Finding a Real Dimension of a Study in Multiple and Joint Correspondence Analysis

Sergio Camiz
Dipartimento di Matematica, Sapienza Università di Roma
E-mail: sergio.camiz@uniroma1.it,

Gastão Coelho Gomes
Departamento de Métodos Estatísticos,
Universidade Federal do Rio de Janeiro
E-mail: gastao@im.ufrj.br

Abstract

In this work, the problem of the proper dimension of a Multiple Correspondence Analysis (*MCA*) is discussed, based on both the re-evaluation of the explained inertia sensu Benzécri (1979) and Greenacre (2006) and a test proposed by Ben Ammou and Saporta (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 the account. Thus, Greenacre (1988) Joint Correspondence Analysis (*JCA*) is introduced and the results obtained on an application are shown and the quality of reconstruction of both *MCA* and *JCA* solutions are compared to the Simple Correspondence Analysis results of the two-way tables. It results that *JCA*'s reduced-dimensional reconstruction is much better than the *MCA*'s one.

Keywords: Correspondence Analysis, Multiple Correspondence Analysis, Joint Correspondence Analysis.

1 Introduction

In this paper, we deal with the problem of the proper dimension of of Multiple Correspondence Analysis [*MCA*, Benzécri et al. (1973-82); Greenacre (1983)] solution, and the performance of its alternative, the Joint Correspondence Analysis [*JCA*,Greenacre (1988)], whose solution depends on an *a priori* selected dimensionality. The performance is based on the partial reconstruction of the original data that results by the application of both *MCA* and *JCA* reconstruction formulas.

The application of these methods to an example taken from studies in linguistics Nardi (2007) 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*. Indeed, the application to the Burt's table of the chi-square metrics, and the following correspondence analysis, emphasize too much the importance of the block-diagonal matrices, whose interest is practically null, in respect to 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 the basis to define a threshold between relevant information and residuals. The relevant information is also tied to the possibility to interpret the factors, according to the paradigms of the methods at hand: it is usually the percentage of explained inertia the most widely used. Thus, to take into account a large share of inertia is the most evident rough method that may

be used and a higher-dimensional solution is normally preferred to a smaller one only if these values are significantly smaller.

The quality of the results is also of high importance and this is the reason of this paper, in which we show the very bad quality of a very bad reduced dimensional solution of Multiple Correspondence Analysis, in particular in respect to the alternative Joint Correspondence Analysis.

2.1 Singular Value Decomposition and Correspondence Analysis

We may ground our further discussion on the well known Singular Value Decomposition [*SVD*, (Greenacre, 1983; Abdi, 2007)] theorem, that states

Theorem 1. *Any real matrix X may be decomposed as $X = U\Lambda^{1/2}V'$, with Λ the diagonal matrix of the real non-negative eigenvalues of XX', U the orthogonal matrix of the corresponding eigenvectors, and V the matrix of eigenvectors of $X'X$ (with the same eigenvalues), with both constraints $U'U = I$ and $V'V = I$.*

This theorem corresponds to the reconstruction formula of an r-rank matrix

$$x_{ij} = \sum_{\alpha=1}^{r} \sqrt{\lambda_\alpha}\ \ u_{i\alpha}\ v_{j\alpha}$$

on which the Eckart and Young (1936) theorem is based:

Theorem 2. *(Eckart and Young) The s-rank reconstruction of any real matrix X, with $s < r$, the rank of X, once its singular values are sorted in decreasing order,*

$$x_{ij} \approx \sum_{\alpha=1}^{s} \sqrt{\lambda_\alpha}\ \ u_{i\alpha}\ v_{j\alpha}$$

is the best one in the least-squares sense.

In particular, we shall adopt its generalization, more suitable for our purposes:

Theorem 3. *Given two real positive definite matrices M and N, any real matrix X may be decomposed as $X = \widetilde{U}\Lambda^{1/2}\widetilde{V}'$, under constraints $\widetilde{U}'M\widetilde{U} = I$ and $\widetilde{V}'N\widetilde{V} = I$.*

The solution is given by the SVD of the matrix $\widetilde{X} = M^{1/2}XN^{1/2} = F\Lambda^{1/2}G'$, with $F'F = I$, $G'G = I$, $\widetilde{U} = M^{-1/2}F$, and $\widetilde{V} = N^{-1/2}G$. It results that $\widetilde{U}\widetilde{U}' = M^{-1}$ and $\widetilde{V}\widetilde{V}' = N^{-1}$ respectively.

