Reconstruction Ability of Exploratory Analysis Methods for Qualitative Data

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 partial reconstruction of the Burt's table through the decompositions performed by Multiple Correspondence Analysis (*MCA*), Greenacre (1988)'s Joint Correspondence Analysis (*JCA*), and Gower and Hand (1996)'s Extended Matching Coefficient (*EMC*) are compared, in order to check the quality of the methods. In particular, the ability is considered separately for the whole, the diagonal, and the off-diagonal tables, that is to describe either each character's distribution or the interaction between pairs of characters, or both. The theoretical aspects are discussed first, then the results obtained in an application are shown and discussed.

Keywords: Correspondence Analysis, Multiple Correspondence Analysis, Joint Correspondence Analysis, Extended Matching Coefficient, Singular Value Decomposition.

1 Introduction

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: in the linear case, the percentage of explained inertia is 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 its corresponding inertia is significantly larger than the previous one.

Tied to this aspect, the reconstruction of the original data table according to a lower rank matrix is of relevance, since it corresponds to the explained inertia; thus the examination of partial reconstructions is helpful to better understand to what extent the reduction in dimension, through the use of factors, provides a reasonable approximation of the original data.

In this paper, we consider the special case of qualitative data, that are usually summarized by the so-called Burt's matrix, the super-contingency table that cross-tabulates all characters taken into account. It is well known that an exploratory factor analysis of such a matrix is possible through Multiple Correspondence Analysis (*MCA*, Benzécri et al., 1973-82; Greenacre, 1983), but also other methods are proposed in literature, able to overcome some limitations of *MCA* itself. Indeed, the most significant one, albeit not seriously considered in literature as such, is the use of the chi-square metrics. The chi-square

metrics is based on the deviation from the expectation, a measure that cannot work on such table, in which the subtables along the diagonal cross each character with itself so that they too are diagonal. Another point is the low amount of explained inertia by the main factors, that depends also on the high number of eigenvalues that result. Tied with this, a problem results in the unpredictability of the partial data reconstruction of the Burt's table as it will be shown in the following.

In this paper, we take into account two different proposed alternatives: the Joint Correspondence Analysis (*JCA*, Greenacre, 1988), whose solution depends on an *a priori* selected dimensionality, and the Principal Component Analysis (*PCA*, Jolliffe, 2002) of the Extended Matching Coefficient matrix (*EMC*, Gower and Hand, 1996), which are proposed by the corresponding authors as a solution.

At the end, these methods will be applied to a very small table, taken from studies in linguistics (Nardi, 2007).

2 Theoretical framework

2.1 Singular Value Decomposition

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,*

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

is the best one in the least-squares sense.

This means that, for every $s < r$, the matrix $H_s = (h_{s,ij})$ solves the problem to approximate an $n \times p$ matrix X by another one H_s of lower rank s at the best in the least-squares sense, thus by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - h_{s,ij})^2 = \text{trace}\,((X - H_s)(X - H_s)') \tag{1}$$

It is well known that Principal Component Analysis (*PCA,* Jolliffe, 2002) finds its rationale on this theorem, once the data table is standardized according to $z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{n}\sigma_j}$,

with $\bar{x}_j$ and σ_j the average and the standard deviation of the j-th character; indeed, this way $Z'Z = cor(X) = C$, the matrix of correlations between the columns of X. Thus, given the *PCA* on the correlation matrix C, with Λ and V their diagonal matrix of eigenvalues and unit matrix of eigenvectors respectively, and given U the unit eigenvectors of ZZ', the reconstruction formula becomes

$$x_{ij} = \left(\sum_{\alpha=1}^{r} \sqrt{\lambda_\alpha} \;\; u_{i\alpha} \; v_{j\alpha} \right) \sqrt{n}\sigma_j + \bar{x}_j \tag{2}$$

For correspondence analysis, we shall adopt the Generalized Singular Values Decomposition(*GSVD*, Greenacre, 1983; Abdi, 2007), in which two other matrices are involved:

