TYING UP THE LOOSE ENDS IN SIMPLE, MULTIPLE AND JOINT CORRESPONDENCE ANALYSIS

Michael Greenacre
Departament d'Economia i Empresa,
Universitat Pompeu Fabra,
Ramon Trias Fargas 25-27 ,
08005 Barcelona. SPAIN
michael@upf.es

Summary. This paper considers several aspects of simple, multiple and joint correspondence analysis that have been misleading, controversial or lacking proper solutions or clarifications. In each case these "loose ends" have been tied up with specific proposals or explanations.

Keywords: Biplot, bootstrap, convex hull, correspondence analysis, explained variance, outliers, rotation, singular-value decomposition.

1 INTRODUCTION

Correspondence analysis (CA) is now no longer a "neglected multivariate method" (Hill 1974) and has found acceptance and application by a wide variety of researchers in different disciplines, notably the social and environmental sciences. The method has also appeared in the major statistical software packages, for example SPSS, Minitab, Stata, SAS and Statistica, and several implementations in R are freely available. My own involvement with CA stretches over 33 years since I arrived in Paris in 1973 to embark on my doctoral studies with Jean-Paul Benzécri. In my opinion, and with experience of the whole range of MDS and biplot methods, CA is the most versatile of them all for ratio-scale data, thanks to its inherent concepts of dimension- and point-weighting. There is only one method that I think can compete with CA, and that is the *spectral map* of Lewi (1976), or weighted log-ratio analysis (Greenacre and Lewi 2005). The spectral map, based on double-centring the log-transformed data, surpasses CA as far as theoretical properties are concerned, but is problematic when the data contain many zero values, as often encountered in research in the social and environmental sciences.

In spite of the vast number of theoretical and applied publications of CA, there are still several issues that remain unsettled and which are often the basis for misconceptions and controversy about the method's properties and interpretation: for example, the measure of variance in CA and multiple CA (MCA), the influence of outlying points, the scaling of row and column coordinates in the maps, whether solutions should be rotated, the statistical significance of the results, and the "horseshoe" effect. In this paper I shall attempt to address these issues and – hopefully – lay them to rest with well-motivated clarifications and solutions.

Although appearing in different but equivalent forms such as "reciprocal averaging", "dual scaling" and "canonical analysis of contingency tables", (simple) CA is generally accepted as a way of visually displaying the association between two categorical variables, based on their cross-tabulation, while MCA is the extension of this method to more than two variables. Categories are depicted as points in a spatial map where certain distances or scalar products may be interpreted as approximations to the original data. I first give a summary of the theory of CA and then tie up the various "loose ends" one by one.

2 Basic CA theory

CA is a particular case of weighted principal component analysis (PCA) (Benzécri 1973, Greenacre 1984: chapter 3). In weighted PCA, a set of multidimensional points exists in a high-dimensional space in which distances and scalar products are measured in a weighted Euclidean sense and the points themselves have differential weights, called "masses" to distinguish them from the dimension weights. A two-dimensional solution (in general, low-dimensional) is obtained by determining the closest plane to the points in terms of weighted least squares, and then projecting the points onto the plane for visualization and interpretation. The original dimensions of the points can also be represented in the plane by projecting unit vectors onto the plane, discussed further in Sect. 7. The following theory shows how to obtain the coordinates of the projected points, called *principal coordinates*, and the coordinates of the projected unit vectors, called *standard coordinates*.

Suppose that $\mathbf{N}$ is an $I \times J$ table of nonnegative data (usually a two-way contingency table but extended to general ratio-scale data). As in PCA, the idea is to reduce the dimensionality of the matrix and visualize it in a subspace of low-dimensionality, usually two- or three-dimensional. The solution was shown by Greenacre (1984: Chapter 2 and Appendix) to be

neatly encapsulated in the singular-value decomposition (SVD) of a suitably transformed matrix. To summarize the theory, first divide **N** by its grand total n to obtain the so-called correspondence matrix $\mathbf{P} = (1/n)\,\mathbf{N}$. Let the row and column marginal totals of **P**, i.e. the row and column *masses*, be the vectors **r** and **c** respectively, and $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices of these masses. Row profiles are calculated by dividing the rows of **P** by their row totals: $\mathbf{D}_r^{-1}\mathbf{P}$. Then CA is a weighted PCA of the row profiles in $\mathbf{D}_r^{-1}\mathbf{P}$, where distances between profiles are measured by the *chi-square metric* defined by $\mathbf{D}_c^{-1}$, and the profiles are weighted by the row masses in $\mathbf{D}_r$. The centroid (weighted average) of the row profiles turns out to be exactly the vector $\mathbf{c}^\mathsf{T}$ of marginal column totals, hence CA of the row profiles analyses the centred matrix $\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}^\mathsf{T}$. The dual CA of column profiles is obtained by simply interchanging rows with columns, i.e. transposing the matrix **P** and repeating all the above.

In both row and column analyses, the weighted sum of chi-square distances of the profile points to their respective centroids is equal to:

$$\text{Inertia} = \phi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{p_{ij} - r_i c_j}{r_i c_j}\right)^2 \tag{1}$$

This quantity, called the (*total*) *inertia*, measures the dispersion of the row profile points and the column profile points in their respective spaces. It is identical to the measure of association known as (Pearson's) mean-square contingency ϕ^2 (square of the "phi-coefficient"), which is Pearson's chi-squared statistic χ^2 divided by the grand total n: $\phi^2 = \chi^2/n$.

The computational algorithm for CA, using the SVD, is as follows:

- Calculate standardized residuals matrix: $\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-1/2}$ (2)
- Calculate SVD: $\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T}$ where $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$ (3)
- Principal coordinates of rows: $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha$ (4)
- Principal coordinates of columns: $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\alpha$ (5)
- Standard coordinates of rows: $\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U}$ (6)
- Standard coordinates of columns: $\mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V}$ (7)

The rows of the coordinate matrices in (4)–(7) above refer to the rows or columns, as the case may be, of the original table, while the columns of these matrices refer to the principal axes, or dimensions, of the solution. Notice that the row and column principal coordinates are scaled in such a way that $\mathbf{F}\mathbf{D}_r\mathbf{F}^\mathsf{T} = \mathbf{G}\mathbf{D}_c\mathbf{G}^\mathsf{T} = \mathbf{D}_\alpha^2$, i.e. the weighted sum-of-squares of the coordinates on the k-th dimension (i.e., their inertia in the direction of this

dimension) is equal to the *principal inertia* (or eigenvalue) α_k^2, the square of the k-th singular value, whereas the standard coordinates have weighted sum-of-squares equal to 1: $\mathbf{X}\mathbf{D}_r\mathbf{X}^\mathsf{T} = \mathbf{Y}\mathbf{D}_c\mathbf{Y}^\mathsf{T} = \mathbf{I}$. Notice further that the only difference between the principal and standard coordinates is the matrix $\mathbf{D}_\alpha$ of scaling factors along the principal axes.

A two-dimensional solution, say, would use the first two columns of the coordinate matrices. The three most common versions for plotting rows and columns jointly are as follows (Sect. 7 treats this topic in more detail):

1. *Symmetric map*: joint plot of row principal and column principal coordinates **F** and **G**.
2. *Asymmetric map of the rows*: joint plot of row principal coordinates **F** and column standard coordinates **Y**.
3. *Asymmetric map of the columns*: joint plot of column principal coordinates **G** and row standard coordinates **X**.

The joint plot of row and column standard coordinates **X** and **Y** has little justification from the point of view of geometric interpretation. The total inertia (1) is equal to the sum of all principal inertias $\alpha_1^2+\alpha_2^2+\dots$. The inertia accounted for in a two-dimensional solution, for example, is the sum of the first two terms $\alpha_1^2+\alpha_2^2$, while the inertia not accounted for is the remainder: $\alpha_3^2+\alpha_4^2+\dots$. These parts of inertia are usually expressed as percentages of inertia explained by each dimension, as in PCA.

