Journal of Statistical Software

July 2019, Volume 90, Issue 4. doi: 10.18637/jss.v090.i04

MultipleCar: A Graphical User Interface MATLAB Toolbox to Compute Multiple Correspondence Analysis

Urbano Lorenzo-Seva
Universitat Rovira i Virgili

Michel van de Velden
Erasmus University

Abstract

In this paper we present the toolbox **MultipleCar**, which is a general program for computing multiple correspondence analysis and which was designed using a graphical user interface. The procedures implemented in **MultipleCar** are the usual ones that are already available in other applications, plus some additional procedures. **MultipleCar** makes it possible to compute (1) joint correspondence analysis, and (2) orthogonal and oblique rotation of coordinates. Although **MultipleCar** was developed in MATLAB, we compiled it as a standalone application for Windows operative systems based on graphical user interfaces. The users can decide whether to use the advanced MATLAB version of **MultipleCar**, or the standalone version (which does not require any programming skills).

Keywords: multiple correspondence analysis, joint correspondence analysis, orthogonal rotation, oblique rotation, exploratory analysis, simple structure, qualitative methods, MATLAB.

1. Introduction

In this paper we present the program **MultipleCar**, which was designed using a graphical user interface. It is a general program for computing multiple correspondence analysis. **MultipleCar** uses traditional procedures and indices as well as more recent developments not included in commercial or free shared packages.

Correspondence analysis (CA) is a popular method that displays the associations between two categorical variables. The method was developed in France by Benzécri in 1973 but only gained popularity outside France after textbooks were published in English 10 years later (Greenacre 1984; and Lebart, Morineau, and Warwick 1990). Multiple correspondence analysis (MCA) can be seen as an extension of correspondence analysis that simultaneously analyzes more than two categorical variables. As well as displaying the associations between two categorical variables, MCA makes it possible to study (bivariate) relationships between

several categorical variables and display the relationships between observations.

Mathematically, MCA and its variations can be defined in several ways. Although the differences between formulations are small and relatively straightforward from a mathematical point of view, practitioners may get confused when interpreting different MCA outcomes, or deciding on how to analyze data themselves. In fact, there appears to be some ambiguity surrounding the formulation of MCA. We briefly introduce CA and MCA and some of its variants.

2. Correspondence analysis (CA)

Correspondence analysis is a well-known, established method. Excellent descriptions can be found in Greenacre (1984), and Lebart *et al.* (1990). Here we give a concise overview of some of the most important mathematical relationships that are useful for understanding the software presented and its options.

Using Greenacre's notation, we can summarize CA as follows. Let $\mathbf{P}$ denote an $n_r \times n_c$ data matrix with nonnegative elements that sum to 1. That is, $\mathbf{1}_{n_r}^\top \mathbf{P} \mathbf{1}_{n_c} = 1$, where, generically, $\mathbf{1}_q$ denotes a q dimensional vector of ones. Correspondence analysis amounts to the following least-squares approximation problem:

$$\min_{\mathbf{R},\mathbf{C}} \left\| \tilde{\mathbf{P}} - \mathbf{D}_r^{1/2} \mathbf{R} \mathbf{C}^\top \mathbf{D}_c^{1/2} \right\|^2, \tag{1}$$

subject to

$$\mathbf{C}^\top \mathbf{D}_c \mathbf{C} = \mathbf{I}_k,$$

where $\tilde{\mathbf{P}} = \mathbf{D}_r^{-1/2} \left(\mathbf{P} - \mathbf{r}\mathbf{c}^\top \right) \mathbf{D}_c^{-1/2}, \mathbf{r} = \mathbf{P}\mathbf{1}_{q_c}$, $\mathbf{c} = \mathbf{P}^\top \mathbf{1}_{q_r}$, $\mathbf{D}_r$ and $\mathbf{D}_c$ are corresponding diagonal matrices (i.e., $\mathbf{D}_r \mathbf{1}_{n_r} = \mathbf{r}$ and $\mathbf{D}_c \mathbf{1}_{n_c} = \mathbf{c}$). The so-called row and column coordinate matrices $\mathbf{R}$ and $\mathbf{C}$ are of rank k, where k is the dimensionality of the approximation. This rank must be chosen by the user. Often $k = 2$ is chosen as this allows for immediate graphical displays of the data.

This least-squares problem can be solved by using singular value decomposition

$$\tilde{\mathbf{P}} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^\top, \tag{2}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal and $\mathbf{\Lambda}$ is a diagonal matrix with the singular values on its diagonal in descending order. By selecting only the first k columns of $\mathbf{U}$ and $\mathbf{V}$ and the corresponding singular values, a k-dimensional least-squares approximation of $\tilde{\mathbf{P}}$ is obtained. The resulting coordinate matrices are

$$\mathbf{R} = \mathbf{D}_r^{-1/2} \mathbf{U} \mathbf{\Lambda}^\alpha \quad \text{and} \quad \mathbf{C} = \mathbf{D}_c^{-1/2} \mathbf{V} \mathbf{\Lambda}^{1-\alpha}, \tag{3}$$

so that

$$\mathbf{R}^\top \mathbf{D}_r \mathbf{R} = \mathbf{\Lambda}^{2\alpha} \quad \text{and} \quad \mathbf{C}^\top \mathbf{D}_c \mathbf{C} = \mathbf{\Lambda}^{(2-2\alpha)}. \tag{4}$$

Depending on the choice of α, the row (column) coordinates are referred to as principal (standard) coordinates if $\alpha = 1$, standard (principal) coordinates if $\alpha = 0$ or symmetrical coordinates if $\alpha = 1/2$. Furthermore, the quality of the k-dimensional approximation can be assessed by considering the so-called inertias. That is,

$$\sum_{i=1}^{k} \lambda_i^2 / \sum_{i=1}^{\kappa} \lambda_i^2, \tag{5}$$

where κ denotes the rank of $\mathbf{P}$. The row and column coordinates are closely related through so-called transition formulae. This implies that rather than using (3) to separately construct the coordinate matrices, one set of coordinates (or singular vectors) can be used to obtain the other set. For example,

$$\mathbf{R} = \mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{V}. \tag{6}$$

The set of coordinates as defined in (3) constitutes a so-called biplot as the inner-product $\mathbf{D}_r^{1/2}\mathbf{R}\mathbf{C}^\top\mathbf{D}_c^{1/2}$ approximates the data.

