  $\phi_{\mathbb{P}}(H) \in \Phi(I, J, K')$  and then  $H' \in \Phi(I, J, K')$ . We complete the proof.

**Definition 12.** Let  $H = (H_{ijk})$  and  $H' = (H'_{ijk})$  be an  $I \times J \times K$  table and an  $I' \times J' \times K'$  table, respectively, with  $H_{i_1 j_1 k_1} = H'_{i_1 j_1 k_1} = 0$ . We call that  $H$  is *subordinated* to  $H'$  if there are tables  $G$  and  $G'$  such that  $H$  is  $I$ -subordinated to  $G$ ,  $G$  is  $J$ -subordinated to  $G'$ , and  $G'$  is  $K$ -subordinated to  $H'$ .

Since the subordination does not depends on the order of  $I$ -,  $J$ -,  $K$ -subordination we have the following theorem.

**Theorem 3.** *If  $H$  is subordinated to  $H' \notin \Phi(I, J, K')$  then  $H \notin \Phi(I, J, K)$ .*

### 3 Non-movable tables

Sumi and Sakata (2009b) obtained that the minimal set of tables to detect the non-movableness for the set  $\Omega(3, 3, K)$  of  $3 \times 3 \times K$  contingency tables.

**Theorem 4 (Sumi and Sakata (2009b)).** *Let  $H \in \Omega(3, 3, K)$ . If  $H \notin \Phi(3, 3, K)$  then  $H$  is subordinated to one of the following tables and their permuting tables for permutations preserving 1 on each coordinate:*

$$\begin{array}{ccc} \begin{array}{c} 0* \quad ** \\ ** \quad *0 \end{array}, & \begin{array}{c} 0*0 \quad *** \quad **0 \\ *** \quad *0* \quad *00 \\ 0** \quad 00* \quad *** \end{array}, & \begin{array}{c} 0*0 \quad *** \quad 0** \\ *** \quad *0* \quad 00* \\ **0 \quad *00 \quad *** \end{array}, \\ (2a) & (3a) & (3b) \end{array}$$

$$\begin{array}{ccc} \begin{array}{c} 0** \quad *** \quad 00* \\ *** \quad *00 \quad *0* \\ 0*0 \quad **0 \quad *** \end{array}, & \begin{array}{c} 0*0 \quad *** \quad 0** \\ *** \quad *00 \quad *** \\ *** \quad *0* \quad 00* \end{array}, & \begin{array}{c} 0** \quad *** \quad 0*0 \\ *** \quad *00 \quad **0 \\ 00* \quad *0* \quad *** \end{array}. \\ (3c) & (3d) & (3e) \end{array}$$

Here  $*$  means a sufficient large integer which is sufficient to be  $\max_{i,j,k} H_{ijk}$  for  $H$ .

Let  $\mathfrak{N}(1, 1, 1)$  be the set consisting of the following 26 tables and their permuting tables for permutations preserving 1 on each coordinate:

$$\begin{array}{l} \begin{array}{c} 0**0 \quad **** \quad 0*** \quad 00*0 \\ **** \quad *00* \quad 000* \quad 00** \\ **00 \quad *000 \quad **0* \quad **** \end{array}, \begin{array}{c} 0**0 \quad **** \quad ***0 \quad 00*0 \\ **** \quad *00* \quad *000 \quad *0*0 \\ 0*0* \quad 000* \quad **0* \quad **** \end{array}, \begin{array}{c} 0*** \quad **** \quad 000* \quad 00** \\ ***0 \quad *000 \quad **** \quad *0*0 \\ 0*00 \quad **00 \quad 0*0* \quad **** \end{array}, \\ \begin{array}{c} 0**0 \quad *0*0 \quad **** \quad 00*0 \\ **** \quad *000 \quad *00* \quad *0** \\ 0*00 \quad **** \quad 0*0* \quad 0*** \end{array}, \begin{array}{c} 0**0 \quad **** \quad 00** \quad 00*0 \\ **** \quad *00* \quad 000* \quad *0** \\ 0*00 \quad **00 \quad **** \quad ***0 \end{array}, \begin{array}{c} 0*00 \quad **** \quad **0* \quad 0*0* \\ **** \quad 00*0 \quad *0** \quad 00** \\ **00 \quad *0*0 \quad *000 \quad **** \end{array}, \\ \begin{array}{c} 0*00 \quad **** \quad 0**0 \quad 0*0* \\ **** \quad *0*0 \quad 00*0 \quad **** \\ **0* \quad *00* \quad **** \quad 000* \end{array}, \begin{array}{c} 0*00 \quad **** \quad ***0 \quad **00 \\ **** \quad 000* \quad *0** \quad *00* \\ 0**0 \quad 00** \quad 00*0 \quad **** \end{array}, \begin{array}{c} 0*00 \quad **00 \quad **** \quad 0*0* \\ **** \quad *000 \quad *0*0 \quad **** \\ 0*** \quad **** \quad 00** \quad 000* \end{array}, \\ \begin{array}{c} 0*10 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *000 \quad *0** \quad *00* \\ 0**0 \quad *0*0 \quad 00*0 \quad **** \end{array}, \begin{array}{c} 0*10 \quad **** \quad 0*** \quad 0**0 \\ **** \quad *000 \quad *0** \quad *0*0 \\ 0*0* \quad *00* \quad 000* \quad **** \end{array}, \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *000 \quad *0** \quad *00* \\ 1**0 \quad *0*0 \quad 00*0 \quad **** \end{array}, \\ \begin{array}{c} 0*00 \quad **** \quad 1*0* \quad 0*** \\ **** \quad *000 \quad *00* \quad *0** \\ 0**0 \quad *0*0 \quad **** \quad 00*0 \end{array}, \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *000 \quad *0** \quad *00* \\ 0**1 \quad *0*0 \quad 00*0 \quad **** \end{array}, \begin{array}{c} 0*00 \quad **** \quad 1*** \quad 0*0* \\ **** \quad *000 \quad *0** \quad *00* \\ 0*0* \quad *0*0 \quad 00*0 \quad **** \end{array}, \\ \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *000 \quad *0** \quad *10* \\ 0**0 \quad *0*0 \quad 00*0 \quad **** \end{array}, \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *000 \quad *0** \quad *01* \\ 0**0 \quad *0*0 \quad 00*0 \quad **** \end{array}, \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *000 \quad *1** \quad *00* \\ 0**0 \quad *0*0 \quad 00*0 \quad **** \end{array}, \\ \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *000 \quad *0** \quad *00* \\ 0**0 \quad *0*0 \quad 00*1 \quad **** \end{array}, \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *000 \quad *0** \quad *00* \\ 0**0 \quad *0*0 \quad 00*1 \quad **** \end{array}, \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *001 \quad *0** \quad *00* \\ 0**0 \quad *0*0 \quad 00*0 \quad **** \end{array}, \\ \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *000 \quad *0** \quad *00* \\ 0**0 \quad *1*0 \quad 00*0 \quad **** \end{array}, \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *000 \quad *0** \quad *00* \\ 0**0 \quad *0*1 \quad 00*0 \quad **** \end{array}, \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *000 \quad *0** \quad *00* \\ 0**0 \quad *0*0 \quad 01*0 \quad **** \end{array}, \\ \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*1* \\ **** \quad *000 \quad *0** \quad *00* \\ 0**0 \quad *0*0 \quad 00*0 \quad **** \end{array}, \begin{array}{c} 0*00 \quad **** \quad 0*** \quad 0*0* \\ **** \quad *010 \quad *0** \quad *00* \\ 0**0 \quad *0*0 \quad 00*0 \quad **** \end{array}. \end{array}$$

The first 9 tables are isolated and the others are movable to each other. Here, for a contingency table  $H$ , we say that  $H$  is *isolated* if there is no Markov move transforming  $H$ .

**Theorem 5.** *A contingency table  $H$  is not reachable to a table  $H'$  with  $H'_{111} > 0$  by Markov moves if and only if  $H$  is subordinate of a table in (2a), (3a)–(3e) or  $\mathfrak{N}(1, 1, 1)$ .*



## 4 Movable tables

By  $M = M^{(1)} \triangleright M^{(2)} \triangleright \dots \triangleright M^{(s)}$  we denote an operation by transformations plusing  $M^{(1)}$ , next  $M^{(2)}$ , step by step, and finally  $M^{(s)}$ . Let  $\mathfrak{M}(1, 1, 1)$  be the set including the below 105 tables and their permuting tables for permutations preserving 1 on each coordinate:

$222_4(12, 12, 12)$ ,  $233_6(12, 132, 123)$ ,  $332_6(123, 213, 21)$ ,  $323_6(132, 12, 123)$ ,  
 $343_8(312, 4321, 321)$ ,  $343_8(312, 2314, 132)$ ,  $343_8(213, 1342, 213)$ ,  
 $343_8(123, 1234, 123)$ ,  $334_8(312, 321, 3421)$ ,  $334_8(123, 123, 1243)$ ,  
 $334_8(312, 132, 2413)$ ,  $334_8(213, 213, 1432)$ ,  $344_9(132, 3241, 2134)$ ,  
 $344_9(123, 1432, 1342)$ ,  $344_{10}(231, 3421, 1234)$ ,  $344_{10}(132, 3412, 1324)$ ,  
 $344_9(132, 2134, 3241)$ ,  $344_{10}(123, 2341, 2341)$ ,  $344_{10}(132, 1324, 3412)$ ,  
 $344_{10}(231, 1243, 4321)$ ,  $244_8(12, 1342, 1234)$ ,  
 $222_4(13, 13, 32) \triangleright 222_4(12, 12, 13)$ ,  $222_4(32, 23, 32) \triangleright 222_4(12, 12, 12)$ ,  
 $222_4(13, 32, 13) \triangleright 222_4(12, 13, 12)$ ,  $222_4(23, 13, 31) \triangleright 222_4(13, 12, 12)$ ,  
 $222_4(13, 32, 12) \triangleright 233_6(12, 143, 123)$ ,  $222_4(13, 34, 43) \triangleright 233_6(12, 132, 124)$ ,  
 $222_4(32, 24, 34) \triangleright 233_6(12, 132, 124)$ ,  $222_4(13, 13, 42) \triangleright 233_6(12, 142, 143)$ ,  
 $222_4(32, 34, 24) \triangleright 233_6(12, 142, 123)$ ,  $222_4(13, 23, 31) \triangleright 233_6(12, 143, 124)$ ,  
 $222_4(23, 13, 41) \triangleright 233_6(13, 142, 123)$ ,  $222_4(13, 13, 32) \triangleright 332_6(123, 213, 31)$ ,  
 $222_4(13, 14, 32) \triangleright 332_6(132, 123, 13)$ ,  $222_4(12, 32, 13) \triangleright 332_6(123, 314, 21)$ ,  
 $222_4(12, 13, 42) \triangleright 323_6(132, 12, 143)$ ,  $222_4(12, 23, 34) \triangleright 323_6(123, 21, 214)$ ,  
 $222_4(13, 32, 13) \triangleright 323_6(123, 31, 214)$ ,  $222_4(13, 13, 42) \triangleright 343_8(312, 4321, 341)$ ,  
 $222_4(12, 13, 42) \triangleright 343_8(123, 1234, 134)$ ,  $222_4(13, 12, 32) \triangleright 343_8(312, 2314, 143)$ ,  
 $222_4(13, 13, 42) \triangleright 343_8(213, 1342, 413)$ ,  $222_4(13, 13, 42) \triangleright 343_8(132, 2143, 413)$ ,  
 $222_4(13, 13, 32) \triangleright 343_8(312, 2314, 143)$ ,  $222_4(13, 14, 42) \triangleright 343_8(132, 2143, 413)$ ,  
 $222_4(12, 43, 23) \triangleright 343_8(213, 1342, 214)$ ,  $222_4(32, 14, 42) \triangleright 343_8(213, 1342, 213)$ ,  
 $222_4(23, 12, 31) \triangleright 343_8(213, 2314, 142)$ ,  $222_4(12, 32, 13) \triangleright 343_8(123, 1324, 124)$ ,  
 $222_4(12, 14, 42) \triangleright 334_8(132, 213, 4132)$ ,  $222_4(12, 24, 43) \triangleright 334_8(213, 213, 1342)$ ,  
 $222_4(23, 34, 42) \triangleright 334_8(312, 421, 4321)$ ,  $222_4(13, 23, 41) \triangleright 334_8(312, 431, 3421)$ ,  
 $222_4(13, 23, 41) \triangleright 334_8(123, 134, 1234)$ ,  $222_4(13, 23, 41) \triangleright 334_8(312, 143, 2413)$ ,  
 $222_4(13, 32, 14) \triangleright 334_8(213, 314, 1432)$ ,  $222_4(13, 23, 21) \triangleright 334_8(312, 143, 2413)$ ,  
 $222_4(12, 34, 41) \triangleright 334_8(123, 124, 1243)$ ,  $222_4(13, 23, 41) \triangleright 344_{10}(231, 4231, 1243)$ ,  
 $222_4(13, 14, 42) \triangleright 344_{10}(231, 1243, 3241)$ ,  $222_4(13, 43, 34) \triangleright 344_{10}(123, 2341, 2431)$ ,  
 $233_6(13, 124, 321) \triangleright 222_4(12, 13, 13)$ ,  $233_6(13, 123, 431) \triangleright 222_4(12, 13, 12)$ ,  
 $233_6(13, 134, 423) \triangleright 222_4(12, 12, 14)$ ,  $233_6(32, 423, 432) \triangleright 222_4(12, 12, 12)$ ,  
 $233_6(13, 234, 413) \triangleright 222_4(12, 13, 12)$ ,  $233_6(32, 143, 143) \triangleright 222_4(13, 12, 12)$ ,  
 $233_6(13, 123, 421) \triangleright 233_6(12, 143, 143)$ ,  $233_6(13, 123, 321) \triangleright 233_6(12, 143, 124)$ ,  
 $233_6(13, 134, 423) \triangleright 332_6(123, 213, 41)$ ,  $233_6(13, 134, 423) \triangleright 343_8(132, 2143, 413)$ ,  
 $233_6(13, 123, 421) \triangleright 233_6(12, 142, 143)$ ,  $343_8(213, 4123, 431) \triangleright 222_4(12, 13, 12)$ ,  
 $343_8(213, 4123, 341) \triangleright 323_6(132, 13, 124)$ ,  $343_8(213, 4123, 421) \triangleright 233_6(12, 143, 123)$ ,  
 $334_8(213, 134, 2143) \triangleright 222_4(12, 12, 14)$ ,  $334_8(213, 341, 3124) \triangleright 222_4(12, 12, 14)$ ,  
 $334_8(213, 321, 4123) \triangleright 233_6(12, 142, 134)$ ,  $334_8(213, 341, 3124) \triangleright 332_6(123, 214, 41)$ ,  
 $344_9(132, 2413, 2314) \triangleright 222_4(13, 12, 12)$ ,  $344_9(123, 4213, 3124) \triangleright 222_4(12, 13, 12)$ ,  
 $344_9(123, 4123, 3214) \triangleright 222_4(12, 12, 14)$ ,  
 $222_4(13, 23, 31) \triangleright 222_4(23, 34, 23) \triangleright 222_4(12, 13, 12)$ ,  
 $222_4(12, 14, 32) \triangleright 222_4(23, 13, 21) \triangleright 222_4(13, 12, 13)$ ,  
 $222_4(13, 13, 42) \triangleright 222_4(32, 23, 34) \triangleright 222_4(12, 12, 14)$ ,  
 $222_4(12, 32, 13) \triangleright 222_4(32, 12, 14) \triangleright 222_4(13, 13, 12)$ ,

$$\begin{aligned}
&222_4(12, 13, 32) \triangleright 222_4(13, 14, 43) \triangleright 222_4(12, 12, 14), \\
&222_4(12, 34, 43) \triangleright 222_4(13, 13, 42) \triangleright 222_4(12, 12, 14), \\
&222_4(13, 13, 42) \triangleright 222_4(32, 42, 43) \triangleright 222_4(12, 12, 14), \\
&222_4(12, 24, 34) \triangleright 222_4(32, 23, 42) \triangleright 222_4(12, 12, 12), \\
&222_4(12, 43, 23) \triangleright 222_4(32, 23, 42) \triangleright 222_4(12, 12, 12), \\
&222_4(13, 14, 43) \triangleright 222_4(23, 13, 31) \triangleright 222_4(13, 12, 12), \\
&222_4(12, 43, 34) \triangleright 222_4(13, 32, 14) \triangleright 222_4(12, 13, 12), \\
&222_4(12, 14, 42) \triangleright 222_4(32, 13, 13) \triangleright 222_4(13, 12, 14), \\
&222_4(13, 14, 42) \triangleright 222_4(13, 32, 13) \triangleright 222_4(12, 13, 14), \\
&222_4(13, 32, 13) \triangleright 222_4(23, 34, 24) \triangleright 222_4(12, 13, 12), \\
&222_4(13, 34, 31) \triangleright 222_4(32, 13, 14) \triangleright 222_4(13, 12, 12), \\
&222_4(12, 32, 13) \triangleright 222_4(13, 34, 41) \triangleright 222_4(12, 14, 12), \\
&222_4(12, 23, 31) \triangleright 222_4(23, 14, 41) \triangleright 222_4(13, 13, 12), \\
&222_4(12, 34, 43) \triangleright 222_4(13, 13, 42) \triangleright 332_6(123, 213, 41), \\
&222_4(13, 13, 42) \triangleright 222_4(13, 23, 41) \triangleright 233_6(12, 143, 143), \\
&222_4(12, 14, 32) \triangleright 222_4(13, 13, 43) \triangleright 332_6(123, 213, 41), \\
&222_4(13, 14, 42) \triangleright 222_4(12, 23, 31) \triangleright 332_6(132, 134, 14), \\
&222_4(12, 14, 42) \triangleright 233_6(13, 123, 341) \triangleright 222_4(12, 13, 12), \\
&222_4(12, 32, 13) \triangleright 233_6(13, 134, 421) \triangleright 222_4(12, 12, 14).
\end{aligned}$$

Here, the symbol  $pqr_d(\cdots)$  means a move with degree  $d$  of the minimal Markov basis obtained by Aoki and Takemura (2003), which is essentially a  $p \times q \times r$  table.

**Theorem 6.** *Suppose that  $\Omega_t$  is obtained from the previous frame  $\Omega_{t-1}$  by adding 1 at the  $(1, 1, 1)$ -cell. Let  $\varphi$  be a map from  $\Omega_{t-1}$  to  $\Omega_t$  by simply adding 1 at the  $(1, 1, 1)$ -cell. Then*

$$\Omega_t = \{\varphi(H) \mid H \in \Omega_{t-1}\} \cup \{\varphi(H) - M \in \Omega_t \mid H \in \Omega_{t-1}, M \in \mathfrak{M}(1, 1, 1)\}.$$

For  $I \times J \times K$  tables  $H$  and  $H'$  we denote by  $H \geq H'$  if  $H_{ijk} \geq H'_{ijk}$  for each  $i, j, k$ . For an operation  $F = M^{(1)} \triangleright M^{(2)} \triangleright \cdots \triangleright M^{(s)}$ , we let  $\psi(F)$  be a table whose  $(i, j, k)$  cell has

$$\max_{u=1, \dots, s} \left( - \sum_{a=1}^u M_{ijk}^{(a)}, 0 \right).$$

The image  $\psi(\mathfrak{M}(1, 1, 1))$  by  $\psi$  gives a necessary condition for a table  $H$  with  $H_{111} = 0$  to reach to a table  $H'$  with  $H'_{111} > 0$  as follows.

**Theorem 7.** *A contingency table  $H$  with  $H_{111} = 0$  is reachable to a table  $H'$  with  $H'_{111} > 0$  by Markov moves if and only if  $H \geq G$  for some  $G$  in  $\psi(\mathfrak{M}(1, 1, 1))$ .*

## 5 Computational algorithm

In this section, we show the algorithm to obtain the theorems in sections 3 and 4. We use the following theorem.

**Theorem 8 (Sakata and Sumi (2008)).** *Let  $H$  be a contingency table. An operation  $M^{(1)} \triangleright \dots \triangleright M^{(\ell)}$  is applicable for  $H$  if and only if*

$$H \geq \psi(M^{(1)} \triangleright \dots \triangleright M^{(\ell)}).$$

The minimal Markov basis  $\mathfrak{B}$  for  $3 \times 3 \times 4$  contingency tables consists of 6750 elements.

$\mathfrak{B}_d^p$  means the set of tables  $H$  of degree  $d$  with  $H_{111}$  being  $p$  (zero or positive). So, it has many times to compute straightforwardly the set  $\mathfrak{H}$  consisting of  $\psi(M^{(1)} \triangleright M^{(2)} \triangleright M^{(3)})$  which is for searching a 3-neighborhood, where  $M^{(1)}, M^{(2)} \in \mathfrak{B}^0 \cup \{O\}$  and  $M^{(3)} \in \mathfrak{B}^1$ .

Let  $H$  be a contingency table in  $\Phi(3, 4, 4)$ . Take an operation  $F = M^{(1)} \triangleright \dots \triangleright M^{(\ell)}$  which are applicable for  $H$ . Then by the above theorem, it holds that  $H \geq \psi(F)$ .

Suppose that there is  $i_0$  such that  $\psi(F)_{i_0jk} = 0$  for arbitrary  $j$  and  $k$ .  $H$  is movable by preserving entries at  $(i_0, j, k)$  for arbitrary  $j$  and  $k$ . Since the set of  $2 \times 4 \times 4$  contingency tables has 1-neighborhood property (cf. Sakata and Sumi (2008)),  $H$  is moved to a table  $H'$  with  $H'_{i_1j_1k_1} = 1$  by some move. Next suppose that there is  $j_0$  such that  $\psi(F)_{ij_0k} = 0$  for arbitrary  $i$  and  $k$ . Recall that the set of  $3 \times 3 \times 4$  contingency tables has 3-neighborhood property by Sumi and Sakata (2009b). Since  $H$  is movable by preserving entries at  $(i, j_0, k)$  for arbitrary  $i$  and  $k$ , the table  $H$  is moved to a table  $H'$  with  $H'_{111} = 1$  by at most 3 moves. Similarly if there is  $k_0$  such that  $\psi(F)_{ijk_0} = 0$  for arbitrary  $i$  and  $j$ , the table  $H$  is moved to a table  $H'$  with  $H'_{111} = 1$  by at most 3 moves. Therefore we may assume that there are no  $i_0, j_0, k_0$  as above.

**Lemma 6.** *a. Let  $H$  be a table accessible to a table  $H'$  with  $H'_{111} > 0$ . If  $G \geq H$  then  $G$  is also accessible to a table  $H''$  with  $H''_{111} > 0$ .  
b. Let  $H$  be a table not accessible to any table  $H'$  with  $H'_{111} > 0$ . If  $G \leq H$  then  $G$  is also not accessible to any table  $H''$  with  $H''_{111} > 0$ .*

By this lemma we only need the set of minimal tables for movability. Thus we use the following algorithm. In the processing, we also need the set of non-movable tables. Let  $d$  be the maximal integer among the entries of Markov moves in  $\mathfrak{B}$ . For a nonnegative integer  $v$  and a  $3 \times 4 \times 4$  table  $H$ , we define  $\varphi_v(H)$  as

$$\varphi_v(H)_{ijk} = \begin{cases} v + d, & \text{if } H_{ijk} \geq v \\ H_{ijk}, & \text{otherwise} \end{cases}$$

and a  $3 \times 3 \times 4$  table  $\phi_{3,4}^J(H)$  by

$$\phi_{3,4}^J(H)_{ijk} = \begin{cases} H_{i3k} + H_{i4k}, & \text{if } j = 3 \\ H_{ijk}, & \text{if } j = 1, 2 \end{cases}.$$

### Algorithm

- a. Initial setting:
  - Let  $\mathfrak{M} = \psi(\mathfrak{B}^+)$ , and let  $v$  be an integer so that  $v - 1$  is the maximal integer among the entries of tables in  $\mathfrak{M}$ .
  - Let  $\mathfrak{N}$  be the set of tables of type (2a)–(3e) where a sufficient large integer is set as  $d + v$ .
  - Put  $r = 3$ .
  - Let  $\mathfrak{C}$  be the set of  $3 \times 4 \times 4$  tables  $H$  such that  $H_{111} = 0$ ,  $0 \leq H_{ijk} \leq v$  for each  $i, j, k$ , and  $\phi_{3,4}^J(H) \in \Phi(3, 3, 4)$ .
- b. If  $\mathfrak{C}$  is empty stop the process. Otherwise, take  $H \in \mathfrak{C}$ , remove it from  $\mathfrak{C}$ , and classify it into the following three scenarios:
  - a. If there is  $G \in \mathfrak{M}$  such that  $H \geq G$  or there is  $G \in \mathfrak{N}$  such that  $H$  is subordinate to  $G$ , then go to the step 2.
  - b. If  $H$  is reachable by at most  $r$  moves  $F$  to the set  $\Omega^1$  of tables  $G$  such that  $G_{111} = 1$  and  $G$  has the marginals as same as those of  $H$ , add the moves  $\phi(F)$  into  $\mathfrak{M}$  and slimize  $\mathfrak{M}$ : remove a table  $G$  from  $\mathfrak{M}$  if  $G \geq \phi(F)$ . If  $v$  is changed, we reset  $v$ ,  $\mathfrak{C}$  and go to the step 2.
  - c. Make a maximal table  $\varphi_v(G)$  such that  $G \geq H$  and  $G$  is not within  $r$ -neighborhood of  $\Omega^1$ , add it into  $\mathfrak{N}$ , and slimize  $\mathfrak{N}$ , that is, remove a table  $G'$  from  $\mathfrak{N}$ , if  $G' \leq \varphi_v(G)$ . Let

$$\mathfrak{N}_1 = \bigcup_{H \in \mathfrak{N}, M \in \mathfrak{B}} \{\varphi_v(H) \mid (H + M) \geq O\} \cup \mathfrak{N}.$$

If there is  $G' \in \mathfrak{N}_1$  such that  $G'_{111} > 0$  then replace  $r$  by  $r + 1$  and go to the scenario b. If  $\mathfrak{N}_1 \neq \mathfrak{N}$ , we replace  $\mathfrak{N}$  by  $\mathfrak{N}_1$  and go to the scenario c again and otherwise go to the step 2.

### References

- AGRESTI, A. (2007): An introduction to categorical data analysis. Wiley Series in Probability and Statistics, 2nd edition, *Wiley-Interscience [John Wiley & Sons]*
- AOKI, S. (2004): Exact methods and Markov chain Monte Carlo methods of conditional inference for contingency tables. *Doctor Thesis, Tokyo University*.
- AOKI, S. and TAKEMURA, A. (2003): Minimal basis for connected Markov chain over  $3 \times 3 \times K$  contingency tables with fixed two-dimensional marginals. *Australian and New Zealand Journal of Statistics* 45, 229–249.
- SAKATA, T. and SUMI, T. (2008): Lifting between the sets of three-way contingency tables and  $r$ -neighborhood property. *Electronic Proceedings of COMP-STAT '2008, Contributed Papers, Categorical Data Analysis*, 87–94.
- STURMFELS, B. (1996): Gröbner bases and convex polytopes. *American Mathematical Society, University Lecture Series* 8.
- SUMI, T. and SAKATA, T. (2009a): A proof of 2-neighborhood theorem for  $3 \times 3 \times 3$  tables. *preprint*.
- SUMI, T. and SAKATA, T. (2009b): The set of  $3 \times 3 \times K$  contingency tables for  $K \geq 4$  has 3-neighborhood property. *preprint*.

# Visualization Techniques for the Integration of Rank Data

Michael G. Schimek<sup>1</sup> and Eva Budinská<sup>2</sup>

<sup>1</sup> Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation

Auenbruggerplatz 2/V, 8036 Graz, Austria, *michael.schimek@medunigraz.at*

<sup>2</sup> Swiss Institute of Bioinformatics, Bioinformatics Core Facility  
Quartier Sorge - Batiment Genopode, 1015 Lausanne, Switzerland,  
*eva.budinska@isb-sib.ch*

**Abstract.** In consumer preference studies, in Web-based meta-search or in meta-analysis of microarray experiments, we are confronted with ranked lists representing the same set of distinct objects. All these applications have in common that one is interested in the top-ranked objects with considerable overlap in their rankings across the lists. This requires the estimation of the truncation point beyond which the ordering of the objects is dominated by noise. The point of degeneration into noise can be obtained from an inference procedure due to Hall and Schimek (2008) which even works for large or huge data sets. Before its execution, it is essential to specify the distance parameter  $\delta$ . In this paper, graphical approaches for the  $\delta$ -choice, as well as for the integration of the top-ranked objects, are introduced for the first time. Finally, the new graphical tools are applied to the integration of microarray data from several experiments.

**Keywords:** data integration, microarray data, ranked list, statistical graphics, top- $k$  list

## 1 Introduction

The statistical integration (aggregation) of several long lists of common distinct objects in rank order is of increasing relevance in various fields of application. Even now, there are computational limits because combinatorial approaches would be NP-hard (Fagin et al., 2003). However, there is a tradition in taking alternative approaches such as multistage ranking models (Fligner and Verducci, 1988). Historically the primary interest has been on the consolidation of preferences of consumers for certain products, companies or institutions. The analyzed data used to be of modest size. Most recently, due to a dramatic increase of computer power, new application fields emerged such as search engine technology for the Web (Mamoulis et al., 2007) and high throughput laboratory devices in biotechnology (DeConde et al., 2006). In both areas, one has to cope with enormous amounts of data in rank order, but there is one important difference between the two: For Web search

a “black box” approach is adopted because the user does not wish to see intermediate results, only the final integrated list of objects (links to Web pages) is relevant, whereas, in the meta-analysis of microarray experiments, the scientist needs to understand the peculiarities of the aggregated result. Such a result has to be verified biologically, otherwise it is useless for further genomic experiments.

Here, in this paper, we tackle the problem of meta-analysis in genomic research. We combine an inference procedure that yields truncated, so-called top- $k$  lists, consisting of objects of high conformity, and graphical techniques that visualize the relationships between the selected objects and the remainder list items. As a direct consequence, the implications and the plausibility of obtained aggregation result can be checked, which is highly relevant for its biological interpretation. Moreover, we introduce an exploratory tool for the selection of the distance parameter  $\delta$  required for the inference procedure of Hall and Schimek (2008), and most useful for list truncation. All procedures were implemented in R. Finally, our integration approach is illustrated on well-known microarray data combining several experiments.

## 2 Inference for top- $k$ lists

Let us assume a discrete space  $O$  that contains all  $N$  objects and that the rank assignment in each list is independent of the assignment in the other lists. Let us have  $\ell$  such lists for which each object  $o$  is associated with a unique label such that  $O$  can be viewed as a space  $O = \{1, 2, \dots, N\}$ . Let us denote the rank of element  $o$  in  $O$  by  $R(o)$  under a particular assignment. We refer to  $\tau(O)$  as a full ranked list, and to  $R_\tau(o)$  as the rank of object  $o$  under the assignment mechanism  $\tau$  to distinguish it from the rank of  $o$  under another assignment mechanism.

In the above mentioned applications, we aim at partial lists (sub-spaces)  $O'_l \subset O$  of lengths  $k_l$  ( $l = 1, 2, \dots, \ell$ ). Note that the lengths of the lists do not necessarily need to be the same. Without loss of generality, we assume that the partial ranked list  $O' = \{o'_1, o'_2, \dots, o'_k\}$  is ordered according to their ranks such that  $R(o'_i) \leq R(o'_j)$  for  $i < j$ . It is implicitly assumed that all the elements that are in  $O$  but not in  $O'$  are ranked equal to or lower than  $k$ . When the number  $N$  of objects is large or even huge, it is unlikely that consensus prevails. Typically, we can observe a general decrease, not necessarily monotone, in the probability for consensus rankings with increasing distance from the top rank position. A top-ranked object is not necessarily a member of all partial lists. Hence, we have to cope with incomplete rankings.

Hall and Schimek (2008) have developed a computationally efficient moderate deviation-based inference procedure for random degeneration in paired rank lists under the above assumptions, yielding an estimate of  $j_0$  (point of degeneration;  $j_0 = k + 1$ , where  $k$  denotes the length of the top list). Their approach allows for various types of rank irregularities and list lengths in the

magnitude of thousands of objects. The necessary reduction of computational costs can be achieved only when the rankings are considered in a pair wise manner, and when a simple distance (shift) parameter  $\delta$ , as defined in the next section, is applied instead of a more demanding permutation metric (for an overview see Fagin et al., 2003). For each combination of two full lists, an estimate  $\hat{k}$  is required.

Let us define a sequence of indicators, where  $I_j = 1$  if the ranking, given by the second assessor to the object ranked  $j$  by the first assessor, is not more than  $\delta$  index positions distant from  $j$ , and otherwise  $I_j = 0$ . Further, let us assume (i) independent Bernoulli random variables  $I_1, \dots, I_N$ , with  $p_j \geq \frac{1}{2}$  for each  $j \leq j_0 - 2$ ,  $p_{j_0-1} > \frac{1}{2}$ , and  $p_j = \frac{1}{2}$  for  $j \geq j_0$ ; (ii) a “general decrease” of  $p_j$  for increasing  $j$  that does not need to be monotone. The index  $j_0$  is the point of degeneration into noise and needs to be estimated ( $\hat{j}_0 - 1 = \hat{k}$ ).

Inference is based on the moderate-deviation bound  $z_\nu \equiv (C \nu^{-1} \log \nu)^{1/2}$ , where  $\nu$  is the pilot sample size and  $C > \frac{1}{4}$  a constant (for its derivation see Hall and Schimek, 2010). For testing, the null hypothesis  $H_0$  is  $p_k = \frac{1}{2}$  for  $\nu$  consecutive index values of  $k$ . The alternative  $H_1$  is  $p_k > \frac{1}{2}$  for at least one of the values of  $k$ . Under the assumption that  $H_0$  applies to the  $\nu$  consecutive values of  $k$  in the respective series of  $\hat{p}_j^\pm$  calculated from the  $\nu$  data pairs  $I_\ell$  for which  $\ell$  lies immediately to the right of  $j$ , or immediately to the left of  $j$ , we reject  $H_0$  if and only if  $\hat{p}_j^\pm - \frac{1}{2} > z_\nu$ . Under  $H_0$ , the variance of  $\hat{p}_j^\pm$  equals  $(4\nu)^{-1}$  (this implies  $C > \frac{1}{4}$ ).

The complex decision problem is solved via an iterative algorithm, adjustable for irregularity in the rankings. It is executed for all possible pairs  $L = (\ell^2 - \ell)/2$  of lists  $\tau_\ell$ , thus we obtain  $L$  values  $\hat{k}_j$  ( $j = 1, 2, \dots, L$ ). The overall top- $k$  list length for the final graphical data integration is defined by  $\hat{k}^* = \max_j(\hat{k}_j)$ .

Before running the iterative algorithm, the distance parameter  $\delta$  needs to be specified for each list pair. A graphical procedure facilitating the  $\delta$  choice will be introduced in the following section.

### 3 The $\Delta$ -matrix

The input for the moderate deviation-based inference procedure is a sequence of  $I$ 's either taking zero or one (i.e. an indicator variable), forming a data stream representing the concordance of the paired ranks of an object  $o$ . The data stream depends on the distance parameter  $\delta$ .  $\delta$  is defined by the shift in index positions of a particular object  $o$  in one list, say  $\tau_i$ , with respect to the other list, say  $\tau_j$ . This means that we assume concordance (i.e.  $I = 1$ ) for an arbitrary object characterized by rank positions in  $\tau_i$  versus  $\tau_j$ , maximal  $\delta$  index values apart.

For the identification of an appropriate  $\delta$  in real data analysis, we suggest the following strategy: Compute all data streams for  $\delta \in [0, 1, 2, \dots, N - 1]$ . Order the data stream vectors column-wise according to increasing  $\delta$  values.