3 Multiple and joint correspondence analysis

Multiple correspondence analysis (MCA) is the application of CA to cross-tabulations of Q (>2) categorical variables. There are two almost equivalent forms of MCA: (i) the CA of the rectangular cases-by-categories *indicator matrix* **Z** which codes individual responses in a 0/1 indicator form; and (ii) the CA of the square categories-by-categories Burt matrix **B** of all two-way cross-tabulations of the Q variables, including the "diagonal" cross-tabulations of each variable with itself, hence Q^2 cross-tables of which $\frac{1}{2}Q(Q-1)$ tables are unique. Joint correspondence analysis (JCA) is a variant of the second form where the "diagonal"cross- tables are not fitted, i.e. the $\frac{1}{2}(Q-1)(Q-2)$ cross-tables of pairs of different variables are visualized.

More details as well as algorithms for MCA and JCA are given in the forthcoming edited volume by Greenacre and Blasius (2006).

4 Data sets used as illustrations

Three data sets will mainly be used to illustrate the issues discussed in this paper:

1. Data set "author", available in the program R (R Development Core Team 2005): a 12×26 matrix with the rows = 12 texts, which form six pairs, each pair by the same author, and columns = 26 letters of the alphabet, *a* to *z*. The data are the counts of these letters in a sample of text from each of the books (or chapters), approximately 8000-10000 letter counts for each.
2. Data set "benthos", from a North Sea environmental monitoring survey: a 10×13 matrix with the rows = 10 species s1, …, s10, and columns = 13 sites (1 to 11 are polluted sites close to an oilfield, R1 and R2 are reference sites lying far away). The data are species counts at each site in a fixed-volume sample from the sea-bed.
3. Data set "mother", from the International Social Survey Program on Family and Changing Gender Roles II (ISSP 1994): the responses of 33123 respondents to four questions on whether mothers should work or stay at home at four respective stages of their married lives.

The symmetric CA maps of data sets "author" and "benthos" are given in Figs. 1 and 2 respectively. In Fig. 1, even though the total inertia is tiny (0.0184) there is still a surprisingly clear pattern in the positions of the 12 books, where each pair of texts by the same author tends to lie in the same area. In Fig. 2, the reference sites are separated from the polluted sites, which themselves form a diagonal spread from site 11 in the upper left to sites 2 and 4 in the lower right, with corresponding spread of species.

5 Measuring variance and comparing different tables

If two tables have the same number of rows and columns, then their inherent variances can be compared using their respective total inertias, but it is not obvious how to proceed when the tables are of different sizes. This matter is of crucial importance in CA when tables from different sources are analysed jointly, in which case some type of table standardization is necessary. For example, Pagés and Bécue-Bertaut (2006) apply the strategy common in multiple factor analysis (MFA) of using the first principal inertia (i.e., eigenvalue) λ_1 of each table as a measure of variance, but the reason for this choice appears to be arbitrary.

If it is assumed that $I \geq J$, then for a fixed number J of columns, the total inertia is again a reasonable measure of variance, since the dimensionality of such tables is a constant J–1. The real problem is when the

dimensionalities of tables being compared are different. If we knew which dimensions reflected "signal" as opposed to "noise" we could compare the inertias of signal, but this decision is often subjective.

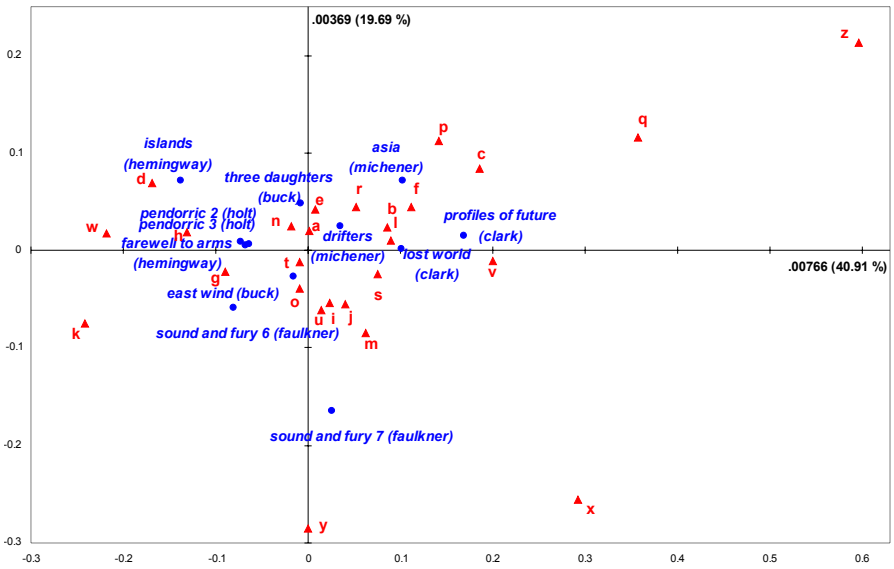

Fig. 1 Symmetric CA map of "author" data: first two principal axes; total inertia = 0.01874

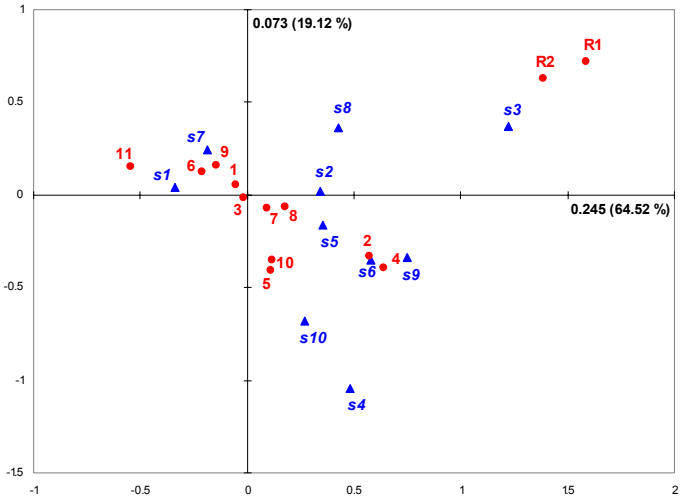

Fig. 2. Symmetric CA map of "benthos" data: first two principal axes; total inertia=0.3798

Greenacre (2006a) demonstrates with several examples that a reasonable compromise is to compare accumulated inertias on K^* dimensions for each table, where K^* is the minimum of the dimensionalities of the tables being compared. In the case of our two frequency tables, the "author" data has 11 dimensions and the "benthos" data 9, so we compare accumulated inertias in 9 dimensions in both, calculated as 0.01836 and 0.3798 respectively. Since 0.3798/0.01836 = 20.7 we can say that the "benthos" data has just over 20 times more variability than the "author" data.

6 The myth of the influential outlier

Many authors have criticized CA, in particular the use of the chi-square distance, for being too sensitive to rare categories. For example, Rao (1995: p.45) says that "since the chi-square distance uses the marginal proportions in the denominator, undue emphasis is given to the categories with low frequencies in measuring affinities between profiles". Legendre (2001: p. 271) says that "a difference between abundance values for a common species contributes less to the distance than the same difference for a rare species, so that rare species may have an unduly large influence on the analysis." My view is that in almost all cases this criticism is unfounded, in fact it is the method's ability to handle large sparse data matrices which has made it so popular in fields such as archeology and ecology. What gives rise to these criticisms is the fact that rare categories usually lie far out on the CA map, and the phenomenon of outliers is generally associated with high influence. But in CA each point has a mass and these outlying points – being established by very low frequencies – have very low mass, which reduces their influence. An inspection of the contributions to inertia of individual points gives the true story about influential points.