Note that the choice of α in (3) influences the interpretation of the sets of points in a biplot. In particular, distances between the principal coordinate points are (approximated) chi-squared distances. On the other hand, the standard coordinate points, which are scaled to have weighted squared length equal to one, merely indicate directions. The biplot relationship ensures that the principal coordinate points can be projected onto the directions indicated by the standard coordinates to retrieve the approximated data values (i.e., the low-dimensional approximation of $\tilde{\mathbf{P}}$). In a symmetric biplot, the projections can be used in a similar fashion. See Greenacre (1993) or Van de Velden and Kiers (2005) for details on the relationship of correspondence analysis and biplots.

In exploratory factor analysis (EFA) coordinates are usually inspected to explain the meaning of the k dimensions, and the best possible solution is the one which is easiest to interpret. In order to maximize the simplicity of coordinates, they are usually (orthogonally or obliquely) rotated. In this biplot scenario, row and column points can both be rotated without influencing the approximation. To see this, let $\mathbf{T}$ denote a rotation matrix for which $\mathbf{T}^{-1}$ exists. Post-multiplication of the row coordinates $\mathbf{R}$ by $\mathbf{T}$ and $\mathbf{C}$ by $\mathbf{T}^{-1}$ does not change the approximation. This rotational freedom was exploited by Van de Velden and Kiers (2005) and Lorenzo-Seva, Van de Velden, and Kiers (2009) to increase interpretability of the solutions using orthogonal and oblique rotations, respectively.

Before rotations are computed, it is not uncommon in the context of EFA for the loading matrices to be weighted. After rotation, the original distances of points to the origin are then re-established. In the context of CA, weighting can also be used. In fact, due to the specific weighting of coordinates in CA, infrequently observed points may be positioned relatively far from the origin. Consequently, these points may have a significant impact on the rotation angle. Pre-multiplying the coordinate matrices by $\mathbf{D}_r$ (for the row points) and $\mathbf{D}_c$ (for the column points) before rotation, removes this effect. Alternatively, as is common practice in EFA, a row-wise normalization can be used. See Lorenzo-Seva *et al.* (2009) for details on these alternative options for weighting CA coordinates for rotation.

Note that, rather than calculating the singular value decomposition (2), the solution can also be found by using

$$\tilde{\mathbf{P}}^\top\tilde{\mathbf{P}} = \mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^\top\right)\mathbf{D}_c^{-1}\left(\mathbf{P} - \mathbf{r}\mathbf{c}\right)\mathbf{D}_r^{-1/2} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^\top. \tag{7}$$

The column coordinates can then be obtained directly from the second formula in (3) whereas the solution for the rows can be obtained by applying (6). This procedure can be useful when the number of rows (columns) of the original data matrix is much larger than the number of columns (rows).

3. Multiple correspondence analysis (MCA)

Multiple correspondence analysis (MCA) is a method that allows the researcher to analyze data on more than two categorical variables. The name may be a little misleading because it suggests that the method is not the same as CA. In fact, MCA is CA applied to a so-called indicator matrix. That is, the categorical data are coded by constructing so-called indicator matrices. For the jth categorical variable we define $\mathbf{Z}_j$ to be the $n \times p_j$ indicator matrix where n denotes the number of observations, p_j the number of categories for variable j and the ijth element of $\mathbf{Z}_j$ is one if individual i selected category j. All other elements are zero. We can construct these indicator matrices for all categorical variables and collect them in a so-called super-indicator matrix $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_p)$. For each categorical variable, an observation corresponds to exactly one category. Hence,

$$\mathbf{Z}\mathbf{1}_{n_c} = p\mathbf{1}_n,$$

and

$$\mathbf{1}_n^\top \mathbf{Z}\mathbf{1}_{n_c} = np.$$

Let

$$\mathbf{z} = \mathbf{Z}^\top \mathbf{1}_n = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_p)^\top,$$

and we define a $n_c \times n_c$ diagonal matrix $\mathbf{D}_z$ satisfying

$$\mathbf{D}_z \mathbf{1}_{n_c} = \mathbf{z}.$$

Inserting $\mathbf{Z}$ for $\mathbf{P}$ in the CA equations of the previous section, we get

$$\tilde{\mathbf{P}}^\top \tilde{\mathbf{P}} = \frac{1}{np}\mathbf{D}_z^{-\frac{1}{2}} \left(\mathbf{Z} - \frac{1}{n}\mathbf{1}\mathbf{z}^\top\right)^\top \left(\mathbf{Z} - \frac{1}{n}\mathbf{1}\mathbf{z}^\top\right) \mathbf{D}_z^{-\frac{1}{2}} = \frac{1}{np}\mathbf{D}_z^{-\frac{1}{2}} \left(\mathbf{Z}^\top\mathbf{Z} - \frac{1}{n}\mathbf{z}\mathbf{z}^\top\right) \mathbf{D}_z^{-\frac{1}{2}}.$$

Alternatively, we can write

$$\tilde{\mathbf{P}}^\top \tilde{\mathbf{P}} = \frac{1}{np}\mathbf{D}_z^{-\frac{1}{2}} \mathbf{Z}^\top \mathbf{M}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}}, \tag{8}$$

where $\mathbf{M}$ denotes the centering matrix. Hence, using (7) to obtain the MCA solution we get

$$\frac{1}{np}\mathbf{D}_z^{-\frac{1}{2}} \mathbf{Z}^\top \mathbf{M}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}} = \mathbf{V}\boldsymbol{\Lambda}^2\mathbf{V}^\top, \tag{9}$$

and it is clear that MCA is closely related to PCA as it amounts to finding the eigendecomposition of the (normalized) covariance matrix.

Although MCA is defined as CA applied to a super-indicator matrix, (9) shows that the calculations can be based directly on the so-called Burt matrix: $\mathbf{B} = \mathbf{Z}^\top\mathbf{Z}$. The Burt matrix is a square symmetric matrix consisting of all cross-tabulations for all combinations of the variables. That is, the Burt matrix contains counts of co-occurrences for all combinations of categories. The diagonal blocks, $\mathbf{Z}_j^\top\mathbf{Z}_j$ (for $j = 1$ to n), of $\mathbf{B}$ contain the marginal frequencies (i.e., counts for each category) for all variables.