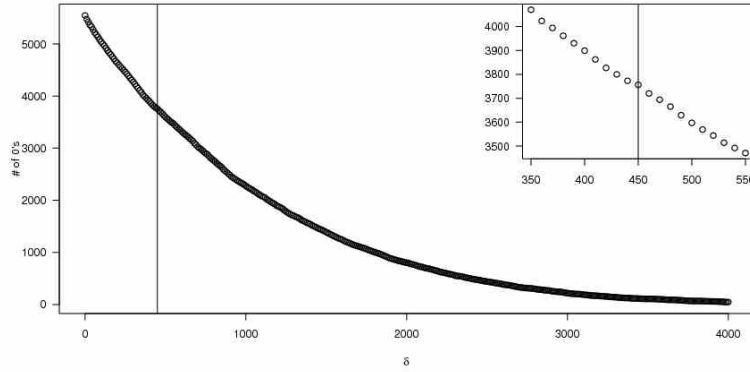
**Table 1.** Two rankings of  $N = 15$  objects, and the data streams and sums of zeros for increasing  $\delta$  values

Object	$\tau_1$	$\tau_2$	$\delta = 0$	$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$	$\delta = 5$	$\delta = 6$	$\delta = 7$	$\delta = 8$	$\dots$	$\delta = 14$
$o_1$	1	1	1	1	1	1	1	1	1	1	1	$\dots$	1
$o_2$	2	8	0	0	0	0	0	0	1	1	1	$\dots$	1
$o_3$	3	5	0	0	1	1	1	1	1	1	1	$\dots$	1
$o_4$	4	3	0	1	1	1	1	1	1	1	1	$\dots$	1
$o_5$	5	2	0	0	0	1	1	1	1	1	1	$\dots$	1
$o_6$	6	4	0	0	1	1	1	1	1	1	1	$\dots$	1
$o_7$	7	6	0	1	1	1	1	1	1	1	1	$\dots$	1
$o_8$	8	7	0	1	1	1	1	1	1	1	1	$\dots$	1
$o_9$	9	13	0	0	0	0	1	1	1	1	1	$\dots$	1
$o_{10}$	10	11	0	1	1	1	1	1	1	1	1	$\dots$	1
$o_{11}$	11	9	0	0	1	1	1	1	1	1	1	$\dots$	1
$o_{12}$	12	12	1	1	1	1	1	1	1	1	1	$\dots$	1
$o_{13}$	13	14	0	1	1	1	1	1	1	1	1	$\dots$	1
$o_{14}$	14	10	0	0	0	0	1	1	1	1	1	$\dots$	1
$o_{15}$	15	15	1	1	1	1	1	1	1	1	1	$\dots$	1
$\#(\text{zeros})$			12	7	4	3	1	1	0	0	0	$\dots$	0

In this way, we obtain a  $N \times N$  matrix  $\Delta$ . The ordered sequence of column sums (i.e. the  $\#(\text{zeros})$  for  $\delta \in [0, 1, 2, \dots, N - 1]$ ) is the information we take advantage of in the so-called  $\Delta$  plot. It represents the reduction of discordance as a function of  $\delta$ . When all column sums remain zero, complete concordance is attained. A reasonable choice of the distance parameter is associated with a distinct decline of the  $\#(\text{zeros})$ . Of course, prior information about the ranking mechanisms  $\tau$  involved and the nature of the data is also relevant for the selection of  $\delta$ . In Table 1, we display a toy example consisting of  $N = 15$  objects in two assignments (rankings)  $\tau_1$  and  $\tau_2$  (no missing values). According to the table, an adequate distance parameter choice would be  $\delta = 7$ . For the purpose of illustration, we refer to Figure 1. Based on two selected rankings of the Sorlie et al. (2003) data set (for details see later), we show a real  $\Delta$  plot. There is a slight bump at  $\delta = 450$  (see also the top right subplot). This value proved to be the best overall choice for all combinations of  $\ell = 3$  lists in our data integration example described in Section 5.

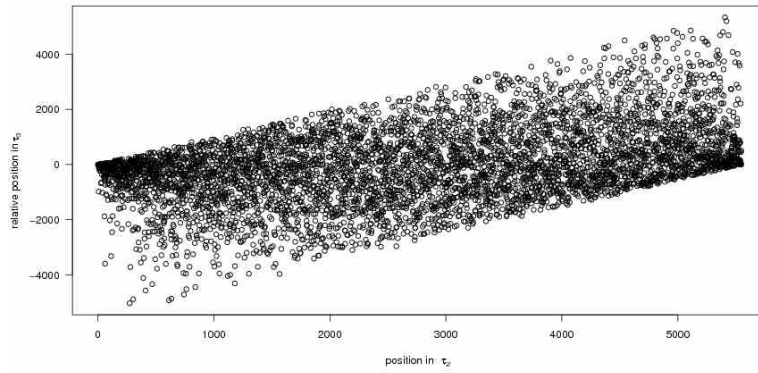
An additional type of graphic we suggest, is the simultaneous plotting of the empirical distances between the individual rank positions of the complete set of objects. Such a plot allows us to check the assumption of decreasing concordance of rankings with increasing distance from the top rank position. In Figure 2, we can see the result for the same two rankings. The assumption of a general decrease of the probability for consensus rankings should be reflected in a dense point cloud on the left concentrated around the zero distance (i.e. the respective rank positions of specific objects are very close to each other), fading out towards the right. In our case, characteristic for gene expression data, there is another, usually less distinct, concentration





**Fig. 1.**  $\Delta$  plot of the second and the third list of gene expression measurements of the Sorlie et al. (2003) study

of points at the highest rank positions (due to a subgroup of genes with extremely low intensities in all analyzed experiments).



**Fig. 2.** Distances of rank positions per object for the second and the third list of gene expression measurements of the Sorlie et al. (2003) study

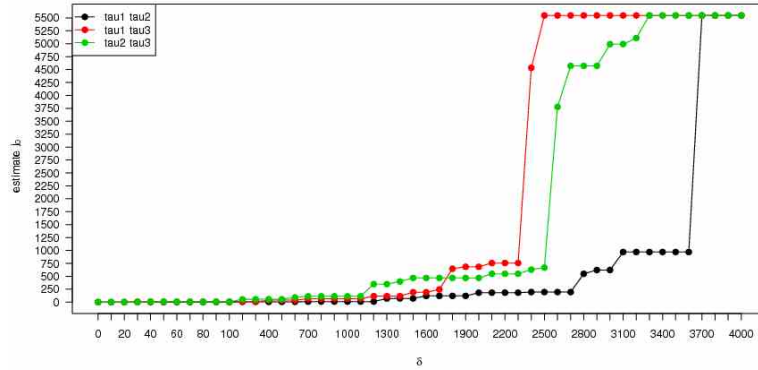
## 4 Graphical integration of partial lists

Our goal is to identify a subset of objects  $o_i$  that is characterized by high rank conformity across the lists. From the Hall and Schimek algorithm, we obtain  $(\ell^2 - \ell)/2$  values  $\hat{k}_j$  for a pre-specified distance parameter  $\delta$ . For the integration of all  $\ell$  lists  $\tau_j$  ( $j = 1, 2, \dots, \ell$ ) truncated at the individual  $\hat{k}_j$ 's, we introduce a heatmap-like graphical aid called “aggregation map” (procedure **aggmap**). A consolidated result based on irregular rankings can never

be unique, hence we need such a tool. Let us have an index  $p = 1, 2, \dots$ . We combine  $\ell - 1$  aggregation levels (groupings of partial lists) in one display: For each group of  $\ell - p$  truncated lists down to the smallest group consisting of just one pair of lists, we (a) select an arbitrary reference list  $\tau^0$  under the condition that it comprises  $\max_j(\hat{k}_j)$  objects among all pairwise comparisons in the group of rankings, (b) print the names of its  $\max_j(\hat{k}_j)$  objects vertically from the highest to the lowest rank position, and (c) add the aggregation information for all remaining  $\ell - p$  rankings (pairwise list combinations) in the group, ordered according to descending list length.

The aggregation information per group and object consists of two measures represented by colored triangles outlined in an array, (i) **membership** in the top- $k$  list, *yes* is denoted by color 'grey' and *no* by color 'white', (ii) **distance** of the rank of an object  $o_i \in \tau^0$  from its position in the other list, visualized by means of a color scale from 'red' *identical* to 'yellow' *far distant*. In addition, an integer value gives the numerical distance between the object's rank positions, a negative sign means ranked lower, and a positive sign means ranked higher in  $\tau$  with respect to  $\tau^0$ .

The procedure **aggmap** was developed by us in R utilizing the **grid** add-on package of Murrell (2006) and will be part of our **TopkLists** package for statistical data integration.

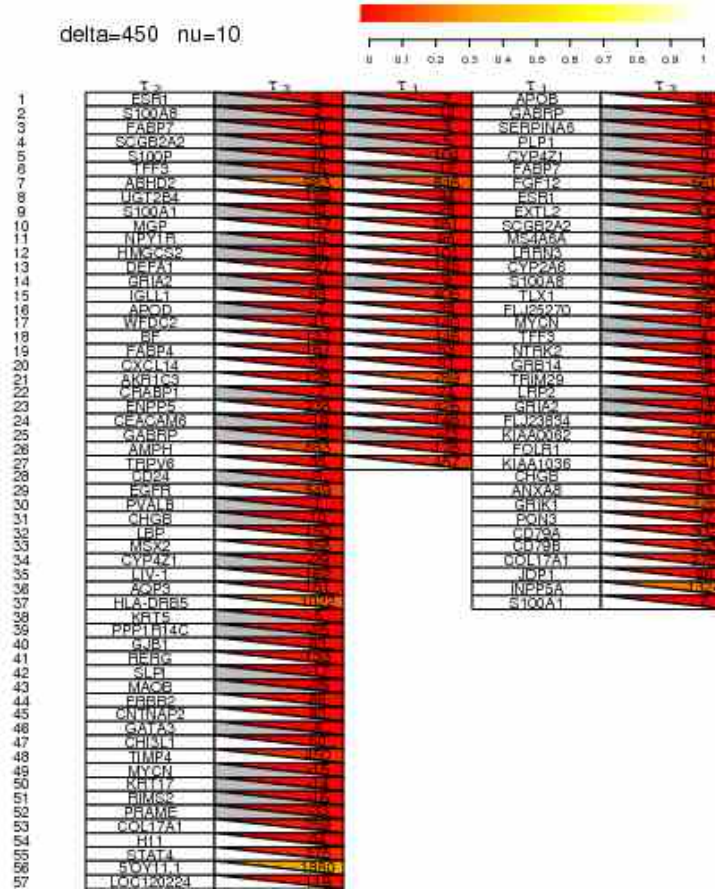


**Fig. 3.** Estimates of  $j_0$  for a range of  $\delta$  values, combining pairwise the lists  $\tau_1$ ,  $\tau_2$ , and  $\tau_3$  of the Sørbye et al. (2003) study

## 5 Meta-analysis of microarray data

We illustrate the graphical integration via **aggmap** on genomic data from a famous paper by Sørbye et al. (2003). The aim of their research was the identification of breast tumor subtypes from gene expression measured by

microarrays. In their study, data from other laboratories were also included. Because they could not estimate  $\hat{j}_0$ , which we can, they defined a top list of 534 genes heuristically.



**Fig. 4.** Aggregation map for three top- $k$  lists of gene expression measurements of the Sørli et al. (2003) study

Here, we considered selected expression data from three patient cohorts called *Norway*, *Norway FU*, and *Stanford*, hybridized on different platforms. Only genes (unique gene symbols) common to all platforms were analyzed. In each dataset, the genes were ordered decreasingly according to their standard deviation. Thus, we obtained three ranked lists,  $\tau_1$ ,  $\tau_2$ , and  $\tau_3$ , each of length  $N = 5812$ . Our task was to identify a subset of genes supported by all three cohorts that can be used for further unsupervised analysis of subtypes of breast cancer. The inference procedure allowed us to truncate the full lists.

Based on Figure 1, we selected  $\delta = 450$ . A summary of estimated  $\hat{j}_0$ 's (pilot sample size  $\nu = 10$ ) for all combinations of lists is given in Figure 3. The plot makes clear that  $\delta < 200$  is too small and  $\delta > 1100$  is too large, hence, the beforehand chosen value of 450 is a good compromise. Shape similarity of the curves indicates concordance among the lists.

Finally, we present the result of the **aggmap** procedure in Figure 4. The numbers on the left of the map are the rank positions of the selected expressed genes (objects) in the reference lists  $\tau^0$  (here from 1 to 57; unique gene symbols in the white boxes). It can easily be seen that there is reasonable conformity among the top-ranked genes (solid block of grey lower triangles up to rank 6). Moreover, the upper triangles mostly in dark colors indicate closeness (for distance details see the numbers) of the ranks of the individual genes identified as top-ranking. Thanks to the data-driven selection of partial lists, a robust set of top-ranked genes could be identified that conforms with the findings of Sørli et al. (2003). Moreover, our approach was able to find an additional gene, TIMP4, now proven to play a putative role in the initiation of breast cancer invasion. We believe that the inference procedure of Hall and Schimek (2008), together with the **aggmap** graphical representation, form a more reliable tool for selection of objects with additional reference to their positioning in the full lists, than any other approach.

## References

- DECONDE, R.P., HAWLEY, S., FALCON, S., CLEGG, N., KNUDSEN, B. and ETZIONI, R. (2006): Combined results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology* 5, 1, article 15.
- FAGIN, R., KUMAR, R. and SIVAKUMAR, D. (2003): Comparing top- $k$  lists. *SIAM J. Discrete Math.* 17, 134-160.
- FLIGNER, M.A. and VERDUCCI, J.S. (1988): Multistage ranking models. *J. Amer. Statist. Assoc.*, 83, 892-901.
- HALL, P. and SCHIMEK, M.G. (2008): Inference for the top- $k$  rank list problem. In: Brito, M.P. (Ed.): *COMPSTAT 2008. Proceedings in Computational Statistics*. Physica, Heidelberg, 433-444.
- HALL, P. and SCHIMEK, M.G. (2010): Moderate deviation-based inference for random degeneration in paired rank lists. *Submitted paper*.
- MAMOULIS, N., YIU, M.L., CHENG, K.H. and CHEUNG, D.W. (2007): Efficient top- $k$  aggregation of ranked inputs. *ACM Transactions on Database Systems*, 32, 3, article 19.
- MURRELL, P. (2006): *R Graphics*. Chapman & Hall/CRC, Boca Raton.
- SØRLIE, T., TIBSHIRANI, R., PARKER, J., HASTIE, T., MARRON, J.S., NOBEL, A., DENG, S., JOHNSEN, H., PESICH, R., GEISLER, S., DEMETER, J., PEROU, C.M., LØNNING, P.E., BROWN, P.O., BØRRESEN-DALE, A.L. and BOTSTEIN, D. (2003): Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA*, 100, 8418-8423.

# Comprehensive Assessment on Hierarchical Structures of DNA markers Using Echelon Analysis

Makoto Tomita<sup>1</sup> and Koji Kurihara<sup>2</sup>

<sup>1</sup> Clinical Research Center, Tokyo Medical and Dental University Hospital  
Faculty of Medicine, 1-5-45 Yushima, Bunkyo-ku, Tokyo, 113-8519, Japan.  
[tomita.crc@tmd.ac.jp](mailto:tomita.crc@tmd.ac.jp)

<sup>2</sup> Faculty of Environmental Science and Technology, Okayama University,  
700-8530, Japan

**Abstract.** A domain where recombination does not occur often, yet maintained linkage disequilibrium exists on DNA sequence is known as a “haplotype block” or “LD block”. Many methods are available to identify LD blocks using disequilibrium parameters, such as the well-known Gabriel’s method on Haploview, and so on. After identifying LD blocks, we can also select tagging SNPs for these LD blocks such as Tagger on Haploview, etc. We considered that Echelon analysis can be applied to identify LD block and to select tagging SNPs, and report herein that the comprehensive method can be applied according to our new method using Echelon analysis. The results of numerical examples are also provided.

**Keywords:** spatial data analysis, DNA data, haplotype, tagging SNP

## 1 Introduction

A major goal of current human genome-wide studies is to identify the genetic basis of complex disorders. (Reich *et al.*, 2001) Variation in the human genomic sequence plays a powerful but poorly understood role in the etiology of common medical conditions.

Single nucleotide polymorphisms (SNPs) are the most abundant form of genetic variation. SNPs are promising tools for mapping susceptibility mutations that contribute to complex diseases. Most SNPs can be used as surrogate markers for positional cloning of genetic loci, because of the allelic association which is well-known as linkage disequilibrium (LD).

In early research, linkage disequilibrium analysis for SNP data is particularly important. Methods of trait mapping based on theories of linkage disequilibrium analysis have been developing quickly in recent years. In DNA sequences, domain “hotspots” exist at which recombinations have occurred briskly. Conversely, large domains with infrequent recombinations in which linkage disequilibrium is maintained also exist. Such domain called a “haplotype block” or “LD block”. Although the value of  $D'$  represents one of

the disequilibrium parameters important for identifying haplotype blocks. (Gabriel *et al.*, 2002)

Gabriel *et al.* (2002) defined “strong linkage disequilibrium” as a state where the upper bound of the 95% confidence interval for  $D'$  exceeds 0.98. Zhu *et al.* (2003) have evaluated haplotypes in which each relative frequency is more than 0.04. Furthermore, the ideas underlying these methods have been combined to identify haplotype blocks in another study by Kamatani *et al.* (2004).

On the other hand, there are many methods for selection of tagging SNPs. Tagger is a tool for the selection and evaluation of tag SNPs from genotype data, packaging into Haploview (Barrett *et al.*, 2005). Then Avi-Itzhak *et al.* (2003) produced “lossless mode” for minimum subsets of SNPs to capture LD blocks based on haplotype estimation. Kamatani *et al.* (2004) have approached selection of tagging SNPs improved the methods of Avi-Itzhak *et al.* .

Echelon analysis (Myers *et al.*, 1997) is a very new analysis method for spatial data that as developed to identify spatial position and to understand the structure of space data both systematically and objectively for one variable of one dimension on the map. Tomita *et al.* (2008) have approached the new method to identify LD blocks using Echelon analysis. We considered that Echelon analysis can be applied not only to identify LD blocks but also to select tagging SNPs, and introduce these comprehensive methods with numerical examples.

## 2 Echelon analysis

### 2.1 Classification of one-dimensional spatial data

An echelon dendrogram is a graph of spatial data expressed hierarchically by echelon structure. The rough flow is as follows. A cross-sectional diagram is made using spatial data for one dimension (Table 1). The echelon dendrogram (Figure 2) is completed by hierarchizing the cross-sectional diagram (Figure 1). The structure of data is clearly shown by the echelon dendrogram.

A more detailed mathematical explanation is warranted. One-dimensional spatial data has the position ( $x$ ) and the value ( $h(x)$ ) on the horizontal and vertical lines, respectively. For  $k$  divided lattice (interval) data, values are taken at the interval  $l_1(i) = (i - 1, i]$ ,  $i = 1, 2, \dots, k$ . Table 1 shows the 25 intervals named from  $A$  to  $Y$  in order and associated values. To make use of the information on spatial positions, a cross-sectional view of a topographical map is made like Figure 1. Nine numbered parts display the same topological structure in these hillforms. These parts are called echelons, comprising peaks, foundations of peaks and foundations of foundations. Numbers 1, 2, 3, 4 and 5 are the peaks of hillforms. Numbers 6 and 7 are the foundations of two peaks. Number 8 is the foundation of two foundations. Number 9 is the foundation of

a foundation and peak and is also called the root. We then have nine clusters  $G(i)$ ,  $i = 1, 2, \dots, 9$  for specified intervals based on these echelons. These are the following clusters of five peaks.

$$G(1) = \{Q, P, R\}, G(2) = \{N\}, G(3) = \{G, F, H\}, G(4) = \{D\}, G(5) = \{X\} \quad (1)$$

and clusters of foundations,

$$G(6) = \{M, O, S, L, T, K, U\}, G(7) = \{C, E, I\}, G(8) = \{B, J, V\} \quad (2)$$

and cluster of the root.

$$G(9) = \{A, W, Y\}. \quad (3)$$

The relationship among clusters is expressed as  $9 ( 8 ( 7 ( 4 3 ) 6 ( 2 1 ) ) 5 )$  using the numbers of clusters. Cluster  $G(6)$  is a parent of  $G(2)$  and  $G(1)$ , and  $G(6)$  has two children in  $G(2)$  and  $G(1)$ . We then define the posterity children  $CH(G(i))$  and family  $FM(G(i))$  for cluster  $G(i)$ .

$$CH(G(8)) = G(7) \cup G(4) \cup G(3) \cup G(6) \cup G(2) \cup G(1) \quad (4)$$

$$FM(G(8)) = G(8) \cup CH(G(8)) \quad (5)$$

A graphical representation is given in the dendrogram shown in Figure 2.

**Table 1.** One dimensional spatial interval data.

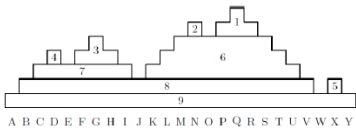
$i$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
ID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
$h(i)$	1	2	3	4	3	4	5	4	3	2	3	4	5	6	5	6	7	6	5	4	3	2	1	2	1

**Table 2.** Digital values over a 5 by 5 array.

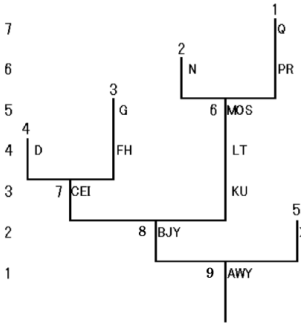
	A	B	C	D	E
1	10	24	10	15	10
2	10	10	14	22	10
3	10	13	19	23	25
4	20	21	12	11	17
5	16	10	10	18	10

**Table 3.** Clusters of 5 by 5 array data.

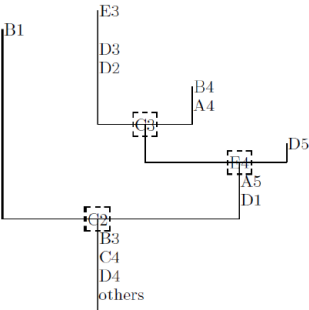
Stage	$i$	$G(i)$	$CH(G(i))$	$FM(G(i))$
1	1	E3 D3 D2	$\phi$	$G(1)$
1	2	B1	$\phi$	$G(2)$
1	3	B4 A4	$\phi$	$G(3)$
1	4	D5	$\phi$	$G(4)$
2	5	C3	$G(j) \ j = 1, 3$	$G(j) \ j = 1, 3, 5$
2	6	E4 A5 D1	$G(j) \ j = 1, 3, 4, 5$	$G(j) \ j = 1, 3, 4, 5, 6$
2	7	C2 B3 C4 D4 and others	$G(j) \ j = 1, 2, 3, 4, 5, 6$	$G(j) \ j = 1, 2, 3, 4, 5, 6, 7$



**Fig. 1.** The hypothetical set of hillforms in one dimensional spatial data.



**Fig. 2.** Echelon dendrogram



**Fig. 3.** Relation of clusters for 5 by 5 array.



## 2.2 Classification of two-dimensional spatial data

Two-dimensional spatial data has the value  $(h(x, y))$  for response variable at position  $(x, y)$ . Spatial data such as remote-sensing data are given as pixels of digital values over the  $M$  by  $N$  lattice area  $l_2(i, j) = \{(x, y) | x_i - 1 < x < x_i, y_j - 1 < y < y_j\}, i = 1, 2, \dots, N, j = 1, 2, \dots, M$ .

For the illustration, we will apply the digital values over a 5 by 5 array shown in Table 2. Posterity children  $CH(G(i))$  and families  $FM(G(i))$  for  $G(i)$  are calculated in Table 3. A graphical representation of these array data is shown as a dendrogram in Figure 3.

## 3 Methods applied by Echelon analysis

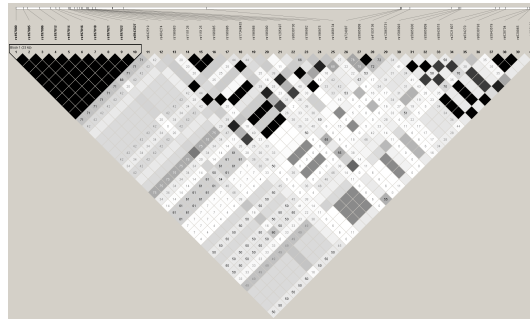
Echelon analysis is applied to the identification of the LD block. (See Tomita *et al.* (2008) in detail.) For selecting tagging SNPs with Echelon analysis, we draw upon the value of  $r^2$  to select tagging SNPs, since Tagger used the condition  $r^2 > 0.8$ . Thus, we think Echelon analysis can be applied to the selection of tagging SNPs. The procedure involves the following steps.

- Step 1'* Loci with minor allele frequencies less than 0.1 are excluded from genotype data, by same reason as identifying LD block (Tomita *et al.*, 2008).
- Step 2'* An Echelon dendrogram is generated from the  $LD(r^2)$  map of data after Step 1' is finished.
- Step 3'* We color the  $LD(r^2)$  map according to the Echelon dendrogram.
- Step 4'* We pull out the part where all combinations in the  $n \times n$  matrix of the LD map are colored, and also marginal cells with high values.
- Step 5'* We select tagging SNPs from a square matrix colored by Step 4'.

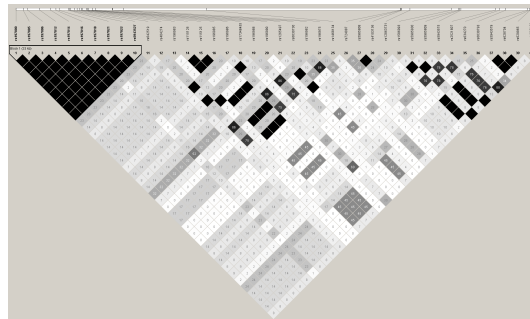
Our method of tagging SNPs is almost similar to the procedure to identify LD blocks. It is the difference between these methods that we perceive a square matrix colored on *Step 5'*. We show details in numerical examples.

## 4 Numerical examples and results

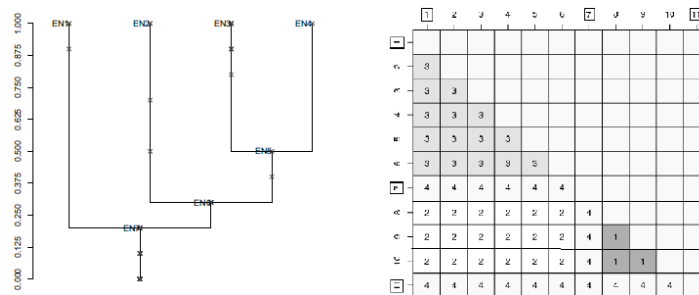
The actual data set was downloaded from the Hapmap Project. Data for region (107,189 loci) of the X chromosome were obtained for 44 subjects. Since the data set was too large, we selected data for 40 loci with linkage disequilibrium as a following research. Tomita *et al.* (2008) have researched this area of the X chromosome from Hapmap Project. The input data was unphased and we did not regenerated any missing data. Figure 4 shows the loci and  $LD(D')$  map with chromosome position for data. Various linkage disequilibriums can be seen to exist from Figure 4.  $LD(D', r^2)$  maps have been made using "Haploview" (Barrett *et al.*, 2005). They showed regions of



**Fig. 4.** LD( $D'$  values) map with chromosome position for data



**Fig. 5.** LD( $r^2$  values) map with chromosome position for data



**Fig. 6.** (Left) Echelon dendrogram ( $r^2$ ) of data; (Right) Coloring Echelon numbers ( $r^2$ ) of data on Step 2' with tag SNP numbers enclosed by square. Coloring Echelon numbers ( $r^2$ ) of data on Step 2' with tag SNP numbers enclosed by square.

the peak displaying extremely high values were colored in Figure (See Tomita *et al.* (2008) in detail.).

Figure 5 shows the loci and  $LD(r^2)$  map with chromosome position for same data. For the first LD block with loci {1-11}, we select tagging SNPs applied by Echelon analysis. Echelon dendrogram made for  $r^2$ s for only first LD block is Figure 6(left hand-side), we see the structure that the peak involves Echelon numbers {1-3}. We pull out the parts where all combinations in the square matrix are colored from Figure 6(right hand-side). We then identify patterns comprising loci {1-6} and {8-10} on Step 4' of Section 3. These patterns mean same sequences within themselves. Thus, it is good enough that we select only one locus from among each pattern for a tagging SNP. On the other hand, Table 4 is a result of tagging SNPs as above same loci by the method of Kamatani *et al.*. The result means htSNPs are loci {1, 7, 11}, however it also means that this LD block has patterns {1-6, 7, 8-10, 11}. It can be understood the almost same results between ours and theirs.

**Table 4.** Haplotypes' expressions by all SNPs and tagging SNPs.

all SNPs												tagging SNPs				
locus	1	2	3	4	5	6	7	8	9	10	11	1	-----	7	---	11
	A	A	T	G	T	A	T	A	A	C	G	A	-----	T	---	G
	G	G	C	T	C	C	C	G	G	T	A	G	-----	C	---	A
	G	G	C	T	C	C	C	G	G	T	G	G	-----	C	---	G
	G	G	C	T	C	C	T	G	G	T	G	G	-----	T	---	G
	G	G	C	T	C	C	T	G	G	T	A	G	-----	T	---	A

## 5 Discussion

Results of block identification and tagging SNPs selection were obtained by each method for actual data from the Hapmap Project.

When the results were obtained, blocks from our method, the method of Kamatani *et al.* and Haploview that were identical or differed little were identified. As application of the Echelon analysis method seems possible, this method is convincing even for data points that differ slightly.

We got results of the selection of tagging SNPs, which are Table 4 for Kamatani *et al.* and Figure 6 for our method of Echelon analysis. To select tagging SNPs is to find some patterns for allele sequences. Both methods have the same result that the first block with loci {1-11} has patterns {1-6, 7, 8-10, 11}. Thus, we think our method has also good performance for a problem of the selection of tagging SNPs.

When blocks are identified to a nearby high peak seen in the Echelon dendrogram, we have to consider as a block of the addition in our own method.

It is also clear that we got the same result for tagging SNPs selection. We can know a structure of linkage disequilibrium by using our R program (Echelon analysis), it is a large advantage. Therefore, good results seem to have been obtained even though the correction has not been considered in detail for this data. We omitted it due to the large amounts of data, although sufficient data was obtained for other results. We conclude that our method is very effective.

## Acknowledgements

This work was partly supported by KAKENHI (21700317; Grant-in-Aid for Young Scientists (B)).

## References

- Avi-Itzhak H.I., Su X., De La Vega F.M. (2003). Selection of minimum subsets of single nucleotide polymorphism to capture haplotype block diversity. *Pacific Symposium on Biocomputing*. 8, 466-477.
- Barrett J. C., Fry B., Maller J., Daly M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 21(2), 263-265.
- Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*. 296, 2225-2229.
- Kamatani, N., Sekine, A., Kitamoto, T., Iida, A., Saito, S., Kogame, A., Inoue, E., Kawamoto, M., Harigai, M. and Nakamura, Y. (2004). Large-scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP Maps, of 199 drug-related genes in 752 subjects: the analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs. *American Journal of Human Genetics*. 75(2), 190-203.
- Myers, W. L., Patil, G. P. and Joly, K. (1997). Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics*. 4, 131-152.
- Reich, D.E. and Lander, E.S. (2001). On the allelic spectrum of human disease, *Trends Genetics*. 17, 502-510.
- Tomita, M., Hatsumichi M. and Kurihara K. (2008). Identify LD Blocks Based on Hierarchical Spatial Data. *Computational Statistics and Data Analysis*. 52, 1806-1820.
- Zhu, X., Yan, D., Cooper, R.S., Luke, A., Ikeda, M.A., Chang, Y.P., Weder, A. and Chakravarti, A. (2003). Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: findings from the family blood pressure program. *Genome Research*. 13, 173-181.

# Non-Hierarchical Clustering for Distribution-Valued Data

Yoshikazu Terada<sup>1</sup> and Hiroshi Yadohisa<sup>2</sup>

<sup>1</sup> Graduate School of Culture and Information Science, Doshisha University  
Kyoto 610-0394, Japan, *dij0030@mail4.doshisha.ac.jp*

<sup>2</sup> Department of Culture and Information Science, Doshisha University  
Kyoto 610-0394, Japan, *hyadohis@mail.doshisha.ac.jp*

**Abstract.** A lot of clustering algorithms and methods for symbolic data has been proposed. However, the clustering methods for distribution-valued data, that can consider relationships between variables in each object, have been not really proposed. In this paper, we present a non-hierarchical clustering method for the distribution-valued data involving the histogram-valued data as an empirical frequency distribution function and joint distributions. In our method, we can consider relationships (e.g. dependency) among variables in each object by using joint distributions. We define the “centroid distribution” of probability distributions. By using the centroid distributions, we propose a new non-hierarchical clustering method for symbolic objects described by probability distributions.

**Keywords:** symbolic data analysis,  $k$ -means clustering, large and complex data

## 1 Introduction

With the rise of the computer, datasets have become increasingly larger and more complex in a variety of fields. Since it is difficult to analyze them with most classical statistical methods, some new methods for analyzing them are required. One of the ways to aggregate large datasets into more manageable datasets is to deal with more complex data table called “symbolic data table” (Billard and Diday, 2006; Bock and Diday, 2000). Whereas a cell of the classical data table contains only a single quantitative or categorical value, that of the symbolic data table can contain a more complicated type of value, such as several values, intervals and distributions. The symbolic data table makes it possible to describe more complex information and redescribe very large datasets as smaller ones. Extended standard data analysis to be able to deal with the symbolic data table is called “Symbolic Data Analysis”. The symbolic data analysis has been studied as one of various useful methods for analyzing large and complex datasets. Various clustering methods for symbolic data have been proposed (e.g. Billard and Diday, 2006; Bock and Diday, 2000; Verde, 2004).

In this paper, we deal with the distribution-valued data including the histogram-data as empirical distribution functions. The distribution-valued data are made, when we aggregate individuals into classes or describe more complex information for these classes. Several clustering methods for histogram-data also have been proposed. For example, Irpino and Verde (2006) and Irpino et al. (2006) present the hierarchical clustering method and the dynamic clustering method for histogram-data using a new distance based on Wasserstein metric, respectively. Verde and Irpino (2008) presents the Mahalanobis-Wasserstein distance, which is a generalization of the classical Mahalanobis distance, for comparing histograms and apply it to the dynamic clustering. These methods assume a case that only the marginal distributions of the multivariate distribution for each object are obtained and then it is described by independent distributions. In De Souza et al. (2007) and De Carvalho and De Souza (2010), dynamic clustering methods for mixed feature-type symbolic data are proposed. These methods involve a pre-processing step for transforming mixed feature-type symbolic data into the modal symbolic data and a dynamic clustering algorithm for the modal symbolic data is obtained in the pre-processing step. In these methods, it is not considered that relationships among variables in each object (e.g. class or category). Moreover, in these methods, if we deal with histogram-data, we must create histograms based on the same interval partition.

A non-hierarchical clustering method for distribution-valued data, represented as joint (or one-dimensional) density functions or distribution functions, is described in this paper. When we deal with the dissimilarity between objects, we can consider various relationships (e.g. dependency) among variables in each object by representing the objects as joint distributions. First, we define the centroid distribution on a set of probability distributions with a dissimilarity between two distributions. Secondly, we propose a  $k$ -means clustering method with the centroid distribution and apply it to an artificial data (MacQueen, 1967). In addition, we compare the proposal method and the classical method by analyzing the same dataset.

## 2 Centroid distribution and Dissimilarities between two distributions

There are many dissimilarity measures between two probability distributions (Gibbs and Su, 2002). We consider a dissimilarity using cumulative distribution functions or probability density functions. First, we define the centroid distribution of probability distributions on a set of probability distributions with a dissimilarity measure.

### **Definition 13.** Centroid distribution

Let  $\mathcal{P}$  be a set of probability distributions and  $d$  be a dissimilarity measure on  $\mathcal{P}$ . The set  $\mathcal{P}$  with the dissimilarity  $d$  will be denoted by  $(\mathcal{P}, d)$ .  $P_i$  ( $i = 1, 2, \dots, n$ ) are probability distributions in  $\mathcal{P}$ . We assume  $\mathcal{Q} = \{Q \in \mathcal{P} \mid$

$d(P_i, Q) < \infty$  ( $i = 1, 2, \dots, n$ )  $\neq \emptyset$  and define the centroid distribution  $P_C$  of probability distributions  $P_i$ , satisfying

$$\sum_{i=1}^n d^2(P_i, P_C) = \inf_{Q \in \mathcal{P}} \sum_{i=1}^n d^2(P_i, Q).$$

Next, we consider the dissimilarity using cumulative distribution functions and give the centroid distribution of probability distributions on a set of probability distributions with the dissimilarity. A set of all probability distributions on  $k$  dimensional Borel measurable space will be denoted by  $\mathcal{P}_r$ . We consider the dissimilarity  $d_C$  on  $\mathcal{P}_r$  defined by

$$d_C(P, Q) = \left( \int_{-\infty}^{\infty} (F(\mathbf{x}) - G(\mathbf{x}))^2 d\mathbf{x} \right)^{\frac{1}{2}} \quad (P, Q \in \mathcal{P}_r),$$

where  $F$  and  $G$  are the cumulative distribution functions of  $P$  and  $Q$ , respectively (Bock and Diday, 2000). Then, the centroid distribution of probability distributions  $P_i \in \mathcal{P}_r$  is given by as the follows.

**Proposition 19.** *Let  $F_i$  be the cumulative distribution functions of  $P_i \in \mathcal{P}_r$ . If  $\mathcal{Q} = \{Q \in \mathcal{P} \mid d_C(P_i, Q) < \infty$  ( $i = 1, 2, \dots, n$ )  $\neq \emptyset$ , then the centroid distribution  $P_C$  of probability distributions  $P_i$  is given by the distribution that has the cumulative distribution function satisfying*

$$F_C(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n F_i(\mathbf{x}) \quad (\forall \mathbf{x} \in \mathbb{R}^r).$$

Next, we consider the dissimilarity using probability density functions. Let  $\mathcal{P}_r^*$  be a set of all absolutely continuous probability distributions defined on  $k$  dimensional Borel measurable space. We define the dissimilarity  $d_D$  on  $\mathcal{P}_r^*$  by

$$d_D(P, Q) = \left( \int_{-\infty}^{\infty} (f(\mathbf{x}) - g(\mathbf{x}))^2 d\mathbf{x} \right)^{\frac{1}{2}} \quad (P, Q \in \mathcal{P}_r^*),$$

where  $f$  and  $g$  are the probability density functions of  $P$  and  $Q$ , respectively. In a similar way, the centroid distribution  $P_C$  of probability distributions  $P_i \in \mathcal{P}_r^*$  ( $i = 1, 2, \dots, n$ ) is given as the follows.

