Tying up the loose ends in simple correspondence analysis

Michael Greenacre

Department of Economics and Business
Universitat Pompeu Fabra
Ramon Trias Fargas, 25-27
08005 Barcelona
SPAIN

E-mail: `michael@upf.es`

Abstract: Although correspondence analysis is now widely available in statistical software packages and applied in a variety of contexts, notably the social and environmental sciences, there are still some misconceptions about this method as well as unresolved issues which remain controversial to this day. In this paper we hope to settle these matters, namely (i) the way CA measures variance in a two-way table and how to compare variances between tables of different sizes, (ii) the influence, or rather lack of influence, of outliers in the usual CA maps, (iii) the scaling issue and the biplot interpretation of maps, (iv) whether or not to rotate a solution, and (v) statistical significance of results.

Keywords: biplot, bootstrap, canonical correlation, chi-square distance, confidence ellipse, contingency table, convex hull, correspondence analysis, inertia, permutation test, rotation, singular value decomposition.

This research has been supported by the Fundación BBVA, Madrid, Spain.

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. Nevertheless, there are still several issues that remain unsettled and which are often the basis for misconceptions and controversy over the method's application and applicability. 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", CA is generally accepted as a way of visually displaying the association between two discrete variables, based on their cross-tabulation in the form of a two-way table of frequencies. The row and column categories are depicted in a spatial map where certain distances or scalar products may be interpreted as approximations to the original data. The most widespread source of confusion in CA is the scaling used to define the coordinates of the joint display of row and column points, and this has led to various misconceptions and doubts about CA's usefulness. Apart from this aspect, there are also differences of opinion about such issues as: the measure of variance used by CA and its use of the chi-square distance, which is at the heart of the method's theory; the apparently excessive influence of outlying points in the map; whether solutions should be rotated or not; and the statistical "significance" of the results.

After a summary of the method's theory, mainly to define terminology and notation, I shall address all these issues and propose explanations and/or solutions. Attention is restricted in this paper to "simple" (two-way) CA. Multiple correspondence analysis (MCA), which visualizes the associations among more than two categorical variables, has its own set of misunderstandings and controversies that are beyond the scope of the present paper, but which will be addressed in a follow-up publication.

2. Summary of correspondence analysis theory

In this section the theory of CA is summarised in order to define the terms and notation for the later sections. CA is a particular case of weighted principal component analysis (PCA) (Benzécri 1973, Greenacre 1984: chapter 3). In this general scheme, a set of multidimensional points exists in a high-dimensional space in which distance is measured by a weighted Euclidean metric 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 – these are often depicted as arrows rather than points, since they may be considered as directions in the biplot style of interpretation (Gower & Hand 1996; Greenacre 1993a, 2004), discussed further in Section 6. 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 frequencies, usually a two-way contingency table. Each row of $\mathbf{N}$ can be expressed relative to its respective total as a vector of relative frequencies, called a row *profile*. The row profiles thus define I points in a Euclidean space. Each profile point is weighted by its *mass*, the row margin of that row relative to the grand total of the table. Distances between row profiles are defined as *chi-square distances*, a weighted Euclidean metric where each squared difference between profile elements is weighted inversely by the average profile element, that is the corresponding column margin relative to the grand total of the table. In a completely symmetric fashion, the columns of $\mathbf{N}$ can be expressed relative to their totals as column profiles, each with a mass and with inter-profile chi-square distances, as if we applied the same description above to the transpose $\mathbf{N}^{\mathsf{T}}$ of $\mathbf{N}$. However, as we discuss later, the table is usually considered as a set of rows or a set of columns depending on the context, and this determines whether we represent row or column profiles.

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 $\mathbf{N}$ 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 $\mathbf{P}$ be the vectors $\mathbf{r}$ and $\mathbf{c}$ respectively, that is the vectors of row and column masses, and $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices of these masses. Thinking of the table as a set of rows, say, we calculate the row profiles by dividing the rows of $\mathbf{P}$ 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-squared 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 $\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}^{\mathsf{T}}$, since it can be shown that the principal axes of the profiles necessarily pass through the centroid. The dual analysis of column profiles can be defined simply by interchanging rows with columns and all associated entities, i.e. transposing the matrix $\mathbf{P}$ and repeating all the above. Thus CA can be equivalently defined as the weighted PCA of the column profiles, contained in the rows of $\mathbf{D}_c^{-1}\mathbf{P}^{\mathsf{T}}$, where distances between profiles are measured by the chi-squared metric defined by $\mathbf{D}_r^{-1}$ and the profiles are weighted by the column masses in $\mathbf{D}_c$. The centroid of the column profiles is the vector $\mathbf{r}^{\mathsf{T}}$ of row masses, hence CA of the column profiles analyses the centred matrix $\mathbf{D}_c^{-1}\mathbf{P}^{\mathsf{T}} - \mathbf{1}\mathbf{r}^{\mathsf{T}}$.

In both row and column analyses, the weighted sum of (chi-squared) distances of the profile points to their respective centroids is the same, and 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 divided by the grand total n: $\phi^2 = \chi^2/n$.

The computational algorithm to obtain coordinates of the row and column profiles with respect to principal axes, using the SVD, is as follows:

1. Calculate the matrix of standardized residuals:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{rc}^\mathsf{T})\mathbf{D}_c^{-1/2} \tag{2}$$

2. Calculate the 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)

3. Principal coordinates of rows: $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha$ (4)

4. Principal coordinates of columns: $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\alpha$ (5)

5. Standard coordinates of rows: $\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U}$ (6)

6. Standard coordinates of columns: $\mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V}$ (7)

The rows of the coordinate matrices in (4)–(7) 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* α_k^2, the square of the k-th singular value (or eigenvalue, denoted by λ_k), 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 on the principal axes.

A two-dimensional solution, say, would use the first two columns of the coordinate matrices. The following are the three most common versions of maps where rows and columns are plotted jointly, with a distance (PCA) or scalar-product (biplot) interpretation as the case may be:

1. *Symmetric map*: joint plot of principal row and column coordinates **F** and **G**

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

The joint plot of row and column standard coordinates **X** and **Y** has no justification from the point of view of interpretation. The specific interpretations of the maps will be given when discussing the scaling problem in Section 6. Since the interpretation is in terms of distances and projections, the *aspect ratio* should always be respected, i.e. a unit on the horizontal axis should be physically equal to a unit on the vertical axis. The total inertia (1) is equal to the sum of all principal inertias $\lambda_1+\lambda_2+\lambda_3\ldots$, and the inertia accounted for in a two-dimensional solution, for example, is the sum of the first two terms $\lambda_1+\lambda_2$, while the inertia not accounted for is the remainder: $\lambda_3+\lambda_4+\ldots$. These parts of inertia are usually expressed as percentages of inertia explained by each dimension, as in PCA.

3. Empirical data: “author” data and “benthos” data

Two data sets will mainly be used to illustrate the issues discussed in the remainder of this paper. The first is the “author” data set available in the program R (R Development Core Team 2005). It is obtained in R from the MASS package, by issuing the following commands:

```
library(MASS)
data(author)
```

The data form a 12 × 26 matrix with the rows representing 12 texts which form six pairs, each pair by the same author (Table 1). The columns are the 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. There are approximately 8000-10000 letter counts for each book or chapter.

The second data set is a typical set of counts in an environmental survey, where several species are counted at a set of sampling locations (Table 2). The columns represent 13 sampling sites (labelled 1 to 11, R1 and R2), and the rows represent 10 species (labelled s1, s2, …, s10). The context is in marine sampling of benthic (sea-bed) species near an oilfield in the North Sea, where the first 11 sites, the polluted sites, lie in an approximate grid around the oilfield and the unpolluted reference sites R1 and R2 lie far away from the oilfield.

The selection of the two data sets was made specifically to have one data set with very low inertia (data set “author”) and one with very high inertia (data set “benthos”). The symmetric CA maps of the two tables are given in Figures 1 and 2 respectively. In Figure 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 books by the same author tends to lie in the same area of the map. In Figure 2, the reference sites are well 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.

4. Measuring the variance and comparing different tables

In CA the variance in a table is quantified by the total inertia, measured in the following equivalent ways, starting with the definition (1) of Section 2.

a. Sum of squared standardized residuals of relative frequencies $$\sum_i\sum_j\left(\frac{p_{ij}-r_ic_j}{\sqrt{r_ic_j}}\right)^2$$

b. Weighted sum of squared differences between contingency ratios and 1 (weights r_ic_j) $$\sum_i\sum_j r_ic_j\left(\frac{p_{ij}}{r_ic_j}-1\right)^2$$

c. Inertia of the row profiles: weighted sum of squared χ^2 distances of the I row profiles to row centroid (weights r_i) $$\sum_i r_i\sum_j\frac{\left(\frac{p_{ij}}{r_i}-c_j\right)^2}{c_j}$$

d. Weighted sum of squared χ^2 distances between all ½$I(I–1)$ pairs of row profiles (weights $r_ir_{i'}$) $$\sum_i\sum_{i'<i} r_ir_{i'}\sum_j\frac{\left(\frac{p_{ij}}{r_i}-\frac{p_{i'j}}{r_{i'}}\right)^2}{c_j}$$

e. Inertia of the column profiles: weighted sum of squared χ^2 distances of the J column profiles to column centroid (weights c_j) $$\sum_j c_j\sum_i\frac{\left(\frac{p_{ij}}{c_j}-r_i\right)^2}{r_i}$$

f. Weighted sum of squared χ^2 distances between all ½$J(J–1)$ pairs of column profiles (weights $c_jc_{j'}$) $$\sum_j\sum_{j'<j} c_jc_{j'}\sum_i\frac{\left(\frac{p_{ij}}{c_j}-\frac{p_{ij'}}{c_{j'}}\right)^2}{r_i}$$

Definitions a and b are symmetric with respect to rows and columns, while the remaining definitions are orientated either to row profiles (c and d) or column profiles (e and f).

This section intends to answer the following question: if we have analysed two different tables, how can we compare their variances? If the two tables have the same number of rows and columns, the answer is simply to compare the values of their respective total inertias. The answer is not obvious, however, when the tables are of different sizes. This problem is of much less importance in PCA where variables are usually standardized to have variance 1, so that the total variance is equal to the number of variables. But it is of crucial importance in CA when tables from different sources are compared or analysed jointly (see, for example, Pagés and Bécue-Bertraut, 2006), in which case some type of table standardization is necessary that involves an equitable measure of table variance.

