In Neil Salkind (Ed.), ***Encyclopedia of Research Design.***
Thousand Oaks, CA: Sage. 2010

Correspondence Analysis

Hervé Abdi · Lynne J. Williams

1 Introduction

Correspondence analysis (CA) is a generalized principal component analysis tailored for the analysis of qualitative data. Originally, CA was created to analyze contingency tables, but, CA is so versatile that it is used with a lot of other data table types.

The goal of correspondence analysis is to transform a data table into two sets of *factor scores*: One for the rows and one for the columns. The factor scores give the best representation of the similarity structure of the rows and the columns of the table. In addition, the factors scores can be plotted as maps, which display the essential information of the original table. In these maps, rows and columns are displayed as points whose coordinates are the factor scores and where the dimensions are called *factors*. Interestingly, the factor scores of the rows and the columns have the same variance and, therefore, both rows and columns can be conveniently represented in one single map.

The modern version of correspondence analysis and its geometric interpretation comes from 1960s France and is associated with the French school of "data analysis" (*analyse des données*) and flourished under the tutelage of Jean-Paul Benzécri (1973, see also Escofier & Pagès, 1998, or Lebart & Fénelon, 1971; a comprehensive English reference is Greenacre, 1984, 2007; see also Weller & Romney, 1990; and Clausen, 1998).

As a technique, it was often discovered (and re-discovered) and so variations of correspondence analysis can be found under several different names such as "*dual-scaling*," "*optimal*

Hervé Abdi
The University of Texas at Dallas

Lynne J. Williams
The University of Toronto Scarborough

Address correspondence to:
Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
E-mail: herve@utdallas.edu http://www.utd.edu/~herve

Table 1: The punctuation marks of six French writers (from Brunet, 1989).

	Period	Comma	All the other marks
Rousseau	7836	13112	6026
Chateaubriand	53655	102383	42413
Hugo	115615	184541	59226
Zola	161926	340479	62754
Proust	38177	105101	12670
Giraudoux	46371	58367	14299

scaling," or "*reciprocal averaging*," (see, *e.g.,* Greenacre, 1984; Nishishato, 1994). The multiple identities of correspondence analysis are a consequence of its large number of properties: It can be defined as an optimal solution for a lot of apparently different problems.

2 Notations

Matrices are denoted with upper case letters typeset in a boldface font, for example $\mathbf{X}$ is a matrix. The elements of a matrix are denoted with a lower case italic font matching the matrix name with indices indicating the row and column positions of the element, for example $x_{i,j}$ is the element located at the i-th row and j-th-column of matrix $\mathbf{X}$. Vectors are denoted with lower case letters typeset in a boldface font, for example $\mathbf{c}$ is a vector. The elements of a vector are denoted with a lower case italic font matching the vector name with an index indicating the position of the element in the vector, for example c_i is the i-th element of $\mathbf{c}$.

3 An Example: How writers punctuate

This example comes from Brunet (1989), who analyzed the way punctuation marks were used by six French writers: Rousseau, Chateaubriand, Hugo, Zola, Proust, and Giraudoux. In the paper, Brunet gave a table indicating the number of times each of these writers use three punctuation marks: the period, the comma, and all the other marks (*i.e.,* interrogation mark, exclamation mark, colon, and semi-colon) grouped together. These data are reproduced in Table 1. From these data we can build the original data matrix which is denoted $\mathbf{X}$. It has $I = 6$ rows and $J = 3$ columns and is equal to

$$\mathbf{X} = \begin{bmatrix} 7836 & 13112 & 6026 \\ 53655 & 102383 & 42413 \\ 115615 & 184541 & 59226 \\ 161926 & 340479 & 62754 \\ 38177 & 105101 & 12670 \\ 46371 & 58367 & 14299 \end{bmatrix} . \tag{1}$$

In the matrix $\mathbf{X}$, the rows represent the authors and the columns represent types of punctuation marks. At the intersection of a row and a column, we find the number of a given punctuation mark (represented by the column) used by a given author (represented by the row).

4 Analyzing the rows

Suppose that the focus is on the *authors*, and that we want to derive a map that reveals the similarities and differences in punctuation style between authors. In this map, the authors are points and the distances between authors reflect the proximity of style of the authors. So, two authors close to each other punctuate in a similar way and two authors who are far away punctuate differently.

4.1 A first (bad) idea: doing PCA

Figure 1: PCA analysis of the Punctuation. Centered Data. Aloz is a supplementary element. Even though Aloz punctuates the same way as Zola, Aloz is further away from Zola than any other author. The first dimension explains 98% of the variance. It reflects mainly the number of punctuation marks produced by the author.

A first idea is to perform a principal component analysis on $\mathbf{X}$. The result is shown in Figure 1. The plot suggests that the data are quite unidimensional. And, in fact, the first component of this analysis explains 98% of the inertia. How to interpret this component? It seems related to the *number* of punctuation marks produced by each author. This interpretation is supported by creating a fictitious alias for Zola. Suppose that, unbeknown to most historians of French literature, Zola wrote a small novel under the (rather transparent) pseudonym of Aloz. In this novel, he kept his usual way of punctuating, but because this is a short novel, he obviously produced a smaller number of punctuation marks than he did in his complete *œuvre*. Here is the (row) vector recording the number of occurrences of the punctuation marks for Aloz:

$$\begin{bmatrix} 2699 \; 5675 \; 1046 \end{bmatrix} . \tag{2}$$

For ease of comparison, Zola's row vector is reproduced here:

$$\begin{bmatrix} 161926 \; 340479 \; 62754 \end{bmatrix} . \tag{3}$$

So Aloz and Zola have the same punctuation style but differ only in their prolixity. A good analysis should reveal such a similarity of style, but as Figure 1 shows, PCA fails to reveal this similarity. In this figure, we have projected Aloz (as a supplementary element) in the analysis of the authors and Aloz is, in fact, further away from Zola than any other author. This example shows that using PCA to analyze the *style* of the authors is not a good idea because a PCA is sensitive mainly to the number of punctuation marks rather than to way punctuation is used.