Thus, the exploratory analysis paradigm states that the most relevant information is tied to the largest eigenvalues and the non-relevant to the least ones. The problem of distinguishing among them, that is to identify at least a tentative cutpoint of either the singular- or the eigenvalues sequence, remains a crucial issue, that seems more easily solved in the case of Simple Correspondence Analysis (*SCA*, Benzécri et al., 1973-82; Greenacre, 1983), since the special chi-square metrics adopted allows some useful solutions and an easy interpretation of the results.

Let N an $r \times c$ contingency table, with $n = n_{..}$ the table grand total, $\vec{r} = (p_{1.}, ..., p_{r.})'$ the vector of row marginal profile (with $p_{ij} = n_{ij}/n$), $\vec{c} = (p_{.1}, ..., p_{.c})'$ the vector of column marginal profile, and $D_r = \mathrm{diag}(\vec{r}), D_c = \mathrm{diag}(\vec{c})$ the corresponding diagonal matrices. The *SCA* of N results from the application of *GSVD* to the contingency table N with the constraints given by the diagonal matrices D_r and D_c. As a result, the reconstruction formula of N is:

$$n_{ij} = nr_ic_j \left(1 + \sum_{\alpha=1}^{\min(r,c)-1} \sqrt{\lambda_\alpha}\ f_{i\alpha}\ g_{j\alpha}\right).$$

This results from the formulation of the problem in terms of the best weighed least-squares approximation of the matrix N by another matrix H of lower rank which minimizes

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-h_{ij})^2}{e_{ij}}=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-h_{ij})^2}{nr_ic_j}=n^{-1}\text{trace}\left(D_r^{-1}(N-H)D_c^{-1}(N-H)'\right) \quad (1)$$

where the weights are the inverse of the expected frequencies. Thus, the reconstruction formula may be well synthesized as

$$N=n\ \vec{r}\ \vec{c}\,'+D_rF\Lambda^{1/2}G'D_c. \quad (2)$$

As a matter of fact, in order to produce a simultaneous graphical representation, *SCA* eigenvectors are usually rescaled, by defining as *coordinates* the quantities $\Phi=F\Lambda^{1/2}$ and $\Psi=G\Lambda^{1/2}$. With this transformation, and applying the Eckart and Young's theorem, any reduced rank approximation obtained by limiting the sum above to the r largest eigenvalues is the best approximation in the weighed least-squares sense:

$$n_{ij}\approx nr_ic_j\left(1+\sum_{\alpha=1}^{r}\frac{1}{\sqrt{\lambda_\alpha}}\ \phi_{i\alpha}\ \psi_{j\alpha}\right).$$

It results that the inertia along each dimension α equals ${\chi_\alpha}^2=n\lambda_\alpha$. As in *SCA* the eigenvalues sum, up to the grand total, to the table chi-square, namely

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

the cutting problem is simply solved by using the classical test for goodness of fit (Kendall and Stuart, 1961) or more easily through the Malinvaud (1987) test. The test may be applied, as, for each α-dimensional partial reconstruction, the residuals correspond to

$$Q_\alpha=\sum_{ij}\frac{(n_{ij}-\widetilde{n}_{\alpha ij})^2}{\widetilde{n}_{\alpha ij}},$$

asymptotically chi-square-distributed with $(r-\alpha-1)\times(c-\alpha-1)$ degrees of freedom. In the formula, $\widetilde{n}_{\alpha ij}$ is the cell value estimated by the α-dimensional solution, and the table chi-square test results when $\alpha=0$ and $\widetilde{n}_{0ij}=\frac{n_{i\cdot}\ n_{\cdot j}}{n_{\cdot\cdot}}$ is the expected value under independence. Now, Malinvaud (1987) showed that, by substituting the estimated cell values with the expected ones under independence hypothesis, the formula may be approximated by

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

that may be more easily used to check for nullity of the residuals. It is interesting to observe that to the same property may be associated the partial chi-square test for significance associated to each eigenvalue, $\chi_\alpha^2=n_{\cdot\cdot}\lambda_\alpha$, with $df=(r+c-2\alpha-1)$ (Kendall and Stuart, 1961), to detect if there are linear ordinations of both rows and column levels that explain the deviation from expectation (Orlóci, 1978).

2.2 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, whose rows are the units and the columns are all the levels of the considered variables, or the so-called Burt's table $B=Z'Z$ that gathers all contingency tables obtained by crosstabulating all the variables in Z, including the diagonal tables obtained by crossing each variable with itself. We drop here other definitions and formulas of both *SCA* and *MCA* and

their relations, that may be found, e.g., in Greenacre (1983). Suffice here to remind that, in both cases, the chi-square metrics is adopted so that the interpretation of results ought to be done once again in terms of deviations from expectation. It is easy to see that in this case the total inertia of Z is $I_z = \frac{J-Q}{Q}$, where Q is the number of variables and J the total number of levels, that is $J = \sum_{i=1}^{Q} l_i$ where $l_i$is the number of levels of the i-th character and that the eigenvectors in SCA of both Z and B are the same, whereas the B's eigenvalues are the squares of Z's: $\mu_\alpha^2 = \nu_\alpha$. Thus, it makes no difference to perform MCA on either matrix.