As a first case let us consider several tables where the number of columns is fixed, and then calculate total inertias as the number of rows is increased. For example, we considered a separate data set of 6371 cases in a health survey, using the variables age (in years) and perceived health status (five categories) (see Greenacre 2002, for an extensive analysis of these data). In order to cross-tabulate age with health status, age was categorized separately into 5, 7, 9 and 11 groups, respectively, with approximately equal numbers of cases in each group for each categorization. The total inertias of the four analyses, as well as the part of inertia and percentage on the first dimension of the CA in each case, were as follows:

Table:	5 × 5	7 × 5	9 × 5	11 × 5
Total inertia:	0.1389	0.1441	0.1462	0.1479
1st dimension:	0.1363	0.1392	0.1407	0.1416
(Percent):	(98.1%)	(96.6%)	(96.2%)	(95.8%)

As the number of age groups increases, inertia must be added to the matrix, so this is not a surprising result. In any case, the increases are not great, suggesting that the essential information is already contained in the 5 × 5 table. In CA the dimensionality of the solution is equal to $\min\{I–1, J–1\}$, which is equal to 4 for all four cases listed above. Hence each table lies exactly in four dimensions, and the effect of increasing the number of age groups is just to add a little more inertia each time. The solution is strongly one-dimensional, and we can see from the decreasing percentages of inertia on the first dimension that the increase in inertia is mainly in the form of noise in the minor dimensions. In summary, it seems reasonable to compare the total inertias of tables which have the same dimensionality, but we can also compare their inertias in principal subspaces of a common dimensionality, for example the first dimension in the above case. The essential point is that the dimensionalities in the comparison be the same.

Next, to illustrate the case of matrices of different dimensionalities, we generated a bivariate continuous distribution from which we constructed three tables, of order 3 × 3, 5 × 5 and 7 × 7 respectively. That is, we categorized the continuous variables first into three categories each and then using narrower intervals of five and seven categories. The total inertias in the respective cases were 0.1754, 0.2448 and 0.3333, illustrating the large increase in inertia with increasing dimensionality of the table, even though the underlying distribution is the same. The tables have respective dimensionalities of 2, 4 and 6. A common way of measuring association between two categorical variables is Cramer's V coefficient. In terms of CA quantities, this coefficient is equal to $\sqrt{\text{inertia}/\min\{I-1, J-1\}}$, the quantity in the square root being the average inertia per dimension of the solution:

3 × 3 table : average inertia = 0.1754/2 = 0.0877, hence $V = \sqrt{0.0877}$ = 0.296

5 × 5 table : average inertia = 0.2448/4 = 0.0612, hence $V = \sqrt{0.0612} = 0.247$

7 × 7 table : average inertia = 0.3333/6 = 0.0555, hence $V = \sqrt{0.0555} = 0.236$.

This shows that Cramer's *V* is not a suitable way of comparing variability, since most of the relevant inertia is invariably concentrated in a few major dimensions, with the result that averaging inertia over the dimensions leads to the measure decreasing as variability is added in the minor dimensions of the larger tables.

If one looks more closely at these last analyses, the principal inertias (eigenvalues) and their percentages are:

3 × 3 :	0.1716	0.0038				
	97.8%	2.2%				
5 × 5 :	0.2069	0.0220	0.0158	0.0001		
	84.5%	9.0%	6.5%	0.0%		
7 × 7 :	0.2392	0.0526	0.0218	0.0188	0.0008	0.0002
	71.8%	15.8%	6.5%	5.6%	0.3%	0.0%

In line with our previous remark, a comparative measure across tables would be the amount of inertia in spaces of the same dimensionality. Since the first analysis is two-dimensional, we could compare the inertias in the first two dimensions of each solution. The accumulated inertias in the two solutions up to dimension 2 are as follows:

3 × 3 :	0.1716	0.1754
5 × 5 :	0.2069	0.2289
7 × 7 :	0.2392	0.2918

The last column gives a fairer idea of how much inertia is introduced into the table by the coding. If we were just comparing the 5 × 5 and 7 × 7 tables, then we would compare the accumulated inertias in the four-dimensional principal subspaces, which have values 0.2448 and 0.3323 respectively. Ratios of inertias can be calculated to quantify the increase, so that the increase from the 3 × 3 to the 5 × 5 table is 0.2289/0.1754 = 1.305, an increase of 30.5%, and from the 5 × 5 to the 7 × 7 table it is 0.3323/0.2448 = 1.357, an increase of 35.7%. From the 3 × 3 to the 7 × 7 table there is a multiplicative increase of 0.2918/0.1754 = 1.664, that is 66.4%.

Thus our proposal, which is more linked to the dimension-reducing objective of CA, is to compare the sum (or average) of the inertias calculated in principal subspaces of the same dimensionality, where the dimensionality is that of the smallest table. This proposal obviously holds for the first case studied above, where the dimensionalities of the tables being compared were equal.

Let us apply this strategy to our two examples, “author” and “benthos”. The author data has 11 dimensions and the benthos data 9, so we compare accumulated inertias in the first 9 dimensions of each: for the author data, it is 0.01836, for the benthos data it is 0.3798. 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.

To conclude this section, there is another way of thinking about this question, namely in terms of “signal” and “noise”. Even in the cases of matrices of the same size, one matrix might have a one-dimensional signal which dominates the variance with the rest being noise, while another matrix might have a two-dimensional signal, say, plus noise. If we knew that, then comparing the “true” underlying variances would involve a comparison of the first inertia of the first matrix and the first two inertias of the second matrix. In practice, however, there is no hard-and-fast decision about how many “significant” dimensions a matrix has, so this approach might be too subjective. Our proposal, which compares solutions of the same dimensionality, would probably include some noise along with signal variance in the inertia calculation of each table, but at least our approach does not bias the comparison just because the matrices are of different sizes.

5. The myth of the influential outliers

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. In this section I will show that an inspection of the numerical contributions, which are an integral part of the CA results, reveals the exact contribution, and thus an indication of the influence, of each category to the CA solution. In almost all cases, rare categories play a minor role in the determination of the solution and thus can be omitted without noticeably changing the solution. What gives rise to the above 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. This problem is bound up with the scaling issue in CA and in the next section I shall deal with the scaling of maps and propose alternative ways to obtain maps which visually “tone down” the rare categories in CA displays.

Both our examples contain some very low frequency columns. For example, in the author data the rarest letters are: *q* (0.07%), *j* (0.08%), *z* (0.08%) and *x* (0.1%), with all other letters occurring

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 first two 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 to the axes are the following: 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 such a rare letter, with a frequency of occurrence of 2.2%, and plays an important role in the analysis by separating out the Faulkner texts on the second axis). Thus, if we removed *q, z* and *x* from the analysis, 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 Figure 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 their squared chi-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%. It is clear that 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) are so rare that they jointly 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 the map 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.

6. The scaling problem in CA

The scaling problem in CA has much in common with the scaling problem in the biplot, which we summarize briefly here. In a biplot a matrix **M** (I×J) is approximated by the product of two matrices $\mathbf{AB}^\mathsf{T}$, which we can write as: $\mathbf{M} \approx \mathbf{AB}^\mathsf{T}$. Usually the approximation is by least-squares and the solution is conveniently obtained using the singular value decomposition of the **M**: $\mathbf{M} = \mathbf{UD}_\sigma\mathbf{V}^\mathsf{T}$, where **U**, **V** and $\mathbf{D}_\sigma$ are respectively the matrices of left and right singular vectors, and the diagonal matrix of singular values $\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_R$ in descending order, and R is the rank of **M**. The scaling problem is illustrated here for a rank-2 approximation of **M**, but applies in the same way to any low-rank approximation. For the rank-2 (or 2-dimensional) case, **A** (I×2) (and **B** (J×2) have two columns each and are obtained from the first two columns of **U** and **V**, respectively, and the corresponding singular values, the first two elements on the diagonal of $\mathbf{D}_\sigma$. In scalar notation, we can write the rank-2 approximation 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 biplot represents row i and column j by the coordinates $[a_{i1}, a_{i2}]$ and $[b_{j1}, b_{j2}]$ respectively, with the scaling "problem" being how to partition σ_1 and σ_2 between the left and right vectors. In general this partitioning is as follows:

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

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 the singular value – this is the row asymmetric map (see Section 2), or "row-metric-preserving" (RMP) biplot by Gabriel (1971); or (ii) $\gamma = 0$, i.e. scale the column coordinates by the singular value – this is the column asymmetric map, 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): $[a_{i1}, a_{i2}] = [u_{i1}, u_{i2}]$ $[b_{j1}, b_{j2}] = [\sigma_1 v_{i1}, \sigma_2 v_{i2}]$

In the row asymmetric biplot the distances between row points approximate the Euclidean distances between the original rows of **M** (hence "RMP"), whereas in the column asymmetric it is the inter-column distances that are approximated in the map (hence "CMP"). When the rows are sampling units and the columns are variables, these two biplots have also been called the *form biplot* and the *covariance biplot* respectively (Aitchison and Greenacre 2002). An alternative scaling, which is 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 our terminology here, the symmetric map scales both rows and columns by the singular values, and is thus 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 can be represented with respect to axes using the same scale, the sum-of-squares of the row coordinates being equal to the sum-of-squares of the column coordinates on each axis k, which is in turn 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). In asymmetric maps, the coordinates which have been scaled by the singular values (i.e., principal coordinates), are usually drawn as points , whereas the unscaled coordinates (i.e., 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) .

The above scheme can be carried over to the CA case, with several nuances because of the masses assigned to rows and columns. The generalized form of the SVD in the case of CA, described in the formulation (2)–(7) of Section 2, leads to the following form for (8), called the reconstitution formula since it effectively estimates the data values from the map:

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

On the left hand side of (9) are the differences between the contingency ratios $p_{ij}/(r_i c_j)$ and 1; on the right hand side 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 group the singular values with the standard coordinates in **X**, we obtain the row principal coordinates **F** defined in (4), and hence the asymmetric row map of the CA, which is RMP in that the chi-square distances between row profiles are approximated. On the other hand, if we combine the singular values with the standard coordinates in **Y**, we obtain the column principal coordinates and thus the column asymmetric map, which is CMP in that the chi-square distances between column profiles are approximated.

If we scale both row and column standard coordinates by the singular values then we obtain the symmetric map, shown in Figures 1 and 2, where both row and column configurations approximate the chi-square distances but the specific scalar product property of the biplot is sacrificed, as pointed out above. Confusingly, this is now called the "symmetric normalization" in SPSS Categories, which does not give the option of a joint display of what I call the symmetric map, also known as the "French scaling" or "Benzécri scaling" since it is the scaling of choice of the Benzécri school of French data analysts. Notice that Gabriel (2002) has shown that the scalar product property, although not specifically satisfied, is usually not severely degraded in the symmetric map.

There are two aspects peculiar to CA which distinguish it from the the general biplot scheme described previously. The first aspect is that in CA the standard coordinates represent actual points which are theoretically possible to observe, namely the unit profile vectors (called *vertices* of the simplex space of the profiles): [1 0 0 ... 0], [0 1 0 ... 0], etc. For example, in the row asymmetric map the projections of the row profiles are represented by the row principal coordinates, and the standard coordinates are projections of these extreme profiles *in the same space as the row profiles.* This explains why in CA the vertex points in standard coordinates describe a cloud of points which is much more dispersed than the cloud of profile points in principal coordinates. This geometric result can also be deduced from the fact that the singular values α_k in CA are always less than 1, being canonical correlations (see, for example, Greenacre, 1984). As before, the directions indicated by the vertex points can be considered as axes onto which profile points can be projected to estimate the elements of the matrix on the left hand side of (9).

