CAR

Lorenzo-Seva, Urbano; van de Velden, Michel; Kiers, Henk A.L.

Published in:
Journal of Statistical Software

DOI:
10.18637/jss.v031.i08

Publication date:
2009

Journal of Statistical Software

September 2009, Volume 31, Issue 8. http://www.jstatsoft.org/

CAR: A MATLAB Package to Compute Correspondence Analysis with Rotations

Urbano Lorenzo-Seva
Rovira i Virgili University

Michel van de Velden
Erasmus University

Henk A. L. Kiers
University of Groningen

Abstract

Correspondence analysis (CA) is a popular method that can be used to analyse relationships between categorical variables. Like principal component analysis, CA solutions can be rotated both orthogonally and obliquely to simple structure without affecting the total amount of explained inertia. We describe a MATLAB package for computing CA. The package includes orthogonal and oblique rotation of axes. It is designed not only for advanced users of MATLAB but also for beginners. Analysis can be done using a user-friendly interface, or by using command lines. We illustrate the use of **CAR** with one example.

Keywords: correspondence analysis, biplot, orthogonal rotation, oblique rotation, simple structure, MATLAB.

1. Introduction

This paper provides a MATLAB (The MathWorks Inc 2007) package that implements correspondence analysis. The package aims to be useful for both beginners and advanced users of MATLAB. The package incorporates a user-friendly interface that helps to control the whole analysis by (a) reading data, (b) selecting variables and the corresponding labels, (c) specifying the axis model and rotations, and (d) configuring the numerical and graphical outcome. The analysis can also be carried out using just a few MATLAB command lines.

Other packages to compute CA are available in the literature (for example, Venables and Ripley 2002; Nenadić and Greenacre 2007; or de Leeuw and Mair 2009). However, these packages were not developed in MATLAB. In addition, our package is the first one that implements orthogonal and oblique rotation of axes.

The remaining sections briefly describe the methodological background to CA and axis rotation (Section 2), and the main technical features of **CAR** (Section 3). In Section 4, an illustrative example is provided to demonstrate the use of **CAR**. Finally, Section 5 provides a brief summary.

2. Methodological background in a nutshell

In this section we give a brief overview of the general concepts of correspondence analysis and axis rotation. If the reader is not familiar with CA, the handbook by Greenacre and Blasius (1994) is advisable reading. Particularly important is the chapter on how to compute CA (Greenacre 1994). It should be noted that **CAR** outcomes are based on the same notation as that proposed in this book. For a more profound understanding of the methodology of axis rotation in the context of CA, we refer to (a) Van de Velden and Kiers (2003, 2005), and (b) Lorenzo-Seva, Van de Velden, and Kiers (2009) for orthogonal and oblique rotation, respectively.

2.1. Correspondence analysis

Let $\mathbf{F}$ be an $n \times p$ contingency table divided by the total number of observations, $\mathbf{1}_i$ an $i \times 1$ vector of ones, $\mathbf{r} = \mathbf{F}\mathbf{1}_p$, $\mathbf{c} = \mathbf{F}'\mathbf{1_n}$ and $\mathbf{D}_r$ and $\mathbf{D}_c$ are diagonal matrices with the elements of $\mathbf{r}$ and $\mathbf{c}$ on the diagonal, respectively. The idea is to analyze a standardized contingence matrix with deviations from independence defined as,

$$\tilde{\mathbf{F}} = \mathbf{D}_r^{-1/2}(\mathbf{F} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2}. \tag{1}$$

The weights (i.e., matrices $\mathbf{D}_r^{-1/2}$ and $\mathbf{D}_c^{-1/2}$) in expression (1) are taken in such a way that rows and columns corresponding to relatively few occurrences receive larger weights than rows and columns corresponding to a large number of occurrences. The aim in correspondence analysis is to find k dimensional coordinate matrices $\mathbf{X}$ and $\mathbf{Y}$, for row and column points, respectively, in such a way that the loss function

$$\phi(\mathbf{X}, \mathbf{Y}) = \|\tilde{\mathbf{F}} - \mathbf{D}_r^{1/2}\mathbf{X}\mathbf{Y}'\mathbf{D}_c^{1/2}\|^2 \tag{2}$$

is minimized subject to $\mathbf{X}'\mathbf{D}_r\mathbf{X} = \mathbf{I}$ and $\mathbf{Y}'\mathbf{D}_c\mathbf{Y} = \mathbf{I}$, where $\|\mathbf{H}\|^2$ denotes the sum of squared elements of $\mathbf{H}$. Let

