C

Correspondence Analysis

Hervé Abdi[1] and Michel Béra[2]
[1]School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA
[2]Centre d'Étude et de Recherche en Informatique et Communications, Conservatoire National des Arts et Métiers, Paris, France

Synonyms

Dual scaling; Hayashi's quantification theory type III; Homogeneity analysis; Method of simultaneous linear regressions; Optimal scaling; Reciprocal averaging

Glossary

CA	Correspondence analysis
Component	A linear combination of the variables of a data table also called dimension or factor
Dimension	See component
Factor	See component
GSVD	Generalized singular value decomposition
MCA	Multiple correspondence analysis
PCA	Principal component analysis
SVD	Singular value decomposition

Definition

Correspondence analysis (CA) is a generalization of principal component analysis tailored to handle nominal variables. CA is traditionally used to analyze contingency tables, but is also often used with data matrices that comprise only nonnegative data. CA decomposes the chi-square statistics associated to the data table into two sets of orthogonal components that describe, respectively, the pattern of associations between the elements of the rows and between the elements of the columns of the data table. When the data table is a set of observations described by a set of nominal variables, CA becomes multiple correspondence analysis (MCA).

Introduction

Correspondence analysis (CA; Cordier 1965; Benzécri 1973; Beh and Lombardo 2015; Lebart and Fénelon 1975; Lebart et al. 1984; Escofier and Pagès 1990; Greenacre 1984, 2007; Abdi and Valentin 2007; Hwang et al. 2010; Abdi 2003; Abdi and Williams 2010b) is an extension of principal component analysis (PCA; for details, see Abdi and Williams 2010a) tailored to handle nominal variables. Originally, CA was developed to analyze contingency tables in which a sample of observations is described by two nominal variables, but it was rapidly extended to the analysis of any data matrices with nonnegative entries.

R. Alhajj, J. Rokne (eds.), *Encyclopedia of Social Network Analysis and Mining*,
https://doi.org/10.1007/978-1-4614-7163-9_140-2

The origin of CA can be traced to the early work of Pearson (1901) or Fisher, but the modern version of correspondence analysis and its geometric interpretation came from the 1960s in France and is associated with the French school of "data analysis" (*analyse des données*) and was developed under the leadership of Jean-Paul Benzécri. As a technique, it was often discovered (and rediscovered), and so variations of CA can be found under several different names such as "dual scaling," "optimal scaling," "homogeneity analysis," or "reciprocal averaging." The multiple identities of correspondence analysis are a consequence of its large number of properties: Correspondence analysis can be defined as an optimal solution for a lot of apparently different problems.

Key Points

CA transforms a data table into two sets of new variables called *factor scores* (obtained as linear combinations of, respectively, the rows and columns): one set for the rows and one set for the columns. These factor scores give the best representation of the similarity structure of, respectively, the rows and the columns of the table. In addition, the factor scores can be plotted as maps that optimally display the information in the original table. In these maps, rows and columns are represented as points whose coordinates are the factor scores and where the dimensions are also called *factors*, *components* (by analogy with PCA), or simply *dimensions*. Interestingly, the factor scores of the rows and the columns have the same variance and, therefore, the rows and columns can be conveniently represented in one single map.

In correspondence analysis, the total variance (often called inertia) of the factor scores is proportional to the independence chi-square statistic of this table and, therefore, the factor scores in CA decompose this χ^2 into orthogonal components.

Historical Background

Correspondence analysis can be traced back to the work of Karl Pearson (1901, 1906) and seems to have been independently discovered and rediscovered many times by different authors, in part because CA is the solution to different optimization problems (see for more details and references the historical note of Lebart and Saporta 2014). For example (among others), Haley in 1935, Fisher in 1940, Maung in 1941, Guttman in 1941, Burt in 1950, Hayashi in 1950, and Kendall and Stuart in 1961, all proposed methods that were essentially equivalent to CA. However, the modern approach of CA was first exposed in the 1965 Ph.D. dissertation of Brigitte Cordier (later better known as Brigitte Escofier; part of her dissertation was also published as Escofier 1969) written under the direction of Jean-Paul Benzécri. Since then, CA has been declined in multiple different but related methods and is becoming a very popular method particularly adapted to very large data sets such as those found in Big Data problems.

Correspondence Analysis: Theory and Practice

Notations

Matrices are denoted by uppercase bold letters, vectors are denoted by lowercase bold, and their elements are denoted by lowercase italic. Matrices, vectors, and elements from the same matrix all use the same letter (e.g., $\mathbf{A}$, $\mathbf{a}$, a). The transpose operation is denoted by $^\top$; the inverse operation is denoted by $^{-1}$. The identity matrix is denoted $\mathbf{I}$, vectors or matrices of ones are denoted $\mathbf{1}$, and matrices or vectors of zeros are denoted $\mathbf{0}$. When provided with a square matrix, the diag operator gives a vector with the diagonal elements of this matrix. When provided with a vector, the diag operator gives a diagonal matrix with the elements of the vector as the diagonal elements of this matrix. When provided with a square matrix, the trace operator gives the sum of the diagonal elements of this matrix.

The data table to be analyzed by CA is a contingency table (or at least a data table with nonnegative entries) with I rows and J columns. It is represented by the $I \times J$ matrix $\mathbf{X}$, whose generic element $x_{i,j}$ gives the number of observations that belong to the ith level of the first nominal variables (i.e., the rows) and the jth level of the second nominal variables (i.e., the columns). The grand total of the table is noted N.