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

The solution is given by the SVD of the matrix $\widetilde{X} = L^{1/2}XM^{1/2} = F\Lambda^{1/2}G'$, with $F'F = I$, $G'G = I$, $\widetilde{U} = L^{-1/2}F$, and $\widetilde{V} = M^{-1/2}G$. It results that $\widetilde{U}\widetilde{U}' = L^{-1}$ and $\widetilde{V}\widetilde{V}' = M^{-1}$ respectively. In this case the minimization problem (1) becomes

$$\sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{l=1}^{p} m_{ki}^{-1}(x_{ij} - h_{s,ij})^2 n_{jl}^{-1} = \text{trace}\left(L^{-1}(X - H_s)M^{-1}(X - H_s)'\right) \tag{3}$$

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 eigen-values sequence, remains a crucial issue, that did not find a univocal solution so far (for *PCA* see, e.g., Jackson, 1993; Peres-Neto *et al.*, 2005). In Simple Correspondence Analysis (*SCA*, Benzécri et al., 1973-82; Greenacre, 1983), it seems more easily solved, since the special chi-square metrics adopted allows some useful solutions and an easy interpretation of the results, and for *MCA* both Ben Ammou and Saporta (1998, 2003) propose an interesting method.

2.2 Simple Correspondence Analysis

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 $D_r = \text{diag}(\vec{r}), D_c = \text{diag}(\vec{c})$ 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. It results the reconstruction formula of N:

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

where 1 is the first trivial eigenvalue, that ties the origin to the centroid of data and represents the independence. This time, as both D_r and D_c are diagonal and represent the table marginal frequencies of rows and columns respectively, the minimization problem (3) results simplified and takes an interesting aspect

$$\begin{aligned} \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - h_{ij})^2}{e_{ij}} &= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - h_{ij})^2}{nr_ic_j} \\ &= n^{-1}\text{trace}\left(D_r^{-1}(N - H)D_c^{-1}(N - H)'\right) \end{aligned} \tag{4}$$

that is the sum of squared deviations from the observed values divided by the expected ones. Thus, the reconstruction formula may be well synthesized as

$$N = n\ \vec{r}\ \vec{c}\,' + D_r F \Lambda^{1/2} G' D_c. \tag{5}$$

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:

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

As it results that 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 inertia along each dimension α equals ${\chi_\alpha}^2 = n\lambda_\alpha$. Thus, 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, since, for each α-dimensional partial reconstruction, the residuals correspond to

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

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 α= 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

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

that may be more easily used to check for nullity of the residuals. Moreover, it is interesting to observe that the partial chi-square associated to each eigenvalue, $\chi^2_\alpha = n_{\cdot\cdot}\lambda_\alpha$, may be checked for significance 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).

2.3 Multiple Correspondence Analysis

Let us consider now a qualitative data table X with n observations, Q nominal characters 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. It is well known that *MCA* of such a matrix 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 J levels of the Q considered variables, or the so-called Burt's table $B = Z'Z$ that gathers all contingency tables obtained by cross-tabulating all the variables in Z, including the diagonal tables obtained by crossing each variable with itself. The idea is to adopt for both matrices the same optimized decomposition of *SCA*, namely the *GSVD* of either $Q^{-1}ZD_r^{-1}$ or

$$Q^{-2}D_r^{-1}BD_r^{-1} = Q^{-2}D_r^{-1}Z'ZD_r^{-1}. \tag{6}$$

Indeed, it is evident that the latter is the square of the previous, so that they share the eigenvectors, and the singular values in Burt's case are the squares of those of the indicator matrix case: $\nu_\alpha = \mu_\alpha^2$ (note that the μ_α are called in literature *eigenvalues* of the Burt's matrix). This identity allows identical interpretation of the resulting factors. Thus, it makes no difference to perform *MCA* on either matrix. On the other side, whereas the total inertia of Z is $I_z = \frac{J-Q}{Q}$, the one of B equals $\sum \nu_\alpha = \sum \mu_\alpha^2$. 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. This point deserves some special attention, since the deviation refers to all contingency tables gathered in the Burt's table, including the diagonal ones. The problem is that such diagonal matrices, that "theoretically" would indicate maximum deviation since the are diagonal, in this case are just the expected ones, as they cross each character with itself.