Considering the eigenvalue decomposition of the (appropriately scaled) Burt matrix rather than the super-indicator matrix, is much more efficient. In particular when the number of

observations, n, is large. Moreover, although coordinates for the observations do not follow from the eigendecomposition of the Burt matrix, it is straightforward to calculate these coordinates using the transition formula (6). Note, however, that in the analysis of $\mathbf{B}$ the squared singular values are obtained. This poses some issues concerning the scaling of coordinates as well as the calculation of explained inertia. For the calculation of the explained inertia (cf. (5)) as well as for the appropriately scaled coordinates (i.e., the coordinates satisfying (4)) this needs to be taken into account.

As Chavent, Kuentz-Simonet, and Saracco (2012) pointed out, despite the close relationship between EFA and MCA, rotation in MCA has not received much attention. In the general context of rotation in PCAMIX, a principal component method for the mixture of qualitative and quantitative variables, Kiers (1991) proposed orthogonal rotation applied to MCA. An application using a real data set that illustrates the advantages of using orthogonal rotation in MCA can be found in Chavent *et al.* (2012). In addition, Adachi (2004) studied the applicability of oblique rotations in multiple correspondence analysis.

4. Joint correspondence analysis (JCA)

The analysis of the Burt matrix reveals an important property of MCA: The cross-tabulations between all variables are simultaneously approximated (in a least-squares sense). Using the definitions for $\mathbf{Z}$ and $\mathbf{z}$, the left hand side of (9) becomes

$$\mathbf{S} = \frac{1}{np}\mathbf{D}_z^{-\frac{1}{2}} \begin{pmatrix} \mathbf{Z}_1^\top\mathbf{Z}_1 - \frac{1}{n}\mathbf{z}_1\mathbf{z}_1^\top & \mathbf{Z}_1^\top\mathbf{Z}_2 - \frac{1}{n}\mathbf{z}_1\mathbf{z}_2^\top & \cdots & \mathbf{Z}_1^\top\mathbf{Z}_p - \frac{1}{n}\mathbf{z}_1\mathbf{z}_p^\top \\ \mathbf{Z}_2^\top\mathbf{Z}_1 - \frac{1}{n}\mathbf{z}_2\mathbf{z}_1^\top & \mathbf{Z}_2^\top\mathbf{Z}_2 - \frac{1}{n}\mathbf{z}_2\mathbf{z}_2^\top & \cdots & \mathbf{Z}_2^\top\mathbf{Z}_p - \frac{1}{n}\mathbf{z}_2\mathbf{z}_p^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}_p^\top\mathbf{Z}_1 - \frac{1}{n}\mathbf{z}_p\mathbf{z}_1^\top & \mathbf{Z}_p^\top\mathbf{Z}_2 - \frac{1}{n}\mathbf{z}_p\mathbf{z}_2^\top & \cdots & \mathbf{Z}_p^\top\mathbf{Z}_p - \frac{1}{n}\mathbf{z}_p\mathbf{z}_p^\top \end{pmatrix} \mathbf{D}_z^{-\frac{1}{2}}. \tag{10}$$

In MCA, this matrix is approximated. However, the approximation of the diagonal blocks, may not be of interest. To remedy this, Greenacre (1988) proposed joint correspondence analysis (JCA). In JCA, coordinates are obtained for the categories of the categorical variables in such a way that the off-diagonal blocks of the Burt matrix (i.e., the cross-tabulations for all pairs of variables) are approximated in a least-squares sense whilst the approximation of the diagonal blocks is ignored.

The matrix $\mathbf{S}$ can be decomposed as: $\mathbf{S} = \mathbf{S}_u + \mathbf{S}_d$ where $\mathbf{S}_d$ is an $n_c \times n_c$ block diagonal matrix with the $p_j \times p_j$ diagonal blocks of $\mathbf{S}$ as diagonal blocks. Then, to find this least-squares solution, the following five-step algorithm can be used:

(0) Select an initial block diagonal matrix $\mathbf{\Phi}_d^t$, of the same order as $\mathbf{S}_d$. For example,

$$\mathbf{\Phi}_d^t = \left(\mathbf{W}\mathbf{W}^\top\right)_d,$$

where $\mathbf{W}$ is a $q \times k$ matrix whose columns are eigenvectors of $\mathbf{S}$ associated with the k largest eigenvalues and whose squared length is equal to the associated eigenvalue.

(1) Compute $\mathbf{S}^t = \mathbf{S}_u + \mathbf{\Phi}_d^t$.

(2) Compute the eigenvalues and associated eigenvectors of $\mathbf{S}^t$ and determine $\mathbf{A}$ and $\mathbf{\Gamma}$ in such a way that the columns of $\mathbf{A}$ are appropriately scaled eigenvectors of $\mathbf{S}^t$ associated with the k largest eigenvalues of $\mathbf{S}^t$.

(3) In Step (0), replace $\mathbf{\Phi}_d^t$ by $\left(\mathbf{A}\mathbf{A}^\top\right)_d$.

(4) Iterate (1)–(3) until stability has been reached.

This algorithm is easily implemented and appears to converge sufficiently fast in practice. Unlike in MCA, the JCA solutions for different choices of dimensionality are not nested.

In order to assess the quality of the JCA solution, we compare the sum of squared residuals (Rss) with the total variation between the categories of the different variables (Tss). Let $\hat{\mathbf{S}}_u = \mathbf{A}\mathbf{A}^\top$ denote the final approximation of $\mathbf{S}$. Then, $\mathbf{E}_u = \mathbf{S}_u - \hat{\mathbf{S}}_u$,

$$\text{Rss} = \psi = \text{trace}\left(\mathbf{E}_u^\top \mathbf{E}_u\right),$$

and

$$\text{Tss} = \text{trace}\left(\mathbf{S}_u^\top \mathbf{S}_u\right).$$

Hence, the quantity can be expressed as

$$1 - \frac{\text{Rss}}{\text{Tss}}.$$

Just as MCA can be regarded as the principal component analysis of categorical variables, JCA can be regarded as a multigroup factor analysis of categorical data (Van de Velden 2000). Once again, despite the close relationship between EFA and JCA, rotation in JCA has received no attention by researchers.

5. Software packages available to compute MCA

CA and MCA are available in most statistical software packages. However, the implementation in these functions is typically kept to a minimum, and JCA is frequently omitted. In addition, none of the major commercial programs offers recent methodological developments.