Computations

The first step of the analysis is to transform the data matrix into a probability matrix (i.e., a matrix comprising nonnegative numbers and whose sum is equal to 1) denoted $\mathbf{Z}$ and computed as $\mathbf{Z} = N^{-1}\mathbf{X}$. We denote $\mathbf{r}$ the vector of the row totals of $\mathbf{Z}$ (i.e., $\mathbf{r} = \mathbf{Z1}$, with $\mathbf{1}$ being a conformable vector of 1 s), $\mathbf{c}$ the vector of the column totals (i. e., $\mathbf{c} = \mathbf{Z}^\top\mathbf{1}$), and $\mathbf{D_c} = \mathrm{diag}\{\mathbf{c}\}$, $\mathbf{D_r} = \mathrm{diag}\{\mathbf{r}\}$. The factor scores are obtained from the following generalized singular value decomposition (GSVD; for details on the singular value decomposition, see Abdi 1988, 2007a, 2007b, Good 1969, Takane 2002, Hotelling 1933, Eckart and Young 1936, and Stewart 1993):

$$\begin{aligned}\left(\mathbf{Z} - \mathbf{rc}^\top\right) &= \mathbf{P\Delta Q}^\top \text{with} \\ \mathbf{P}^\top\mathbf{D_r}^{-1}\mathbf{P} &= \mathbf{Q}^\top\mathbf{D_c}^{-1}\mathbf{Q} = \mathbf{I}. \end{aligned} \quad (1)$$

Note that the subtraction of the matrix $\mathbf{rc}^\top$ from $\mathbf{Z}$ is equivalent to a double centering of matrix $\mathbf{Z}$ (Abdi 2007c, 2007e; Abdi et al. 1997). The matrix $\mathbf{P}$ (respectively, $\mathbf{Q}$) contains the left (respectively, right) generalized singular vectors of $(\mathbf{Z} - \mathbf{rc}^\top)$, and the diagonal elements of the diagonal matrix $\mathbf{\Delta}$ give its singular values. The squared singular values, which are called eigenvalues, are denoted λ_ℓ and stored into the diagonal matrix $\mathbf{\Lambda}$. Eigenvalues express the variance extracted by the corresponding factor, and their sum is called the total inertia (denoted $\mathcal{I}$) of the data matrix. With the so-called triplet notation (Escoufier 2007) that is sometimes used as a general framework to formalize multivariate techniques, CA is equivalent to the analysis of the triplet $\left((\mathbf{Z} - \mathbf{rc}^\top), \mathbf{D_c}^{-1}, \mathbf{D_r}^{-1}\right)$.

From the GSVD, the row and (respectively) column factor scores are obtained as

$$\mathbf{F} = \mathbf{D_r}^{-1}\mathbf{P\Delta} \quad \text{and} \quad \mathbf{G} = \mathbf{D_c}^{-1}\mathbf{Q\Delta}. \quad (2)$$

Note that the factor scores of a given set (i.e., the rows or the columns) are pairwise orthogonal when they describe different dimensions and that the variance of the factor scores for a given dimension is equal to the eigenvalue associated with this dimension. So, for example, the variance of the row factor scores is computed as

$$\begin{aligned}\mathbf{F}^\top\mathbf{D_r}\mathbf{F} &= \mathbf{\Delta P}^\top\mathbf{D_r}^{-1}\mathbf{D_r}\mathbf{D_r}^{-1}\mathbf{P\Delta} \\ &= \mathbf{\Delta P}^\top\mathbf{D_r}^{-1}\mathbf{P\Delta} = \mathbf{\Delta}^2 = \mathbf{\Lambda}. \end{aligned} \quad (3)$$

What Does Correspondence Analysis Optimize?

In CA, the criterion that is maximized is the variance of the factor scores (see Lebart et al. 1984; Greenacre 1984). For example, the row first factor $\mathbf{f}_1$ is obtained as a linear combination of the columns of the matrix $(\mathbf{Z} - \mathbf{rc}^\top)$ taking into account the constraints imposed by the matrices $\mathbf{D_r}^{-1}$ and $\mathbf{D_c}^{-1}$. Specifically, this means that we are searching for the vector $\mathbf{q}_1$ containing the weights of the linear combination such as $\mathbf{f}_1$ is obtained as

$$\mathbf{f}_1 = \mathbf{D_r}^{-1}\left(\mathbf{Z} - \mathbf{rc}^\top\right)\mathbf{D_c}^{-1}\mathbf{q}_1, \quad (4)$$

such that

$$\mathbf{f}_1 = \arg\max_{\mathbf{f}} \mathbf{f}^\top\mathbf{D_r}\mathbf{f}, \quad (5)$$

under the constraint that

$$\mathbf{q}_1^\top\mathbf{D_c}^{-1}\mathbf{q}_1 = 1. \quad (6)$$

The subsequent row factor scores will maximize the residual variance under the orthogonality constraint imposed by the matrix $\mathbf{D_r}$ $\left(\text{i.e.,} \mathbf{f}_2^\top\mathbf{D_r}\mathbf{f}_1 = 0\right)$.

How to Identify the Elements Important for a Factor

In CA, the rows and the columns of the table have a similar role (and variance) and therefore we can use the same statistics to identify the rows and the columns important for a given dimension.

Because the variance extracted by a factor (i.e., its eigenvalue) is obtained as the weighted sum of the squared factor scores for this factor of either the rows or columns of the table, the importance of a row (respectively, a column) is reflected by the ratio of its squared factor score to the eigenvalue of this factor. This ratio is called the contribution of the row (respectively, column) to the factor. Specifi-cally, the contributions of row i to component ℓ and of column j to component ℓ are obtained, respectively, as

$$\text{ctr}_{i,\ell} = \frac{r_i f_{i,\ell}^2}{\lambda_\ell} \quad \text{and} \quad \text{ctr}_{j,\ell} = \frac{c_j g_{j,\ell}^2}{\lambda_\ell} \tag{7}$$

(with r_i being the ith element of $\mathbf{r}$ and c_j being the jth element of $\mathbf{c}$). Contributions take values between 0 and 1, and their sums for a given factor are equal to 1 for either the rows or the columns. A convenient rule of thumb is to consider that contributions larger than the average (i.e., $\frac{1}{I}$ for the rows and $\frac{1}{J}$ for the columns) are important for a given factor.