Let us illustrate this first property using our data sets, first the benthos data where the principal inertias α_k^2 are relatively high. Figure 3 shows the column asymmetric map, with columns in principal coordinates and rows in standard coordinates. In this figure the column configuration is identical to that of Figure 2 (inter-column chi-square distances are approximated) whereas the row configuration is considerably stretched out, and stretched out more in the vertical than the horizontal direction (the row standard coordinates in Figure 3 are the row principal coordinates in Figure 2 *divided* by the singular values on axes 1 and 2, equal to $\sqrt{0.245}$ and $\sqrt{0.073}$ respectively). The joint plot in Figure 3 still looks acceptable since the column profile points in principal coordinates are still reasonably spread out compared to the row vertex points in standard coordinates. The situation is completely different for the author data, however, since the principal inertias are very small. In Figure 4, the row asymmetric map, the points representing the 12 texts are a small smudge at the centre of the map, a striking geometric demonstration of the very low inertia of these data.

It is clear that some scale change would be necessary in Figure 4 in order to map the two sets of points together, for example change the scale of the column points by multiplying the standard coordinates by a value considerably less than 1 to bring them down to the scale of the row principal coordinates. But this scale change would destroy the property that the standard coordinates represent actual profiles (vertex points), so the question is whether we can rescale the standard coordinates to give them meaningful lengths, while still being comparable in scale to the row profiles in principal coordinates. To answer this question, we should first consider what the lengths of the column vectors signify in Figure 4 (this argument applies similarly to the lengths of the row point vectors in standard coordinates in Figure 3). This brings us to the second particular aspect of CA, namely the presence of the masses r_i and c_j in the matrix being represented in the asymmetric maps, as given by (9).

We can write (9) from the "row profile point of view" as follows:

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

that is, the asymmetric map biplots the differences between the row profile elements and their averages, expressed relative to the averages. In order to recover an element on the left, we project the row profile point onto the biplot axis defined by the column vector and multiply this projection by the length of the vector. So if we compare two column points in Figure 4 such as *z* and *f* which define biplot axes practically in the same direction, the projections of the book profiles onto this common direction will be the same. The estimated values for *z*, however, will be more than five times those for *f*, because of the much longer distance of *z* to the origin. This makes little sense, because *z* is the less frequent letter: the reason for this apparent anomaly is the division on the left of (10) by c_j, which for *z* (j=26) gives high values because the letter *z* is so rare. Instead of recovering profiles values relative to the average c_j, we might propose to 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) leads to a biplot using the principal row coordinates $[\alpha_1 x_{i1}, \alpha_2 x_{i2}]$ and the column standard coordinates rescaled by the respective column masses $[c_j y_{j1}, c_j y_{j2}]$. This biplot is shown in Figure 5. Now the column points have been pulled in by different amounts, depending on the values of their relative frequencies (masses) c_j, so that the letter *z* is practically at the origin, while a common letter such as *e* is now more prominent. This biplot scaling for CA is, in fact, the one proposed by Gabriel & Odoroff (1990).

But Figure 5 is not satisfactory either, since it 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:

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

which means that the standard column coordinates are rescaled by the *square roots* of the column masses, using expected relative frequency c_j as a surrogate for the variance of column *j*. This map is shown in Figure 6 and it is clear that the common scale for rows and columns is adequate for the joint visualization. Moreover, a long vector such as that of letter "*y*" in Figure 6, tells us that there is more variance in the percentages of "*y*" than you would expect using the above standardization, compared to the letter "*e*", for example (one could also introduce arguments of over- and under-dispersion in this situation). The same conclusion is arrived at by recalling that 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 together than on the other vectors.

In Figure 7 we show the benthos data similarly scaled, this time with the columns in principal and the rows in rescaled standard coordinates, the column version of (12):

$$\left(\frac{p_{ij}}{c_j} - r_i\right) \Big/ r_i^{1/2} \approx (r_i^{1/2} x_{i1})(\alpha_1 y_{j1}) + (r_i^{1/2} x_{i2})(\alpha_2 y_{j2}) \tag{13}$$

It is clear from Figures 6 and 7 that this scaling of the coordinates 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 Figure 6, nor between the row points (species) in Figure 7 – it is the direction and lengths of these point vectors that have meaning. In fact, the squared lengths $c_j y_{jk}^2$ and $r_i x_{ik}^2$ of the point vectors in these respective standard CA biplots are proportional to the contributions to inertia on the k-th principal axis. Hence longer vectors in the direction of a principal axis indicate higher contributions to that axis' inertia.

Notice also how the differences in total inertia can be observed in the standard biplots: in Figure 6 there is a relatively small dispersion of the books in Figure 6 compared to the fan of letter vectors (low total inertia), while in Figure 7 there is a relatively large dispersion of the sites in Figure 7 compared to the fan of species vectors (high total inertia).

As a final remark in this section, whatever map is chosen amongst those recommended above, at least one set of points is in principal coordinates to show the form of the corresponding profile cloud – this set is usually the set regarded as the "cases" of the study, for example the books, sites, age groups...

7. 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 family of factorial analyses can be rotated, but two further questions need answering first: (i) why is rotation necessary in the context of the data? (ii) which CA coordinates need to be rotated and how, taking into account that each point in CA has an assigned mass?

First, why would a solution need rotation? The idea behind a rotation is that if subsets of variables are made to coincide more closely with the dimensions of the solution subspace, a process called "reification", then the interpretation is simplified. 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. The analogy between variables in PCA/factor analysis and the categories of a single variable in CA is tenous to say the least. Another point, perhaps even more important, is that the full CA space is not the unlimited vector space of real numbers but the simplex space of the profiles, which are vectors [$v1$ $v2$ ...] 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 the variables can point in any direction depending on their correlation structure. From this point of view it is unlikely that some categories would form patterns at right-angles to one another and thus be candidates for rotation to "simple structure". In both examples discussed in this paper there is no benefit at all in rotating the solution (see the vectors for the letters and the species in Figures 3 and 6).

Having said this, there are some occasions in my experience, albeit extremely rare, where rotation would have been useful, but these have been almost entirely in the MCA context. For example, Figure 8 shows an MCA of 10 categorical variables which include a missing data category for each variable. All the missing data categories are in a bunch in the lower right side of the map, opposing all the substantive categories lying in a diagonal band. Since we are not interested in the missing data categories, it would be useful to rotate the solution so that these non-response categories coincided along one dimension – we could then simply ignore that dimension and look at projections on other pairs of dimensions to interpret the substantive categories.

To conclude this section, supposing that rotation is justified in some rare cases, such as the above, how could a formal rotation of the solution be made? Van der Velden (2000) and van der Velden and Kiers (2005) consider 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 or columns define some type of variable on which cases or groups are evaluated (e.g., the letters in the author data, the species in the benthos data) then rotation of that variable's categories can be considered, but not the cases, since it is the variable's categories that are used to interpret the axes. The standard coordinates of the variable's 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, which are affected by the parts of inertia along the original (unrotated) principal axes. As far as joint rotation of row and column coordinates, this would only be justified when both rows and columns are considered to be variables and play symmetric roles, as in the case of multiple correspondence analysis: for example two substantive questions in a questionnaire. There is more justification for rotating coordinates in multiple correspondence analysis (see, for example, Adachi, 2004), especially the constrained form known as non-linear PCA, than in simple correspondence analysis.

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, that is the (*j, s*)-th element of $\tilde{\mathbf{Y}} = \mathbf{YQ}$, for **Q** an orthogonal rotation matrix. Notice that the mass c_j is squared because the objective function involves fourth powers of the coordinates. Since $c_j \tilde{y}_{js}^2 = (c_j^{1/2} \tilde{y}_{js})^2$, an almost identical alternative is suggested, which would be our preferred recommendation since it is a trivial variation of the usual rotation algorithm: perform a rotation (unweighted) on the rescaled standard coordinates $c_j^{1/2} y_{js}$, which are exactly those used in the standard CA biplot defined in Section 6. This gives an additional justification for this scaling and the rotation is effectively reifying the contributions on each principal axis.

8. Stability and statistical significance of maps

CA is primarily a descriptive technique and is frequently criticized for not being inferential, but there are actually several possibilitites for investigating the statistical properties of the results. If the data are in the form of 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, see Greenacre 1984: Section 8.1 for a summary). In addition, when the bilinear model (9) is estimated by maximum likelihood instead of weighted least squares, a whole range of hypotheses can be tested (Gilula and Haberman 1986).

In a context of more general types of data, there are two levels at which the variability of the results can be investigated, called “internal stability” and “external stability” by Greenacre (1984). 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 Section 5. Greenacre (1984) gives some rules about determining the potential affect on 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 variation 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, in its most basic form. In the author data, the rows and columns are fixed but the text has been sampled within each book, although the way this was done has not been disclosed. Let us suppose, for purposes of illustration, that each batch of text was simply chosen at random from the total corpus of text. Meulman (1982) and Greenacre (1984) have 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. The data for each book is regarded as a multinomial population from which as many letters are selected at random as in the original data set. 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.

Figure 9 shows the result of the partial bootstrap after replicating the data matrix through random sampling 100 times, leading to 100 replicates for each letter, projected in principal coordinates onto the map of the original table. Rather than draw all these row and column replicates, their dispersion can be summarized in the map by one of two ways: plotting either convex hulls or confidence ellipses for each subcloud of replicates.

Figure 10 shows the convex hulls of each subcloud. Since the convex hull is sensitive to outlying replicates, it is usually "peeled" once, that is the convex hull of points is removed and the convex hull of the remaining points is drawn. Figure 11 shows the convex hulls after a peeling that removes an average of 8.7 points from each subcloud. In this sense the convex hulls of Figure 11 cover an average of 91.3% of the points. To obtain a convex hull including 95% of the points, one could devise a scheme to remove exactly 5% of the points from each subcloud (in our example, five points from each subcloud). These points should be the most outlying points in each case, for example one proposal would be to peel off those points which are furthest from the centroid of the replicates.

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 for details, also Murdoch and Chow, 1996). Figure 12 shows the 95% confidence ellipses. 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 (see, for example, Greenacre (1993b)), 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.

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 which contains all the points.

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 assumption is not satisfied in either of the examples presented here, we present the results for purposes of comparison (Figure 13). Compared with Figure 12 the ellipses are rounder, indicating less correlation than in the replicates based on bootstrapping, otherwise the two approaches give quite similar results. The convex hull approach with peeling is to be preferred, however, since it gives a more accurate indication of the dispersion of the points.

Finally, permutation testing is possible in specific cases where the data have an inherent structure, for example in the "author" data where the texts are in pairs and in the "benthos" data where the sampling points are grouped as polluted (near oilfield) and unpolluted (reference). For example, we noted before that the pairs of texts by the same author tend to lie close to one another in Figure 1, but is this "significant"? One way to answer this question is to consider the complete set of ways of assigning the labels of the 12 texts to the 12 row points in Figure 1, calculating for each assignment some measure of proximity between pairs of books, for example the sum of the six interpoint distances between pairs of texts. It turns out that, in the over 10000 unique ways of assigning the six pairs of labels, the original correct assignment (as in Figure 1) is found to give the smallest sum-of-distances, hence in this sense the phenomenon is highly significant, with $P<0.00001$.