**Proposition 20.** *Let  $P_i \in \mathcal{P}_r^*$  ( $i = 1, 2, \dots, n$ ) be the distributions that have the probability density functions  $f_i$ , respectively. We assume  $\mathcal{Q} = \{Q \in \mathcal{P} \mid d_D(P_i, Q) < \infty$  ( $i = 1, 2, \dots, n$ )  $\neq \emptyset$ . Then, the centroid distribution  $P_C$  of probability distributions  $P_i$  ( $i = 1, 2, \dots, n$ ) is given by the distribution that has the probability density function defined by*

$$f_C(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}) \quad (\forall \mathbf{x} \in \mathbb{R}^r).$$

### 3 Non-hierarchical clustering algorithm for distribution-valued data

Here, we present the  $k$ -means clustering method for distribution-valued data with the centroid distribution defined in the section 2.

If we are not interested in individuals but in classes of individuals, we aggregate the data into these classes. In the classical data analysis, we analyze the data table whose cells contain only single values (e.g. means). That is, we consider the classes as points in  $n$  dimensional real space  $\mathbb{R}^n$ . Hence, if these classes have different distribution structures and have the same single values, then the classical clustering method cannot classify these classes based on these structures. Using a new clustering method for distribution-valued data, that is, considering these classes as distributions (e.g. empirical distribution functions), we can classify these classes based on these structure. The new clustering algorithm for distribution-valued data proceeds as the follows:

- Step 1:** Initial seeds  $P_{C_j}$  ( $j = 1, 2, \dots, k$ ) are appropriately determined from the objects  $P_i$  ( $i = 1, 2, \dots, n$ ) described by distributions (e.g. by using random numbers).
- Step 2:** Dissimilarity  $d_C(P_i, P_{C_j})$  (or  $d_D(P_i, P_{C_j})$ ) from seed  $P_{C_j}$  to object  $P_i$  is evaluated for  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, k$ .
- Step 3:** The centroid distribution  $P_{C_j}$  of each cluster  $C_j$  ( $j = 1, 2, \dots, k$ ) is decided as a new seed.
- Step 4:** Each object is assigned to the nearest seed.
- Step 5:** If it satisfies a stopping rule (e.g. pre-determined maximum iteration number) then stop, else go to Step 2.

This algorithm aims at optimizing the following objective function

$$Q_C = \sum_{i=1}^k \sum_{j \in C_i} d_C^2(P_j, P_{C_i}) \quad \left( \text{or } Q_D = \sum_{i=1}^k \sum_{j \in C_i} d_D^2(P_j, P_{C_i}) \right).$$

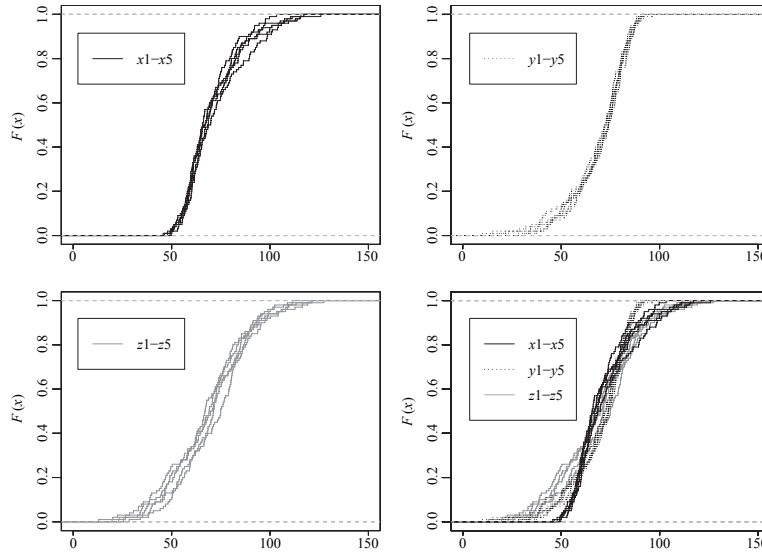
### 4 Numerical example

Here, we apply the new algorithm to an artificial data. First, we apply it to univariate data. Let  $a$ ,  $b$  and  $\varepsilon$  be distributed with the beta distributions  $\beta(1, 5)$ ,  $\beta(5, 1)$  and the normal distribution  $N(0, 25)$ , respectively. Then,  $X = 100 \times a + 55 + \varepsilon$ ,  $Y = 100 \times b - 10 + \varepsilon$  and probability distributions of  $X$ ,  $Y$  are denoted by  $P_X$  and  $P_Y$ , respectively. Random numbers  $\mathbf{x}_i = (x_{i_1}, x_{i_2}, \dots, x_{i_{100}})$ ,  $\mathbf{y}_i = (y_{i_1}, y_{i_2}, \dots, y_{i_{100}})$ ,  $\mathbf{z}_i = (z_{i_1}, z_{i_2}, \dots, z_{i_{100}})$  ( $i = 1, 2, \dots, 5$ ) are generated from  $P_X$ ,  $P_Y$  and  $N(70, 400)$ , respectively. We consider  $\mathbf{x}_i$ ,  $\mathbf{y}_i$ ,  $\mathbf{z}_i$  ( $i = 1, 2, \dots, 5$ ) as 15 classes. Means, standard deviations, and distribution-valued data of these classes are shown in Table 1 and Fig. 1, respectively. We give the cluster-



**Table 1.** Means and standard deviations of  $x_i$ ,  $y_i$ ,  $z_i$ 

Class	mean	SD	Class	mean	SD	Class	mean	SD
$x_1$	69.68	13.66	$y_1$	70.54	14.41	$z_1$	68.23	21.79
$x_2$	70.52	12.17	$y_2$	69.49	14.35	$z_2$	67.72	19.86
$x_3$	72.99	16.24	$y_3$	67.84	16.15	$z_3$	70.02	17.81
$x_4$	70.81	15.32	$y_4$	70.45	13.99	$z_4$	72.12	19.05
$x_5$	70.85	15.69	$y_5$	69.90	13.34	$z_5$	68.74	20.08

**Fig. 1.** Distribution-valued data of  $x_i$ ,  $y_i$ ,  $z_i$ .

ing result of the classical method using the mean and standard deviation of each class in Table 2, and the result of the proposal method is described in Table 3. The structures of these classes can not be considered in the classical clustering method and the classes can not properly be classified by using the method. On the other hand, the classes can properly be classified by using the proposal method.

Next, we apply it to bivariate data. Let  $C_i$  ( $i = 1, 2, 3$ ) be independent and identically distributed chi-square random variables with 1 degree of freedom and  $\varepsilon_1$  be distributed the normal distribution  $N(0, 0.25)$ . Then,

$$\mathbf{X}_1 \sim N_2 \left( \begin{bmatrix} 5 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \right), \quad \mathbf{X}_2 \sim N_2 \left( \begin{bmatrix} 5 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right),$$

**Table 2.** The result of the classical clustering for mean and standard deviation data

Class	Clustering result	Class	Clustering result	Class	Clustering result
$x_1$	Cluster1	$y_1$	Cluster1	$z_1$	Cluster3
$x_2$	Cluster1	$y_2$	Cluster1	$z_2$	Cluster3
$x_3$	Cluster2	$y_3$	Cluster1	$z_3$	Cluster2
$x_4$	Cluster1	$y_4$	Cluster1	$z_4$	Cluster2
$x_5$	Cluster1	$y_5$	Cluster1	$z_5$	Cluster3

**Table 3.** The result of the proposal clustering for distribution-valued data

Class	Clustering result	Class	Clustering result	Class	Clustering result
$x_1$	Cluster1	$y_1$	Cluster2	$z_1$	Cluster3
$x_2$	Cluster1	$y_2$	Cluster2	$z_2$	Cluster3
$x_3$	Cluster1	$y_3$	Cluster2	$z_3$	Cluster3
$x_4$	Cluster1	$y_4$	Cluster2	$z_4$	Cluster2
$x_5$	Cluster1	$y_5$	Cluster2	$z_5$	Cluster3

$$\mathbf{X}_3 = (X_{31} = C_1 - 6, X_{32})^T \quad (X_{32} \sim N(-5, 1)), \quad \mathbf{X}_4 \sim N_2 \left( \begin{bmatrix} -5 \\ -5 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right),$$

$$\mathbf{X}_5 = (X_{51}, X_{52} = X_{51}^2 + \varepsilon_1)^T \quad (X_{51} \sim N(0, 1)),$$

$$\mathbf{X}_6 = (X_{61}, X_{62} = C_3 + \varepsilon_1)^T \quad (X_{61} \sim N(0, 1)),$$

where  $X_{i1}$  and  $X_{i2}$  ( $i = 3, 6$ ) are independent. Probability distributions of  $\mathbf{X}_i$  ( $i = 1, 2, \dots, 6$ ) are denoted by  $P_{\mathbf{X}_i}$ . Random numbers  $\mathbf{x}_{ij_k}$  ( $i = 1, 2, \dots, 6; j = 1, 2, \dots, 5; k = 1, 2, \dots, n_{ij}; 30 \leq n_{ij} \leq 50$ ) are generated from  $P_{\mathbf{X}_i}$ , respectively. We consider  $C_{ij} = \{\mathbf{x}_{ij_k} \mid k = 1, 2, \dots, n_{ij}\}$  ( $i = 1, 2, \dots, 6; j = 1, 2, \dots, 5$ ) as 30 classes. Means, variances and correlation coefficients of these classes are shown in Table 4.  $C_{ij}$  and  $C_{i+1j}$  ( $i = 1, 3, 5; j = 1, 2, \dots, 5$ ) have a similar mean, respectively.

**Table 4.** Means, variances and correlation coefficients of  $C_{ij}$ 

Class	mean1	mean2	variance1	variance2	correlation
$C_{11}$	5.15	6.00	0.55	0.66	0.64
$C_{12}$	5.43	3.47	1.11	0.84	0.89
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C_{65}$	-0.03	-0.20	0.55	2.98	0.31

We give the clustering result of the classical method using mean, variance and correlation coefficient data in Table 5, and that of the proposal method for empirical joint distribution functions of  $C_{ij}$  ( $i = 1, 2, \dots, 6$ ;  $j = 1, 2, \dots, 5$ ) are described in Table 6.

**Table 5.** The result of the classical clustering for Table 4

Class	$k = 3$	$k = 6$	Class	$k = 3$	$k = 6$	Class	$k = 3$	$k = 6$
$C_{11}$	Cluster1	Cluster1	$C_{31}$	Cluster2	Cluster5	$C_{51}$	Cluster3	Cluster6
$C_{12}$	Cluster1	Cluster2	$C_{32}$	Cluster2	Cluster3	$C_{52}$	Cluster3	Cluster6
$C_{13}$	Cluster1	Cluster2	$C_{33}$	Cluster2	Cluster3	$C_{53}$	Cluster3	Cluster6
$C_{14}$	Cluster1	Cluster1	$C_{34}$	Cluster2	Cluster3	$C_{54}$	Cluster3	Cluster6
$C_{15}$	Cluster1	Cluster2	$C_{35}$	Cluster2	Cluster3	$C_{55}$	Cluster3	Cluster6
$C_{21}$	Cluster1	Cluster2	$C_{41}$	Cluster2	Cluster3	$C_{61}$	Cluster3	Cluster6
$C_{22}$	Cluster1	Cluster2	$C_{42}$	Cluster2	Cluster4	$C_{62}$	Cluster3	Cluster6
$C_{23}$	Cluster1	Cluster1	$C_{43}$	Cluster2	Cluster3	$C_{63}$	Cluster3	Cluster6
$C_{24}$	Cluster1	Cluster1	$C_{44}$	Cluster2	Cluster3	$C_{64}$	Cluster3	Cluster6
$C_{25}$	Cluster1	Cluster1	$C_{45}$	Cluster2	Cluster3	$C_{65}$	Cluster3	Cluster6

**Table 6.** The result of the proposal clustering for distribution-valued data

Class	$k = 3$	$k = 6$	Class	$k = 3$	$k = 6$	Class	$k = 3$	$k = 6$
$C_{11}$	Cluster1	Cluster1	$C_{31}$	Cluster2	Cluster3	$C_{51}$	Cluster3	Cluster5
$C_{12}$	Cluster1	Cluster1	$C_{32}$	Cluster2	Cluster3	$C_{52}$	Cluster3	Cluster5
$C_{13}$	Cluster1	Cluster1	$C_{33}$	Cluster2	Cluster3	$C_{53}$	Cluster3	Cluster5
$C_{14}$	Cluster1	Cluster1	$C_{34}$	Cluster2	Cluster3	$C_{54}$	Cluster3	Cluster5
$C_{15}$	Cluster1	Cluster1	$C_{35}$	Cluster2	Cluster3	$C_{55}$	Cluster3	Cluster5
$C_{21}$	Cluster1	Cluster2	$C_{41}$	Cluster2	Cluster4	$C_{61}$	Cluster3	Cluster6
$C_{22}$	Cluster1	Cluster2	$C_{42}$	Cluster2	Cluster4	$C_{62}$	Cluster3	Cluster6
$C_{23}$	Cluster1	Cluster2	$C_{43}$	Cluster2	Cluster4	$C_{63}$	Cluster3	Cluster6
$C_{24}$	Cluster1	Cluster2	$C_{44}$	Cluster2	Cluster3	$C_{64}$	Cluster3	Cluster6
$C_{25}$	Cluster1	Cluster1	$C_{45}$	Cluster2	Cluster4	$C_{65}$	Cluster3	Cluster6

In the case of  $k = 3$ , both method give the same result, in which the classes are clustered by the locations of these classes properly. In the case of  $k = 6$ , the classes can not properly be classified by using the classical method. On the other hand, the classes can properly be classified by using the proposal method. The differences between  $C_{ij}$  and  $C_{i+1j}$  ( $i = 1, 3, 5$ ;  $j = 1, 2, \dots, 5$ ) about the structures of these classes can be considered in the proposal method.

As a consequence, though the structure of these objects is not considered in the result of the classical  $k$ -means clustering method, it is considered in the result of the proposal method. When we are interested in classes of individuals, we need to consider the structures of these classes.

## 5 Conclusion

In this paper, we have defined a centroid distribution on a set of probability distributions with a dissimilarity between two probability distributions and represented the centroid distributions with the dissimilarity using cumulative distribution functions and probability density functions. Further, we have proposed a non-hierarchical clustering method for distribution-valued data using the centroid distribution.

For the future study, we would like to find the centroid distributions for various dissimilarities between two probability distributions and propose other clustering methods for distribution-valued data with these centroid distributions. There is some possibility that new classification structures in a variety of fields are found by using the proposal method.

## References

- BILLARD, L. and DIDAY, E. (2006): *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons, Chichester.
- BOCK, H. H. and DIDAY, E. (2000): *Analysis of Symbolic Data: Exploratory Methods for Extraction Statistical Information from Complex Data*. Springer, Berlin.
- DE CARVALHO, F. A. T. and DE SOUZA, R. M. C. R. (2010): Unsupervised pattern recognition models for mixed feature-type symbolic data. *Pattern Recognition Letters* 30, 430–443.
- DE SOUZA, R. M. C. R., DE CARVALHO, F. A. T. and PIZZATO, D. F. (2007): A Partitioning Method for Mixed Feature-Type Symbolic Data Using a Squared Euclidean Distance. *Lecture Notes in Computer Science* 4314, 260–273.
- GIBBS, A. L. and SU, F. E. (2002): On choosing and bounding probability metrics. *International Statistical Review* 70 (3), 419–435.
- IRPINO, A. and VERDE, R. (2006): A New Wasserstein Based Distance for the Hierarchical Clustering of Histogram Symbolic Data. In: H. H. Bock, W. Gaul and M. Vichi (Eds.): *Data Science and Classification*. Springer, Berlin, 185–192.
- IRPINO, A., VERDE, R. and LECHEVALLIER, Y. (2006): Dynamic clustering of histograms using Wasserstein metric. In: Rizzi, A. and Vichi, M. (Eds.): *COMPSTAT 2006 Proceedings in Computational Statistics*. Physica-Verlag, Berlin, 869–876.
- MACQUEEN, J. B. (1967): Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, 281–297.
- VERDE, R. (2004): Clustering Methods in Symbolic Data Analysis. In: H. H. Bock, W. Gaul and M. Vichi (Eds.): *Classification, Clustering, and Data Mining Applications*. Springer, Berlin, 299–317.
- VERDE, R. and IRPINO, A. (2008): Comparing Histogram Data Using a Mahalanobis-Wasserstein Distance. In: Brito, P. (Eds.): *COMPSTAT 2008 Proceedings in Computational Statistics*. Physica-Verlag, Berlin, 77–89.

# On Composite Pareto Models

Sandra Teodorescu<sup>1</sup> and Raluca Vernic<sup>2</sup>

<sup>1</sup> Faculty of Economic Sciences, Ecological University of Bucharest

1G Vasile Milea Blvd., Bucharest, Romania, *cezarina.teodorescu@yahoo.com*

<sup>2</sup> Faculty of Mathematics and Computer Science, Ovidius University of Constanta  
124 Mamaia Blvd., Constanta, Romania, and

Institute for Mathematical Statistics and Applied Mathematics, Casa Academiei

13 Calea 13 Septembrie, Bucharest, Romania, *rvernic@univ-ovidius.ro*

**Abstract.** To model statistical data generated by two different distributions, Cooray and Ananda (2005) introduced a composite Lognormal-Pareto model, further developed by Scollnik (2007). In this paper, we consider a more general composite Pareto model by replacing the Lognormal distribution with an arbitrary continuous one. The main characteristics of this model, as well as some statistical inference are presented. We will also provide comprehensive and numerical details to illustrate the particular case of the composite Gamma-Pareto model.

**Keywords:** composite distributions, Pareto distribution, Gamma distribution, statistical inference

## 1 INTRODUCTION

Statisticians sometimes encounter data obviously generated by two different models. This often happens in the case of insurance payment data, when actuaries must handle smaller data with high frequencies and occasional larger data with lower frequencies (see e.g. Kaas et al. (2001); Klugman et al. (2004)). Then the corresponding distribution can be modeled as a combination of two distributions, consisting of a less heavy-tailed distribution up to a certain threshold, and a heavy-tailed distribution afterwards. Such distributions are the composite models as suggested by Cooray and Ananda (2005). They constructed a composite model with the probability density function (pdf)

$$f(x) = \begin{cases} cf_1(x), & -\infty < x \leq \theta \\ cf_2(x), & \theta < x < \infty \end{cases}, \quad (1)$$

where  $f_1$  and  $f_2$  are pdf-s, while  $c$  is a normalizing constant resulting from continuity and differentiability conditions imposed in  $\theta$ . In Klugman et al. (2004), (1) is called a two-component spliced model applicable when the tail behavior is inconsistent with the small losses behavior.

In Cooray and Ananda (2005),  $f_1$  is considered to be the Lognormal pdf, while  $f_2$  is the Pareto one. Based on the same distributions, Scollnik (2007) proposed two different composite models, more general than the one studied

by Cooray and Ananda (2005). In the same manner, Teodorescu and Vernic (2009) suggested the composite Exponential-Pareto models. Moreover, Preda and Ciumara (2006) conducted a comparative study on the Weibull-Pareto and Lognormal-Pareto composite distributions.

In this paper, we go even further, considering a more general composite Pareto model: let  $f_1$  be an arbitrary pdf, while  $f_2$  remains Pareto. We choose the Pareto distribution because it is a classical heavy-tailed distribution, mostly used to model actuarial heavy-tailed claims. If we want to model insurance claims, then the choice of  $f_1$  will be restricted to distributions less-heavy tailed than the Pareto one and defined only for positive values.

Therefore, Section 2 is dedicated to the study of general composite models. In Section 2.1 we recall a mixture model equivalent to model (1), focusing on its characteristics and on two parameters estimation methods. Section 2.2 is dedicated to a similar study of the composite Pareto model. In Section 3 we introduce two new composite Gamma-Pareto models discussing some of their properties based on the general theory presented in Section 2. The suggested estimation methods are illustrated on the generated data.

In the following, we denote  $\mathbb{R}$  the set of real numbers and  $\mathbb{N}$  the set of all positive integers. Also, if  $f$  is a real function, then  $f'$  denotes its first derivative, while  $f''$  its second derivative.

## 2 COMPOSITE MODELS

### 2.1 Properties of a mixture model

**The model.** Scollnik (2007) noticed that the pdf of the composite model (1) can also be written as

$$f(x) = \begin{cases} rf_1^*(x), & -\infty < x \leq \theta \\ (1-r)f_2^*(x), & \theta < x < \infty \end{cases}, \quad (2)$$

where  $r \in [0, 1]$ , while  $f_1^*$ ,  $f_2^*$  are adequate truncations of the pdf-s  $f_1$  and  $f_2$ . More precisely, if  $F_i$  is the cumulative distribution function (cdf) of  $f_i$ , then

$$\begin{aligned} f_1^*(x) &= f_1(x)/F_1(\theta), & -\infty < x \leq \theta, \text{ and } 0 \text{ otherwise,} \\ f_2^*(x) &= f_2(x)/(1 - F_2(\theta)), & \theta < x < \infty, \text{ and } 0 \text{ otherwise.} \end{aligned} \quad (3)$$

It is easy to see that the pdf (2) can be interpreted as a two-component mixture model with mixing weights  $r$  and  $1-r$ , i.e.

$$f(x) = rf_1^*(x) + (1-r)f_2^*(x), \quad r \in [0, 1]. \quad (4)$$

In general, we prefer a smooth pdf (2), so we impose continuity and differentiability conditions in  $\theta$ . Combining the respective conditions for  $r$ , we obtain a restriction for  $\theta$ , i.e.  $f_1(\theta)/f_2(\theta) = f_1'(\theta)/f_2'(\theta)$ .

Teodorescu and Vernic (2009) identified certain properties of model (2) that we present below. If  $X$  is a random variable (r.v.) with pdf  $f$ , we denote its

$n$ -th order initial moment by  $E_n(f) = E(X^n)$ , and its characteristic function by  $\varphi_f(t) = \varphi_X(t) = E(e^{itX})$ .

**Proposition 2.1**

a) Let  $F$  denote the cdf of the pdf given in (2). Then

$$F(x) = \begin{cases} rF_1(x)/F_1(\theta), & -\infty < x \leq \theta \\ r + (1-r)(F_2(x) - F_2(\theta))/(1 - F_2(\theta)), & \theta < x < \infty \end{cases}.$$

b) Assuming that all the quantities involved exist, the  $n$ -th order initial moment of the density function (2) is  $E_n(f) = rE_n(f_1^*) + (1-r)E_n(f_2^*)$ , while its characteristic function is  $\varphi_f(t) = r\varphi_{f_1^*}(t) + (1-r)\varphi_{f_2^*}(t)$ ,  $t \in \mathbb{R}$ .

**Generating random values from the composite density (2)**

To this end, we suggest using the inversion method, assuming that both  $F_1, F_2$  admit inverse functions. Therefore, if  $u$  is a value generated from the uniform distribution  $U(0,1)$ , we obtain a value  $x$  generated from (2) as

- a. If  $u \leq F(\theta) = r$ , then solve  $u = rF_1(x)/F_1(\theta)$  for  $x$ , i.e.  $x = F_1^{-1}(uF_1(\theta)/r)$ ;
- b. If  $u > r$ , then solve  $u = r + (1-r)(F_2(x) - F_2(\theta))/(1 - F_2(\theta))$  for  $x$ , i.e.  $x = F_2^{-1}((u-r)/(1-r) + F_2(\theta))$ .

**Statistical inference**

**I. A first algorithm.** In Teodorescu and Vernic (2009), the following algorithm based on the maximum likelihood (ML) method was suggested. Assume that the pdf (2) depends on the real parameters  $\delta_1, \dots, \delta_s, \theta$ , where  $s \in \mathbb{N}$ , and consider the random data sample  $(x_1, \dots, x_n)$ . Without loss of generality, assume that it is an ordered sample, i.e.  $x_1 \leq x_2 \leq \dots \leq x_n$ . To apply the ML method, we must know the integer value  $m$  so that the unknown parameter  $\theta$  is in between the  $m$ -th and  $(m+1)$ -th observations, i.e.  $x_m \leq \theta \leq x_{m+1}$ . If somehow we know this  $m$ , the likelihood function is

$$L(x_1, \dots, x_n; \delta_1, \dots, \delta_s, \theta) = r^m (1-r)^{n-m} \prod_{i=1}^m f_1^*(x_i) \prod_{j=m+1}^n f_2^*(x_j). \quad (5)$$

Unfortunately, in general, we don't know the exact value of  $m$ ; moreover, if  $m$  changes, the ML estimation also changes. Therefore, we suggest the following estimation algorithm that takes into consideration all possible values of  $m$  so that  $x_m \leq \theta \leq x_{m+1}$ :

- Step 1.* For each  $m = 1, 2, \dots, n-1$ , do: let  $\hat{\delta}_1, \dots, \hat{\delta}_s, \hat{\theta}$  be the solutions of the ML system  $\left\{ \frac{\partial \ln L}{\partial \delta_i} = 0, i = 1, \dots, s; \frac{\partial \ln L}{\partial \theta} = 0 \right\}$ . If  $x_m \leq \hat{\theta} \leq x_{m+1}$ , then check the restriction in  $\theta$  for the current estimations, and if it is verified, then take  $\hat{\delta}_i^{ML} = \hat{\delta}_i, i = 1, \dots, s; \hat{\theta}^{ML} = \hat{\theta}$ .
- Step 2.* If Step 1 doesn't give any solution for  $\theta$ , then  $m=n$  or  $m=0$ , hence we recommend using only  $f_1$  and, respectively,  $f_2$  for the likelihood function.

**Remark 2.1** In this case, we must check  $n-1$  intervals. If more candidates for  $\theta$  emerge from Step 1, then a statistical test or a selection criterion can be used to choose the one that gives the best fit. Also, since Step 1 is based on the ML method, the resulting estimators are consistent and asymptotically normal distributed. Note that, if  $n$  is large and the ML system very complex, it might be difficult to implement the algorithm. So, we suggest an alternative algorithm based on MM (method of moments) and on quantiles matching.

**II. A second algorithm.** As already seen, our main problem is that we don't know anything about the magnitude of the parameter  $\theta$ . Therefore, assuming as before that (2) depends on  $s+1$  unknown parameters, we suggest the following alternative algorithm:

*Step 1.* Let  $q_1$  and  $q_3$  be the first and, respectively, the third empirical quantiles of the data sample. Assume that  $q_1 < \theta < q_3$ . Then we apply the MM to match the first  $s-1$  empirical moments with the corresponding theoretical ones, adding two more equations obtained by matching the two quartiles, i.e. add 
$$\begin{cases} rF_1(q_1)/F_1(\theta) = 0.25 \\ r + (1-r)(F_2(q_3) - F_2(\theta)) / (1 - F_2(\theta)) = 0.75 \end{cases}.$$

If the resulting system has no solution, then go to Step 2.

*Step 2.* Assume that  $q_1, q_3 < \theta$  and follow Step 1. The two respective equations are now 
$$\begin{cases} rF_1(q_1)/F_1(\theta) = 0.25 \\ rF_1(q_3)/F_1(\theta) = 0.75 \end{cases}.$$
 If, again, there is no corresponding solution, go to Step 3.

*Step 3.* Finally, let assume that both  $q_1, q_3 > \theta$  and follow Step 1. The two equations are 
$$\begin{cases} r + (1-r)(F_2(q_1) - F_2(\theta)) / (1 - F_2(\theta)) = 0.25 \\ r + (1-r)(F_2(q_3) - F_2(\theta)) / (1 - F_2(\theta)) = 0.75 \end{cases}.$$

**Remark 2.2** Once we have a solution for this second algorithm, we can use the first one to improve it, since now we have information on  $\theta$ .

## 2.2 Some properties of a composite Type II Pareto model

In this section, let  $f_1$  be a general pdf, while  $f_2$  is a Pareto density.

The composite Type II Pareto model will be developed starting from model (1), based on a version of the generalized Pareto distribution (GPD) exceeding the threshold value  $\theta$ , i.e. we consider

$$f_2(x) = \alpha(\gamma + \theta)^\alpha / (\gamma + x)^{\alpha+1}, \text{ where } \gamma > -\theta, \theta, \alpha > 0, x > \theta. \quad (6)$$

This is also known as the Lomax or Type II Pareto distribution, see e.g. Johnson et al. (1994).

**Proposition 2.2** *The composite Type II Pareto pdf is*

$$f(x) = \begin{cases} rf_1(x)/F_1(\theta), & -\infty < x \leq \theta \\ (1-r)\alpha(\gamma + \theta)^\alpha / (\gamma + x)^{\alpha+1}, & \theta < x < \infty \end{cases}, \quad (7)$$



where  $\alpha, \theta > 0, \gamma > -\theta$  and  $0 \leq r \leq 1$  satisfy the conditions

$$\alpha + 1 = -(\gamma + \theta) f_1'(\theta) / f_1(\theta), \quad (8)$$

$$r = \alpha f_1'(\theta) F_1(\theta) / \left( \alpha f_1'(\theta) F_1(\theta) - (\alpha + 1) f_1^2(\theta) \right). \quad (9)$$

*Proof.* The pdf (7) is obtained from (1) and (2) using  $f_2$  from (6), and noting that  $F_2(\theta) = 0$ . This pdf has at least four parameters:  $\alpha, \theta, \gamma$  and  $r$ . We impose a continuity condition in the threshold  $\theta$ , i.e.  $f(\theta - 0) = f(\theta + 0)$ . This gives

$$r = \alpha F_1(\theta) / (\alpha F_1(\theta) + (\gamma + \theta) f_1(\theta)). \quad (10)$$

We should ensure that the respective pdf is smooth if we impose a differentiability condition in  $\theta$ , i.e.  $f'(\theta - 0) = f'(\theta + 0)$ . This gives

$$r = \alpha(\alpha + 1) F_1(\theta) / \left( \alpha(\alpha + 1) F_1(\theta) - (\gamma + \theta)^2 f_1'(\theta) \right). \quad (11)$$

From (10) and (11) we easily obtain condition (8). Inserting  $\gamma + \theta$  into (10) we obtain the expression (9) of  $r$ .  $\square$

**Remark 2.3** Because of conditions (8)-(9), the number of unknown parameters was decreased by two. To further reduce it, we impose the condition  $f''(\theta - 0) = f''(\theta + 0)$ , so  $r f_1''(\theta) / F_1(\theta) = (1 - r) \alpha(\alpha + 1)(\alpha + 2)(\gamma + \theta)^{-3}$ .

Using Proposition 2.1, the cdf of (7) becomes

$$F(x) = \begin{cases} r F_1(x) / F_1(\theta), & -\infty < x \leq \theta \\ 1 - (1 - r)(\gamma + \theta)^\alpha / (\gamma + x)^\alpha, & \theta < x < \infty \end{cases}. \quad (12)$$

From Proposition 2.1 and (6), we can get the  $n$ -th order moment of the composite Type II Pareto distribution as

$$E_n(f) = r E_n(f_1^*) + (1 - r) \alpha \sum_{k=0}^n \binom{n}{k} \frac{(-\gamma)^{n-k} (\gamma + \theta)^k}{\alpha - k}, \quad \alpha > n. \quad (13)$$

### 3 COMPOSITE GAMMA-PARETO MODELS

#### 3.1 The composite Gamma-Type II Pareto model

The composite Gamma-Type II Pareto model will be developed based on model (7). We denote  $\Gamma(\nu) = \int_0^\infty x^{\nu-1} e^{-x} dx$  the Gamma function and  $\Gamma(\nu; t) = \int_0^t x^{\nu-1} e^{-x} dx$ ,  $\nu, t > 0$ , the lower incomplete Gamma function. For details on the Gamma distribution see e.g. Johnson et al. (1994).

**Proposition 3.1** *The composite Gamma-Type II Pareto pdf is*

$$f(x) = \begin{cases} r \beta^\delta x^{\delta-1} e^{-\beta x} / \Gamma(\delta; \beta \theta), & 0 < x \leq \theta \\ (1 - r) \alpha (\gamma + \theta)^\alpha / (\gamma + x)^{\alpha+1}, & \theta < x < \infty \end{cases}, \quad (14)$$

where  $\beta, \delta, \alpha, \theta > 0$ ,  $\gamma > -\theta$  and  $0 \leq r \leq 1$  satisfy the conditions

$$\alpha + 1 = (\gamma + \theta) (\beta\theta - \delta + 1) / \theta, \quad (15)$$

$$r = \frac{\alpha (\beta\theta - \delta + 1) \Gamma(\delta; \beta\theta)}{\alpha (\beta\theta - \delta + 1) \Gamma(\delta; \beta\theta) + (\alpha + 1) (\beta\theta)^\delta e^{-\beta\theta}}. \quad (16)$$

*Proof.* We obtain pdf (14) based on (7) assuming  $f_1(x) = \beta^\delta x^{\delta-1} e^{-\beta x} / \Gamma(\delta)$ ,  $x > 0$  (i.e. a Gamma pdf) and noting that  $F_1(\theta) = \Gamma(\delta; \beta\theta) / \Gamma(\delta)$ . Hence, the truncated Gamma pdf is  $f_1^*(x) = \beta^\delta x^{\delta-1} e^{-\beta x} / \Gamma(\delta; \beta\theta)$ ,  $0 < x \leq \theta$ . Inserting now  $f_1(\theta)$ ,  $f_1'(\theta)$  and  $F_1(\theta)$  into (8) and (9), we can easily obtain (15) and (16).  $\square$

**Remark 3.1** Because of conditions (15)-(16), the number of unknown parameters decreased from six to four (e.g. expressing  $r$  and  $\alpha$  in terms of  $\beta, \gamma, \delta$  and  $\theta$ ). From Remark 2.3 and (15)-(16), we obtain  $\gamma(1 - \delta) = \beta\theta^2$ . Thus, the number of unknown parameters can be further reduced to three.

Since  $F_1(x) = \Gamma(\delta; \beta x) / \Gamma(\delta)$ , from (12) we obtain the cdf of (14) as

$$F(x) = \begin{cases} r \Gamma(\delta; \beta x) / \Gamma(\delta; \beta\theta), & 0 < x \leq \theta \\ 1 - (1 - r) (\gamma + \theta)^\alpha / (\gamma + x)^\alpha, & \theta < x < \infty \end{cases}. \quad (17)$$

Moreover, from Proposition 3.1 and (13), we get the  $n$ -th order initial moment of the composite Gamma-Type II Pareto model for  $\alpha > n$ ,

$$E_n(f) = r \frac{\Gamma(n + \delta; \beta\theta)}{\beta^n \Gamma(\delta; \beta\theta)} + (1 - r) \alpha \sum_{k=0}^n \binom{n}{k} \frac{(-\gamma)^{n-k} (\gamma + \theta)^k}{\alpha - k}. \quad (18)$$

### 3.2 The composite Gamma-Pareto model

This model represents a particular case of the composite Gamma-Type II Pareto model for  $\gamma = 0$ . From Proposition 3.1, we can easily obtain the composite Gamma-Pareto pdf

$$f(x) = \begin{cases} r \beta^\delta x^{\delta-1} e^{-\beta x} / \Gamma(\delta; \beta\theta), & 0 < x \leq \theta \\ (1 - r) \alpha \theta^\alpha / x^{\alpha+1}, & \theta < x < \infty \end{cases}, \quad (19)$$

where  $\alpha, \beta, \delta, \theta > 0$  and  $0 \leq r \leq 1$  satisfy the conditions

$$\alpha = \beta\theta - \delta, \quad r = \frac{(\beta\theta - \delta) \Gamma(\delta; \beta\theta)}{(\beta\theta - \delta) \Gamma(\delta; \beta\theta) + (\beta\theta)^\delta e^{-\beta\theta}}. \quad (20)$$

Moreover, the cdf of the composite Gamma-Pareto model is

$$F(x) = \begin{cases} r \Gamma(\delta; \beta x) / \Gamma(\delta; \beta\theta), & 0 < x \leq \theta \\ 1 - (1 - r) (\theta/x)^{\beta\theta - \delta}, & \theta < x < \infty \end{cases}, \quad (21)$$

while its  $n$ -th order initial moment is, for  $\alpha > n$ ,

$$E_n(f) = r\Gamma(n + \delta; \beta\theta) / (\beta^n \Gamma(\delta; \beta\theta)) + (1 - r)\alpha\theta^n / (\alpha - n). \quad (22)$$

**Remark 3.2** Due to (20), we have three unknown parameters. We can't decrease their number based on the equality of the second derivatives because for  $\gamma = 0$ , Remark 3.1 leads to  $\beta\theta^2 = 0$ , which is impossible.