As SCA, given a Burt matrix B, MCA may be defined as the weighted least-squares approximation of B by another matrix H of lower rank, minimizing

$$n^{-1}Q^{-2}\text{trace}\left(D_r^{-1}(B-H)D_r^{-1}(B-H)'\right). \tag{3}$$

Notice how (3) derives from (1). In terms of the subtables, this may be rewritten as

$$\begin{aligned} &n^{-1}\text{trace}\left(D^{-1}(B-H)D^{-1}(B-H)'\right) = \\ &= n^{-1}\sum_{i=1}^{Q}\sum_{j=1}^{Q}\text{trace}\left(D_i^{-1}(N_{ij}-H_{ij})D_j^{-1}(N_{ij}-H_{ij})'\right), \end{aligned}$$

where H is the supermatrix of the H_{ij}. Introducing the norm notation

$$\|A-B\|_{ij}^2 = \text{trace}\left(D_i^{-1}(A-B)\ D_j^{-1}\ (A-B)'\right)$$

the minimization can be written as

$$n^{-1}\sum_{i=1}^{Q}\sum_{j=1}^{Q}\|N_{ij}-H_{ij}\|_{ij}^2. \tag{4}$$

In MCA the identification of the true dimension is particularly difficult, despite the MCA is a SCA of a particular table, because the chi-square test has no sense. Indeed, for B a chi-squared statistic may again be calculated as if it were a contingency table, and this simplifies as

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

where χ_{ij}^2 is the chi-squared statistic for the off-diagonal subtable $N_{ij} = Z_i'Z_j$ crossing the i-th and the j-th characters, but without the possibility to make a test. Unfortunately neither Q_α nor $\widetilde{Q}_\alpha$ computed on the indicator matrix Z are chi-square distributed (Ben Ammou and Saporta, 1998), since Z is composed by 0's and 1's.

Usually, the high number of eigenvalues of the MCA, and their corresponding low values, was criticized by the same Benzécri (1979) that suggests to reevaluate them. Indeed, if we compare SCA and MCA applied to the same two characters contingency table, a relation between the eigenvalues may be found. Indeed, by partitioning a two-characters Burt's table $Z'Z$ into submatrices it can be shown (ibid.) the relation $\mu_\alpha = \frac{1\pm\sqrt{\lambda_\alpha}}{2}$ that holds among the eigenvalues of Z and those of the SCA of the contingency table crossing the two characters. In this case, it is evident that to the eigenvalues $\lambda_\alpha = 0$ of SCA correspond eigenvalues $\mu_\alpha = \frac{1}{2}$ of Z and $\nu_\alpha = \frac{1}{4}$ of B, whereas to the others two correspond, one of which larger and the other smaller than $\frac{1}{2}$ and $\frac{1}{4}$ respectively. Generalizing this argument to several characters results in admitting to limit attention in MCA only to the eigenvalues larger than their mean, that is $\mu \geq \overline{\mu}_\alpha = \frac{1}{Q}$.

The argument is discussed in detail by both Benzécri (1979) and Greenacre (1988, 2006). Both authors suggest, in order to get a measure of relative importance of each factor, to re-evaluate the eigenvalues larger than the mean (equal to $\frac{1}{Q}$) according to the formula

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

Greenacre (1988) suggests to consider as total inertia the sum of the re-evaluated eigenvalues and consider as percentage of explained inertia the ratio $\frac{\rho(\mu_\alpha)}{\sum_\alpha \rho(\mu_\alpha)}$. This results in a dramatic re-evaluation of the relative importance of the first eigenvalues. On the opposite, Greenacre bases his arguments on the unusefulness to take into account the diagonal block matrices and the utility to limit attention only to the total off-diagonal inertia of the table, that is the sum of squared (non-re-evaluated) eigenvalues minus the diagonal inertia: that is

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

Experiments show that the Greenacre's reevaluation is always limited to a share of the total inertia of Burt's table even by taking into account all the eigenvalues larger than the mean.

