

# Joint Correspondence Analysis Versus Multiple Correspondence Analysis: A Solution to an Undetected Problem

Sergio Camiz and Gastão Coelho Gomes

**Abstract** The problem of the proper dimension of the solution of a Multiple Correspondence Analysis (*MCA*) is discussed, based on both the re-evaluation of the explained inertia *sensu* Benzécri (Les Cahiers de l'Analyse des Données 4:377–379, 1979) and Greenacre (Multiple correspondence analysis and related methods, Chapman and Hall (Kluwer), Dordrecht, 2006) and a test proposed by Ben Ammou and Saporta (Revue de Statistique Appliquée 46:21–35, 1998). This leads to the consideration of a better reconstruction of the off-diagonal sub-tables of the Burt's table crossing the nominal characters taken into account. Thus, Greenacre (Biometrika 75:457–467, 1988) Joint Correspondence Analysis (*JCA*) is introduced, the results obtained on an application are shown, and the quality of reconstruction of both *MCA* and *JCA* solutions are compared to that of a series of Simple Correspondence Analyses run on the whole set of two-way tables. It results that *JCA*'s reduced-dimensional reconstruction is much better than the *MCA*'s one, that reveals highly biased and non-monotone, but also than the *MCA*'s re-evaluation, as suggested by Greenacre (Multiple correspondence analysis and related methods, Chapman and Hall (Kluwer), Dordrecht, 2006).

## 1 Introduction

The identification of the dimension of a data table under study is a crucial issue in most multidimensional scaling techniques, in particular in the linear methods, since most of the analyses that follow the scaling depend on this choice. To quote

---

S. Camiz (✉)

Sapienza Università di Roma, Rome, Italy

e-mail: [sergio.camiz@uniroma1.it](mailto:sergio.camiz@uniroma1.it)

G.C. Gomes

Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

e-mail: [gastao@im.ufrj.br](mailto:gastao@im.ufrj.br)

only some, the number of factors to be interpreted, those on which to attempt a classification, the dimension in which to search for a non-linear solution or for a factor analysis, etc.

In this paper, we focus on this problem in Multiple Correspondence Analysis (*MCA*, [Benzécri et al., 1973–1982](#); [Greenacre, 1984](#)), in particular considering its alternative, the Joint Correspondence Analysis (*JCA*, [Greenacre, 1988](#)), whose solution depends on an a priori selected dimensionality, and on the partial reconstruction of the original data that results by the application of reconstruction formulas.

The application of these methods to a small example taken from a recent study ([Camiz and Gomes, 2009](#)) will show unexpected results when comparing the reconstruction: even if *JCA* was supposed to perform better, the results of *MCA*, in comparison with those of *JCA*, would seriously get questionable its use, unless without some adjustments. Indeed, the application to the Burt's table of the chi-square metrics, and the following correspondence analysis, biases the results, by improving the reconstruction of the diagonal blocks while raising the bias of the off-diagonal ones that contain the most interesting information.

## 2 Theoretical Framework

In exploratory multidimensional scaling the identification of the proper dimension of the solution is strictly tied to the crucial distinction between relevant and non-relevant information, something similar to the identification of errors in classical statistics, but not the same. For metric scaling, the percentage of explained inertia is usually taken as a measure of information, also tied to its interpretability. Thus, taking into account a large share of inertia is the most often used rule of thumb, but without a good theoretical grounding. Indeed, in literature stopping rules may be found: for Principal Component Analysis, [Jackson \(1993\)](#) compared some of the existing ones. For Simple Correspondence Analysis (*SCA*, [Benzécri et al., 1973–1982](#); [Greenacre, 1984](#)) a classical test for goodness of fit ([Kendall and Stuart, 1961](#)) may be applied as approximated by the [Malinvaud \(1987\)](#) test (see also [Saporta and Tambrea, 1993](#)):

$$\tilde{Q}_\alpha = \sum_{ij} \frac{(n_{ij} - \tilde{n}_{\alpha ij})^2}{n r_i c_j} = \chi^2 - \sum_{\beta=1}^{\alpha} \chi_\beta^2 = \sum_{\gamma=\alpha+1}^{\min(r,c)-1} \lambda_\gamma,$$

