Variants of Simple Correspondence Analysis

by Rosaria Lombardo and Eric J. Beh

Abstract

This paper presents a description of the R package **CAvariants**. It performs six variants of correspondence analysis on a two-way contingency table. The main function that shares the same name as the package - `CAvariants` - allows the user to choose (via a series of input parameters) from six different correspondence analysis procedures. These include the classical approach to (symmetrical) correspondence analysis, singly ordered correspondence analysis, doubly ordered correspondence analysis, non symmetrical correspondence analysis, singly ordered non symmetrical correspondence analysis and doubly ordered non symmetrical correspondence analysis. The code provides the flexibility for constructing either a classical correspondence plot or a biplot graphical display. It also allows the user to consider other important features that allow for one to assess the reliability of the graphical representations, such as the inclusion of algebraically derived elliptical confidence regions. This paper provides R functions that elaborates more fully on the code presented in Beh and Lombardo (2014).

Introduction

Computational procedures for detecting the association between two or more categorical variables are important aspects of statistical theory and practice. In particular, correspondence analysis provides a quick and simple graphical summary of how categories and variables are associated with one another. The theoretical aspects of the analysis are well documented in the statistical and allied disciplines; see, for example, Benzécri (1973), Greenacre (1984), Lebart et al. (1984), Beh (2004a), Nishisato (2007), and Beh and Lombardo (2014). Despite the necessity for programs and functions that allow its user to perform correspondence analysis, their availability for many of the varied approaches is generally limited. Commercially available statistical software, such as MATLAB, Minitab, SAS and SPSS provide a means of carrying out correspondence analysis, although their procedures often provide only the most basic of features as part of their output. Generally nothing beyond the calculation of principal inertia values, profile coordinates, contribution to inertia and a two-dimensional correspondence plot are provided. Other popular statistical languages, such as R, provide some packages for performing simple and multiple correspondence analysis of the classical (symmetrical) type, (Murtagh, 2005; Nenadic and Greenacre, 2007; Alberti, 2015; De Leeuw, 2006; De Leeuw and Mair, 2009a; Ringrose, 2012; Kostov et al., 2015). Nevertheless, at present, no popular statistical packages provide functions to perform ordered variants of symmetrical and non symmetrical correspondence analysis.

Overview of correspondence analysis in R

Since the mid 2000's the programming environment of R has proven to be extremely popular in all areas of theoretical and applied statistics. This is due in part to the free availability of the programme from the Comprehensive R Archive Network (CRAN; `http://cran.r-project.org/`), the versatility of the coding environment and the ever increasing number of packages that are now available on the CRAN.

Various R packages have received a great deal of attention for their contribution to the computing of correspondence analysis (CA). One of the first is the **MASS** package (Venables and Ripley, 1999; Ripley et al., 2002). It provides the user with a means of performing simple and multiple correspondence analysis with the option of including supplementary points onto a display. More recently the **ca** package of Nenadic and Greenacre (2007) includes functions for performing simple, multiple and joint correspondence analysis using two and three dimensions for the graphical displays. Supplementary points were incorporated into the R code of Murtagh (2005) while the **anacor** package of De Leeuw and Mair (2009a) allows the user to perform classical and canonical correspondence analysis with missing values (De Leeuw, 2006; De Leeuw and Mair, 2009b). Further, one may refer to the `CA` or `MCA` functions in the **FactoMineR** package by Lé et al. (2008). For lexical tables, the `CaGalt` function incorporated into the **FactoMineR** package by Kostov et al. (2015) may be used. Another recent package - **cabootcrs** - by Ringrose (2012) checks the reliability of association by superimposing onto a plot bootstrap confidence regions. The **CAinterprTools** package by Alberti (2015) makes use of graphical features to enrich a visual interpretation of CA results. Alternatively, De Leeuw and Mair (2009a) prepared the **homals** package for performing Gifi's approach to correspondence analysis. As well, Clavel et al. (2014) presented **dualScale** package for doing dual scaling (i.e. multiple correspondence analysis) of multiple

Table 1: R packages and some CA variants. CA: simple CA; NSCA: non symmetrical CA; MCA: multiple CA; JCA: joint CA; SOCA: singly ordered CA; DOCA: doubly ordered CA; SONSCA: singly ordered NSCA; DONSCA: doubly ordered NSCA; CCA: canonical CA; CNSCA: canonical NSCA; DCA: discriminant CA

	Variants of Correspondence Analysis										
package	CA	NSCA	MCA	JCA	SOCA	DOCA	SONSCA	DONSCA	CCA	CNSCA	DCA
ade4	x	x	x						x		x
anacor	x		x						x		
ca	x		x	x							
cabootcrs	x										
CAinterprTools	x		x								
CAvariants	x	x			x	x	x	x			
cncaGUI									x	x	
dualScale	x		x								
ExPosition	x		x								x
FactoMineR	x		x								
homals	x		x								
MASS	x		x								
MUDICA	x		x								
PTAk	x										
vegan	x		x						x		

choice data. Baxter and Cool (2010) and Alberti (2015) provide a good overview of correspondence analysis using R with an archaeological focus. Another R based package that can be downloaded freely from the CRAN is **ExPosition**. It is written by Herve Abdi and his team and performs a variety of different multivariate data analysis techniques, including correspondence analysis and multiple correspondence analysis. Abdi's group has also been responsible for other variations of correspondence analysis including *multi-block discriminant correspondence analysis* **(MUDICA)** (**?**) and *discriminant correspondence analysis* (**?**). Furthermore, another suite of R functions that enable the user to perform a variety of correspondence analysis techniques is **vegan** (Oksanen et al., 2013), which was developed primarily for vegetation ecologists. It includes functions that provide the user with a large array of techniques to choose from including classic correspondence analysis, canonical correspondence analysis and detrended correspondence analysis. One may also consider the **ade4** package (Dray and Dufour, 2007; Chessel et al., 2004; Thioulouse et al., 1997), which also includes non symmetric correspondence analyse, to analyse ecological and environmental data in the framework of numerous euclidean exploratory methods. Further, the **cncaGUI** package (Nieto Librero et al., 2015) allows canonical correspondence analysis and canonical non symmetrical correspondence analysis providing inferential results by using bootstrap methods. The **PTAk** package includes (Leibovici, 2010, 2015) functions for doing multiway data decomposition, and in particular, it also allows simple correspondence analysis and a generalisation of correspondence analysis for k-way tables. Lastly, but certainly not least, the R code of Murtagh (2005) for performing simple and multiple correspondence analysis may also be considered.