How to Identify the Important Factors for an Element

The factors important for a given row or column are identified by computing statistics called squared cosines. These statistics are obtained by decomposing the squared distance of an element along the factors of the analysis. Specifically, the vector of the squared (χ^2) distance from the rows and columns to their respective barycenter (i.e., average or center of gravity) is obtained as

$$\mathbf{d_r} = \text{diag}\left\{\mathbf{F}\mathbf{F}^\top\right\} \quad \text{and} \quad \mathbf{d_c} = \text{diag}\left\{\mathbf{G}\mathbf{G}^\top\right\}. \tag{8}$$

Recall that the total inertia ($\mathcal{I}$) in CA is equal to the sum of the eigenvalues and that, in CA, this inertia can also be computed as the weighted sum of the squared distances of the rows or the columns to their respective barycenter. Formally, the inertia can be computed as

$$\mathcal{I} = \sum_{l}^{L} \lambda_\ell = \mathbf{r}^\top \mathbf{d_r} = \mathbf{c}^\top \mathbf{d_c}. \tag{9}$$

The squared cosines between row i and component ℓ and column j and component ℓ are obtained, respectively, as

$$\cos^2_{i,\ell} = \frac{f_{i,\ell}^2}{d_{r,i}^2} \quad \text{and} \quad \cos^2_{j,\ell} = \frac{g_{j,\ell}^2}{d_{c,j}^2}. \tag{10}$$

(with $d_{r,i}^2$ and $d_{c,j}^2$ being, respectively, the ith element of $\mathbf{d_r}$ and the jth element of $\mathbf{d_c}$). The sum of the squared cosines over the dimensions for a given element is equal to 1, and so the cosine can be seen as the proportion of the variance of an element that can be attributed to a given dimension.

Correspondence Analysis and the Chi-Square Test

CA is intimately related to the independence χ^2 test. Recall that the (null) hypothesis stating that the rows and the columns of a contingency table $\mathbf{X}$ are independent can be tested by computing the following χ^2 criterion:

$$\chi^2 = N \sum_{i}^{I} \sum_{j}^{J} \frac{\left(z_{i,j} - r_i c_j\right)^2}{r_i c_j} = N\phi^2, \tag{11}$$

where ϕ^2 is called the mean square contingency coefficient (for a 2×2 table, it is called the squared coefficient of correlation associated with the χ^2 test; in this special case it takes values between 0 and 1). For a contingency table, under the null hypothesis, the χ^2 criterion follows a χ^2 distribution with $(I-1)(J-1)$ degrees of freedom and can, therefore, be used (under the usual assumptions) to test the independence hypothesis.

The total inertia of the matrix $(\mathbf{Z} - \mathbf{r}\mathbf{c}^\top)$ (under the constraints imposed on the SVD by the matrices $\mathbf{D}_\mathbf{r}^{-1}$ and $\mathbf{D}_\mathbf{c}^{-1}$) can be computed as the sum of the eigenvalues of the analysis or directly from the data matrix as

$$\begin{aligned}\mathcal{I} &= \text{trace}\left\{\mathbf{D_c}^{-\frac{1}{2}}(\mathbf{Z}-\mathbf{r}\mathbf{c}^\top)^\top\mathbf{D_r}^{-1}(\mathbf{Z}-\mathbf{r}\mathbf{c}^\top)\mathbf{D_c}^{-\frac{1}{2}}\right\}\\ &= \sum_i^I\sum_j^J\frac{\left(z_{i,j}-r_ic_j\right)^2}{r_ic_j} = \phi^2.\end{aligned} \tag{12}$$

This shows that the total inertia is proportional to the independence χ^2 (specifically $\mathcal{I} = N^{-1}\chi^2$) and therefore that the factors of CA perform an orthogonal decomposition of the independence χ^2 where each factor "explains" a portion of the deviation to independence.

The Transition Formula

In CA, rows and columns play a symmetric role and their factor scores have the same variance. As a consequence of this symmetry, the row factor scores (respectively, the column factor scores) can be derived from the column factor scores (respectively, the row factor scores). This can be seen by rewriting Equation 2, taking into account Equations 3 to 6. For example, the factor scores for the rows can be computed as

$$\begin{aligned}\mathbf{F} &= \mathbf{D_r}^{-1}\mathbf{P}\boldsymbol{\Delta}\\ &= \mathbf{D_r}^{-1}(\mathbf{Z}-\mathbf{r}\mathbf{c}^\top)\mathbf{D_c}^{-1}\mathbf{Q}\\ &= \mathbf{D_r}^{-1}(\mathbf{Z}-\mathbf{r}\mathbf{c}^\top)\mathbf{G}\boldsymbol{\Delta}^{-1}\\ &= \mathbf{D_r}^{-1}\mathbf{Z}\mathbf{G}\boldsymbol{\Delta}^{-1}-\mathbf{D_r}^{-1}\mathbf{r}\mathbf{c}^\top\mathbf{G}\boldsymbol{\Delta}^{-1}.\end{aligned} \tag{13}$$

Because the matrix $(\mathbf{Z}-\mathbf{r}\mathbf{c}^\top)$ contains deviations to its row and column barycenters, the row and column sums are equal to 0 and therefore the matrix $\mathbf{D_r}^{-1}\mathbf{r}\mathbf{c}^\top\mathbf{G}\boldsymbol{\Delta}^{-1}$ is equal to $\mathbf{0}$ and Equation 13 can be rewritten as

$$\mathbf{F} = \mathbf{D_r}^{-1}\mathbf{Z}\mathbf{G}\boldsymbol{\Delta}^{-1}. \tag{14}$$