The software for computing the most recent methodological developments seems to have been developed mainly in R (R Core Team 2019). The most elaborate package is **ca** (Nenadić and Greenacre 2007): It is an R package that makes it possible to include supplementary points and adjust eigenvalues for improved fit. It also allows for the corresponding adjustments of contributions, joint correspondence analysis (JCA) and subset analysis. The main drawback of **ca** is that it was not developed with a graphical user interface. Lê, Josse, and Husson (2008) offer to compute MCA in R as a part of the **FactoMineR** package, which is dedicated to multivariate data analysis that is implemented within the **Rcmdr** environment (Fox 2005; Fox and Bouchet-Valat 2019) so that a graphical user interface can be used. The only R package that includes the rotation of coordinates is **PCAmixdata** (Chavent *et al.* 2012), and it is limited to orthogonal rotation.

In the context of MATLAB (The MathWorks Inc. 2019), two basic functions are available for computing MCA. They are available at the MATLAB file exchange site: `mcorran1` (MCA based on the indicator matrix; Trujillo-Ortiz 2008), and `mcorran2` (MCA based on the Burt matrix; Trujillo-Ortiz 2009). Again, the implementation in these functions is kept to a minimum. A more elaborate function can be obtained from Pinti, Rambaud, Griffon, and Ahmed (2010). However, their toolbox focuses on the multiple correspondence analysis of fuzzy coded data sets.

In conclusion, although in the context of R there are several packages that perform CA and MCA, MATLAB users currently only have a few basic functions at their disposal. We therefore decided to develop **MultipleCar** as a MATLAB application. Also, to make the procedures accessible to researchers without R or MATLAB programming skills, we created a compiled release of our software.

6. MultipleCar: A MATLAB toolbox to compute MCA

6.1. Overall description of MultipleCar

We have developed **MultipleCar** in MATLAB 2019a, and compiled it to be run in Microsoft Windows 64-bit operating systems. We provide the source code to be used as a typical MATLAB toolbox, but also the standalone version to be run under Windows. We tested **MultipleCar** on several computers with different versions of Windows (7/8/10) and found that it works correctly.

The main characteristics of **MultipleCar** are:

1. MATLAB advanced users can use **MultipleCar** as a typical toolbox, and they can analyze their data using the command line, or the graphical user interface.

2. **MultipleCar** can be used as a standalone Windows application, and the user does not need to have MATLAB installed on the computer. The **MultipleCar** graphical user interface can be used to control the whole toolbox, and no command lines are needed.

3. **MultipleCar** implements the most important features already available in the various R packages: MCA based on the indicator matrix, MCA based on the Burt matrix, the inclusion of supplementary points, adjustment of eigenvalues for improved fit, and JCA. In addition, it is the only toolbox that implements the orthogonal and oblique rotation of coordinates.

These characteristics make **MultipleCar** the most advanced MATLAB toolbox for computing MCA and JCA. In addition, the graphical user interface makes it very helpful for applied researchers with no knowledge of MATLAB or R data analysis programming.

6.2. Procedures implemented in MultipleCar

MultipleCar has been developed to compute multiple correspondence analysis. Below we describe the procedures used in detail. Multiple correspondence analysis can be computed from two different kinds of input matrix: the indicator matrix or the Burt matrix. The suitability of the matrix to be analyzed is assessed by three tests: chi-square, total inertia, and Cramer's V index. In **MultipleCar** the number of dimensions to be retained must be specified. The program computes the principal inertias and adjusted principal inertias from the Burt matrix.

MultipleCar can compute both, MCA and JCA. We regard multiple correspondence analysis as pure exploratory data analysis. From this point of view, the user should be able to inspect the data set from any point of view. To allow this flexibility in the exploratory analysis, users

can switch between different principal coordinates settings, that is, different choices for α in (3). For MCA, the program allows four coordinate configurations:

1. variables as principal coordinates;
2. observations as principal coordinates;
3. both variables and observations as principal coordinates; (Sometimes referred to as the French model; in this case, the row and column points do not constitute a so-called biplot: One set of points cannot be projected on the other to retrieve the approximated data values.)
4. biplot symmetrical coordinates. (Both variables and observations coordinates are scaled to constitute a so-called biplot.)

When **MultipleCar** is used to compute JCA, only one configuration is allowed: variables as principal coordinates. Users can decide to graphically represent just the variable coordinates or to include the subject points, too.

The quality of the coordinates is assessed by computing a variety of indices. The absolute contributions indicate how much a coordinate contributed to the inertia described along the corresponding axis. A relatively high absolute contribution for a particular row indicates that the row had an important influence on determining the position of the axis. The relative contributions are the squared correlations between a variable category and the principal axes. The relative contributions indicate how well a certain point (i.e., the coordinate of a variable category) is represented by a particular axis. They can be interpreted as the amount of inertia that an axis contributed to the inertia of a point.

MultipleCar allows orthogonal varimax rotation and oblique quartimin rotation. In addition, in the context of EFA, loading matrices are frequently weighted before rotations are computed. After rotation, the original distances of points from the origin are re-established, so that the interpretation is not affected by the weights applied. This common practice in EFA is also applied to the context of correspondence analysis (see Lorenzo-Seva *et al.* 2009). Re-scaling coordinates using the masses of the corresponding categories prevents infrequently observed points from playing an important role in determining the rotation angle. When the row-wise normalization of the matrix to be rotated is computed, all the coordinates have the same influence on the final position of the axes. **MultipleCar** allows users to decide which one of these weighting schemes they wish to use (if any).

6.3. Input and output

To run the standalone program, **MultipleCar** must be used on the Windows operating system. To run the graphical user interface with MATLAB as a toolbox, the following command line must be executed in the MATLAB prompt

```
>> MultipleCar
```

Once the main window of **MultipleCar** has been opened, the data can be loaded and the analysis configured. Figure 1 shows the graphical user interface of **MultipleCar**. The input consists of an ASCII file containing scores on the variables, the number of categories in