An overview of the broad areas of correspondence analysis that these packages cover is summarised in Table 1. While non symmetrical correspondence analysis for nominal variables is included in some of the R packages on the CRAN that perform correspondence analysis, the remaining ordered variants have not yet been made available in any R package. However, fragments of R code for some of these CA variants are available in Beh and Lombardo (2014). Therefore, this paper provides a comprehensive description of R code that enhances, beyond the classics, the type of correspondence analysis that one may use. The advantages of these variants is that they enable the user to incorporate categorical predictor/response associations and the ordinal structure of a variable. For ordered variables we can easily identify any linear and non-linear sources of association that may exist in the data. The ordered variants also provide a visualisation of non-linear trends of association; the classical approaches to correspondence analysis do not encompass these features.

The theoretical aspects underlying all the six variants of correspondence analysis considered in this paper can be found in Beh and Lombardo (2014) and Lombardo et al. (2016). However, here we will provide the reader with a brief overview of the theoretical aspects of these analyses. We also describe how the algebraic confidence ellipses for polynomial biplots can be derived; this aspect of the analysis has not been described elsewhere.

Some theory

Symmetrical and non symmetrical correspondence analysis

Consider a two-way contingency table $\mathbf{N}$ of dimension $I \times J$ such that it is a cross-classification of two variables consisting of I row categories and J column categories, respectively. Denote the matrix of the joint relative frequencies by $\mathbf{P} = \left(p_{ij}\right)$ so that $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$. Let $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ and $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$ be the ith marginal row proportion and the jth marginal column proportion, general elements of the diagonal matrices, $\mathbf{D}_I$ and $\mathbf{D}_J$, respectively.

There are many ways that correspondence analysis can be performed and Nishisato (2007, Chapter 2) provides an excellent overview of some of them. Here, we present the chi-squared statistic expressed in terms of the weighted sum-of-squares of the centred column profiles since this alternative expression of X^2 is useful when comparing symmetrical correspondence analysis with its non symmetrical variant. Therefore, consider the chi-square statistic of $\mathbf{N}$ which is defined as

$$X^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{\bullet j}}{p_{i\bullet}}\left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}\right)^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{\bullet j}}{p_{i\bullet}}\pi_{ij}^2\,.$$

where $\mathbf{\Pi} = \left(\pi_{ij}\right)$ is the $I \times J$ matrix of centred column profiles. In this case, the weight matrices in $\Re^I$ and $\Re^J$ are defined by the elements of the matrices $\mathbf{D}_I^{-1}$ and $\mathbf{D}_J$, respectively.

Suppose we now treat the column variable as a predictor variable and the row variable as its response variable. When such an asymmetric association structure exists between the two categorical variables one may consider non symmetrical correspondence analysis (Lauro and D'Ambra, 1984; D'Ambra and Lauro, 1989; Kroonenberg and Lombardo, 1999). To quantify this asymmetric association, consider the Goodman-Kruskal (1954) tau index

$$\tau = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{\bullet j}\left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}\right)^2}{1 - \sum_{i=1}^{I} p_{i\bullet}^2} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{\bullet j}\pi_{ij}^2}{1 - \sum_{i=1}^{I} p_{i\bullet}^2} = \frac{\tau_{num}}{\tau_{den}}\,.$$

For this asymmetric case, the weight matrices are $\mathbf{I}$ (an $I \times I$ identity matrix) and $\mathbf{D}_J$ respectively. Notice that the denominator can be treated as a constant term since it does not depend on the predictor variable. For this reason it can be neglected without losing any information about the structure of the association. Therefore τ_{num} is the measure of association considered in non symmetrical correspondence analysis.

In order to graphically depict the association or the prediction of the rows given the columns in a low dimensional space, we may consider the generalized singular value decomposition of the centred column matrix $\mathbf{\Pi}$ using the suitable weight matrices (Kroonenberg and Lombardo, 1999).

Suppose we consider a general framework for the symmetrical and non symmetrical variants of CA (Lombardo et al., 2016), that considers generic weight matrices, $\mathbf{V}_I$ and $\mathbf{W}_J$, in $\Re^I$ and $\Re^J$. This general framework may be defined by considering the weighted centred column profile matrix

$$\tilde{\mathbf{\Pi}} = \mathbf{V}^{1/2}\mathbf{\Pi}\mathbf{W}^{1/2}$$

Therefore, symmetric (or classical) correspondence analysis may be performed by considering $\mathbf{V} = \mathbf{D}_I^{-1}$ and $\mathbf{W} = \mathbf{D}_J$, while non symmetrical correspondence analysis is defined when $\mathbf{V} = \mathbf{I}$ and $\mathbf{W} = \mathbf{D}_J$. Doing so leads to the generalized singular value decomposition (GSVD) of

$$\text{GSVD}\left(\tilde{\mathbf{\Pi}}\right) = \mathbf{A}\mathbf{\Lambda}\mathbf{B}^T\,.$$

where the right and left singular vectors are $\mathbf{A}(= a_{im})$ and $\mathbf{B}(= b_{jm})$, respectively. These quantities have the orthonormality properties with metrics $\mathbf{D}_I^{-1}$ or $\mathbf{I}$ (identity) (in $\Re^I$, depending on the symmetric or asymmetric relationship between the rows and columns) and $\mathbf{D}_J$ (in $\Re^J$), respectively. As usual, the elements of the diagonal matrix of singular values, $\mathbf{\Lambda} = \text{diag}\left(\lambda_m\right)$, are arranged in descending order.

Ordered symmetrical and non symmetrical correspondence analysis

When both variables are ordered, we adapt SVD by using basis vectors for the row and column spaces by performing the bivariate moment decomposition (BMD) on the matrix $\tilde{\mathbf{\Pi}}$. The BMD of $\tilde{\mathbf{\Pi}}$ is expressed as

$$\text{BMD}\left(\tilde{\mathbf{\Pi}}\right) = \mathcal{A}\mathbf{Z}\mathcal{B}^T$$

where $\mathcal{A}$ and $\mathcal{B}$ are the row and column polynomial matrices defined by Emerson (1968), respectively, and $\mathbf{Z}$ is the matrix of the generalized correlations (Rayner and Beh, 2009). The construction of polynomials $\mathcal{A}$ and $\mathcal{B}$ requires the specification of *a priori* scores, $s_X(i)$ and $s_Y(j)$ (defined by `mi` and `mj` in `CAvariants`, respectively), to reflect the ordinal structure of the row and column variables. These polynomials are orthonormal with respect to the weight matrices. For the analysis of nominal variables, when a symmetrical association between the variables is considered, the weights in $\Re^I$ and $\Re^J$ are $\mathbf{D}_I^{-1}$ and $\mathbf{D}_J$, respectively. When an asymmetric association is considered, the weights are given by $\mathbf{I}$ and $\mathbf{D}_J$, respectively.