**Statistical inference.** We start from the second algorithm in Section 2.1. The unknown parameters are  $\beta, \delta$  and  $\theta$ . In addition to the quartiles  $q_1, q_3$ , we need one more MM equation. We choose  $\bar{x} = E_1(f)$ , where  $\bar{x}$  is the empirical mean of the data sample. We take  $n = 1$  in (22), denote  $\xi = \beta\theta$  and replace  $\theta = \xi/\beta$ . Then, using the expressions of  $r$  and  $\alpha$  given in (20), we obtain

$$E_1(f) = \frac{\xi - \delta}{\beta((\xi - \delta)\Gamma(\delta; \xi) + \xi^\delta e^{-\xi})} \left( \Gamma(\delta + 1; \xi) + \frac{\xi^{\delta+1} e^{-\xi}}{\xi - \delta - 1} \right), \quad \alpha > 1.$$

Following Step 1 of the second algorithm, let assume that  $q_1 < \theta < q_3$  and add two more equations for these quartiles. From (21), these equations are

$$\begin{cases} 0.25 = \frac{(\xi - \delta)\Gamma(\delta; \beta q_1)}{(\xi - \delta)\Gamma(\delta; \xi) + \xi^\delta e^{-\xi}} \\ 0.25 = \frac{\xi^\xi e^{-\xi}}{((\xi - \delta)\Gamma(\delta; \xi) + \xi^\delta e^{-\xi})(\beta q_3)^{\xi - \delta}} \end{cases}. \quad (23)$$

The two equations must be numerically solved along with  $\bar{x} = E_1(f)$ . Steps 2 and 3 are carried out in a similar way.

When using the first algorithm based on the ML method, things get more complicated because one equation involves the derivative of the incomplete Gamma function,  $\frac{\partial \Gamma(\delta; \xi)}{\partial \delta}$ , not implemented in the usual software. In other words, assuming that the data sample is ordered and that  $x_m \leq \theta \leq x_{m+1}$  (information based on e.g. the algorithm above), denoting  $S_{ab} = \sum_{i=a}^b x_i$ ,  $P_{ab} = \prod_{j=a}^b x_j$ , and using the ML system, we have

$$\begin{cases} \beta = (\delta n + \xi(m - n))/S_{1m} \\ \frac{n\xi^\delta e^{-\xi}}{(\xi - \delta)((\xi - \delta)\Gamma(\delta; \xi) + \xi^\delta e^{-\xi})} + (n - m) \ln \frac{\xi}{\beta} - \ln P_{m+1n} = 0 \\ \frac{\xi^\delta e^{-\xi} (1 + (\xi - \delta) \ln \xi) + (\xi - \delta)^2 \frac{\partial \Gamma(\delta; \xi)}{\partial \delta}}{(\xi - \delta)((\xi - \delta)\Gamma(\delta; \xi) + \xi^\delta e^{-\xi})} - \ln \beta - \frac{1}{n} \ln P_{1n} = 0 \end{cases}.$$

We suggest avoiding the last equation which might cause problems, combining instead the first two equations with e.g. the MM equation  $\bar{x} = E_1(f)$ .

**Numerical example.** To illustrate the two estimation methods described above, we used  $n=5000$  data (generated by inversion method starting from density (19)). The real values of the parameters are  $\theta = 3, \beta = 2, \delta =$

$4 \Rightarrow \alpha = 2, r = 0.7602, \xi = 6$ , while the empirical mean and quartiles are  $\bar{x} = 2.6785, q_1 = 1.3260, q_3 = 2.8793$ . We solved system (23) and the equation  $\bar{x} = E_1(f)$  using Mathematica software, and obtained the solution  $\tilde{\xi} = 6.1962, \tilde{\beta} = 2.1478, \tilde{\delta} = 4.2016$ , so  $\tilde{\theta} = 2.8849, \tilde{\alpha} = 1.9946, \tilde{r} = 0.7509$ .

Starting from this value of  $\tilde{\theta}$ , we applied the ML based algorithm suggested above to see if the parameters values can be improved. This was indeed the case, and we obtained the new slightly improved solutions  $\hat{\xi} = 6.0687, \hat{\beta} = 2.0607, \hat{\delta} = 4.0662$ , so  $\hat{\theta} = 2.9449, \hat{\alpha} = 2.0025, \hat{r} = 0.7579$ . We started from  $m=3753$ , but finally obtained  $m=3834$ .

We also applied the  $\chi^2$  test to check the distribution fitting for both sets of parameters, and to see which fit is best. For the first set of parameters,  $\chi^2=11.7719$ , while for the ML parameters  $\chi^2=11.7153$ . The corresponding 0.05  $\chi^2$  quantile is 14.0671, hence both composite Gamma-Pareto distributions fit the respective data, but since the  $\chi^2$  distance is smaller for the ML parameters, their fit is best, as expected.

## References

- COORAY, K. and ANANDA, M.A. (2005): Modeling actuarial data with a composite Lognormal-Pareto model. *Scandinavian Actuarial Journal* 5, 321–334.
- JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1994): *Continuous Univariate Distributions*. Wiley, New York.
- KAAS, R., GOOVAERTS, M., DENUIT, M. and DHAENE, J. (2001): *Modern Actuarial Risk Theory*. Kluwer Academic Publishers, Boston.
- KLUGMAN, S.A., PANJER, H.H. and WILLMOT, G.E. (2004): *Loss Models: from Data to Decisions* (2nd edition). New York, John Wiley & Sons, Inc.
- PREDĂ, V. and CIUMARA, R. (2006): On composite models: Weibull-Pareto and Lognormal-Pareto. A comparative study. *Romanian Journal of Economic Forecasting* 3 (2), 32–46.
- SCOLLNIK, D.P.M. (2007): On composite Lognormal-Pareto models. *Scandinavian Actuarial Journal* 1, 20–33.
- TEODORESCU, S. and VERNIC, R. (2009): Some composite Exponential-Pareto models for actuarial prediction. *Romanian Journal of Economic Forecasting* 12 (4), 82–100.

# Visualisation of Large Sized Data Sets: constraints and improvements for graph design

Jean-Paul Valois<sup>1</sup>

TOTAL Exploration Production, F64018 Pau Cedex, *jean-paul.valois@total.com*

**Abstract.** Background and history of Data Visualisation are actualised taking into account recent workings from several fields (ergonomy, neurophysiology). The paper outlines the constraints resulting from Large Sized Data Sets and highlights some suitable improvements of software or graphs. Displaying density of points appears as an important challenge, namely for 2D scatter plots and parallel coordinates plots.

**Keywords:** data visualisation, exploratory data analysis, histogram, scatter plot, parallel coordinates, statistical population density

## 1 Introduction and background

In the end of the eighteenth century arose the idea of using graphical display for rendering salient information from data (e.g. by Playfair, (1786)). This idea has been magnified by many authors (e.g. Bertin (1967), Tukey (1977), Cleveland (1993)). Instead of displaying predefined information, they used graphs as “partners” (Tukey, (1990)) for obtaining answer to questions: graphs are not constrained by distribution hypotheses; besides algorithmic models cannot always reflect all the aspects of data. Graphs should be used to first verify whether a model is suitable or not (Cleveland (1993)). ‘Graphical Data Analysis’ also called ‘Visual Data Mining’ is so ranked among the exploratory methods, this last word has been preferred since Tukey (1977).

Playfair himself perfectly expressed fundamentals of this technique (Spence and Wainer (2005)). Tukey (1977) wanted what the graph “forces us to notice what we never expected to see”; nevertheless he did not really propose any keys to define this impact. To obtain it, Bertin (1977) used visual variables and emphasized the interest of data reordering in agreement with previous works (Wilkinson and Friendly (2009)). More details about the history of graphs in the statistics field can be found in Beniger and Robin (1978), some more points are mentioned hereafter to actualize this background.

Experiments of cognitive psychology completed the first Bertin’s intuitions concerning the visual variables (e.g. Kolata (1984), Lewandowski and Spence (1989)). Neurophysiological researches drawn attention again into the preattentive vision (Julesz (1991)): the first steps of perception influences all the subsequent perceptive process (Zeki (1992)). Besides, Dehaene et al.

(1999) postulated that inherited brain structures give us capabilities to manage quickly simple computations; the way for perceiving numbers is today conceived as a “mental line”. So the help of spatial perception for appreciating quantities seems in agreement with neurophysiological modern concepts.

In the other hand, ergonomic researches (Pomeranz et al. 1977, Carswell and Wickens (1987), Danek and Koubek (1995)), concluded there is better to display together the pieces of information which are needed in a decision process. Configural effects (like symmetry) could help to pick up more quickly salient information (Benett and Flach (1992); this can help to conceive new types of graphs (Massie et al. (2009)). In any way, a close connection between the visual task in graphs and the asked question seems to be a key point for the power (“impact”) of graphs (Mitchell and Byers (1992).

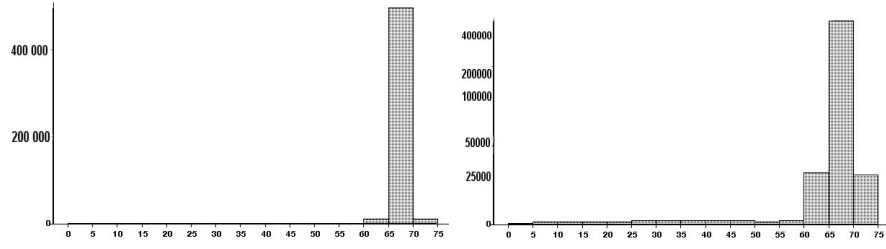
In summary, Graphical Analysis deals with two aspects (Benett and Flach (1992). The semantic of the graph often needs suitable data preprocessing to exhibit what the data have to say (Tufte (1983)). Graph design or visual syntax manages what the reader can perceive (see practical advices in Kosslyn (1994, 2006) and Gillian et al. (1998)); intuitions for a better design occurred in fact about one century before Playfair (Palsky (1996)). So we have both to transfer requested information and to organise the visual elements in such a way that suitable information could be pick up at once by inherited brain structures, the higher level process of perception being thus more devoted to thinking about data structure himself.

In this paper we consider the new challenge we are faced to using Large Sized Data Sets (chap. 2), and wonder whether improvements are needed for graphical techniques. The following purpose is organized according to the input variables (Valois (2000)): one variable (chap. 3), two variables (chap. 4) and more (chap. 5). The topic is limited to the more classical graphs encountered in the statistic field.

## 2 The challenge of Large Sized Data Sets (LSDS)

In the last decades, data dimension exploded in many industries. This first involves the size of the data base. Besides categorical variables became also more numerous and may be hierarchically organised. In oil industry, pressure and temperature measured at several check points of producing wells each 5 seconds result into 5 millions of lines during only 6 months (10 to 30 variables); in the other hand, a world country can provide hundreds of fields, each of them including some hundreds of wells ; hydrocarbons can come from tens of dynamic layers or sedimentary bodies. Because measurements can come from several physical origins, variables can be weakly correlated and a lot of artefacts or outliers can be found.

Using LSDS, discussion whether tables or graphs are more suitable is back and treillis plots could provide hundreds of graphs side by side. An alternative way is to switch the content of a window from one item to another one by



**Fig. 1.** Distribution of 520 000 Temperature measurements, left: histogram, right: parametric rootogram.

mouse clicking. In both cases we need to inspect graphs ranked according to mixed items or properties (e.g. the wells classified per geological layer for instance, in each of them the more producing first).

Dynamic link between graphs is a compromise between the need to look at all aspects of data and to provide selected elements to the preattentive vision. So we need software which provide both dynamic link, no limitation in memory and allow a free mixing of categorical variables. Speed is a prerequisite to agree with the brain behaviour.

### 3 One or two quantitative variables

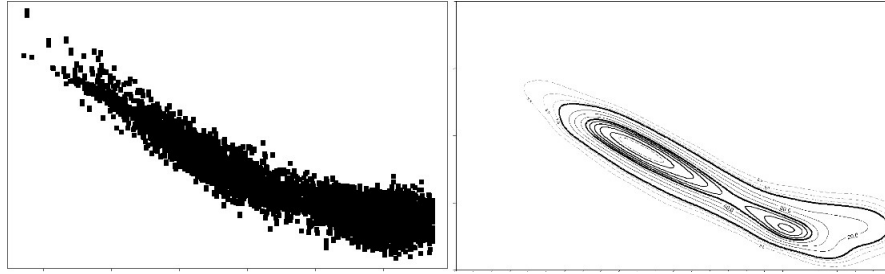
Using one variable, histogram is an old experimental approach of population density. In case of unbalanced frequencies, the smallest bars can be not visible any more (Figure 1, left). The “rootogram” (Chernoff (1973)) displays as y axis the square root of frequencies; this correction can not be enough for LSDS. So we propose a parametric rootogram where y values are the k-root of the observed counts:

$$y = count^{(1/k)}, k = (1/\log_{10}(Yr) * (\log_{10}(count_{max}) - \log_{10}(count_{min}))) \quad (1)$$

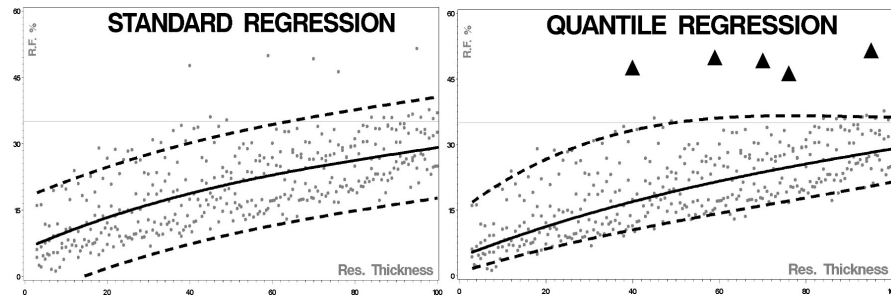
in which Yr is the suitable ratio between the size of extreme bars, say 100 for a graph of 10 cm; countmax and countmin are the extrema found in bar frequencies. Values of  $k$  less than 1 are anyway ceiled to 1, resulting into a standard histogram when the vertical height is suitable.

With two variables, the history of scatter-plot has been reviewed by Friendly and Denis (2005). Indeed the x,y axes can result from prior computations (PCA, MDS). In the LSDS context, improvements like markers, color and size of symbols cannot be considered because overlapped points. Displaying the density is therefore the main challenge, either using the experimental density or an estimated value of it (kernel method for instance).

Instead of computing the cumulative density from the file ranked by increasing values (like for CDF curve), Bowman and Foster (1993) rank the file by decreasing density, so outliers and queues of population are integrated in



**Fig. 2.** Pressure (x) vs Depth (y), 20 000 obs., left: scatter plot, right: contour curves of density.



**Fig. 3.** Relation Recovery Factor (y) vs Oil Thickness (x), 450 obs.

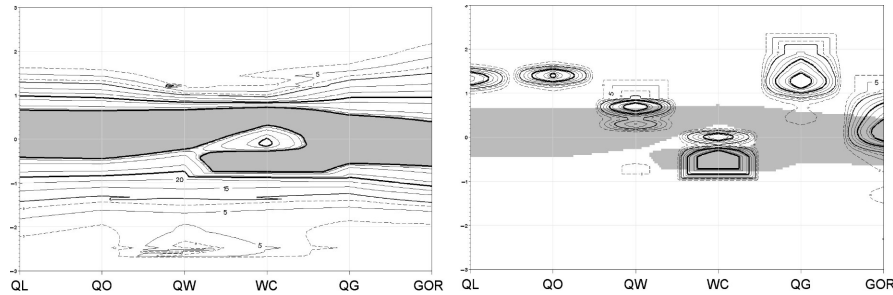
the last computational steps. In Figure 2, the scatter plot (left) could suggest irregularities in the trend, density contour curves (right) show a perfect linear trend for the most part of the population. This method highlights more details in the middle of population than do ellipses proposed as 2D box-plot. Such a nested surfaces can be used to flag the outlier points, as pointed by Mazeika et al. 2008, these propose a more comprehensive discussion of surface definition.

When a (x,y) structure is suspected, the more general approach seems today the quantile regression (Koenker, 2005). Figure 3 (right) shows a linear relationship in the lower y values, the higher ones being thresholded with an asymptotic trend; in this case, outlier values were presumed by the end user and are nicely found. The LOESS algorithm by Cleveland (1993) is today a well established method; as it involves an L2 regression, resulting interval of confidence is symmetric, so it cannot be applied to all the cases.

#### 4 More variables

Inselberg (1985) proposed a system with parallel y axes and pointed that correlations between variables are translated into segments which converge either outer of related axes (positive correlations), or inner them (negative correlations). If any, a prior variable normalization acts in fact as a metric





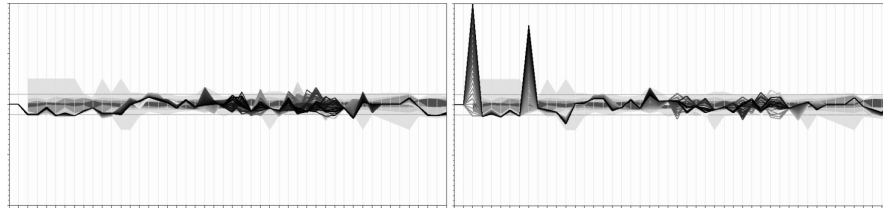
**Fig. 4.** Parallel coordinates with a Data Base of Oil Fields.

which modifies the location of such crossing points. Indeed correlations are only visible for the side by side variables; these can be ranked according to PCA order or equivalent techniques.

LSDS can result in a very large amount of broken lines, so the figure can become hard to read. Fua et al.(1999) proposes a multiresolutional view, a clustering of data groups, and finally a band representation - with possible effects of variable-width density. Ergonomics tells us to design graphs such a way where visual task agrees with that we first see: in this case the lines representing observations are seen first, and applications rather use this technique to select and describe observations or subsets of data (Steed et al. (2009)) rather than variables interpretation. Some more diversified visual effects have been proposed by Wegman and Luo (2002).

Our example Figure 4 displays some properties of wells (monthly rates of oil, water...) from a data base of 600 000 monthly lines. The challenge is to quickly highlight the features of some fields among a lot (each field include tens of wells). Cumulative density contour curves (computed as in section 3) provide a more detailed description than classical Tukey boxes, see for instance the bimodal distribution of WC. The left sub-figure displays the whole data base (400 000 obs.); the 50% contour curve is overlapped as a grey background in the right subfigure. This last one displays the density curves for one field from the data base: this provides higher values for QO, resulting in lower values in  $WC(= QW/QO+QW)$ . So these two ways of display used together allow to highlight the properties of a subset of data (here one selected field).

Our second example deals with alarming methods (Valois et al., (2007)); the challenge is both to display a very large number of observations (normal situation), and to exhibit a limited set of these when occurs an anomalous event. Parallel coordinates provide a way to control the 30 available variables in one view, Figure 6 displays both density curves corresponding to the normal situation and broken lines corresponding to observations taken at the beginning of alarm (left and right: two successive time slices). X and Y axes exclude the time, so we use width and transparency for suggesting the time evolution of the anomalous lines.



**Fig. 5.** Alarming control display in parallel coordinates, Xi: Temp. and Pressures at different check points.

## 5 Discussion

Difficulties for computing statistical inference with Large Sized Data Sets are often considered, the question of considering graph design in this context is less frequently pointed. The paper is devoted to consider the constraints in this context. It does not provide really new type of graphs, we rather suggest some tracks for a more suitable design. The examples show in fact that these constraints can surprisingly be effective even using a few variables (histogram and scatter-plot), or with medium sized data sets (scatter plot, e.g. from 1000 observations if overlapped items), so our suggestions could perhaps be useful in a broader context.

Rootogram can challenge histogram, both differing only by the graph design; the suggested parametric rootogram could be proposed in software because it converges with histogram in low sized Data Sets and can adapt to the size of data set. Another way to better display histograms could indeed use outlier detection tools, these were not considered here: in an exploratory perspective, outliers should be first observed before perhaps excluded.

Although displaying density is became the most common way when considering one variable (histogram, or Tukey box plot), most often scatter plot is used for displaying each observation with two variables. Density contour curves or quantiles regression appear as powerful and underemployed tools, Loess algorithm is often the one mentioned in an EDA context or in software. All these tools involve data preprocessing to extract suitable information and transfer these into graph elements so the preattentive vision can capture the salient information at once. There are indeed a lot of computational methods for estimating the density, a discussion of this point was not in the scope of this paper. We only outlined the interest of cumulative density computed from higher to lower density slices of the variable.

For LSDS, the main challenge lies perhaps into the high dimensional plots like radar plot or parallel coordinates. As a centred figure, radar plot can be used on geographical maps (Valois (2000)). Parallel coordinates offer conversely more possibilities for interactive dynamic actions (Steed et al. (2009)); they can be used in different ways for comparing data subsets. As for the scatter plot, a hot point is to compute and display the population density, and make the main population visually as distinct as possible from the subset of

interest or outlier data. According to ergonomic advices, as far as more complex information from LSDS is concerned, we need to superimpose elements, this results in more complex graphs. Far from the ideal of simplicity promoted by Tufte (1983), the concept of perceptual units (Kosslyn (1994)) becomes now a hot point. This context and principles can suggest new types of graphs, as pointed by the background section of this paper, on this point we agree several authors; nevertheless our topic was limited to the more traditional graphs.

## References

- BENETT, K.B. and FLACH, J.M. (1992): Graphical displays: implications for divided attention, focused attention and problem solving, *Human Factors*, 34 (5), 513-533.
- BENIGER J.R., ROBYN D.L. (1978): Quantitative Graphics in Statistics : A brief history, *The American Statistician*, 32, 1,1-11.
- BERTIN, J. (1967): *La semiologie graphique*, Paris, Gauthier Villars.
- BERTIN, J. (1977): *La graphique et le traitement graphique de l'information*, Flammarion.
- BOWMAN, A. W. and FOSTER, P. J. (1993): Density Based Exploration of Bivariate Data, *Statistics and Computing*, 3, 171-177.
- CARSWELL, C.M. and WICKENS C.D. (1987): Information integration and the object display, *Ergonomics*, 30, 511-527.
- CHERNOFF H. (1973): The use of faces to represent points in k-dimensionnal space graphically, *Journal of the american statistical association*, 68, 342, 361-368.
- CLEVELAND, W.S. (1993): *Visualizing Data*, Hobart Press.
- DANEK, A.M. and KOUBEK R.J. (1995): Mapping perceptual and cognitive processing for the effective use of graphical displays in shop floor scheduling tasks, *The intern. Journal of Human factors in manufacturing*, 5, 4, 401-415.
- DEHEANE, S., SPELKE, E., PINEL, P., STANESCU, R. and TSIVKIN S. (1999): Sources of mathematical thinking: behavioral and brain-imaging evidence, *Science*, 284, 970-4.
- FRIENDLY, M. and DENIS, D. (2005): The early origins and development of the scatterplot, *Journal of the history of the behavioral sciences*, (2005), 41(2), 103-130, refs. 1 p.1/2.
- FUA, Y.H., WARD M.O., RUDENSTEINER E.A., Hierarchical Parallel Coordinates for Exploration of Large Datasets, *iee vis*, pp.4, 10th IEEE Visualization 1999 (Vis'99), 1999, 9pp.
- GILLIAN, D.J., WICKENS, C.D., HOLLANDS, J.G., CARSWELL, C.M. (1998): Guidelines for Presenting Quantitative data in HFES Publications, *Human Factors*, 40, 1, 28-412, march 1998.
- INSELBERG, A. (1985): The plane with parallel coordinates, Special issue on computational geometry, *The visual Computer*, I, 69-97.
- JULESZ, B. (1991): Early vision and focal attention, *Reviews of Modern physics*, 63, 735-772.
- KOENKER (2005): Quantile Regression, *Series Economic Society Monographs*, 38; Cambridge University Press, 366 pp.

- KOLATA, G. (1984): The proper display of data, *Science*, 226, 156-157.
- KOSSLYN, S.M. (1994): *Elements of graph Design*, Freeman and Co., New-York, 309 pp.
- KOSSLYN, S.M. (2006): *Graph Design for the Eye and Mind*, Oxford University Press, 2006.
- LEWANDOWSKI, S. and SPENCE, I. (1989): Discriminating strata in scatterplots, *Journ. Of Am. Stat. Ass.*, 84, 407, 682-688.
- MASSIE, A.E., KOPYLOV, I., CUMMINGS, M.L. (2009): Supporting system Acquisition Decisions through Ecological Perception, 47th AIAA Aerospace Science Meeting Including The New Horizons Forum and Aerospace Exposition, 5-8 Jan 2009, Orlando, Florida. Ref AIAA 2009-1196.
- MAZEIKA, A., BHLEN M.H., MYLOV, P. (2008): Using Nestes Surfaces for Visual Detection of Structures in Databases, in *Visual Data Mining*, Simoff S.J. ed., Springer Verlag, 2008, 91-102.
- MITCHELL, J.A. and BIERER D.W. (1992): Decision statistic mapping and number of information dimensions on decision making with graphical displays, in: *Proceedings of the human factor society, 36th. annual meeting, 1503-1507*.
- PALSKY, G., (1996): *Des chiffres et des cartes, naissance et dveloppement de la cartographie quantitative au XIXme sicle*. Paris, CTHS, Mmoire de la section de gographie, n19, 332pp.
- PLAYFAIR, W. (1786): *The commercial and political atlas*, London, Carry. Rd; 2005, avec : Introduction by Wainer H., Spence I, ISBN-13:9780521855549 — ISBN-10:0521855543.
- POMERANZ, J.R., SAGER, L.C. and STOEVEER, R.J. (1977): Perception of wholes and of their components parts: some configural superiority effects; *Journal of experimental psychology, Human perception and preformance*, 3, 422-435.
- SPENCE,I and WAINER, H (2005): Introduction to the reprint of *The Commercial and Polical Atlas and statistical breviary*, by Playfair, W. 1786, Cambridge University Press, 1-35.
- STEED, C.A., FITZPATRICK, P.J., JANKUN-KELLY, T.J., YANCEY, A.N., SWANN, J.E. (2009): An interactive parallel coordinates technique applied to a tropical cyclone climate analysis, *Computer and Geosciences*, 35, 1529-1539.
- TUFTE, E. (1983): *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut.
- TUKEY, J.W. (1977): *Exploratory data analysis*, Addison Wesley Publishing Company.
- TUKEY, J.W. (1990): Data-Based Graphics: Visual Display in the decades to come, *Statistical Science*, 5, 3, 327-339.
- VALOIS, J.-P. (2000): Approche graphique en analyse des donnees, *Journal Soc. Fr. de Statistique*, 141,4, 6-40. (english text on request to: valois.jp@orange.fr).
- VALOIS, J.-P., BLONDEAU, Ch., DOSSOU-GBETE, S., BORDES, L. (2007): Smart Alarming Methods: an overview, highlight on statistical methods, *European Workshop on Data Stream Analysis*, Caserta, Italie, 14-16 mars, 2007.
- WEGMAN, E.J., LUO, Q. (2002): On Methods of Computer Graphics for Vizualizing Densities, *Computational and Graphical Statistics*, 11, 1, 137-162.
- WILKINSON, L. and FRIENDLY, M. (2009): The History of the Cluster Heat Map, *The American statistician*, (2009), 63(2), 179-184.
- ZEKI, S. (1992): The visual image in mind and brain, *Sci. American*, 266, 8, 43-50.

# Selecting Variables in Two-Group Robust Linear Discriminant Analysis

Stefan Van Aelst and Gert Willems

Department of Applied Mathematics and Computer Science, Ghent University  
Krijgslaan 281 S9, B-9000 Gent, Belgium.  
*Stefan.VanAelst@UGent.be, Gert.Willems@UGent.be*

**Abstract.** We consider two-group robust linear discriminant rules that are obtained by replacing the empirical means and covariance in the classical discriminant rules by S or MM-estimates of location and scatter. We consider the problem of selecting the variables that are relevant for separating the two groups. We propose to use a fast and robust bootstrap method to test which variables contribute significantly to the canonical variate, and thus the discrimination of the classes. This is useful since classical bootstrap methods may be unstable as well as extremely time-consuming when robust estimates are involved. Based on this test, the least relevant variables can then be removed from the model in e.g. a stepwise manner.

**Keywords:** bootstrap, linear discriminant analysis, robustness, variable selection

## 1 Robust linear discriminant analysis

We consider two  $p$ -dimensional populations,  $\Pi_1$  and  $\Pi_2$ , that have different means  $\boldsymbol{\mu}_1$  and  $\boldsymbol{\mu}_2$  but share a common covariance matrix  $\Sigma$ . Moreover, we also assume equal prior probabilities. The Bayes rule then classifies an observation  $\mathbf{x} \in \mathbb{R}^p$  into population  $\Pi_1$  if  $d_1^L(\mathbf{x}) > d_2^L(\mathbf{x})$ , where

$$d_j^L(\mathbf{x}) = \boldsymbol{\mu}_j^t \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_j^t \Sigma^{-1} \boldsymbol{\mu}_j; \quad j = 1, 2,$$

and into population  $\Pi_2$  otherwise. The direction  $\mathbf{a}$  that best separates the two populations is given by  $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \Sigma^{-1}$ . The corresponding projection  $\mathbf{a}^t \mathbf{x}$  is also called the canonical variate or discriminant coordinate. Since  $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$  and  $\Sigma$  are unknown, they need to be estimated from an available sample of the form  $\mathcal{Z}_n = \{\mathbf{x}_{11}, \dots, \mathbf{x}_{1n_1}, \mathbf{x}_{21}, \dots, \mathbf{x}_{2n_2}\} \subset \mathbb{R}^p$ . Robust linear discriminant analysis methods can be obtained by using robust estimates of the two centers and common scatter matrix (see e.g. Croux et al. (2008) and Bianco et al. (2008) and references therein). Here, we use the highly robust two-group S-estimators (He and Fung (2000)) and corresponding MM-estimators.

Consider a loss function  $\rho_0 : [0, \infty[ \rightarrow [0, \infty[$  which is bounded, increasing and sufficiently smooth, then the two-group S-estimates of the locations and

common scatter matrix are defined as the solution  $\tilde{\boldsymbol{\mu}}_{1n}$ ,  $\tilde{\boldsymbol{\mu}}_{2n}$  and  $\tilde{\Sigma}_n$  that minimizes  $|C|$  subject to

$$\frac{1}{n_1 + n_2} \sum_{j=1}^2 \sum_{i=1}^{n_j} \rho \left( [(\mathbf{x}_{ji} - T_j)^t C^{-1} (\mathbf{x}_{ji} - T_j)]^{\frac{1}{2}} \right) = b$$

among all  $T_1, T_2 \in \mathbb{R}^p$  and  $C \in \text{PDS}(p)$ . Here,  $\text{PDS}(p)$  denotes the set of positive definite symmetric matrices of size  $p$  and  $|C|$  denotes the determinant of the square matrix  $C$ .

In this paper, the loss function  $\rho_0$  is taken from the common class of Tukey biweight functions, given by  $\rho_c(t) = \min(t^2/2 - t^4/(2c^2) + t^6/(6c^4), c^2/6)$ . The constant  $b$  is usually chosen to ensure consistency of the S-estimates at the normal model. The constant  $c$  in the Tukey biweight loss function  $\rho_c$  can be tuned to achieve the desired degree of robustness, but at the same time it affects the efficiency of the S-estimators. Therefore, highly robust S-estimators can have quite low efficiency (see e.g. Salibián-Barrera et al. (2006)). To remedy this, MM-estimators (Tatsuoka and Tyler (2000)) apply an M-step starting from the highly robust S-estimator.

Let  $\tilde{\Sigma}_n$  be the S-estimate of scatter and denote  $\hat{\sigma}_n := |\tilde{\Sigma}_n|^{1/(2p)}$  the corresponding S-estimate of multivariate scale. Let  $\rho_1$  be a loss function from the same class as  $\rho_0$ . Then, the multivariate MM-estimates of the two locations and common shape  $(\hat{\boldsymbol{\mu}}_{1n}, \hat{\boldsymbol{\mu}}_{2n}, \hat{\Gamma}_n)$  minimize

$$\frac{1}{n_1 + n_2} \sum_{j=1}^2 \sum_{i=1}^{n_j} \rho_1 \left( [(\mathbf{x}_{ji} - T_j)^t G^{-1} (\mathbf{x}_{ji} - T_j)]^{\frac{1}{2}} / \hat{\sigma}_n \right)$$

among all  $(T, G) \in \mathbb{R}^p \times \text{PDS}(p)$  for which  $|G|=1$ . The corresponding MM-estimator for the common scatter matrix is given by  $\hat{\Sigma}_n = \hat{\sigma}_n^2 \hat{\Gamma}_n$ . It is well-known that the MM-estimates inherit the robustness (breakdown point) of the initial S-estimate of multivariate scale (determined by the loss function  $\rho_0$ ). Hence, the loss function  $\rho_1$  can be tuned to guarantee a high efficiency, e.g. 95%, at the normal model.

## 2 Bootstrapping S and MM-estimators

Bootstrap is an attractive nonparametric method to obtain inference for estimators. However, applying the standard bootstrap on robust estimators poses both a computational issue and a robustness issue. Calculating robust estimates is complex and requires a high computation time, which makes it infeasible to obtain a large number of recalculated robust estimates in a reasonable amount of time. Because bootstrap samples are drawn with replacement, the amount of outliers varies between bootstrap samples and can exceed the breakdown point of the estimator in some samples. For multivariate S and MM-estimators, both issues can be solved at once by the fast and

robust bootstrap (FRB) as studied in Salibian-Barrera et al. (2006). Here, we use the FRB to obtain many recalculations of the robust S or MM-estimates of the two locations and common scatter matrix in a linear discriminant analysis. These FRB estimates can then be used to estimate the sampling distribution of the canonical variate coefficients  $\mathbf{a}$ . Based on this distribution, we can investigate which variables contribute significantly to the canonical variate.

Both S and MM-estimates can be written as the solution of a set of smooth fixed point equations as follows. Let  $\hat{\boldsymbol{\theta}}_n$  be a vector that collects all parameter estimates of interest. In our case, for S-estimates  $\hat{\boldsymbol{\theta}}_n$  contains the two location estimates and the common scatter estimate in vectorized form. In case of MM-estimates, the vector  $\hat{\boldsymbol{\theta}}_n$  additionally contains the MM location and vectorized shape estimates. Then, the estimates can be written as the solution of

$$\hat{\boldsymbol{\theta}}_n = \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n) \quad (1)$$

where the function  $g_n : \mathbb{R}^d \rightarrow \mathbb{R}^d$  depends on the sample  $\mathcal{Z}_n$ . Given a bootstrap sample  $\mathcal{Z}_n^*$  (i.e. a sample of size  $n_1 + n_2$  drawn with replacement from  $\mathcal{Z}_n$ ), the recalculated estimate  $\hat{\boldsymbol{\theta}}_n^*$  then is the solution of the corresponding fixed point equation  $\hat{\boldsymbol{\theta}}_n^* = \mathbf{g}_n^*(\hat{\boldsymbol{\theta}}_n^*)$ , where the function  $\mathbf{g}_n^*$  now depends on  $\mathcal{Z}_n^*$ . Instead of calculating  $\hat{\boldsymbol{\theta}}_n^*$ , we consider the simple approximation  $\hat{\boldsymbol{\theta}}_n^{1*} := \mathbf{g}_n^*(\hat{\boldsymbol{\theta}}_n)$ . Hence,  $\hat{\boldsymbol{\theta}}_n^{1*}$  is a one-step approximation of  $\hat{\boldsymbol{\theta}}_n^*$  starting from the initial value  $\hat{\boldsymbol{\theta}}_n$ , obtained for the original sample.

Note that since we are keeping the estimates  $\hat{\boldsymbol{\theta}}_n$  fixed when calculating the approximations  $\hat{\boldsymbol{\theta}}_n^{1*}$ , these approximations will likely underestimate the variability of the estimator. To remedy this, a linear correction can be applied as follows. Consider a Taylor expansion about  $\hat{\boldsymbol{\theta}}_n$ 's limiting value  $\boldsymbol{\theta}$ ,

$$\hat{\boldsymbol{\theta}}_n = \mathbf{g}_n(\boldsymbol{\theta}) + \nabla \mathbf{g}_n(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + R_n,$$

where  $R_n$  is the remainder term and  $\nabla \mathbf{g}_n(\cdot) \in \mathbb{R}^{m \times m}$  is the matrix of partial derivatives. Assuming that the remainder term is negligible, this can be rewritten as

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \sim [\mathbf{I} - \nabla \mathbf{g}_n(\boldsymbol{\theta})]^{-1} \sqrt{n}(\mathbf{g}_n(\boldsymbol{\theta}) - \boldsymbol{\theta}),$$

where  $\sim$  denotes that both sides have the same limiting distribution. Under certain conditions we will have that  $\sqrt{n}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}) \sim \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$  and  $\sqrt{n}(\mathbf{g}_n^*(\boldsymbol{\theta}) - \boldsymbol{\theta}) \sim \sqrt{n}(\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n) - \hat{\boldsymbol{\theta}}_n)$ . If we furthermore approximate  $[\mathbf{I} - \nabla \mathbf{g}_n(\boldsymbol{\theta})]^{-1}$  by  $[\mathbf{I} - \nabla \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)]^{-1}$  we obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) \sim [\mathbf{I} - \nabla \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)]^{-1} \sqrt{n}(\mathbf{g}_n^*(\hat{\boldsymbol{\theta}}_n) - \hat{\boldsymbol{\theta}}_n).$$

We now define the fast and robust bootstrap estimates as

$$\hat{\boldsymbol{\theta}}_n^{R*} := \hat{\boldsymbol{\theta}}_n + [\mathbf{I} - \nabla \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)]^{-1}(\hat{\boldsymbol{\theta}}_n^{1*} - \hat{\boldsymbol{\theta}}_n).$$

It can be shown that the distribution of these fast and robust bootstrap estimates  $\hat{\boldsymbol{\theta}}_n^{R*}$  is consistent in the sense that it estimates the same limiting distribution as the sampling distribution of  $\hat{\boldsymbol{\theta}}_n^*$  does (see Salibian-Barrera et al. (2006), Theorem 2). Moreover, the FRB estimates  $\hat{\boldsymbol{\theta}}_n^{R*}$  are easy to calculate for every bootstrap sample and they inherit the robustness of the solution  $\hat{\boldsymbol{\theta}}_n$  for the original sample. Indeed, if an observation is downweighted in the original sample, then this observation receives the same low weight when calculating the FRB estimate of a bootstrap sample, no matter how many outliers occur in the bootstrap sample.