where  $\tilde{n}_{\alpha ij}$  is the cell value estimated by the  $\alpha$ -dimensional solution.  $\tilde{Q}_\alpha$ , asymptotically chi-square distributed with  $(r - \alpha - 1) \times (c - \alpha - 1)$  degrees of freedom, tests the independence of the residuals in respect to the  $\alpha$ -dimensional representation. This is possible because the eigenvalues of *SCA* sum, up to the grand total, to the table chi-square, namely

$$\chi^2 = n \sum_{\alpha=1}^{\min(r,c)-1} \lambda_\alpha = \sum_{\alpha=1}^{\min(r,c)-1} \chi_\alpha^2.$$

## 2.1 Multiple Correspondence Analysis

It is well known that *MCA* is but a generalization of *SCA* and it is based on *SCA* of either the indicator matrix  $Z$ , gathering all characters involved, or the Burt's table  $B = Z'Z$ , that includes the diagonal tables with the marginals. The eigenvectors of both  $Z$  and  $B$  are the same, whereas the  $B$ 's eigenvalues are the squares of  $Z$ 's (also called  $B$ 's singular values):  $\mu_\alpha^2 = \nu_\alpha$ . As *SCA*, it may be shown that, given a Burt matrix  $B$ , *MCA* may be defined as the weighted least-squares approximation of  $B$  by another matrix  $H$  of lower rank, that minimizes

$$n^{-1} Q^{-2} \text{trace} \left( D_r^{-1} (B - H) D_r^{-1} (B - H)' \right).$$

that is, considering the subtables of  $B$ , that minimizes

$$n^{-1} \sum_{i=1}^Q \sum_{j=1}^Q \|N_{ij} - H_{ij}\|_{ij}^2. \quad (1)$$

where the norm  $\|A_{ij}\|_{ij}^2 = \text{trace} \left( D_i^{-1} A_{ij} D_j^{-1} A_{ij}' \right)$  is the usual chi-square. Indeed, in *SCA* this is limited to only one table.

In *MCA* the identification of the dimensionality is particularly difficult: indeed, for  $B$ , crossing  $Q$  characters with  $J = \sum_{i=1}^Q l_i$  pooled levels (with  $l_i$  the number of levels of the  $i$ -th character) a statistic may again be calculated as if it were a contingency table

$$\chi_B^2 = 2 \sum_{i=2}^Q \sum_{j=1}^{i-1} \chi_{ij}^2 + n(J - Q), \quad (2)$$

where  $\chi_{ij}^2$  is the chi-squared statistic for the off-diagonal subtable  $N_{ij} = Z_i' Z_j$ , and  $n(J - Q)$  is that of the diagonal subtables. As  $\chi_B^2$  is not chi-square distributed, no test is possible. Thus, the current users refer to the total inertia of  $Z$ :  $I_z = \frac{J-Q}{Q}$ , and consider its share explained by the highest level eigenvectors, although it is very low, due to their high number of pooled levels. In practice, they are satisfied when the first factors are enough larger than the following, regardless of the figures involved, as it is generally admitted that the explained inertia is “highly underestimated”. This underestimation was raised by [Benzécri \(1979\)](#) argued by the arbitrary number of levels and by the relation between the eigenvalues issued by either *SCA* or *MCA* of  $Z$  applied on a two characters table: the relation  $\mu_\alpha = \frac{1 \pm \sqrt{\lambda_\alpha}}{2}$  is thus interpreted to limit attention to the eigenvalues larger than the trivial average  $\frac{1}{2}$ , the smaller considered as “artifacts”. This argument is generalized to consider in *MCA* only the eigenvalues larger than their mean, that is  $\mu \geq \bar{\mu}_\alpha = \frac{1}{Q}$ . As a consequence, each factor inertia is re-evaluated as the average deviation from the mean eigenvalue, according to the formula

$$\rho(\mu_\alpha) = \left( \frac{Q}{Q-1} \right)^2 (\mu_\alpha - \bar{\mu})^2, \quad \mu_\alpha \geq \bar{\mu} = \frac{1}{Q}. \quad (3)$$

and its share of total inertia is based on the inertias sum, thus taking the ratio  $\frac{\rho(\mu_\alpha)}{\sum_{\alpha > \frac{1}{Q}} \rho(\mu_\alpha)}$ . [Greenacre \(1988, 2006\)](#) too suggests to re-evaluate the inertia according to (3), but compares each one to the total off-diagonal inertia of the table, that is

$$\frac{Q}{Q-1} \left( \sum_{\mu_\alpha} \mu_\alpha^2 - \frac{J-Q}{Q^2} \right),$$

a share that results always lower than Benzécri's one.