As for *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

$$n^{-1}Q^{-2}\text{trace}\left(D_r^{-1}(B-H)D_r^{-1}(B-H)'\right). \tag{7}$$

Notice how (7) derives from (4). 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

$$n^{-1}\sum_{i=1}^{Q}\sum_{j=1}^{Q}\|N_{ij}-H_{ij}\|_{ij}^2. \tag{8}$$

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-like 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, and it is dramatically inflated by the diagonal tables without any real meaning.

The high number of eigenvalues of *MCA*, and their corresponding low explained inertia, were criticized by Benzécri (1979) that suggests a reevaluation. 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

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

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. This does not affect the interpretation of the factors, that essentially depends upon the eigenvectors, thus to the contributions of both levels and characters to them, but more the quality of representation of these elements on the factor subspaces, that varies according to the percentage of inertia attributed to each one. Indeed, this is a point that would deserve some specific consideration, in particular in deciding which reevaluation may be better taken into account. In the following, we shall call *adjusted MCA* the one with re-evaluated inertia, thus with the coordinates recalculated accordingly.

The reduction in number of the dimension, thanks to both Benzécri's and Greenacre's reevaluations, does not solve the problem of the true dimension of the table. To this question, an answer comes by Ben Ammou and Saporta (1998, 2003): they suggest to estimate the significance of the eigenvalues of *MCA* according to their distribution under independence. In this case, $E(\mu) = 1/Q$, thus $\sum_{\beta=1}^{J-Q} \mu_\beta = \frac{J-Q}{Q}$ 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%.

2.4 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 $\sum \sum_{q \neq s} \chi^2_{qs}$. 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, since 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, 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 simultaneously

$$n^{-1} \sum_{i=1}^{Q} \sum_{j=1}^{i-1} \|N_{ij} - H_{ij}\|_{ij}^2 , \tag{9}$$

instead of (8) with the corresponding $\chi^2_J = \sum_{i=1}^{Q} \sum_{j=1}^{i-1} \chi^2_{ij}$, 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 (9) as follows:

$$n^{-1} \sum_{i=1}^{Q} \sum_{j=1}^{i-1} \|N_{ij} - H_{ij}\|_{ij}^2 = n^{-1} \sum_{i=1}^{Q} \sum_{j=1}^{i-1} \|N_{ij} - n\, \vec{r}_i\, \vec{r}_j\,' - L_{ij}\|_{ij}^2 \tag{10}$$

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 (10) with the rank-$(K+1)$ matrix $H = \vec{r}\,\vec{r}\,' + L$ which satisfies minimizing (9), 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 $D = \mathrm{diag}(\vec{r})$. 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 (5). 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

$$\begin{aligned} B - n\ \vec{r}\ \vec{r}\,' = 2n^{-1} \sum_{i=1}^{Q} \sum_{j=1}^{i-1} \|N_{ij} - n\ \vec{r}_i\ \vec{r}_j\,' - L_{ij}\|_{ij}^2 \\ +\ n^{-1} \sum_{k=1}^{Q} \|N_{kk} - n\ \vec{r}_k\ \vec{r}_k\,' - L_{kk} - C_{kk}\|_k^2 . \end{aligned} \tag{11}$$

is equivalent to minimizing (10) because the latter set of terms in (11) 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 (11) 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 *MCA* of the full Burt matrix B, that is

$$B - n\ \vec{r}\ \vec{r}\,' \approx nDXD_\beta X'D. \tag{12}$$

3. limiting attention to the first K dimensions, say the first K columns of X $\vec{x}_{(1)}, \cdots, \vec{x}_{(K)}$, (12) 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 of 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.