It has been shown by Salibian-Barrera et al. (2006) that the FRB commutes with smooth functions. In our setting this implies for instance that the sampling distribution of the coefficients  $\mathbf{a} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\Sigma^{-1}$  of the canonical variate is estimated consistently by the distribution of  $\mathbf{a}^{R*} = (\boldsymbol{\mu}_1^{R*} - \boldsymbol{\mu}_2^{R*})(\Sigma^{R*})^{-1}$ . Based on the FRB distribution of these coefficients we can thus examine which variables contribute significantly to the discrimination of the two groups.

### 3 FRB test

To investigate how well the FRB can estimate the sampling variability of the canonical variate and thus can be used to select the variables that are relevant for discriminating the two groups, we ran a small simulation study. We considered two equally sized samples with  $p = 4$  variables where the sample sizes were 25, 50 and 100. Both samples were drawn from a multivariate normal distribution with identity covariance matrix. The center of the first group was  $\boldsymbol{\mu}_1 = (-1, 1, 0, 0)^t$  while the center of the second group was  $\boldsymbol{\mu}_2 = (1, -1, 0, 0)^t$ . The coefficients of the (normalized) canonical variate at the population level then equal  $(-1/\sqrt{2}, 1/\sqrt{2}, 0, 0)^t$ . Hence, the first two components carry discriminatory power, while the last two variables do not aid in separating the two populations. To examine the robustness of the FRB, we also consider contaminated data sets with 20% of outliers in the second group. The outliers were obtained by shifting the center to  $\boldsymbol{\mu}_{\text{out}} = (-3, 3, -3, 3)^t$ . The simulation results shown in Table 1 are based on  $M = 500$  randomly generated samples for each setting.

For each of the four variables, we test whether the corresponding coefficient  $a_j$  ( $j = 1, \dots, 4$ ) in the canonical variate equals zero or not. Based on the distribution of the FRB estimates  $\mathbf{a}^{R*}$ , we can easily obtain an estimate of the corresponding p-value for each of the four variables. Table 1 shows for each of the four variables the average of the FRB estimated p-values based on  $B = 999$  bootstrap samples. From the top panel of Table 1, we see that for clean data the FRB estimated p-values are small for the first two variables and large for the last two variables, as expected. Comparing the results for clean data (Eps=0%) with the results for contaminated data (Eps=20%)



in Table 1, we see that the accuracy of the FRB estimates p-values is not much affected by the outliers, which illustrates the robustness of the FRB test procedure.

Eps	Var	25	50	100
0%	1	0.005 (0.017)	0.000 (0.001)	0.000 (0.001)
	2	0.004 (0.019)	0.000 (0.000)	0.000 (0.000)
	3	0.495 (0.283)	0.494 (0.280)	0.481 (0.280)
	4	0.518 (0.290)	0.512 (0.310)	0.491 (0.310)
20%	1	0.038 (0.144)	0.007 (0.071)	0.002 (0.071)
	2	0.033 (0.125)	0.007 (0.070)	0.001 (0.070)
	3	0.523 (0.288)	0.507 (0.292)	0.484 (0.292)
	4	0.494 (0.284)	0.494 (0.276)	0.489 (0.276)

**Table 1.** Average FRB estimates (with standard deviations) of the p-value for testing whether the coefficient of each variable in the canonical variate equals zero or not. Results are shown for both clean data (Eps=0%) and contaminated data (Eps=20%) based on samples of size 25, 50 and 100.

To investigate further the accuracy of the distribution of the test statistic as estimated by FRB, we show for each of the four variables the percentage of times that the test rejects the null hypothesis if the test is performed respectively with 5% nominal significance level (Table 2) and with 1% nominal significance level (Table 3). Since the null hypothesis does not hold at the population level for the first two variables, the corresponding results provide us information about the power of the test. For the last two variables the null hypothesis holds at the population level, so for these variables we obtain information on the level of the test. The results for contaminated data provide us information on the robustness of the power and level, respectively.

From the results for variables 3 and 4 in Tables 2 and 3 we can see that for clean data the FRB based tests maintain the nominal level quite well and, as expected, the accuracy increases with increasing sample size. The corresponding results for variables 1 and 2 show that in this setting the test also has high power and thus succeeds well in identifying the variables with discriminatory power. The results for contaminated data show that the contamination affects the level of the test (variables 3 and 4), with empirical levels that become less accurate, as well as the power (variables 1 and 2) of the test. The loss of power is quite large for small samples ( $n_1 = n_2 = 25$ ), but the difference in power decreases quickly if the sample size grows. Both the loss of power and less precise observed level of the test under contamination can at least partly be explained by the fact that the second sample contains much less useful information (uncontaminated observations) which leads to an increased imprecision of the robust estimates of its group center and the common scatter matrix, compared to clean data.

Eps	Var	25	50	100
0%	1	0.980	1.000	1.000
	2	0.988	1.000	1.000
	3	0.048	0.050	0.046
	4	0.052	0.084	0.054
20%	1	0.894	0.980	0.988
	2	0.906	0.980	0.988
	3	0.042	0.052	0.066
	4	0.044	0.038	0.038

**Table 2.** Percentage of samples for which the test that the coefficient of each variable in the canonical variate equals zero or not, rejects the null hypothesis at the 5% nominal significance level. Results are shown for both clean data (Eps=0%) and contaminated data (Eps=20%) based on samples of size 25, 50 and 100.

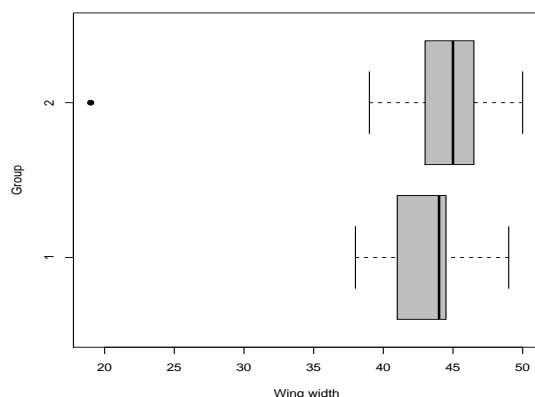
Eps	Var	25	50	100
0%	1	0.898	0.998	1.000
	2	0.908	1.000	1.000
	3	0.008	0.014	0.008
	4	0.018	0.020	0.010
20%	1	0.770	0.960	0.986
	2	0.798	0.964	0.986
	3	0.012	0.010	0.012
	4	0.004	0.012	0.014

**Table 3.** Percentage of samples for which the test that the coefficient of each variable in the canonical variate equals zero or not, rejects the null hypothesis at the 1% nominal significance level. Results are shown for both clean data (Eps=0%) and contaminated data (Eps=20%) based on samples of size 25, 50 and 100.

## 4 Example

The FRB test for significance of the canonical variate coefficients, can be used for instance in variable selection procedures for two-group linear discriminant analysis. To illustrate variable selection based on the FRB test, we consider the Biting Flies data taken from Johnson and Wichern (2002). The data set consists of two groups of 35 flies (*Leptoconops torrens* and *Leptoconops carteri*) and we consider the measurements `wing length`, `wing width`, `third palp length`, `third palp width`, and `fourth palp length`. The variable `wing width` contains a clear outlier in the second group as can be seen in Figure 1. Hence, a robust discriminant analysis is advisable to reduce the possible effect of outliers.

Figure 2 shows the FRB distribution of each of the robust estimates of the canonical variate coefficients. The vertical line in these histograms is drawn at the null hypothesis that the coefficient equals zero. This plot already suggests



**Fig. 1.** Boxplot of wing width for both groups of biting flies.

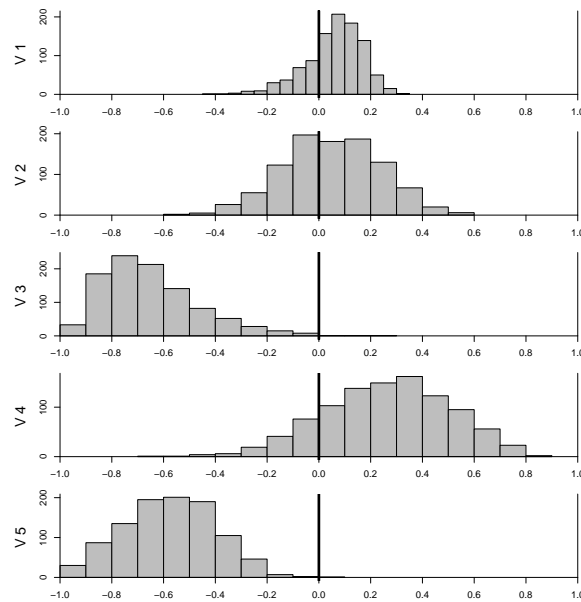
that variables 3 and 5 contain the most discriminatory power, but the other variables are less relevant. Using backward elimination, each time removing the least significant variable if its p-value as estimated by FRB is smaller than 5%, we get the series of models shown in Table 4. From this table, we see that the final model (last line) indeed only contains variables 3 and 5. In this example a classical analysis yields the same final model, but due to the outlier in variable 2, variable 1 is removed before variable 2.

Model	Variable				
	1	2	3	4	5
1	0.490	0.817	0.006	0.296	0.002
2	0.306	-	0.016	0.216	0.000
3	-	-	0.016	0.096	0.000
4	-	-	0.006	-	0.000

**Table 4.** Estimated p-values for testing, based on FRB, whether each canonical variate coefficient equals zero. In each step the least significant variable is removed if its p-value exceeds 0.05.

## 5 Conclusion

We showed that the FRB can be used to test whether variables contribute significantly to a robust two-group discriminant analysis. We illustrated that this test can be used in variable selection procedures for such a discriminant analysis. In future research we will investigate whether a similar procedure



**Fig. 2.** FRB distribution of each of the robust estimates of the standardized coefficients of the canonical variate in the Biting Flies data.

can be used for variable selection in multi-group robust discriminant analysis, by using a robust version of Wilks Lambda as in Todorov (2007).

## References

- BIANCO, A., BOENTE, G., PIRES, A. M. and RODRIGUES, I. M. (2008): Robust discrimination under a hierarchy on the scatter matrices. *Journal of Multivariate Analysis* 99, 1332–1357.
- CROUX, C., FILZMOSER, P. and JOOSSENS, K. (2008): Classification efficiencies for robust linear discriminant analysis. *Statistica Sinica* 18, 581–599.
- HE, X. and FUNG, W. K. (2000): High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis* 72, 151–162.
- JOHNSON, R. A. and WICHERN, D. W. (2002): *Applied Multivariate Statistical Analysis*. Prentice Hall, Upper Saddle River.
- SALIBIAN-BARRERA, M., VAN AELST, S. and WILLEMS, G. (2006): PCA based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association* 101, 1198–1211.
- TATSUOKA, K. S. and TYLER, D. E. (2000): The uniqueness of S and M-functionals under non-elliptical distributions. *Annals of Statistics* 28, 1219–1243.
- TODOROV, V. (2007): Robust selection of variables in linear discriminant analysis. *Statistical Methods and Applications* 15, 395–407.

# How to Take into Account the Discrete Parameters in the BIC Criterion?

Vincent Vandewalle

Département STID, IUT C Roubaix, Université Lille 2  
25-27, rue du Maréchal Foch, 59100 Roubaix, France,  
*vincent.vandewalle@univ-lille2.fr*

**Abstract.** When using the BIC criterion to select one model among several models, only the continuous parameters are taken into account in the penalization. However, when considering models with discrete parameters to be estimated, this criterion can lead to select too simple models, not taking into account the possible over-fitting caused by the estimation of the discrete parameters. Ideally we would like to integrate the likelihood on every possible value of the discrete parameter. In this article we study how this integral can be approximated from a practical and theoretical point of view in the particular case of the parsimonious multinomial distribution.

**Keywords:** model selection, discrete parameters, parsimonious models, Bayesian integrated criterion

## 1 Introduction

In many models some discrete parameters must be estimated. It is for instance the case in the models proposed by Celeux and Govaert (1991) where the location of the modal modality must be estimated for each variable in each class. From a Bayesian perspective it is natural to integrate the likelihood on the whole parameter space, *i.e.* to integrate over both continuous and discrete parameters. However the BIC approximation relies on Taylor expansions which require the parameter space to be continuous. A natural heuristic here is to consider the standard BIC criterion without taking into account the discrete parameters in the BIC approximation. However this approach would neglect over-fitting of the likelihood due to the estimation of the discrete parameters. Consequently, this can lead to favor the model with discrete parameters when considering the model choice issue. We study this problem on a toy model (see Section 2), for which numerical integration can be done easily.

The outline of this article is the following. In Section 2 we will define the modal modality model. In Section 3 we will focus on possible approximations of the integrated likelihood when considering this model. In Section 4 we will compare these different approximations on simulated and real data. In section 5 we will conclude and discuss the perspectives of this work.

## 2 Modal modality model

Here we will take the example of the parsimonious models of Celeux and Govaert (1991) in the case of products of binomial distributions, latter extended by Biernacki et al. (2006) in the case of products of multinomial distributions. For sake of simplicity, we will suppose that a random variable  $\mathbf{X}$  comes from the multinomial distribution  $\mathcal{M}(1, \alpha_1, \dots, \alpha_m)$  ( $\sum_{h=1}^m \alpha_h = 1$ ,  $\alpha_h > 0$ ), and that an  $n$  independent and identically distributed sample  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  coming from  $\mathbf{X}$  is observed. We can impose constraints on  $(\alpha_1, \alpha_2, \dots, \alpha_m)$  in order to get parsimonious models, as for instance  $\alpha_1 = \alpha_2 = \dots = \alpha_m$ . The constraint proposed by Biernacki et al. (2006) is:

$$\alpha_h = \begin{cases} 1 - \varepsilon & \text{if } h = h^* \\ \frac{\varepsilon}{m-1} & \text{otherwise,} \end{cases}$$

where  $h^*$  is the location of the modal modality. For  $h^*$  to be the modal modality, we need to have  $\varepsilon \leq \frac{m-1}{m}$ . In this setting, two parameters must be estimated,  $\varepsilon$  which is a continuous parameter ( $\varepsilon \in [0, \frac{m-1}{m}]$ ), and  $h^*$  which is a discrete parameter ( $h^* \in \{1, 2, \dots, m\}$ ).

## 3 Integrated likelihood and BIC approximations

### 3.1 Integrated likelihood

Supposing that  $\varepsilon$  and  $h^*$  are *a priori* independent and that every value of the discrete parameter has the same probability, the prior distribution of  $(\varepsilon, h^*)$  is the following

$$p(\varepsilon, h^*) = \frac{1}{m} p(\varepsilon).$$

So that when integrating the likelihood we have

$$p(\mathbf{x}) = \frac{1}{m} \sum_{h^*=1}^m \int_0^{\frac{m-1}{m}} p(\mathbf{x}|\varepsilon, h^*) p(\varepsilon) d\varepsilon.$$

For instance, we can choose a truncated Dirichlet prior for  $p(\varepsilon)$

$$p(\varepsilon) = C \varepsilon^{-\frac{1}{2}} (1 - \varepsilon)^{-\frac{1}{2}} \mathbf{1}_{[0, \frac{m-1}{m}]}(\varepsilon),$$

where  $C$  is some normalization constant.

Noting  $n_h = \sum_{i=1}^n x_{ih}$ , the logarithm of the integrated likelihood (IL) is

$$\text{IL} = \log \left( \frac{1}{m} \sum_{h=1}^m \int_0^{\frac{m-1}{m}} (1 - \varepsilon)^{n_h} \left( \frac{\varepsilon}{m-1} \right)^{n-n_h} C \varepsilon^{-\frac{1}{2}} (1 - \varepsilon)^{-\frac{1}{2}} d\varepsilon \right) \quad (1)$$

This sum is rather easy to compute in practice since it requires the use of incomplete beta function which is implemented efficiently in many softwares (for numerical experiments the R software have been used). This value will be compared to the hereafter approximations.

### 3.2 Standard BIC approximation

In the standard BIC approximation we just take the maximum likelihood estimator of the parameters

$$(\hat{\varepsilon}, \widehat{h^*}) = \arg \max_{\varepsilon, h} (1 - \varepsilon)^{n_h} \left( \frac{\varepsilon}{m - 1} \right)^{n - n_h},$$

this maximization is easily performed by taking  $\widehat{h^*} = \arg \max_h n_h$ , then  $\hat{\varepsilon} = 1 - \frac{n_{\widehat{h^*}}}{n}$ . When the discrete parameters are not taken into account, the BIC criterion is:

$$\text{BIC}_1 = \log \left( (1 - \hat{\varepsilon})^{n_{\widehat{h^*}}} \left( \frac{\hat{\varepsilon}}{m - 1} \right)^{n - n_{\widehat{h^*}}} \right) - \frac{1}{2} \log n,$$

this one will be noted  $\text{BIC}_1$ . However this approximation is not justified when considering discrete parameters.

### 3.3 BIC approximation with discrete parameters

We will now consider the BIC approximation when discrete parameters are taken into account. But in a first time we will need to consider the case where the maximum of the likelihood will be reached on the constraint, since this will be the case for at least one term in the sum equation (1). We will first use the following proposition (Vandewalle (2009)).

**Proposition 21.** *Let  $L : [a, b] \mapsto \mathbb{R}$ , such that  $L$  be one time differentiable on  $[a, b]$  and that it reaches its maximum at  $b$  with  $L'(b) > 0$ . Then*

$$\log \left( \int_a^b e^{nL(u)} du \right) = nL(b) - \log n + O(1).$$

For a comparison note that

$$\log \left( \int_a^b e^{nL(u)} du \right) = nL(c) - \frac{1}{2} \log n + O(1),$$

if  $L$  would reach its maximum for  $c \in ]a, b[$ .

This result stays valid for functions  $L$  which depends on  $n$  under some regularity conditions (Lebarbier and Mary-Huard (2006)). We see that the penalty here is  $-\log n$  where it would be  $-\frac{1}{2} \log n$  in the standard BIC approximation. So we can interpret the constraint saturation as the addition of one parameter. This can seem quite natural, since we expect to select models for which constraints are not saturated when estimating the parameters.

Applying Proposition 21 we get

$$\log \left( \int_0^{\frac{m-1}{m}} (1-\varepsilon)^{n_h} \left( \frac{\varepsilon}{m-1} \right)^{n-n_h} C \varepsilon^{-\frac{1}{2}} (1-\varepsilon)^{-\frac{1}{2}} d\varepsilon \right) = \log p(\mathbf{x}|\hat{\varepsilon}, h) - \frac{1+s_h}{2} \log n + O(1)$$

where  $s_h = 1$  if the constraint is saturated (*i.e.*  $\hat{\varepsilon} = \frac{m-1}{m}$ ) and 0 otherwise.

Then replacing these approximations in equation (1) we get

$$\text{BIC}_2 = \log \left( \frac{1}{m} \sum_{h=1}^m (1-\hat{\varepsilon}_h)^{n_h} \left( \frac{\hat{\varepsilon}_h}{m-1} \right)^{n-n_h} n^{-\frac{1+s_h}{2}} \right)$$

where  $\hat{\varepsilon}_h$  is the maximum likelihood estimator of  $\varepsilon$  when  $h$  is constrained to be the modal modality. So that when it is numerically feasible we recommend to use the above expression, since it avoids numerical integration on the continuous parameter while keeping the sum in equation (1) on every possible state of the discrete variable.

We can also simplify  $\text{BIC}_2$  to avoid the integration on the states of the discrete variable, which gives the alternative criterion

$$\text{BIC}_3 = \log \left( (1-\hat{\varepsilon}_{h^*})^{n_{h^*}} \left( \frac{\hat{\varepsilon}_{h^*}}{m-1} \right)^{n-n_{h^*}} \right) - \frac{1}{2} \log n - \log m.$$

Here, we have the standard BIC criterion penalized by the logarithm of the number of possible states of the discrete variable. If one term dominates in the sum, the over-penalization is justified, which is the case for  $n$  large enough. Otherwise, if all the terms in the sum have about the same value, we should not penalize by  $\log m$ .

## 4 Numerical experiments

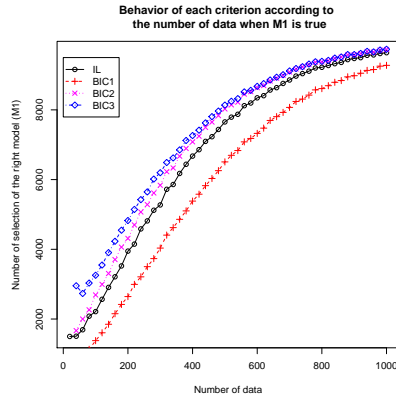
### 4.1 Study in the simple case of the modal modality model

In order to compare the various criteria we consider two models :  $M_1$  the full model,  $M_2$  the parsimonious model with modal modality.

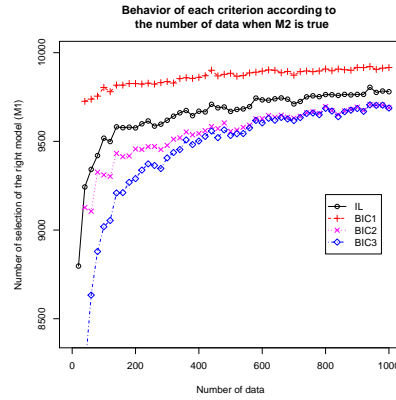
Let  $\mathbf{X} \sim \mathcal{M}(1, 0.40, 0.35, 0.25)$ .  $M_1$  and  $M_2$  are put into competition, we are interested in the number of times where the true model, here  $M_1$ , is selected by the criteria defined above on 10,000 replicates and for  $n$  from 20 to 1,000. Results are illustrated Figure 1. We see that the nearest criterion of IL is the  $\text{BIC}_2$  criterion. The usual BIC criterion ( $\text{BIC}_1$ ) seems to under-penalized the parsimonious model whereas the over-penalized BIC criterion ( $\text{BIC}_3$ ) over-penalizes it. This over-penalization is all the more large as  $n$  is small.



Let now  $\mathbf{X} \sim \mathcal{M}(1, 0.40, 0.30, 0.30)$ .  $M_1$  and  $M_2$  are put into competition, we are interested in the number of times where the true model, here  $M_2$ , is selected by the criteria defined above on 10,000 replicates and for  $n$  from 20 to 1,000. Results are illustrated Figure 2. We see that the nearest criterion of IL is the  $\text{BIC}_2$  criterion. The usual BIC criterion ( $\text{BIC}_1$ ) selects more often the model  $M_2$  since this one is under-penalized. The over-penalized BIC criterion ( $\text{BIC}_3$ ) over-penalizes the model  $M_2$ . This over-penalization is all the more large as  $n$  is small.



**Fig. 1.** Number of times where the full model is selected by the various criteria given  $n$ .



**Fig. 2.** Number of times where the parsimonious model is selected by various criteria given  $n$ .

So the best approximation of the integrated likelihood is the  $\text{BIC}_2$  criterion that we recommend when numerical integration is not possible. Here the number of possible states of the discrete variable is rather small, whereas in practice this number states can be quite large. In practice, when the number of possible states of the discrete parameter is large it could be desirable to approach the sum by Monte-Carlo, the main issue is to choose a good proposal. This question is discussed below.

## 4.2 Monte-Carlo study for the binary case

Now we consider the case where we have binary data in the multivariate case (in dimension  $d$ ), so that a data  $\mathbf{x}_i$  ( $i \in \{1, \dots, n\}$ ) is written  $\mathbf{x}_i = (\mathbf{x}_i^1, \mathbf{x}_i^2, \dots, \mathbf{x}_i^d)$  and  $\mathbf{x}_i^j$  is supposed to be drawn from a Bernoulli distribution. In this case Celeux and Govaert (1991) have proposed to impose the equality of  $\varepsilon$  for each variable. Then the vector of modal positions, noted  $\mathbf{h}$ , can take  $2^d$  different values. As a consequence, when  $d$  is large it is not

possible to perform the integration over all these states. Here, we propose to approach this sum by Monte-Carlo using importance sampling. In order to get a good proposal, the full product of independent Bernoulli model is learned, then the position of the modal modality is drawn from this learned model. Then given  $\mathbf{h}$ ,  $\varepsilon$  is easily computed as seen in Section 3.1.

On this experiment we only want to compare the different approximations of the integrated likelihood without considering the model choice issue. Let consider the cases  $d = 5$ ,  $d = 10$  and  $d = 20$  variables, and  $\varepsilon = 0.45$  for each variable. We have generated 100 datasets, and for each dataset we performed the importance sampling generating 10,000 modal positions. Results are presented in Table 1.

Criterion \ $n$	20	50	100	1000
$d = 5$ dimensions				
IL	-70.91 (0.9)	-174.96 (1.2)	-347.77 (1.7)	-3448.18 (7.2)
BIC <sub>1</sub>	-68.77 (1.4)	-172.59 (1.7)	-345.03 (2.2)	-3444.38 (7.2)
BIC <sub>2</sub>	-70.50 (0.8)	-174.59 (1.2)	-347.41 (1.6)	-3447.84 (7.2)
BIC <sub>3</sub>	-72.23 (1.4)	-176.05 (1.7)	-348.49 (2.2)	-3447.85 (7.2)
$d = 10$ dimensions				
IL	-140.24 (1.0)	-348.15 (1.2)	-693.71 (2.4)	-6891.66 (10)
BIC <sub>1</sub>	-135.98 (2.1)	-343.32 (2.1)	-688.22 (3.3)	-6884.02 (10)
BIC <sub>2</sub>	-139.49 (1.0)	-347.44 (1.2)	-693.01 (2.3)	-6890.97 (10)
BIC <sub>3</sub>	-142.91 (2.1)	-350.25 (2.1)	-695.15 (3.3)	-6890.95 (10)
$d = 20$ dimensions				
IL	-279.01 (0.8)	-694.51 (1.4)	-1385.87 (2.4)	-13795.88 (14)
BIC <sub>1</sub>	-271.06 (2.6)	-685.31 (3.2)	-1374.98 (3.5)	-13765.95 (11)
BIC <sub>2</sub>	-277.93 (0.8)	-693.46 (1.4)	-1384.84 (2.4)	-13794.85 (14)
BIC <sub>3</sub>	-284.93 (2.6)	-699.18 (3.2)	-1388.85 (3.5)	-13779.81 (11)

**Table 1.** Mean value of the criterion according the values of  $n$  and  $d$ , the standard deviation is given into parenthesis.

We see that the proposed approach performs well, it is all the more important when the number of data is small. We see that the standard BIC approximation (BIC<sub>1</sub>) over-estimates the integrated likelihood favoring the parsimonious model. So that when it is possible, numerical integration should be performed on the discrete parameters.

We now consider binary data from the UCI database repository and the Statlog database. The parsimonious product of binary distributions model is used, and the same strategy as above is used to perform numerical integration. As in the previous study we compare the different approximations of the integrated likelihood without considering the model choice issue. If the initial data are continuous they are discretized using the Fisher algorithm (Fisher (1958)). Results are presented in Table 4.2.

Dataset	$n$	$d$	IL	BIC <sub>1</sub>	BIC <sub>2</sub>	BIC <sub>3</sub>
SPECT Heart (Test)	187	23	-2759.1	-2742.5	-2758.0	-2758.5
SPECT Heart (Train)	80	23	-1015.5	-999.0	-1014.5	-1014.9
Acute Inflammations	120	7	-572.7	-568.1	-572.2	-572.9
Abalone	34	7	-164.1	-159.6	-163.6	-164.4
Breast Cancer Diagnostic	569	30	-9978.9	-9958.5	-9977.6	-9979.3
Crab	200	5	-695.9	-693.6	-695.5	-697.1
Cushings	27	2	-23.7	-22.5	-23.8	-23.9
Fglass	214	9	-947.6	-940.7	-947.0	-946.9

**Table 2.** Comparison of the approximations of the log-likelihood value for binary data of the UCI and Statlog databases.

We see that on real data which do not come from the proposed model, criteria BIC<sub>2</sub> and BIC<sub>3</sub> are good approximations of the integrated likelihood (IL) whereas the criterion BIC<sub>1</sub> does not penalizes enough the discrete parameters.

## 5 Conclusion and perspectives

The presented work is in progress. We have asked the question of taking into account the discrete parameters into the BIC approximation. We have shown that in order to get a good approximation of the integrated likelihood the sum over the possible values of the discrete parameter should be performed. In practice when the number of possible states of the discrete parameter is too large the exact computation of this sum cannot be performed, and must be approximated by Monte-Carlo. For the modal modality model, we have shown that application to real issues is possible using Monte-Carlo integration. As a conclusion we recommend to use the BIC<sub>2</sub> approximation since it is less expensive than IL but produce similar results.

The proposed approach is very useful when putting into competition models with many discrete parameters since it allows to take into account the over-fitting of the likelihood due to the discrete parameters estimation. In future work we will use this approach on large scale issues. For instance, in the classification framework by putting into competition many parsimonious models like the models proposed by Bernacki et al. (2006) and compare the proposed strategy to the standard BIC criterion.

## References

- BIERNACKI, C., CELEUX, G., GOVAERT, G. and LANGROGNET, F. (2006): Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. *Computational Statistics and Data Analysis* 51 (2), 587-600.
- CELEUX, G. and GOVAERT, G. (1991): Clustering Criteria for Discrete Data and Latent Class Models. *Journal of Classification*, 8, 157-176.

- FISHER, W. D. (1958): On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53, 789-798.
- LEBARBIER, E. and MARY-HUARD, T. (2006): Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la SFdS*, 147 (1), 39-58.
- SCHWARTZ, G. (1978): Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- VANDEWALLE, V. (2009): Estimation et choix sélection en classification semi-supervisée. *Thèse de doctorat, Université Lille 1*, 156 pages.

# Analysis of Breath Alcohol Measurements Using Compartmental and Generalized Linear Models

Chi Ting Yang<sup>1</sup>, Wing Kam Fung<sup>2</sup>, and Thomas Wai Ming Tam<sup>3</sup>

<sup>1</sup> Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong. *yang@graduate.hku.hk*

<sup>2</sup> Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong. *wingfung@hku.hk*

<sup>3</sup> Forensic Science Division, Government Laboratory, 88 Chung Hau Street, Ho Man Tin, Kowloon, Hong Kong. *wmtam@govtlab.gov.hk*

**Abstract.** Pharmacokinetic parameters are important for clinical application. Traditionally, compartmental model together with non-compartmental approach are commonly adopted to deal with the analysis of pharmacokinetic data. In this study, we apply the alternative generalized linear model as proposed by Wakefield (2004) to the breath alcohol measurements of Chinese subjects in Hong Kong. Both compartmental and generalized linear models are fitted to each subject involved. Four core pharmacokinetic parameters including the time to maximum blood alcohol concentration (BAC) level, the BAC level attained at peak, the rate of clearance and elimination half-life are examined. The parameter estimates under the two models are then compared. Finally, we also extend the concept from the individual to the population level.

**Keywords:** blood alcohol concentration, one-compartment model, generalized linear model, nonlinear mixed model, generalized linear mixed model

## 1 Introduction

Pharmacokinetics is a vital topic in drug development which involves the course of absorption, distribution, metabolism, and elimination in human body (Gibaldi and Perrier, 1982; Kwon, 2001; Norberg et al., 2003). Most of the time, kinetics of the metabolite are learnt by dealing with its measured concentration level in blood over time after drug administration. The process is informative in the determination of pharmacokinetic parameters which are of interest to forensic scientists.

Traditionally, there are two different approaches, compartmental and non-compartmental, in the analysis of pharmacokinetic data collected from a wide variety of biological, biomedical and other aspects (Kwon, 2001; Ritschel and Kearns, 1999; Holford, 1987). The non-compartmental approach is commonly adopted for clinical application, in which the pharmacokinetic parameters are

estimated directly from the collected data. For compartmental method, different mathematical models and fitting techniques are developed to trace the biological process and, in the meantime, quantify the parameters of interest accordingly (Davidian and Giltinan, 1995; Lindsey, 2001; Piantadosi, 1997). The one-compartment model is the simplest one, and it assumes the entire body as a single kinetically homogeneous compartment (Gibaldi and Perrier, 1982). Multi-compartment models have also been developed. Researchers like Crowder and Tredger (1981), Crowder (1983) and Nelder (1991) suggested other alternatives which are in polynomial form to describe the profile of drug concentration level in blood. In fact, many non-linear models arose from the study of chemical kinetics also provide a natural setting for the analysis of pharmacokinetic data.

Compartmental model is a very useful device for describing drug distribution and excretion. The assumption on the variability of data following the normal or log-normal distribution, however, may not always be appropriate for pharmacokinetic data. Moreover, the mean function of such models is a sum of exponentials for which inference and computation may not be straightforward (Wakefield, 2004; Salway and Wakefield, 2008). Lindsey et al. (2000) compared different probability distributions for pharmacokinetic data and suggested the skewed gamma distribution for such kind of data analysis. Wakefield (2004), and Salway and Wakefield (2008) extended the concept to generalized linear model.

Analogous to other drug pharmacokinetics, compartmental models are often adopted in alcohol study to describe the biological process and, in the meantime, a similar concern encounters as described. This paper is based on the alcohol project on the Chinese population (Tam et al., 2005). We apply the alternative generalized linear model as proposed by Wakefield (2004) to the breath alcohol measures. This model is easier to handle and well known to statisticians. We are interested in investigating whether the alternative method can provide a good approximation to the one-compartment model. Both compartmental and generalized linear models are fitted to each subject involved and their corresponding pharmacokinetic parameters are examined then. Apart from the individual fitting, we also explore the flexibility of extending both approaches to the population level in order to capture variability among subjects.

## 2 Methods and Materials

### 2.1 Subjects and Conditions

Tam et al. (2005) collected a total of 184 healthy Chinese subjects for the alcohol breath analysis. Chemists from Forensic Division of Government Laboratory, Hong Kong Special Administrative Region, in accordance with their general practice of breath alcohol measurements, conducted the experiment. Ethical approval was obtained from the Human Research Ethics Committee

for Non-Clinical Faculties of the University of Hong Kong and participation of the subjects was on voluntary basis recruitment with consent given at the time of joining this study. Two alcohol beverages were served including beer containing 5% (v/v) of ethanol and wine (red wine and white wine) with 13% (v/v) of ethanol. Each subject was allowed to drink at a rate they found most comfortable, however, all were required to finish their drinks within an hour. Mixing of different alcohol concentrations was restricted strictly to avoid complexity in subsequent data interpretation. The volume of liquor administered by each subject was accurately recorded and the amount of alcohol administered ranged from 0.4 to 0.8 g/kg was generally acceptable by majority participants. High alcohol administration or overdose was avoided to ensure the experiment smoothly being conducted.

## 2.2 Sampling of Breath Alcohol

Breath alcohol monitoring was preferred instead of blood sampling because the non-invasive nature (Jones, 2000) of the former method allows frequent sampling. Alcotest 7110 evidential breathalyser (Drger Company, Lubeck, Germany) was selected for breath alcohol monitoring because of its high specificity for ethyl alcohol (Gullberg, 2000). To prevent inaccuracy originating from residual alcohol (Gullberg, 1992) the first breath alcohol measurement was made about 10-15 min after drinking stopped. During the initial rapid absorption phase, measurements at 10-15 min intervals were usually made for the first 60 min so as not to miss alcohol concentration at peak. During the elimination phase, about 20-30 min intervals between data collection were sufficient for the plotting of the alcohol concentration-time profile. The subject had to provide two separate exhalations into the breathalyser at an interval of 2-5 min apart, and the mean reading was collected for subsequent data treatment. For the sake of simplicity, each city or country has its own legitimate breath/blood factor with 1:2300 being adopted in Hong Kong.

## 2.3 Statistical Analysis

### 2.3.1 Compartmental Models

The simplest one-compartment model is adopted in our analysis which can be well described in terms of bi-exponentials. Here we assume the alcohol metabolism follows compartmental characteristics with first-order kinetics applying to absorption and elimination phases, and the formula is as follows,

$$C_t = \frac{D \cdot ka \cdot kel}{Cl(ka - kel)} [\exp\{-kel \cdot t\} - \exp\{-ka \cdot t\}] + \varepsilon_t \quad (1)$$

where the subscript  $t$  is the time measure (hr);  $C_t$  is the observed BAC (mg/100ml) level at time  $t$ ; and  $D$  refers to the dose administered by the

subject in gram;  $Cl$  is the alcohol clearance measure (mg/100ml/hr); while  $ka$  and  $kel$  determine the rates of alcohol absorption and elimination, respectively; and  $\varepsilon_t$  are normal errors. To ensure positivity of the measurement, logarithm transformation applies to the estimates of clearance, absorption and elimination rates. The derived pharmacokinetic parameters of interest are as follows,

- the time to maximum BAC level:  $t_{\max} = \frac{1}{ka - kel} \log\left(\frac{ka}{kel}\right)$  ;
- the BAC level attained at peak:  $C_{\max} = \mu(t_{\max}) = \frac{D}{V} \left(\frac{kel}{ka}\right)^{kel/(ka - kel)}$  where  $V$  measures the central compartment volume of distribution,  $V = Cl/kel$ ;
- the clearance:  $Cl$ ;
- the elimination half-life:  $t_{1/2} = \frac{\log 2}{kel}$ .