Rather than the absolute number, the "style" of the authors, is, in fact, expressed by the *relative frequencies* of their use of the punctuation marks. This suggests the data matrix should be transformed such that each author is described by the *proportion* of his usage of the punctuation marks rather than by the number of marks used. The transformed data matrix is called a *row profile* matrix. In order to obtain the row profiles, we divide each row by its sum. This matrix of row profiles is denoted $\mathbf{R}$. It is computed as:

$$\mathbf{R} = \mathsf{diag}\left\{\mathbf{X}\underset{J\times 1}{\mathbf{1}}\right\}^{-1}\mathbf{X} = \begin{bmatrix} .2905 & .4861 & .2234 \\ .2704 & .5159 & .2137 \\ .3217 & .5135 & .1648 \\ .2865 & .6024 & .1110 \\ .2448 & .6739 & .0812 \\ .3896 & .4903 & .1201 \end{bmatrix} \tag{4}$$

(where diag transforms a vector into a diagonal matrix with the elements of the vector on the diagonal, and $\underset{J\times 1}{\mathbf{1}}$ is a J by 1 vector of ones).

The "average writer" would be someone who uses each punctuation mark according to its proportion in the sample. The profile of this average writer would be the *barycenter* (also called *centroid*, *center of mass*, or *center of gravity*) of the matrix. Here, the barycenter of $\mathbf{R}$ is a vector with $J = 3$ elements, it is denoted $\mathbf{c}$, and computed as

$$\mathbf{c}^T = \underbrace{\left(\underset{1\times I}{\mathbf{1}} \times \mathbf{X} \times \underset{J\times 1}{\mathbf{1}}\right)^{-1}}_{\text{Inverse of the total of } \mathbf{X}} \times \underbrace{\underset{1\times I}{\mathbf{1}}\mathbf{X}}_{\text{Total of the columns of } \mathbf{X}} = \begin{bmatrix} .2973 & .5642 & .1385 \end{bmatrix} . \tag{5}$$

If all authors punctuate the same way, they all punctuate like the average writer therefore, in order to study the differences between authors, we need to analyze the matrix of *deviations* to the average writer. This matrix of deviations is denoted as $\mathbf{Y}$ and it is computed as:

$$\mathbf{Y} = \mathbf{R} - \left(\underset{I\times 1}{\mathbf{1}} \times \mathbf{c}^T\right) = \begin{bmatrix} -.0068 & -.0781 & .0849 \\ -.0269 & -.0483 & .0752 \\ .0244 & -.0507 & .0263 \\ -.0107 & .0382 & -.0275 \\ -.0525 & .1097 & -.0573 \\ .0923 & -.0739 & -.0184 \end{bmatrix} . \tag{6}$$

4.2 Masses (rows) and weights (columns)

In correspondence analysis, we assign a mass to each row and a weight to each column. The mass of each row reflects its importance in the sample. In other words, the mass of each row is the proportion of this row in the total of the table. The masses of the rows are stored in a vector denoted $\mathbf{m}$, which is computed as

$$\mathbf{m} = \underbrace{\left(\underset{1\times I}{\mathbf{1}} \times \mathbf{X} \times \underset{J\times 1}{\mathbf{1}}\right)^{-1}}_{\text{Inverse of the total of } \mathbf{X}} \times \underbrace{\mathbf{X}\underset{J\times 1}{\mathbf{1}}}_{\text{Total of the rows of } \mathbf{X}} = \left[.0189\ .1393\ .2522\ .3966\ .1094\ .0835\right]^T . \quad (7)$$

From the vector $\mathbf{m}$ we define the matrix of masses as $\mathbf{M} = \mathrm{diag}\{\mathbf{m}\}$.

The weight of each column reflects its importance for *discriminating* between the authors. So the weight of a column reflects the information this columns provides to the identification of a given row. Here, the idea is that columns that are used often do not provide much information, and columns that are used rarely provide much information. A measure of how often a column is used is given by the proportion of times it is used, which is equal to the value of this column component of the barycenter. Therefore the weight of a column is computed as the inverse of this column component of the barycenter. Specifically, if we denote by $\mathbf{w}$ the J by 1 weight vector for the columns, we have:

$$\mathbf{w} = [w_j] = \left[c_j^{-1}\right] \quad (8)$$

For our example, we obtain:

$$\mathbf{w} = [w_j] = \left[c_j^{-1}\right] = \begin{bmatrix} \frac{1}{.2973} \\ \frac{1}{.5642} \\ \frac{1}{.1385} \end{bmatrix} = \begin{bmatrix} 3.3641 \\ 1.7724 \\ 7.2190 \end{bmatrix} . \quad (9)$$

From the vector $\mathbf{w}$, we define the matrix of weights as $\mathbf{W} = \mathrm{diag}\{\mathbf{w}\}$.

4.3 Generalized singular value decomposition of $\mathbf{Y}$

Now that we have defined all these notations, correspondence analysis boils down to a *generalized singular value decomposition* (GSVD, see Abdi, 2007) problem. Specifically, matrix $\mathbf{Y}$ is decomposed using the GSVD under the constraints imposed by the matrices $\mathbf{M}$ (masses for the rows) and $\mathbf{W}$ (weights for the columns):

$$\mathbf{Y} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T \quad \text{with:} \quad \mathbf{P}^T\mathbf{M}\mathbf{P} = \mathbf{Q}^T\mathbf{W}\mathbf{Q} = \mathbf{I} . \quad (10)$$

And that way we get:

$$\mathbf{Y} = \underbrace{\begin{bmatrix} 1.7962 & 0.9919 \\ 1.4198 & 1.4340 \\ 0.7739 & -0.3978 \\ -0.6878 & 0.0223 \\ -1.6801 & 0.8450 \\ 0.3561 & -2.6275 \end{bmatrix}}_{\mathbf{P}} \times \underbrace{\begin{bmatrix} .1335 & 0 \\ 0 & .0747 \end{bmatrix}}_{\mathbf{\Delta}} \times \underbrace{\begin{bmatrix} 0.1090 & -0.4114 & 0.3024 \\ -0.4439 & 0.2769 & 0.1670 \end{bmatrix}}_{\mathbf{Q}^T} . \tag{11}$$

The rows of the matrix $\mathbf{X}$ are now represented by their *factor scores* (which are the projections of the observations onto the singular vectors). The row factor scores are stored in an $I = 3$ by $L = 2$ (L stands for the number of non-zero singular values) matrix denoted $\mathbf{F}$. This matrix is obtained as

$$\boxed{\mathbf{F} = \mathbf{P\Delta}} = \begin{bmatrix} 0.2398 & 0.0741 \\ 0.1895 & 0.1071 \\ 0.1033 & -0.0297 \\ -0.0918 & 0.0017 \\ -0.2243 & 0.0631 \\ 0.0475 & -0.1963 \end{bmatrix} . \tag{12}$$