9. Summary

I have considered five "loose ends" in simple CA, all of which are issues of contention and frequently debated, but without a satisfactory resolution of the problem. In the following I summarize them along with my proposals for "tying them up".

(i) *Comparing inertias of different sized tables*: the proposal is to identify the lowest dimensionality of the set of tables, suppose this is equal to p, and then compare the sum-of-inertias on the first p dimensions in the CA of each table.

(ii) *Outliers*: in CA outliers are not necessarily influential, and in fact they are seldom very influential at all because they usually have low masses – the only way to judge their influence is to look at the tables of contributions to inertia.

(iii) *Scaling of joint plots*: in general, the symmetric map, used routinely by most French data analysts, is the best default option. When, from a substantive point of view, the table is regarded as clearly asymmetric in nature, where the rows, say, are the "subjects" and the columns the "variables" of the table, then a biplot display may be preferable, and the standard CA biplot is the best choice with the following scaling: (a) represent the rows in principal coordinates, because the inter-subject distances are important to interpret, and (b) represent the columns in standard coordinates multiplied by the square-root of the corresponding column mass. This scaling gives row point projections onto biplot axes equal to standardized values of the row profiles.

(iv) *Rotations*: in general, these are not justified by the nature of the simplex geometry of CA; in the few cases where maps turn out in such a way that rotation would simplify the interpretation, perform a rotation on standard coordinates multiplied by the square-root of their respective masses (as in (iii) above), i.e. taking the masses into account by weighting all squared coordinates by the corresponding point masses.

(v) *Stability of maps*: at the internal level, look at the point contributions to inertia to see which points influence the principal axes the most; at the external level, data can be replicated using bootstrapping according to the original sampling scheme which was used to establish the data, followed by projection of the replicated row profiles and/or column profiles onto the original CA map, indicating the dispersion of each set of replicates by their concentration ellipse or, preferably, their convex hull, optionally peeled.

References

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

AITCHISON, J. and GREENACRE, M.J. (2002). Biplots of compositional data. *Applied Statistics 51*, 375–392.

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*. Chichester, UK: Wiley

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*. London: Chapman and Hall.

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*, London: Chapman and Hall, forthcoming.

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

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

GREENACRE, M.J. (1993b). *Correspondence Analysis in Practice.* London: Academic Press.

GREENACRE, M.J. (2002). Correspondence analysis of the Spanish national health survey. *Gaceta Sanitaria 16*, 160-170.

GREENACRE, M.J. (2004). Weighted metric multidimensional scaling. Working paper number 777, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona.

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

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

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

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

MURDOCH, D.J. and CHOW, E.D. (1996). A graphical display of large correlation matrices. *American Statistician 50*, 178–180.

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*. London: Chapman and Hall, in press.

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*. New York:W.H. Freeman and Co.

VAN DE VELDEN, M. (2000). *Some Topics in Correspondence Analysis*. PhD Thesis, University of Amsterdam.

VAN DE VELDEN, M. and KIERS, H.A.L. (2005). Rotation in correspondence analysis. *Journal of Classification 22*, 251–271.

Table 1 Data set “author”: letter counts in 12 samples of texts from books by six different authors (R Development Core Team 2005).

Abbrev.	***a***	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	***j***	***k***	***l***	***m***
TD-Buck	550	116	147	374	1015	131	131	493	442	2	52	302	159
EW-Buck	557	129	128	343	996	158	129	571	555	4	76	291	247
Dr-Mich	515	109	172	311	827	167	136	376	432	8	61	280	146
As-Mich	554	108	206	243	797	164	100	328	471	4	34	293	149
LW-Clark	590	112	181	265	940	137	119	419	514	6	46	335	176
PF-Clark	592	151	251	238	985	168	152	381	544	7	39	416	236
FA-Hem	589	72	129	339	866	108	159	449	472	7	59	264	158
Is-Hem	576	120	136	404	873	122	156	593	406	3	90	281	142
SF7-Faul	541	109	136	228	763	126	129	401	520	5	72	280	209
SF6-Faul	517	96	127	356	771	115	189	478	558	6	80	322	163
Pen3-Holt	557	97	145	354	909	97	121	479	431	10	94	240	154
Pen2-Holt	541	93	149	390	887	133	154	463	518	4	65	265	194

Abbrev.	***n***	***o***	***p***	***q***	***r***	***s***	***t***	***u***	***v***	***w***	***x***	***y***	***z***
TD-Buck	534	516	115	4	409	467	632	174	66	155	5	150	3
EW-Buck	479	509	92	3	413	533	632	181	68	187	10	184	4
Dr-Mich	470	561	140	4	368	387	632	195	60	156	14	137	5
As-Mich	482	532	145	8	361	402	630	196	66	149	2	80	6
LW-Clark	403	505	147	8	395	464	670	224	113	146	13	162	10
PF-Clark	526	524	107	9	418	508	655	226	89	106	15	142	20
FA-Hem	504	542	95	0	416	314	691	197	64	225	1	155	2
Is-Hem	516	488	91	3	339	349	640	194	40	250	3	104	5
SF7-Faul	471	589	84	2	324	454	672	247	71	160	11	280	1
SF6-Faul	483	617	82	8	294	358	685	225	37	216	12	171	5
Pen3-Holt	417	477	100	3	305	415	597	237	64	194	9	140	4
Pen2-Holt	484	545	70	4	299	423	644	193	66	218	2	127	2

Abbreviations: TD (Three Daughters), EW (East Wind) – Buck (Pearl S. Buck)

Dr (Drifters), As (Asia) – Mich (James Michener)

LW (Lost World), PF (Profiles of Future) – Clark (Arthur C. Clarke)

FA (Farewell to Arms), Is (Islands) – Hem (Ernest Hemingway)

SF7 and SF6 (Sound and Fury, chapters 7 and 6) – Faul (William Faulkner)

Pen3 and Pen2 (Bride of Pendorric, chapters 3 and 2) – Holt (Victoria Holt)

Table 2 Data set “benthos”: abundances of 10 marine species near an oilfield in the North Sea at 13 sites (sites 1 to 11 are polluted, R1 and R2 are unpolluted reference sites).

	Sites												
Species	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***R1***	***R2***
s1	193	79	150	72	141	302	114	136	267	271	992	5	12
s2	49	30	57	34	39	63	58	71	39	68	76	25	48
s3	19	39	11	38	18	20	11	22	30	40	3	55	65
s4	9	26	5	30	35	2	11	13	5	63	1	0	1
s5	17	7	15	8	10	13	21	10	8	18	5	8	3
s6	2	12	4	12	6	7	3	10	8	12	4	2	6
s7	4	2	0	3	4	11	8	1	3	3	29	2	3
s8	7	1	6	1	3	4	2	1	8	6	6	4	6
s9	4	5	2	11	1	2	3	3	2	2	2	3	1
s10	1	5	7	1	5	4	0	1	0	4	0	0	0

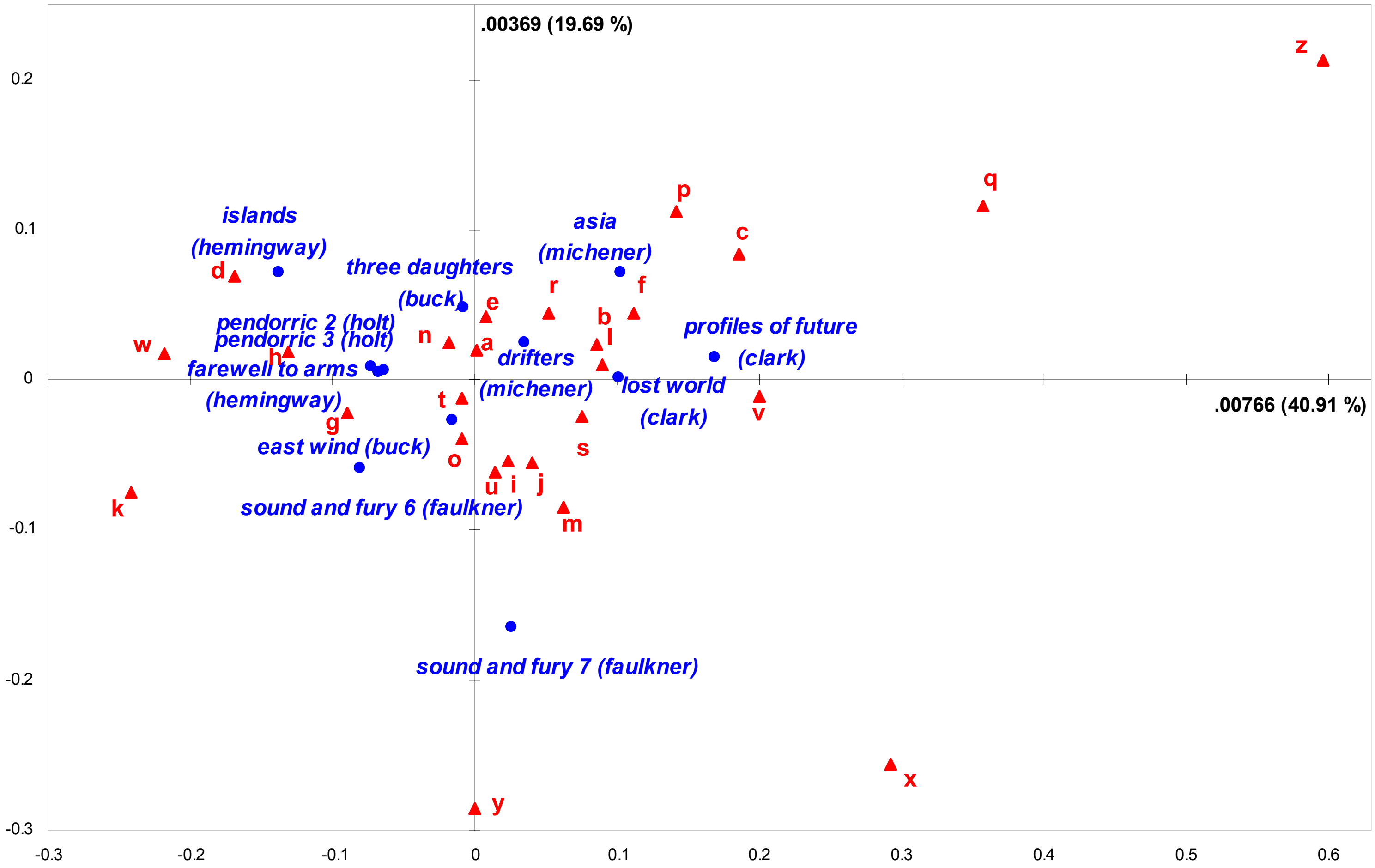

Figure 1: Symmetric CA map of "author" data: first two principal axes; total inertia=0.01874