$$\tilde{\mathbf{F}} = \mathbf{K}\boldsymbol{\Gamma}\mathbf{V}' \tag{3}$$

be the singular value decomposition of matrix $\tilde{\mathbf{F}}$, where $\boldsymbol{\Gamma}$ is a diagonal matrix with singular values on the diagonal, in weakly descending order, and $\mathbf{K}'\mathbf{K} = \mathbf{V}'\mathbf{V} = \mathbf{I}$. As Van de Velden and Kiers (2005) pointed out, $\phi(\mathbf{X}, \mathbf{Y})$ is minimized by

$$\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{K}_k\boldsymbol{\Gamma}_k^{\alpha} \tag{4}$$

and

$$\mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V}_k\boldsymbol{\Gamma}_k^{1-\alpha}, \tag{5}$$

where $\mathbf{K}_k$ and $\mathbf{V}_k$ are, respectively, the $n \times k$ and $p \times k$ matrices of singular vectors corresponding to the k largest singular values in the $k \times k$ diagonal matrix $\mathbf{\Gamma}_k$. Finally, α is the parameter that determines the type of coordinates in $\mathbf{X}$ and $\mathbf{Y}$. Three choices are usually considered for α:

1. $\alpha = 1$: The column coordinates $\mathbf{Y}$ are referred to as *principal coordinates* and the row coordinates $\mathbf{X}$ as *standard coordinates.*

2. $\alpha = 0$: The column coordinates $\mathbf{Y}$ are referred to as *standard coordinates* and the row coordinates $\mathbf{X}$ as *principal coordinates.*

3. $\alpha = 0.5$: Both column and row coordinates are referred to as *symmetrical coordinates.*

An important feature of $\mathbf{X}$ and $\mathbf{Y}$ is that for any choice of α, the matrix product $\mathbf{D}_r^{1/2}\mathbf{X}\mathbf{Y}'\mathbf{D}_c^{1/2}$ optimally approximates $\tilde{\mathbf{F}}$, in the sense that the sum of squared differences between this product and $\tilde{\mathbf{F}}$ is as small as possible. In addition, this equals the biplot model. Hence, these solutions can be interpreted as so-called biplots.

The models corresponding to $\alpha = 0$ and $\alpha = 1$ are often referred to as *asymmetric models*, whereas the model corresponding to $\alpha = 0.5$ is referred to as the *symmetric model.* Another common model in correspondence analysis involves two asymmetric models: the rows of $\mathbf{X}$ when $\alpha = 1$, and the rows of $\mathbf{Y}$ when $\alpha = 0$. Such a model is often referred to as the *French symmetrical model.* It must be noted that the French symmetrical model does not satisfy the biplot requirements.

The distinction between principal, standard and symmetrical coordinates is fundamental in CA:

1. *Principal coordinates* are the coordinates of the set of (column or row) variables that are studied. If $\alpha = 0$, these coordinates are related to column variables, and if $\alpha = 1$, to row variables.

2. *Standard coordinates* are the coordinates of the set of (column or row) variables that help to describe the set of variables studied. If $\alpha = 0$, these coordinates are related to row variables; and, if $\alpha = 1$, to column variables.

3. When *symmetrical coordinates* are chosen, both column and row coordinates are described.

2.2. Supplementary rows and columns

An important and useful property of the CA solution concerns the close relationship between row and column points. In particular, one can calculate one set of coordinates from the other. Let $\mathbf{X}_s$ denote the set of standard row coordinates (i.e., row coordinates as in (4) with $\alpha = 0$). Then, principal coordinates for the columns, $\mathbf{Y}_p$, can be obtained from

$$\mathbf{Y}_p = \mathbf{D}_c^{-1}\mathbf{F}'\mathbf{X}_s, \tag{6}$$

and, using similar notation, principal row coordinates can be obtained from

$$\mathbf{X}_p = \mathbf{D}_r^{-1}\mathbf{F}\mathbf{Y}_s. \tag{7}$$

These formulae are often referred to as transition formulae. A proof and derivation can be found in Van de Velden (2000).

The transition formulae can be used to plot additional rows or columns into the CA map. Let $\mathbf{f}_{sp}$ denote the $p \times 1$ vector with the frequencies for the p column categories for the supplementary row. The k-dimensional coordinates for this supplementary row can be calculated as:

$$\mathbf{x}_{sp} = \left[1/\left(\sum_{j=1}^{p} f_j\right)\right]\mathbf{f}'_{sp}\mathbf{Y}_s. \tag{8}$$

For supplementary column points we can derive a similar formula. Obviously, supplementary points do not play a role in the determination of the CA map and they are therefore also referred to as passive points.

2.3. Diagnostics

To assess the quality of a CA solution, the percentage of total variance, or, as one calls it in CA, inertia, accounted for by the k-dimensional solution can be considered by calculating γ_k/γ_T where $\gamma_k = \sum_{i=1}^{k}\gamma_i^2$, $\gamma_T = \sum_{i=1}^{K}\gamma_i^2$, and k denotes the rank of $\tilde{\mathbf{F}}$.