For multiple individuals, random effects are introduced to the compartmental model in order to account for the variability among subjects. Here the pharmacokinetic parameters are taken as follows,

$$\begin{aligned} Cl &= \exp\{\alpha_1 + a_1\}, \text{ the alcohol clearance;} \\ ka &= \exp\{\alpha_2 + a_2\}, \text{ the absorption rate;} \\ kel &= \exp\{\alpha_3 + a_3\}, \text{ the elimination rate;} \end{aligned}$$

where  $\alpha_1$ ,  $\alpha_2$  and  $\alpha_3$  denote the fixed-effects of the measured pharmacokinetic parameters correspondingly; while  $a_1$ ,  $a_2$  and  $a_3$  are the measurements of random-effects towards each parameter of interest, and follow the multivariate normal distribution.

### 2.3.2 Gamma Generalized Linear Models

Wakefield (2004) simulated the one-compartment behavior of theophylline via the generalized linear approach which is a log-linear fractional polynomial model given as follows,

$$E(C_{ij}) = D^* \exp(\beta_0 + \beta_1 t_{ij} + \beta_2 / t_{ij}) \quad (2)$$

where  $\beta_2$  determines the absorption; and the condition  $\beta_0 > 0$ ,  $\beta_1 < 0$  and  $\beta_2 < 0$  are required in order to make the model interpretable, an increasing absorption phase and a decreasing elimination phase. The data are assumed to follow the gamma distribution;  $C_{ij} \sim Ga\{\phi^{-1}, (\mu_{ij}\phi)^{-1}\}$  and the  $\phi^{1/2}$  is the coefficient of variation. The standard GENMOD procedure available in SAS/STAT module is adopted to conduct the analysis (SAS Institute, 2008). Here we invoke the option of GEE (Generalized Estimating Equations) method in the statement to deal with correlated data structure arising from the repeated measurements. The derived parameters of interest are as follows,

- the time to maximum BAC level:  $t_{\max} = (\frac{\beta_2}{\beta_1})^{1/2}$ ;



- the BAC level attained at peak:  $C_{\max} = D^* \exp[\beta_0 - 2(\beta_1\beta_2)^{1/2}]$ ;
- the clearance:  $Cl = \frac{(\beta_1/\beta_2)^{1/2}}{2 \exp(\beta_0) K_1 [2(\beta_1\beta_2)^{1/2}]}$ ; where  $K_n[x]$  denotes a modified Bessel function of the second kind of order  $n$ ;
- the elimination half-life:  $t_{1/2} = -\frac{\log 2}{\beta_1}$ .

For multiple individuals, a generalized linear mixed model (GLMM) is proposed to capture the variability among subjects. All parameters  $\beta$ 's in formula (2) are taken as random and follow the multivariate normal distribution.

### 3 Results

#### 3.1 Individual Fitting

Excluding those subjects who failed to satisfy the convergence criterion in model fitting, there are a total of 145 subjects involved in our subsequent analysis. Both compartmental and generalized linear models are fitted to each subject involved and their corresponding pharmacokinetic parameter estimates are given in the upper panel of Table 1.

Parameters like the BAC level attained at peak are dose dependent, and the mean estimate may not reflect the full picture of biological process well because of various alcohol dose taken by the participants. Here we make a comparison between the compartmental and generalized linear models. According to the compartmental method, it takes about 0.485 hour on average for alcohol concentration to reach its maximum level in blood after dose administration. The generalized linear model obtains almost the same result. Moreover, the two models give very close estimates for the average BAC level attained at peak  $C_{\max}$ . In the post absorption phase, the pharmacokinetic clearance parameter helps measure the volume of BAC clearing off from the body per unit time and it is found that the clearance estimates of the two models are similar to each other. The estimates of the elimination half-life, however, are rather different in the two models. To illustrate the fitting of the two models, we plot in Figure 1 the BAC-time profile of a female subject under a full stomach condition. Both compartmental and generalized linear models are fitted to the subject in order to trace her biological process of alcohol administration. The fitted curves in Figure 1 under the two models look very similar to each other.

#### 3.2 Population Fitting

In the preceding section, we consider model fitting at the individual level. Here we extend the concept from the individual to the population level by taking all parameters as random. The one-compartment model with random effects now is in the form of a nonlinear mixed representation while the generalized linear mixed model is the alternative method to trace the variability

among subjects. The lower panel of Table 1 summaries the findings of pharmacokinetic parameters obtained from the population mixed models. Similar to the individual fitting, consistent findings can be observed between the two population models in the estimates of the time to the maximum BAC level and the BAC level attained at peak. However, for elimination half-time, as in the individual fitting, the estimate obtained under the generalized linear model is much smaller than that obtained under the compartmental model. This suggests that the former model may not provide a good approximation to the compartmental model in the estimation of elimination half-time.

## 4 Discussion

The pharmacokinetics of alcohol has widely been studied since the 1930s. Usually, differential calculus is adopted to describe the drug flow between the compartments and hence formulates mathematical models to calculate alcohol concentration level at any time-profile after dose administration. Pharmacokinetic modeling via a compartmental approach offers a convenient visualization towards the biological process of alcohol administration and this is particularly helpful in drink driving investigation when the time BAC retrospect extrapolation is needed. Despite the BAC time-profile outlining, the compartmental model also serves as a good reference in pharmacokinetic parameters qualifying which is essential to clinical application.

In this study, we compare the performance of the traditional one compartment model to the alternative generalized linear model proposed by Wakefield (2004), in analysing the Chinese breath alcohol measurements. Four core pharmacokinetic parameters are examined accordingly. Except the elimination half-life parameter, our results show that consistent estimates can be obtained for the other three parameters under the compartmental and the generalized linear models. Some participants (about 20%) were excluded from our analysis due to the convergence issue. On closer examination, it is noted that their BAC time-profiles are somehow different from the others with a much higher initial concentration followed by a slow decay in the post absorption phase. Further research is needed to tackle such limitations. The major benefit of the generalized linear model is its computational convenience comparing to the compartmental method (Wakefield, 2004). It takes the form of a standard generalized linear model and the inference is also convenient that makes the method attractive to the analysis of pharmacokinetic data.

## 5 Acknowledgement

The authors would like to thank the three reviewers for their helpful comments and Dr T.L. Ting, the Government Chemist of the Hong Kong Special

Administration Region, for his support and permission to publish this article. This work was partially supported by the HKU Research Output Prize Funding.

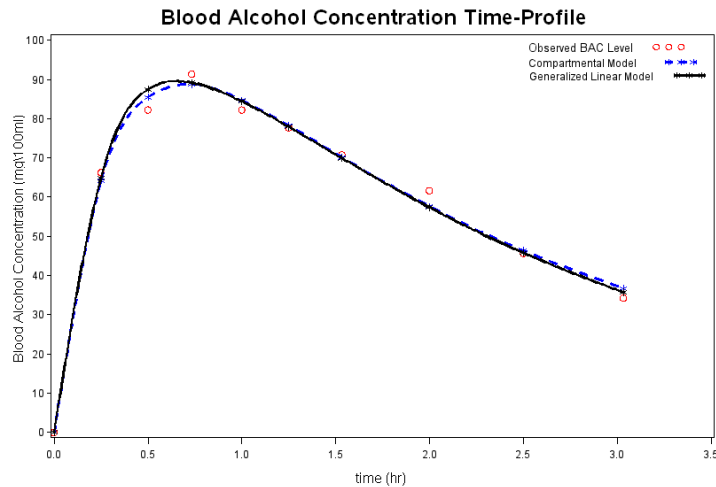
## References

- CROWDER, M. J. (1983): A growth curve analysis for EDP curves. *Appl. Statist.* 32, 15-18.
- CROWDER, M. J. and TREDER, J. A. (1981): The use of exponentially damped polynomials for biological recovery data. *Appl. Statist.* 30, 147-152.
- DAVIDIAN, M. and GILTINAN, D. M. (1995): *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, London.
- GIBALDI, M. and PERRIER, D. (1982): *Pharmacokinetics*, 2nd Edition. Dekker, New York.
- GULLBERG, R. G. (1992): The elimination rate of mouth alcohol: mathematical modelling and implications in breath alcohol analysis. *Journal of Forensic Science* 37, 1363-1372.
- HOLFORD, N. H. G. (1987): Clinical Pharmacokinetics of Ethanol. *Clinical Pharmacokinetics* 13, 273-292.
- JONES, A. W. (1990): Status of alcohol absorption among drinking driver. *Journal of Analytical Toxicology* 14, 198-200.
- KWON, Y. (2001): *Handbook of Essential Pharmacokinetics, Pharmacodynamics, and Drug Metabolism for Industrial Scientists*. Kluwer Academic/Plenum Publishers, New York.
- LINDSEY, J. K., BYROM, W. D., WANG, J., JARVIS, P. and JONES, B. (2000): Generalized Nonlinear Models for Pharmacokinetic Data. *Biometrics* 56, 81-88.
- LINDSEY, J. K. (2001): *Nonlinear Models in Medical Statistics*. Oxford University Press, U.K.
- NELDER, J. A. (1991): Generalized Linear Models for enzyme-kinetics data. *Biometrics* 47, 1605-1615.
- NORBERG, A., JONES, W. A., HAHN, R. G. and GABRIELSSON, J. L. (2003): Role of Variability in Explaining Ethanol Pharmacokinetics: Research and Forensic Applications. *Clin Pharmacokinet* 42(1), 1-31.
- PIANTADOSI, S. (1997): *Clinical Trials: A Methodologic Perspective*. Wiley, New York.
- RITSCHER, W. A. and KEARNS, G. L. (1999): *Handbook of Basic Pharmacokinetics (Including Clinical Applications)*, 5th Edition. American Pharmaceutical Association, Washington.
- SALWAY, R. and WAKEFIELD, J. (2008): Gamma Generalized Linear Models for Pharmacokinetic Data. *Biometrics* 64, 620-626.
- SAS Institute Inc. (2008): SAS/STAT 9.2 User's Guide. SAS Institute Inc., Cary, N.C.
- TAM, T. W. M., YANG, C. T., FUNG, W. K., MOK, V. K. K. (2005): Widmark factors for local Chinese in Hong Kong: A statistical determination on the effects of various physiological factors. *Forensic Science International* 151, 23-29.

WAKEFIELD, J. (2004): Non-linear regression modeling and inference. In *Methods and Models in Statistics*. Adams, N., Crowder, M., Hand, D., Stephens, D. (eds), pp. 119-153. Imperial College Press, London.

**Table 1.** Summary statistics of four core pharmacokinetic parameters including the time to maximum BAC level, the BAC level attained at peak, the rate of clearance and elimination half-life for the individual to the population fittings.

Parameter	Compartmental Model			Generalized Linear Model		
	Mean	S.D.	95% Confidence Intervals for Mean	Mean	S.D.	95% Confidence Intervals for Mean
<i>Individual Fitting</i>						
$t_{\max}$	0.485	0.214	(0.450, 0.520)	0.484	0.198	(0.451, 0.516)
$C_{\max}$	60.37	21.68	(56.81, 63.93)	60.52	21.34	(57.02, 64.03)
Clearance	21.90	12.17	(19.90, 23.90)	23.43	12.63	(21.35, 25.50)
$t_{1/2}$	2.090	1.100	(1.909, 2.270)	1.757	0.804	(1.625, 1.889)
<i>Population Fitting</i>						
$t_{\max}$	0.457	0.167	(0.430, 0.485)	0.487	0.152	(0.462, 0.512)
$C_{\max}$	61.64	25.83	(57.40, 65.88)	59.90	21.12	(56.43, 63.37)
Clearance	19.72	8.985	(18.24, 21.19)	23.39	12.18	(21.39, 25.39)
$t_{1/2}$	2.316	1.152	(2.127, 2.505)	1.731	0.692	(1.617, 1.844)



**Fig. 1.** A female subject's BAC time profile after alcohol administration under a full stomach. Both compartmental and generalized linear models are fitted.

# Fisher Scoring for Some Univariate Discrete Distributions

Thomas W. Yee

Department of Statistics, University of Auckland, Private Bag 92019,  
Auckland Mail Centre, Auckland 1142, New Zealand, *t.yee@auckland.ac.nz*

**Abstract.** The classes of vector generalized linear and additive models (VGLMs and VGAMs; Yee and Wild (1996), Yee and Hastie (2003)) enables maximum likelihood estimation of many models and distributions including categorical data analysis, survival analysis, time series, data, nonlinear least-squares models, and scores of standard and nonstandard univariate and continuous distributions. Usually Fisher scoring is used for these. This paper focusses on univariate discrete distributions, e.g., the negative binomial, zero-inflated and zero-altered Poisson and negative binomial distributions, the zeta and Zipf distributions, etc. A selection of topics are chosen, e.g., the choice of initial values that are robust to outliers is often as much an art as it is a science. The author's VGAM package for R is used for illustration.

**Keywords:** maximum likelihood estimation, Fisher scoring, univariate discrete distributions, vector generalized linear and additive models, VGAM package for R

## 1 Introduction

In the area of applied statistics and data analysis the estimation of parameters of univariate discrete distributions is common, e.g., the Poisson and negative binomial (NB) distributions for counts. Sometimes zeros require special treatment and this leads to zero-inflated and zero-altered (hurdle) variants. Positive versions of these also exist, i.e., zero-truncated. Others such as the logarithmic and zeta distributions are based on mathematical functions or series expansions. Books such as Johnson et al. (2005) give authoritative treatments on many such distributions.

In R there are now many packages for parameter estimation. However, many of them are limited by not allowing the parameters to be modelled as functions of covariates other than an intercept term, i.e., parameter  $\theta_j$  is scalar only. Some packages simply feed the problem into a general optimizer such as `optim()`.

The VGAM package differs from other implementations in that its framework is much wider. This makes usage much easier for the practitioner, e.g., no need to switch between several packages. The key algorithm is iteratively reweighted least squares (IRLS) and Fisher scoring (FS). In this paper we focus on univariate discrete distributions for brevity rather than because the

package/framework is limited to such. Table 1 summarizes some of these families. For brevity we also concentrate on VGLMs and largely omit VGAMs.

The NB distribution is used a lot in this paper so it is helpful to describe it now. One NB parameterization has probability function

$$P(Y = y) = \binom{y+k-1}{y} \left( \frac{\mu}{\mu+k} \right)^y \left( \frac{k}{k+\mu} \right)^k, \quad y = 0, 1, \dots, \quad (1)$$

where  $k > 0$ . Some software implementations are restricted to an intercept-only estimate of  $k$ . In contrast, `family = negbinomial(zero = NULL)` fits (4)–(5) in VGAM. The positive, zero-inflated and zero-altered variants are readily handled; see Table 1.

Type	Name	Distribution
Exponential family	<code>binomialff</code> <code>poissonff</code>	Binomial Poisson
Others distributions	<code>borel.tanner</code> <code>felix</code> <code>genpoisson</code> <code>geometric</code> <code>hzeta</code> <code>invbinomial</code> <code>logff</code> <code>negbinomial</code> <code>yulesimon</code> <code>zetaff</code> <code>zipf</code>	Borel-Tanner Felix Generalized Poisson Geometric Haight's zeta Inverse binomial Logarithmic Negative binomial Yule-Simon Zeta Zipf
Positive, zero-altered, zero-inflated	<code>zanegbinomial</code> <code>zapoisson</code> <code>zibinomial</code> <code>zinegbinomial</code> <code>zipoisson</code> <code>posbinomial</code> <code>posnegbinomial</code> <code>pospoisson</code>	Zero-altered negative binomial Zero-altered Poisson Zero-inflated binomial Zero-inflated negative binomial Zero-inflated Poisson Positive binomial Positive negative binomial Positive Poisson
Misc. and variants	<code>amlbinomial</code> <code>amlpoisson</code> <code>cenpoisson</code> <code>mbinomial</code> <code>mix2poisson</code>	Asymmetric maximum likelihood—for binomial Asymmetric maximum likelihood—for Poisson Censored Poisson Matched binomial Mixture of two Poissons

**Table 1.** Some VGAM family functions for univariate discrete responses. They have been grouped informally. Zero-altered are also known as hurdle models.

## 2 The VGLM framework

Suppose the data is  $(\mathbf{x}_i, y_i)$ ,  $i = 1, \dots, n$ , where the observed response is  $y$ . VGLMs are defined as a model for which the conditional distribution of  $Y$  given explanatory  $\mathbf{x}$  is of the form

$$f(y|\mathbf{x}; \mathbf{B}) = h(y, \eta_1, \dots, \eta_M) \quad (2)$$

for some known function  $h(\cdot)$ , where  $\mathbf{B} = (\boldsymbol{\beta}_1 \boldsymbol{\beta}_2 \cdots \boldsymbol{\beta}_M)$  is a  $p \times M$  matrix of unknown regression coefficients and the  $j$ th linear predictor is

$$\eta_j = \eta_j(\mathbf{x}) = \boldsymbol{\beta}_j^T \mathbf{x} = \sum_{k=1}^p \beta_{(j)k} x_k, \quad j = 1, \dots, M. \quad (3)$$

Here  $\mathbf{x} = (x_1, \dots, x_p)^T$  with  $x_1 = 1$  if there is an intercept. In this paper (2) is a distribution whose parameters are modelled using linear predictors, e.g.,

$$\log \mu = \eta_1 = \boldsymbol{\beta}_1^T \mathbf{x}, \quad (4)$$

$$\log k = \eta_2 = \boldsymbol{\beta}_2^T \mathbf{x}, \quad (5)$$

for the NB distribution (1). Another example is a distribution with a location parameter  $\xi$  and a scale parameter  $\sigma > 0$ , where we may take  $\eta_1 = \xi$  and  $\eta_2 = \log \sigma$ . In general,  $\eta_j = g_j(\theta_j)$  for some parameter link function  $g_j$  and parameter  $\theta_j$ , e.g., a positive parameter such as  $k$  in (1) is best dealt with using a log link to give (5).

VGLMs are estimated using IRLS. We have a log-likelihood  $\ell = \sum_{i=1}^n w_i \ell_i\{\eta_1(\mathbf{x}_i), \dots, \eta_M(\mathbf{x}_i)\}$  with known positive prior weights  $w_i$ . Let  $\mathbf{U}(\boldsymbol{\beta}) = \partial \ell / \partial \boldsymbol{\beta}$  denote the score vector and  $\mathcal{J}(\boldsymbol{\beta}) = -\partial^2 \ell / (\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T)$  the observed information matrix (OIM). The Newton-Raphson algorithm for maximizing the likelihood is  $\boldsymbol{\beta}^{(a+1)} = \boldsymbol{\beta}^{(a)} + \mathcal{J}(\boldsymbol{\beta}^{(a)})^{-1} \mathbf{U}(\boldsymbol{\beta}^{(a)})$  which, in this case, can be written in IRLS form as

$$\boldsymbol{\beta}^{(a+1)} = \left( \sum_{i=1}^n \mathbf{X}_i^T \mathbf{W}_i^{(a)} \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{X}_i^T \mathbf{W}_i^{(a)} \mathbf{z}_i^{(a)} \right). \quad (6)$$

Here  $\mathbf{X}_i = \text{Diag}(\mathbf{x}_i^T, \dots, \mathbf{x}_i^T)$ ,  $\mathbf{W}_i$  has  $(j, k)$ th element

$$(\mathbf{W}_i)_{jk} = -E \left[ \frac{\partial^2 \ell_i}{\partial \eta_j \partial \eta_k} \right], \quad (7)$$

(usually) and the adjusted dependent vector  $\mathbf{z}_i^{(a)}$  is an  $M$ -vector given by  $\mathbf{z}_i^{(a)} = \mathbf{X}_i \boldsymbol{\beta}^{(a)} + \left( \mathbf{W}_i^{(a)} \right)^{-1} \mathbf{d}_i^{(a)}$ , where  $\mathbf{d}_i$  has  $j$ th element  $(\mathbf{d}_i)_j = \partial \ell_i / \partial \eta_j$ .

The iteration number is  $a$ . In IRLS, the  $\mathbf{z}_i^{(a)}$  are regressed upon a large model matrix at each iteration. The *working weight matrices*  $\mathbf{W}_i^{(a)}$  here may correspond to the OIM, else the expected information matrices (EIM) as in (7). The EIM means a Fisher scoring algorithm whereas the OIM means the Newton-Raphson algorithm. Since *each*  $\mathbf{W}_i^{(a)}$  needs to be positive-definite FS is almost always used.

### 3 Software design

As all the distributions of Table 1 reside within the general framework of Section 2 the software is able to exploit its structure and make it more easy for the user. This comes about by using common arguments, drawing from a pool of link functions and re-using several types of algorithms for computing initial values. Here are some details.

#### Some user-oriented topics

In R, VGLMs may be fitted in a manner very similar to GLMs, viz.

```
vglm(yvector ~ x2 + x3, family = VGAMfamilyFunction, data = mydataframe)
```

The function assigned to the `family` argument is known as a VGAM *family function*. Table 1 lists some relevant to this paper. Written in S4, VGAM has methods functions for standard generics such as `coef()`, `fitted()`, `predict()`, `summary()` and `vcov()`.

#### Random variates and EIM estimation

Many `dpqr`-functions exist to return the density, distribution function, quantile function and random generation of their respective VGAM family function. For example, there are the `[dpqr]zipois()` functions corresponding to the zero-inflated Poisson distribution VGAM family function `zipoisson()`.

Section 4.1 describes approximating the EIM using simulation. For this, the argument `nsimEIM` controls the number of random variates generated. This number holds true at each  $i$ , and if the model is intercept-only then the working weight can be averaged over all  $i$ .

#### Link functions, Initial values and convergence

Most VGAM family functions allow a flexible choice of link functions to be assigned to each  $\eta_j$ . Some VGAM link functions currently available such as `cloglog`, `loge`, `logit`, `loglog`. There are about a dozen available currently.

Parameter link functions are used to increase the chances of successful convergence by avoiding numerical problems, e.g., positive parameters have a log link. This also improves the quadratic approximation to the log-likelihood function which means a higher chance of convergence and at a speedier rate.

The argument `shrinkage.init` is used for one type of algorithm for computing initial values; see Section 4.2 and Section 5.3 for an example.

## 4 Selected topics

### 4.1 Working weight matrix estimation

While the EIM is tractable for many common univariate discrete distributions, it is very commonly untractable for others. Several methods have been proposed in the literature. Probably the most common method implemented in VGAM is to use simulated FS to estimate the working weights



$w_i \text{Var}(\partial \ell_i / \partial \boldsymbol{\eta}_i)$ . It requires an expression for the score vector and the ability to generate random variates from that distribution. Both of these are often not too difficult. As a specific example, the VGAM family function `negbinomial()` uses `rnbinom()` (from the `stats` package) for computing the working weights in the `weights` slot. The 1-1 element of the EIM is tractable and therefore requires no approximation. The EIM is diagonal and

$$\frac{\partial \ell_i}{\partial k} = \psi(y_i + k) - \psi(k) - \frac{y_i + k}{k + \mu_i} + 1 + \log \frac{k}{k + \mu_i}. \quad (8)$$

The 2-2 element of the EIM involves an infinite series, although it can easily be calculated using `pnbinom()` if  $\mu \approx 0$ .

## 4.2 Initial values

Initial values can be quite critical for certain distributions. Consequently, convergence for some distributions is intrinsically harder to achieve than in others. The choice of initial values is often as much an art as it is a science.

All VGAM family functions are self-starting. We want initial values which are easy to compute, robust to outliers, and close to the solution. Some ideas are to use method-of-moments estimators (where available) and grid search methods.

The VGAM package offers a wide selection of options for users to input initial values. Here is a short description.

- a. Arguments `coefstart`, `etastart` and `mustart` are available for `vglm()` and `vgam()`. These are the most general and apply to any VGAM family function. They are  $\beta_i^{(0)}$ ,  $\eta_i^{(0)}$ ,  $\mu_i^{(0)}$  respectively.
- b. Argument such as `iscale` and `ishape` are available for many VGAM family functions. The “i” in their name means for initial values.
- c. The argument `method.init` is available many VGAM family functions, and can be assigned an integer from one upwards. The value 1 is always the default, meaning the ‘best’ method—the one most likely to lead to successful convergence. If failure occurs then the user can try `method.init = 2`, and/or `method.init = 3`, etc. One of the values might mean using a method-of-moments estimator, else a grid search, etc.

One of the examples (Section 5.3) uses the `shrinkage.init` argument. This represents a value  $s$ , say, where  $0 \leq s \leq 1$ . Then in general, the formula used is something like  $\mu_i^{(0)} = s\tilde{\mu} + (1 - s)y_i$  where  $\tilde{\mu}$  is some measure of central tendency such as a weighted mean or median. The initial values can be thought of as the mean or median plus slight perturbations towards the actual data. A default value typically might be  $s \approx 0.95$ .

## 4.3 Half-stepping

Half-stepping is a method for increasing the chance of successful convergence. Assuming the model has a log-likelihood function, half-stepsizing ensures that

an improvement is made at each iteration. This extension takes a half step if the next iteration ‘overshoots’ the solution because the quadratic approximation to the log-likelihood function is inaccurate. If necessary an even smaller step is taken (e.g.,  $\frac{1}{4}$  or  $\frac{1}{8}$  etc.), and an improvement is guaranteed since the step direction is ascending (because the  $\mathbf{W}_i^{(a)}$  are positive-definite).

Half-stepsizing is particularly useful at early stages of the iteration if the initial values are poor. Furthermore, a very large negative value of the log-likelihood can be returned to prevent an iteration from falling outside the parameter space.

## 5 Examples

Here are some simple examples illustrating a few of the concepts of this paper.

### 5.1 Example 1: zeta distribution

The zeta distribution is sometimes used in insurance to model the number of policies held by an individual in an insurance portfolio. The long tailed distribution has the probability function  $P(Y = y) = 1/[y^{p+1}\zeta(p+1)]$ , for  $p > 0$  (so a log link is the default) and  $y = 1, 2, \dots$ . Then  $E(Y) = \mu = \zeta(p)/\zeta(p+1)$  and  $\text{Var}(Y) = \zeta(p-1)/\zeta(p+1) - \mu^2$  provided  $p > 1$  and  $p > 2$  respectively. The following data comes from Knight (2000).

```
> zdata <- data.frame(y = 1:5, w = c(63, 14, 5, 1, 2))
> fit <- vglm(y ~ 1, zetaff, zdata, weight = w)
> (phat <- Coef(fit))
      pp
1.682557
> with(zdata, cbind(round(dzeta(y, phat) * sum(w), 1), w))
      w
[1,] 66.4 63
[2,] 10.3 14
[3,]  3.5  5
[4,]  1.6  1
[5,]  0.9  2
> with(zdata, c(SampleMean = weighted.mean(y, w), EstimatedMean =
      head(fitted(fit), 1)))
      SampleMean EstimatedMean
      1.411765      1.633302
```

That is,  $\hat{p} \approx 1.68$  so that the estimated mean but not the variance are finite. A goodness of fit test could be done to check model fit.

### 5.2 Example 2: half-stepping

The following shows the benefit of half-stepping in a binary regression. Ridout (1990) gives this miniature data set where FS fails. The setting is dilution series data where the log-likelihood is close to quadratic. The dilution series

method is used to estimate the density of organisms in a sample when direct counting is not possible, but the presence/absence of the organism in a subsample can be determined. The original sample is progressively diluted and at the  $i$ th dilution  $n_i$  subsamples are tested, of which, say,  $Y_i$  are found to contain the organism. Let  $v_i$  denote the volume of the original sample which is present in each of the subsamples at the  $i$ th dilution. The parameter of interest is the density of organisms per unit volume,  $\lambda$ , say. If the organisms are randomly distributed then the number of organisms in a subsample at the  $i$ th dilution  $\sim \text{Poisson}(\mu_i = \lambda v_i)$ . In particular, let the probability  $\pi_i$  be  $P(\text{the organism is present in the subsample}) = 1 - \exp(-\lambda v_i)$ . So  $Y_i \sim \text{Bin}(n_i, \pi_i)$  (assuming independence between the subsamples at each dilution). We wish to estimate  $\theta = \log \lambda$ . It is easy to show  $\log(-\log(1 - \pi_i)) = \eta_i = \theta + \log v_i$ . But

```
ridout <- data.frame(v = 10^(3:1), r = c(4, 3, 3), n = rep(5, 3))
ridout <- transform(ridout, logv = log(v))
glm.fail = glm(r/n ~ offset(logv), weight = n, binomial(link =
  cloglog), ridout, trace=TRUE, maxit = 25)
```

fails because the iterations oscillates between two local solutions. Instead,

```
> library(VGAM)
> vglm.ok <- vglm(r/n ~ offset(logv) + 1, weight = n,
+   binomialfff(link = cloglog), ridout, trace = TRUE)
VGLM    linear loop 1 : deviance = 22.7279
VGLM    linear loop 2 : deviance = 28.2982
Taking a modified step..
VGLM    linear loop 2 : deviance = 18.0853
VGLM    linear loop 3 : deviance = 17.8097
VGLM    linear loop 4 : deviance = 18.2077
Taking a modified step..
VGLM    linear loop 4 : deviance = 17.4608
VGLM    linear loop 5 : deviance = 17.4662
Taking a modified step..
VGLM    linear loop 5 : deviance = 17.4367
VGLM    linear loop 6 : deviance = 17.4367
Taking a modified step..
VGLM    linear loop 6 : deviance = 17.4366
VGLM    linear loop 7 : deviance = 17.4366
Taking a modified step..
VGLM    linear loop 7 : deviance = 17.4366
```

This effectively converges to the maximum likelihood estimate (MLE),  $\hat{\theta} = -5.4007$ , successfully.

### 5.3 Example 3: negative binomial initial values

We show how robust `negbinomial()` can be to outliers.

```
set.seed(123); x <- runif(n <- 500)
y1 <- rnbinom(n, mu = exp(3+x), size = exp(1)) # k is size
```

That is,  $n = 500$ ,  $X_i \sim \text{Unif}(0, 1)$ ,  $\mu = \exp(3 + x)$ ,  $k = e^1 \approx 2.72$  in (1). Then

```
> y1[1] <- 1e8
> nbfit <- vglm(y1 ~ x, negbinomial, trace=TRUE)
VGLM    linear loop  1 : loglikelihood = -3364.296
VGLM    linear loop  2 : loglikelihood = -3340.707
VGLM    linear loop  3 : loglikelihood = -3323.534
... <edited> ...
VGLM    linear loop  9 : loglikelihood = -3241.135
VGLM    linear loop 10 : loglikelihood = -3241.058
VGLM    linear loop 11 : loglikelihood = -3241.058
```

Here the response might typically have a maximum of around 200, but one of the values is replaced by  $10^8$ . Convergence is still achieved.

## 6 Discussion

This paper has described a general regression framework to compute MLEs of parameters from univariate discrete distributions. The software design takes advantage of the theoretical structure so that the user can more easily fit a whole class of models.

Being a large project, there is much more work to do. There are scores and scores of univariate discrete distributions currently unimplemented. Like all software, writing reliable and efficient code requires skill, as well as a lot of time to test, maintain and cater for future improvements.

Outside the small arena of univariate discrete distributions there are hundreds of unimplemented models and distributions. These potentially cover a wide range of multivariate response types and models, including univariate and multivariate distributions, categorical data analysis, quantile/expectile regression, time series, survival analysis, extreme value analysis, mixture models, correlated binary data and nonlinear least-squares problems. Implementing more of these within VGAM is a perpetual task!

## Acknowledgements

The author wishes to thank Prof Ridout for a copy of his paper.

## References

- JOHNSON, N. L., KEMP, A. W., and KOTZ, S. (2005): *Univariate Discrete Distributions, Third Edition*. John Wiley & Sons, Hoboken, NJ, USA.
- KNIGHT, K. (2000): *Mathematical Statistics*. Chapman & Hall/CRC Press, Boca Raton, FL, USA.
- RIDOUT, M. S. (1990): Non-convergence of Fisher's method of scoring—a simple example. *GLIM Newsletter* 20(6).
- YEE, T. W., HASTIE, T. J. (2003): Reduced-rank vector generalized linear models. *Statistical Modelling* 3(1), 15–41.
- YEE, T. W., WILD, C. J. (1996): Vector generalized additive models. *Journal of the Royal Statistical Society B*, 58(3), 481–493.

# Constructing Economic Summary Indexes via Principal Curves

Mohammad Zayed<sup>1,2</sup> and Jochen Einbeck<sup>1</sup>

<sup>1</sup> Department of Mathematical Sciences, Durham University, Science Laboratories, South Rd., Durham, DH1 3LE, UK, [jochen.einbeck@dur.ac.uk](mailto:jochen.einbeck@dur.ac.uk)

<sup>2</sup> Applied Statistics and Insurance Department, Mansoura University, Mansoura, 35516, Egypt, [m.a.zayed@dur.ac.uk](mailto:m.a.zayed@dur.ac.uk)

**Abstract.** Index number construction is an important and traditional subject in both the statistical and the economical sciences. A novel technique based on *localized principal components* to compose a single summary index from a collection of indexes is proposed, which is implemented by fitting a (local) principal curve to the multivariate index data. We exploit the ability of principal curves to extract robust low-dimensional ‘features’ (corresponding to the summary index) from high-dimensional data structures, yielding further useful analytic tools to study the behaviour and composition of the summary index over time.

**Keywords:** summary indexes, feature extraction, principal component analysis, smoothing

## 1 Introduction

A standard problem in economics is the question of how to construct a single (summary) index from a series of individual (sub-)indexes. For instance, the main measure of inflation for national macro-economic purposes is the Consumer Price Index (CPI), which covers essentially the monetary expenditures on all goods and services by all households of a certain economy (for instance, the UK). This index, say  $X_0$ , is usually computed from sub-indexes  $X = (X_1, \dots, X_p)'$  by weighted averaging of type

$$X_0 = w_1X_1 + \dots + w_pX_p = w'X \quad (1)$$

where  $w = (w_1, \dots, w_p)'$  is a set of weights relating to the composition of expenditure, which is allowed to vary over time, i.e.  $w = w(t)$ . Economists have taken substantial efforts to derive formulas which give appropriate or ‘representative’ weights for a certain economy. The actual process of averaging in (1) is rather crude from a statistical perspective. It is highly dependent on outlying (potentially erroneous) data, it is not able to deal with missing data, it does not allow an analysis of the relative contribution of the sub-indexes over time, and does not take into account the differing variability

(information) contained in the indexes at different time points (other than through the weights, perhaps). A potential alternative addressing these issues was already suggested by Tintner (1946) and Moser (1984) in the context of production and price indexes, and labour market indicators, respectively. They proposed to construct a linear summary index by finding that linear combination  $\gamma'X$  of  $X_1, \dots, X_p$  with maximal variance  $\text{Var}(\gamma'X)$  among all unit vectors  $\gamma$ . The solution to this problem is found via principal component analysis (PCA), and is given by the first eigenvector  $\gamma$  of the covariance matrix  $\Sigma = \text{Cov}(X)$  of  $X$ . Assuming the existence of a ‘price line’  $X = aX_0 + \epsilon$ , with  $a \in \mathbb{R}^p$ , Theil (1960) developed a variant of PCA to estimate  $a$  and  $\gamma$  simultaneously. Neither of these authors used any additional weighting, though (external) weights  $w$  could be easily accommodated by considering  $X_w = (w_1X_1, \dots, w_pX_p)'$  instead of  $X$  itself.

If we have a set of variables, each can be represented as a mix of a systematic component and an error, applying PCA to these variables results in constructing a number of independent factors, usually less in number than data dimension, which capture most of the total variance in the data set. This is done by finding some linear function of the variables in the data set, which is least subject to errors. Principal components are of interest mainly in cases where the variables under consideration, the values of which formulate the data cloud, are considered to be symmetric, rather than one or more variable being generated from the remaining ones.

PCA-based approaches have not yet found widespread application in the context of economic index data. One reason for that is that PCA will find that line through the multidimensional cloud of indexes which gives *globally* the best fit in terms of squared orthogonal distances; in other words ‘one line has to fit it all’. The approximation done this way may be good in some parts of the data cloud but poor in others. As a consequence, the loadings  $\gamma = (\gamma_1, \dots, \gamma_p)'$  will reflect the contribution of the subindexes  $1, \dots, p$  towards the overall index not equally well over the full data range — actually, the amount of information that individual indexes contribute towards the overall index may vary greatly; an example for this is provided later in this article. Hence, what would be needed is a tool to maximize the variance locally, providing at each point the best local approximation to the data cloud. This implies that we need to fit a sequence of localized principal components, rather than one global principal component. The statistical concept corresponding to this viewpoint is a (local) principal curve.