2.5 The Extended Matching Coefficient

JCA has been introduced by Greenacre (1988) as a way to drop the excessive attention given to the diagonal of the Burt's matrix by *MCA*, that indeed does not deserve any interest. In addition, it is our opinion that the use of the chi-square metrics, that finds its rationale in *SCA*, since its factors partition correctly the chi-square of the contingency data under study, does not find any theoretical justification in the case of Burt's table. Thus, it was interesting to take into account the proposal by Gower and Hand (1996) to drop the chi-square metrics in favor of a simpler one: the *Extended Matching Coefficient*. Indeed, for two units, they define it as the number of common levels across all characters. Thus, given the indicator matrix Z, ZZ' would give us a similarity matrix to deal with; indeed, given its size, the Burt's table B as such *is* its corresponding in the dual space, so that it is sufficient to perform the *SVD* of the centered Burt's matrix, that is

$$Q^{-2}B = Q^{-2}Z'Z \tag{13}$$

to be compared with (6). Now, the reconstruction formula (2) holds, but this time the layers may not be interpreted in terms of deviations from expectation, that is not taken into account by the method, but merely as contribution to the reconstruction, that is in this case the frequencies in the Burt's cells.

3 An Application to the kind of words

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

To show in detail the different behavior of the different analyses in practice, 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), Dissemination (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. In Table 1 the Burt's table that results by crossing the three characters is reported. Note that the abbreviations are the levels' labels that appear on the graphics.

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 - Levels			Publications - Words			Publications - 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 Table 2 are reported the eigenvalues of the three *SCA*s of the contingency data tables that cross the three characters two by two: 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. In two cases, the tests do not give the second factors any real meaning, since the p-value is larger than 5%, whereas in the case of the table type of publication - kind of words the second factor is also significant.

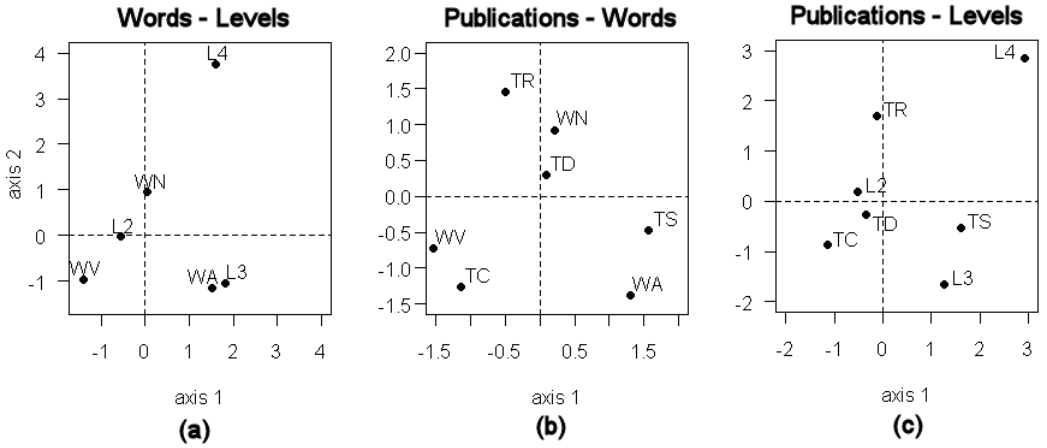


Figure 1: *Words' type example: The pair of characters levels on the three two-way SCAs: (a) Words vs. Levels; (b) Publications vs. Words; (c) Publications vs. Levels.*

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); the same, the interpretation is straightforward: for the first table, both verbs and nouns seem to have in general less syllables than the adjectives; for the second, the variation in use of the words according to the higher complexity of the publication: verbs for the childish, nouns for reviews and disseminations, adjectives for

scientific summaries; for the third, the more complicated words (3 and more syllables) in scientific summaries than in all others. In the second table it is noteworthy the opposite pattern of verbs and adjectives, the first reducing while the publication is of higher level and the second raising: this explains clearly the observed Guttman effect. The position of long words very elongated on the second axis of both the first and the third analyses, in the latter case also with reviews, is explained by the shortness of the verbs and its scarce presence in childish publications, but we said that it is not significant. We may ground our comparisons on this interpretation of the data. Running *MCA*, the pattern of eigenvalues is represented in Table 3, in which are reported the eigenvalues, their percentage to the total (equal to $\frac{J-Q}{Q} = 2.33$), the cumulate percentage, the singular values 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. Then re-evaluated inertia and percentages according to both Benzécri (1979) and Greenacre (1988).*