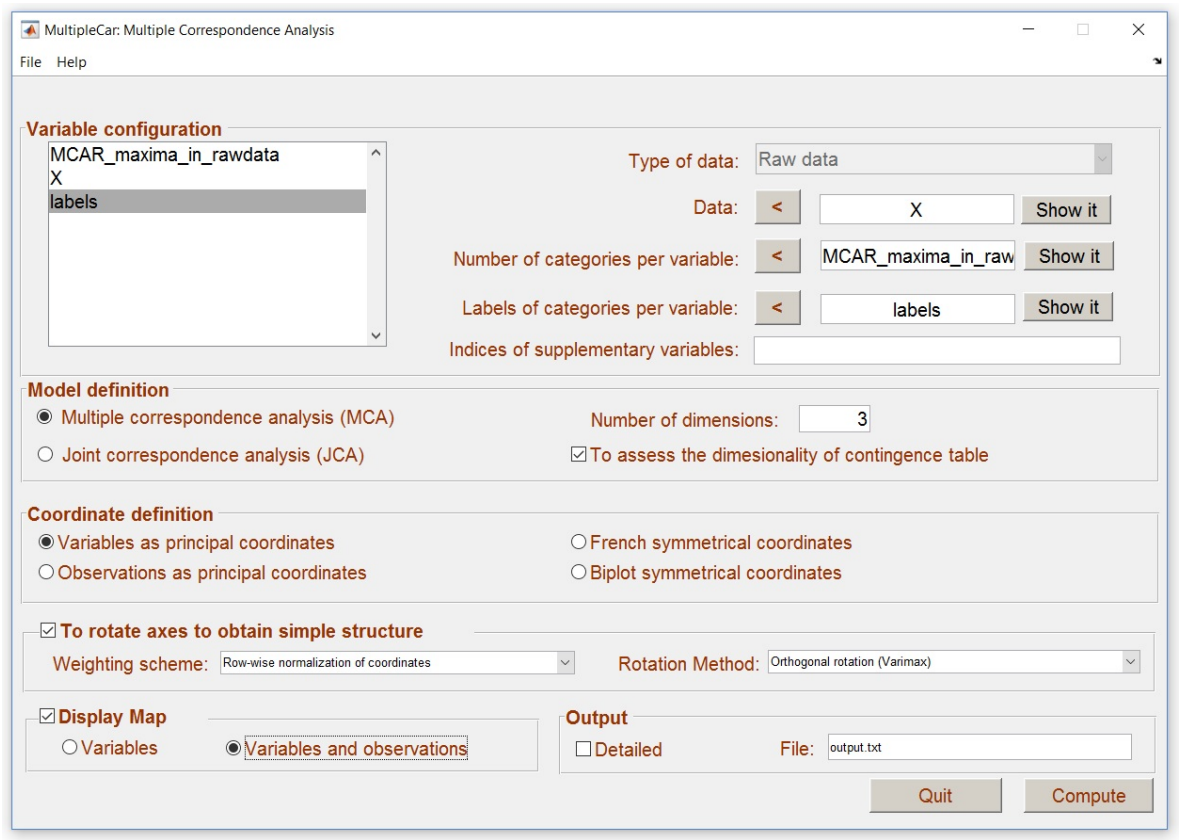

Figure 1: Main window of **MultipleCar** after a multiple correspondence analysis has been configured. This interface can be used to control the whole package, and no further use of the command line interface is in fact needed.

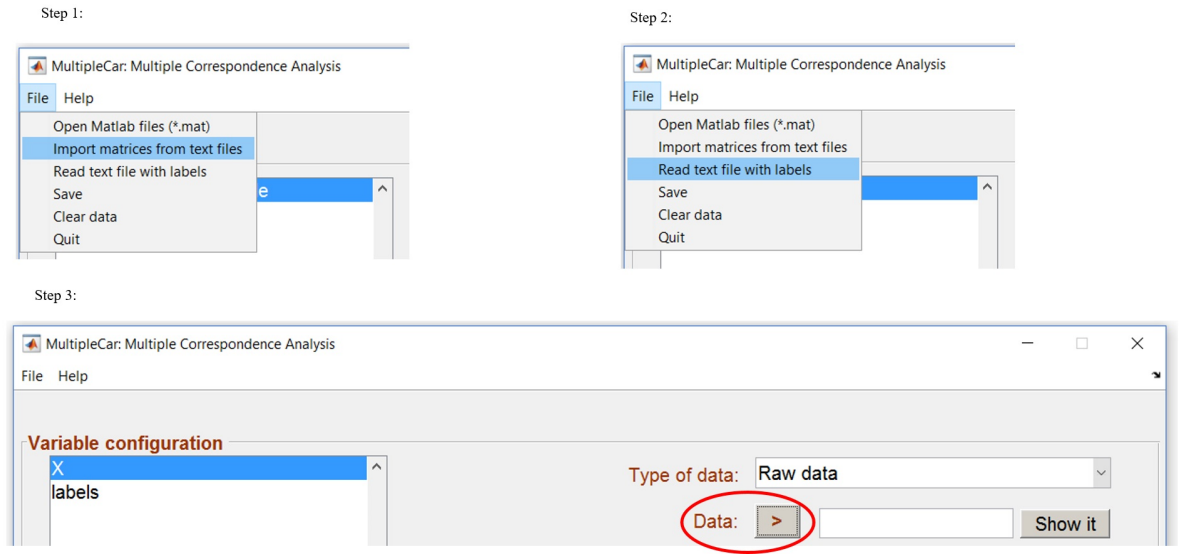

Figure 2: Reading raw data and selecting the data matrix.