An alternative is proposed by Ben Ammou and Saporta (1998, 2003): they suggest to estimate the significance of the eigenvalues of *MCA* according to their distribution. If the characters are independent, $\sum_{\beta=1}^{J-Q}\mu_\beta = \frac{J-Q}{Q}$ and $S_{\mu^2} = \sum_{\beta=1}^{J-Q}\mu_\beta^2 = \frac{J-Q}{Q^2} + \frac{\sum_{i\neq j}\phi_{ij}^2}{Q^2}$ with $n_{..}\phi_{ij}^2 \approx \chi^2_{(l_i-1)(l_j-1)}$, thus,

$$E[n_{..}\phi_{ij}^2] = E[\chi_{ij}^2] = (l_i-1)(l_j-1)$$

so the expectation of the variance S_μ^2 of the eigenvalues is

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

Roughly, one may assume that the interval $\frac{1}{Q}\pm 2\sigma$ should contain about 95% of the eigenvalues. Indeed, since the kurtosis of the set of eigenvalues is lower than for a normal distribution, the actual proportion is larger than 95%.

2.3 Joint Correspondence Analysis

Greenacre (1988) criticizes *MCA* approach since it is not a natural generalization of *SCA* and proposes his *Joint Correspondence Analysis* (*JCA*) as its natural generalization. Moreover, in *MCA* no justification exist for fitting the diagonal subtables B which contribute the term $n(J-Q)$ to the total variation. A more natural measure of total variation is the sum $\sum\sum_{q\neq s}\chi_{qs}^2$. This suggests an alternative generalization of correspondence analysis which fits only the off-diagonal contingency tables, analogous to factor analysis where values on the diagonal of the covariance or correlation matrix are of no direct interest.'

Indeed, the proposed redefinition of the total variation, by removing the diagonal block-matrices, would fix an important bias due to the application to the Burt's table of the chi-square metrics, as the diagonal structure of the diagonal block-matrices represents a very high deviation from the expected values, that *MCA* analyzes as if it were a true deviation. On this basis, on the opposite to the current use, this kind of analysis is not really suitable.

So, Greenacre (1988) proposes his *Joint Correspondence Analysis* (*JCA*) as a weighed least-squares approximation aiming at minimizing

$$n^{-1}\sum_{i=1}^{Q}\sum_{j=1}^{i-1}\|N_{ij}-H_{ij}\|_{ij}^2\,, \tag{5}$$

instead of (4) with the corresponding $\chi_J^2 = \sum_{i=1}^{Q}\sum_{j=1}^{i-1}\chi_{ij}^2$, sum of the chi-squares of all off-diagonal tables, that unfortunately may not be checked for significance.

In order to get the solution, he proposes an alternating least-squares algorithm, based on the reformulation of (5) as follows:

$$n^{-1}\sum_{i=1}^{Q}\sum_{j=1}^{i-1}\|N_{ij}-H_{ij}\|_{ij}^2 = n^{-1}\sum_{i=1}^{Q}\sum_{j=1}^{i-1}\left\|N_{ij}-n\,\vec{r}_i\,\vec{r}_j\,'-L_{ij}\right\|_{ij}^2 \tag{6}$$

with $\vec{r}_i$ the diagonal of the i-th block-diagonal matrix. Calling H and L the supermatrices gathering the H_{ij} and L_{ij} respectively, Greenacre (1988) states the equivalence of the rank-K solution of L which satisfies the normal equations in the minimization of the second term of (6) with the rank-$(K+1)$ matrix $H = \vec{r}\,\vec{r}\,' + L$ which satisfies minimizing (5), with $\vec{r}$ the supervector gathering the Q vectors $\vec{r}_i$.

The matrix approximation L of rank K is of the form $L = nDXD_\beta X'D$, where the $J \times K$ matrix X is normalized as $X'DX = QI$, with $D = \text{diag}(\vec{r})$. The matrix X of parameters has rows corresponding to the categories of the variables and columns corresponding to the dimensions of the solution, that must be chosen in advance. The diagonal matrix D_β contains a scale parameter for each dimension. This form of L and the normalization conditions are chosen to generalize the bivariate case (2). The parameter matrix X is partitioned row-wise according to the variables as $X_1, \cdots, X_Q$, where X_q is $J_q \times K$, so that the submatrices of L are $L_{qs} = nD_qX_qD_\beta X_s'D_s$. There are also inherent centering constraints on X of the form $X'r = 0$ due to the orthogonality with the dimension defined by the trivial solution. It is evident that the dimension of the solution must be chosen in advance.