So, if we denote by $\mathbf{R}$ the matrix $\mathbf{R} = \mathbf{D_r}^{-1}\mathbf{Z}$ in which each row (whose sum is 1) is called a *row profile* and $\mathbf{C}$ the matrix $\mathbf{C} = \mathbf{D_c}^{-1}\mathbf{Z}^\top$ in which each column (whose sum is 1) is called a *column profile*, the transition formulas are

$$\mathbf{F} = \mathbf{R}\mathbf{G}\boldsymbol{\Delta}^{-1}\text{ and }\mathbf{G} = \mathbf{C}\mathbf{F}\boldsymbol{\Delta}^{-1}. \tag{15}$$

This shows that the factor scores of an element of one set (e.g., a row) are computed as the barycenter of the expression of this element in the other set (e.g., the columns) followed by an expansion (as expressed by the $\boldsymbol{\Delta}^{-1}$ term) that is inversely proportional to the singular value of each factor.

In CA, the Singular Values Are Never Larger than One

Note that, together, the two transition formulas from Equation 15 imply that the diagonal terms of $\boldsymbol{\Delta}^{-1}$ are larger than one (because otherwise the range of each set of factor scores would be smaller than the other one which, in turn, would imply that all these factor scores are null) and, therefore, that the singular values in CA are always equal to or smaller than one.

Distributional Equivalence

An important property of correspondence analysis is to give the same results when two rows (respectively, two columns) that are proportional are merged together. This property, called *distributional equivalence* (or also "distributional equivalency"), also applies approximately: The analysis is only changed a little when two rows (or columns) that are almost proportional are merged together.

How to Interpret Point Proximity

In a CA map when two row (respectively, column) points are close to each other, this means that these points have similar profiles, and when two points have the same profile, they will be located exactly at the same place (this is a consequence of the distributional equivalence principle). The proximity between row and column points is more delicate to interpret because of the barycentric principle (see section on "The Transition Formula"): The position of a row (respectively, column) point is determined from its barycenter on the column (respectively, row), and therefore, the proximity between a row point and one column point cannot be interpreted directly.

Asymmetric Plot: How to Interpret Row and Column Proximity

CA treats rows and columns symmetrically, and so their roles are equivalent. In some cases, however, rows and columns can play different roles, and this symmetry can be misleading. As an illustration, in the example used below (see section "Example: The Colors of Sounds"), the participants were asked to choose the color that would match a given piece of music. In this framework, the colors can be considered as a dependent variable and the pieces of music as an independent variable. In this case the roles are asymmetric and the plots can reflect this asymmetry by normalizing one set such that the variance of its factor scores is equal to 1 for each factor. For example, the normalized to 1 column factor scores, denoted $\tilde{\mathbf{G}}$ would be computed as (compare with Equation 2)

$$\tilde{\mathbf{G}} = \mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q}. \tag{16}$$

In the asymmetric plot obtained with $\mathbf{F}$ and $\tilde{\mathbf{G}}$, the distances between rows and columns can now be interpreted meaningfully: The distance from a row point to a column point reflects their association (and a row is positioned exactly at the barycenter of the columns).

Centered Versus Noncentered Analysis

CA is obtained from the generalized singular value decomposition of the centered matrix $(\mathbf{Z} - \mathbf{r}\mathbf{c}^{\top})$, but it could be also obtained from the same singular value decomposition of matrix $\mathbf{Z}$. In this case, the first pair of singular vectors is equal to $\mathbf{r}$ and $\mathbf{c}$ and their associated singular value is equal to 1. This property is easily verified from the following relations:

$$\begin{aligned} \mathbf{c}\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Z} &= \mathbf{1}\mathbf{Z} = \mathbf{r} \text{ and} \\ \mathbf{r}\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{Z}^{\top} &= \mathbf{1}\mathbf{Z}^{\top} = \mathbf{c}. \end{aligned} \tag{17}$$

Because in CA, the singular values are never larger than one, $\mathbf{r}$ and $\mathbf{c}$ having a singular value of 1 are the first pair of singular vectors of $\mathbf{Z}$. Therefore, the generalized singular value decomposition of $\mathbf{Z}$ can be developed as

$$\mathbf{Z} = \mathbf{r}\mathbf{c}^{\top} + \left(\mathbf{Z} - \mathbf{r}\mathbf{c}^{\top}\right) = \mathbf{r}\mathbf{c}^{\top} + \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\top}. \tag{18}$$

This shows that the ℓth pair of singular vectors and singular value of $(\mathbf{Z} - \mathbf{r}\mathbf{c}^{\top})$ are the $(\ell + 1)$th pair of singular vectors and singular value of $\mathbf{Z}$.

Supplementary Elements

Often in CA we want to know the position in the analysis of rows or columns that were not actually analyzed. These rows or columns are called *illustrative, supplementary*, or *out of sample* rows or columns (or supplementary observations or variables). In contrast with the appellation of supplementary (which are not used to compute the factors), the active elements are used to compute the factors.