Both our examples contain some very low frequency columns. For example, in the author data the rarest letters are: *q* (0.07% occurrence), *j* (0.08%), *z* (0.08%) and *x* (0.1%), while all other letters occur 1% or more. Of these Figure 1 shows *q, z* and *x* to be outlying, which might suggest that these three letters have high influence in the map. However, an inspection of the contributions of these letters to the axes shows that they have contributions of 1.1%, 3.7% and 1.3% respectively to the first axis and 0.2%, 1.0% and 2.1% to the second. The major contributors are: to the first axis *d* (17.0%), *w* (16.1%), *h* (14.6%), and *c* (10.2%), and to the second axis *y* (48.5%) (note that *y* is not so rare, with a frequency of

occurrence of 2.2%). Thus, if we removed *q, z* and *x*, the map would hardly change, thus countering the belief that these outlying points have high influence.

The argument that rare categories greatly affect the chi-square distance between rows is similarly dispelled. In Fig. 1 we can see that the two books Islands (Hemingway) and Profiles of Future (Clarke) lie the furthest apart on the first axis, so their interprofile distance should be the most affected by these rare outlying letters. We calculated the square of their chi-square distance in the full space to be 0.1020, with the sum of the contributions of the letters *q*, *z* and *x* to this distance equal to 0.0077, which is a modest percentage contribution of 7.6%. Hence these two books will still be far apart even if these letters were removed from the analysis.

There is a similar result in the case of the benthos data. The first five species account for 93.5% of the counts in the table, while the last five species (*s6* to *s10*) account for the remaining 6.5%. The contributions of these five rare species to the first and second axes are jointly 6.2% and 12.5% respectively, even though in Fig. 2 their positions appear as spread out as the more commonly occurring species.

The phenomenon nevertheless remains that low frequency points are often situated in outlying positions in the map because of their unusual profiles – this is an issue that is bound up with the decision how to scale a CA map, which is the subject of the next section.

7 The scaling problem in CA

The scaling problem in CA has much in common with that of the biplot, summarized briefly here. In a biplot a matrix $\mathbf{M}$ ($I \times J$) is approximated by the product of two matrices $\mathbf{AB}^{\mathsf{T}}$, which we can write as: $\mathbf{M} \approx \mathbf{AB}^{\mathsf{T}}$. In our context the approximation is by least squares and the solution is encapsulated in the SVD: $\mathbf{M} = \mathbf{U}\mathbf{D}_{\sigma}\mathbf{V}^{\mathsf{T}}$. For the two-dimensional (rank-2) case, $\mathbf{A}$ ($I \times 2$) and $\mathbf{B}$ ($J \times 2$) are obtained from the first two columns of $\mathbf{U}$ and $\mathbf{V}$ and corresponding singular values, written in scalar notation as:

$$m_{ij} \approx a_{i1}b_{j1} + a_{i2}b_{j2} = \sigma_1 u_{i1}v_{j1} + \sigma_2 u_{i2}v_{j2} \tag{8}$$

The scaling "problem" is how to partition σ_1 and σ_2 between the left and right vectors. In general, this partitioning is as follows:

$$a_{i1} = \sigma_1^{\gamma} u_{i1} \quad a_{i2} = \sigma_2^{\gamma} u_{i2} \qquad b_{j1} = \sigma_1^{1-\gamma} v_{j1} \quad b_{j2} = \sigma_2^{1-\gamma} v_{j2}$$

i.e., a γ power of the singular value is assigned to the left singular vector and a $(1-\gamma)$ power to the right singular vector. Gower (2006) calls solutions with such scalings the "γ-family".

In the practice of biplots there are two common choices: (i) $\gamma = 1$, i.e. scale the row coordinates by respective singular values – this is the row asymmetric map, also called "row principal" in SPSS, or "row-metric-preserving" (RMP) biplot by Gabriel (1971); or (ii) $\gamma = 0$, i.e. scale the column coordinates by the singular values – this is the column asymmetric map, or "column principal", or "column-metric-preserving" (CMP):

row asymmetric (RMP): $[a_{i1}, a_{i2}] = [\sigma_1 u_{i1}, \sigma_2 u_{i2}]$ $[b_{j1}, b_{j2}] = [v_{i1}, v_{i2}]$

column asymmetric (CMP): $[a_{i1}, a_{i2}] = [u_{i1}, u_{i2}]$ $[b_{j1}, b_{j2}] = [\sigma_1 v_{i1}, \sigma_2 v_{i2}]$

When the matrix **M** is of a cases-by-variables form, these two biplots have been called the *form biplot* and *covariance biplot* respectively (Aitchison and Greenacre, 2002). An alternative scaling, seldom used, is to scale both row and column coordinates by the square root of the singular values (i.e., γ=½) , but this is neither RMP nor CMP. In my terminology (Greenacre 1984, 1993a) symmetric scaling is when both rows and columns are scaled by the singular values, giving a map that is both RMP and CMP but not in the γ-family and thus, strictly speaking, not a biplot:

symmetric (RMP & CMP): $[a_{i1}, a_{i2}] = [\sigma_1 u_{i1}, \sigma_2 u_{i2}]$ $[b_{j1}, b_{j2}] = [\sigma_1 v_{i1}, \sigma_2 v_{i2}]$

The symmetric map is a convenient choice since both row and column points have the same sum-of-squares on each axis k, equal to the part of the variance along that axis: $\sum_i (\sigma_k u_{ik})^2 = \sum_j (\sigma_k v_{jk})^2 = \sigma_k^2$. When drawing the asymmetric map, however, the sum-of-squares of each set of coordinates can be very different, in which case two different scales have to be used (see, for example, the function `biplot` in the R package `MASS`).

In asymmetric maps, the coordinates which have been scaled by the singular values (i.e., principal coordinates), are drawn as points, whereas the unscaled coordinates (standard coordinates), are often depicted using arrows drawn from the origin of the map. As a general rule, points in a map have an interpoint distance interpretation, whereas arrows indicate directions, or "biplot axes" onto which the other set of points (in principal coordinates) can be projected to obtain estimations of the data values m_{ij}. These biplot axes can be calibrated in the units of the data (see Gabriel and Odoroff 1990, Greenacre 1993a, Gower and Hand 1996).

NOTE: Since the interpretation is in terms of distances and projections, an *aspect ratio* of 1 should be respected when drawing the maps, i.e. a unit on the horizontal axis should be physically equal to a unit on the vertical axis.

The above scheme can be carried over to the CA case, with several nuances as we shall see (note that in SPSS categories, the "symmetric normalisation" refers to the case γ=½ and not to what I call the symmetric

map, which in SPSS is called "principal normalisation"). The generalized form of the SVD in the case of CA (see Eqs. (2)–(7)), leads to the following form for (8), called the "reconstitution formula" since it estimates the data values from the map:

$$\left(\frac{p_{ij} - r_i c_j}{r_i c_j} \right) \approx \alpha_1 x_{i1} y_{j1} + \alpha_2 x_{i2} y_{j2} \tag{9}$$