Thus Greenacre (1988) proposes the approximate reconstruction of the whole matrix $B - n\,\vec{r}\,\vec{r}\,'$, namely

$$B - n\,\vec{r}\,\vec{r}\,' \approx nDXD_\beta X'D + C,$$

where C is a block diagonal matrix with submatrices C_{qq}, $q = 1, ..., Q$ down the diagonal and zeros elsewhere. Here, each C_{qq} is composed by dummy parameters which effectively allow perfect fitting of the submatrices on the diagonal of $B - n\,\vec{r}\,\vec{r}\,'$, thereby eliminating their influence on the model of interest. The minimization of

$$\begin{aligned} B - n\,\vec{r}\,\vec{r}\,' = 2n^{-1} \sum_{i=1}^{Q} \sum_{j=1}^{i-1} \left\| N_{ij} - n\,\vec{r}_i\,\vec{r}_j\,' - L_{ij} \right\|_{ij}^2 \\ +\ n^{-1} \sum_{k=1}^{Q} \left\| N_{kk} - n\,\vec{r}_k\,\vec{r}_k\,' - L_{kk} - C_{kk} \right\|_k^2 . \end{aligned} \tag{7}$$

is equivalent to minimizing (6) because the latter set of terms in (7) can always be made zero by setting $C_{ii} = N_{ii} - n\,\vec{r}_i\,\vec{r}_i\,' - L_{ii}$.

The algorithm proposed by Greenacre (1988) to minimize (7) can be performed iteratively by alternating between the variables in C and those in X and D_β as follows:

1. fix the dimension K of the solution.

2. initiate the algorithm with an analysis of the full Burt matrix B, that is

$$B - n\,\vec{r}\,\vec{r}\,' \approx nDXD_\beta X'D. \tag{8}$$

3. limiting attention to the first K dimensions, say the first K columns of X $\vec{x}_{(1)}, \cdots, \vec{x}_{(K)}$, (8) can be rewritten as

$$B - n\,\vec{r}\,\vec{r}\,' \approx \sum_{k=1}^{K} n\beta_k D\vec{x}_{(k)}\vec{x}\,'_{(k)}D.$$

   so that, if all quantities except the β_k (k = 1,$\cdots$, K) are regarded as fixed, the problem reduces to a simple weighted least-squares regression (see Greenacre, 1988, for further details).

4. Keeping X and D_β fixed, set

$$C_{ii} = N_{ii} - n\,\vec{r}_i\,\vec{r}_i\,' - nD_iX_iD_\beta X_i'D_i \quad (i = 1, \cdots, Q).$$

5. Keeping C fixed, minimize with respect to X and D_β: this is achieved by performing a correspondence analysis on the table $B* = B - C$, that is the Burt matrix with modified submatrices on its diagonal, setting X equal to the first K vectors of optimal row or column parameters and the diagonal of D_β equal to the square roots of the first K principal inertias respectively.

6. Iterate the last two steps until convergence.

In the special case Q = 2, where the problem reduces to fitting the single off-diagonal submatrix N_{12}, the initial solution described above is optimal and provides the simple correspondence analysis of $N = N_{12}$ exactly. $N = N_{12}$ exactly.

3 An Application

To show in detail the different behavior of the different correspondence analyses, we refer to a data set taken from Nardi (2007), consisting in 2000 words taken from four different kind of periodic reviews (*Childish (TC), Review (TR), Divulgation (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: *Burt's table of the words' type example.*

	L2	L3	L4	WN	WV	WA	TC	TR	TD	TS
L2	1512	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	1053	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

In Table 1 the Burt's table that results by crossing the three characters is reported. In Table 2 are represented the first results of the *SCA*s of the three contingency data tables, crossing the three characters two by two, limited to the first two eigenvalues, namely, the eigenvalues, the percentage of corresponding inertia, and the p-value associated to the chi-square calculated for the corresponding one-dimensional reconstruction, that in this case is identical to the Malinvaud's test, since each solution is 2-dimensional.

Table 2: *SCA of the three contingency data tables of the three characters two by two. In the columns, the eigenvalues, the percentage of inertia, and the p-value of the chi-square associated to the factors.*

	words vs. levels			publications vs. words			publications vs. levels		
N.	eigen	%	p-value	eigen	%	p-value	eigen	%	p-value
1	.0925	99.98	.0000	.0253	80.53	.0000	.0619	98.82	.0000
2	.0000	0.02	.8625	.0061	19.47	.0022	.0007	1.18	.4771

In two cases, the chi-squares test that the second factor has no real meaning, since the p-value is larger than 5%, whereas for the case of the table crossing the type of publication and the kind of words the second factor is also significant. In Figure 1 the results of the three *SCA*s are represented too: it must be pointed out that the vertical position of the items is significant only for the second graphic. Indeed, the inspection of this factor plane shows an arch pattern due to a Guttman effect (Guttman, 1941; Camiz, 2005).