The projection formula uses the transition formula (see Equation 15) and is specific to correspondence analysis. Specifically, let $\mathbf{i}_{\text{sup}}^{\top}$ be an illustrative row and $\mathbf{j}_{\text{sup}}$ be an illustrative column to be projected (note that in CA, prior to projection, an illustrative row or column is rescaled such that its sum is equal to 1). The coordinates of the illustrative rows (denoted $\mathbf{f}_{\text{sup}}$) and columns (denoted $\mathbf{g}_{\text{sup}}$) are obtained as

$$\begin{aligned} \mathbf{f}_{\text{sup}} &= \left(\mathbf{i}_{\text{sup}}^{\top}1\right)^{-1}\mathbf{i}_{\text{sup}}^{\top}\mathbf{G}\tilde{\boldsymbol{\Delta}}^{-1} \text{ and} \\ \mathbf{g}_{\text{sup}} &= \left(\mathbf{j}_{\text{sup}}^{\top}1\right)^{-1}\mathbf{j}_{\text{sup}}^{\top}\mathbf{F}\tilde{\boldsymbol{\Delta}}^{-1} \end{aligned} \tag{19}$$

(Note that the scalar terms $\left(\mathbf{i}_{\text{sup}}^{\top}\mathbf{1}\right)^{-1}$ and $\left(\mathbf{j}_{\text{sup}}^{\top}\mathbf{1}\right)^{-1}$ are used to ensure that the sum of the elements of $\mathbf{i}_{\text{sup}}$ or $\mathbf{j}_{\text{sup}}$ is equal to 1; if this is already the case, these terms are superfluous).

Programs and Packages

CA is implemented in most statistical packages (e.g., SAS, XLSTAT, SYSTAT) with R giving the most comprehensive implementation. Several packages in R are specifically dedicated to correspondence analysis and its variants. The most

popular are the packages CA, FactoMineR (Husson et al. 2011), ade4, vegan and ExPosition (Beaton et al. 2014, this last package was used to analyze the example presented below). MATLAB programs are also available for download from www.utdallas.edu/~herve.

Example: The Colors of Sounds

To illustrate CA we have collected data (as part of a science project!) from 22 participants who were presented with 9 "pieces of music" and asked to associate 1 of 10 colors to each piece of music. The pieces of music were (1) the music of a video game (video), (2) a jazz song (jazz) (3) a country and Western song (country), (4) a rap song (rap), (5) a pop song (pop), (6) an extract of the opera Carmen (opera), (7) the low *F* note played on a piano, (8) the middle *F* note played on the same piano, and, finally, (9) the high *F* note still played on the same piano. The data are shown in Table 1 where the columns are the pieces of music, the rows are the colors, and the numbers at the intersection of the rows and the columns give the number of times the color in the row was associated with the piece of music in the column. A graphics representation of the data from Table 1 is given by the "heat map" displayed in Fig. 1.

A CA of Table 1 extracted eight components. We will, here, only consider the first two components which together account for 65% of the total inertia (with eigenvalues of 0.287 and 0.192, respectively). The factor scores of the observations (rows) and variables (columns) are shown in Tables 2 and 3, respectively. The corresponding map is displayed in Fig. 2.

We can see from Figs. 2 (symmetric plot) and 3 (asymmetric plot) that the first component isolates the color black from all the other colors and that black is mostly associated with two pieces of music: rap and the low *F*. The association with the low note reflects a standard association between pitch and color (high notes are perceived as bright and low notes as dark); by contrast, the association of rap music with the black color is likely to reflect a semantic association. The squared cosines show that the first component accounts for most of the variance of the black color (93%, see Table 2). The second component separates the colors brown and (to a lesser extent) green from the other colors (in particular purple) and that brown and green as associated with Pop and Middle *F*. On the other side of the second dimension, we find the color purple and a quartet of pieces of music (Video, High *F*, Opera, and Jazz).

Key Applications

CA, being a very versatile method, has been applied to many different domains such as sensory evaluation, brain imaging, genetic, sociology, psychology, ecology, graph theory, etc. (see Beh and Lombardo 2015, for a recent review of applications).

Future Directions

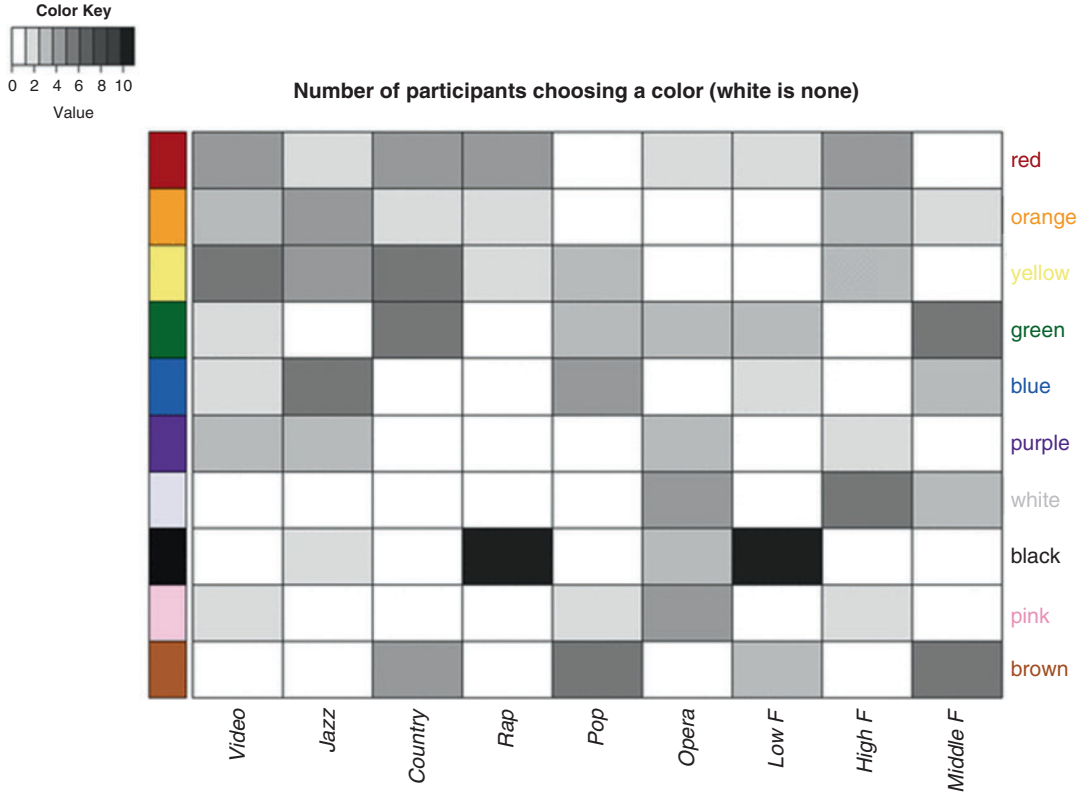
CA is still mostly a descriptive method, future directions of research are likely to integrate computationally based cross-validation methods to identify significant dimensions and significant contributions for rows and columns (see, e.g., Beaton et al. 2016). Another likely development would be to develop adapted sparsification methods (such as the LASSO, see, e.g., Allen et al. 2014). Also, the development of multiblock and multi-table extensions of CA will make CA even more adapted to the analysis of large heterogonous data sets.