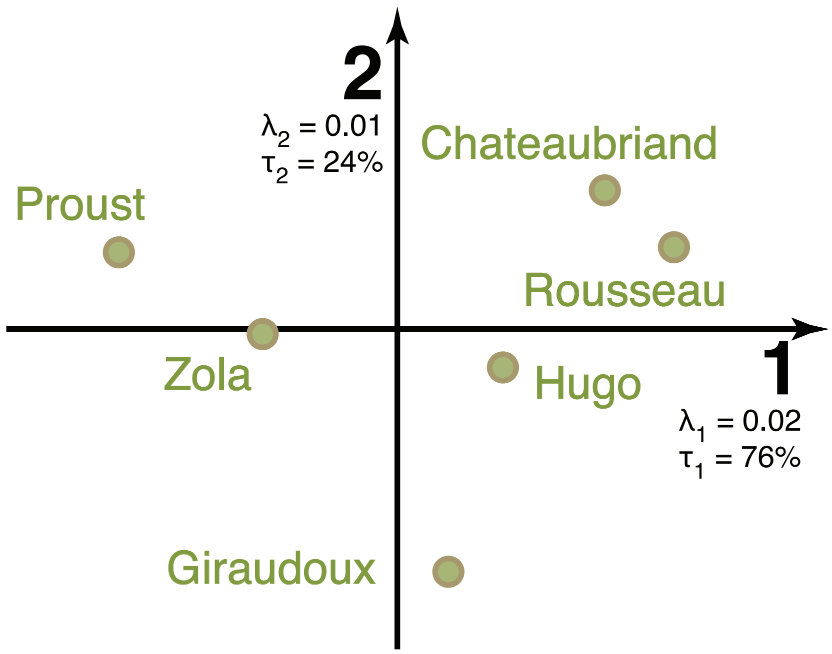

Figure 2: Plot of the correspondence analysis of the rows of matrix $\mathbf{X}$. The first two factors of the analysis for the rows are plotted (*i.e.*, this is the matrix $\mathbf{F}$). Each point represents an author. The variance of each factor score is equal to its eigenvalue.

The variance of the factor scores for a given dimension is equal to the squared singular value of this dimension (*Nota Bene:* the variance of the observations is computed taking into account their masses). Or, equivalently, we say that the variance of the factor scores is equal to the eigenvalue of this dimension (*i.e.,* the eigenvalue is the square of the singular value). This can be checked as follows:

$$\mathbf{F}^T\mathbf{M}\mathbf{F} = \mathbf{\Delta}^2 = \mathbf{\Lambda} = \begin{bmatrix} 0.1335^2 & 0 \\ 0 & 0.0747^2 \end{bmatrix} = \begin{bmatrix} 0.0178 & 0 \\ 0 & 0.0056 \end{bmatrix} . \tag{13}$$

We can display the results by plotting the factor scores as a map where each point represents a row of the matrix $\mathbf{X}$ (*i.e.,* each point represents an author). This is done in Figure 2. On this map, the first dimension seems to be related to time (the rightmost authors are earlier authors, the leftmost authors are more recent), with the exception of Giraudoux who is a very recent author. The second dimension singularizes Giraudoux. These factors will be easier to understand after we have analyzed the columns. This can be done by analyzing the matrix $\mathbf{X}^T$. Equivalently this can be done by doing what is called the *dual analysis.*

5 Geometry of the GSVD

CA has a simple geometric interpretation. For example, when a row profile is interpreted as a vector, it can be represented as a point in a multidimensional space. Because the sum of a profile is equal to one, row profiles are, in fact points in a $J - 1$ dimensional space. Also, because the components of a row profile take value in the interval [0 1], the points representing these row-profiles can only lay in the subspace whose "extreme points" have one component equal to one and all other components equal to zero. This subspace is called a *simplex.* For example, Figure 3 shows the 2-dimensional simplex corresponding to the subspace of all possible row profiles with three components. As an illustration, the point describing Rousseau (with coordinates equal to [.2905 .4861 2234]) is also plotted. For this particular example, the simplex is an equilateral triangle and, so the three dimensional row profiles can conveniently be represented as points on this triangle as illustrated in Figure 4a which shows the simplex of Figure 3 in two dimensions. Figure 4b shows all six authors and the barycenter.

The weights of the columns, which are used as constrains in the GSVD have also a straightforward geometric interpretation. As illustrated in Figure 5, each side of the simplex is stretched by a quantity equal to the square root of the dimension it represents (we use the square root because we are interested in squared distances but not in squared weights so using the square root of the weights insures that the *squared* distances between authors will take into account the weight rather than the squared weights).

The masses of the rows are taken into account to find the dimensions. Specifically, the first factor is computed in order to obtain the maximum possible value of the sum of the masses times the squared projections of the authors points (*i.e.,* the projections have the largest possible variance). The second factor is constrained to be orthogonal (taking into account the masses) to the first one and have the largest variance for the projections. The remaining factors are computed with similar constraints. Figure 6 shows the stretched simplex, the