When only one of the two variables consists of ordered categories, rather than considering the BMD or the GSVD of $\tilde{\mathbf{\Pi}}$, one may consider instead its hybrid decomposition (HD) (Beh, 2001, 2008; Lombardo et al., 2016). This method of decomposition consists of singular vectors for the nominal variable and orthogonal polynomials for the ordered variable. Consider the case, as does the package **CAvariants**, where the column variable consists of ordered categories and the row variable consists of nominal categories. Then the HD of $\tilde{\mathbf{\Pi}}$ takes the form

$$\text{HD}\left(\tilde{\mathbf{\Pi}}\right) = \mathbf{A}\mathbf{Z}\mathcal{B}^T$$

where $\mathbf{A}$ is the column matrix of singular vectors for the nominal row categories and $\mathcal{B}$ is the column matrix of orthogonal polynomials for the ordered column categories. The generic elements of $\mathbf{Z}$, z_{mv}, are the *hybrid* generalised correlations; for further details on these elements see Beh and Lombardo (2014) and Lombardo et al. (2016).

Generalized correlations in ordered CA variants

The generalized correlation matrix, $\mathbf{Z}$, in the BMD of $\tilde{\mathbf{\Pi}}$ reflects the various sources of association between the variables and are derived using orthogonal polynomials (Best and Rayner, 1996; Beh, 1997; Rayner and Beh, 2009). For example, when the row and column scores are defined as consecutive integers such that $s_X(i) = i$ for $i = 1, ..., I$ and $s_Y(j) = j$ for $j = 1, ..., J$, then z_{11} is Pearson's product moment correlation of $\mathbf{N}$. A simple generalization of this correlation is z_{12} which is a measure of the correlation between any change in the location of the row categories and dispersion of the column categories. For this reason, z_{12} is a generalized correlation describing the linear-by-quadratic association between the row and column categories.

For ordered CA variants, the total inertia is

$$\text{Inertia}\left(\tilde{\mathbf{\Pi}}\right) = \sum_{u=1}^{I-1}\sum_{v=1}^{J-1} z_{uv}^2.$$

which can also be written in matrix form as

$$\text{Inertia}\left(\tilde{\mathbf{\Pi}}\right) = \text{trace}\left(\mathbf{Z}^T\mathbf{Z}\right) = \text{trace}\left(\mathbf{Z}\mathbf{Z}^T\right) = \text{trace}\left(\mathbf{\Lambda}^2\right).$$

From the matrix of generalized correlations $\mathbf{Z}$, we can obtain the inertia of each polynomial axis by considering the sum-of-squares of z_{uv} over either u or v. Using BMD or HD, the symmetric and asymmetric measures of association (X^2 and τ) can be partitioned into polynomial components that reflect various sources of variation for each of the categories. The inertias of the polynomial components will henceforth be referred to as *sources of inertia* and are akin to the principal inertia values in (symmetrical or non symmetrical) correspondence analysis.

A formal statistical test of the X^2 or τ index can be made. To test the statistical significance of the total inertia in the symmetrical and non symmetrical case, we can compare the chi-squared statistic, or the C-statistic, $C = \tau * (n-1) * (I-1)$ (Light and Margolin, 1971), with the χ^2 variable distribution with $(I-1)(J-1)$ degree of freedom; see, for example, Beh and Lombardo (2014) for further details.

Unequal inertias of the row and column polynomials. When considering the BMD of $\tilde{\mathbf{\Pi}}$, the total inertia of the row and column spaces ($\Re^I$ and $\Re^J$, respectively) will be identical. However,the inertia associated with each of the row and column polynomials will often be different. For the row categories, there are $I-1$ row inertia values - one for each of the axes - where the inertia of the uth polynomial axis is $z^2_{u\bullet}$. Similarly, there are $J-1$ column inertia values - one for each of the axes - where the inertia of the vth axis is $z^2_{\bullet v}$. For this reason, we recommend constructing polynomial biplots for the ordered variants of correspondence analysis instead of the traditional correspondence plots constructed using principal coordinates. See Beh and Lombardo (2014) and Lombardo et al. (2016) for more details on these features.

For the HD of $\tilde{\mathbf{\Pi}}$, the interpretation and properties of the $\mathbf{Z}$ matrix are a mixture (or hybrid) of $\mathbf{\Lambda}$ from the GSVD and $\mathbf{Z}$ from the BMD. When considering the space $\Re^J$, calculating $\sum_{m=1}^{M^*} z^2_{m1} = z^2_{\bullet 1}$ gives the location component of the ordered columns and represents the principal inertia for this variable along the first polynomial axis. Similarly in $\Re^I$, computing $\sum_{v=1}^{J-1} z^2_{1v} = z^2_{1\bullet} = \lambda^2_1$ yields the principal inertia of the first principal axis for the nominal row variable. Like BMD, HD yields different sets of inertia values for each axis in the $\Re^I$ and $\Re^J$ spaces.

Polynomial biplots and elliptical confidence regions

When constructing a polynomial biplot, the ordered row and column categories can be displayed in a single plot since the row and column coordinates are computed with respect to the same set of polynomial axes. For example, in a polynomial *row metric preserving* (or row isometric) biplot, the column standard polynomial coordinates are

$$\mathbf{G} = \mathcal{B} \quad \left(g_{jv} = \beta_{jv}\right)$$

while the principal polynomial coordinates for the row categories are

$$\mathbf{F} = \mathcal{A}\mathbf{Z} = \tilde{\mathbf{\Pi}}\mathbf{W}_J\mathcal{B} \quad \left(f_{iv} = \alpha_{iu}z_{uv} = \sum_{j=1}^{J} w_{\bullet j}\tilde{\pi}_{ij}\beta_{jv}\right).$$

In practice, the coordinates for both the row and column categories are computed using the same orthonormal polynomial axes, i.e. the column polynomials.

The `plot.CAvariants` function provides the user with the option of constructing parametric (or algebraic) elliptical confidence regions for all the six CA variants not only for the nominal CA variants as originally proposed by Beh (2010). We compute the semi-major and semi-minor axis lengths of the elliptical region for the row and column categories. Here, we provide the ellipse axes lengths for the ordered symmetric variants of correspondence analysis. For example, the semi-major axis length of the confidence ellipse for the ith row category is

$$x_{i(\alpha)} = z^2_{11}\sqrt{\frac{\chi^2_\alpha}{n \times \text{trace}(\mathbf{Z}'\mathbf{Z})}\left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{I-1} a^2_{im}\right)} \tag{1}$$

while the semi-minor axis length for this row is

$$y_{j(\alpha)} = z^2_{22}\sqrt{\frac{\chi^2_\alpha}{n \times \text{trace}(\mathbf{Z}'\mathbf{Z})}\left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{I-1} a^2_{im}\right)}. \tag{2}$$

Similar semi-axis lengths can also be derived for the column categories and for the non-symmetrical CA variants. Furthermore, note that ellipsoids can be constructed for three- or higher- dimensional correspondence plots by considering the input parameter `M >2` in the `plot.CAvariants` function. For further details on this issue see Beh (2010); Beh and Lombardo (2014).