Conclusion

CA is a very versatile and popular technique in multivariate analysis. In addition to the basics presented here, CA includes numerous variants such as multiple correspondence analysis (to analyze several nominal variables, see Abdi

Correspondence Analysis, Table 1 Twenty-two participants associated one of ten colors to nine pieces of music. The column labeled $\mathbf{x}_{i+}$ gives the total number of choices made for each color. N is the grand total of the data table. The vector of mass for the rows, r, is the proportion of choices made for each color ($r_i = x_{i+}/N$). The row labeled x_{+j} gives the total number of times each piece of music was presented (i.e., it is equal to the number of participants). The centroid row, $\mathbf{c}^{\mathrm{T}}$, gives these values as proportions ($c_j = x_{+j}/N$)

Color	Video	Jazz	Country	Rap	Pop	Opera	Low F	High F	Middle F	x_{i+}	r
Red	4	2	4	4	1	2	2	4	1	24	0.121
Orange	3	4	2	2	1	1	0	3	2	18	0.091
Yellow	6	4	5	2	3	1	1	3	0	25	0.126
Green	2	0	5	1	3	3	3	1	5	23	0.116
Blue	2	5	0	1	4	1	2	1	3	19	0.096
Purple	3	3	1	0	0	3	0	2	1	13	0.066
White	0	0	0	0	1	4	1	5	3	14	0.071
Black	0	2	0	11	1	3	10	1	1	29	0.146
Pink	2	1	1	0	2	4	0	2	0	12	0.061
Brown	0	1	4	1	6	0	3	0	6	21	0.106
x_{+j}	22	22	22	22	22	22	22	22	22	$N = 198$	1.000
$\mathbf{c}^{\mathrm{T}}$	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11		



Correspondence Analysis, Fig. 1 CA The Colors of Music. A heat map of the data from Table 1

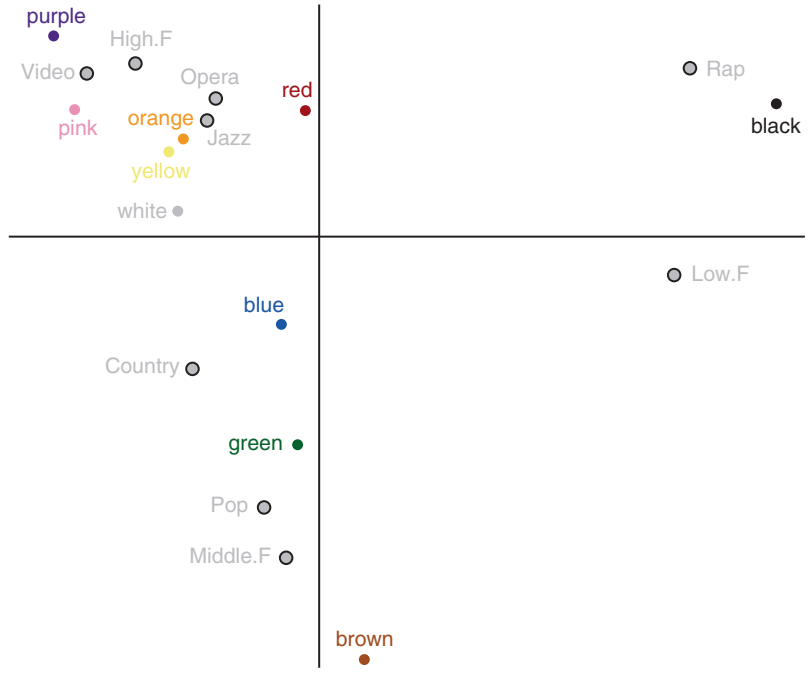
Correspondence Analysis, Table 2 CA The Color of Music. Factor scores, contributions, mass, mass × squared factor scores, inertia to barycenter, and squared cosines for the rows. For convenience, squared cosines and contributions have been multiplied by 1000 and rounded

	F_1	F_2	Ctr_1	Ctr_2	r_i	$(r_i \times F_1^2)$	$(r_i \times F_2^2)$	$\left(r_i \times d_{r,i}^2\right)$	$(\cos_1^2)$	$(\cos_2^2)$
Red	−0.026	0.299	0	56	0.121	0.000	0.011	0.026	3	410
Orange	−0.314	0.232	31	25	0.091	0.009	0.005	0.030	295	161
Yellow	−0.348	0.202	53	27	0.126	0.015	0.005	0.057	267	89
Green	−0.044	−0.490	1	144	0.116	0.000	0.028	0.048	5	583
Blue	−0.082	−0.206	2	21	0.096	0.001	0.004	0.050	13	81
Purple	−0.619	0.475	87	77	0.066	0.025	0.015	0.050	505	298
White	−0.328	0.057	26	1	0.071	0.008	0.000	0.099	77	2
Black	1.195	0.315	726	75	0.146	0.208	0.014	0.224	929	65
Pink	−0.570	0.300	68	28	0.061	0.020	0.005	0.053	371	103
Brown	0.113	−0.997	5	545	0.106	0.001	0.105	0.108	12	973
$\sum$	–	–	1000	1000	–	0.287	0.192	0.746		
						λ_1	λ_2	$\mathcal{I}$		
						39%	26%			
						τ_1	τ_2			