Regardless of the re-evaluation, to decide the number of factors to take into account, the only test currently available is proposed by [Ben Ammou and Saporta \(1998\)](#), based on the distribution of the average eigenvalue under the null hypothesis of independence: its expected variance is

$$\sigma^2 = E[S_\lambda^2] = \frac{1}{n_{..} Q^2 (J-Q)} \sum_{i \neq j} (l_i - 1)(l_j - 1),$$

so that one may assume for  $\frac{1}{Q}$  the confidence interval at 95% level  $\frac{1}{Q} \pm 2\sigma$ . Indeed, since the kurtosis is lower than for a normal distribution, the actual proportion is larger than 95%.

## 2.2 Joint Correspondence Analysis

In order to remove the bias due to the diagonal submatrices, [Greenacre \(1988\)](#) proposes the *Joint Correspondence Analysis (JCA)* as a better generalization of *SCA*. *JCA* fits only the off-diagonal contingency tables by minimizing, instead of (1),

$$n^{-1} \sum_{i=1}^Q \sum_{j=1}^{i-1} \|N_{ij} - H_{ij}\|_{ij}^2, \quad (4)$$

and considers as measure of inertia, instead of (2), the sum of the chi-squares of all off-diagonal tables

$$\chi_J^2 = \sum_{i=1}^Q \sum_{j=1}^{i-1} \chi_{ij}^2,$$

that unfortunately may not be tested for significance. *JCA* is an alternating weighed least-squares algorithm that reminds the *MINRES* method for least-squares

**Table 1** Burt’s table of the three-characters data set of 2,000 words

	L2	L3	L4	WN	WV	WA	TC	TR	TD	TS
L2	1,512	0	0	788	483	241	433	385	399	295
L3	0	375	0	203	23	149	64	82	86	143
L4	0	0	113	62	9	42	3	29	21	60
WN	788	203	62	1,053	0	0	229	284	273	267
WV	483	23	9	0	515	0	174	133	125	83
WA	241	149	42	0	0	432	97	79	108	148
TC	433	64	3	229	174	97	500	0	0	0
TR	385	82	29	284	133	79	0	496	0	0
TD	399	86	21	273	125	108	0	0	506	0
TS	295	143	60	267	83	148	0	0	0	498
	L2	L3	L4	WN	WV	WA	TC	TR	TD	TS

**Table 2** First one-dimensional layer of the layers by kind of words table, one-dimensional reconstruction, and corresponding residuals of SCA

	Layer				Reconstruction				Residual			
	TC	TR	TD	TS	TC	TR	TD	TS	TC	TR	TD	TS
L2	57	7	17	−80	435	382	400	296	−2	3	−1	−1
L3	−33	−4	−10	47	60	89	85	141	4	−7	1	2
L4	−23	−3	−7	33	5	25	22	61	−2	4	−1	−1

factor analysis, where the off-diagonal elements of a correlation matrix are fitted (Thomson, 1934). In the special case  $Q = 2$ , the solution is exactly the SCA of the off-diagonal table  $N = N_{12}$ .

### 3 An Application

To show the different behavior of the different correspondence analyses, we refer to a data set taken from Camiz and Gomes (2009), consisting in 2,000 words taken from four different kind of periodic reviews (*Childish* (TC), *Review* (TR), *Divulgateion* (TD), and *Scientific Summary* (TS)), classified according to their grammatical kind (*Verb* (WV), *Noun* (WN), and *Adjective* (WA)) and the number of internal layers (*Two-* (L2), *Three-* (L3), and *Four and more layers* (L4)), as a measure of the word complexity (Table 1). All the computations have been performed with the *ca* package (Nenadic and Greenacre, 2006, 2007) contained in the *R* environment (R-project, 2009).