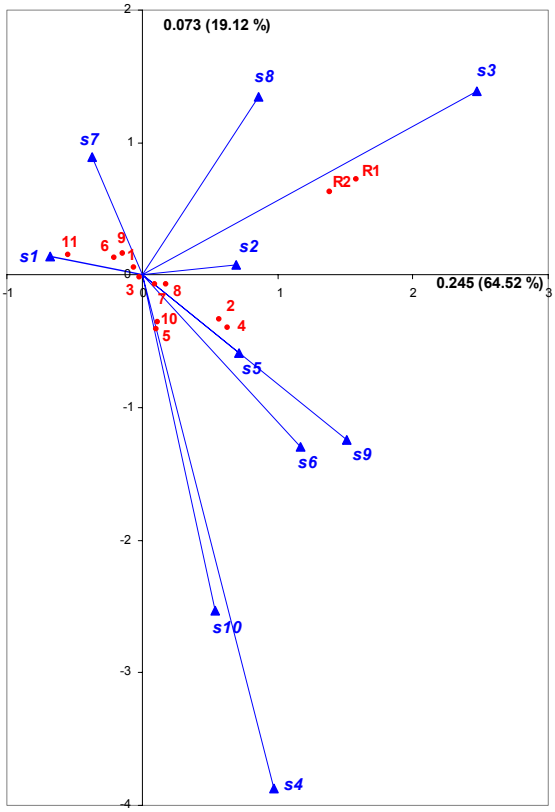

Fig. 3 Column asymmetric CA map of "benthos" data: first two principal axes, with column (profile) points in principal coordinates and row (vertex) points in standard coordinates connected to the origin as biplot axes**.**

On the right hand side of (9) we have the singular values from (3) and the elements x_{ik} and y_{ik} (k = 1,2) of the first two columns of the standard coordinate matrices **X** and **Y** defined by (6) and (7). Hence, if we assign the singular values to **X**, we obtain the row principal coordinates **F** defined in (4), and hence the asymmetric row map of the CA, which is approximating the chi-square distances between row profiles. On the other

hand, if we assign the singular values to **Y**, we obtain the column principal coordinates (5) and thus the column asymmetric map, which approximates the chi-square distances between column profiles. If we scale both row and column standard coordinates by the singular values then we obtain the symmetric map, shown in Figs. 1 and 2, but no scalar product property as in (9). However, Gabriel (2002) has shown that the scalar product property is not severely degraded in the symmetric map.

There are two aspects specific to CA which distinguish it from the general biplot scheme described above. The first aspect is that in CA the standard coordinates represent actual points which are theoretically possible to observe, namely the unit profile vectors which are *vertices* of the simplex space of the profiles: [1 0 0 ... 0], [0 1 0 ... 0], etc. For example, Fig. 3 shows the column asymmetric map of the "benthos" data, with column profiles in principal coordinates, identical to Fig. 2, and row vertices in standard coordinates. The rows are necessarily more dispersed than the columns and, compared to Fig. 2, the row configuration is stretched out more in the vertical than the horizontal direction (positions of the row vertices in Fig. 3 are those of the row profiles in Fig. 2 *divided* by the singular values on axes 1 and 2, $\sqrt{0.245}$ and $\sqrt{0.073}$ respectively).

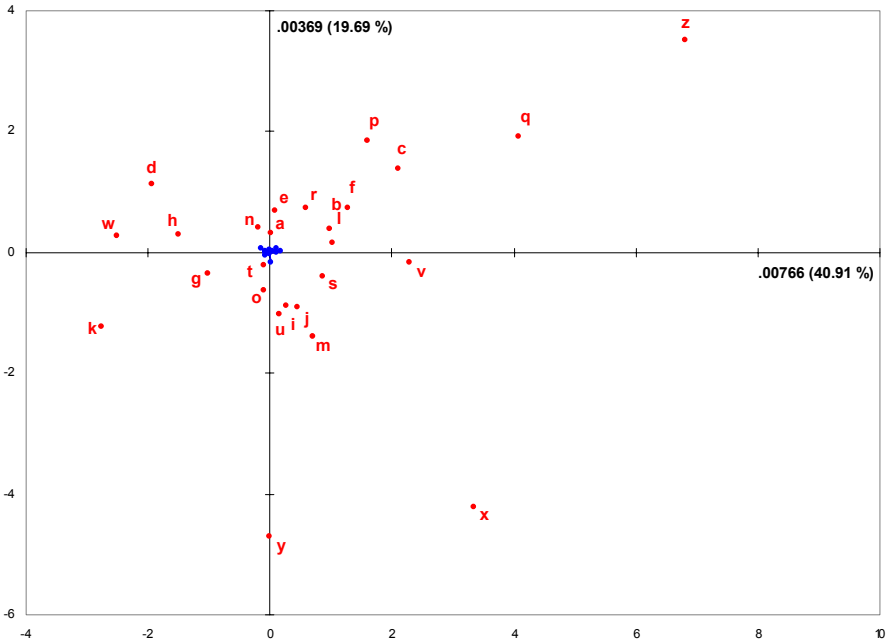

Fig. 4 Row asymmetric CA map of "author" data: first two principal axes, with book profiles in principal coordinates close to origin of the map and letter vertices in standard coordinates, showing very low inertia in the data**.**

The map in Fig. 3 looks acceptable since the principal inertias 0.245 and 0.073 are relatively high. The situation is completely different for the "author" data, however, since the principal inertias are tiny. In the asymmetric map of Fig. 4, the book profiles form a small smudge at the centre of the map compared to the letter vertices, a striking geometric demonstration of the very low inertia of these data. Hence, for good visualization of the books in a biplot, some change of scale of the column points is required. This brings us to the second specific aspect of CA, namely the presence of the masses r_i and c_j in the matrix being represented, as given by (9).

We can write (9) from the "row profile point of view" as follows, grouping the principal and standard coordinates:

$$\left(\frac{p_{ij}}{r_i} - c_j\right) \Big/ c_j \approx (\alpha_1 x_{i1}) y_{j1} + (\alpha_2 x_{i2}) y_{j2} \tag{10}$$

that is, the asymmetric map biplots the differences between the row profile elements and their averages, expressed relative to the averages c_j. As an alternative, we could recover actual profile values directly, in which case the mass c_j is carried over to the right hand side of (10) and absorbed in the standard coordinates as follows:

$$\left(\frac{p_{ij}}{r_i} - c_j\right) \approx (\alpha_1 x_{i1})(c_j y_{j1}) + (\alpha_2 x_{i2})(c_j y_{j2}) \tag{11}$$

(note that the symbol ≈ is used repetitively and signifies the weighted least-squares approximation in the original SVD). The form (11) suggests a biplot using principal row coordinates $[\alpha_1 x_{i1},\ \alpha_2 x_{i2}]$ and column standard coordinates rescaled by the column masses $[c_j y_{j1},\ c_j y_{j2}]$. In this biplot (not shown here – see Greenacre 2006a) the column points have been pulled in by varying amounts depending on the values of their relative frequencies (masses) c_j. Thus the rare letter *z* is practically at the origin, while the common letter *e* is now more prominent. This biplot scaling for CA is exactly the one proposed by Gabriel and Odoroff (1990).

But this scaling goes to the other extreme of pulling in the column points too much and, in any case, we already know that the deviations between the profile elements and their average on the left hand side of (11) will be high for frequent letters and low for rare letters, so the lengths of the vectors are still without interest. An obvious compromise between (10) and (11) is to represent standardized differences:

$$\left(\frac{p_{ij}}{r_i} - c_j\right) \Big/ c_j^{1/2} \approx (\alpha_1 x_{i1})(c_j^{1/2} y_{j1}) + (\alpha_2 x_{i2})(c_j^{1/2} y_{j2}) \tag{12}$$

i.e., the standard column coordinates are rescaled by the *square roots* of the column masses. This map is shown in Fig. 5 and it is clear that the

common scale for rows and columns is adequate for the joint visualization. This scaling mimmicks the idea in PCA where standardized values are recovered in the biplot. The distance between tic-marks on a biplot vector is inversely related to the length of the vector (Greenacre 1993a, 1993b; Aitchison and Greenacre 2002), so the tic marks on the "*y*" vector will be closer, indicating a higher variance in the profile values of this letter (i.e., overdispersion compared to the variance estimated from the mean). Another advantage of the above scaling is that the squared lengths of the column vectors are related to their respective contributions to principal inertias, both along axes and in the plane.