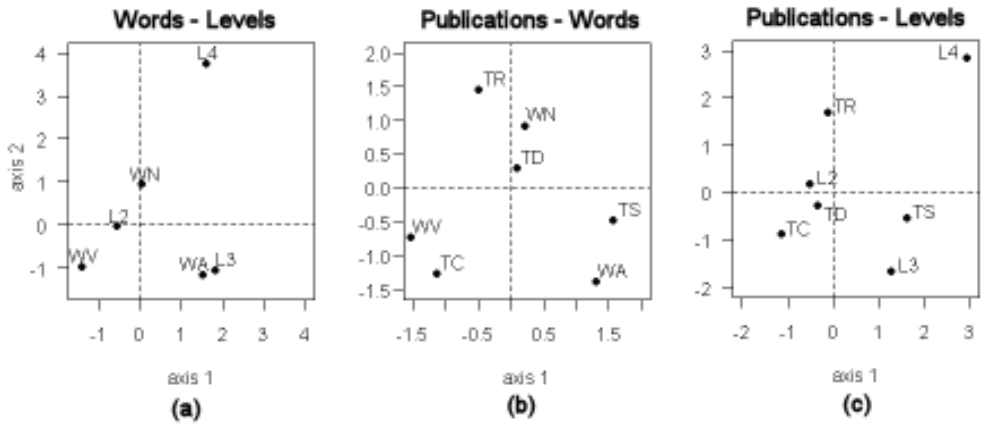

Figure 1: *Words' type example: The pair of characters levels on the three two-way SCAs: (a) Words vs. Levels; (b) Publications vs. Words; (c) Publications vs. Levels.*

Running *MCA*, the pattern of eigenvalues is represented in Table 3, in which are reported the singular values of Z, their percentage to their total (that equals $\frac{J-Q}{Q} = 2.33$), the cumulate percentage, the eigenvalues of the Burt's matrix, corresponding to the explained inertia, and the cumulate inertia.

Table 3: *MCA singular values, percentage to the total and cumulate percentage, eigenvalues, and cumulate inertia of the Burt's table of words' type example.*

Number	Singular value	Percentage	Cumulate %	Eigenvalue	Cumulate inertia
1	0.4896	20.98	20.98	0.239688	0.239688
2	0.3640	15.60	36.58	0.132472	0.372160
3	0.3434	14.72	51.30	0.117930	0.490090
4	0.3300	14.14	65.44	0.108885	0.598975
5	0.3084	13.22	78.66	0.095100	0.694076
6	0.2728	11.69	90.35	0.074431	0.768507
7	0.2252	9.65	100.00	0.050713	0.819220

Indeed, according to both Benzécri (1979) and Greenacre (1988), only three singular values are larger than $1/Q = 1/3$, so that the re-evaluations, reported in Table 4, are referenced to only three dimensions. In both cases, the first dimension re-evaluated inertia is by far larger than the others.

Table 4: *Inertia re-evaluation according to both Benzécri (1979) and Greenacre (1988) of words' type example.*

	Benzécri's Re-evaluation			Greenacre's Re-evaluation		
Number	Inertia	%	Cum.%	*Inertia*	%	Cum.%
1	0.0549	95.91	95.91	0.2344	88.36	88.36
2	0.0021	3.69	99.60	0.0460	3.40	91.76
3	0.0002	0.40	100.00	0.0151	0.37	92.13
Total	0.0572	100.00		0.2954	92.13	

If we apply the Ben Ammou and Saporta (1998, 2003) estimation of the average singular value distribution under independence, we find that the standard deviation is $\sigma = 0.0159364$, so that the confidence interval at 95% level is $(0.30146 < \lambda < 0.36521)$.

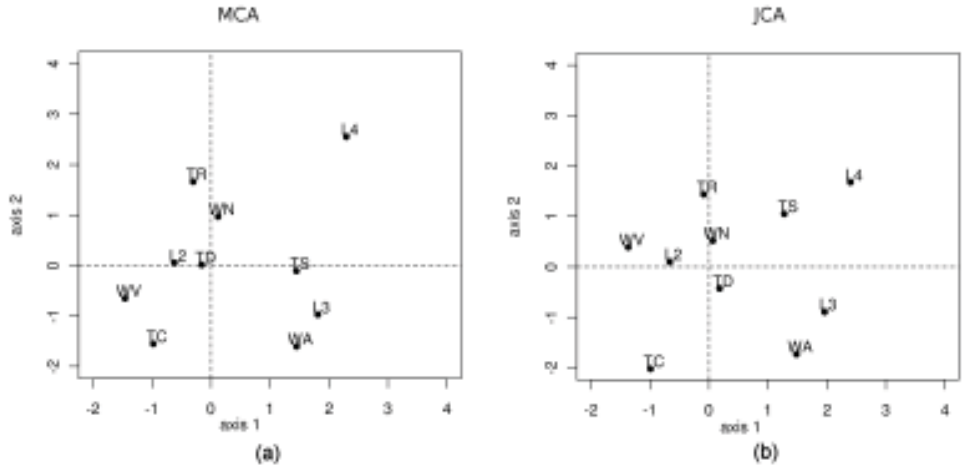

Figure 2: *Words' type example: representation of the three-characters levels on the plane spanned by the first two factors: (a) MCA; (b) JCA.*