We first limit attention to the table crossing Layers by Kind of words, with a chi-square = 125.262 with six degrees of freedom, thus highly significant (test value = 10.177). According to Malinvaud (1987) its SCA gives only one significant eigenvalue (0.061891, test-value = 10.439) summarizing 98.82 of total inertia. The one-dimensional reconstruction is reported in Table 2, with a reduction of absolute

**Table 3** Results of *MCA* on the Burt’s table crossing two characters: singular values and eigenvalues, percentages of inertia, total and off-diagonal residuals of the corresponding reconstruction, re-evaluated inertia and percentages, total and off-diagonal residuals of the corresponding reconstruction

N.	Singular value	Eigen value	Perc. Inertia	Cumul. Perc.	Reconstruction		Re-evaluation		Reconstruction	
					Total	Off-diag	Inertia	Perc.	Total	Off-diag
0					5,215	328			5,215	328
1	0.389863	0.624390	24.98	24.98	4,357	483	0.061891	98.82	4,125	29
2	0.263783	0.513598	20.54	45.52	3,978	730	0.000740	1.18	4,026	0
3	0.250000	0.500000	20.00	65.52	3,102	730				
4	0.236587	0.486402	19.46	84.98	1,946	487				
5	0.141083	0.375610	15.02	100.00	0	0				

residuals from 328, in respect to independence, to only 29. Indeed, the two-dimensional solution has no residuals and identical results are found with *JCA*, as expected.

The *MCA*, applied to the corresponding  $2 \times 2$  Burt’s table, gives the results shown in Table 3. In the table, both singular values and eigenvalues are reported with their percentage to the trace ( $=2.5$ ), the absolute residuals of the total and off-diagonal reconstructions, then the re-evaluated inertias with the corresponding reconstructions, limited to the two main eigenvalues larger than the mean (0.5). According to Ben Ammou and Saporta (1998) only the first factor should be taken into account, since the confidence interval for the mean eigenvalue is  $0.47658 < \lambda < 0.52342$ .

In the last two columns of Table 3 the absolute residuals for the re-evaluated *MCA*, both total and off-diagonal, are reported according to the dimension, the 0 corresponding to the deviation from independence: the results are identical to those of *SCA*. Instead, looking at the columns 6 and 7, we have a surprise: whereas the total residuals of the reconstruction decrease monotonically to zero, the off-diagonal ones immediately increase, until the mean eigenvalue, then monotonically decrease, with a better approximation only at the last step. That is, only the total reconstruction is better than the independent table in estimating the table itself.

If we apply both *MCA* and *JCA* to the three-characters data table from which the previous table was extracted, we find a similar but worst pattern. Here, only 3 out of 7 *MCA* eigenvalues are above the mean, with only one significant, as the confidence interval at 95% level is now  $(0.30146 < \lambda < 0.36521)$ , and a second one non-significant but very close to its upper bound. This is in agreement with the Malinvaud (1987) test applied to the three two-way tables, only one of which has a significant second factor. In Table 4 total and off-diagonal absolute residuals for normal *MCA*, *JCA*, and re-evaluated *MCA* inertias are reported according to the dimension (the 0 corresponds to the independence).

Observing the table one may note the same pattern of the residuals of *MCA* as before: a monotone reduction of the total residuals and an increase of the off-diagonal ones until the average eigenvalue, then a reduction of the latter, so that only a six-dimensional solution shows off-diagonal residuals lower than the

**Table 4** Total and off-diagonal absolute residuals of normal *MCA*, *JCA*, and re-evaluated *MCA* on the Burt’s table crossing three characters

Dim	MCA		JCA		Re-evaluated MCA	
	Total	Off-diag.	Total	Off-diag.	Total	Off-diag.
0	8,905	954	8,905	954	8,905	954
1	7,557	1,044	6,629	240	6,885	311
2	7,378	1,537	6,206	145	6,581	232
3	7,089	1,813	5,836	18	6,509	214
4	5,949	1,572				
5	3,675	977				
6	2,335	729				
7	0	0				

independence. On the opposite, the re-evaluated inertias get a monotone pattern but far from the quality of adjustment of *JCA*, that performs quite well. Indeed, the re-evaluated *MCA* needs two dimensions to approach the one-dimensional solution of *JCA*, but never reaching the two-dimensional one.