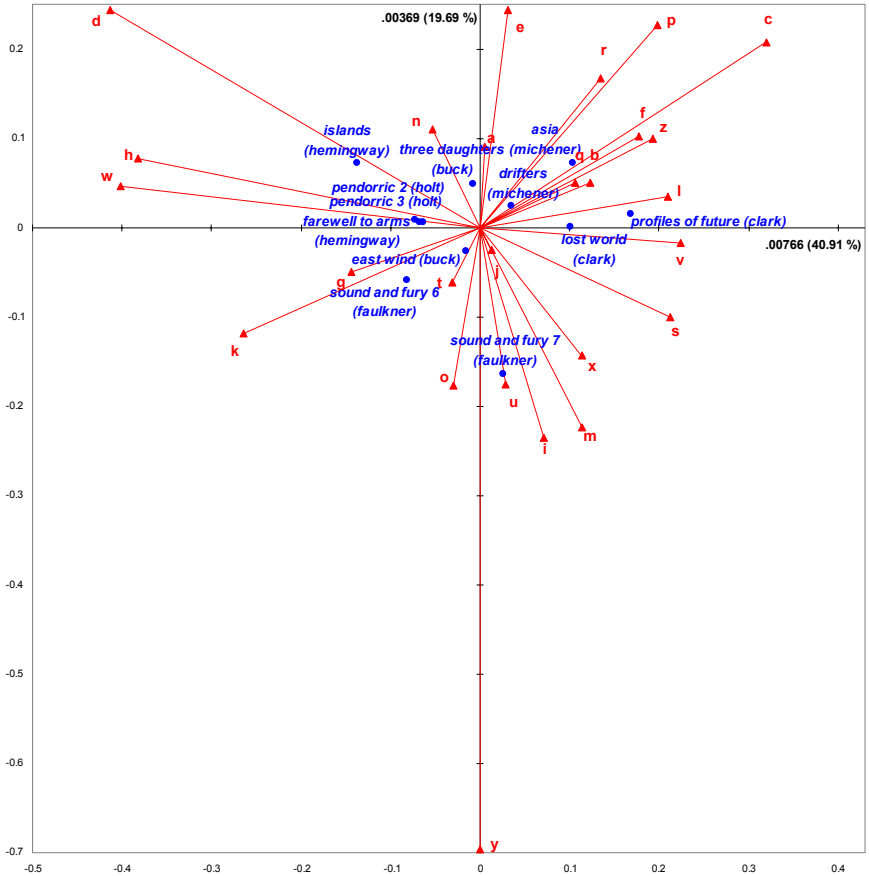

Fig. 5 Standard CA biplot of the "author" data, with letter (column) points in standard coordinates rescaled by square roots of masses, and book (row) points in principal coordinates.

In Fig. 6 we show the "benthos" data in a similarly scaled CA map, this time with the columns in principal and the rows in rescaled standard

coordinates, the column version of (12). It is clear from Figs. 5 and 6 that this scaling functions well irrespective of the large difference in the total inertias of the two data sets. Since these are biplots of standardized profile values, we call these maps *standard CA biplots.* It should be emphasized that there is no distance interpretation between the column points (letters) in Fig. 5, neither between the row points (species) in Fig. 6 – it is the direction and lengths of these point vectors that have meaning.

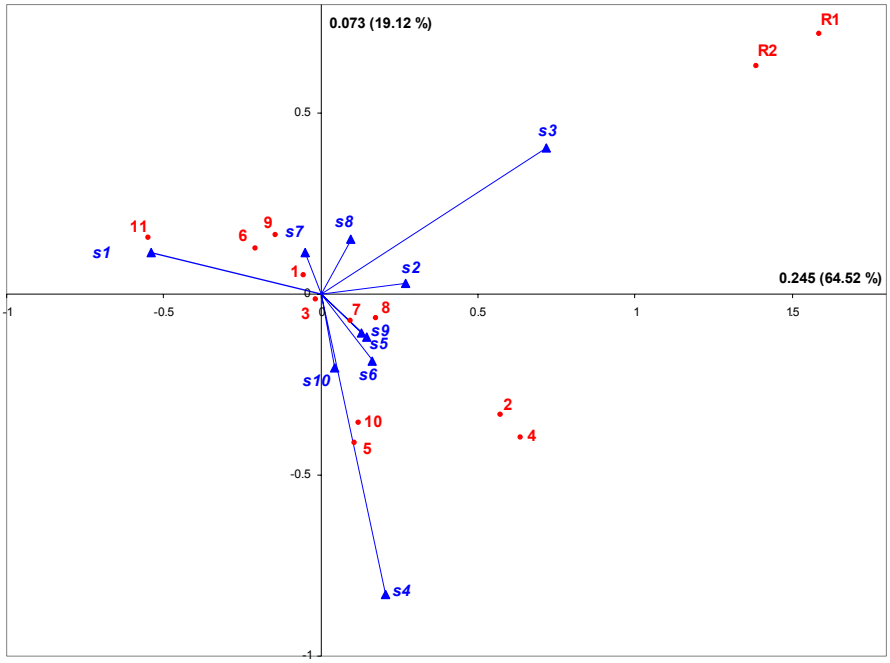

Fig. 6 Standard CA biplot of the "benthos" data, with sites (columns) in principal coordinates and species (rows) in standard coordinates rescaled by square roots of their masses.

8 To rotate or not to rotate

The short answer to this question is "yes, but why?". Rotation of CA solutions is possible, just as the solution of any of the factorial family of analyses can be rotated, but three questions need answering before we attempt it in CA: (i) does rotation have any meaning in CA geometry? (ii) is rotation necessary in the context of the data? (ii) if rotation is justified, which CA coordinates need to be rotated and how?

First, what does rotation mean in the case of CA? In general, rotation is applied so that subsets of "variables" coincide more closely with the dimensions of the solution subspace, leading to a simpler interpretation of the dimensions. The only consequence is that the percentages of variance explained are redistributed along the newly rotated axes, while still conserving all the variance explained by the solution as a whole. In simple CA we do not have a set of variables as such, but rather a multicategory row "variable" and a multicategory column "variable". These often have different roles, one serving as a variable in the usual sense, used to interpret the solution space, the other defining groups whose positions are depicted in the "variable" space. But the analogy between variables in PCA/factor analysis and categories of a single variable in CA is tenous to say the least. The full CA space is not the unlimited vector space of real numbers but the simplex space of the profiles, i.e. vectors $[v_1\ v_2\ \ldots\ ]$ of nonnegative numbers with the unit constraint: $\Sigma_j\ v_j = 1$, delimited by the unit profiles as vertices of a simplex. Row and column points are both centred within this space so we obtain for each set a fan of points radiating out from the centre in all directions, a situation far different from the usual one in PCA/factor analysis, where only the cases are centred and the variables are free to point in any direction depending on their correlation structure. From this point of view it seems unlikely that some categories would form patterns at right-angles to one another and thus be candidates for rotation to "simple structure".

Second, in my over 30 years' experience of CA, I have hardly ever encountered a situation where rotation would have been necessary or useful. In both examples discussed in this paper there is no benefit at all in rotating the solution (see the vectors for the species and letters in Figs. 3 and 5). The rare exceptions have invariably been in an MCA context. For example, Fig. 7 shows an MCA of 10 categorical variables which include a missing data category for each variable. All 10 missing categories are in a bunch in the upper right side of the map, opposing all the substantive categories lying in a diagonal band. If we are not interested in the missing data categories, it would be very convenient if these categories lay close to a principal axis, since then we could simply ignore that axis and look at projections on other pairs of axes to interpret the substantive categories.