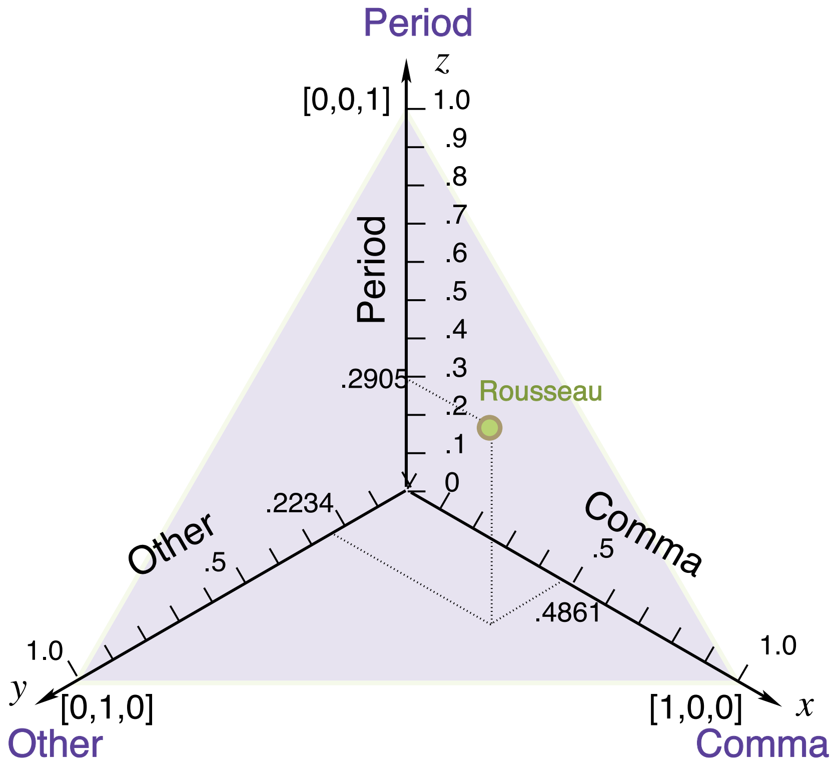

Figure 3: In three dimensions, the simplex is a two dimensional triangle whose vertices are the vectors $[1\ 0\ 0]$, $[0\ 1\ 0]$ and $[0\ 0\ 1]$. The point describing Rousseau (with coordinates $[.2905\ .4861\ 2234]$) is also plotted.

author points, and the two factors (note that the origin of the factors is the barycenter of the authors).

The "stretched simplex" shows the whole space of the possible profiles. Figures 6 shows that the authors occupy a small portion of the whole space: They do not vary much in the way they punctuate. Also the "stretched simplex" represents the columns as the vertices of the simplex: The columns are represented as row profiles with the column component being one and all the other components being zeros. This representation is called an *asymmetric* representation because the rows always have a dispersion smaller than (or equal to) the columns.

6 Distance, Inertia, Chi-square, and CA

6.1 Chi-squared distances

In CA, the Euclidean distance in the "stretched simplex" is equivalent to a weighted distance in the original space. For reasons which will be made more clear later, this distance is called the χ^2 *distance*. The χ^2 distance between two row profiles i and i' can be computed from

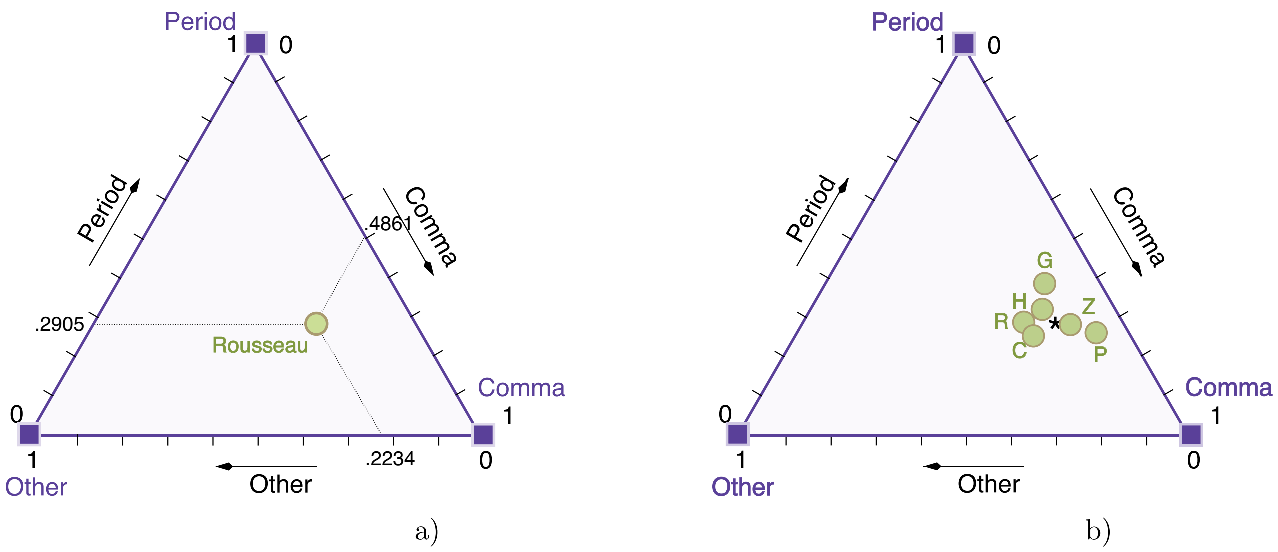

Figure 4: The simplex as a triangle. *a)* With Rousseau (compare with Figure 3) *b)* With all six authors and their barycenter.

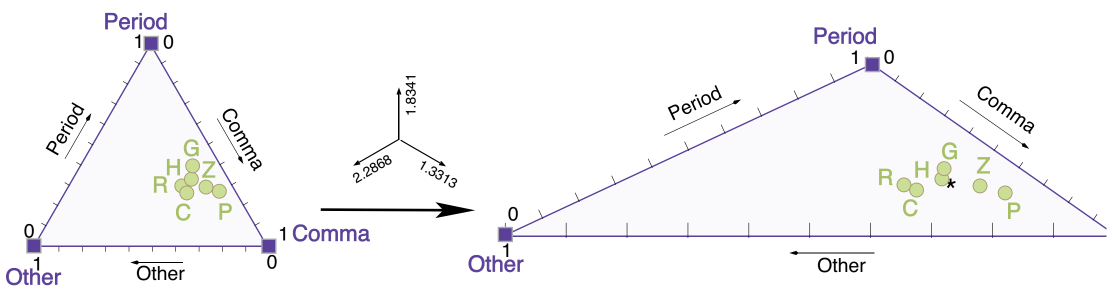

Figure 5: Geometric interpretation of the columns weights. Each side of the simplex is stretched by a factor equal to the square root of the weights.

the factor scores as

$$d^2_{i,i'} = \sum_{\ell}^{L} \left(f_{i,\ell} - f_{i',\ell}\right)^2 \tag{14}$$

or from the row-profiles as

$$d^2_{i,i'} = \sum_{j}^{J} w_j \left(r_{i,j} - r_{i',j}\right)^2 \quad . \tag{15}$$

6.2 Inertia

The variability of the row profiles relative to their barycenter is measured by a quantity—akin to variance—called *inertia* and denoted $\mathcal{I}$. The inertia of the rows to their barycenter is computed as the weighed sum of the squared distances of the rows to their barycenter. We