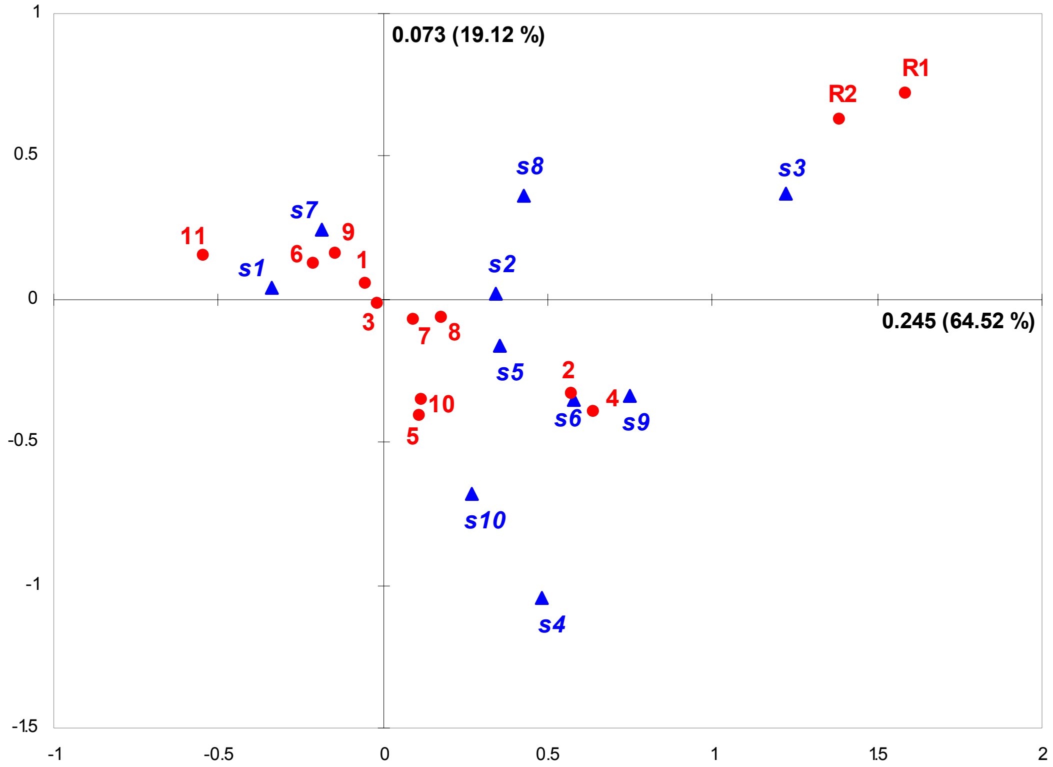

Figure 2: Symmetric CA map of “benthos” data: first two principal axes; total inertia=0.380

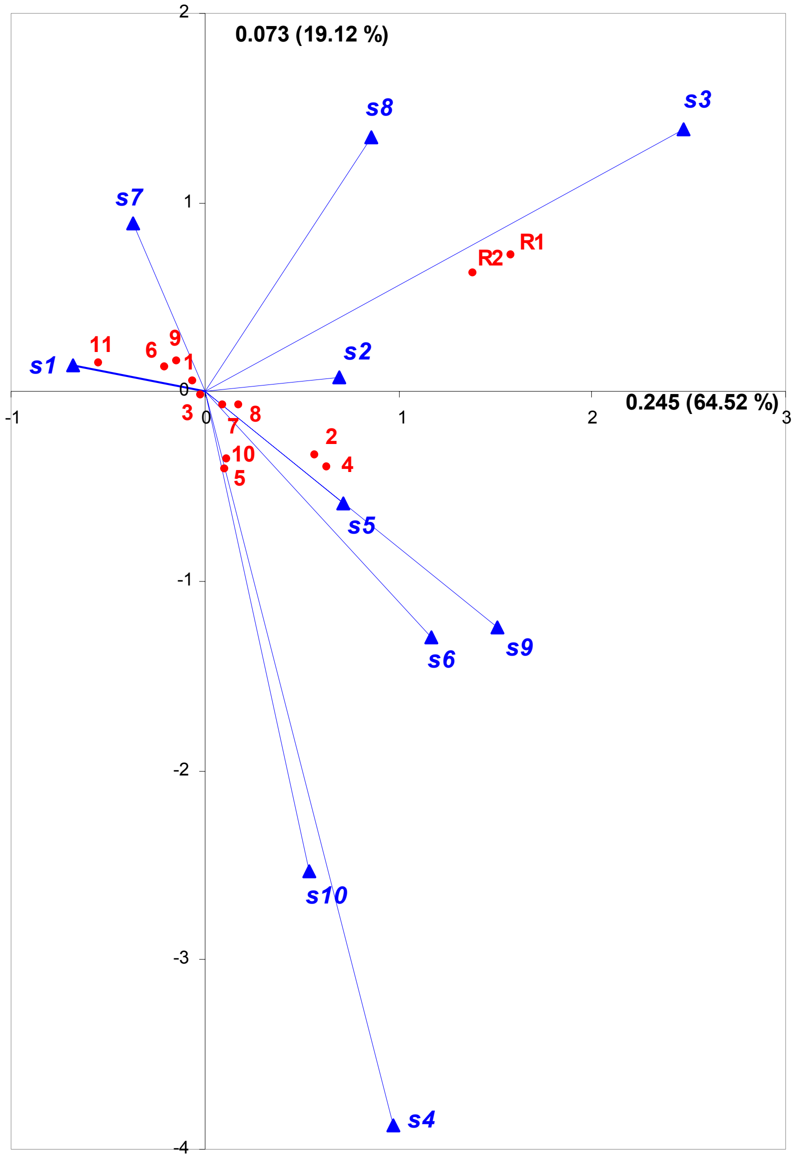

Figure 3: Column asymmetric CA map of "benthos" data: first two principal axes, showing the column (profile) points in principal coordinates and row (vertex) points in standard coordinates; the vertex points are joined to the origin.

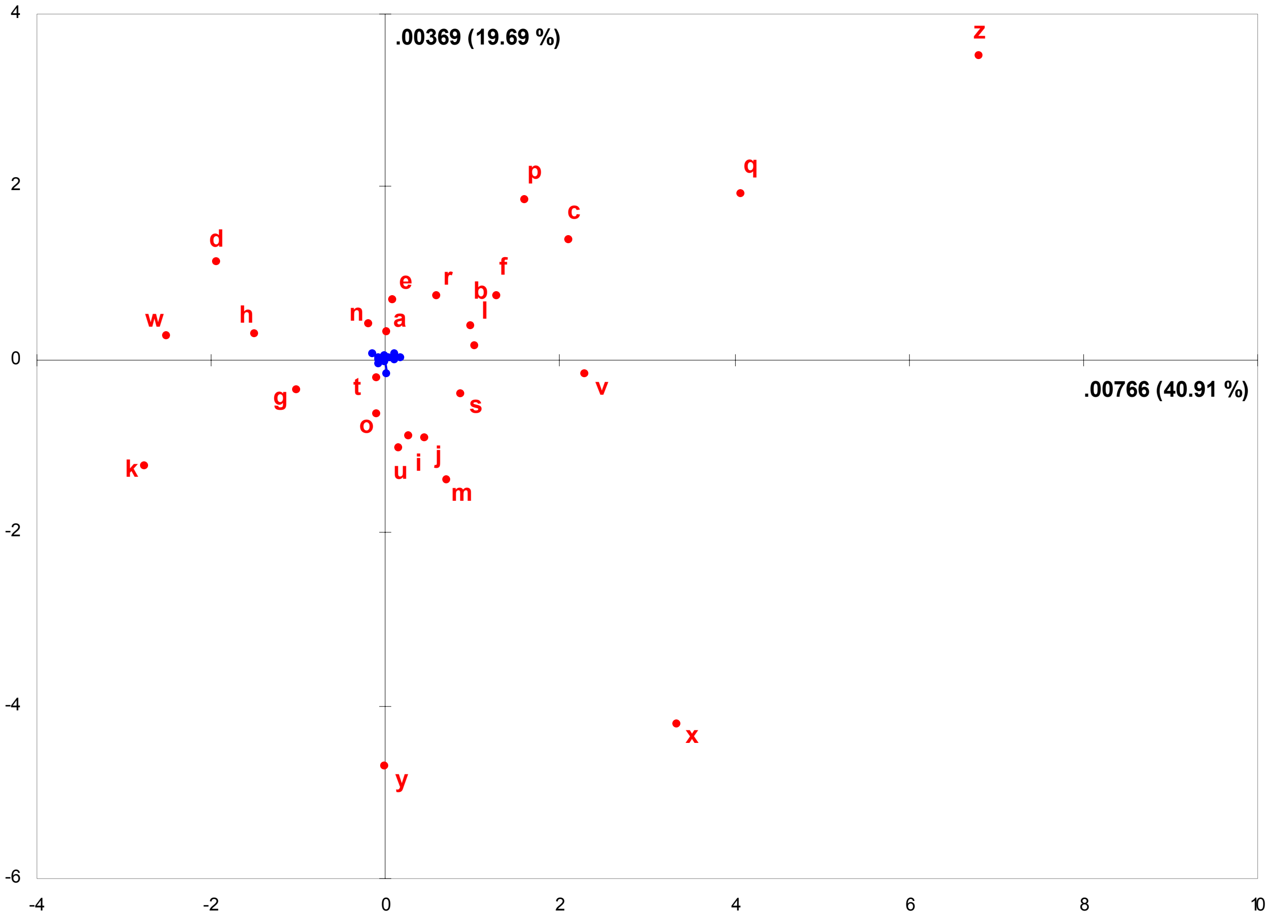

Figure 4: Row asymmetric CA map of "author" data: first two principal axes, showing the letter (column) points in standard coordinates and book (row) points in principal coordinates; the book points have such a small total inertia that they are all practically at the origin of the map.