N	Eigen values	% Inertia	Cumul. %	Sing. values	Cum. Inertia	Re-ev. Inertia	Benzécri's %	Benzécri's Cum.%	Greenacre's %	Greenacre's Cum.%
1	0.4896	20.98	20.98	0.2397	0.2397	0.0549	95.91	95.91	88.36	88.36
2	0.3640	15.60	36.58	0.1325	0.3722	0.0021	3.69	99.60	3.40	91.76
3	0.3434	14.72	51.30	0.1179	0.4901	0.0002	0.40	100.00	0.37	92.13
4	0.3300	14.14	65.44	0.1089	0.5990	0.0572	100.00		92.13	
5	0.3084	13.22	78.66	0.0951	0.6941					
6	0.2728	11.69	90.35	0.0744	0.7685					
7	0.2252	9.65	100.00	0.0507	0.8192					

In addition, on the table are reported the re-evaluated inertia and its percentages and cumulated ones according to both Benzécri (1979) and Greenacre (1988), limited to the only three singular values larger than $1/Q = 1/3$, with the totals in the following row. In both cases, the first dimension's re-evaluated inertia is by far larger than the others. If we apply the Ben Ammou and Saporta (1998, 2003) estimation of the average eigenvalue distribution under independence, we find that the standard deviation is $\sigma = 0.0159364$, so that the confidence interval at 95% level is $(0.30146 < \lambda < 0.36521)$. As a consequence, only the first eigenvalue 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.

In Figure 2*a* the distribution of all character levels on the plane spanned by the first two factors of *MCA* is represented. Indeed, the patterns of all the characters' levels repeat fairy well the same in the three two-way tables: thus it may be taken as a sign of coherence between the individual *SCA*s and *MCA*. It may be observed that the similarity is good even on the second dimension, albeit not significant, whereas on the plane the Guttman effect appears again in good evidence. This may also depend upon the magnitude of the first two eigenvalues, that is sufficiently high to state that the three characters share around either 48% or 36% of the first and second factor respectively. Concerning the inertia reevaluation, this does not affect the interpretation of the single factors but if anything the spaces, since it acts as different multiplicative constants on the factors.

Let us now discuss the results of the *JCA* carried out on the same example. In the 2-dimensional solution[1] the axes inertias are 0.2488 and 0.0272, with a proportion of 90.15% and 9.85%, respectively: considering significant only the first axis, we may observe in Figure 2*b* a pattern of levels nearly identical to the one of *MCA*.

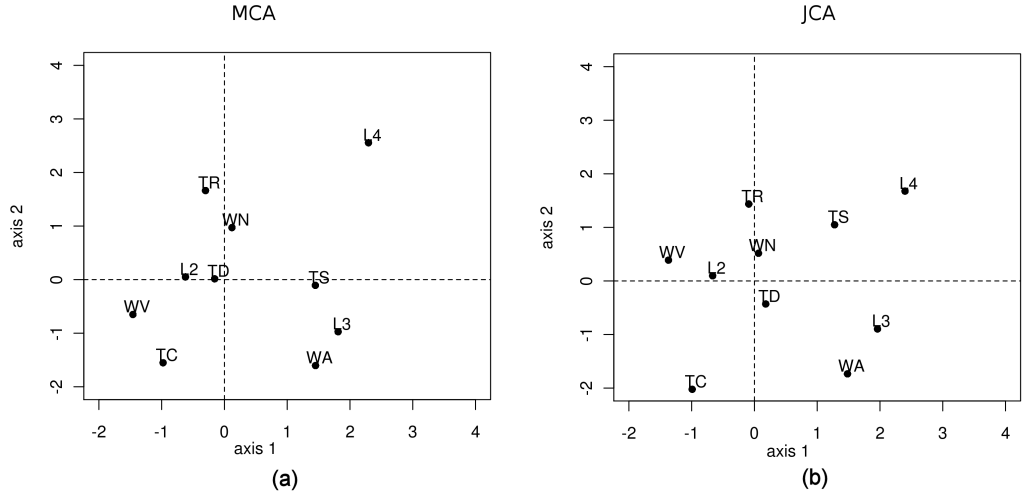
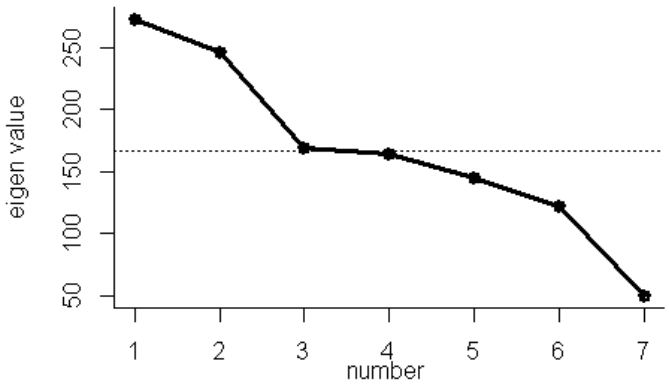