Unlike the confidence circles of Lebart et al. (1984) and the more computationally intensive bootstrap techniques proposed in the literature (Markus, 1994; Linting et al., 2007; Ringrose, 2012; Greenacre, 1984; Lombardo and Ringrose, 2012), constructing confidence ellipses in this manner takes into consideration the contribution of the ith row principal polynomial coordinate in dimensions higher than the second. In fact, since all I dimensions are reflected in the semi-major and semi-minor

axis lengths, all of the contribution that a point to the symmetrical or asymmetrical association can be accounted for in a two-dimensional plot using equations (1) and (2). Additional information for how to compute the p-values of each category point can be easily found by considering a similar theoretical development of the p-values described in (Beh and Lombardo, 2014, 2015) for a correspondence analysis of a contingency table with nominal variables.

An Overview of the CAvariants Package

The primary function discussed in this paper is `CAvariants`. It allows the user to select which analysis to perform from a suite six correspondence analysis techniques. These include symmetrical (or classical) correspondence analysis, non symmetrical correspondence analysis and their ordered variants, described in (Beh and Lombardo, 2014, 2015).

The six variations of simple correspondence analysis included in the package **CAvariants** are

- The classical approach to simple correspondence analysis (the default analysis and is defined by the input parameter `catype = "CA"`)
- Two-way, or doubly ordered, symmetrical correspondence analysis (the user can perform this analysis by defining the input parameter `catype = "DOCA"`)
- One-way, or singly ordered, correspondence analysis for tables of symmetrically related variables, where the column variable is ordered (the user can perform this analysis by defining the input parameter `catype = "SOCA"`).
- Non symmetrical correspondence analysis where the nominal column variable is a predictor of the nominal row variable (the user can perform this analysis by defining the input parameter `catype = "NSCA"`)
- Two-way, or doubly ordered, non symmetrical correspondence analysis where the ordered column variable is a predictor of the ordered row variable (the user can perform this analysis by defining the input parameter `catype = "DONSCA"`).
- One-way, or singly ordered, non symmetrical correspondence analysis, where the ordered column variable is a predictor of the nominal row variable (the user can perform this analysis by defining the input parameter `catype = "SONSCA"`)

The input parameters of the function `CAvariants` are:

- The two-way contingency table, `Xtable`.
- The assigned ordered scores for the row categories. By default, `mi = NULL` which gives consecutive integer valued (natural) scores.
- The assigned ordered scores for the column categories. By default, `mj = NULL` which gives consecutive integer valued (natural) scores.
- The horizontal polynomial or principal axis. By default `firstaxis = 1`.
- The vertical polynomial or principal axis. By default `lastaxis = 2`.
- The input parameter for specifying what variant of correspondence analysis is considered. By default `catype = "CA"`, other possible values are: `catype = "SOCA"`,`catype = "DOCA"`,`catype = "NSCA"`,`catype = "SONSCA"`,`catype = "DONSCA"`.
- The input parameter, `ellcomp`, ensures that the characteristics of the algebraic confidence ellipses are computed and stored. When `ellcomp = TRUE` (which is by default), the output includes the characteristics of the ellipses. The eccentricity of the confidence ellipses is summarised by the quantity `eccentricity`, this is the distance between the center and either of its two foci, which can be thought of as a measure of how much the conic section deviates from being circular (when it is equal to zero then the region becomes circular). The semi-major axis length of the ellipse for each row and column point is given by `HL Axis 1` while `HL Axis 2` gives the semi-minor axis length of the points along the second axis. The area of the ellipse for each row and column category is given by `Area` while the p-value of each category is defined by `P-value`.
- The number of axes `Mell` considered in determining the structure of the elliptical confidence regions. By default, `Mell = min(nrow(Xtable),ncol(Xtable)) -1`, i.e. the rank of the data matrix.
- The confidence level, `alpha`, of the elliptical regions. By default, `alpha = 0.05`.

To visually portray and assess the statistical significance of the categories to the association between the variables of a contingency table, the graphical function called `plot.CAvariants` can be called by the user. As well as displaying the classic correspondence plot or biplot, this function allows one to

superimpose onto the plot algebraically derived elliptical confidence regions for each of the principal coordinates (Lebart et al., 1984; Beh, 2010; Lombardo and Ringrose, 2012; Beh and Lombardo, 2015) for all CA variants. There are other features of the plot, that through `plot.CAvariants`, the user may utilise. Some of these are applicable to all of the analyses and some are applicable to only a few. The input parameters of the function `plot.CAvariants` are:

- The name of the output object, for example say `res`, used with the main function `CAvariants`.
- The horizontal polynomial or principal axis, `firstaxis`. By default, `firstaxis = 1`.
- The vertical polynomial or principal axis, `lastaxis`. By default, `lastaxis = 1`.
- The size of characters, `cex`, displayed on the correspondence plot or biplot. By default, `cex = 0.8`.
- The parameter, `cex.lab` that specifies the size of character labels of axes in graphical displays. By default, `cex.lab = 0.8`.
- The scaling parameter, `prop`, for specifying the limits of the plotting area. By default, `prop = 1`.
- The type of graphical display required (either a classical correspondence plot or a biplot). The user can look at a classical correspondence plot by defining the input parameter `plottype = "classic"`. When `plottype = "biplot"`, it produces biplot graphical displays, or polynomial biplots in case of an ordered analysis. Note that for ordered analysis only polynomial biplots are suitable. In particular for the singly ordered variants, only row isometric polynomial biplots make sense, as we assume that the ordered variable is the column variable (the column coordinates are standard polynomial coordinates and the row coordinates are principal polynomial coordinates). By default, `plottype = "biplot"`.
- For a biplot, one may specify that it be a row-isometric biplot (`biptype = "row"`) or a column-isometric biplot (`biptype = "column"`). This feature is available for the nominal symmetrical and the non symmetrical correspondence analyses. By default, a row-isometric biplot, `biptype = "row"`, is produced.
- The parameter for scaling the biplot coordinates, `scaleplot`, originally proposed in Section 2.3.1 of Gower et al. (2011) and described on page 135 of Beh and Lombardo (2014). By default, `scaleplot = 1`.
- The parameter `posleg` for specifying the position of the legend when portraying trends of ordered categories in ordered variants of correspondence analysis. By default, `posleg = "topleft"`.
- The parameter `pos` for specifying the position of point symbols in the graphical displays. By default, `pos = 2`.
- The logical parameter, `ell` which specifies whether algebraic confidence ellipses are to be included in the plot or not. Setting the input parameter to `ell = TRUE` will allow the user to assess the statistical significance of each category to the association between the variables. The ellipses will be included when the plot is constructed using principal coordinates (being either row and column principal coordinates or row and column principal polynomial coordinates). By default, this input parameter is set to `ell = FALSE`. See also the input parameter `ellcomp` of the function `CAvariants` for a description of the numeric characteristics of the confidence ellipses (eccentricity, area, etc.), as well as the input parameter `ellprint` of the function `print.CAvariants` for getting a print of these parameters.
- The number of axes `Mell` considered when portraying the elliptical confidence regions. By default, it is equal to `Mell = min(nrow(Xtable),ncol(Xtable))-1`, i.e. the rank of the data matrix. This parameter is identical to the input parameter `Mell` of the function `CAvariants`.
- The confidence level of the elliptical regions. By default, `alpha = 0.05`.