## 2 Principal curves

The concept of *principal curves* was introduced in the Statistics literature by Hastie and Stuetzle (1989) (hereafter: HS) as a nonparametric extension of PCA. A principal curve is descriptively defined as a smooth curve  $f : \mathbb{R} \rightarrow \mathbb{R}^p$ ,  $\lambda \mapsto f(\lambda)$  that passes through the ‘middle’ of a  $p$ -variate data set, providing a nonlinear summary of the data. For HS curves, the notion of the

‘middle’ of the data cloud is implemented via the concept of self-consistency (Tarpey & Flury, 1996), meaning that each point on the curve is the average of all points that project there.

Principal curves have recently attracted interest particularly in the engineering literature (Ming-Ming et al., 2010) due to their ability to extract low-dimensional ‘features’ from high-dimensional data structures via the curve parametrization  $\lambda$ . In particular, for  $X \in \mathbb{R}^p$ , one defines the *projection index* as the parameter of the closest point on the curve to  $X$ , i.e.

$$\lambda_f(X) = \sup_{\lambda} \{\lambda : \|X - f(\lambda)\| = \inf_{\eta} \|X - f(\eta)\|\}. \quad (2)$$

In our context, the extracted feature  $\lambda_f(X)$  would be corresponding to the summary index of  $X$ , as we will illustrate in the following section. However, we are not only interested in this overall index, but also in the local contributions of the individual sub-indexes, for which we need to determine loadings in terms of localized eigenvectors. The original algorithm by HS does not compute these, neither explicitly nor implicitly, so it is of limited use for our development. An alternative concept, which is explicitly based on localized PCA, is the local principal curve (LPC) algorithm (Einbeck et al., 2005):

Given  $n$  replicates of  $X$ , forming a  $p$ -variate data cloud  $x_i, i = 1, \dots, n$ , where  $x_i = (x_{i1}, \dots, x_{ip})'$ , a smooth curve which passes through the middle of the data cloud is found as follows:

- a. Choose a suitable starting point  $x_{(0)} \in \mathbb{R}^p$ , either by hand or at random from the data cloud. Set  $x = x_{(0)}$ .
- b. Calculate  $\mu^x$ , a local mean around  $x$ .
- c. Perform a principal component analysis locally at  $x$ , yielding a localized eigenvector  $\gamma^x$ .
- d. Find a new value for  $x$  by following  $\gamma^x$  a predetermined step size, starting at  $\mu^x$ .
- e. Repeat steps 2 to 4 until  $\mu^x$  remains (approximately) constant.

The local principal curve is determined by the series of the  $\mu^x$  values. The actual localization in 2. and 3. is performed through multivariate kernel functions, see Einbeck et al. (2005) for details on these steps. After termination of the algorithm, the parametrization  $\lambda$  is calculated retrospectively through the Euclidean distances between neighboring  $\mu^x$ , and interpolated between the  $\mu^x$  through linear segments or cubic splines (Einbeck et al., 2009), yielding a fully parametrized one-dimensional curve  $f(\lambda)$  through  $p$ -dimensional space, which passes precisely through all the local means  $\mu^x$ . Note that the algorithm is robust to outlying data points due to the localized way of averaging.

There is one important adjustment that is useful to be made for index data: Normally, there is some reference date for which all sub-indexes take a baseline value, say 100, and also the overall index takes this value. Hence, also the parametrized principal curve has to reflect this property and this can be

realized through an *anchor*: This is a point of predetermined coordinates, say  $x_{(0)} = (100, \dots, 100)'$ , and predetermined parameter value ('index')  $\lambda = 100$ , through which the curve *is forced to pass*. This is implemented by inverting steps 2 and 3 above, and recalculating  $\lambda$  by integrating over the arc length of the curve starting with the anchor point. Of course, this method is only feasible when the baseline time point is part of the time interval considered. We illustrate this algorithm, and its functionality as a 'feature extractor' for the summary index, in the subsequent section.

### 3 Analysis of CPI data

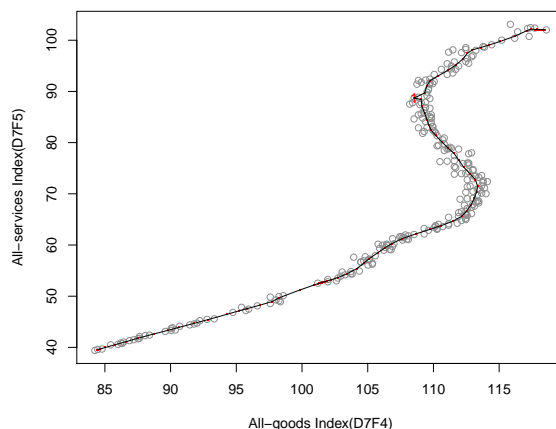
In the applied part of this work, two sets of consumer price indexes have been used, the first, as an introductory example, is a two dimensional set, and the second is a twelve dimensional set. All data are monthly UK data published through 'National Statistics Online' covering the period from January 1988 until December 2008. Both sets of indexes are complemented subsets of the same total summary index, which is the total CPI for 'All Items'. The indexes used for analysis are: (2005=100 for all indexes)

D7BT: CPI INDEX 00 : ALL ITEMS  
 D7BU: CPI INDEX 01 : FOOD AND NON-ALCOHOLIC BEVERAGES  
 D7BV: CPI INDEX 02 : ALCOHOLIC BEVERAGES, TOBACCO & NARCOTICS  
 D7BW: CPI INDEX 03 : CLOTHING AND FOOTWEAR  
 D7BX: CPI INDEX 04 : HOUSING, WATER AND FUELS  
 D7BY: CPI INDEX 05 : FURN, HH EQUIP & ROUTINE REPAIR OF HOUSE  
 D7BZ: CPI INDEX 06 : HEALTH  
 D7C2: CPI INDEX 07 : TRANSPORT  
 D7C3: CPI INDEX 08 : COMMUNICATION  
 D7C4: CPI INDEX 09 : RECREATION & CULTURE  
 D7C5: CPI INDEX 10 : EDUCATION  
 D7C6: CPI INDEX 11 : HOTELS, CAFES AND RESTAURANTS  
 D7C7: CPI INDEX 12 : MISCELLANEOUS GOODS AND SERVICES  
 D7F4: CPI INDEX: ALL GOODS  
 D7F5: CPI INDEX: ALL SERVICES

#### 3.1 Index construction from two sub-indexes

We aim to reconstruct the overall index (CPI INDEX 00: ALL ITEMS) using two sub-indexes: the CPI INDEX: ALL GOODS and the CPI INDEX: ALL SERVICES. We use the modified LPC algorithm using an anchor at  $x_{(0)} = (100, 100)'$  and  $\lambda = 100$  (corresponding to the reference point January 2005), as outlined in Section 2. For simplicity, a constant weight  $w = 1/500 * (547, 453)'$  for all years is used. Now, applying this adjusted LPC algorithm to fit a summary curve through the two weighted indexes, one obtains the fit produced in Figure 1. It seems to give a reasonable summary for the two-dimensional data set in the form of a one-dimensional curve.





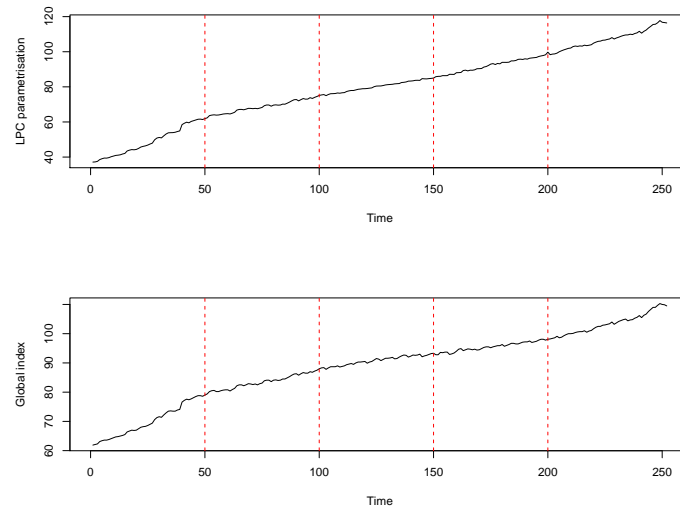
**Fig. 1.** LPC fit for 2D CPI data.

A first property of interest when using this statistical approach in CPI context could be: compared to the original overall index, how well is the resulting fit capturing the overall index behaviour? Figure 2 compares how the projection indexes  $\lambda_f(X)$  and the original **CPI INDEX 00** change over time. Figure 2 suggests that the statistically fitted overall index captures most movements in the true index, which is a desirable situation. Also, it can be seen that the fitted index looks smoother than the original index, due to the underlying smoothing properties implied by using the LPC algorithm.

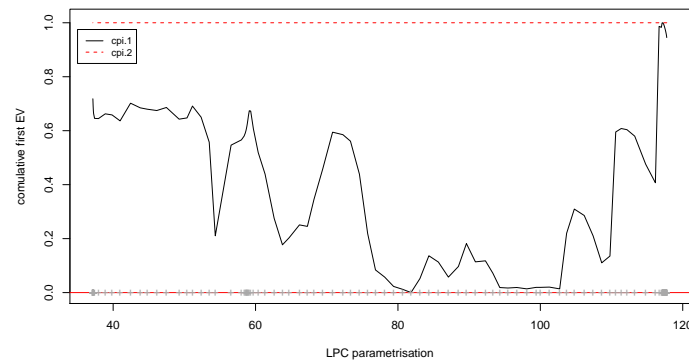
The other useful informative tool accompanying the use of LPCs is related to the total variance explained by the curve and how each variable (sub-index) contributes to the fitted overall index. This is statistically measured through ‘loadings’, i.e. the entries of the (local) eigenvectors. At every point on the curve, the sum of squared loadings of the first eigenvector should be equal to one. This ‘unity’ property of eigenvectors provides a good tool to indicate how the sub-indexes influence the fitted overall index at each point (time). Figure 3 shows the cumulative squared loadings of first eigenvectors for our example. Useful interpretations could be derived from such a figure, for instance, around the fitted curve’s parameter values of 80 and 100, the second sub-index has a dominating effect on the fitted overall index.

### 3.2 Index construction from twelve sub-indexes

Adopting the same techniques used in the previous example, the LPC algorithm was applied to fit the overall consumer price index from the twelve sub-indexes (**INDEX 01**, **INDEX 02**, ..., **INDEX 12**). Main indicators from the resulting fit are shown in Figure 4. We can study the index behaviour and the dominating underlying factors affecting it over time.

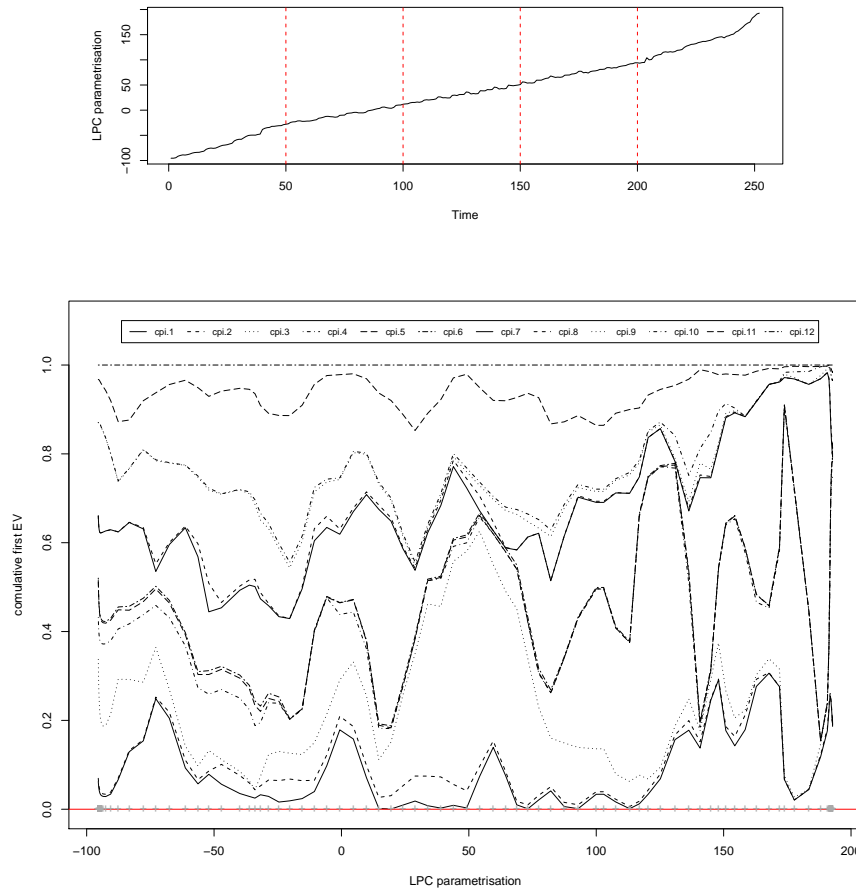


**Fig. 2.** LPC-based (top) and average-based (bottom) CPI behaviour over time.



**Fig. 3.** Cumulative squared loadings of first eigenvectors - 2D fit.

The bottom part of Figure 4 allows to assess the contributions of the 12 sub-indexes over time. For example, it can be seen that the third index has the largest effect on the fitted overall index around the LPC parameter value 50 (which corresponds to some time point near 150), and the same can be said about index four around parameter values of 120 and 178 (times: 19 and 249) and that the first index alone contributes by 30% in the fitted index around parameter value of 169 (time near 246), and so on. All such interpretations can have useful meanings in the econometrics context.



**Fig. 4.** A 12-D example. Top: reconstructed summary index (LPC parametrisation over time); bottom: cumulative squared loadings (first eigenvector) over time.

One remaining important feature of the proposed technique is the ability to predict missing data points at any given time (discrete or continuous) within the data range. This is achieved, technically, through ‘calibration’ of time and the LPC parametrisation (by plotting them against each other and using a nonparametric smoother to find the functional relationship). Having done this, if we assume that we want to predict the data point that corresponds to, say, time = 220.5, we plug this value in the calibrated object which gives a parameter value of 126.3425, then we get the corresponding estimated 12-dimensional weighted point on the fitted curve, and applying a simple adverse-weight formula to each index ( $cpi = weighted.cpi * average.weight / current.index.weight$ ), we get the real time

estimated sub-indexes' values (101.78, 102.53, 96.37, 108.04, 99.48, 102.38, 102.93, 100.1, 98.87, 104.82, 102.84, 103.38). This could be useful in handling missing data as well as predicting any assumed in-between data points (for instance, holidays).

## 4 Conclusion

The work presented in this paper is merely a statistically-based approach to fit and analyse main economic indexes. The computed index using the LPC algorithm has the ability to capture the basic trend of the original corresponding index over time. Being based upon principal component analysis, it allows to detect the influence of all variables (sub-indexes) on the fitted index at all points (time), and would furthermore allow to assess the degree of 'local linearity' of the index, in terms of total local variance explained, at each point in time by looking at the localized first eigenvalues. The main novel feature of the proposed technique is that it is nonlinear and even non-parametric, while the traditional PCA-based methods are linear, which may be of limited accuracy in particular if the time range considered is quite large.

It should be noted that the proposed technique, just as PCA itself and the modified version by Theil (1960), is an 'ex-post' algorithm, i.e. one needs to have the full data available in order to reconstruct the indexes retrospectively. However, unlike other principal curve algorithms, the LPC methodology would in principle allow for an updating algorithm, which would enable to extend the previously fitted curve and the associated statistics once new data have come in. This is a matter of future research.

## References

- EINBECK, J., TUTZ, G. and EVERS, L. (2005): Local principal curves. *Statistics and Computing* 15 (4), 301-313.
- EINBECK, J., EVERS, L., and HINCHLIFF, K. (2009): Data compression and regression based on local principal curves. In: Fink et al. (Eds): *Advances in Data Analysis, Data Handling and Business Intelligence*, Springer, Heidelberg, 701-712.
- HASTIE, T. and STUETZLE, W. (1989): Principal curves. *Journal of the American Statistical Association* 84 (406), 502-516.
- MING-MING, Y., JIAN, L., CHUAN-CAI, L. and JING-YU, Y. (2010): Similarity preserving principal curve: an optimal one-dimensional feature extractor for data representation. *IEEE Transactions on Neural Networks*, to appear.
- MOSER, J. W. (1984): A principal component analysis of labor market indicators. *Eastern Economic Journal* X (3), 243-257.
- TARPEY, T. and FLURY, B. (1996): Self-consistency: a fundamental concept in statistics. *Statistical Science* 11 (3), 229-243.
- THEIL, H. (1960): Best linear index numbers of prices and quantities. *Econometrica* 28 (2), 464-480.
- TINTNER, G. (1946): Some applications of multivariate analysis to economic data. *Journal of the American Statistical Association* 41 (236), 472-500.

# Censored Survival Data: Simulation and Kernel Estimates

Jiří Zelinka

Department of Mathematics and Statistics, Faculty of Science, Masaryk University  
Kotlářská 2, Brno, Czech Republic, [zelinka@math.muni.cz](mailto:zelinka@math.muni.cz)

**Abstract.** Non-parametric estimates of survival and hazard function belongs to the basic instruments in survival analysis. In previous papers methods of kernel estimates involving growth models of cancer cells were designed by author's colleagues. To verify the quality of these the tests on the simulated data were suggested.

During the test procedure some theoretical problems appeared. They concerned especially additional requests for distribution of simulated censoring data. The problems were largely resolved and estimation procedures were successfully tested on simulated data. This paper summarizes the achievements.

**Keywords:** hazard function, censoring, simulation, kernel estimate

## Survival and hazard functions

The survival time or lifetime, i.e. the random variable  $T$  is the time from the beginning of follow-up to the death (or to any event under consideration).

Let us denote the cumulative distribution function of  $T$  by  $F$  and the survival function by  $\bar{F} = 1 - F$ .

Hazard function  $\lambda = \lambda(x)$  describes an intensity of survival probability:

$$\lambda(x) = \frac{f(x)}{\bar{F}(x)} = -\frac{\bar{F}'(x)}{\bar{F}(x)} = -\log'(\bar{F}(x)) \quad (1)$$

if the density  $f$  exists. From (1) we have

$$\bar{F}(x) = e^{-\int_0^x \lambda(t) dt}. \quad (2)$$

## Random censorship model

Let  $T_1, T_2, \dots, T_n$  be independent and identically distributed lifetimes with distribution function  $F$ .

Let  $C_1, \dots, C_n$  be independent and identically distributed censoring times with distribution function  $G$  which are usually assumed to be independent of the lifetimes.

In the random censorship model we observe

$$(X_i, \delta_i), i = 1, \dots, n, \text{ where } X_i = \min(T_i, C_i)$$

and  $\delta_i = 1_{\{X_i=T_i\}}$  indicates whether the observations is censored or not.  $\{X_i\}$  are independent and identically distributed with survival function  $\bar{L}$ :  $\bar{L}(x) = \bar{F}(x)\bar{G}(x)$ .

## Kernel estimates of the hazard function

Let  $[0, \tau]$ ,  $\tau > 0$ , be an interval for which  $L(T) < 1$  and  $\lambda \in C^2[0, T]$  and let  $K$  be a continuous and symmetric function on  $R$  called a kernel satisfying conditions:

- a.  $\text{supp } K = [-1, 1]$
- b.  $K \in \text{Lip}[-1, 1]$
- c.  $\int_{-1}^1 x^k K(x) dx = \begin{cases} 1, & k = 0 \\ 0, & k = 1 \\ \beta_2 \neq 0, & k = 2. \end{cases}$

The well-known kernels:

$$K(x) = \frac{3}{4}(1 - x^2)1_{[-1,1]} \quad \text{Epanechnikov kernel}$$

$$K(x) = \frac{15}{16}(1 - x^2)^2 1_{[-1,1]} \quad \text{quartic kernel}$$

The kernel estimate of the hazard function is given as

$$\hat{\lambda}_{h,K}(x) = \frac{1}{h} \sum_{i=1}^n K\left(\frac{x - X_{(i)}}{h}\right) \frac{\delta_{(i)}}{n - i + 1}. \quad (3)$$

The parameter  $h$  is called bandwidth or smoothing parameter.

Let us denote

$$V(K) = \int_{-1}^1 K^2(x) dx, \quad \beta_2 = \int_{-1}^1 x^2 K(x) dx, \\ \Lambda = \int_0^T \frac{\lambda(x)}{L(x)} dx, \quad D_2 = \int_0^T \left(\lambda^{(2)}(x)\right)^2 dx.$$

The global quality of the estimate – Mean Integrated Square Error:

$$MISE(\hat{\lambda}_{h,K}) = \int_0^T MSE(\hat{\lambda}_{h,K}(x)) dx = \int_0^T E(\hat{\lambda}_{h,K}(x) - \lambda(x))^2 dx,$$

The leading term  $\overline{MISE}(\hat{\lambda}_{h,K})$  of  $MISE(\hat{\lambda}_{h,K})$  takes the form

$$\overline{MISE}(\hat{\lambda}_{h,K}) = \frac{1}{4}h^4\beta_2^2D_2 + \frac{V(K)\Lambda}{nh}$$

The asymptotically optimal bandwidth minimizing  $\overline{MISE}(\hat{\lambda}_{h,K})$  with respect to  $h$  is given by the formula

$$h_{opt} = n^{-1/5} \left( \frac{\Lambda V(K)}{\beta_2^2 D_2} \right)^{1/5} \quad (4)$$

The estimate of  $h_{opt}$  will be denoted with  $\hat{h}_{opt}$ . See Horová & Zelinka (2006) for method of evaluating the appropriate estimate  $\hat{h}_{opt}$ .

## Simulation study

### Simulation of lifetimes

For given hazard function  $\lambda$  we have (see (2))

$$F(x) = 1 - e^{-\int_0^x \lambda(t) dt}$$

We can see that the lifetimes  $T_1, \dots, T_n$  can be evaluated numerically by resampling random variables  $U_1, \dots, U_n$  uniformly distributed on interval  $[0, 1]$ .

### Simulation of censoring times

Real situation: Let's have a clinical study dealing with some disease. The research begins in time  $t_0$  (we can suppose  $t_0 = 0$ ). Patients come to the study randomly in interval  $[t_0, t_1]$ , the begin of treatment is given by random variable  $B$  with cumulative distribution function  $H$ . The coming of patients is broken in time  $t_1$ , but the study may continue to some time  $t_2 \geq t_1$  when it is finished.

The censorship time is  $C = t_2 - B$ . For the survival function  $\bar{G}$  we have

$$\bar{G}(x) = H(t_2 - x),$$

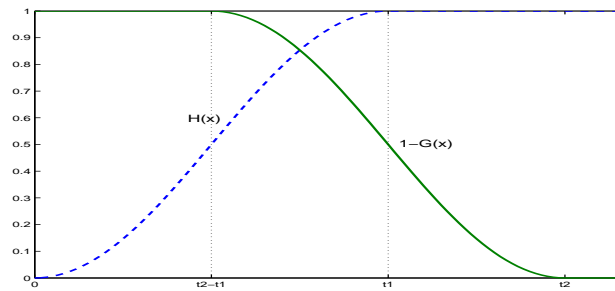
$$h_{opt}^5 = \frac{V(K)\Lambda}{n\beta_2^2 D_2},$$

for

$$V(K) = \int_{-1}^1 K^2(x) dx, \beta_2 = \int_{-1}^1 x^2 K(x) dx, \\ \Lambda = \int_0^\tau \frac{\lambda(x)}{\bar{F}(x)\bar{G}(x)} dx, D_2 = \int_0^\tau \left( \lambda^{(2)}(x) \right)^2 dx.$$

Regarding the choice of values  $\tau$ , the natural choice is  $\tau = t_2$ , but this yields problem with counting  $\Lambda$  as  $\bar{G}(t_2) = 0$ . The following procedure can be solution of this problem: for given  $\bar{G}$  let us take such  $\lambda$  that

$$\bar{F}(t_2) > 0, \quad \frac{\lambda(x)}{\bar{G}(x)} = O(1), \text{ for } x \rightarrow t_2.$$



**Fig. 1.** Cumulative distribution function for coming of patients ( $H$ ) and survival function of censoring times ( $\bar{G} = 1 - G$ ).

As a result of this property we have  $\lambda(t_2) = 0$  and for  $\lambda \in C^2[0, T]$  also  $\lambda'(t_2) = 0$  as  $\lambda$  is non-negative function.

In all simulations let the begins of treatment  $B$  be uniformly distributed on  $[0, t_1]$ . Due to this fact the cumulative distribution function  $C$  is uniformly distributed on  $[t_2 - t_1, t_2]$ .

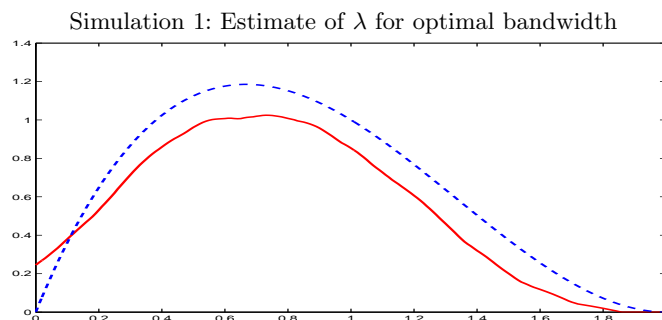
### Simulation 1

We use unimodal hazard function  $\lambda$  on  $[0, t_2]$ :  $\lambda(x) = x(2 - x)^2$ , i.e.  $F(x) = 1 - e^{\frac{x^2}{12}(3x^2 - 16x + 24)}$ . The shape of the hazard function was chosen as for real data – the probability of death first increases and then decreases. Let  $K$  be the Epanechnikov kernel and  $n = 100$

Case A:  $t_1 = 1, t_2 = 2, h_{opt} = 0.4437$

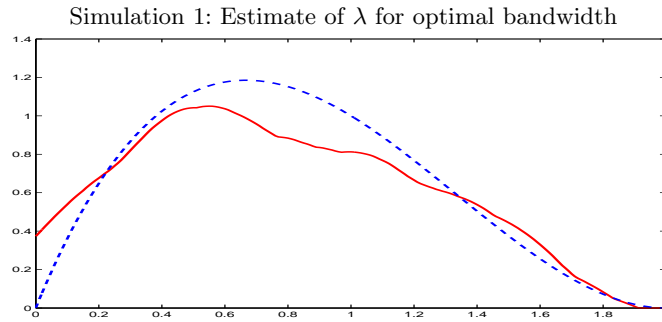
Case B:  $t_1 = 1.5, t_2 = 2, h_{opt} = 0.4721$

Case C:  $t_1 = 2, t_2 = 2, h_{opt} = 0.4993$

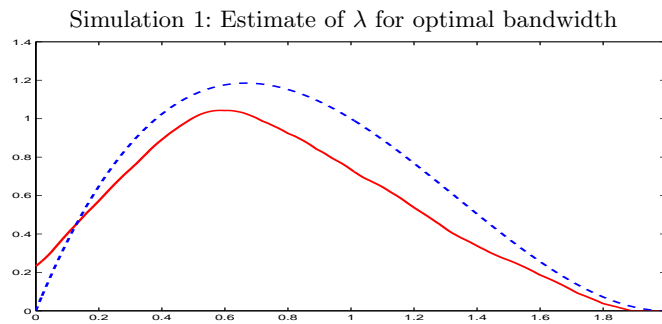


**Fig. 2.** Case A:  $\lambda$  – dashed line, estimate – solid line

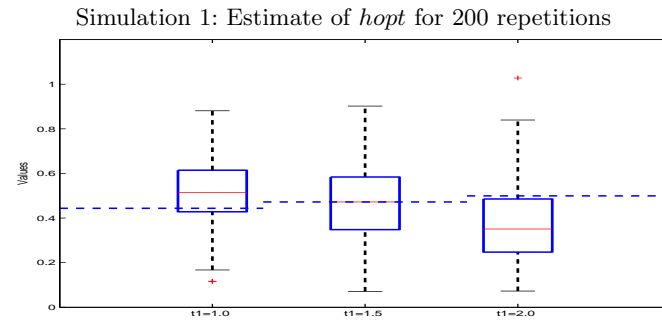




**Fig. 3.** Case B:  $\lambda$  – dashed line, estimate – solid line



**Fig. 4.** Case C:  $\lambda$  – dashed line, estimate – solid line



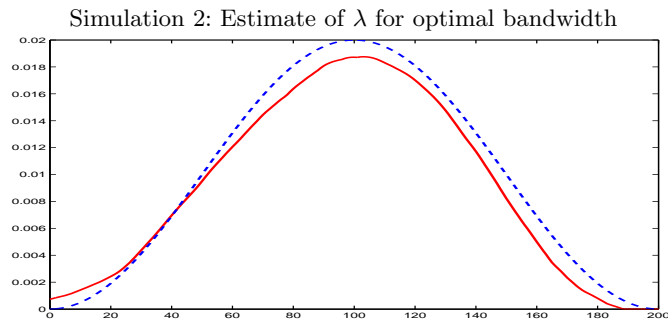
**Fig. 5.** Dashed lines: optimal bandwidths

## Simulation 2

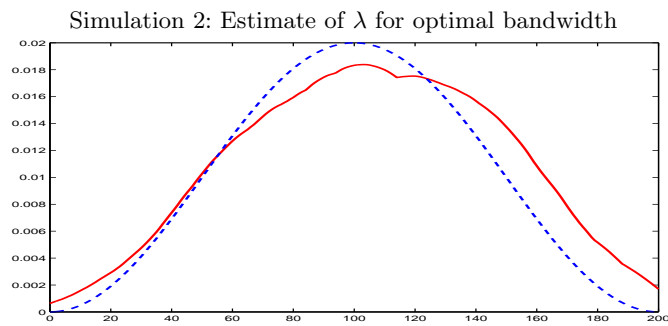
We use unimodal hazard function  $\lambda$  on  $[0, t_2]$ :  $\lambda(x) = \frac{1}{100} \left( 1 - \cos \frac{2\pi}{t_2} x \right)$ , i.e.  $F(x) = 1 - e^{\frac{1}{100} \left( \frac{t_2}{2\pi i} \sin \frac{2\pi i}{t_2} x - x \right)}$ . The shape of the hazard function is similar as in Simulation 1 but we use non-polynomial function. Let  $K$  be the Epanechnikov kernel and  $n = 100$

Case A:  $t_1 = 100$ ,  $t_2 = 200$ ,  $h_{opt} = 43.703$

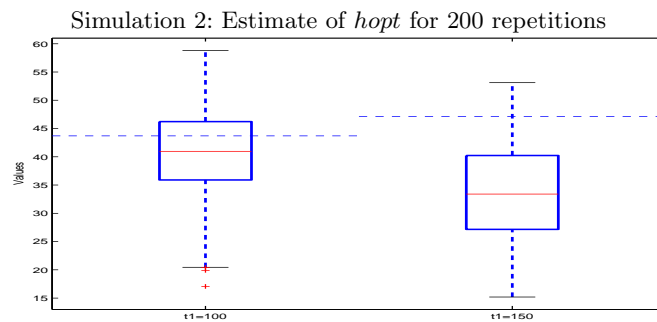
Case B:  $t_1 = 150$ ,  $t_2 = 200$ ,  $h_{opt} = 47.122$



**Fig. 6.** Case A:  $\lambda$  – dashed line, estimate – solid line



**Fig. 7.** Case B:  $\lambda$  – dashed line, estimate – solid line



**Fig. 8.** Dashed lines: optimal bandwidths

### Simulation 3

We use bimodal hazard function  $\lambda$  on  $[0, t_2]$ :  $\lambda(x) = \frac{1}{100} \left( 1 - \cos \frac{4\pi}{t_2} x \right)$ , i.e.  $F(x) = 1 - e^{\frac{1}{100} \left( \frac{t_2}{4\pi} \sin \frac{4\pi}{t_2} x - x \right)}$ . This case is also common for real data – the probability of death increases again after a decline. Let  $K$  be the Epanechnikov kernel and  $n = 200$

Case A:  $t_1 = 100$ ,  $t_2 = 200$ ,  $h_{opt} = 23.443$

Case B:  $t_1 = 150$ ,  $t_2 = 200$ ,  $h_{opt} = 25.255$

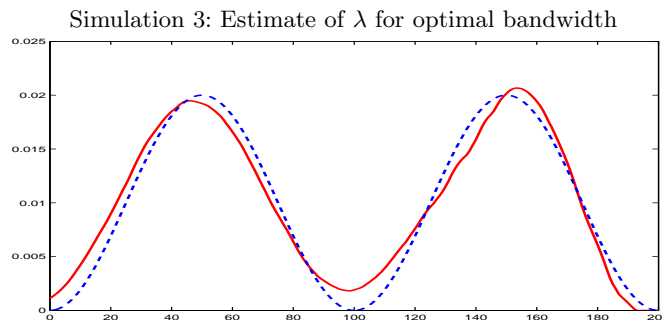


Fig. 9. Case A:  $\lambda$  – dashed line, estimate – solid line

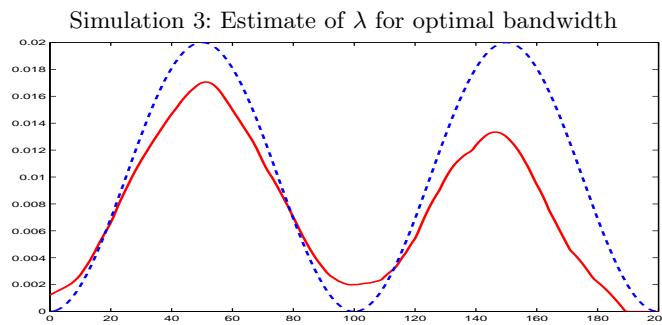


Fig. 10. Case B:  $\lambda$  – dashed line, estimate – solid line

### Conclusion

The simulations indicate that the proposed method of generating random censored data for given cumulative distribution function  $C$  and hazard function  $\lambda$  can be well applied for testing the algorithms of survival analysis.

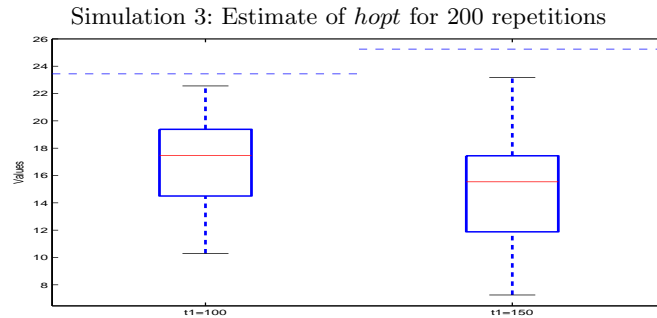


Fig. 11. Dashed lines: optimal bandwidths

At the same time the simulations show that the method of bandwidth choice proposed in Horová & Zelinka (2006) gives worse results for the greater frequency of censored data, but the estimates of optimal bandwidth are still well usable. For more detailed results a larger study would be to compile but that goes beyond this paper.

## Acknowledgement

Research supported by MŠMT of Czech Republic, no. LC06024.

## References

- COLLETT D.: *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC: Boca Raton-London-New York-Washington, D.C., 2003.
- HOROVÁ I., ZELINKA J. and BUDÍKOVÁ M.: Estimates of Hazard Functions for Carcinoma Data Sets. *Environmetrics*, **17**, 239–255, 2006.
- HOROVÁ I. and ZELINKA J.: (2006) Kernel Estimates of Hazard Functions for Biomedical Data Sets. In *Applied Biostatistics: Case studies and Interdisciplinary Methods*, Springer, 2006.
- HOROVÁ I., POSPÍŠIL Z. and ZELINKA J.: Semiparametric Estimation of Hazard Function for Cancer Patients, *Sankhya*, **69**, 494–513, 2008.
- HOROVÁ I., POSPÍŠIL Z. and ZELINKA J.: Hazard function for cancer patients and cancer cell dynamics, *Journal of Theoretical Biology*, **258**, 437–443, 2009.
- MÜLLER H.G. and WANG J.L.: Nonparametric Analysis of Changes in Hazard Rates for Censored Survival Data: An alternative Change-Point Models. *Biometrika*, **77**(2), 305–314, 1990.
- RAMLAU-HANSEN H.: Counting Processes Intensities by Means of Kernel Functions. *The Annals of Statistics*, **11**(2), 453–466, 1983.
- TANNER M.A. and WONG W.H.: The Estimation of the Hazard Function from Randomly Censored Data by the Kernel Method. *The Annals of Statistics*, **11**(3), 989–993, 1983.
- UZUNOGULLARI U. and WANG J.L.: A comparison of Hazard Rate Estimators for Left Truncated and Right Censored Data. *Biometrika*, **79**(2), 297–310, 1992.
- WAND, I.P. and JONES, I.C.: *Kernel smoothing*. Chapman & Hall, London, 1995.