Third, supposing that rotation were justified in some rare cases, how could a formal rotation of the solution be made? Van der Velden (2003) considers rotations of principal coordinates or standard coordinates of the rows or columns, and even the simultaneous rotation of row and column coordinates. In my opinion the choice is entirely dependent on the substantive nature of the data. If the rows and columns can be considered in a cases-by-variables format (e.g., books=cases and letters=variables for

"author" data, or sites=cases and species=variables for "benthos" data) then rotation of the "variables" can be considered, but not the "cases", since it is the "variable" categories that are used to name the axes. The standard coordinates of the "variable" categories are analogous to the projections of unit vectors onto the principal axes (cf. factor loadings in PCA/factor analysis) and could be candidates for rotation to simple orthogonal or oblique structure. There seems to be little justification for rotating principal coordinates. As far as joint rotation of row and column coordinates, this would only be justified when both variables play symmetric roles, as in the case of MCA: for example two questions in a questionnaire such as in the "mother" example. There is more justification for rotating coordinates in MCA (Adachi 2004), especially the constrained form known as non-linear principal component analysis, than in CA.

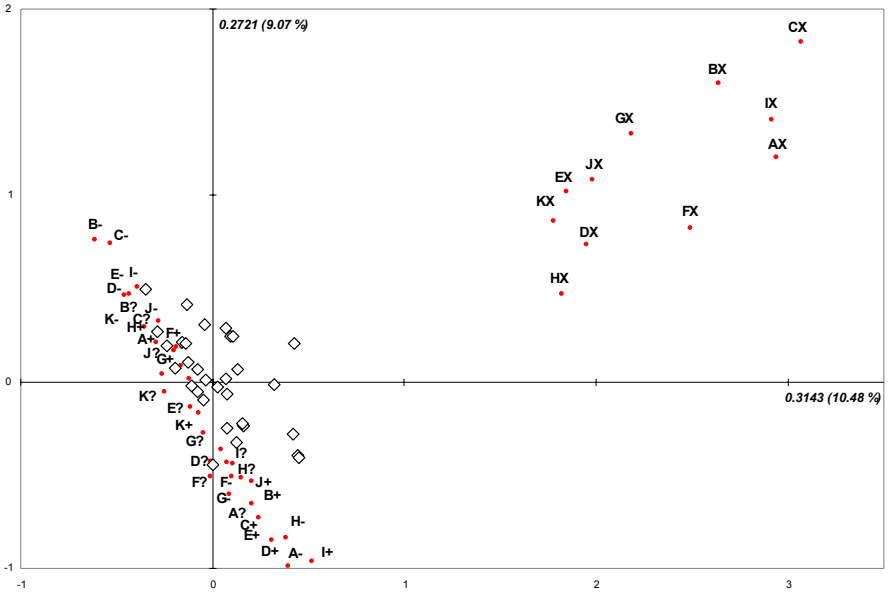

Fig. 7 MCA map of response categories to 11 questions, labelled A B, C, ..., K, plus a character "+" (agree), "?" (unsure), "-" (disagree) or "X" (missing). The diamonds correspond to supplementary demographic categories (abbreviations not given here).

A technical issue in rotating CA solutions is how the masses should be taken into account in an axis rotation, since we are less interested in how well a low-frequency category coincides with an axis than a high-frequency category. Thus, the rotation criterion should be weighted: for example, a (weighted) varimax rotation of the J column standard coordinates would maximize the following function:

$$\sum_j \sum_s c_j^2 (\tilde{y}_{js}^2 - \frac{1}{J} \sum_{j'} \tilde{y}_{j's}^2)^2 \qquad (14)$$

where $\tilde{y}_{js}$ is the rotated standard coordinate, i.e. an element of $\tilde{\mathbf{Y}} = \mathbf{YQ}$, where **Q** is an orthogonal rotation matrix. The form (14) suggests that an unweighted rotation could be performed on the rescaled coordinates $c_j^{1/2} \tilde{y}_{js}$ used in the standard CA biplot defined in Sect. 7, giving an additional justification for this scaling.

9 Statistical significance of results

Although CA is primarily a descriptive technique, often criticized for not being inferential, there are several possibilitites for investigating the statistical variability of the results. If the data are a contingency table, arising from multinomial random sampling, principal inertias can be formally tested for significance, using the multivariate normal approximation to the multinomial and consequent distribution of eigenvalues of the covariance matrix (Lebart 1976, Greenacre 1984: Sect. 8.1). In addition, when the bilinear model (9) is estimated by maximum likelihood rather than weighted least squares, a range of hypotheses can be tested (Gilula and Haberman 1986, Vermunt and Anderson 2005).

Greenacre (1984) introduced the notions of "internal stability" and "external stability" for visualization methods such as CA. Internal stability refers to the data set at hand, without reference to the population from which the data might come, and is thus applicable in all situations, even for population data or data obtained by convenience sampling. Here we are concerned how our interpretation is affected by the particular mix of row and column points determining the map. Would the map change dramatically (and thus our interpretation too) if one of the points is omitted, for example one of the species in our second example? Such a question is bound up with the concept of influence and how much each point influences the rotation of the principal axes in determining the final solution. The numerical results of CA known as "inertia contributions" provide indicators of the influence of each point. The principal inertia $\lambda_k = \alpha_k^2$ on the k-th principal axis can be decomposed into parts for the row points and, separately, into parts for each column point. If a point contributes highly to an axis, then it is influential in the solution. Of particular interest are points with low mass that have high influence: these would be influential outliers, as opposed to the non-influential outliers described in Sect. 6. Greenacre (1984) gives some rules about determining

the potential rotation of principal axes if a point were removed, which is one way of quantifying the influence in graphical terms.

External stability is equivalent to the sampling variability of the map, and is applicable when the data arise from some random sampling scheme. In order to investigate this variation, we need to know the way the data were collected. Meulman (1982) proposed using a bootstrapping procedure to calculate several, say N, replicates of the data matrix, where N is typically chosen to be of the order of 100 to 500. For example, in the "author" case, the data for each book is regarded as a multinomial population from which as many letters are selected at random, with replacement, as originally sampled. Having established N replicates, there are two ways to proceed. Greenacre (1984) proposed using the replicates as supplementary row and column points in the analysis of the original matrix, leading to a sub-cloud of N points for each row and column; this strategy is called the "partial bootstrap" by Lebart (2006). The alternative, proposed by Meulman (1982) is to re-run the CA on each of the replicate matrices and put all solutions together using, for example, Procrustes analysis, with the original configuration as a target, or alternatively using generalized Procrustes of all the replicate configurations.

The partial bootstrap was performed on the "author" data, with $N = 100$, so that 100 replicates of each book's profile are projected onto the map of the original table. Rather than draw all the replicates, the dispersion of each subcloud is summarized in Fig. 8 by its convex hull. Since the convex hull is sensitive to outlying replicates, it is usually "peeled", that is the convex hull of points is removed and the convex hull of the remaining points is drawn (see Greenacre 2006a for an example). To obtain a convex hull including 95% of the points, 5% of the most outlying points need to be removed from each subcloud. Alternatively, confidence ellipses with 95% coverage can be calculated by finding the principal axes of each subcloud of points and then drawing an ellipse with axes having major and minor radii equal to a scale factor times the standard deviation (square root of eigenvalue) on each axis, where the scale factor depends on the sample size (see Sokal and Rohlf 1981: pp. 504–599). The confidence ellipse approach assumes that the pair of coordinates for each subcloud of replicates follows a bivariate normal distribution, an assumption which is not necessarily true. When profiles are at the extremes of the profile space, which is an irregular simplex, replicated profiles can lie on one of the faces of the simplex, generating straight lines in their projections onto subspaces. In this case, confidence ellipses would exceed the permissible profile region and include points that are impossible to realize. Convex hulls would include these straight line "barriers" in the space and would thus be more realistic.