In addition to this overall measure of fit, which is equal to Pearson's chi squared statistic for testing independence in a contingency table times the total number of observations, we can assess the quality of the row and column coordinates by further decomposing the inertia.

Let $\mathbf{X}$ denote the k dimensional matrix of principal row coordinates. By dividing the weighted squared principal coordinates through the inertias, we obtain so-called absolute contributions for the row coordinates. Thus, the absolute contribution of the i-th row to the j-th axis is defined as

$$\omega_{ij} = (r_i/\gamma_j^2)x_{ij}^2. \tag{9}$$

The term *absolute* refers to the weights r_i, which are equal to the total number of observations in a row, that play a role in the calculation of the contributions of points. The absolute contributions indicate how much a coordinate *contributed* to the inertia described along the corresponding axis. They are often used to assign appropriate labels to the k axes in the CA approximation (see, for example, Greenacre 2007). A relatively high absolute contribution for a certain row indicates that the row had an important influence on determining the position of the axis. Hence, the axes can be labeled in terms of subsets of variables that have a high contribution.

In addition, by considering the squared principal coordinates relative to the weighted sum of squared coordinates over the k dimensions, we obtain the so-called relative contributions for the rows. That is, the relative contribution for the j-th axis to the i-th row is

$$\sigma_{ij} = \frac{x_{ij}^2}{\sum_{l=1}^{k} x_{il}^2}. \tag{10}$$

In this case, the weights r_i are divided out. The relative contributions are the squared correlations between a (in this case) row and the principal axes. They can be interpreted geometrically as the squared cosines of the angles between each row profile and each principal axis. The relative contributions indicate how well a certain point 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. The sum of the first k relative contributions gives an indication of the quality of the representation of a point in the k dimensions.

Absolute and relative contributions for the column coordinates can be obtained in a similar fashion. For a more elaborate treatment, as well as a geometrical interpretation of the contributions, see Greenacre (2007).

2.4. Orthogonal and oblique rotation

Row and column coordinates, $\mathbf{X}$ and $\mathbf{Y}$, are usually inspected in order to explain the meaning of the k dimensions. As in exploratory factor analysis (EFA), the best possible solution is the one which is easiest to interpret. Note that expressions (4) and (5) do not ensure simplicity in the coordinates. However, rotation can be used to maximize the simplicity in rotated row coordinates, $\tilde{\mathbf{X}}$, and rotated column coordinates, $\tilde{\mathbf{Y}}$. If simplicity is maximized in $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$, the interpretation of the dimensions may also be simplified.

Van de Velden (2000) and Van de Velden and Kiers (2005) proposed an orthogonal rotation procedure for matrices $\mathbf{X}$ and $\mathbf{Y}$. As has been pointed out above, $\mathbf{X}$ and $\mathbf{Y}$ satisfy the biplot requirement that $\mathbf{D}_r^{1/2}\mathbf{X}\mathbf{Y}'\mathbf{D}_c^{1/2}$ optimally approximates $\tilde{\mathbf{F}}$. In the orthogonal rotation proposed by Van de Velden and Kiers (2005), the rotated matrices still satisfy the requirement: if $\mathbf{T}$ is an orthogonal rotation matrix ($\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}$), then $\mathbf{D}_r^{1/2}\mathbf{X}\mathbf{T}\mathbf{T}'\mathbf{Y}'\mathbf{D}_c^{1/2} = \mathbf{D}_r^{1/2}\mathbf{X}\mathbf{Y}'\mathbf{D}_c^{1/2}$. Note that both matrices $\mathbf{X}$ and $\mathbf{Y}$ are post-multiplied by $\mathbf{T}$, so the same rotation matrix is used to rotate both matrices. This can be exploited in such a way that only one or both matrices are rotated to simplicity using a joint criterion.

In oblique rotation, it is more straightforward to rotate either $\mathbf{X}$ or $\mathbf{Y}$ to simplicity than both $\mathbf{X}$ and $\mathbf{Y}$. This is because, unlike orthogonal rotation, $\mathbf{X}$ and $\mathbf{Y}$ are not rotated by the *same* rotation matrix if the biplot requirement is to be satisfied.