The print method included in the package, **CAvariants**, is `print.CAvariants` and displays the main results of the analysis specified by the user. The results displayed depends on the type of analysis being performed. The principal inertia values, total inertia and p-values are included as part of its ouput when `catype = "CA"`, `catype = "SOCA"` or `catype = "DOCA"` and are based on Pearson's chi-squared statistic. The Goodman Kruskal tau-index is the association measure of interest when `catype = "NSCA"`, `catype = "SONSCA"` or `catype = "DONSCA"`. When an ordered analysis is specified - such as when `catype = "DOCA"`, `catype = "SOCA"`, `catype = "SONSCA"` or `catype = "DONSCA"` - a table describing the significant polynomial components of inertia will also be reported.

The input parameters of `print.CAvariants` are:

- The name of the output object, for example say `res`, used with the main function `CAvariants`.
- The number of dimensions, `printdims`, that are used to generate the correspondence plot, or biplot, and for summarising the numerical output of the analysis. By default, `printdims = 2`.

- The flag parameter, `ellprint`, allows that the characteristics of the confidence ellipses (eccentricity, semi-axis, area, p-values) are displayed. By default, `ellprint = TRUE`.
- The number of axes, `Mell`, used for the construction of the confidence ellipses. By default, it is equal to its maximum value, `Mell = min(nrow(Xtable),ncol(Xtable)) -1`, i.e. the rank of the data matrix. This input parameter is identical to the parameter `Mell` of both the functions `CAvariants` and `plot.CAvariants`.
- The level of significance used for the construction of the elliptical regions, `alpha`. By default, `alpha = 0.05`.
- The minimum number of decimal places, `digits`, used for displaying the numerical summaries of the analysis. By default, `digits = 3`.

In general, this function produces the following output:

- The two-way contingency table, `Xtable`.
- The matrix of row weights, `Row weights: Imass`. These weights depend on the type of analysis performed.
- The matrix of column weights, `Column weights: Jmass`. These weights are equal to the data column margins for all types of analysis performed.
- The total inertia, `Total inertia`, of the analysis performed. For example, when considering the variants of non symmetrical correspondence analysis, the numerator of the Goodman-Kruskal tau index, the associated C-statistic and its p-value is produced.
- The inertia values, their percentage contribution to the total inertia and the cumulative percent inertias of the row and column space`Inertias`. When performing an ordered correspondence analysis, this output summary describes both the row and column spaces for each principal or polynomial axis. When `catype` is `"CA"` or `"NSCA"`, the associated inertia values in the row and column spaces are identical.
- The generalized correlation matrix `Generalized correlation matrix`, when performing an ordered correspondence analysis, `catype` should be `"DOCA"`, `"DONSCA"`, `"SOCA"` or `"SONSCA"`.
- The row principal coordinates, `Row principal coordinates`, when `catype` is `"CA''` or `"NSCA"`.
- The column principal coordinates, `Column principal coordinates`, when `catype` is `"CA"` or `"NSCA"`.
- The row standard coordinates, `Row standard coordinates` , when `catype` is `"CA"` or `"NSCA"`.
- The column standard coordinates, `Column standard coordinates`, when `catype` is `"CA"` or `"NSCA"`.
- The row principal polynomial coordinates, `Row principal polynomial coordinates`, when performing an ordered correspondence analysis.
- The column principal polynomial coordinates, `Column principal polynomial coordinates`, when performing a doubly ordered correspondence analysis.
- The `Row standard polynomial coordinates`, i.e. standard polynomial coordinates for the row categories when performing a doubly ordered correspondence analysis.
- The `Column standard polynomial coordinates`, i.e. standard polynomial coordinates for the column categories when performing an ordered correspondence analysis.
- The Euclidean distance of the row categories from the origin of the plot, `Row distances from the origin of the plot.`
- The Euclidean distance of the column categories from the origin of the plot, `Column distances from the origin of the plot.`
- The polynomial components of the total inertia and their p-values, `Polynomial components`. The total inertia of the column space is partitioned to identify polynomial components when `catype` is `"SOCA"` or `"SONSCA"`. When `catype` is `"DOCA"` or `"DONSCA"`, the total inertia of both the row and column space is partitioned to identify of polynomial components.
- The inner product, `Inner product`, of the biplot coordinates (concerning the first two axes when `firstaxis = 1` and `lastaxis = 2`).
- When the input flag parameter is `ellprint = TRUE`, then the print includes the eccentricity of the confidence ellipses, the semi-major axis length of the ellipse for each row and column point, `HL Axis 1`, the semi-minor axis length for the ellipse for each row and column point, `HL Axis 2`, the area of the ellipse for each row and column point, `Area` and the p-value for each row and column point, `P-value`, see also the parameter `ellcomp` of the function `CAvariants` for a detailed description of these parameters.

Furthermore, the summary method included in the package, **CAvariants**, is `summary.CAvariants`. The main input parameter of this function is given by the name of the output object of the main function `CAvariants`. The output of `summary.CAvariants` provides the list of the objects names of the output and a selection of the main output objects described in `print.CAvariants`.

Numerical outputs

As an example of the complete set of numerical results that is obtained from performing a particular variant of correspondence analysis, consider the case where a singly ordered non symmetrical correspondence analysis is performed on the data table `shopdataM` available in the package **CAvariants**. This object is the contingency table being analysed and is described more fully in the Application section. The output object name of the main function is called `res` and is the execution of the `CAvariants` function on the `shopdataM`. The object `res` is defined such that

```
R> res <- CAvariants(shopdataM, catype = "SONSCA")
R> names(res)
```

The results are available in the following entries

```
1 res$Xtable
2 res$rows
3 res$cols
4 res$r
5 res$rowlabels
6 res$collabels
7 res$Rprinccoord
8 res$Cprinccoord
9 res$Rstdcoord
10 res$Cstdcoord
10 res$Cstdcoord
11 res$tauden
12 res$tau
13 res$inertiasum
14 res$inertias
15 res$inertias2
16 res$comps
17 res$catype
18 res$mj
19 res$mi
20 res$pcc
21 res$Jmass
22 res$Imass
23 res$Trend
24 res$Z
25 res$ellcomp
26 res$risell
27 res$Mell
```

These results may be printed to the screen by using

```
R> print.CAvariants(res)
```

while a summary of each of these numerical features is produced by using

```
R> summary.CAvariants(res)
```