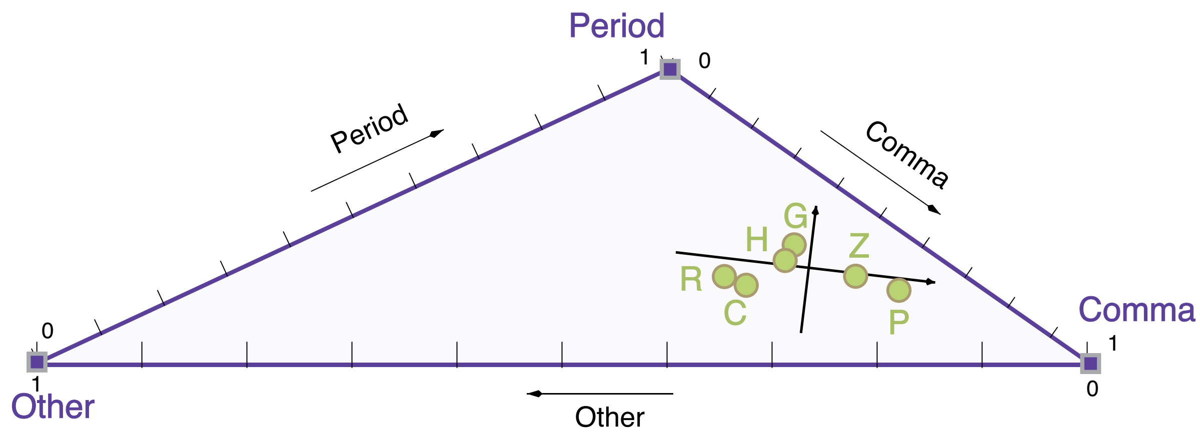

Figure 6: Correspondence analysis: The "stretched simplex" along with the factorial axes. The projections of the Authors' point onto the factorial axes give the factor scores.

denote by $d^2_{\mathbf{c},i}$ the (squared) distance of the i-th row to the barycenter, it is computed as

$$d^2_{\mathbf{c},i} = \sum_j^J w_j \left(r_{i,j} - c_j\right)^2 = \sum_\ell^L f^2_{i,\ell} \tag{16}$$

where L is the number of factors extracted by the CA of the table, [this number is always smaller or equal to $\min(I,J)-1$]. The inertia of the rows to their barycenter is then computed as

$$\mathcal{I} = \sum_i^I m_i d^2_{\mathbf{c},i} \ . \tag{17}$$

The inertia can also be expressed as the sum of the eigenvalues (see Equation 13):

$$\mathcal{I} = \sum_\ell^L \lambda_\ell \ . \tag{18}$$

This shows that in CA, each factor extracts a portion of the inertia, with the first factor extracting the largest portion, the second factor extracting the largest portion left of the inertia, etc.

6.3 Inertia and the Chi-squared test

Interestingly, the inertia in CA is closely related to the chi-square test. This test is traditionally performed on a contingency table in order to test the independence of the rows and the columns of the table. Under independence, the frequency of each cell of the table should be proportional to the product of its row and column marginal probabilities. So if we denote by $x_{+,+}$ the grand total of matrix $\mathbf{X}$, the expected frequency of the cell at the i-th row and

j-th column is denoted $E_{i,j}$ and computed as:

$$E_{i,j} = m_i c_j x_{+,+} \ . \tag{19}$$

The chi-squared test statistic, denoted χ^2 is computed as the sum of the squared difference between the actual values and the expected values weighted by the expected values:

$$\chi^2 = \sum_{i,j} \frac{(x_{i,j} - E_{i,j})^2}{E_{i,j}} \ . \tag{20}$$

When rows and columns are independent, χ^2 follows a chi-squared distribution with $(I-1)(J-1)$ degrees of freedom. Therefore, χ^2 can be used to evaluate the likelihood of the row and columns independence hypothesis. The statistics χ^2 can be rewritten to show its close its relationship with the inertia of CA, namely:

$$\chi^2 = \mathcal{I} x_{+,+} \ . \tag{21}$$

This shows that CA analyzes—in orthogonal components— the pattern of deviations to independence.

7 Dual Analysis: The column space

In a contingency table, the rows and the columns of the table play a similar role, and therefore the analysis that was performed on the rows can also be performed on the columns by exchanging the role of the rows and the columns. This is illustrated by the analysis of the columns of matrix $\mathbf{X}$, or equivalently by the rows of the transposed matrix $\mathbf{X}^T$. The matrix of column profiles for $\mathbf{X}^T$ is called $\mathbf{O}$ (like cOlumn), it is computed as

$$\mathbf{O} = \mathsf{diag}\left\{\mathbf{X}^T \underset{I \times 1}{\mathbf{1}}\right\}^{-1} \mathbf{X}^T \tag{22}$$

The matrix of the deviations to the barycenter is called $\mathbf{Z}$ and it is computed as:

$$\mathbf{Z} = \mathbf{O} - \left(\underset{I \times 1}{\mathbf{1}} \times \mathbf{m}^T\right) = \begin{bmatrix} -.0004 & -.0126 & .0207 & -.0143 & -.0193 & .0259 \\ -.0026 & -.0119 & -.0227 & .0269 & .0213 & -.0109 \\ .0116 & .0756 & .0478 & -.0787 & -.0453 & -.0111 \end{bmatrix} .$$

Weights and masses of the columns analysis are the inverse of their equivalent for the row analysis. This implies that the punctuation marks factor scores are obtained from the GSVD with the constraints imposed by the two matrices $\mathbf{W}^{-1}$ (masses for the rows) and $\mathbf{M}^{-1}$ (weights for the columns, compare with Equation 10):

$$\mathbf{Z} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T \qquad \text{with:} \qquad \mathbf{U}^T\left(\mathbf{W}^{-1}\right)\mathbf{U} = \mathbf{V}^T\left(\mathbf{M}^{-1}\right)\mathbf{V} = \mathbf{I}\ . \tag{23}$$

This gives:

$$\mathbf{Z} = \underbrace{\begin{bmatrix} 0.3666 & -1.4932 \\ -0.7291 & 0.4907 \\ 2.1830 & 1.2056 \end{bmatrix}}_{\mathbf{U}} \times \underbrace{\begin{bmatrix} .1335 & 0 \\ 0 & .0747 \end{bmatrix}}_{\boldsymbol{\Delta}} \times \underbrace{\begin{bmatrix} 0.0340 & 0.1977 & 0.1952 & -0.2728 & -0.1839 & 0.0298 \\ 0.0188 & 0.1997 & -0.1003 & 0.0089 & 0.0925 & -0.2195 \end{bmatrix}}_{\mathbf{V}^T}\ . \tag{24}$$