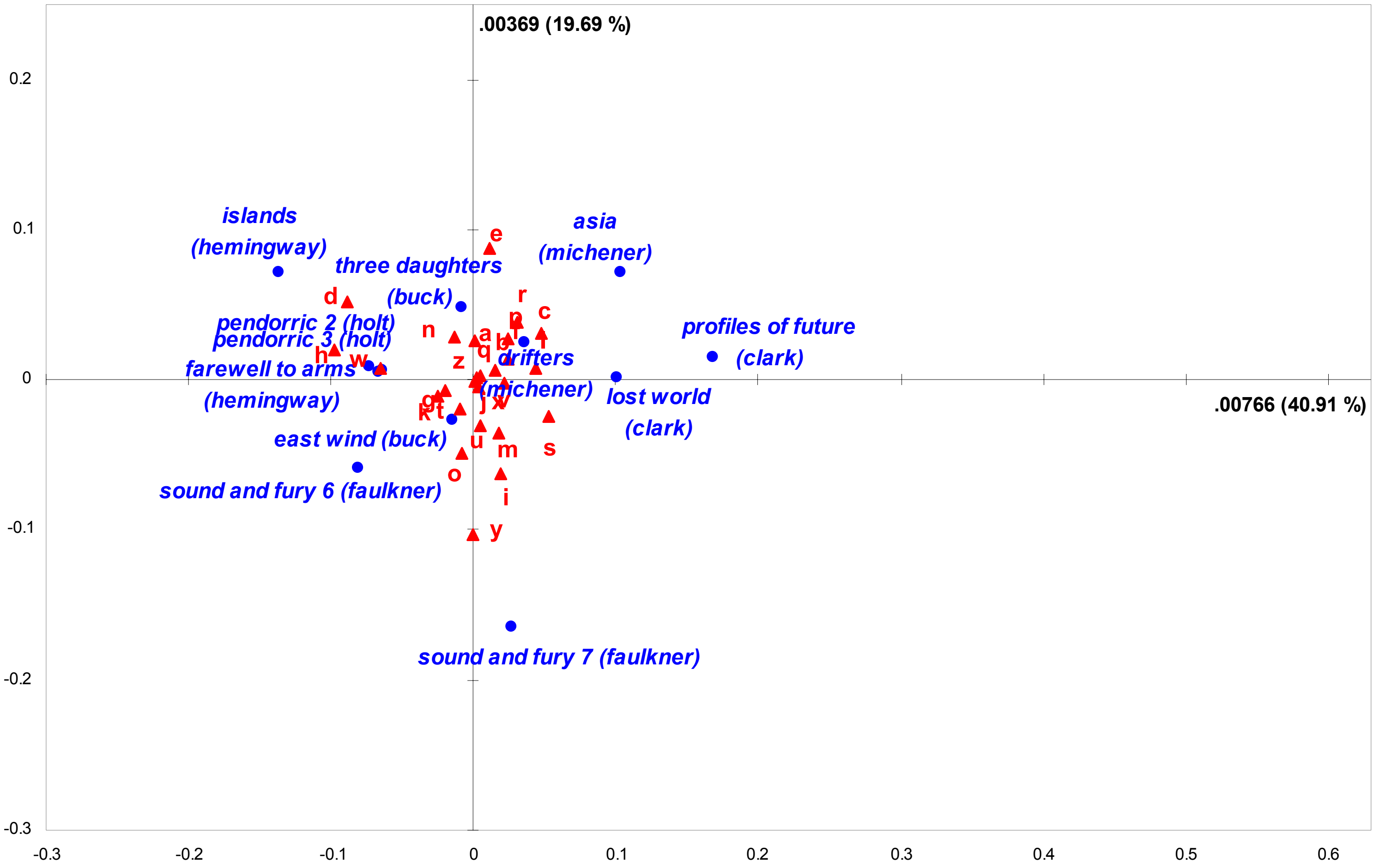

Figure 5: Gabriel biplot of CA of “author” data: first two principal axes, showing the letter (column) points in standard coordinates scaled by column masses and book (row) points in principal coordinates.

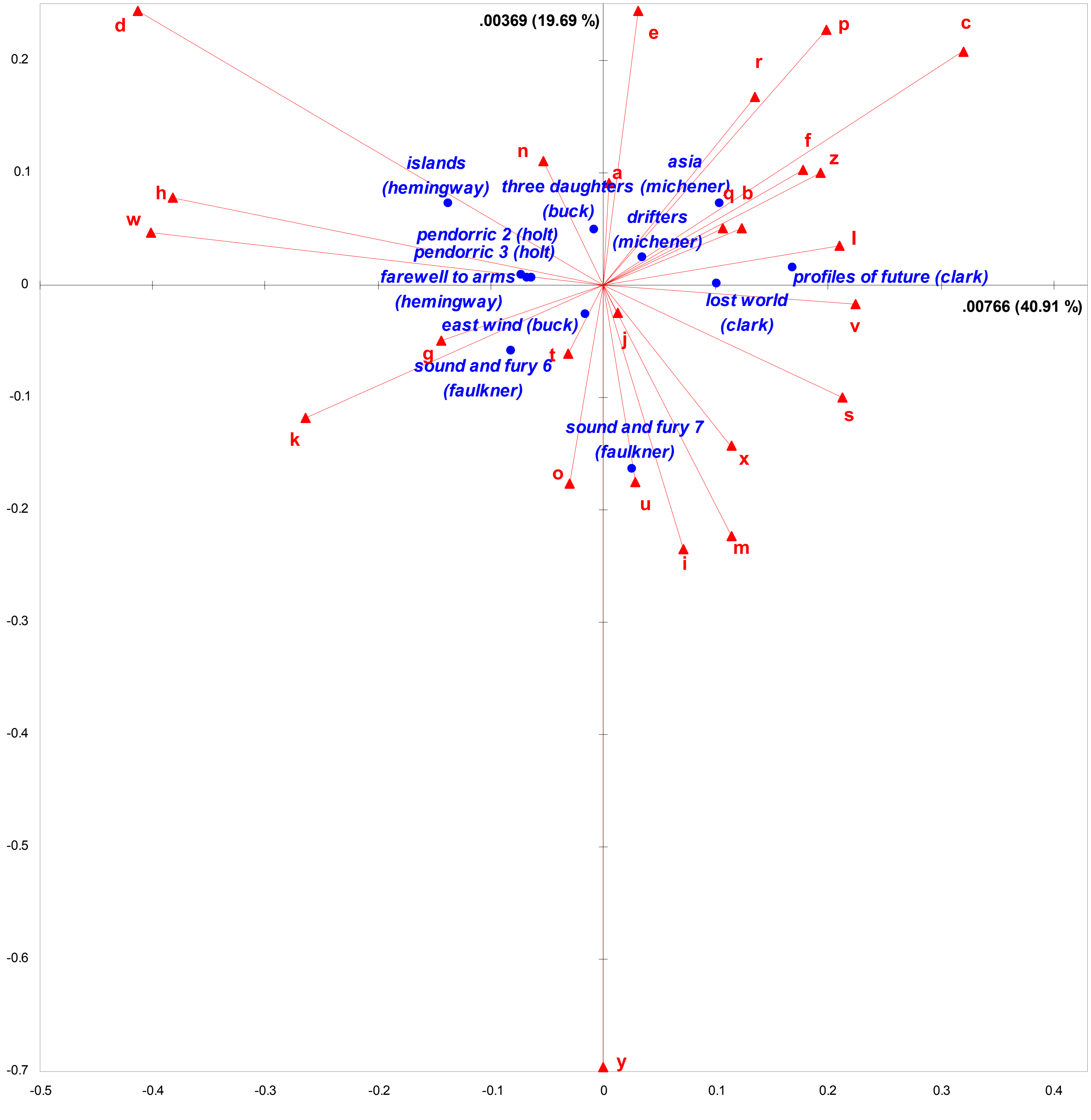

Figure 6: Standard biplot of the CA of the "author" data: first two principal axes, showing the letter (column) points in standard coordinates rescaled by square roots of masses, and book (row) points in principal coordinates.

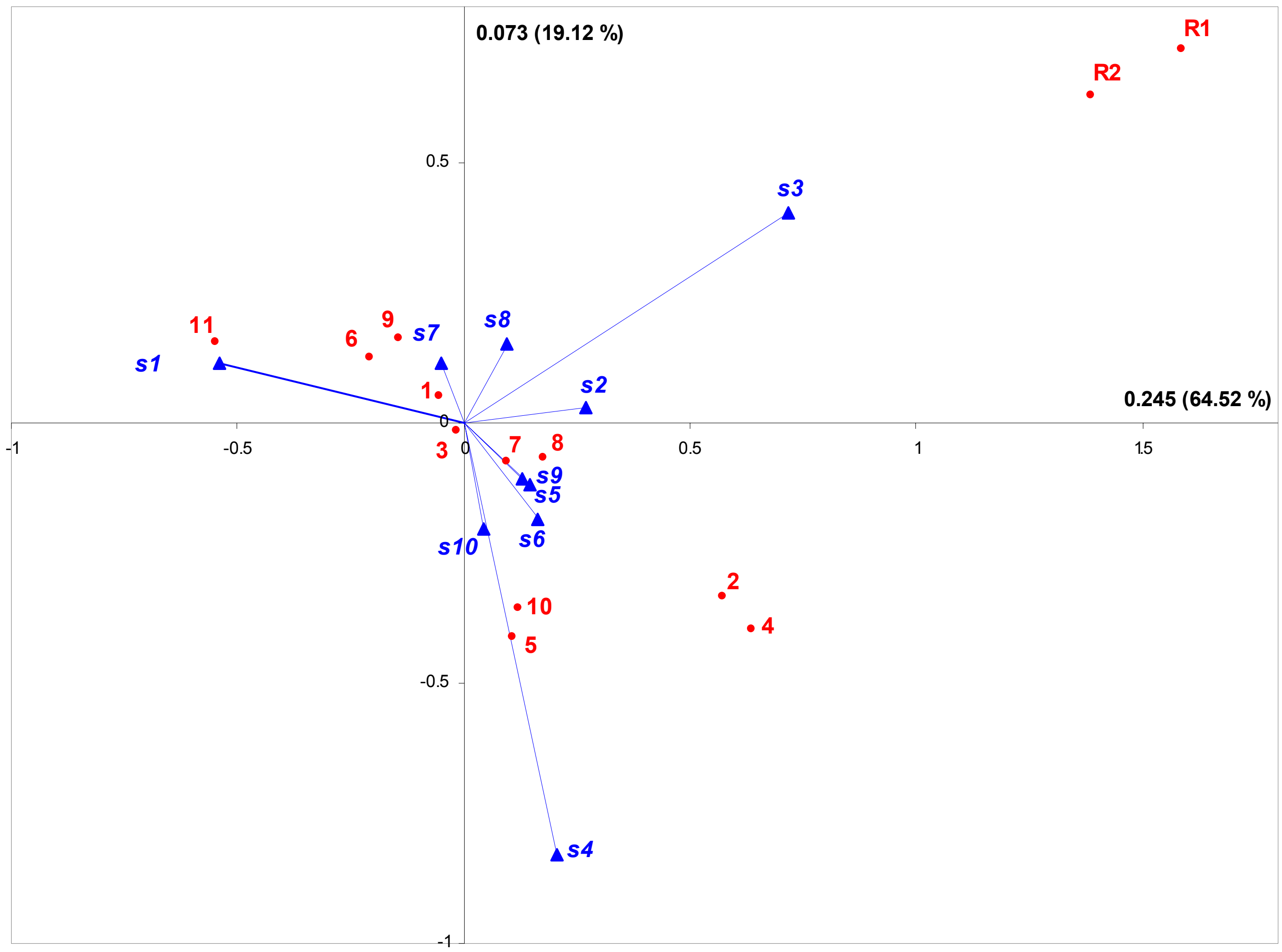

Figure 7: Standard biplot of the column profiles of the “benthos” data: first two principal axes, showing the sites (columns) in principal coordinates and the species (rows) in standard coordinates rescaled by square roots of their masses.