4 Conclusion

The results of this experimentation show that the [Ben Ammou and Saporta \(1998\)](#) test reveals useful for estimating the suitable dimension of an *MCA* solution. Instead, the reconstruction of the Burt’s table performed by normal *MCA* is so biased that it is not the case to keep on using *MCA* as it is normally performed. The re-evaluated inertias avoid the dramatic bias introduced by the diagonal blocks, but its quality of reconstruction, limited to the factors whose eigenvalue is larger than the mean, is far from being acceptable. In particular, it is so poor in respect to *JCA* that one may wonder why not eventually shift to this method. Indeed, some questions may arise whether the chi-square metrics would be really suitable for a Burt’s table, but this is a question that deserves a broader discussion.

**Acknowledgements** This work was mostly carried out during the reciprocal visits of both authors in the framework of the bilateral agreement between Sapienza Università di Roma and Universidade Federal do Rio de Janeiro, of which both authors are scientific responsible. The first author was also granted by his Faculty, the Facoltà d’Architettura ValleGiulia of La Sapienza. All grants are gratefully acknowledged.

References

Ben Ammou, S., & Saporta, G. (1998). Sur la normalité asymptotique des valeurs propres en ACM sous l’hypothèse d’indépendance des variables. *Revue de Statistique Appliquée*, 46(3), 21–35.

Benzécri, J. P. (1979). Sur les calcul des taux d’inertie dans l’analyse d’un questionnaire. *Les Cahiers de l’Analyse des Données*, 4(3), 377–379.

- Benzécri, J. P., et al. (1973–1982). *L'Analyse des données*, Tome 1. Paris: Dunod.
- Camiz, S., & Gomes, G. C. (2009). Correspondence analyses for studying the language complexity of texts. In *VIII Congreso Chileno de Investigación Operativa, OPTIMA, Concepción (Chile)*, on CD-ROM.
- Greenacre, M. J. (1984). *Theory and application of correspondence analysis*. London: Academic.
- Greenacre, M. J. (1988). Correspondence analysis of multivariate categorical data by weighted least squares. *Biometrika*, 75, 457–467.
- Greenacre, M. J. (2006). From simple to multiple correspondence analysis. In M. J. Greenacre, J. Blasius (Eds.), *Multiple correspondence analysis and related methods* (pp. 41–76). Dordrecht: Chapman and Hall (Kluwer).
- Greenacre, M. J., & Blasius, J. (Eds.). (2006). *Multiple correspondence analysis and related methods*. Dordrecht: Chapman and Hall (Kluwer).
- Jackson, D. A. (1993). Stopping rules in principal component analysis: A comparison of heuristical and statistical approaches. *Ecology*, 74(8), 2204–2214.
- Kendall, M. G., & Stuart, A. (1961). *The advanced theory of statistics* (Vol. 2). London: Griffin.
- Malinvaud, E. (1987). Data analysis in applied socio-economic statistics with special consideration of correspondence analysis. In *Marketing science conference*. Joy en Josas: HEC-ISA.
- Nenadic, O., & Greenacre, M. (2006). Computation of multiple correspondence analysis, with code in R. In M. J. Greenacre & J. Blasius (Eds.), *Multiple correspondence analysis and related methods* (pp. 523–551). Dordrecht: Chapman and Hall (Kluwer).
- Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The *ca* package. *Journal of Statistical Software*, 20(3), 1–13.
- R-project (2009). <http://www.r-project.org/>
- Saporta, G., & Tambrea, N. (1993). About the selection of the number of components in correspondence analysis. In J. Janssen & C.H. Skiadas (Eds.), *Applied stochastic models and data analysis* (pp. 846–856). Singapore: World Scientific.
- Thomson, G. H. (1934). Hotelling's method modified to give Spearman's  $\rho$ . *Journal of Educational Psychology*, 25, 366–374.



Classification and Data Mining

Giusti, A.; Ritter, G.; Vichi, M. (Eds.)

2013, XIV, 286 p. 85 illus., 49 illus. in color., Softcover

ISBN: 978-3-642-28893-7