Let us first consider the case in which we wish to rotate $\mathbf{Y}$ to maximal simplicity. We use standard coordinates for $\mathbf{X}$ (i.e., $\alpha = 0$), and search an oblique rotation matrix $\mathbf{U}$, which maximizes simplicity in the matrix $\tilde{\mathbf{Y}} = \mathbf{Y}(\mathbf{U}')^{-1}$, where $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{U}$. As is customary in oblique factor rotation, we impose the constraint $\operatorname{diag}(\mathbf{U}'\mathbf{U}) = \mathbf{I}$, where $\operatorname{diag}(\mathbf{H})$ denotes the diagonal matrix with the diagonal elements of $\mathbf{H}$ on its diagonal. The rotation matrix $\mathbf{U}$ can be obtained by maximizing simplicity in $\tilde{\mathbf{Y}}$ in terms of, for example, the quartimin criterion (Jennrich and Sampson 1966; see also Clarkson and Jennrich 1988). Analogously, when we wish to have simplicity in $\tilde{\mathbf{X}}$, we take standard scores for $\mathbf{Y}$ (i.e., set $\alpha = 1$) and we search for an oblique rotation matrix $\mathbf{U}$ that maximizes simplicity in $\tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{U}')^{-1}$, $\tilde{\mathbf{Y}} = \mathbf{Y}\mathbf{U}$ and $\operatorname{diag}(\mathbf{U}'\mathbf{U}) = \mathbf{I}$ so that the rotated standard coordinates $\tilde{\mathbf{Y}}$ are standardized. Note that, by following this procedure, the biplot requirement is still satisfied after rotation.

Now we turn to the procedure for rotating both $\mathbf{X}$ and $\mathbf{Y}$ to maximal joint simplicity. This procedure is appropriate when $\mathbf{X}$ and $\mathbf{Y}$ have similar roles: that is, when $\alpha = 0.5$ and coordinates are symmetrical. In this case, we wish to find an oblique rotation matrix $\mathbf{U}$ such that a joint simplicity criterion in terms of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ is optimized. Lorenzo-Seva *et al.* (2009) proposed maximizing the joint simplicity of the above rotated loading matrices $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$, subject to the constraint $\mathrm{diag}(\mathbf{\Phi}) = \mathrm{diag}(\mathbf{U}'\mathbf{U}) = \mathbf{I}$, where $\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{U}')^{-1}$ and $\tilde{\mathbf{Y}} = \mathbf{Y}(\mathbf{U}')^{-1}$ denote rotated loading matrices, while $\mathbf{\Phi}$ is a matrix with unit diagonal. After the oblique rotation of axes, $\mathbf{D}_r^{1/2}\mathbf{X}\mathbf{Y}'\mathbf{D}_c^{1/2} = \mathbf{D}_r^{1/2}\tilde{\mathbf{X}}\mathbf{\Phi}\tilde{\mathbf{Y}}'\mathbf{D}_c^{1/2}$.

After the coordinate matrices have been rotated, $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ can be used to name the dimensions, and to decide which (row and column) variables are best related to each dimension. In the oblique rotation, matrix $\mathbf{\Phi}$ reveals how strongly the dimensions are correlated to one another. It is important to consider which coordinate values in the rotated loading matrices must be interpreted in order to name the dimensions. Lorenzo-Seva *et al.* (2009) proposed computing the mean of the squared coordinates for each dimension in each rotated loading matrix. Then, each squared coordinate is compared to its corresponding mean: only coordinates whose squared values are larger than the mean are considered to be salient coordinates. Labels can be assigned to the dimensions depending on the characteristics of the variables with salient coordinates in the dimension.

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 reestablished, so the interpretation is not affected by the weights applied. For example, Kaiser (1958) proposed to weight the rows of the pattern matrix so that all the rows have the same influence in the final position of the rotated axes. This is a usual practice nowadays when computing orthogonal Varimax rotation, and it is computed as a default in most statistical packages. This kind of weighting schemes is also applied in the context of oblique rotation (see, for example, Lorenzo-Seva 2000).

In the context of CA, other weighting schemes may also be applicable. Let $\mathbf{W}_x$ and $\mathbf{W}_y$ be diagonal matrices, with weights on the diagonal and zeros elsewhere. $\mathbf{W}_x$ and $\mathbf{W}_y$ are weighting matrices related to the coordinate matrices $\mathbf{X}$ and $\mathbf{Y}$, respectively. The aim is to weight the rows of $\mathbf{X}$ and $\mathbf{Y}$ during the rotation, so the products $\mathbf{W}_x\mathbf{X}$ and $\mathbf{W}_y\mathbf{Y}$ are rotated (instead of $\mathbf{X}$ and $\mathbf{Y}$). Three options for $\mathbf{W}_x$ and $\mathbf{W}_y$ can be considered:

1. Each weighting matrix is defined as an identity matrix. With this option, no weight is actually applied.

2. Due to the specific weighting in correspondence analysis, infrequently observed points are sometimes positioned relatively far from the origin. In this situation, these particular points may play an important role in determining the rotation angle. To prevent this from happening, the coordinates can be rescaled using the corresponding masses, i.e., $\mathbf{W}_x = \mathbf{W}_r^{1/2}$ and $\mathbf{W}_y = \mathbf{W}_c^{1/2}$ (Greenacre 2006). This weighting procedure places infrequent points close to the origin, while others remain a long way from it.