Figure 2: *Words' type example: representation of the three-characters levels on the plane spanned by the first two factors: (a) MCA; (b) JCA.*

Some differences appear on the second axis, in which are noticeable the very different positions of verbs and childish publications on the negative side and of long words and summaries on the positive one, but, once again, this may not be considered significant.

Table 4: *Eigenvalues, percentages of explained and cumulate inertia of the analysis of EMC on Words' type example. On the left the pattern of the eigenvalues.*

N.	eigen	%	cum %
1	272.7187	23.38	23.38
2	245.3787	21.03	44.41
3	168.3971	14.43	58.84
4	163.5518	14.02	72.86
5	145.3435	12.46	85.32
6	121.7007	10.43	95.75
7	49.5341	4.25	100.00



Eventually, we got the results of *EMC*. The seven $(J - Q)$ non-zero eigenvalues and the corresponding percentages of explained and cumulate inertia are reported on Table 4. They are reported also in the figure nearby, where the average corresponds to the

[1]The *R* package *ca*, that we used, gave a diagnostic when asked of running the 1-dimensional solution.

dotted line. Thus, one may identify two major eigenvalues that summarize 44% of the total inertia, three others around the mean and a minor one. As this time, no method is known to decide which is the true Burt's table dimension, according to this method, so that we can only compare the results with the previous ones, thus considering the first dimension as the "true" one, but also taking into account the second one at least for the graphical representation. In Figure 3 all levels are plotted on the plane spanned by the first two factors: indeed, the pattern of levels along the first axis is somehow similar to the ones resulting from both *MCA* and *JCA* but not so much: both *L4* and *L3* and even more *WA* and *WN* are exchanged, slightly modifying the interpretation of the results.

An interpretation of the pattern is possible, considering that, opposite to correspondence analysis, in principal component analysis the lower frequencies results close to the center and the higher far away. Indeed, this is the case of both *L2* and *WN* that have the highest marginal values, whereas *4L*, with the lowest, is toward the center.