The factor scores for the punctuation marks are stored in a $J = 3 \times L = 2$ matrix called $\mathbf{G}$ which is computed in the same way $\mathbf{F}$ was computed (see Equation 12 on page 6). So, $\mathbf{G}$ is computed as:

$$\mathbf{G} = \mathbf{U}\boldsymbol{\Delta} = \begin{bmatrix} 0.0489 & -0.1115 \\ -0.0973 & 0.0367 \\ 0.2914 & 0.0901 \end{bmatrix}\ . \tag{25}$$

7.1 Transition formula: from the rows to the columns and back

Comparing Equation 24 with Equation 11, shows that the singular values are the same for both analyses. This means that the inertia (*i.e.,* the square of the singular value) extracted by each factor is the same for both analyses. Because the variance extracted by the factors can be added, to obtain the total inertia of the data table, this also means that each analysis is decomposing the same inertia which, here, is equal to:

$$\mathcal{I} = .1335^2 + .747^2 = .0178 + .0056 = 0.0234\ . \tag{26}$$

Also, the generalized singular decomposition of one set (say the columns) can be obtained from the other one (say the rows). For example the generalized singular vectors of the analysis of the columns can be computed directly from the analysis from the rows as

$$\mathbf{U} = \mathbf{W}\mathbf{Q}\ . \tag{27}$$

Combining Equations 27 and 25 shows that the factors for the rows of $\mathbf{Z}$ (*i.e.,* the punctuation marks) can be obtained directly from the singular value decomposition of the authors matrix (*i.e.,* the matrix $\mathbf{Y}$) as

$$\mathbf{G} = \mathbf{W}\mathbf{Q}\boldsymbol{\Delta}\ . \tag{28}$$

As a consequence, we can, in fact, find directly the factor scores of the columns from their profile matrix (*i.e.,* the matrix $\mathbf{O}$), and from the factor scores of the rows. Specifically, the equation which gives the values of $\mathbf{O}$ from $\mathbf{F}$ is

$$\mathbf{G} = \mathbf{OF}\boldsymbol{\Delta}^{-1} , \tag{29}$$

and conversely $\mathbf{F}$ could be obtained from $\mathbf{G}$ as

$$\mathbf{F} = \mathbf{RG}\boldsymbol{\Delta}^{-1} . \tag{30}$$

These equations are called "*transition formulas from the rows to the columns*" (and vice versa) or simply the *transition formulas.*

7.2 One single GSVD for CA

Because the factor scores obtained for the rows and the columns have the same variance (*i.e.,* they have the same "scale"), it is possible to plot them in the same space. This is illustrated in Figure 7. The symmetry of the rows and the columns in CA is revealed by the possibility of *directly* obtaining the factors scores from one single GSVD. Specifically, let $\mathbf{D_m}$ (respectively $\mathbf{D_c}$) denote the diagonal matrices with the elements of $\mathbf{m}$ (respectively $\mathbf{c}$ on the diagonal, and $\mathbf{N}$ denote the matrix $\mathbf{X}$ divided by the sum of all its elements. This matrix is called a *stochastic* matrix, all its elements are larger than zero and their sum is equal to one. The factors scores for the rows and the columns are obtained from the following GSVD:

$$\left(\mathbf{N} - \mathbf{mc}^T\right) = \mathbf{S}\boldsymbol{\Delta}\mathbf{T}^T \quad \text{with} \quad \mathbf{S}^T\mathbf{D}_{\mathbf{m}}^{-1}\mathbf{S} = \mathbf{T}^T\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{T} = \mathbf{I} . \tag{31}$$

The factor scores for the rows ($\mathbf{F}$) and the columns ($\mathbf{G}$) are obtained respectively as

$$\mathbf{F} = \mathbf{D}_{\mathbf{m}}^{-1}\mathbf{S}\boldsymbol{\Delta} \quad \text{and} \quad \mathbf{G} = \mathbf{D}_{\mathbf{c}}^{-1}\mathbf{T}\boldsymbol{\Delta} . \tag{32}$$

8 Supplementary elements

Often in CA we want to know the position in the analysis of rows or columns that were not analyzed. These rows or columns are called illustrative or supplementary rows or columns (or supplementary observations or variables). By contrast with the appellation of supplementary (which are *not* used to compute the factors) the *active* elements are those used to compute the factors. Table 2 shows the punctuation data table with four additional columns giving the detail of the "other punctuation marks" (*i.e.,* the exclamation mark, the interrogation mark, the semi-colon, and the colon). These punctuation marks were not analyzed for two reasons: first these marks are too rare and therefore they would distort the factor space and second the "Other" marks comprises all these other marks and therefore to analyze them with "Other" would be redundant. There is also a new author in Table 2: We counted

Table 2: Number of punctuation marks used by six major French authors (from Brunet, 1989). The exclamation point, question mark, semicolon, and colon are supplementary columns. Abdi (1994) Chapter 1 is a supplementary row. Notations x_{i+}: sum of the i-th row; x_{+j}: sum of the j-th column; x_{++}: grand total.

	Active Elements					Supplementary Elements			
	Period	Comma	Other Marks	x_{i+}	$\mathbf{m}$ $\frac{x_{i+}}{x_{++}}$	Exclamation	Question	Semicolon	Colon
Rousseau	7836	13112	6026	**26974**	**.0189**	413	1240	3401	972
Chateaubriand	53655	102383	42413	**198451**	**.1393**	4669	4595	19354	13795
Hugo	115615	184541	59226	**359382**	**.2522**	19513	9876	22585	7252
Zola	161926	340479	62754	**565159**	**.3966**	24025	10665	18391	9673
Proust	38117	105101	12670	**155948**	**.1094**	2756	2448	3850	3616
Giraudoux	46371	58367	14229	**119037**	**.0835**	5893	5042	1946	1418
x_{+j}	**423580**	**803983**	**197388**	**1424951**					
$\mathbf{w}^{\mathsf{T}} = \frac{x_{++}}{x_{+j}}$	**3.3641**	**1.7724**	**7.2190**	x_{++}					
$\mathbf{c}^{\mathsf{T}} = \frac{x_{+j}}{x_{++}}$	**.2973**	**.5642**	**.1385**						
Abdi (Chapter 1)	216	139	26						

Table 3: Factor scores, contributions, and cosines for the Rows. Negative contributions are shown in italic. Abdi (1994) Chapter 1 is a supplementary row.