3. In the context of EFA, it is common practice to carry out a row-wise normalization of the matrix to be rotated. In CA, this procedure involves rescaling the coordinates using $\mathbf{W}_x = \mathrm{diag}(\mathbf{X}'\mathbf{X})^{-1/2}$ and $\mathbf{W}_y = \mathrm{diag}(\mathbf{Y}'\mathbf{Y})^{-1/2}$. With this scheme, all the rows have the same influence on the final position of the axes.

3. Main technical features of the CAR package

The **CAR** package comprises these three main components:

1. `Car()`: it starts the user-friendly interface that controls the other components in the package. If this function is used, the user does not need either of the other two components in the package.

2. `Canalysis()`: it computes CA. As output, it produces a structure matrix which contains all the matrices related to the analysis. This output can be conveniently printed using `PrintDescriptives()` and `PrintCoordinates()` functions. It should be noted that these functions print the matrices that are relevant to the coordinate model specified by the user in the `Canalysis()` function. In addition, a graphical representation of the dimensions can be obtained using the `Map()` function.

3. `ComputeRotation()`: it computes the orthogonal and oblique rotation of axes. Again, as output, it produces a structure matrix which contains all the matrices related to the rotation. This output can be conveniently printed using the function `PrintRotation()`. This function prints the matrices that are relevant to the rotation specified by the user in the `ComputeRotation()` function.

To use the **CAR** package, one basic step must be carried out in MATLAB: the folder in which the package is stored must be defined as the *current directory*. This can be done by using the current directory window in the MATLAB environment, or the following command line in the MATLAB prompt:

```
>> cd C:\users\desktop\car
```

if **CAR** is stored in the folder `C:`\users\desktop\car . After this, the user can decide whether to use the user-friendly interface, or the MATLAB command lines to execute `Canalysis()` and `ComputeRotation()` functions.