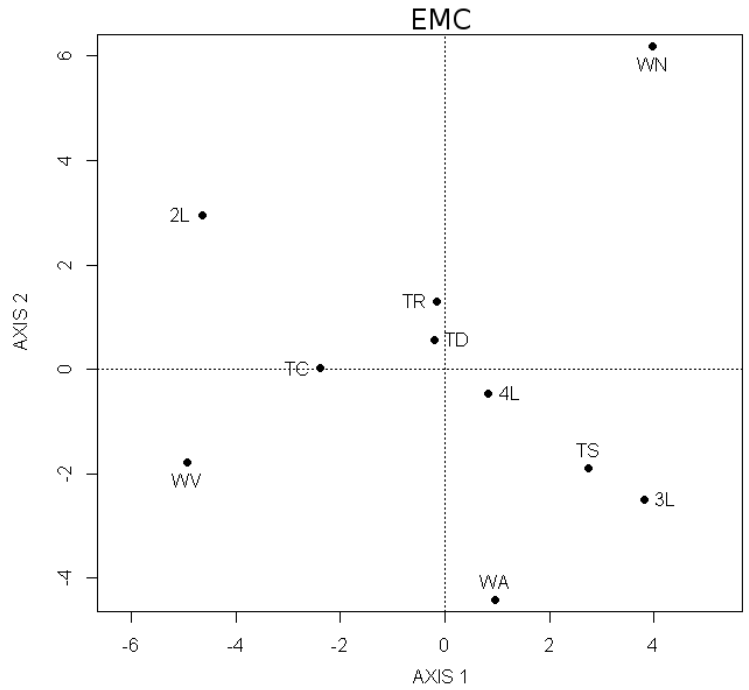


Figure 3: *Words' type example: representation of the three-characters levels on the plane spanned by the first two factors of the centered PCA on the Burt's table, corresponding to the Extended Matching Coefficient.*

Let us look now at the one-dimensional reconstruction, as resulting by the *SCA*s of the three individual tables, by both *MCA* and *adjusted MCA*, by Greenacre's *JCA*, and by *EMC* 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. 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, that here is not taken into account. At first glance, it is evident the high difference in the cumulate absolute residuals of both *MCA* and *EMC* in respect to the other solutions, that is an important sign of their limits in respect to the other

methods. It is noteworthy how the adjusted *MCA*, that is the one with the re-evaluated inertia applied to the reconstruction, works better than *MCA*.

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

Original Contingency Tables

	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		29		134

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

	304		342		397

Adjusted *MCA* First Layer

	WN	WV	WA		TC	TR	TD	TS		TC	TR	TD	TS
L2	783	471	258	L2	433	385	399	295	WN	229	284	273	267
L3	206	39	130	L3	64	82	86	143	WV	174	133	125	83
L4	63	6	44	L4	3	29	21	60	WA	97	79	108	148

Adjusted *MCA* cumulate absolute residuals

	78		67		166

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

	44		64		134

EMC First Layer

	WN	WV	WA		TC	TR	TD	TS		TC	TR	TD	TS
L2	630	595	287	L2	477	381	391	262	WN	178	256	259	360
L3	334	-72	114	L3	12	88	88	187	WV	234	134	139	7
L4	89	-8	32	L4	10	27	27	49	WA	88	106	108	131

EMC cumulate absolute residuals

	631		219		390

Indeed, the quality of *JCA* one-dimensional reconstruction is in all cases much better, meaning that its graphical simultaneous representation of the three tables is the one that approximates best. Finally, looking at the first layer obtained by *EMC* we find a behavior somehow comparable with the first layer of *MCA*: much worst for the first table, much better for the second and relatively equal for the third. This may also depend on the different way that this method uses to reconstruct the data table, as each layer does not represent a deviation from expectation but rebuilds the table anew. Thus a better reconstruction must be expected through a larger number of factors. We did it, by comparing the sum of the absolute differences in the partial reconstructions obtained by increasing the solution dimension: this could be done for the whole 7 factors of both *MCA* and *EMC* and only for the first 3 above the mean for both adjusted *MCA* and *JCA*. The results are given in Table 6.

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*, adjusted MCA and* JCA *respectively: to 0 correspond the deviations from independence.*

	MCA			*Adjusted MCA*			*JCA*			*EMC*		
N	*Total*	*Diag.*	*Off*	*Total*	*Diag.*	*Off*	*Total*	*Diag.*	*Off*	*Total*	*Diag.*	*Off*
	8906	7000	953	8906	7000	953	8906	7000	953			
1	7557	5470	1044	6879	6263	308	6629	6149	240	7849	5363	1243
2	7378	4303	1537	6588	6116	236	6206	5916	145	5950	3907	1022
3	7089	3463	1813	6510	6080	215	5836	5800	18	5185	3129	1028
4	5949	2805	1572							3961	2172	895
5	3675	1720	977							2143	1080	531
6	2335	877	729							513	394	60
7	0	0	0							0	0	0