# Index

## A

*Abad Montes, F.*, 1151  
ABC, 57  
acceleration, 1247  
acceptance sampling, 1231  
actuar, 145  
*Adams, N.M.*, 167  
adaptive learning, 831  
adaptive prediction, 189  
*Adelfio, G.*, 625  
affine equivariance, 79  
*Afonso, F.*, 633, 1621  
aggregate models, 145  
*Agostinelli, C.*, 69  
*Agró, G.*, 1557  
*Aguilera, A.M.*, 641  
*Aguilera-Morillo, M.C.*, 641  
*Ahn, S.K.*, 297  
Aitchison geometry, 79  
*Albrecher, H.*, 135  
*Alexandrov, T.*, 649  
*Alfö, M.*, 369  
*Alibrandi, A.*, 657  
alignment, 1605  
alternating least squares algorithm, 1247  
*Álvarez-Verdejo, E.*, 665, 1589  
*Ambroladze, A.*, 231  
amino-acids, 745  
*Ammar, S.*, 673  
*Amouh, T.*, 681  
*Anagnostopoulos, C.*, 167  
*Aneiros, G.*, 689  
anomalies, 33  
*Antoniadis, A.*, 697  
*Aparicio-Pérez, F.*, 705  
application, 1111  
approximate Bayesian computation, 47  
*Arcos, A.*, 665, 1351  
*Arlt, J.*, 713, 761  
*Arltová, M.*, 713, 721, 761  
ARMA, 1215  
ARMAX model, 887  
*Artiles, J.*, 1573

asymmetry, 315  
*Audrino, F.*, 729  
autocovariances, 705  
auxiliary information, 665  
*Aziz, N.*, 381

## B

B-spline expansion, 641  
backfitting, 1517  
*Badez, N.*, 633  
bagging, 729  
*Balduzzi, S.*, 737  
*Balzani, L.*, 737  
bandwidth selection, 509  
bank performance, 1477  
*Bardinet, E.*, 1343  
*Barth, E.*, 327  
*Bartkowiak, A.*, 745  
*Bartošová, J.*, 753, 1007, 1023  
basal ganglia, 1343  
basis pursuit algorithm, 1461  
*Bašta, M.*, 713, 761  
Bayes estimator, 1597  
Bayesian analysis, 277, 315  
Bayesian hierarchical models, 155  
Bayesian inference, 437, 1413  
Bayesian integrated criterion, 1685  
Bayesian local regression, 47  
Bayesian networks, 549  
Bayesian semiparametric models, 1437  
Bayesian statistics, 469  
Beale test, 657  
beanplot, 967  
*Benaglia, T.*, 769  
*Benali, H.*, 1343  
*Benammou, S.*, 777, 785  
*Benner, A.*, 19  
*Bernau, C.*, 793  
Bernstein von-Mises's theorem, 1183  
*Berro, A.*, 89  
*Biernacki, C.*, 801  
bilinear models, 887  
*Billard, L.*, 1621  
*Bína, V.*, 753

- binary diagnostic test, 533
  - Bini, M.*, 809
  - bivariate binary data, 1119
  - bivariate censoring, 1367
  - block matrices, 817
  - blood alcohol concentration, 1693
  - Blum, M.G.B.*, 47
  - Bolla, M.*, 817
  - Bologna, S.*, 825
  - boolean factor analysis, 1047
  - boosting, 1135
  - bootstrap, 199, 509, 1015, 1207, 1343, 1383, 1677
  - Bordes, L.*, 243
  - Bottou, L.*, 177
  - Bougeard, S.*, 389
  - Boulesteix, A.-L.*, 793
  - Bouveyron, C.*, 831
  - Bouzas, P.R.*, 839
  - branching, 1477
  - Bravo, M.C.*, 1063
  - breast disease, 1159
  - Brechmann, E.C.*, 1477
  - Broccoli, S.*, 397
  - Brossat, X.*, 697
  - Bruzzese, D.*, 847
  - Bry, X.*, 405
  - Buckley-James estimators, 384
  - Buckley-James model, 381
  - Budinská, E.*, 1637
- C**
- c-optimal experimental design, 879
  - Cénac, P.*, 421
  - Caballero-Águila, R.*, 855
  - Cabras, S.*, 863
  - calibration weights, 753
  - Cardot, H.*, 413, 421
  - Castellano, R.*, 429
  - categorical variables, 389
  - Cavriani, G.*, 397
  - cellular Potts model, 57
  - censored
    - right, 381
  - censored data, 243, 381
  - censoring, 1717
  - Cerioni, A.*, 871
  - Černý, M.*, 879
  - Ceulemans, E.*, 359
  - change of support, 285
  - change point analysis, 501
  - change-point, 943
  - change-point detection, 557
  - Chaouch, M.*, 421
  - Chauveau, D.*, 243, 769
  - Chauvin, C.*, 389
  - Chen, P.*, 887
  - chi-bar-square distribution, 445
  - Chiodi, M.*, 625
  - Cho, S.*, 297
  - Chuliá, H.*, 1429
  - Chung, Y.-K.*, 437
  - Ciampi A.*, 895
  - circular block bootstrap, 1581
  - circulatory system mortality, 737
  - CLAFIC, 493
  - class prediction, 793
  - classification, 167, 189, 1079, 1143, 1375
  - classification regression tree, 1621
  - climatic factors, 721
  - cluster analysis, 1063, 1533
  - cluster detection, 847, 911
  - cluster number determination, 1533
  - clustering, 265, 349, 697, 1023
  - clustering analysis, 1271
  - clustering of curves, 625
  - clusters number, 657
  - clusterwise regression, 461
  - co-clustering, 369
  - coincidences, 33
  - cointegration, 713
  - Colin R.*, 903
  - Colombi, R.*, 445
  - colon crypt dynamics, 57
  - common and distinctive cluster model, 359
  - comparison of mean curves, 1581
  - Competitiveness, 809
  - complex data, 633
  - component method, 1007
  - composite distributions, 1661
  - compositional data, 1223
  - compound distribution, 145
  - computational statistics, 19
  - computational statistics and data analysis, 19

computer algebra systems, 903  
 concentration, 1501  
 conditional heteroskedasticity, 1055  
 conditional logit, 369  
 conditioning, 1485  
 confidence ellipses, 745  
 conjoint analysis, 777  
 consensus analysis, 1063  
 consensus measure, 1063  
 contagion, 1429  
 contingency tables, 1629  
 control charts, 1207  
 convergence of contingency tables, 817  
 correlated component regression, 1335  
 correlated frailty, 1119  
 correlation, 541  
*Costanzo, G.D.*, 453  
 count data, 1135  
 counting process, 1509  
 Cox process, 839  
 credit default swaps, 429  
 credit risk modelling, 265  
 credit scoring, 167, 1199, 1525  
 cross-validation, 1239  
 cryptography, 879  
 cubic clustering criterion, 657  
*Cubiles de la Vega, M.D.*, 1375  
*Cucala, L.*, 912, 914, 916  
*Jairo Cugliari*, 697  
 cumulative distribution function, 1453  
 customer classification, 1311  
 customer satisfaction, 1311  
*Cuxac, P.*, 1255  
*Czado, C.*, 1477

## D

*D'Alessandro, A.*, 625  
 data dimensionality reduction, 549  
 data integration, 1637  
 data mining, 1311  
 data reconstitution, 1279  
 data visualization, 633, 1669  
 database, 681  
*Davino, C.*, 919  
*De Bartolo, S.*, 453  
*De Carvalho, F.A.T.*, 461, 959, 1271  
*de Falguerolles, A.*, 477  
*de Melo, F.M.*, 1271  
 death rate, 761

*Debruyne M.*, 927  
*Dehon, C.*, 935  
*Dell'Accio, F.*, 453  
*Demeyer, S.*, 469  
 demography, 761, 1151  
 density estimation, 673  
 dependent data., 1095  
 depth function, 1215  
*Derquenne, Ch.*, 943  
 design-based estimation, 413  
*Despeyroux, T.*, 1271  
*Dessertaine, A.*, 413  
 deterministic algorithm, 589  
*Devroye, L.*, 3  
*Di Maso, M.*, 737  
*Di Salvo, F.*, 1557  
 diagnostic analysis, 381  
*Diana, G.*, 951  
 dictionary learning, 327  
*Diday, E.*, 633, 959, 1621  
 difference time series of seawater  
   temperatures, 1191  
 diffusion processes, 1541  
 dimension reduction, 305, 349, 895  
 dimensionality reduction, 337, 1055,  
   1279  
 Dirichlet process, 1303  
 Dirichlet process priors, 1437  
 Dirichlet processes, 277  
 discrete  $k$ -out-of- $n$ : $G$  system, 1565  
 discrete choice, 369  
 discrete optimization, 525  
 discrete parameters, 1685  
 discrete sampling, 1111  
 discriminant analysis, 189, 389  
 discriminative classifiers, 1493  
 disease clustering, 1421  
 disease mapping, 1413  
 distance-based model, 517  
 distribution function, 1351  
 divisive hierarchy tree, 1621  
 DNA code, 745  
 DNA data, 1645  
 documents categorization, 1271  
*Drago, C.*, 967  
 drift term, 209  
 DSA, 1295  
*Duchesnay, E.*, 101

*DuClos, C.*, 285  
*Dyachenko A.*, 895  
 dynamic clustering, 959  
 dynamic specification test, 1287  
 dynamics, 209

**E**

echelon analysis, 1167  
 ecologic regression, 1413  
*Edler, L.*, 19  
 education level, 1007  
 efficiency, 177  
 efficiency measure, 1047  
*Evelyn Eger*, 1079  
*Einbeck, J.*, 1709  
*El-Saied, H.*, 975  
 electrical data, 1391  
 electricity consumption, 413  
 electricity prices, 1055  
 EM algorithm, 565, 769, 801, 991  
 empiric characteristic function, 1375  
 empirical Bayes, 47  
 empirical characteristic function, 501  
 empirical likelihood, 1327  
 empirical mode decomposition, 1391  
 empirical variogram, 1239  
 entropy, 1533  
 environmental justice, 277  
 EOF, 1557  
 epidemiology, 911  
 equivalent models, 1437  
 error bounds, 1485  
 errors in variables, 285  
*Escabias, M.*, 641  
 estimated inclusion probabilities, 951  
 estimating function, 1327  
 EU-SILC, 753  
 European Health database, 737  
 event-study, 429  
 evidence approximation, 47  
 evolutionary algorithms, 1127  
 exogenous factors, 1151  
 exploratory data analysis, 1669  
 exponential distribution, 1597  
 exposure-response relationship, 1319  
 extended Baum Welch, 1493

**F**

*Fabián, Z.*, 983

face recognition, 597  
 factor models, 305  
*Faria, S.*, 991  
 FDA, 1557  
 feature extraction, 1709  
 feature selection, 1079  
*Feinerer, I.*, 999  
*Ferraty, F.*, 689  
*Fiala, T.*, 1007  
*Filipova, K.*, 729  
 filter design, 1501  
*Filzmoser, P.*, 79  
 financial asset allocation, 1263  
 financial portfolio optimization., 265  
 finite mixture model, 1023, 1183  
 finite population, 1351  
*Fiori, A. M.*, 1015  
 first-passage-time location function,  
     1541  
 first-passage-times, 1541  
*Fischer, N.*, 469  
 Fisher Discriminant Analysis, 221  
 Fisher scoring, 1701  
 Fisher's iris data, 1621  
 fMRI, 111, 1079, 1343  
*Forbelská, M.*, 1023  
 forecasting, 967, 1055  
 forgetting factor, 167  
*Fortunato, L.*, 277  
 Fourier series, 525  
 FPCA, 625  
 fractal spectral processes, 1039  
*Francisco-Fernández, M.*, 1031  
 frequentist risk, 863  
*Frías, M.P.*, 1039  
*Fried, R.*, 975  
*Frolov, A.A.*, 1047  
*Frouin, V.*, 101  
*Fujita, T.*, 1167  
 functional data, 189, 453, 641, 689, 697  
 functional linear regression, 199  
 functional networks, 1343  
 functional principal components, 413  
 functional principal components  
     analysis, 199, 839  
*Fung, W.*, 437  
*Fung, W.K.*, 1693  
 fuzzy clustering, 1605



**G**

Gamma distribution, 1661  
 gap filling, 1557  
 GARCH, 1215  
*Garcia-Leal, J.*, 1509  
*García-Martos, C.*, 1055  
*García-Santesmases, J.M.*, 1063  
 gastrointestinal cancer, 1421  
 Gaussian copula, 485  
 Gaussian mixture model, 1493  
 Gaussian process, 209  
 gene expression, 1335  
 generalization prediction, 231  
 generalized linear mixed model, 1693  
 generalized linear model, 1693  
 generalized moments, 983  
 generative model, 1047  
*Genest, Y.*, 633  
 genetic marker dependency modelling, 549  
 genome wide analyses, 101  
 genomics, 793  
*Genuer, R.*, 1079  
 geographical clusters, 1167  
 geometric quantiles, 421  
 geometrical ergodicity, 1183  
 geostatistics, 1239  
*Ghribi, M.*, 1255  
*Giacalone, M.*, 657  
 Gibbs sampler, 887, 1183  
 Gibbs sampling, 469  
*Giebel, S.M.*, 1087  
 gini index, 1525  
 Gini's coefficient of mutability, 1533  
*Giordano, F.*, 1095  
*Giordano, S.*, 445  
 global optimization, 525  
*Gocheva-Ilieva, S.*, 1071  
 Gompertz diffusion process, 1151  
*Gonçalves, F.*, 991  
*González-Manteiga, W.*, 199  
*González, J.J.*, 1573  
*González, S.*, 665  
*González-Carmona, A.*, 1517  
*Goto, M.*, 1453  
*Gotway, C.A.*, 285  
*Goulet, V.*, 145  
*Gozzi, G.*, 1103

*Grady, C.*, 111  
 Granger causality, 445  
 graphical models, 445  
*Grossi, L.*, 1103, 1263  
 group effects, 919  
 grouping structure, 657  
*Gutiérrez, R.*, 1111  
*Gutiérrez-Sánchez, R.*, 1111

**H**

H-matrices, 1485  
*Haas, S.*, 135  
*Haasdonk, B.*, 221  
*Hadj Mbarek, M.*, 189  
*Hand, D.J.*, 33, 167  
 haplotype, 1645  
*Hara, A.*, 1159  
 Hardy-Weinberg equilibrium, 437  
*Hayashi, K.*, 493  
 hazard function, 1717  
 health-related quality of life, 397  
*Helman, K.*, 761  
*Hens, N.*, 1119  
*Hermoso-Carazo, A.*, 855  
*Hernández, C.N.*, 1573  
*Heuchenne, C.*, 509  
 heuristic algorithms, 89  
 hidden forces, 33  
 hierarchical Bayes, 429, 1613  
 hierarchical clustering, 847  
 hierarchical latent class model, 549  
 high dimensional data, 349, 421, 573, 1335  
 high-dimensional integration, 135  
*Hirooka, H.*, 1175  
 histogram, 1669  
 histogram data, 581  
 histopathological diagnosis, 1159  
 history of statistics, 477  
*Hladík, M.*, 879  
*Hochreiter, R.*, 1127  
*Hofer, V.*, 1135  
 Horvitz-Thompson estimator, 413, 951  
*Hoshino, T.*, 1437  
 household incomes, 753, 1023  
*Hron, K.*, 79  
*Hu, Y.-Q.*, 437  
*Hušková, M.*, 501  
 Huber M-estimator, 975

*Hubert, M.*, 589, 1143  
*Huete Morales, M.D.*, 1151  
*Hunter, D.R.*, 769  
*Húsek, D.*, 1047, 1533  
 hypothesis test, 199  
 hypothesis testing, 863

**I**

identifiability, 469  
*Iizuka, M.*, 1247  
*Iliev, I.*, 1071  
 imputation, 485, 1517  
 incomplete data, 485  
 indefinite kernels, 221  
 industrial application, 633  
 inequality, 1015  
 influence function, 1015  
 influential data, 1263  
 information criteria, 297  
 information gain, 1047  
 information value, 1199  
 insertion algorithm, 801  
 inspection by variables, 1231  
 integer factoring, 879  
 integral operator, 135  
 integrated relative lift, 1525  
 International Association of Statistical  
   Computing, 19  
 interval-valued data, 461  
*Irpino, A.*, 581  
 IRT model, 397  
*Ishibashi, Y.*, 1159  
*Ishihara, T.*, 315  
*Ishioka, F.*, 1167  
 italian manufacturing sector, 1103  
 item response modeling, 1361  
*Ito, M.*, 1175

**J**

*Jacques, J.*, 801  
*Jaïdane-Saïdane, M.*, 1391  
*Josserand, E.*, 413

**K**

k-means clustering, 1653  
*Käärik, E.*, 485  
*Käärik, M.*, 485  
*Kamatani, K.*, 1183  
*Kamijo, K.*, 1191

Kaplan-Meier, 1367  
*Karatzoglou, A.*, 999  
*Kearney, G.*, 285  
 kernel density estimation, 769  
 kernel estimate, 1717  
 kernel methods, 221  
 kernel regression, 1095  
 kernel smoothing, 1199  
 kernels, 1501  
*Kerr, J.*, 1477  
*Keszöcze, O.*, 649  
*Kirch, C.*, 501  
*Klufa, J.*, 1231  
*Koláček, J.*, 1199, 1525  
*Komorníková, M.*, 1287  
*Konczak G.*, 1207  
*Kortas, H.*, 785  
*Kosiorowski, D.*, 1215  
*Košmelj, K.*, 1223  
 kriging, 1239  
*Kubota, T.*, 1239  
*Kurihara, K.*, 1159, 1645  
*Kuroda, M.*, 1247

**L**

*Labusch, K.*, 327  
*Lalanne, C.*, 101  
*Lambertini, C.*, 737  
*Lamirel, J.-C.*, 1255  
*Lang, S.*, 155  
*Langhamrová, J.*, 713, 721  
*Larabi Marie-Sainte, S.*, 89  
 large and complex data, 1653  
 large scale, 999  
 LARS, 69, 1405  
 LASSO, 927  
 last observation carried forward, 1175  
 latent class analysis, 1335  
 latent variables, 469  
*Laurent, G.*, 509  
*Laurini, F.*, 1263  
 LDA, 597  
 learning vector quantization, 1159  
 least-squares quadratic estimation, 855  
*Lebarbier, E.*, 557  
*Lechevallier, Y.*, 895, 1271  
*Lee, J.A.*, 337  
*Lee, P.H.*, 517  
 Lee-Carter method, 713

- Lehéricy, S.*, 1343  
*Lelu, A.*, 1279  
*Lenčuchová, J.*, 1287  
*Leray, P.*, 549, 673  
 leverage effect, 315  
*Li, S.*, 597  
*Li, W.*, 1295  
*Li, X.*, 1031  
*Lian, H.*, 1303  
 lifetime data, 253  
 lift, 1525  
 likelihood methods, 1327  
 likelihood ratio, 437  
 likelihood ratio test, 991  
*Lim, E.W.C.*, 525  
*Linares-Pérez, J.*, 855  
 Lindley approximation, 1597  
 linear discriminant analysis, 1677  
 linear mixed models, 1477  
 linear regression, 1405  
 linear smoothing, 1175  
 local asymptotic normality, 1183  
 local influence, 381  
 local polynomial regression, 1031  
 local standard fractal dimension, 1191  
 logitboost, 1573  
 logratio transformations, 79  
*Lombardo, R.*, 1311  
 long-range dependence parameters, 1039  
 longitudinal data, 1319, 1581  
*Lopiano, K.K.*, 285  
*Löster, T.*, 1533  
 low-pass filters, 1501  
 LTPD plans, 1231  
*Lu, X.*, 1319  
*Luati, A.*, 1501  
*Luna del Castillo, J.D.*, 533  
*Lunardon, N.*, 1327  
*Luzio, D.*, 625
- M**
- Müller, H.G.*, 209  
*Macq, B.*, 681  
 macroeconomic variables, 729  
 Made in Italy, 809  
*Magidson, J.*, 1335  
 Mahalanobis distance, 221  
*Malherbe, C.*, 1343  
 manifold learning, 337  
*Marcucci, E.*, 1613  
*Marek, L.*, 1231  
 marginal likelihood, 1437  
 marginal posterior distribution, 863  
 Markov assumptions, 1287  
 Markov basis, 1629  
 Markov chain Monte Carlo, 315, 1303  
 Markov switching models, 429, 1287  
*Márquez, M.D.*, 1429  
*Marrelec, G.*, 1343  
*Martínez-Calvo, A.*, 199  
*Martinetz, T.*, 327  
*Martínez, H.*, 1351  
*Martínez, S.*, 1351  
*Martínez-Miranda, M.D.*, 1517  
 martingales, 1509  
 mass spectrometry, 649  
 matching pursuit, 327  
*Matei, A.*, 1361  
 mathStatistica, 903  
 maximum likelihood estimation, 1701  
 maximum margin learning, 1493  
 MCMC, 277, 1597  
 mean shape, 1087  
*Meintanis, S.G.*, 501  
*Meira-Machado, L.*, 1367  
*Menjoge, R.S.*, 1383  
*Messé, A.*, 1343  
 metric projection, 1605  
*Mhamdi, F.*, 1391  
*Michel, V.*, 1079  
 microarray data, 1637  
 microarrays, 895  
*Minami, H.*, 493  
 minimax, 785  
 missing data, 1557  
 missing values, 777  
*Misumi, T.*, 1175  
*Mittnik, S.*, 541  
*Miwa, T.*, 1399  
 mixed effects models, 1023  
 mixed logit, 1613  
 mixed stain, 437  
 mixture models, 243, 277, 359, 517  
 mixture of distributions, 1613  
 mixture of trees, 673  
 mixture Poisson regression models, 991

- Miyazaki, K., 1437  
 Mizuta, M., 493  
 Mkhadri, A., 1405  
 model selection, 69, 231, 297, 831, 1685  
 model-calibration approach, 1351  
 modified areal unit problem, 285  
 Mohebbi, M., 1413, 1421  
 molecular biomedical research, 19  
 Molitor, J., 277  
 Molitor, N.-T., 277  
 Monte Carlo, 1207  
 Monte Carlo methods, 3, 1421  
 Monte Carlo simulation., 1597  
 Montero Alonso, M.A., 533  
 Moreira, A., 1367  
 Mori, Y., 1247  
 mortality, 713  
 Mourad, R., 549  
 moving average, 1207  
 multi-core, 1295  
 multi-move sampler, 315  
 multi-objective optimization, 265  
 multiblock *PLS*, 389  
 multiblock redundancy analysis, 389  
 multilevel models, 155, 737, 809  
 multinomial logit, 369  
 multiple change point detection, 1461  
 multiple comparisons, 1399  
 multiple correspondence analysis, 1311  
 multiple sensors, 855  
 multiplicative cascade, 453  
 multivariate exponential distributions, 825  
 multivariate mixture, 769  
 multivariate outliers detection, 89  
 multivariate statistical methods, 79  
 multivariate stochastic volatility, 315  
 multivariate type I generalized logistic distributions, 825  
 multivariate volatility, 305, 1429  
 Muñoz, J.F., 665, 1351  
 Muñoz, M.P., 1429
- N**
- Nafidi, A., 1111  
 Nagakubo, T., 1453  
 NAIRU, 123  
 Nakano, J., 1445  
 Navarrete-Alvarez, E., 1509  
 negative binomial, 745  
 neighborhood property, 1629  
 nested cross-validation, 793  
 Neubauer, J., 1461  
 neural gas, 327  
 neural network, 1159  
 neuroimaging, 101, 565  
 New, J.R., 525  
 Ng, P., 1477  
 Niglio, M., 1469  
 Noirhomme-Fraiture, M., 681  
 nonlinear, 1319  
 nonlinear mixed model, 1693  
 nonlinear projection, 337  
 nonlinear regression, 1071  
 nonlinear time series, 887  
 nonparametric, 297  
 nonparametric estimation, 1367  
 nonparametric mixture, 769  
 nonparametric regression, 509  
 nonresponse, 1361  
 normal distribution function, 1399  
 novelty detection, 831  
 number of clusters, 895  
 numerical accuracy, 1295  
 numerical algorithms, 1485
- O**
- Oder, A., 111  
 Okayasu, I., 1159  
 OLAP, 1445  
 Omori, J., 315  
 one-compartment model, 1693  
 online advertising, 1223  
 online estimation algorithm, 421  
 online learning, 177  
 open source, 1549  
 operational risk, 541  
 Opsomer, J., 1031  
 ordinary least squares, 581  
 ORF length, 745  
 Ouhourane, M., 1405  
 outlier patches, 887  
 outliers, 589, 871, 975  
 outlyingness, 927, 1143  
 output gap, 123  
 output laser power, 1071  
 overlapping clustering, 959  
 overparametrized model, 1461

**P**

P-splines, 155, 641  
*Pekalska, E.*, 221  
 PAC Bayes Bound, 231  
*Pacifico, L.D.S.*, 959  
 pairwise likelihood, 1327  
*Pan, J.*, 305  
 panel data, 1103  
 parameter estimation, 745  
 parallel coordinate plot, 1445  
 parallel coordinates, 1669  
 parallelization, 1295  
 parameter tuning, 793  
 parameter uncertainty, 123  
 parameters initialization, 1469  
 Pareto distribution, 1661  
*Parrado-Hernández, E.*, 231  
*Parrella, M.L.*, 1095  
 parsimonious models, 1685  
 partial least squares, 101  
 particle swarm optimization, 89, 525  
*Paterlini, S.*, 265, 541  
 path modeling, 405  
*Paula R. Bouzas*, 1565  
*Pauli, F.*, 1327  
 PCA factors, 1071  
 peak picking, 649  
*Pélegrini-Issac, M.*, 1343  
*Peña, J. M.*, 1485  
 penalized discriminant analysis, 111  
*Perlberg, V.*, 1343  
 permutation tests, 847, 1207  
*Pernkopf, F.*, 1493  
*Perri, P.F.*, 951  
 Perturb and Combine, 673  
 perturbation, 493  
*Petríková, A.*, 1287  
 phase-type distribution, 1565  
*Pino, R.*, 1375  
*Plaia, A.*, 1557  
 PLS, 405  
 PLS regression, 785  
*Poggi, J.-M.*, 697, 1391  
 Poisson distribution, 1421  
 Poisson regression, 1413  
*Poline, J.-B.*, 101, 565  
*Polonik, W.*, 305  
*Polyakov, P.Y.*, 1047

polyhedral cones, 1399  
 polyhedron, 1399  
 polynomial matrices, 705  
 polynomial transformations, 1311  
 population, 1151  
 population drift, 167  
 population projection, 1007  
 population-based algorithms, 265  
 portfolio optimization, 1127  
 posterior distribution of change-points, 557  
 poverty, 753  
 power-law distribution, 1279  
 pps sampling, 951  
*Preda, C.*, 189  
 prediction, 111  
 prediction intervals, 123  
 prediction with principal components, 839  
 predictive values, 533  
 principal component analysis, 573, 641, 1247, 1709  
 probabilistic constraints, 1127  
 probability distribution function, 581  
 product-limit estimator, 381  
 productivity growth, 1103  
*Proietti, T.*, 1501  
 projection pursuit, 89  
 pseudo-likelihood, 863, 1327  
 pyramidal clustering, 959

**Q**

Q-ML estimators, 1469  
*Qannari, E.M.*, 389  
 quality indexes, 1199, 1525  
 quantile plot, 871  
 quantile regression, 919  
 quantiles, 1199  
*Queiroz, D.N.*, 461  
*Quesada-Rubio, J.-M.*, 1509  
 quiz data, 801

**R**

R, 145, 1375  
 R commander plug-ins, 1549  
 R package compositions, 1223  
*Racugno, W.*, 863  
*Ramos-Ábalos, E.M.*, 1111  
 random forests, 1079

- random process, 1191
  - random taste heterogeneity, 1613
  - random variate generation, 3
  - randomization test, 1279
  - rank data, 801
  - ranked list, 1637
  - ranking data, 517
  - ratio and difference estimators, 665
  - Raya-Miranda, R.*, 1517
  - Redont, P.*, 405
  - reduced k-means, 359
  - regression, 477
  - reinsurance, 135
  - relatedness coefficient, 437
  - relational data, 1271
  - relative, 437
  - relative effect, 1453
  - reliability, 243, 1565
  - renal tumours, 1087
  - repeated measures, 1453
  - repeated measures mixed-effects model, 1175
  - reproducibility, 111
  - resampling, 991, 1581
  - restarting criteria, 1247
  - reweighted MCD, 871
  - Řezáč, M.*, 1199, 1525
  - Řezanková, H.*, 1533
  - Richardson, S.*, 277
  - Rigall, G.*, 557
  - right censoring, 509
  - right kurtosis, 1015
  - risk capital, 541
  - risk theory, 145
  - Robbins-Monro, 421
  - Robin, S.*, 557
  - robust distances, 871
  - robust estimation, 1215
  - robust estimators, 983, 1263
  - robust regression, 69
  - robust statistics, 927
  - robustness, 69, 79, 421, 573, 589, 975, 1143, 1677
  - Rodríguez, A.F.*, 123
  - Rodríguez, J.*, 1055
  - Roldán Nofuentes, J.A.*, 533
  - Román-Román P.*, 1541
  - Rosadi, D.*, 1549
  - Rosales-Moreno, M.-J.*, 1509
  - Rousseeuw, P.J.*, 589
  - RPCR, 589
  - RSA, 879
  - RSIMPLS, 589
  - Rueda, M.*, 665, 1351, 1589
  - Ruggieri, M.*, 1557
  - ruin probability, 135
  - Ruiz, E.*, 123
  - Ruiz-Castro, J.E.*, 839, 1565
  - Ruiz-Fuentes N.*, 839
  - Ruiz-Gazen, A.*, 89
  - Ruiz-Medina, M.D.*, 1039
- S**
- S-estimators, 69
  - Saavedra, P.*, 1573
  - Sadeghi, S.*, 1597
  - Sakakihara, M.*, 1247
  - Sakata, T.*, 1629
  - Sakurai, H.*, 1581
  - Salibian-Barrera, M.*, 69
  - Sánchez, M.J.*, 1055
  - Sánchez-Borrego, I. R.*, 1589
  - Santana, A.*, 1573
  - Saporta, G.*, 189, 461, 469, 777
  - Sato-Ilic, M.*, 1605
  - Scaccia, L.*, 429, 1613
  - scan statistics, 911
  - scatterplot, 1383, 1669
  - Scepi, G.*, 967
  - Schenk, J.-P.*, 1087
  - Schiffler, S.*, 649
  - Schiltz, J.*, 1087
  - Schimek, M.G.*, 1637
  - Schwarz's Bayesian information criterion, 1533
  - score moments, 983
  - seasonal cointegrating rank, 297
  - Seck, D.*, 1621
  - SEER, 405
  - segmentation, 943
  - selection bias, 509
  - SEM, 405
  - semi-parametric estimation, 1613
  - semi-parametric regression, 689
  - sensitivity analysis, 493
  - Seong, B.*, 297
  - sequential conditional test, 1629

- sequential sampling, 1303
  - Serrano-Pérez J.J.*, 1541
  - Shawe-Taylor, J.*, 231
  - Shigemasa, K.*, 1437
  - sICA, 1343
  - sign-regular matrices, 1485
  - similarity-based embedding, 337
  - Simon, S.*, 1295
  - simulation, 3, 525, 951, 1717
  - simulations, 19
  - simultaneously comparison, 533
  - Sinoquet, C.*, 549
  - skew normal distribution, 397
  - Skočdoplová, V.*, 879
  - smoothing, 1031, 1709
  - smoothing parameter, 1517
  - social-economic factors, 721
  - software Mathematica, 1231
  - sorting process., 801
  - Sottoriva, A.*, 57
  - Souissi, B.*, 777
  - sparse coding, 327, 649
  - sparse parameter estimation, 1461
  - sparsity prior, 1303
  - spatial autocorrelation, 1421
  - spatial correlation, 1413
  - spatial data, 1167
  - spatial data analysis, 1645
  - spatial heterogeneity, 155
  - spatial marked point processes, 911
  - spatial median, 421
  - spatial models, 565
  - spatial scan statistic, 1167
  - spatiotemporal parametric models, 1039
  - SPC, 253
  - spectrum and stability, 817
  - Spring, R.*, 111
  - state space models, 123
  - stationary signals, 1573
  - statistical computing, 477, 1485
  - statistical graphics, 1637
  - statistical inference, 1661
  - statistical inference in diffusion process, 1111
  - statistical population density, 1669
  - statistical quality control, 1191
  - statistical shape analysis, 1087
  - statistical testing, 565
  - Steinhorst, K.*, 649
  - stem cell modeling, 57
  - stochastic EM algorithm, 243
  - stochastic gradient averaging, 421
  - stochastic gradient descent, 177
  - stochastic optimization, 1127
  - stochastic orderings, 445
  - stochastic process, 453
  - stock returns., 315
  - Strother, S.*, 111
  - structural equation modeling, 469
  - structural modelling, 1319
  - suicide, 721
  - Sumi, T.*, 1629
  - summary indexes, 1709
  - supervised classification, 389, 831
  - Support Vector Machines, 231
  - survey sampling, 1361
  - survival analysis, 1509
  - SVM, 999, 1375
  - symbolic data, 1605
  - symbolic data analysis, 461, 633, 967, 1445, 1653
  - symbolic data management, 681
  - symbolic methods, 903
  - symbolic objects, 1063
  - Szustalewicz, A.*, 745
- T**
- T<sup>2</sup>hotelling, 657
  - t-SNE, 337
  - tagging SNP, 1645
  - Taguri, M.*, 1581
  - tail dependence, 541
  - Takala, E-P.*, 1319
  - Tam, T.W.M.*, 1693
  - Tarumi, T.*, 1239
  - Tasoulis, D.K.*, 167
  - Tavaré, S.*, 57
  - temperature, 761
  - Tenenhaus, A.*, 101
  - Teodorescu, S.*, 1661
  - Terada, Y.*, 1653
  - testable contingency table parameters, 817
  - testing non-linearity, 1287
  - testing remaining non-linearity, 1287
  - text mining, 999

THEME, 405  
*Thirion, B.*, 101, 565, 1079  
 threshold model, 1469  
 threshold regimes, 729  
 thresholding, 785  
 time series, 705, 721, 761, 943, 975  
 time series analysis, 1549  
 time series factor analysis and dynamical conditional correlation, 1429  
 time-varying coefficient model, 1175  
*Timmerman, M.E.*, 359  
*Tomita, M.*, 1167, 1645  
 top-k list, 1637  
*Torres-Ruiz F.*, 1541  
*Toschi, E.*, 737  
 tree-structured models, 729  
 trend extraction, 1391  
 Tribes algorithm, 89  
 trimming, , 573  
 triple of ensembles, 1135  
*Trombetta, G.*, 453  
*Tucholka, A.*, 565  
 Tukey M-estimator, 975  
 two-dimensional data, 597  
 two-per-stratum variance approximation, 1031  
 two-way clustering, 895  
 type I generalized logistic distribution, 825  
 Type II interval censored data, 1119

## U

*Umlauf, N.*, 155  
 uncertain observations, 855  
 uncertainty visualization, 1383  
 univariate discrete distributions, 1701  
 unobserved classes, 831

## V

*Vakili, K.*, 935  
*Valderrama, M.J.*, 641  
*Valois J.-P.*, 1669  
 value-at-risk, 541  
*Van Aelst, S.*, 573, 1677  
*Van der Veeken, S.*, 1143  
*Van Deun, K.*, 349  
*Van Mechelen, I.*, 349  
*Vandervieren, E.*, 573

*Vandewalle, V.*, 1685  
 variable bandwidth selection, 1095  
 variable importance, 1079  
 variable selection, 689, 793, 927, 1335, 1405, 1677  
 variance comparison, 1087  
 variance components, 943  
 vector  $\varepsilon$  algorithm, 1247  
 vector generalized linear and additive models, 1701  
*Velucchi, M.*, 809  
*Ventura, L.*, 863, 1327  
*Verde, R.*, 581  
*Verdonck, T.*, 589  
*Verleysen, M.*, 337  
*Vernic, R.*, 1661  
*Verron, T.*, 405  
*Vesely, V.*, 1461  
 vgam package for R, 1701  
*Vicari, D.*, 369  
*Vieu, P.*, 689  
 VISA algorithm, 1405  
*Vistocco, D.*, 847, 919  
*Vitale, C.D.*, 1469  
 volterra integral equations, 1541

## W

*Wang, D.Q.*, 381  
 warranty, 253  
 Wasserstein distance, 581  
 waveforms, 625  
 wavelet thresholded transform, 1039  
 wavelet transformation, 1159  
 wavelet-PLS, 785  
 wavelets, 697, 761  
 wearout, 253  
*Wehenkel, L.*, 673  
 weighted regression, 777  
*Welsch, R.E.*, 1383  
*Werft, W.*, 19  
 Wiener-Kolmogorov filter, 705  
*Wienke, A.*, 1119  
 wild bootstrap, 1517  
*Willems, G.*, 573, 1677  
*Wohlmayr, M.*, 1493  
*Wolfe, R.*, 1413, 1421  
*Wunder, C.*, 19



**Y**

*Yadohisa, H.*, 1653  
*Yamamoto, Y.*, 1445  
*Yamanouchi, A.*, 1191  
*Yang, C.T.*, 1693  
*Yao, Q.*, 305  
*Yashchin, E.*, 253  
*Yee, T.W.*, 1701  
*Yelnik, J.*, 1343  
*Yener, T.*, 541  
 Yeo-Johnson transformation, 1071

yield curve modeling and forecasting,  
     729

*Young, L.J.*, 285  
*Yu, P.L.H.*, 517, 597

**Z**

*Žabkar, V.*, 1223  
*Zayed, M.*, 1709  
*Zelinka, J.*, 1717  
*Zhao, J.*, 597  
*Zhu, D.-G.*, 437  
*Zied, K.*, 785  
*Zouhaier, D.*, 785