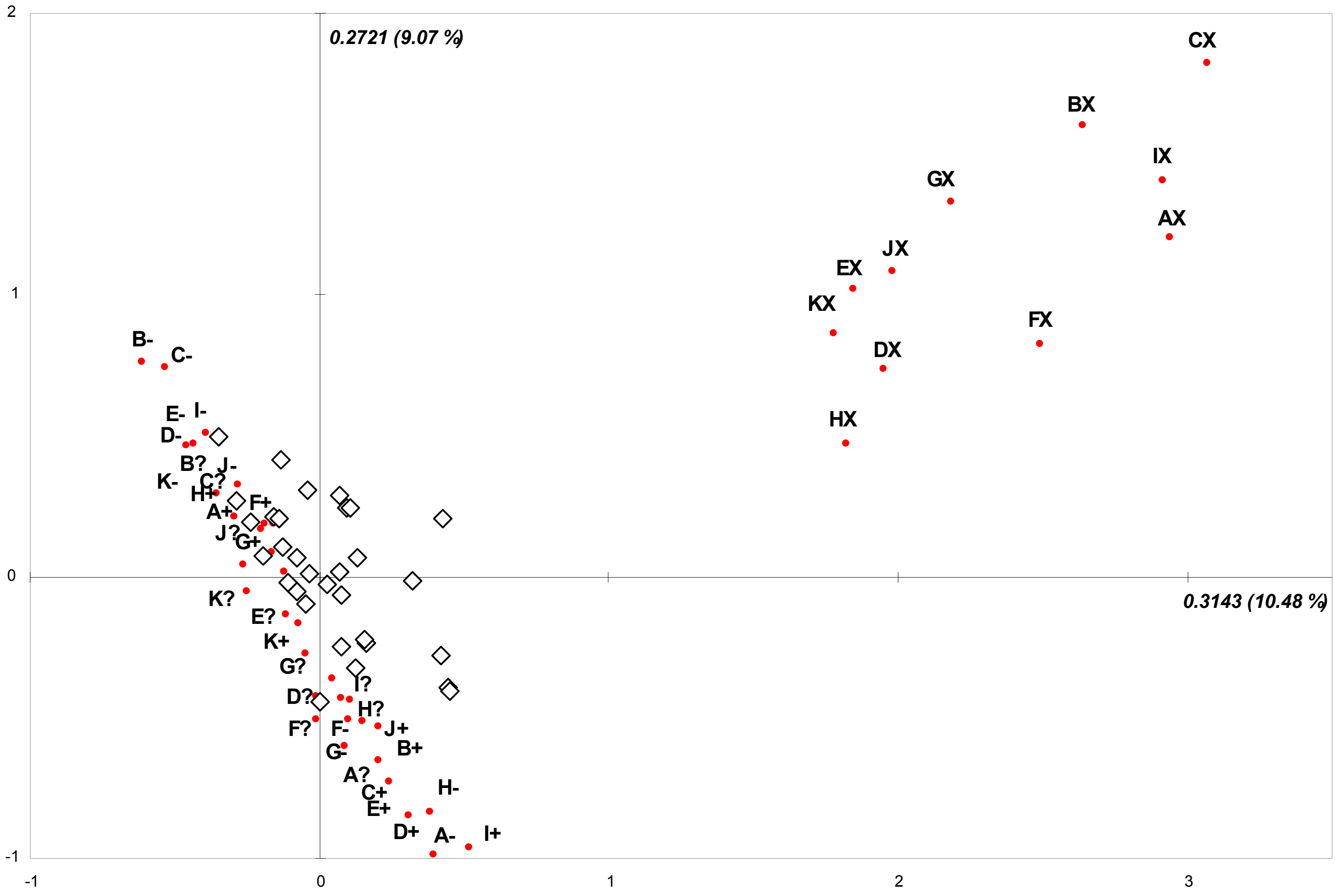

Figure 8: Multiple correspondence analysis map of response categories to 11 questions, labelled A B, C, ..., K, plus a character "**+**" (agree), "**?**" (unsure), "**-**" (disagree) or "**X**" (missing). The diamonds correspond to supplementary demographic categories (abbreviations not given here).

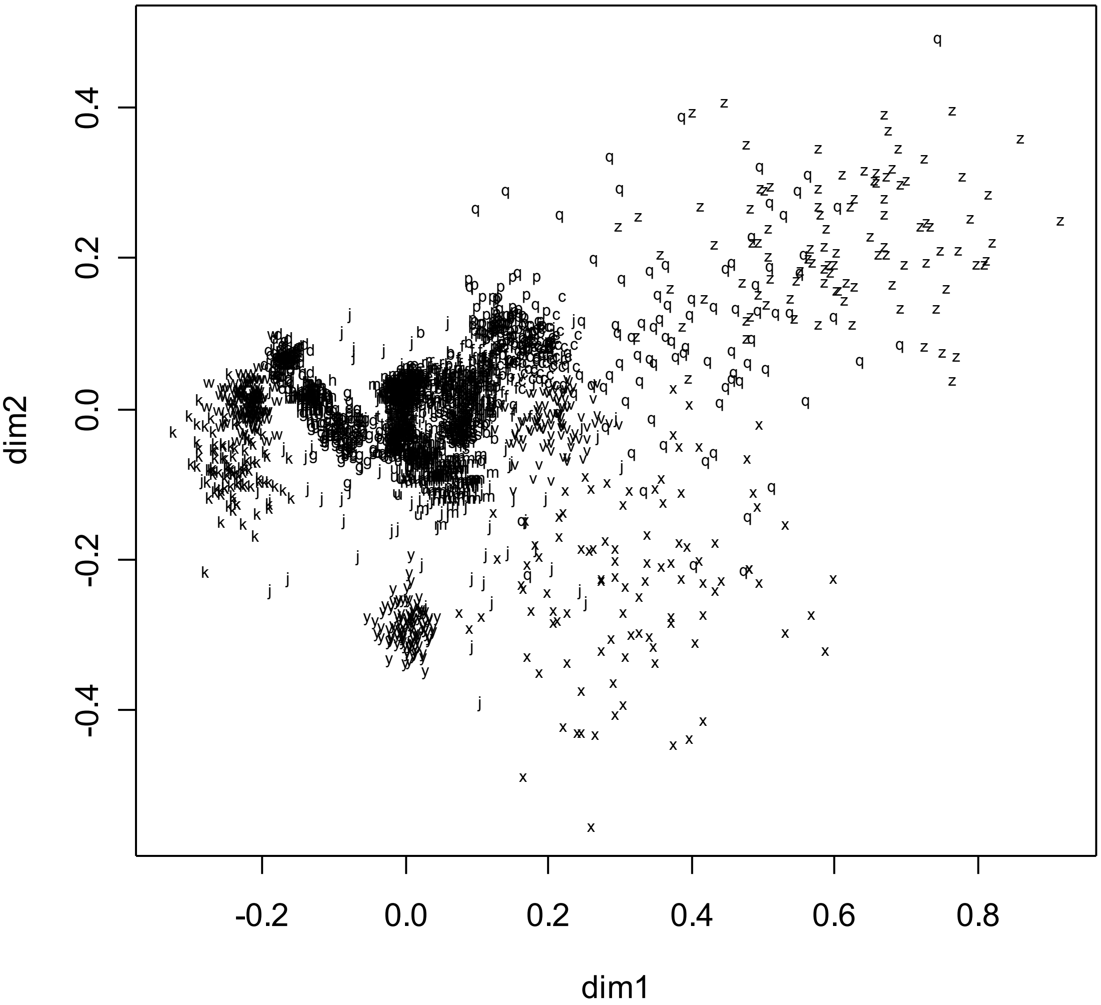

Figure 9: (Partial) bootstrapping of 26 letters, after 100 replications of the data matrix. The more frequent the letter is in the texts, the more concentrated (less variable) are the replicates.

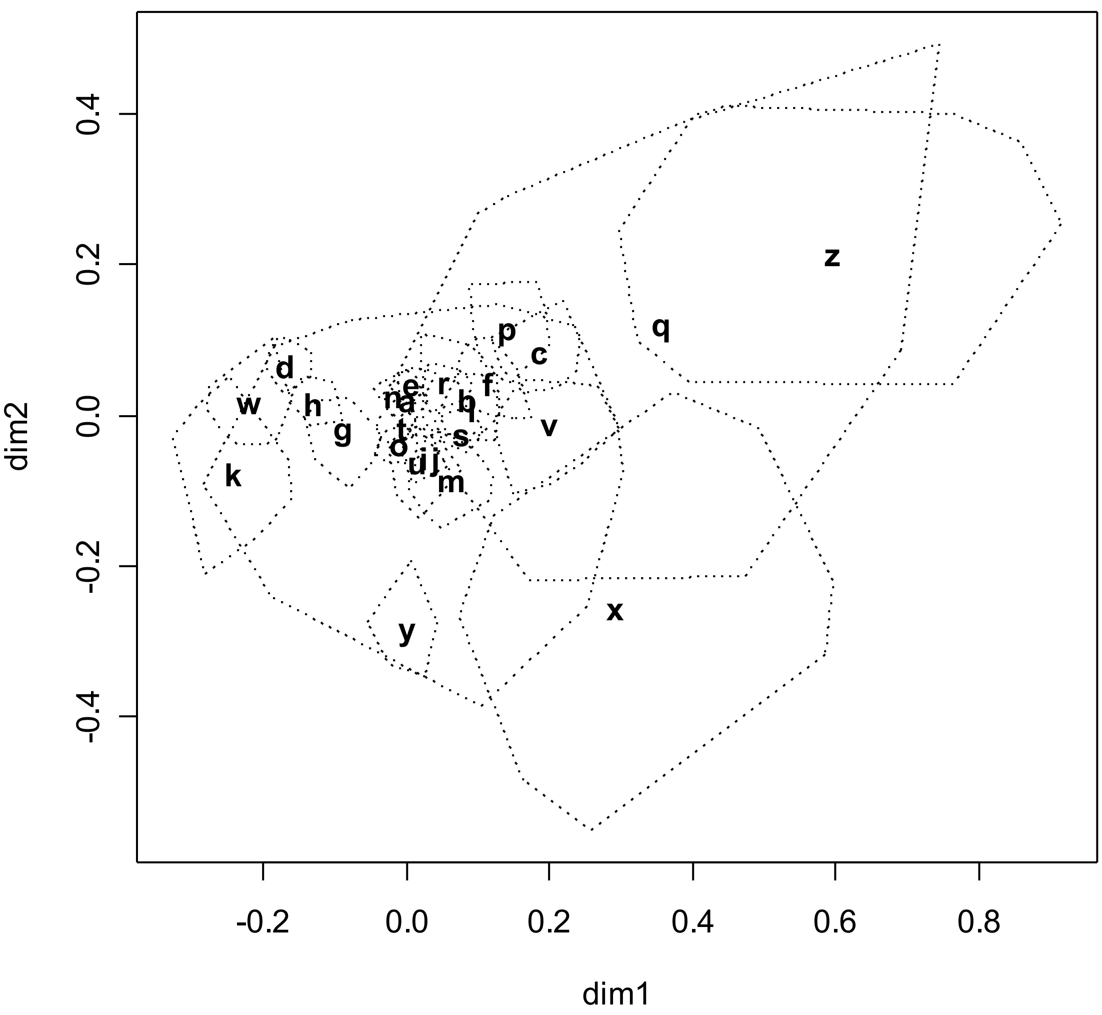

Figure 10: Convex hulls of points in Figure 10, showing letters in their principal coordinate positions in the original map.