In Table 6 are reported the cumulate absolute residuals of reconstructions of both *MCA*s, normal and adjusted, *JCA*, and *EMC*: they are total and partitioned according to the diagonal matrices and to the off-diagonal ones. In this latter case, the residuals are divided by two, that is the sum of the residuals of the individual 2×2 contingency tables, that form either triangular off-diagonal sub-matrix. The residuals for 0-dimension are the deviations from independence and, as said, the following are reported for all the allowed dimensions: $7 = J - Q$ for both *MCA* and *EMC* and only 3 for both adjusted *MCA* and *JCA*, the number of singular values of the Burt's table larger than the mean.

The first row reports the deviations in respect to the independence, that for *EMC* does not make any sense. For each method, the first column represents the variation of the whole Burt's table reconstruction: it is always descendant, as it should be expected, although with different slope: in this respect, *EMC* performs best by far. Indeed, the same occurs for what concerns the reconstruction of the diagonal tables: once again the *EMC*'s performance is the best, albeit not as for the total table. Both *MCA* and *EMC* eventually rebuild totally the Burt's table, as expected. The surprises arise looking at the off-diagonal tables reconstruction: here, the *MCA* reconstruction is dramatically bad and problematic: indeed, all partial reconstructions are worst than the independence, that is the estimated frequencies are further from the observed ones than those due to the independence, but the last one. That is the first 5 dimensions, instead of improving

the estimation, get it even worse! In this respect, *EMC* performs much better, as it is constantly decreasing.

If we look now at both adjusted *MCA* and *JCA*, we notice that, for what concerns the diagonal submatrices, they perform very badly, even worst than *MCA*, but this ought to be expected, specifically for *JCA*, in which the diagonal submatrices are intentionally neglected. On the other side, the improvement in the reconstruction of the off-diagonal ones is incredibly better, with an excellent performance of *JCA*.

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 better 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 whole data table in respect to *EMC*, even in what concerns the diagonal submatrices, but mostly concerning the off-diagonal ones, that are even more biased: the reconstruction of the two-way off-diagonal tables is for the most reduced-dimensional solutions worst than the initial independence table. Indeed, only redefining the coordinates according to the adjusted *MCA*, a suitable reconstruction may be performed, albeit far from optimality. It is interesting to note that the adjusted *MCA* performs much better than *(*EMC), a sign that, despite the theory that would oppose its use for Burt's table, the chi-square metrics is more suited for such kind of data. Eventually, the performance of *JCA* is by no means the most suitable to deal with the off-diagonal tables, that is on the study of the interaction among the levels of the different characters.

The re-evaluations proposed by both Benzécri (1979) and Greenacre (2006) were only quoted in literature so far, but never applied in the daily practice. Indeed, they do not influence the interpretation of the factors, but only the graphics and the quality of representation of the character levels. Eventually, *JCA* seems the most promising development of *JCA* and its properties deserve some further deepening. The same, Greenacre's followers (Tateneni and Browne, 2000; Vermunt and Anderson, 2005; Greenacre, 2006), that propose alternative *JCA* programs, do not quote sufficiently their important improvement.

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.

Gower, J.C., Hand, D.J. (1006). *Biplots*. London, Chapman and Hall.

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.

Jackson, D.A. (1993). Stopping rules in principal component analysis: a comparison of heuristical and statistical approaches. *Ecology*, 74: pp. 22042214.

Jolliffe, I.T. (2002). Principal Components Analysis. Berlin, Springer.

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.

Peres-Neto, P.R., Jackson, D.A., Somers, K.M. (2005), How Many Principal Components? Stopping Rules for Determining the Number of Non-trivial Axes Revisited. *Computational Statistics and Data Analysis*, 49: pp. 974-997.

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

Tateneni, K., Browne, M.W. (2000), A Noniterative Method of Joint Correspondence Analysis. *Psychometrika*, 65(2): pp.157-165.

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.