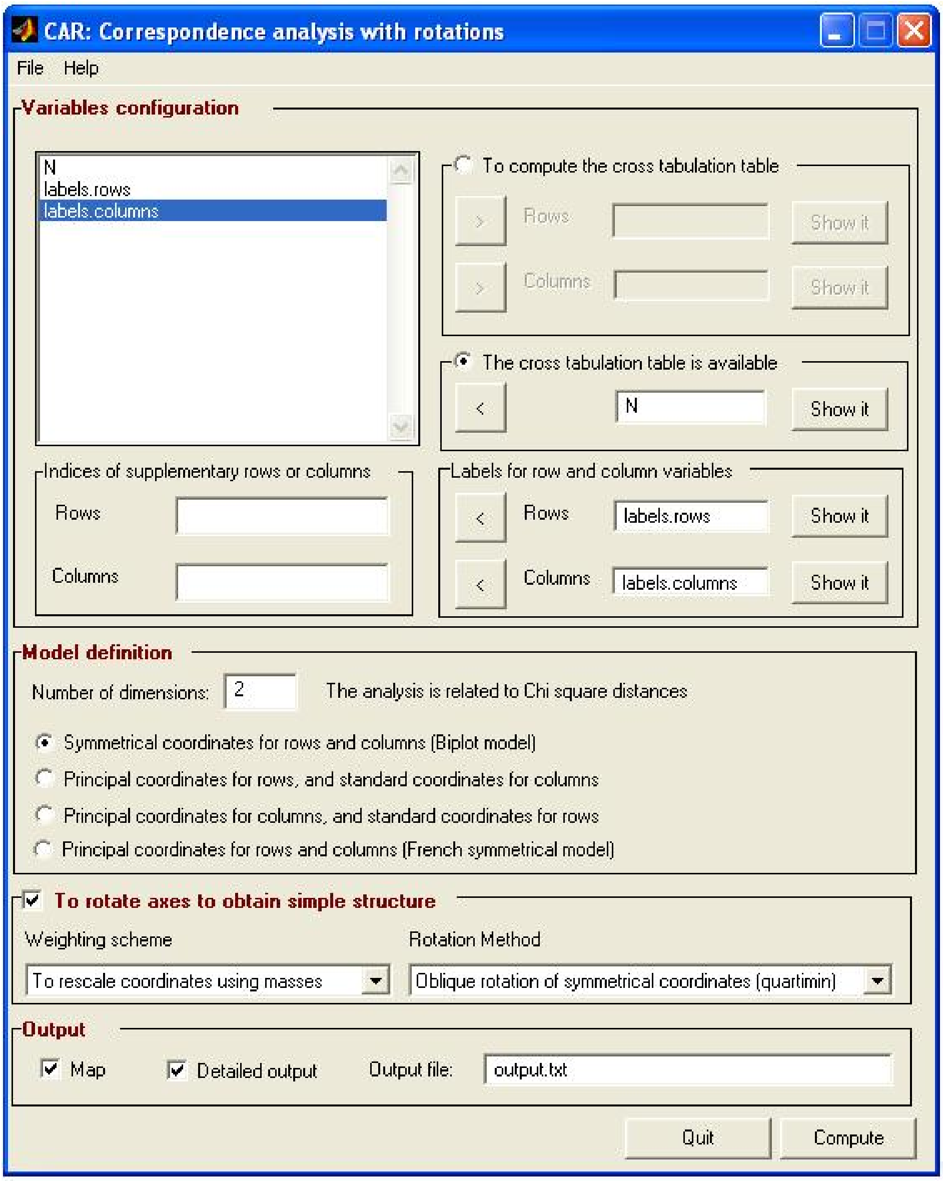
3.1. User interface

To run the user-friendly interface, the following command line must be executed in the MATLAB prompt

```
>> car
```

Figure 3.1 shows the user-friendly interface. This interface can be used to control the whole package, and no further command lines are in fact needed.

To run an analysis, eight steps must be followed: (1) if data are not available in the MATLAB *workspace* when the `Car()` function is initially run, they must be loaded; (2) the cross tabulation matrix must be computed, or the matrix that contains it must be specified in case it is not available in the data loaded; (3) the row and column labels must be defined, in case they are available in the loaded data; (4) supplementary rows or columns in the cross tabulation matrix must be defined in case some of the rows or columns are to be considered as supplementary points; (5) the axis model, which includes the number of dimensions to be retained, and the kind of axes that the user wants to analyze, must be specified; (6) if an axis rotation needs to be computed, the kind of rotation should be specified, and details of the weighting scheme and rotation method be given; (7) the output options should be defined; and, finally,



CAR: Correspondence analysis with rotations

Variables configuration

N
labels.rows
labels.columns

○ To compute the cross tabulation table

Rows:

Columns:

● The cross tabulation table is available: N

Indices of supplementary rows or columns

Rows:

Columns:

Labels for row and column variables

Rows: labels.rows

Columns: labels.columns

Model definition

Number of dimensions: 2

The analysis is related to Chi square distances

● Symmetrical coordinates for rows and columns (Biplot model)

○ Principal coordinates for rows, and standard coordinates for columns

○ Principal coordinates for columns, and standard coordinates for rows

○ Principal coordinates for rows and columns (French symmetrical model)

☑ **To rotate axes to obtain simple structure**

Weighting scheme: To rescale coordinates using masses

Rotation Method: Oblique rotation of symmetrical coordinates (quartimin)

Output

☑ Map ☑ Detailed output Output file: output.txt

Figure 1: User-friendly interface after a particular analysis has been specified.

(8) computing should start. To help users configure an analysis, the interface includes a help menu that guides them through these eight steps. The help that accompanies the package is in CHM and HLP Windows OS formats, and in an RTF document.

3.2. Input data

The analysis can be computed from raw data, or from contingency tables. Labels for rows and/or columns are allowed. Some rows and columns in the cross tabulation matrix can also be considered as supplementary points. The data stored in the MATLAB memory workspace is in fact available in the user-friendly interface. However, data stored in MAT files can also be loaded. In addition, matrices stored in text files (ASCII) can be loaded using the interface. The help that accompanies the package gives an in-depth explanation of the various data formats that can be used.

For example, in Figure 3.1 the interface was configured to compute the analysis from a contingency table that was already available in matrix N. The labels for rows and columns were stored in the matrices `labels.rows` and `labels.columns`, respectively.

3.3. Model definition and rotations

The model definition includes decisions about (1) the number of dimensions to be retained; (2) the kind of coordinates to be interpreted; and (3) whether the axes are to be rotated and, if necessary, the weighting scheme and rotation method to be computed. The rotation methods that can be used depend on the kind of coordinates to be interpreted: the main idea is that, for each kind of coordinate, only rotations methods that have been proposed in the literature are included in **CAR**. For example, when *Symmetrical coordinates for rows and columns (Biplot model)* is specified, only orthogonal Varimax and oblique Quartimin rotations are implemented; however, when *Principal coordinates for rows and columns (French symmetrical model)* is specified, no rotations are implemented because nobody has yet proposed any kind of rotation with this model specification.

3.4. Output

The output is stored in a text file, which, in Windows OS, is automatically shown in the notebook application. To get diagnostic indices and a detailed outcome, you must select the *Detailed* output option on the user interface. The MAP option displays the coordinates in a bidimensional graph. It should be noted that, if more than two dimensions are retained, all the possible bidimensional pairs of coordinates can be displayed using a tool bar in the graph.