Axis	λ	%	Rousseau	Chateaubriand	Hugo	Zola	Proust	Giraudoux	Abdi (Chapter 1)
					Factor Scores				
1	.0178	76	0.2398	0.1895	0.1033	-0.0918	-0.2243	0.0475	-0.0908
2	.0056	24	0.0741	0.1071	-0.0297	0.0017	0.0631	-0.1963	0.5852
					Contributions				
1			0.0611	0.2807	0.1511	*0.1876*	*0.3089*	0.0106	–
2			0.0186	0.2864	*0.0399*	0.0002	0.0781	*0.5767*	–
					Cosines				
1			0.9128	0.7579	0.9236	0.9997	0.9266	0.0554	0.0235
2			0.0872	0.2421	0.0764	0.0003	0.0734	0.9446	0.9765
					Squared Distances to Grand Barycenter				
–	–	–	0.0630	0.0474	0.0116	0.0084	0.0543	0.0408	0.3508

Table 4: Factor scores, contributions, and cosines for the columns. Negative contributions are shown in italic. Exclamation mark, question mark, semicolon, and colon are supplementary columns.

Axis	λ	%	Period	Comma	Other Marks	Exclamation	Question	Semicolon	Colon
					Factor Scores				
1	.0178	76	-0.0489	0.0973	-0.2914	-0.0596	-0.1991	-0.4695	-0.4008
2	.0056	24	0.1115	-0.0367	-0.0901	0.2318	0.2082	-0.2976	-0.4740
					Contributions				
1			*0.0399*	0.2999	*0.6601*	–	–	–	–
2			0.6628	*0.1359*	*0.2014*	–	–	–	–
					Cosines				
1			0.1614	0.8758	0.9128	0.0621	0.4776	0.7133	0.4170
2			0.8386	0.1242	0.0872	0.9379	0.5224	0.2867	0.5830
					Squared Distances to Grand Barycenter				
–	–		0.0148	0.0108	0.0930	0.0573	0.0830	0.3090	0.3853

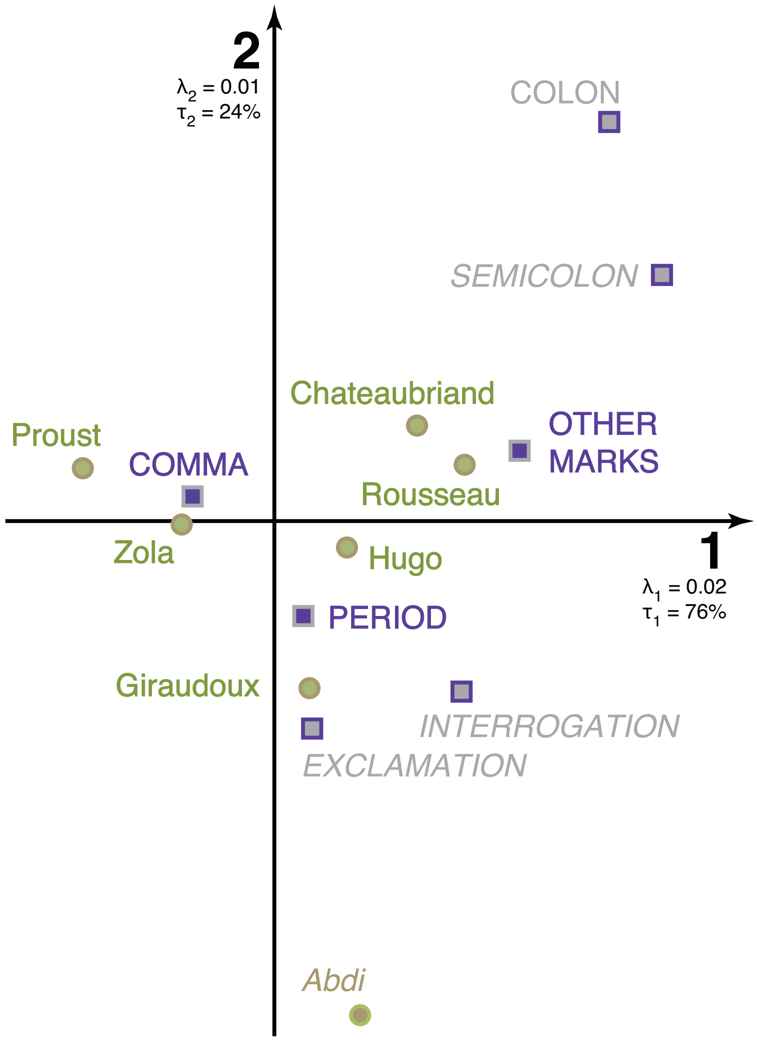

Figure 7: Correspondence analysis of the punctuation of six authors, Comma, Period, and Others are active columns; Rousseau. Chateaubriand, Hugo, Zola, Proust, and Giraudoux are active rows; Colon, Semicolon, Interrogation and Exclamation are supplementary columns; Abdi is a supplementary row.

the marks used by a different author, namely Hervé Abdi in the first chapter of his 1994 book called *"Les réseaux de neurones."* This author was not analyzed because the data are available for only one chapter (not his complete work) and also because this author is not a literary author.

The values of the projections on the factors for the supplementary elements are computed from the *transition formula*. Specifically, a supplementary row is projected into the space defined using the transition formula for the active rows (*cf.* Equation 30) and replacing the active row profiles by the supplementary row profiles. So, if we denote by $\mathbf{R}_{\text{sup}}$ the matrix of the supplementary row profiles, then $\mathbf{F}_{\text{sup}}$—the matrix of the supplementary row factor scores—is computed as:

$$\mathbf{F}_{\text{sup}} = \mathbf{R}_{\text{sup}} \times \mathbf{G} \times \boldsymbol{\Delta}^{-1} \ . \tag{33}$$

For example, the factor scores of the author Abdi is computed as

$$\mathbf{F}_{\text{sup}} = \mathbf{R}_{\text{sup}}\mathbf{G}\mathbf{\Delta}^{-1} = \begin{bmatrix} 0.0908 & -0.5852 \end{bmatrix} . \tag{34}$$

Supplementary columns are projected into the factor space using the transition formula from the active rows (*cf.* Equation 29) and replacing the active column profiles by the supplementary column profiles. If we denote by $\mathbf{O}_{\text{sup}}$ the supplementary column profile matrix, then $\mathbf{G}_{\text{sup}}$, the matrix of the supplementary column factor scores, is computed as:

$$\mathbf{G}_{\text{sup}} = \mathbf{O}_{\text{sup}}\mathbf{F}\mathbf{\Delta}^{-1} . \tag{35}$$

Table 4 gives the factor scores for the supplementary elements.