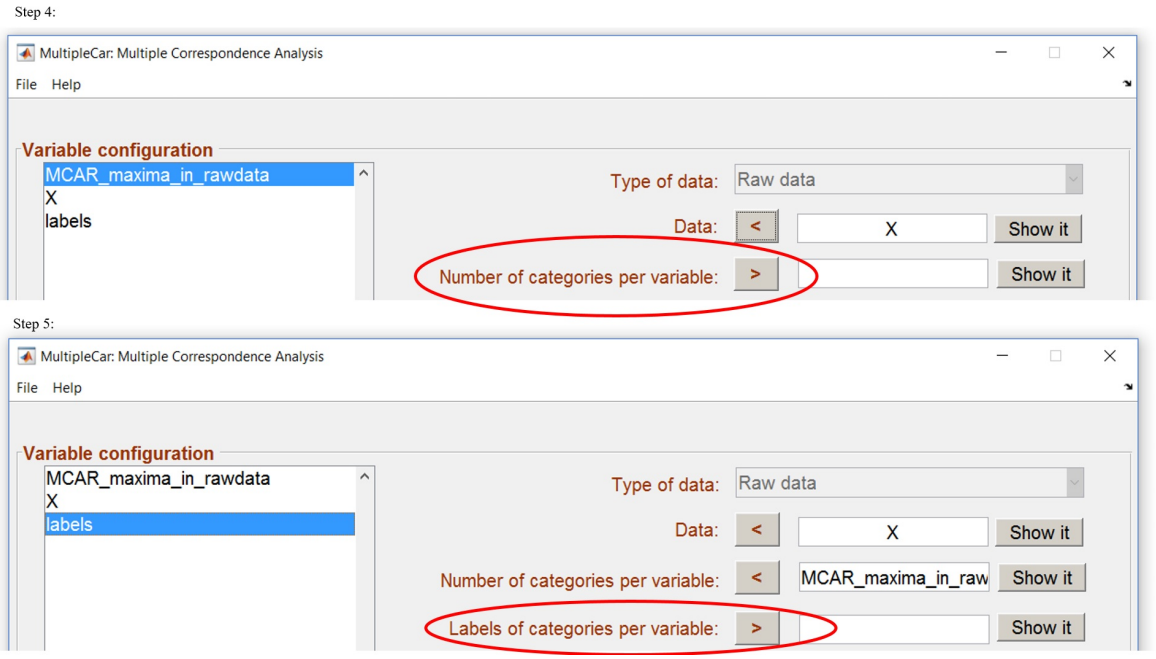

Figure 3: The last two steps consist of reading the variable with the number of categories in the variables and the variable that contains the labels.

each variable, the labels for variable categories (optional), the variables to be considered as supplementary points (optional), and the number of dimensions to retain. Alternatively, an ASCII file containing the indicator matrix or the Burt matrix can be used. Finally, users who are familiar with MATLAB can choose to read the data stored in their own MATLAB files. The output consists of the indices explained above and is stored in the ASCII file `OUPUT.TXT`. If the output information is too detailed, users can choose a simplified output option. In addition, the MAP option displays the coordinates in a bidimensional graph. Figure 1 shows the main window of **MultipleCar**, while Figures 2 and 3 show how to configure the program in order to analyze a data set. We illustrate and explain these steps in the next section by means of an example.

7. An illustrative example

From the Eurostat database (see `http://ec.europa.eu/eurostat`), we recorded three variables for 287 regions in Europe during the year 2013: (1) *Youth employment rate*, (2) *Youth long-term unemployment rate*, and (3) *Country of the region*. The population under study consisted of individuals between 15 and 29 years old, and long-term unemployment meant having been without work for 12 months or longer. The employment rates in the 287 regions ranged from 16.5% (in Dytiki Macedonia, Greece) to 75.2% (in Ostschweiz, Switzerland) and long-term unemployment ranged from 0.5% (in Prague, Czech Republic) to 37.4% (in Ceuta, Spain). We computed a new categorical *Employment* variable coded as: Very low (employment rates between 16.5 and 31.2%), Low (employment rates between 31.3 and 45.9%), High (employment rates between 46.0 and 60.5%), and Very high (employment rates between 60.6 and 75.2%). We computed a new categorical *Long-Term Unemployment* variable coded as:

```
Adjusted principal inertias based on the eigenvalues of the Burt matrix

Dim      Value       %      Cum%   Scree plot

  1     0.6562    54.5      54.5   ****************************************
  2     0.2653    22.1      76.6   ****************
  3     0.0961     8.0      84.6   ******
  4     0.0441     3.7      88.2   ***
  5     0.0139     1.2      89.4   *
  6     0.0002     0.0      89.4   *
  7     0.0000     0.0      89.4   *

        ------  ------
Total:  1.0759    89.4

Average adjusted explained variance:    5.3%
(dimensions explaining less variance should be excluded from the map;
equivalent to Kaiser's One-eigenvalue rule in EFA)
```

Table 1: Adjusted principal inertias based on eigenvalues of the Burt matrix.

Very low (unemployment rates between 0.5 and 9.7%), Low (unemployment rates between 9.8 and 19.0%), High (unemployment rates between 19.1 and 28.2%), and Very high (unemployment rates between 28.3 and 37.4%). Finally, the 287 regions were from one of the 32 countries listed in Table 1.

We analyze the three categorical variables (*Employment*, *Long-Term Unemployment*, and *Country*) using multiple correspondence analysis with **MultipleCar**. In order to read the data and to set the program as presented in Figure 1, the user must follow the five steps presented in Figures 2 and 3. These steps are:

- Step 1: Read the data using the *File* tab. Data can be either read as a text file in ASCII format or a MATLAB (`*.mat`) file and is stored in a data matrix named X. The program expects to find a matrix in which each column corresponds to a categorical variable, and the categories are coded as category numbers (consecutive, starting with 1).
- Step 2: Labels related to each variable in matrix X can be read from a text file in ASCII format and are stored in a text variable named *labels*. Each label is expected to be stored in a row of the text file. The number of rows should be equal to the total number of categories and the order of labels should correspond to the category numbers.
- Step 3: By clicking the arrow next to *Data*, the matrix X can be selected for analysis. When the matrix is selected, the vector `MCAR_maxima_in_rawdata` is computed by the program. This vector contains the maximum value for each variable.
- Step 4: The vector `MCAR_maxima_in_rawdata` is selected to define the number of categories in each variable contained in matrix X. Alternatively, the user can load a vector

using the same procedure described in Step 1, and select it to define the number of categories for each variable.

- Step 5: The text variable *labels* is selected to assign labels to the categories of the variables. If no labels are given, the output will just show the corresponding numerical value.

To compute an MCA solution using the default values, the user needs to click the button *Compute*. The program computes the 247×50 indicator matrix and asks the user to indicate the number of dimensions to retain. This is the dimensionality of the solution. Note that, for CA and MCA the solutions are nested. That is, the choice of the number of dimensions does not affect the (unrotated) solution. It merely removes/adds coordinates. Therefore, some initial choice of the number of dimensions can be used and, based upon the output, a final choice can be made. For JCA, solutions are not nested. In this case, the analysis needs to be re-run for different choices of k to select k based on the explained inertias of solutions of different dimensionalities. In our example, on the basis of the adjusted inertias, we decide to retain three dimensions the explained 84.6% of the total inertia (see Table 1).

MCA outcomes are typically presented as a display of the coordinates in a bidimensional graph. However, when more than two dimensions are retained, the graphical presentation becomes more complex. Figure 4 shows the bidimensional graphs in three panels, one for each pair of dimensions. Understanding the information contained in these graphs is not straightforward. In this case, it is preferable to interpret the rotated principal coordinates. In our example, Bentler's simplicity index (1977) is 0.586 before rotation and 0.965 afterwards. We therefore advise to interpret the rotated coordinates are shown in Table 2. It can be useful to label dimensions by using the rotated coordinate values. Lorenzo-Seva *et al.* (2009) proposed comparing the squared coordinates to the corresponding mean. Only coordinates whose squared values are larger than the mean are considered to be salient coordinates. These salient coordinates can be used to assign labels to the dimensions. The salient coordinates are marked in Table 2 with an $*$ mark.