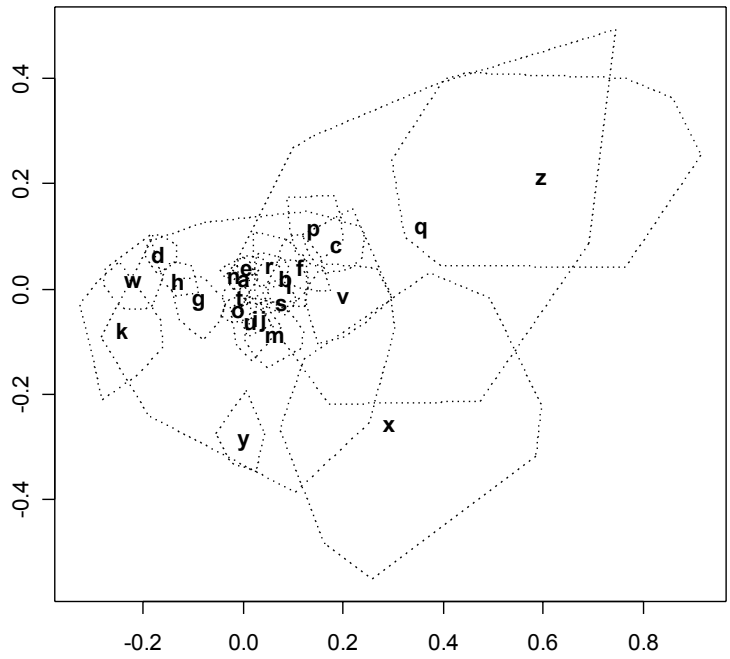

Fig. 8 Convex hulls of points in Figure 10, showing letters in their principal coordinate positions in the original map.

A non-statistical approach for elliptical representation of scatters of points is given by Silverman and Titterington (1980), who describe an algorithm for finding the ellipse with smallest area containing the points.

Finally, Gifi (1990: 408–417) proposes using the delta method for calculating asymptotic variances and covariances of the coordinates, which also leads to confidence ellipses. This methodology, which uses the partial derivatives of the eigenvectors with respect to the multinomial proportions, relies on the assumption of independent random sampling. Although this is not satisfied in either of the examples presented here, we calculated the confidence regions using this approach (not shown here – see Greenacre 2006a). The results are quite similar, giving confidence ellipses of about the same size but more spherical in shape, indicating less correlation than in the replicates based on bootstrapping.

10 Loose ends in MCA and JCA

Fig. 9 shows the optimal two-dimensional MCA of the "mother" data set, after the scale adjustment to be described in Sect. 10.1 below. Along the

first axis the "don't know" (DK) responses labelled *?* oppose all the other responses "stay at home" (*H*) , "work part-time" (*w*) and "work full-time" (*W*) for the four questions. Supplementary points for the 24 countries and the 12 gender-age groups are indicated by diamonds, unlabelled apart from CDN (Canada) which had a relatively high frequency of DK's.

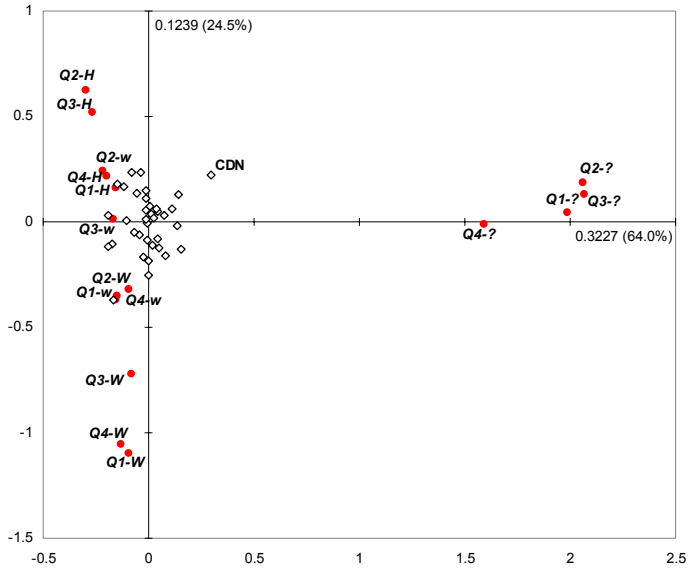

Fig. 9 MCA of "mother" data: first two principal axes, showing active response category points and supplementary points.

Figs. 10 and 11 show the active and supplementary points with respect to axes 2 and 3, in separate maps for easier visualization. The arch ("horseshoe") effect is clearly visible, with attitudes that mothers should stay at home, even without children, in upper right, and should work full-time, even with children, in upper left.

Most of the loose ends in MCA come about because the geometry of simple CA does not generalize easily to the multivariate case. Greenacre (1988) proposed joint correspondence analysis (JCA) as a more natural extension of CA. Space precludes a detailed discussion here, so a summary is given.

10.1 Variance explained in MCA

In Figs. 9 and 10 the percentages of inertia for the first three principal axes are given as 64.0%, 24.5% and 6.8% respectively. The inertias on each axis have been adjusted according to Greenacre (1988), leading to more

realistic percentages of inertia explained along each axis (if this adjustment is not performed, the usual percentages obtained from MCA of the indicator matrix **Z** would be 22.5%, 17.1% and 13.0% respectively). Greenacre (1993c) proved that the adjusted percentages are a lower bound on the percentages obtained using JCA. I also conjecture that all off-diagonal tables of the Burt matrix **B** can be perfectly reconstructed in a JCA of K^* dimensions, where K^* is the number of dimensions for which $\lambda_k > 1/Q$, where λ_k is the k-th inertia (eigenvalue) in the analysis of **Z**. I do not agree with the corrections proposed by Benzécri (1979), since these imply that 100% of inertia of can be explained by the K^*-dimensional MCA solution, which is easily shown to be false by counterexample.

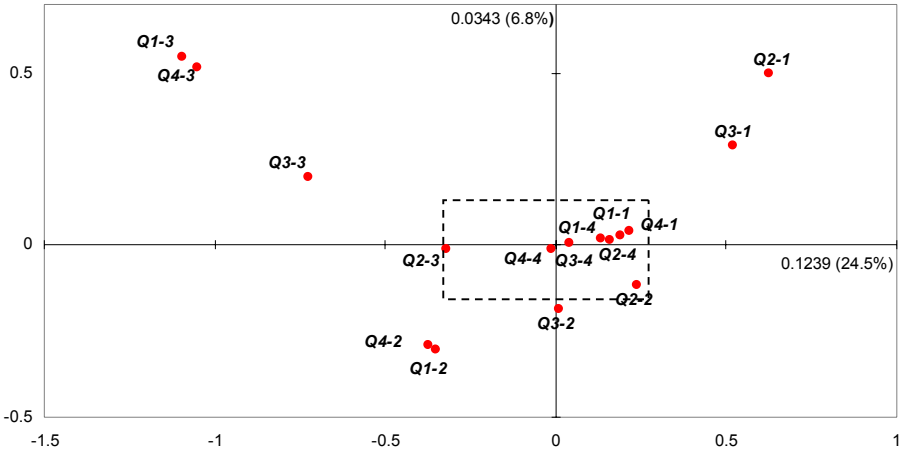

Fig. 10 MCA of "mother" data: principal axes 2 (horizontal) by 3 (vertical), showing response category points.