Application

To demonstrate the application of a variant of simple correspondence analysis described in the **CAvariants** package, we present the following example. We shall confine our attention to the non symmetrical correspondence analysis of a singly ordered contingency table. The contingency table that we are examining is concerned with shoplifting in The Netherlands and summarises, in part, the results of a survey of the Dutch Central Bureau of Statistics (Israëls, 1987). The data considers a sample of 20819 men who were suspected of shoplifting in Dutch stores between 1977 and 1978. The predictor variable consists of the age groups of the perpetrators (less than 12yrs, 12 to 14yrs, 15 to

17yrs, 18 to 20yrs, 21 to 29yrs, 30 to 39yrs, 40 to 49yrs, 50 to 64yrs, 65yrs and over) while the response variable of the table consists of the items stolen. These items are *clothing*, *clothing accessory*, *tobacco and/or provisions*, *stationary*, *books*, *records*, *household goods*, *candy*, *toys*, *jewelry*, *perfume*, *hobby and/or tools* and *other items*. For an extensive description of this example, and the application of correspondence analysis, see Lombardo et al. (2016).

After choosing the suitable variant of correspondence analysis, we create the object `res` that consists of the complete features of the analysis by running the command `res <-CAvariants(shopdataM,catype = "SONSCA")`.

Using the function `print.CAvariants(res)` will return as part of its output the following numerical features

```
    RESULTS for SONSCA Correspondence Analysis

    Data Table:
            M12<  M13 M16 M19 M25 M35 M45 M57 M65+
clothing     81  138 304 384 942 359 178 137  45
accessories  66  204 193 149 297 109  53  68  28
tobacco     150  340 229 151 313 136 121 171 145
stationary  667 1409 527  84  92  36  36  37  17
books        67  259 258 146 251  96  48  56  41
records      24  272 368 141 167  67  29  27   7
household    47  117  98  61 193  75  50  55  29
candy       430  637 246  40  30  11   5  17  28
toys        743  684 116  13  16  16   6   3   8
jewelry     132  408 298  71 130  31  14  11  10
perfumes     32   57  61  52 111  54  41  50  28
hobby       197  547 402 138 280 200 152 211 111
other       209  550 454 252 624 195  88  90  34

    Row Weights: Imass
            clothing accessories tobacco stationary books records household candy toys jewelry
clothing           1           0       0          0     0       0         0     0    0       0
accessories        0           1       0          0     0       0         0     0    0       0
tobacco            0           0       1          0     0       0         0     0    0       0
stationary         0           0       0          1     0       0         0     0    0       0
books              0           0       0          0     1       0         0     0    0       0
records            0           0       0          0     0       1         0     0    0       0
household          0           0       0          0     0       0         1     0    0       0
candy              0           0       0          0     0       0         0     1    0       0
toys               0           0       0          0     0       0         0     0    1       0
jewelry            0           0       0          0     0       0         0     0    0       1
perfumes           0           0       0          0     0       0         0     0    0       0
hobby              0           0       0          0     0       0         0     0    0       0
other              0           0       0          0     0       0         0     0    0       0
            perfumes hobby other
clothing           0     0     0
accessories        0     0     0
tobacco            0     0     0
stationary         0     0     0
books              0     0     0
records            0     0     0
household          0     0     0
candy              0     0     0
toys               0     0     0
jewelry            0     0     0
perfumes           1     0     0
hobby              0     1     0
other              0     0     1

    Column Weights: Jmass
      12<   13    16     19    25     35     45     57    65+
12< 0.137 0.00 0.000 0.0000 0.000 0.0000 0.0000 0.0000 0.0000
13  0.000 0.27 0.000 0.0000 0.000 0.0000 0.0000 0.0000 0.0000
16  0.000 0.00 0.171 0.0000 0.000 0.0000 0.0000 0.0000 0.0000
```

```
19  0.000 0.00 0.000 0.0808 0.000 0.0000 0.0000 0.0000 0.0000
25  0.000 0.00 0.000 0.0000 0.166 0.0000 0.0000 0.0000 0.0000
35  0.000 0.00 0.000 0.0000 0.000 0.0665 0.0000 0.0000 0.0000
45  0.000 0.00 0.000 0.0000 0.000 0.0000 0.0394 0.0000 0.0000
57  0.000 0.00 0.000 0.0000 0.000 0.0000 0.0000 0.0448 0.0000
65+ 0.000 0.00 0.000 0.0000 0.000 0.0000 0.0000 0.0000 0.0255

 Total inertia  0.038

Inertias, percent inertias and cumulative percent inertias of the row space

  inertia inertiapc cuminertiapc
1  0.0300     79.88        79.88
2  0.0037      9.86        89.74
3  0.0032      8.44        98.18
4  0.0003      0.92        99.10
5  0.0003      0.67        99.77
6  0.0001      0.17        99.94
7  0.0000      0.05        99.99
8  0.0000      0.01       100.00
Inertias, percent inertias and  cumulative percent inertias of the column space

  inertia2 inertiapc2 cuminertiapc2
1   0.0225      59.83         59.83
2   0.0096      25.58         85.41
3   0.0028       7.33         92.74
4   0.0019       5.18         97.92
5   0.0003       0.82         98.74
6   0.0003       0.74         99.48
7   0.0001       0.35         99.83
8   0.0001       0.17        100.00

    Predictability Index for Variants of Non symmetrical Correspondence Analysis:

 Numerator of Tau Index predicting the rows given the column categories

[1] 0.038

 Tau Index predicting the rows given the column categories

[1] 0.041

 C-statistic 10331.51 and p-value 0

 Polynomial Components of Inertia

** Column Components **
                 Component Value P-value
Location                6181.536       0
Dispersion              2642.363       0
Cubic                    757.192       0
Error                    750.418       0
** C-Statistic **      10331.509       0

 Generalized correlation matrix of Hybrid Decomposition
       v1     v2     v3     v4     v5     v6     v7     v8
m1 -0.147  0.084  0.018 -0.030  0.011  0.005 -0.005  0.003
m2 -0.028 -0.034 -0.032  0.024  0.005 -0.010  0.003  0.001
m3 -0.013 -0.037  0.036 -0.016 -0.004  0.006 -0.006  0.002
m4 -0.001  0.002  0.006  0.014 -0.010  0.005 -0.001 -0.001
m5 -0.001 -0.001 -0.007 -0.006 -0.007  0.009 -0.004 -0.004
m6  0.000  0.000 -0.001  0.000  0.000 -0.001 -0.006  0.005
m7  0.000  0.000  0.000 -0.001 -0.002 -0.003 -0.001 -0.002
m8  0.000  0.000  0.000  0.000 -0.001  0.000  0.001  0.001
```

```
m9  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000

 Column standard polynomial coordinates = column polynomial axes
      Axis 1  Axis 2
M12<  -1.232   1.352
M13   -0.759   0.142
M16   -0.285  -0.652
M19    0.188  -1.029
M25    0.661  -0.991
M35    1.135  -0.536
M45    1.608   0.336
M57    2.081   1.624
M65+   2.554   3.328

  Row principal polynomial coordinates
             Axis 1  Axis 2
clothing      0.072  -0.056
accessories   0.017  -0.014
tobacco       0.039   0.017
stationary   -0.084   0.033
books         0.012  -0.012
records       0.000  -0.021
household     0.015  -0.004
candy        -0.045   0.027
toys         -0.067   0.049
jewelry      -0.017  -0.006
perfumes      0.014   0.000
hobby         0.030   0.015
other         0.014  -0.030

 Column distances from the origin of the plot
      Axis 1  Axis 2
M12<   0.057   0.002
M13    0.027   0.000
M16    0.000   0.000
M19    0.027   0.002
M25    0.046   0.004
M35    0.041   0.000
M45    0.031   0.005
M57    0.021   0.022
M65+   0.010   0.047

 Row distances from the origin of the plot
             Axis 1  Axis 2
clothing      0.005   0.003
accessories   0.000   0.000
tobacco       0.001   0.000
stationary    0.007   0.001
books         0.000   0.000
records       0.000   0.000
household     0.000   0.000
candy         0.002   0.001
toys          0.005   0.002
jewelry       0.000   0.000
perfumes      0.000   0.000
hobby         0.001   0.000
other         0.000   0.001

 Inner product of coordinates (first two axes when 'firstaxis=1' and 'lastaxis=2')
              M12<    M13    M16    M19    M25    M35    M45    M57   M65+
clothing     0.111  0.097  0.014 -0.112 -0.150 -0.118 -0.065 -0.012  0.044
accessories  0.029  0.022  0.002 -0.024 -0.031 -0.027 -0.019 -0.011 -0.002
tobacco      0.057  0.017 -0.010 -0.003  0.004 -0.023 -0.059 -0.094 -0.122
stationary  -0.132 -0.089 -0.003  0.089  0.114  0.110  0.098  0.085  0.064
```

```
books        0.023  0.015  0.000 -0.014 -0.018 -0.018 -0.018 -0.017 -0.015
records      0.011  0.008  0.000 -0.008 -0.010 -0.010 -0.008 -0.006 -0.004
household    0.023  0.014  0.000 -0.013 -0.016 -0.018 -0.018 -0.018 -0.017
candy       -0.074 -0.049 -0.001  0.048  0.061  0.061  0.055  0.049  0.039
toys        -0.122 -0.070  0.004  0.061  0.074  0.088  0.101  0.113  0.115
jewelry     -0.021 -0.013  0.000  0.012  0.015  0.016  0.016  0.016  0.015
perfumes     0.021  0.010 -0.001 -0.008 -0.009 -0.013 -0.018 -0.023 -0.026
hobby        0.048  0.007 -0.012  0.010  0.021 -0.011 -0.055 -0.098 -0.135
other        0.026  0.030  0.007 -0.039 -0.054 -0.036 -0.010  0.017  0.043

    Eccentricity of ellipses
[1] 0.757

    Ellipse axes, Area, p-values of rows
            HL Axis 1 HL Axis 2 Area P-value
clothing        0.013     0.009    0   0.000
accessories     0.010     0.007    0   0.000
tobacco         0.011     0.007    0   0.000
stationary      0.010     0.007    0   0.000
books           0.012     0.008    0   0.000
records         0.008     0.005    0   0.000
household       0.015     0.010    0   0.000
candy           0.013     0.008    0   0.000
toys            0.011     0.007    0   0.000
jewelry         0.013     0.008    0   0.000
perfumes        0.015     0.010    0   0.297
hobby           0.011     0.007    0   0.000
other           0.011     0.007    0   0.000

    Ellipse axes, Area, p-values of columns
     HL Axis 1 HL Axis 2  Area P-value
M12<     0.034     0.022 0.002       0
M13      0.020     0.013 0.001       0
M16      0.020     0.013 0.001       0
M19      0.023     0.015 0.001       0
M25      0.025     0.016 0.001       0
M35      0.026     0.017 0.001       0
M45      0.031     0.020 0.002       0
M57      0.046     0.030 0.004       0
M65+     0.070     0.046 0.010       0
```