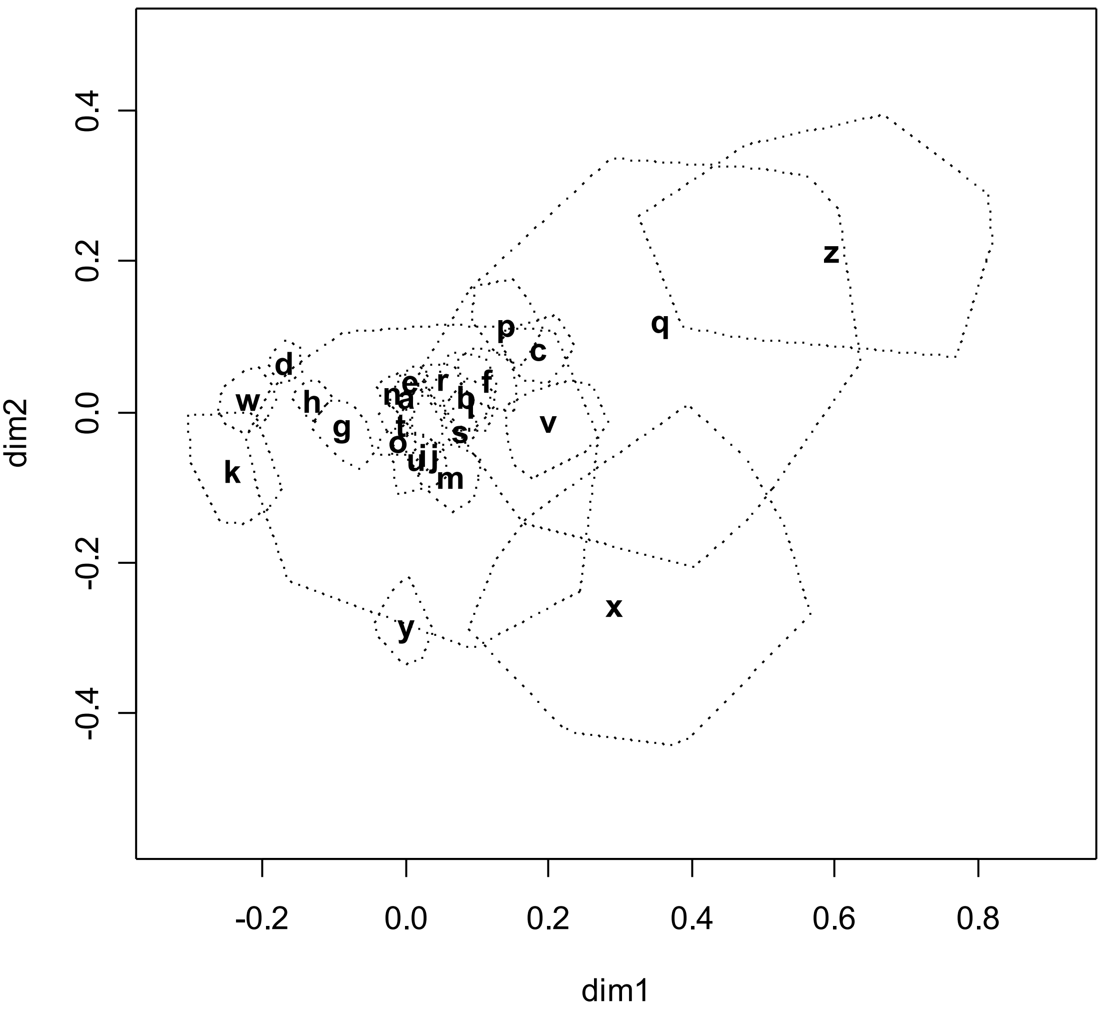

Figure 11: Peeled convex hulls of points in Figure 10, removing an average of 8.7 points (per 100 replicates) from the hulls in Figure 11.

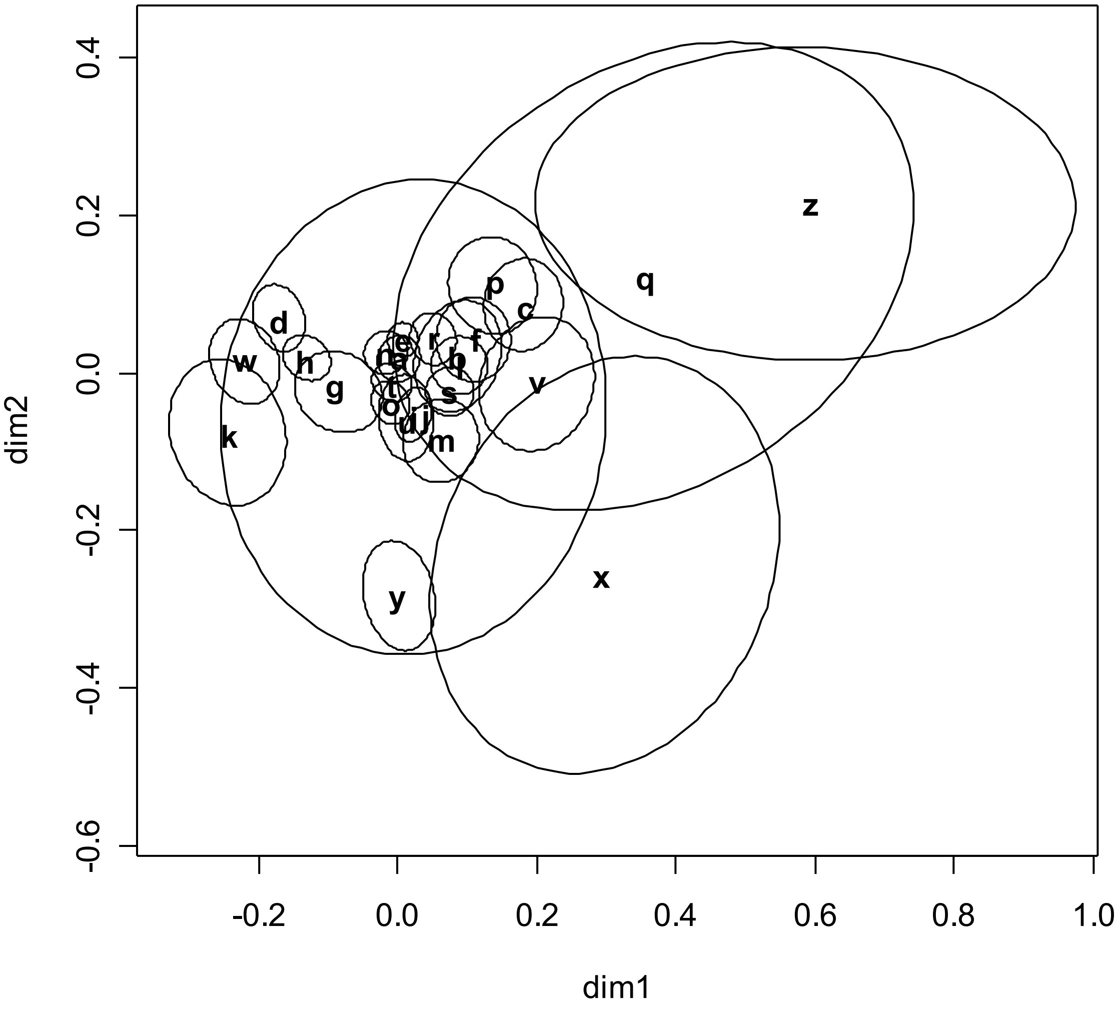

Figure 12: Concentration ellipses which enclose 95% of the points, assuming bivariate normal distribution of each subcloud.

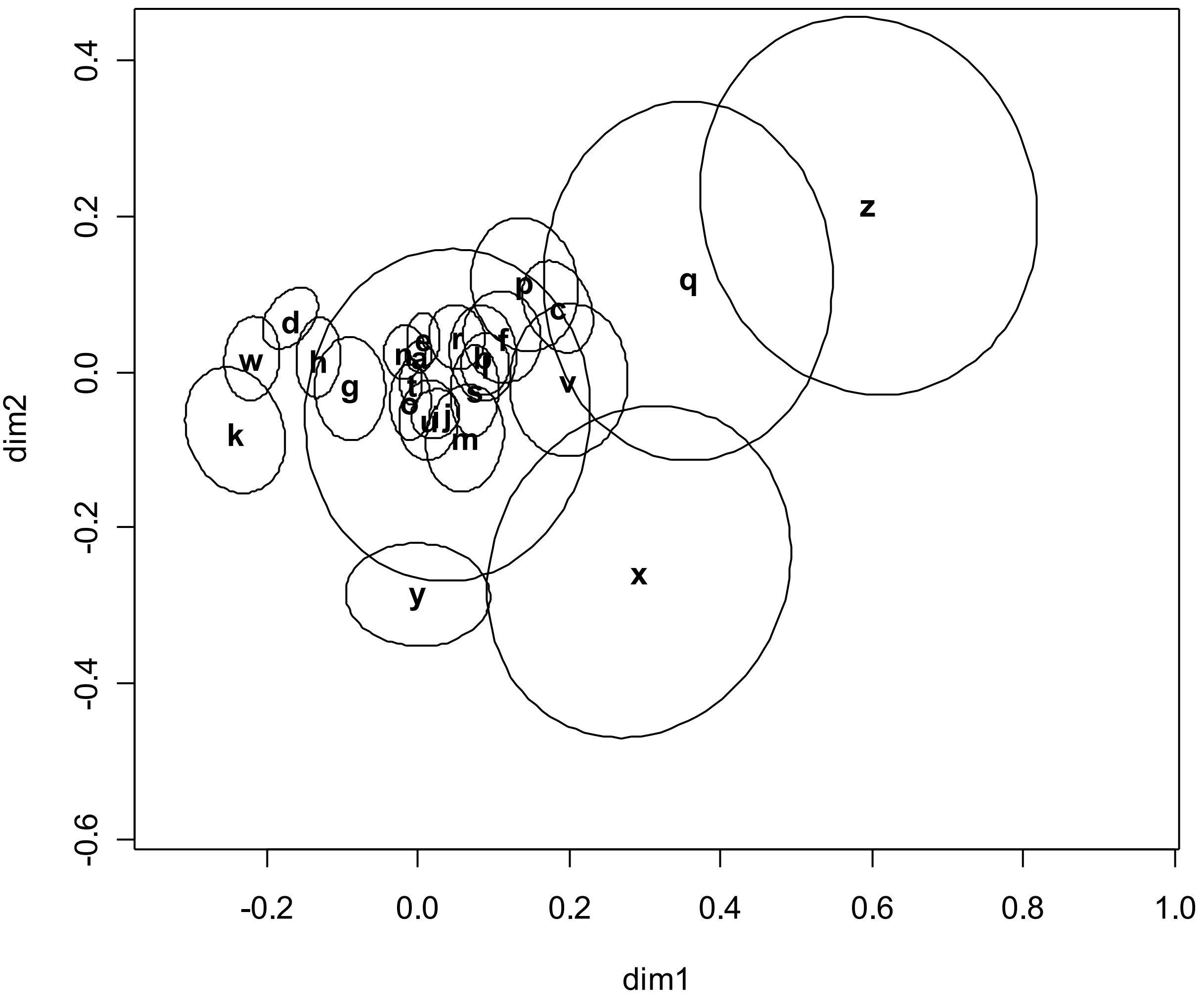

Figure 13: 95% concentration ellipses based on the delta method (Gifi 1990).