As a consequence, only the first singular value is outside the confidence interval and should be considered significant. As a matter of facts, the second one is very close to the threshold (0.3640): this is consistent with the fact that one of the 2-dimensional tables has a significant second eigenvalue.

Table 5: *Original two-way contingency tables of words' type example and their reconstruction according to the first dimension of* SCA*s,* MCA*, and* JCA*, with the corresponding cumulate absolute residuals.*

Original Burt's Matrix

	WN	WV	WA		TC	TR	TD	TS		TC	TR	TD	TS
L2	788	483	241	L2	433	385	399	295	WN	229	284	273	267
L3	203	23	149	L3	64	82	86	143	WV	174	133	125	83
L4	62	9	42	L4	3	29	21	60	WA	97	79	108	148

SCA First Layer

	WN	WV	WA		TC	TR	TD	TS		TC	TR	TD	TS
L2	788	483	241	L2	435	382	400	296	WN	253	257	267	276
L3	204	23	149	L3	60	89	85	141	WV	165	144	127	79
L4	61	9	42	L4	5	25	22	61	WA	82	96	112	142

SCA cumulate absolute residuals

	2				107					2210			

MCA First Layer

	WN	WV	WA		TC	TR	TD	TS		TC	TR	TD	TS
L2	770	559	183	L2	492	409	401	211	WN	249	257	264	283
L3	216	-24	183	L3	13	69	82	211	WV	219	155	145	-3
L4	67	-20	66	L4	-5	18	23	76	WA	32	84	97	219

MCA cumulate absolute residuals

	14440				18972					21183			

JCA First Layer

	WN	WV	WA		TC	TR	TD	TS		TC	TR	TD	TS
L2	783	484	245	L2	435	391	393	293	WN	259	260	266	269
L3	207	29	139	L3	53	82	87	153	WV	160	136	136	82
L4	63	2	48	L4	12	24	25	52	WA	81	100	104	147

JCA cumulate absolute residuals

	280				488					2570			

Let us look now at the one-dimensional reconstruction, as resulting by the *SCA*s of the three individual tables, by the *MCA*, and by Greenacre's *JCA* as reported in Table 5. The comparison of the *SCA* one-dimensional solutions with the original tables shows that the amount of the cumulate absolute residuals is in good agreement with the quality of the solution, as represented by the corresponding chi-square.

Table 6: *Absolute residuals of the reduced dimensional reconstructions of both the Burt's table and the two-way off-diagonal ones according to* MCA*, reevaluated MCA and* JCA *respectively: to 0 correspond the deviations from independence.*

	MCA		*MCA reeval.*		*JCA*	
Dim	*total*	*Off-diag.*	*total*	*Off-diag.*	*total*	*Off-diag.*
0	8906	953	8906	953	8906	953
1	7557	1044	6879	308	6629	240
2	7378	1537	6588	236	6206	145
3	7089	1813	6510	215	5836	18
4	5949	1572				
5	3675	977				
6	2335	729				
7	0	0				

For this reason, the low quality of the reconstruction of the table crossing kind of words with the type of publications depends on the significance of the second dimension of the *SCA* of this table. At first glance, it is evident the high difference in the cumulate absolute residuals of *MCA* in respect to the other solutions, that is an important sign of the limits of *MCA* in respect to *JCA*. Indeed, the quality of *JCA* one-dimensional reconstruction is in all cases acceptable, so that it is possible to observe a synthetical graphical representation of the three tables that is realistic. On the opposite, the *MCA* reconstruction is dramatically bad: in Table 6 are reported the cumulate absolute residuals of reconstructions of both *MCA* and *JCA*, both for the whole Burt's table and for the three off-diagonal two-way tables. The residuals for 0-dimension are the deviations from independence and the following are reported for all the allowed dimensions: $7 = J-Q$ for *MCA* and 3 for *JCA*, that corresponds to the number of singular values of the Burt's table larger than the mean. Looking at the table, we may notice a continuous decrease of the total residuals in both analyses, with a perfect fit for the total reconstruction of *MCA*, decrease that is somehow slower for *JCA*. On the opposite, the off-diagonal reconstruction of *JCA* is fast and effective, with the 3-dimensional solution nearly perfect, whereas the reconstruction of *MCA* follows a very different pattern. Indeed, the off-diagonal residuals increase progressively, instead of diminishing, until the average eigenvalue, then lower, but improving the reconstruction in respect to the deviation from independence only with the last two dimensions.