Finally, the salient coordinates in the first dimension suggest that this dimension is bipolar: One pole is high employment (salient positive coordinates), whereas the other is medium (i.e., values between low and high) and long-term unemployment (salient negative coordinates). The countries are ordered along this bipolar dimension according to their employment/long-term unemployment levels. For example, Estonia, Finland, Latvia, Malta, Sweden, and Germany are on the high employment pole; whereas Ireland, Slovakia, Bulgaria, Portugal, Spain, Croatia, and Cyprus are on the medium and long-term unemployment pole. This dimension could be labeled as *Countries with non-extreme employment/long-term unemployment levels*. Salient coordinates in the second dimension suggest a very low level of employment and a very high long-term unemployment in some countries: Macedonia, Greece, and to some extent Italy. This dimension could be labeled as *Countries with a very difficult employment situation*. Salient coordinates in the third dimension suggest a very high level of employment in some countries: Denmark, the Netherlands, Switzerland, and Norway. This dimension could be labeled as *Countries with very high employment*. Note that the same conclusions can be drawn from the maps in Figure 4. However, as the number of dimensions k was larger than 2, the interpretation of the rotated coordinates is more straightforward.

Labels	Variable categories	D3	D2	D1
E-VL	Employment / Very low	0.738	2.064*	-0.629
E-L	Employment / Low	0.575	-0.636	-0.347
E-H	Employment / High	-1.219*	-0.117	-0.224
E-VH	Employment / Very high	0.002	0.299	2.439*
U-VL	Long-term unemployment / Very low	-0.429	-0.256	0.191
U-L	Long-term unemployment / Low	1.536*	-0.512	-0.405
U-H	Long-term unemployment / High	1.228*	1.548*	-0.446
U-VH	Long-term unemployment / Very high	0.210	3.068*	-0.874
DE	Denmark	0.541	0.398	3.384*
NE	Netherlands	0.541	0.398	3.384*
SW	Switzerland	0.164	0.319	2.701*
NO	Norway	-0.428	0.194	1.627*
MC	Macedonia	-0.250	3.741*	-1.063
GR	Greece	0.320	2.680*	-0.839
IT	Italy	0.880	1.034	0.459
ES	Estonia	-1.720*	-0.078	-0.716
FI	Finland	-1.720*	-0.078	-0.716
LA	Latvia	-1.720*	-0.078	-0.716
ML	Malta	-1.720*	-0.078	-0.716
SW	Sweden	-1.720*	-0.078	-0.716
GE	Germany	-1.343*	0.002	-0.033
AU	Austria	-0.966	0.081	0.651
UK	United Kingdom	-0.856	0.074	0.736
TU	Turkey	-0.303	-0.411	-0.376
FR	France	-0.158	-0.596	-0.423
CZR	Czech Republic	-0.026	-0.658	-0.316
BE	Belgium	0.209	-0.715	-0.367
PL	Poland	0.211	-0.722	-0.336
RO	Romania	0.211	-0.722	-0.336
HU	Hungary	0.215	-0.741	-0.259
LI	Lithuania	0.215	-0.741	-0.259
LU	Luxembourg	0.215	-0.741	-0.259
SL	Slovenia	0.215	-0.741	-0.259
IR	Ireland	1.167*	-1.008	-0.300
SK	Slovakia	1.226*	-0.704	-0.403
BU	Bulgaria	1.404*	-1.075	-0.311
PT	Portugal	1.619*	-0.339	-0.326
SP	Spain	1.696*	0.303	-0.359
CR	Croatia	2.118*	-1.275	-0.342
CY	Cyprus	2.118*	-1.275	-0.342

Table 2: Rotated principal coordinates of variables. Salient coordinates are marked with an $*$ mark.

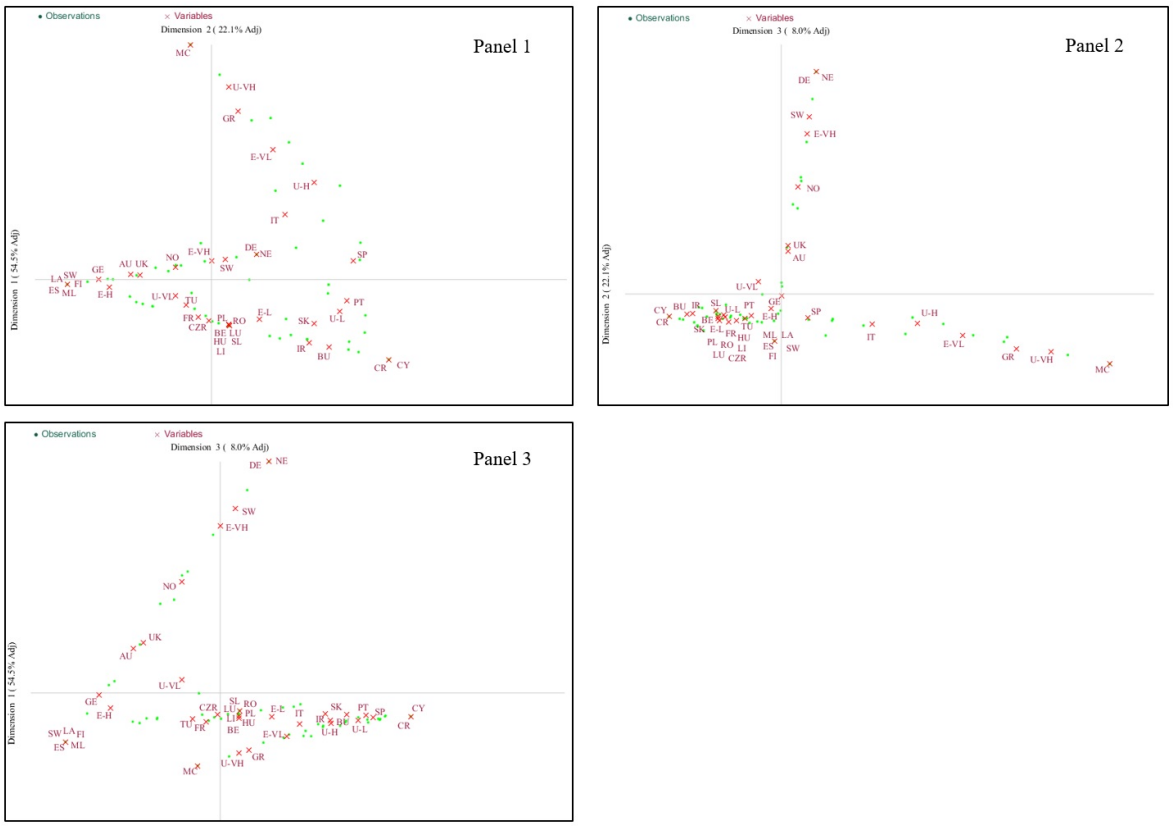

Figure 4: Multiple correspondence analysis maps among pairs of dimensions. The meaning of the labels is printed in Table 2.