3.5. Additional MATLAB command lines

As already pointed out, all the analyses in the **CAR** package can be computed using MATLAB command lines. This means running the `Canalysis()`, `PrintDescriptives()`, `PrintCoordinates()`, `Map()`, `ComputeRotation()`, and `PrintRotation()` functions from the MATLAB prompt. The help included in the package carefully describes how these functions can be used. In addition, there are a few additional commands in the **CAR** package if the user-friendly interface is used. After executing the following two lines,

```
>> output = getappdata(0,'output');
>> rotation = getappdata(0,'rotation');
```

all the output matrices obtained during the analysis are available in `output` and `rotation` MATLAB structures. After executing the following two lines

```
>> help canalysis;
>> help computerotation;
```

the matrices in the `output` and `rotation` structures are described.

Folder	Description
`\car`	This is the main folder and contains the `car.m` file (the function that executes and controls the user-friendly interface) and all the other folders in **CAR** package.
`\car\lib`	This contains all the MATLAB functions implemented in the **CAR** package.
`\car\help`	This contains the help files in CHM and HLP Windows OS format, and a structure of folders that contain the source files for compiling these two help resources. It also contains the `help.rtf` file.
`\car\data`	This contains some data sets that can be useful for learning how to use the **CAR** package.

Table 1: Directory structure of the **CAR** package.

3.6. Directory structure

CAR consists of 106 files organized in four directories (or folders). They are structured as shown in Table 1.

4. Illustrative example: Smoking habits

This example consists of artificial data on the smoking habits of different types of workers in a company (Greenacre 1984, p. 55). Smoking habits were None, Light, Medium, and Heavy; while the types of workers were Senior Managers, Junior Managers, Senior Employees, Junior Employees, and Secretaries. The corresponding contingency table of order 4 × 5 was analysed using the **CAR** package. Traditionally, two dimensions are retained with this data set. Because we aim to describe both variables (smoking habits and employees) and how they are related, we computed symmetrical coordinates (i.e., $\alpha = .5$). In addition, we wanted the axes to be obliquely rotated. As a weighting scheme, we rescaled the coordinates using the corresponding masses. The configuration of the user interface for this analysis is shown in Figure 3.1. Figure 4 shows the map of the unrotated coordinates.

The (detailed) output produced during the analysis consisted of 146 lines stored in an `output.txt` file. To illustrate the output, Table 2 shows an extraction of the output related to the oblique axis rotation. Bentler's simplicity index (1977) is computed in order to assess whether the rotated solution is simpler than the unrotated solution. The index ranges from zero to one, and is maximum only if each variable is generated by a single dimension (i.e., a very simple solution is encountered). Table 2 shows the corresponding values of the index for each matrix. The values of the simplicity index showed that the loading matrices were the simplest matrices and, therefore, the most easily interpreted. Table 2 also shows the rotated coordinates related to the loading matrices. Dimension d1 was a bipolar dimension, and showed that Senior Employees had a tendency to non-smoking habits, whereas Junior

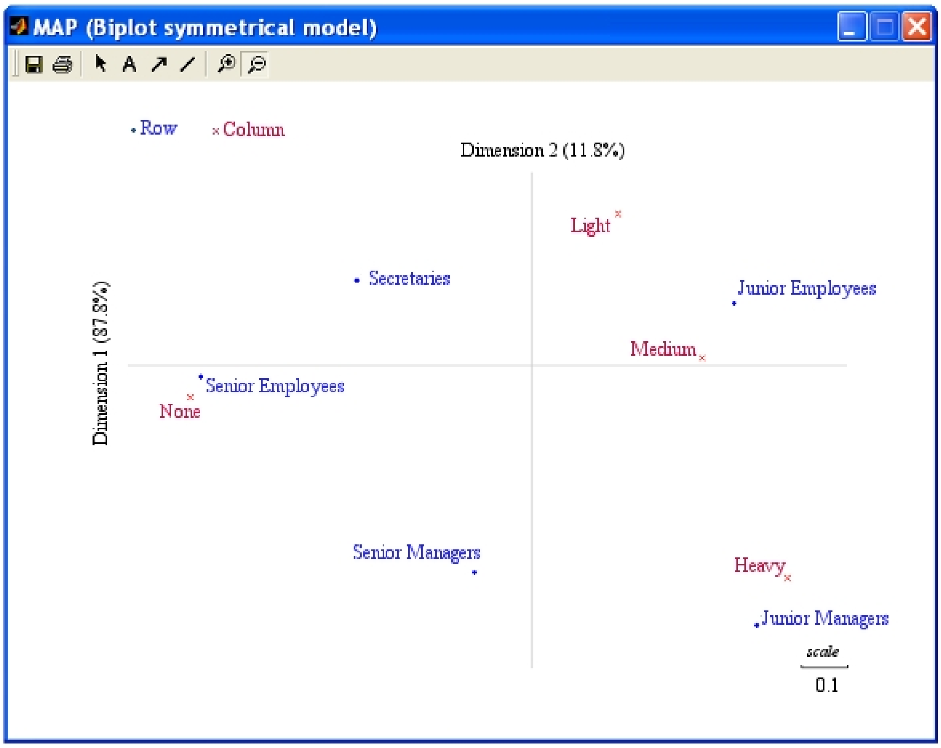


Figure 2: User-friendly interface after a particular analysis has been specified.

