

Preface

This book concerns weighted correlation and applications involving rankings and preferences. For instance, recommendation of data analysis tools, stock trading support, information retrieval, meta-learning, and recommender systems. Instead of using Spearman's rank correlation coefficient, which is not suitable in some applications, some weighted correlation coefficients will be presented and analyzed in the book and will then be used in a number of applications. The first of these coefficients, r_W , was proposed by us in 2001 [61, 63, 93] and it weighs the distance between two ranks using a linear function of those ranks, giving more importance to higher ranks than lower ones. The statistical distribution of r_W together with some motivating applications will be the subject of Chap. 2.

In Chap. 3 another weighted rank correlation coefficient, r_{W2} , introduced in [74] and applied in a bioinformatics context in [73], will be presented. This coefficient is the second of its series, following the coefficient r_W which was motivated by a machine learning problem concerning the recommendation of learning algorithms. Unlike Spearman's r_S , which treats all ranks equally, these coefficients weigh the distance between two ranks using a linear function of those ranks in the case of r_W and a quadratic function in the case of r_{W2} . The presence of ties, which can happen naturally in some of the applications, will also be taken into consideration in Chap. 3, together with a simulation study to compare the three coefficients r_{W2} , r_W and r_S .

In Chap. 4 we describe in the first part the new developments in weighted Principal Component Analysis (PCA) [42] and in the second part a new method to select variables. The focus is on problems where the values taken by each variable do not all have the same importance and where the data may be contaminated with noise and contain outliers, as is the case with microarray data. This kind of data, which contains the expression levels of a large number of genes (variables), measured simultaneously, for a relatively much smaller number of tissue samples, presents many statistical challenges. There we propose the use of a weighted correlation coefficient, as an alternative to Pearsons, leading thus to a so-called weighted PCA (WPCA1). Then, we apply WPCA1 to the problem of analysing gene expression datasets. In the second part of Chap. 4 we propose a new

PCA-based algorithm to iteratively select the most important genes in a microarray dataset. We show that this algorithm produces better results when WPCA1 is used instead of the usual PCA. We also show that this algorithm used together with support vector machines can compete with the significance analysis of microarrays (SAM) supervised algorithm [97, 98].

Another weighted Principal Component Analysis (WPCA2) for time series data, is presented in Chap. 5. First, in some situations the number of observations in each series is too large and so it is of paramount importance to be able to compress the series, thus reducing its dimension. Second, in a time series context, it is frequent that some observation times are more important than others and the usual PCA cannot take this into account. Thus, a weighted PCA specific for time series data, which was introduced in [70], is described in this chapter and then applied to well-known datasets.

In Chap. 6 we will describe a method for the weighted clustering of time series. This method does not give the same importance to all the observations; instead, it lets the most important observations, for instance the most recent, have a larger weight. A fundamental problem in the clustering of time series is the choice of a relevant metric, and in this chapter, we will use a metric, based on Pearson correlation coefficient, which uses the notion of weighted mean and weighted covariance. We present also some motivating applications.

Finally, we thank everyone who has collaborated with us in the subject of weighted correlation and applications and in particular Pavel Brazdil, Carlos Soares, and Luís Roque. We thank also the editor Eva Hiripi.

Porto
January 2015

Joaquim Pinto da Costa

Rankings and Preferences

New Results in Weighted Correlation and Weighted
Principal Component Analysis with Applications

Pinto da Costa, J.

2015, X, 91 p. 12 illus., 4 illus. in color., Softcover

ISBN: 978-3-662-48343-5