The total inertia of data, defined by the Goodman-Kruskal tau index (which may also be referred to as the index of predictability) when performing a non symmetrical correspondence analysisis, is $\tau = 0.0414$; in the output this is reflected by `Tau Index predicting the rows given the column`. To determine whether this index is statistically significant, we compute the C-statistic and find that it is equal to 10331.5 (with 96 degrees of freedom). Therefore, with a p-value that is less than 0.0001, the age of the perpretators is a strong predictor of the items that are stolen. The Goodman-Kruskal tau index and the statistical significance of the C-statistic are summarised as part of the ouput, together with the partition of the C-statistic, which identifies significant sources of variation in the ordered column categories. Indeed, we can look at the inertia explained by each polynomial axis to mark differences with the other non-ordered analysis. We can see that the most dominant contribution to the total inertia of the data is due to the component associatied with the linear polynomial of the columns. This location component is 6182 and explains 59.8% of the total inertia. The next most dominant is the dispersion component of 2642 and reflects that 25.6% of the variation in the column categories is due to their difference in dispersion. Similarly, the cubic component is 757 and accounts for about 7% of the column variation. Even if the remaining, higher order, components are all statistically significant (their associated p-value is less than 0.001), they will be not taken into consideration since polynomials with degree higher than three (and more commonly, four) show limited information about the association structure and variation of the variables. Hence, collectively, components higher than the fourth are referred to as the *error* polynomial component. Note that the first two components (linear and dispersion) explain 85.4% of the total inertia, so the first two polynomial axes will provide a sufficient graphical display of the variation of the categories. Furthermore, with the specification of `ellprint = TRUE` in `print.CAvariants`, the output consists of the eccentricity value of the ellipses, the semi-axis lengths of the ellipse for each of the categories, the area of each ellipse and the associated p-values.

Polynomial biplot: Portraying the predictability

When an ordered analysis is performed, the trend plots of the row and column categories are depicted. For example, when performing a singly ordered NSCA, the variation, or trend, of the row categories is examined by observing how it is affected by the ordered column categories when using a polynomial transformation. Figure 1 shows a parabolic trend of the row category *clothing*. This trend highlights that there is a greater propensity to steal clothing by people aged 25 to 45 years than those of a younger, or older, age. Figure 2 provides an alternative visual display of these trends and is constructed by depicting the row (items) categories using principal coordinates and the column (age) categories using standard coordinates. Hence a row isometric biplot is constructed. Since the analysis also incorporates the ordered nature of the column categories and the nominal structure of the row categories, Figure 2 is referred to as the row isometric polynomial biplot of the data.