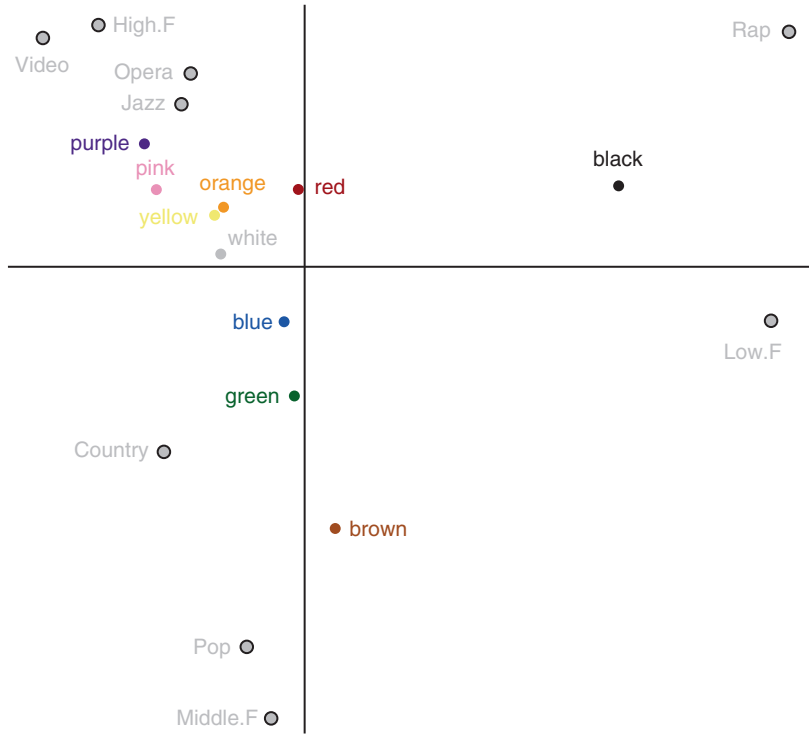
Correspondence Analysis, Table 3 CA The Colors of Music. Factor scores, contributions, mass, mass × squared factor scores, inertia to barycenter, and squared cosines for the columns. For convenience, squared cosines and contributions have been multiplied by 1000 and rounded

	G_1	G_2	$\tilde{G}_1$	$\tilde{G}_2$	ctr_1	ctr_2	c_j	$(c_j \times G_1^2)$	$(c_j \times G_2^2)$	$\left(d_{c,j}^2\right)$	$(\cos_1^2)$	$(\cos_2^2)$
Video	−0.541	0.386	−1.007	0.879	113	86	0.111	0.032	0.017	0.071	454	232
Jazz	−0.257	0.275	−0.478	0.626	25	44	0.111	0.007	0.008	0.069	105	121
Country	−0.291	−0.309	−0.541	−0.704	33	55	0.111	0.009	0.011	0.066	142	161
Rap	0.991	0.397	1.846	0.903	379	91	0.111	0.109	0.017	0.133	822	132
Pop	−0.122	−0.637	−0.227	−1.450	6	234	0.111	0.002	0.045	0.064	26	709
Opera	−0.236	0.326	−0.440	0.742	22	61	0.111	0.006	0.012	0.079	78	149
Low F	0.954	−0.089	1.777	−0.203	351	5	0.111	0.101	0.001	0.105	962	8
High F	−0.427	0.408	−0.795	0.929	70	96	0.111	0.020	0.018	0.074	271	249
Middle F	−0.072	−0.757	−0.134	−1.723	2	330	0.111	0.001	0.064	0.084	7	759
$\sum$	–	–	–	–	1000	1000	–	0.287	0.192	0.746		
								λ_1	λ_2	$\mathcal{I}$		
								39%	26%			
								τ_1	τ_2			

and Valentin 2007, Lebart et al. 1984, and Greenacre 2007), discriminant correspondence analysis (to assign observation to a priori defined groups, see, e.g., Nakache et al. 1977; Saporta and Niang 2006; Abdi 2007d), and multi-block correspondence analysis (when the variables are structured into blocks, see Abdi and Valentin 2007, Lebart et al. 1984, Greenacre 2007, Abdi et al. 2013, Beaton et al. 2016, and Williams et al. 2010).



Correspondence Analysis, Fig. 2 CA The Colors of Music. Symmetric plot: The projections of the rows and the columns are displayed in the same map. $\lambda_1 = 0.287$, $\tau_1 = 39$; $\lambda_2 = 0.192$, $\tau_2 = 26$. In this plot the proximity between rows and columns cannot be directly interpreted



Correspondence Analysis, Fig. 3 CA The Colors of Music. Asymmetric plot: The projections of the rows and the columns are displayed in the same map. The inertia of the projections of the column factor scores is equal to 1 for each dimension and the inertia of the projections of the row factor scores are $\lambda_1 = 0.287$, $\tau_1 = 39$; $\lambda_2 = 0.192$, $\tau_2 = 26$. In this plot the proximity between rows and columns can be directly interpreted

Cross-References

- Clustering Algorithms
- Data Mining
- Distance and Similarity Measures
- Eigenvalues, Singular Value Decomposition
- Matrix Algebra, Basics of
- Matrix Decomposition
- Network Analysis in French Sociology and Anthropology
- Network Models
- Principal Component Analysis
- Probability Matrices
- Similarity Metrics on Social Networks

References

Abdi H (1988) A generalized approach for connectionist auto-associative memories: interpretation, implications and illustration for face processing. In: Demongeot J (ed) Artificial intelligence and cognitive sciences. Manchester University Press, Manchester, pp 149–164