Employees had a tendency towards smoking habits. Dimension d2 was a unipolar dimension, and showed that Managers, especially Junior Managers, had a tendency towards strong smoking habits.

The correlation between dimensions was 0.25. If d1 and d2 are related to non-smoking habits and strong smoking habits, respectively, these two dimensions would be expected to be orthogonal (i.e., a correlation of zero). However, it should be noted that the dimensions are also related to the types of workers in the company, and that each type of worker was not so simple as to be related to only one smoking behaviour. Secretaries were the most complex group: they were between low and non-smoking habits. Senior managers were also a complex group: they were close to strong smoking habits but some of them had no smoking habits.

5. Summary

We have presented the MATLAB package **CAR** for simple correspondence analysis. The package contains all the features of commercially software packages as well as a new feature: orthogonal and oblique axis rotations. It is designed for non-MATLAB users and advanced users. For the former, a user-friendly interface was developed to control the whole analysis. For the latter, the analysis can be performed using just a few command lines.

```
ROTATION OF COORDINATES TO MAXIMIZE SIMPLICITY

Bentler's simplicity index (1977) before and after rotation

                     Before      After
                     rotation    rotation
Row coordinates      0.859       0.973
Column coordinates   0.782       0.968

Rotated row coordinates (pattern matrix)

Value                     Dimensions
                           1          2
Senior Managers            0.384     -0.599
Junior Managers           -0.129     -0.874
Senior Employees           0.697      0.102
Junior Employees          -0.497      0.102
Secretaries                0.252      0.325

Rotated column coordinates (pattern matrix)

Value         Dimensions
                1          2
None            0.747      0.043
Light          -0.372      0.417
Medium         -0.362     -0.047
Heavy          -0.254     -0.740

Inter-dimensions correlation matrix

Dimensions       1          2
  1              1.000
  2              0.250      1.000
```

Table 2: Extraction of the output from the illustrative example.

Acknowledgments

This research was partially supported by a grant from the Spanish Ministry of Science and Technology (PSI2008-00236/PSIC), and a grant from the Catalan Ministry of Universities, Research and the Information Society (2005SGR00017).

References

Bentler P (1977). "Factor Simplicity Index and Transformations." *Psychometrika*, **42**, 277–

295.

Clarkson DB, Jennrich RI (1988). "Quartic Rotation Criteria and Algorithms." *Psychometrika*, **53**, 251–259.

de Leeuw J, Mair P (2009). "Simple and Canonical Correspondence Analysis Using the R Package **anacor**." *Journal of Statistical Software*, **31**(5), 1–18. URL http://www.jstatsoft.org/v31/i05/.

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

Greenacre MJ (1994). "Computation of Correspondence Analysis." In MJ Greenacre, J Blasius (eds.), *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*, pp. 53–78. Academic Press, London.

Greenacre MJ (2006). "Tying Up Some Loose Ends in Simple, Multiple, Joint Correspondence Analysis." In A Rizzi, M Vichi (eds.), *COMPSTAT 2006 – Proceedings in Computational Statistics*, pp. 163–186. Springer-Verlag, Berlin.

Greenacre MJ (2007). *Correspondence Analysis in Practice*. 2nd edition. Chapman & Hall/CRC, London.

Greenacre MJ, Blasius J (1994). *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*. Academic Press, London.

Jennrich RI, Sampson PF (1966). "Rotation for Simple Loadings." *Psychometrika*, **31**, 313–323.

Kaiser HF (1958). "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika*, **23**, 187–200.

Lorenzo-Seva U (2000). "The Weighted Oblimin Rotation." *Psychometrika*, **68**, 49–60.

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**, 583–600.

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. URL http://www.jstatsoft.org/v20/i03/.

The MathWorks Inc (2007). *MATLAB - The Language of Technical Computing, Version 7.5*. The MathWorks, Inc., Natick, Massachusetts. URL http://www.mathworks.com/products/matlab/.

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 (2003). "An Application of Rotation in Correspondence Analysis." In H Yanai, A Okada, K Shigemasu, Y Kano, JJ Meulman (eds.), *New Developments in Psychometrics*, pp. 471–478. Springer-Verlag, Tokyo.

Van de Velden M, Kiers HAL (2005). "Rotation in Correspondence Analysis." *Journal of Classification*, **22**, 251–271.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.

Affiliation:

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

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 31, Issue 8 *Submitted:* 2008-12-18
September 2009 *Accepted:* 2009-07-14