8. Program limitations

The number of variables, categories and observations in the data set are not limited. However, when large data sets are analyzed and depending on the characteristics of the computer (processor chip, memory, etc.), computing the indicator matrix can take some time. To speed up the analysis time, the indicator and the Burt matrices are given to the user afterwards so that these matrices can be saved and used for future analyses without needing to compute them again. In addition, if the analysis is made more than once in the same session, **MultipleCar** does not compute the indicator matrix again but uses the one that has already been computed.

9. Program availability

MultipleCar can be downloaded for free from `http://psico.fcep.urv.cat/utilitats/CorrespondenceAnalysis`. The user can download a standalone version of the program to be run on Windows 64-bit operating systems. Alternatively, the **MultipleCar** toolbox can be run as a MATLAB script. Note that, in order to use **MultipleCar** as a standalone application, users that do not have MATLAB installed on their computer first need to download a MATLAB runtime program. The site also provides a manual that includes video tutorials on how to use **MultipleCar**, and some example data sets.

Acknowledgments

This project has been possible with the support of Ministerio de Economía, Industria y Competitividad, the Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF) (PSI2017-82307-P).

References

Adachi K (2004). "Oblique Promax Rotation Applied to the Solutions in Multiple Correspondence Analysis." *Behaviormetrika*, **31**(1), 1–12. doi:10.2333/bhmk.31.1.

Bentler PM (1977). "Factor Simplicity Index and Transformations." *Psychometrika*, **42**(2), 277–295. doi:10.1007/bf02294054.

Benzécri JP (1973). *Tome 1: La Taxinomie. Tome 2: L'Analyse de Correspondances*. Dunod, Paris.

Chavent M, Kuentz-Simonet V, Saracco J (2012). "Orthogonal Rotation in PCAMIX." *Advances in Data Analysis and Classification*, **6**(2), 131–146. doi:10.1007/s11634-012-0105-3.

Fox J (2005). "The R Commander: A Basic Statistics Graphical User Interface to R." *Journal of Statistical Software*, **14**(9), 1–42. doi:10.18637/jss.v014.i09.

Fox J, Bouchet-Valat M (2019). ***Rcmdr**: R Commander*. R package version 2.5-3, URL http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/.

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

Greenacre MJ (1988). *Correspondence Analysis in the Social Sciences*. Academic Press, London.

Greenacre MJ (1993). "Biplots in Correspondence Analysis." *Journal of Applied Statistics*, **20**(2), 251–269. doi:10.1080/02664769300000021.

Kiers HAL (1991). "Simple Structure in Component Analysis Techniques for Mixtures of Qualitative and Quantitative Variables." *Psychometrika*, **56**(2), 197–212. doi:10.1007/bf02294458.

Lê S, Josse J, Husson F (2008). "**FactoMineR**: An R Package for Multivariate Analysis." *Journal of Statistical Software*, **25**(1), 1–18. doi:10.18637/jss.v025.i01.

Lebart L, Morineau A, Warwick KM (1990). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. John Wiley & Sons, New York.

Lorenzo-Seva U, Van de Velden M, Kiers HAL (2009). "Oblique Rotation in Correspondence Analysis a Step Forward in the Search of the Simplest Interpretation." *British Journal of Mathematical and Statistical Psychology*, **62**(3), 583–600. doi:10.1348/000711008x368295.

Nenadić O, Greenacre MJ (2007). "Correspondence Analysis in R, with Two- and Three-Dimensional Graphics: The **ca** Package." *Journal of Statistical Software*, **20**(3), 1–13. doi:10.18637/jss.v020.i03.

Pinti A, Rambaud F, Griffon JL, Ahmed AT (2010). "A Tool Developed in MATLAB for Multiple Correspondence Analysis of Fuzzy Coded Data Sets: Application to Morphometric Skull Data." *Computer Methods and Programs in Biomedicine*, **98**(1), 66–75. doi:10.1016/j.cmpb.2009.09.009.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

The MathWorks Inc (2019). *MATLAB – The Language of Technical Computing, Version R2019a*. Natick. URL http://www.mathworks.com/products/matlab/.

Trujillo-Ortiz A (2008). *Multiple Correspondence Analysis Based on the Indicator Matrix*. MATLAB Central File Exchange version 1.3.0.0, URL https://de.mathworks.com/matlabcentral/fileexchange/22154-multiple-correspondence-analysis-based-on-the-indicator-matrix.

Trujillo-Ortiz A (2009). *Multiple Correspondence Analysis Based on the Burt Matrix*. MATLAB Central File Exchange version 1.3.0.0, URL https://de.mathworks.com/matlabcentral/fileexchange/22558-multiple-correspondence-analysis-based-on-the-burt-matrix.

Van de Velden M (2000). *Topics in Correspondence Analysis*. Ph.D. thesis, University of Amsterdam. Tinbergen Institute Research Series, 238.

Van de Velden M, Kiers HAL (2005). "Rotation in Correspondence Analysis." *Journal of Classification*, **22**(2), 251–271. doi:10.1007/s00357-005-0016-5.

Affiliation:

Urbano Lorenzo-Seva
CRAMC (Research Center for Behaviour Assessment)
Department of Psychology
Universitat Rovira i Virgili
Ctra de Valls s/n, 43007 Tarragona, Spain
E-mail: urbano.lorenzo@urv.cat

Michel van de Velden
Econometric Institute
Erasmus University
P.O. Box 1738
3000 DR Rotterdam, The Netherlands
E-mail: vandevelden@ese.eur.nl

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
July 2019, Volume 90, Issue 4 *Submitted:* 2017-06-30
doi:10.18637/jss.v090.i04 *Accepted:* 2018-04-09