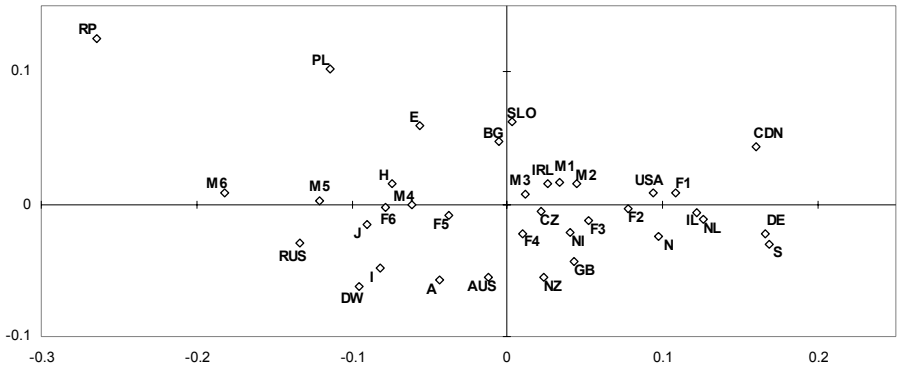

Fig. 11 MCA of "mother" data: supplementary category points as averages for countries and gender-age groups (F1=youngest female,...,F6=oldest female; similarly, M1,...,M6); enlargement of the box in Fig.10.

10.2 Inertia contributions in MCA and JCA

The standard output of CA expresses elemental inertias relative to the principal inertia on an axis (*contributions absolues*) or relative to the inertia of a point (*contributions relatives*, or squared cosines, or squared correlations) – see, for example, Greenacre (1993a: Chap.11). In making the same calculations for MCA and JCA, the former contributions are still valid, although it is useful to add up these contributions over the categories of each variable, in which case the summed contribution can be transformed to squared correlations, the so-called "discrimination measures" of homogeneity analysis (Gifi 1990) – see Greenacre (1993a: Chap.18). For the latter type of contributions, however, the same problem as in Sect.10.1 exists but at the level of a single point.

10.3 Supplementary points in MCA

Fig.11 shows the supplementary points for the 24 countries and the 12 gender-age group categories. Since there are many possible scalings for these points, Greenacre (2006b) justifies the universal use of displaying the average respondent points (i.e., rows) in each subsample. Averages do not actually have to be calculated from scratch, since it can be shown that they are identical to the principal coordinates of the supplementary points as columns in the analysis of the indicator matrix $\mathbf{Z}$, multiplied by the respective square roots of the principal inertias λ_k of $\mathbf{Z}$.

10.4 The arch, or "horseshoe" effect

The arch effect, which is partially or in extreme cases an artefact of the CA simplex geometry, is not a drawback, in my opinion, since it creates a second dimension of what I call "polarization", opposing groups which have a combination of extreme responses against those with middle responses that follow the general gradient along the horizontal axis (see Greenacre and Pardo (2006) for a good illustration of this phenomenon).

Acknowledgments

This research has been supported by the Fundación BBVA in Madrid, Spain, and I wish to express my thanks to the director-general, Prof. Rafael Pardo, for his encouragement with this work. Analyses were performed using XLSTAT 2006, and I appreciated the fruitful collaboration with Thierry Fahmy of Addinsoft to

improve the CA and MCA modules of this statistical package. R routines for CA, MCA and JCA are described by Nenadić and Greenacre (2005), and will soon appear in the CRAN library.

References

Adachi, K. (2004) Oblique Promax rotation applied to the solutions in multiple correspondence analysis. *Behaviormetrika 31*: 1–12.

Aitchison, J. & Greenacre, M.J. (2002) Biplots of compositional data. *Applied Statistics 51*: 375–392.

Benzécri, J.-P. (1973) *L'Analyse des Données. Tôme I: l'Analyse des Correspondances.* Dunod, Paris.

Benzécri, J.-P. (1979) Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Données 3*: 55–71.

Gabriel, K.R. (1971). The biplot-graphical display with applications to principal component analysis. *Biometrika 58*: 453–467.

Gabriel, K.R. (2002). Goodness of fit of biplots and correspondence analysis. *Biometrika 89*, 423–436.

Gabriel, K.R. and Odoroff, C.L. (1990). Biplots in biomedical research. *Statistics in Medicine 9*: 423–436.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, Chichester, UK.

Gilula, Z. and Haberman, S. J. (1986). Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association 81*: 780–788.

Gower, J.C. and Hand, D.J. (1996). *Biplots*. Chapman and Hall, London.

Gower, J.C. (2006). Divided by a common language: analysing and visualizing two-way arrays. In M.J. Greenacre and J. Blasius (eds), *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, London, forthcoming.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.

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

Greenacre, M.J. (1993a) *Correspondence Analysis in Practice.* Academic Press, London.

Greenacre, M.J. (1993b). Biplots in correspondence analysis. *Journal of Applied Statistics 20*: 251–269.

Greenacre, M.J. (1993c).Multivariate generalizations of correspondence analysis. In C.M. Cuadras and C.R. Rao (eds), *Multivariate Analysis: Future Directions 2*, North Holland, Amsterdam, pp.327–340.

Greenacre, M.J. (1998). Diagnostics for joint displays in correspondence analysis. In J Blasius and M.J. Greenacre (eds), *Visualization of Categorical Data*. Academic Press, San Diego, pp. 221–238.

Greenacre, M.J. (2006a). Tying up the loose ends in simple correspondence analysis. Working Paper no. 940, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona.

Greenacre, M.J. (2006b). From simple to multiple correspondence analysis. In M.J. Greenacre and J. Blasius (eds), *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, London, forthcoming.

Greenacre, M.J. and Blasius, J. (2006) (eds) *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, London, forthcoming.

Greenacre, M.J. and Lewi, P.J. (2005) Distributional equivalence and subcompositional coherence in the analysis of contingency tables, ratio-scale measurements and compositional data. Working Paper no. 908, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona.

Greenacre, M.J. and Pardo, R. (2006). Subset correspondence analysis: visualizing relationships among a set of response categories from a questionnaire survey.. *Sociological Methods and Research*, forthcoming.

Hill, M.O. (1974) Correspondence analysis: a neglected multivariate method. *Applied Statistics 23*: 340–354.

ISSP (1994). *International Social Survey Program: Family and Changing Gender Roles II.* Central Archive for Empirical Social Research, Cologne, Germany.

Lebart L. (1976). The significance of eigenvalues issued from correspondence analysis. In J. Gordesch and P. Naeve (eds), *Proceedings in Computational Statistics*, Physica Verlag, Vienna, pp. 38–45.

Lebart, L. (2006). Validation techniques in multiple correspondence analysis. In M.J. Greenacre and J. Blasius (eds), *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, London, forthcoming.

Legendre, P. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia 129*: 271-280

Meulman, J. (1982). *Homogeneity Analysis of Incomplete Data*. DSWO Press, Leiden, The Netherlands.

Nenadić, O. and Greenacre, M.J. (2005) The computation of multiple correspondence analysis, with code in R. Working Paper no. 887, Departament d'Economia I Empresa, Universitat Pompeu Fabra, Barcelona.

Pagès, J. and Bécue-Bertaut, M. (2006). Multiple factor analysis for contingency tables. In M.J. Greenacre and J. Blasius (eds), *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, London, forthcoming.

R Development Core Team (2005). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`

Rao. C.R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Qüestiió 19*: 23–63.

Silverman, B.W. and Titterington, D.M. (1980). Minimum covering ellipses. *SIAM J. Sci. Stat. Comput.* 1: 401–409.

Sokal, R. R. and Rohlf, F.J. (1981). *Biometry: The Principles and Practice of Statistics in Biological Research. 2nd Edition*. W.H. Freeman & Co, New York.

Van de Velden, M. (2003). *Some Topics in Correspondence Analysis*. PhD Thesis, University of Amsterdam

Vermunt, J.K. and Anderson, C.J. (2005). Joint correspondence analysis by maximum likelihood. *Methodology 1*: 18–26.