9 Little Helpers: Contributions and cosines

Contributions and cosines are coefficients whose goal is to facilitate the interpretation. The contributions identify the important elements for a given factor, whereas the (squared) cosines identify the factors important for a given element. These coefficients express importance as the proportion of something into a total. The contribution is the ratio of the weighted squared projection of an element on a factor by the sum of the weighted projections of all the elements for this factor (which happens to be the eigenvalue of this factor). The squared cosine is the ratio of the squared projection of an element on a factor by the sum of the projections of this element on all the factors (which happens to be the squared distance from this point to the barycenter). Contributions and squared cosines are proportions that vary between 0 and 1.

The squared *cosines*, denoted h, between row i and factor ℓ (respectively and column j and factor ℓ) are obtained as:

$$h_{i,\ell} = \frac{f_{i,\ell}^2}{\sum\limits_{\ell} f_{i,\ell}^2} = \frac{f_{i,\ell}^2}{d_{\mathbf{c},i}^2} \qquad \text{and} \qquad h_{j,\ell} = \frac{g_{j,\ell}^2}{\sum\limits_{\ell} f_{j,\ell}^2} = \frac{g_{j,\ell}^2}{d_{\mathbf{r},j}^2} . \tag{36}$$

Squared cosines help locating the factors important for a given observation. The *contributions*, denoted b, of row i to factor ℓ and of column j to factor ℓ are obtained respectively as:

$$b_{i,\ell} = \frac{m_i f_{i,\ell}^2}{\sum\limits_{i} m_i f_{i,\ell}^2} = \frac{m_i f_{i,\ell}^2}{\lambda_\ell} \qquad \text{and} \qquad b_{j,\ell} = \frac{c_j g_{j,\ell}^2}{\sum\limits_{j} c_j f_{j,\ell}^2} = \frac{c_i g_{j,\ell}^2}{\lambda_\ell} . \tag{37}$$

Contributions help locating the observations important for a given factor. An often used rule of thumb is to consider that the important contributions are larger than the average contribution, which is equal to the number of elements (*i.e.,* $\frac{1}{I}$ for the rows and $\frac{1}{J}$ for the columns). A dimension is then interpreted by opposing the positive elements with large contributions to the negative elements with large contributions. Cosines and contributions for the punctuation example are given in Tables 3 and 4.

10 Multiple correspondence analysis

Correspondence analysis works with a contingency table which is equivalent to the analysis of two nominal variables (*i.e.,* one for the rows and one for the columns). Multiple correspondence analysis (MCA) is an extension of CA which allows the analysis of the pattern of relationship among several nominal variables. MCA is used to analyze a set of observations described by a set of nominal variables. Each nominal variable is comprised of several levels, and each of these levels is coded as a binary variable. For example gender (F *vs.* M) is a nominal variable with two levels. The pattern for a male respondent will be [0 1] and [1 0] for a female. The complete data table is composed of binary columns with one and only one column taking the value "1" per nominal variable.

MCA can also accommodate quantitative variables by recoding them as "bins." For example, a score with a range of −5 to +5 could be recoded as a nominal variable with three levels: less than 0, equal to 0, or more than 0. With this schema, a value of 3 will be expressed by the pattern 0 0 1. The coding schema of MCA implies that each row has the same total, which for CA implies that each row has the same *mass.*

Essentially, MCA is computed by using a CA program on the data table. It can be shown that the binary coding scheme used in MCA create artificial factors and therefore artificially reduces the inertia explained the first factors of the analysis. A solution for this problem is to correct the eigenvalues obtained from the CA program (see Greenacre, 2007; Greenacre & Blasius, 2006; Abdi, 2007).

Related entries

Barycentric discriminant analyis (BADIA), Canonical correlation analysis, categorical variables, Chi-square test, Coefficient alpha (Cronbach's alpha), Data mining, Descriptive discriminant analysis, Discriminant analysis, Exploratory data analysis, Exploratory factor analysis, Guttman scaling, Matrix algebra, Principal component analysis, R.

Further readings

1. Abdi, H. (2003). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds): *Encyclopedia for research methods for the social sciences.* Thousand Oaks: Sage.

2. Abdi, H. (2007a). Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD). In N.J. Salkind (Ed.): *Encyclopedia of measurement and statistics*. Thousand Oaks (CA): Sage. pp. 907–912.
3. Abdi, H. (2007b). Distance. In N.J. Salkind (Ed.): *Encyclopedia of measurement and statistics*. Thousand Oaks (CA): Sage. pp. 280–284.
4. Abdi, H. (2007c). Discriminant correspondence analysis. In N.J. Salkind (Ed.): *Encyclopedia of measurement and statistics*. Thousand Oaks (CA): Sage. pp. 270–275.
5. Abdi, H., & Valentin, D. (2007). Multiple correspondence analysis. In N.J. Salkind (Ed.): *Encyclopedia of measurement and statistics*. Thousand Oaks (CA): Sage. pp. 651–657.
6. Benzécri, J.P. (1973). *L'analyse des données (2 vol.)*. Paris: Dunod
7. Brunet, E. (1989). Faut-il pondérer les données linguistiques. *CUMFID*, **16**, 39–50.
8. Clausen, S.E. (1998). *Applied correspondence analysis*. Thousand Oaks (CA): Sage.
9. Escofier, B., & Pagès, J. (1998). *Analyses factorielles simples et multiples*. Paris: Dunod.
10. Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
11. Greenacre, M.J. (2007). *Correspondence analysis in practice (2nd Edition)*. Boca Raton (FL): Chapman & Hall/CRC.
12. Greenacre, M.J. & Blasius, J. (Eds.) (2007). *Multiple correspondence analysis and related methods*. Boca Raton (FL): Chapman & Hall/CRC.
13. Hwang, H., Tomiuk, M. A., & Takane, Y. (in press). Correspondence analysis, multiple correspondence analysis and recent developments. In R. Millsap & A. Maydeu-Olivares (Eds.). *Handbook of quantitative methods in psychology.* London: Sage Publications. pp.
14. Lebart, L, & Fénelon, J.P. (1971). *Statistiques et informatique appliquées*. Paris: Dunod.
15. Weller S.C., & Romney, A.K. (1990). *Metric scaling: Correspondence analysis*. Thousand Oaks (CA): Sage.