Abdi H (2003) Multivariate analysis. In: Lewis-Beck M, Bryman A, Futing T (eds) Encyclopedia for research methods for the social sciences. Sage, Thousand Oaks, pp 699–702

Abdi H (2007a) Singular value decomposition (SVD) and generalized singular value decomposition(GSVD). In: Salkind NJ (ed) Encyclopedia of measurement and statistics. Sage, Thousand Oaks, pp 907–912

Abdi H (2007b) Eigen-decomposition: eigenvalues and eigenvectors. In: Salkind NJ (ed) Encyclopedia of measurement and statistics. Sage, Thousand Oaks, pp 304–308

Abdi H (2007c) Discriminant correspondence analysis. In: Salkind NJ (ed) Encyclopedia of measurement and statistics. Sage, Thousand Oaks, pp 270–275

Abdi H (2007d) Metric multidimensional scaling. In: Salkind NJ (ed) Encyclopedia of measurement and statistics. Sage, Thousand Oaks, pp 598–605

Abdi H (2007e) Z-scores. In: Salkind NJ (ed) Encyclopedia of measurement and statistics. Sage, Thousand Oaks, pp 1057–1058

Abdi H, Valentin D (2007) Multiple correspondence analysis. In: Salkind NJ (ed) Encyclopedia of measurement and statistics. Sage, Thousand Oaks, pp 651–657

Abdi H, Williams LJ (2010a) Principal component analysis. Wiley Interdiscip Rev Comput Stat 2:433–459

Abdi H, Williams LJ (2010b) Correspondence analysis. In: Salkind NJ (ed) Encyclopedia of research design. Sage, Thousand Oaks

Abdi H, Valentin D, O'Toole J (1997) A generalized auto-associator model for face processing and sex categorization: from principal components to multivariate analysis. In: Levine D (ed) Optimality in biological and artificial networks. Lawrence Erlbaum, Mahwah, pp 317–337

Abdi H, Williams LJ, Valentin D (2013) Multiple factor analysis: principal component analysis for multi-table and multi-block data sets. Wiley Interdiscip Rev Comput Stat 5:149–179

Allen G, Grosencik L, Taylor J (2014) A generalized least squares matrix decomposition. J American Stat Assoc, Theory Methods 109:145–159

Beaton B, Chin Fatt CR, Abdi H (2014) An ExPosition of multivariate analysis with the Singular Value Decomposition in R. Computational Statistics & Data Analysis 72:176–189

Beaton D, Dunlop J, ADNI Abdi H (2016) Partial least squares-correspondence analysis: a framework to simultaneously analyze behavioral and genetic data. Psychol Methods 21:621–651

Beh EJ, Lombardo R (2015) Correspondence analysis: theory, practice and new strategies. Wiley, London

Benzécri J-P (1973) L'analyse des données, vol 1, 2. Dunod, Paris

Cordier B (1965) L'analyse des correspondances. PhD dissertation, University of Rennes

Eckart C, Young G (1936) The approximation of a matrix by another of a lower rank. Psychometrika 1:211–218

Escofier B (1969) L'analyse factorielle des correspondances. Cahiers du Bureau Universitaire de Recherche Opérationnelle 13:25–59

Escofier B, Pagès J (1990) Analyses factorielles simples et multiples: objectifs, méthodes, interprétation. Dunod, Paris

Escoufier Y (2007) Operators related to a data matrix: a survey. In: COMPSTAT: 17th symposium proceedings in computational statistics, Rome, Italy, 2006. Physica Verlag, New York, pp 285–297

Good I (1969) Some applications of the singular value decomposition of a matrix. Technometrics 11:823–831

Greenacre MJ (1984) Theory and applications of correspondence analysis. Academic, London

Greenacre MJ (2007) Correspondence analysis in practice, 2nd edn. Chapman & Hall/CRC, Boca Raton

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 25:417–441

Husson F, Lê S, Pagès J (2011) Exploratory multivariate analysis by example using R. Chapman & Hall/CRC, Boca Raton

Hwang H, Tomiuk MA, Takane Y (2010) Correspondence analysis, multiple correspondence analysis and recent developments. In: Millsap R, Maydeu-Olivares A (eds) Handbook of quantitative methods in psychology. Sage, London

Lebart L, Fénelon JP (1975) Statistique et informatique appliquées. Dunod, Paris

Lebart L, Saporta G (2014) Historical elements of correspondence analysis and multiple correspondence analysis. In: Blasius J, Greenacre M (eds) Visualization and verbalization of data. CRC Press, Boca Raton

Lebart L, Morineau A, Warwick KM (1984) Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. Wiley, London

Nakache JP, Lorente P, Benzécri JP, Chastang JF (1977) Aspect pronostics et thérapeutiques de l'infarctus myocardique aigu. Les Cahiers de l'Analyse des Données 2:415–534

Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 6:559–572

Pearson K (1906) On certain points connected with scale order in the case of a correlation of two characters which for some arrangement give a linear regression line. Biometrika 13:25–45

Saporta G, Niang N (2006) Correspondence analysis and classification. In: Greenacre M, Blasius J (eds) Multiple correspondence analysis and related methods. Chapman & Hall, Boca Raton, pp 371–392

Stewart GW (1993) On the early history of the singular value decomposition. SIAM Rev 35:551–566

Takane Y (2002) Relationships among various kinds of eigenvalue and singular value decompositions. In: Yanai H, Okada A, Shigemasu K, Kano Y, Meulman J (eds) New developments in psychometrics. Springer, Tokyo, pp 45–56

Williams LJ, Abdi H, French R, Orange JB (2010) A tutorial on multi-block discriminant correspondence analysis (MUDICA): a new method for analyzing discourse data from clinical populations. J Speech Lang Hear Res 53:1372–1393