To graphically study the results, we can now compare the 2-dimensional graphics obtained by the three *SCA*s, shown in Figure 1, with those obtained by both *MCA* and *JCA*, shown in Figure 2. The position of the levels of each character are represented on the plane spanned by the first two factors. Considering also that the second dimension is limited in significance, we may note that both *MCA* and *JCA* factor planes represent a good compromise among the three 2-dimensional graphics. The reciprocal positions of the items are not so different among *MCA* and *JCA*: only *WV* and *TS*, are more shifted and their position on *JCA* plane seemsbetter reflect their relation with the other levels.

4 Conclusion

This study started with the aim to understand to what extent the *JCA* (Greenacre, 1988) could be of help in identifying the true dimension of an analysis concerning a set of qualitative data. In this sense, the confidence interval proposed by Ben Ammou and Saporta (1998, 2003) seems a useful answer to this problem, in agreement with the most one-dimensional solution of the *SCA*s applied to the two-way tables of the first application. During the study, the problem of the data reconstruction not only showed that *MCA* is bad in reconstructing the data table, due to the inflation in the number of eigenelements, but also that the re-evaluations proposed by both Benzécri (1979) and Greenacre (2006) do not take into account the fact that the reconstruction of the two-way off-diagonal tables is for the most reduced-dimensional solutions worst than the initial independence table. To get closer to the daily use of the graphics, as a help for the description and the interpretation of the data, the higher homogeneity of the ranges of the various characters on factor planes of *JCA* improves the interpretation ability of the graphics themselves. It is very strange that, despite the number of studies developed on *MCA*, no trace results in literature of the serious drawbacks found in *MCA*, nor Greenacre (1988) and the followers (**?**Vermunt and Anderson, 2005; Greenacre, 2006) quote their important improvement. Thus, *JCA* seems a most promising development and its properties deserve some further deepening.

Acknowledgements

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

References

Abdi, H. (2007). Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition ($GSVD$). In: N. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.

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.

Ben Ammou, S., Saporta G. (2003). On the connection between the distribution of eigenvalues in multiple correspondence analysis and log-linear models. *REVSTAT-Statistical Journal*, 1(0), 42-79.

Benzécri, J.P., et coll. (1973-82). *L'Analyse des données*, Tome 2. Paris: Dunod.

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.

Camiz, S. (2005). The Guttman Effect: its Interpretation and a New Redressing Method. *Tetradia Analushsq Dedomenwn (Data Analysis Bulletin)*, 5, 7-34.

Eckart, C., Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1, 211-218.

Greenacre, M.J. (1983). *Theory and Application of Correspondence Analysis*. London: Academic Press.

Greenacre, M.J. (1988). Correspondence analysis of mutlivariate categorical data by weighted least squares. *Biometrika*, 75, 457-467.

Greenacre, M.J. (2006). From Simple to Multiple Correspondence Analysis. In: Greenacre and Blasius (2006) (Eds.), 41-76.

Greenacre, M.J., Blasius, J. (Eds.) (2006). *Multiple Correspondence Analysis and Related Methods*. Dordrecht (The Netherlands): Chapman and Hall (Kluwer).

Guttman, L. (1941). The Quantification of a Class of Attributes: a Theory and Method of Scale Construction. In P. Horst (Ed.) *The Prediction of Personal Adjustment*. New York, Social Science Research Council.

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. Marketing Science Conference, Joy en Josas: HEC-ISA.

Nardy, M.N.S. (2007). A sintaxe no interior das palavras - efeitos de gênero na lingua escrita contemporânea. PhD Thesis in Linguistics. Rio de Janeiro, Facultade de Leteras da Universidade Federal de Rio de Janeiro.

Nenadic, O., Greenacre, M. (2006). *Computation of multiple correspondence analysis, with code in R*. In: Greenacre and Blasius (2006) (Eds.), 523-551.

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.

Orlóci, L. (1978). *Multivariate Analysis in Vegetation Research*, 2nd ed.. Den Haag: Junk.

R-project (2009), http://www.r-project.org/

Vermunt, J.K., Anderson, C. (2005). Joint Correspondence Analysis (JCA) by Maximum Likelihood, *European Journal of Research Methods for the Behavioral and Social Sciences*, 1(1), 18-26.