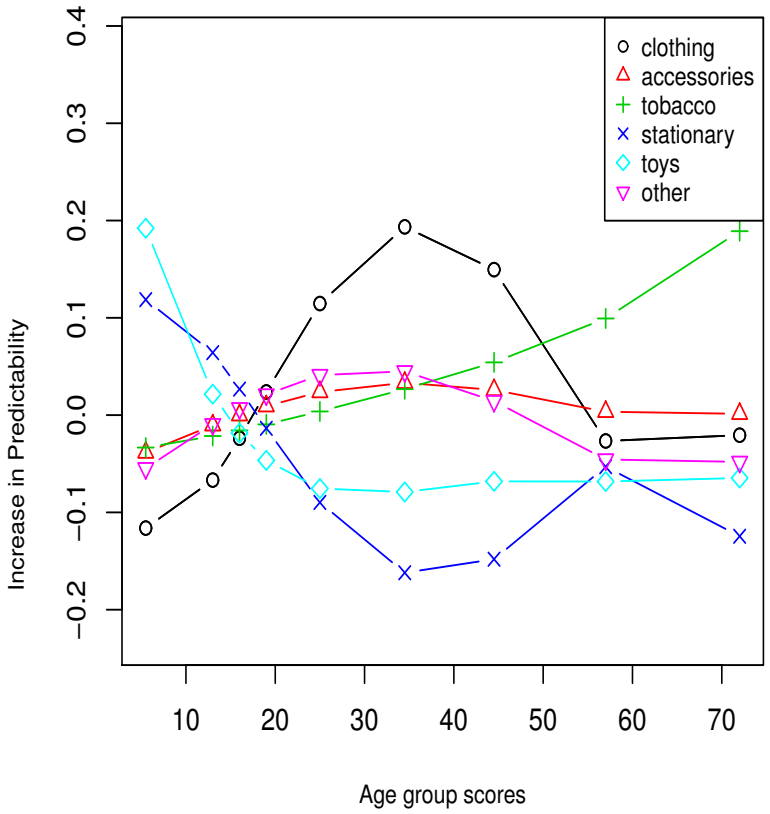

Figure 1: Trend of rows: a selection of rows of the centred column profile table reconstructed by using the first two polynomials.

The trend plot of Figure 1 and the polynomial biplot given by Figure 2 can be obtained using the following command

```
R> plot.CAvariants(res,  plottype = "biplot", biptype = "row", scaleplot = 5, pos = 1)
```

When the first two polynomial axes are used to construct the biplot of Figure 2, the resulting configuration has a parabolic shape. Observe that the explained inertia of the polynomial axes is as follows: the first polynomial axis accounts for 59.8% of the inertia and the second polynomial axis for 25.6% of the inertia. We can therefore see that the novelty of the polynomial biplot is based on the polynomial representation of the predictor variable. The first linear polynomial axis represents the deviation from the mean centred profile accounting for the ordered structure of the age groups, which is reflected in the correct ordering of the age categories along the first polynomial axis. The second polynomial axis shows a parabolic shape of the categories with positive concavity. Furthermore, note that the left-hand side of the first axis is dominated by the young age groups with adolescents and

young adults at the centre of the display (who steal items consistent with the average number of thefts of all items). The mid-adult and older age groups are on the right-hand side of Figure 2.

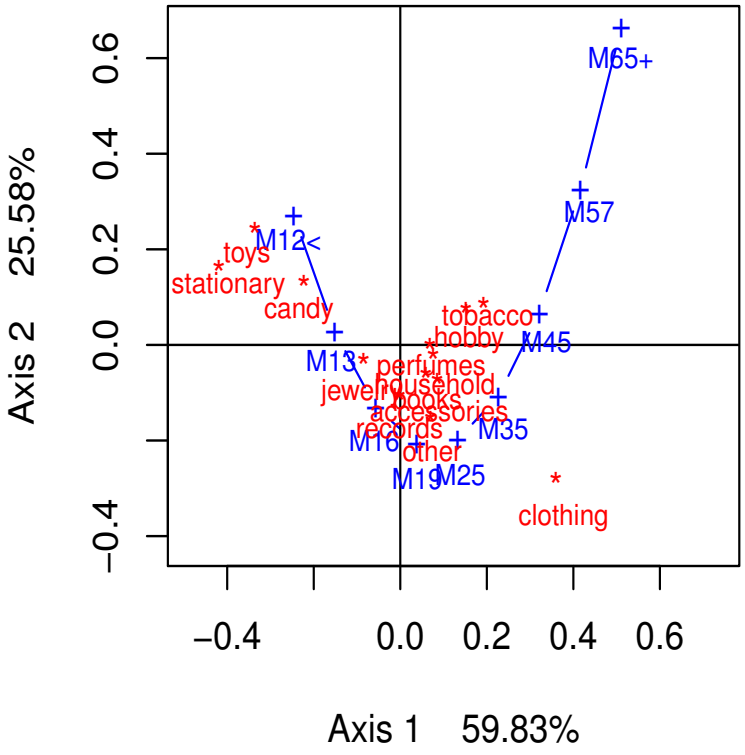

Figure 2: Row-isometric polynomial biplot of singly ordered NSCA of shoplifting data: first two polynomial components, *Stolen goods* and *Age*.

The magnitude of the coordinates indicate the importance of the first two polynomial components for modelling the trends of the items. In particular, we see that the first two polynomial coordinates are sufficient to model the trends for most stolen goods. The reliability of the graphical representation can be assessed by constructing elliptical confidence regions for the row categories which are depicted using row principal polynomial coordinates. These ellipses can be obtained using the `plot.CAvariants` function such that

```
R> plot.CAvariants(res, scaleplot = 1, ell = TRUE, alpha = 0.05)
```

Figure 3 gives the 95% confidence ellipses for the row categories and are constructed so that the weights of the semi-axes are expressed in terms of the hybrid generalized correlations rather than the squared singular values associated to each axis. These ellipses are constructed so that the information contained in all of the dimensions is depicted so that, for `plot.CAvariants`, `M = 8`. Since this figure does not show clearly ellipses for a scale problem of coordinates, we can focus our attention more closely to those points closer the origin of Figure 3 by specifying that

```
R> plot.CAvariants(res, scaleplot = 1, ell = TRUE, alpha = 0.05, prop = 60)
```

By zooming closer to the origin, the configuration of points near the origin is given by Figure 4. It shows the overlap of the confidence region for perfumes with the origin. It means that all of the items, except *perfumes*, are important contributors to the asymmetric association since their confidence ellipse do not overlap the origin of the plot.

The contribution of the all items to the association structure is also reflected in the p-values that are summarised as part of the `print.CAvariants` output with `M = 8` and appear in the last column of the table, titled `Ellipse axes,Area,p-values of rows` where `alpha = 0.05`. These results show that the only non-statistically significant row category is *perfumes*, as expected from its ellipse, with

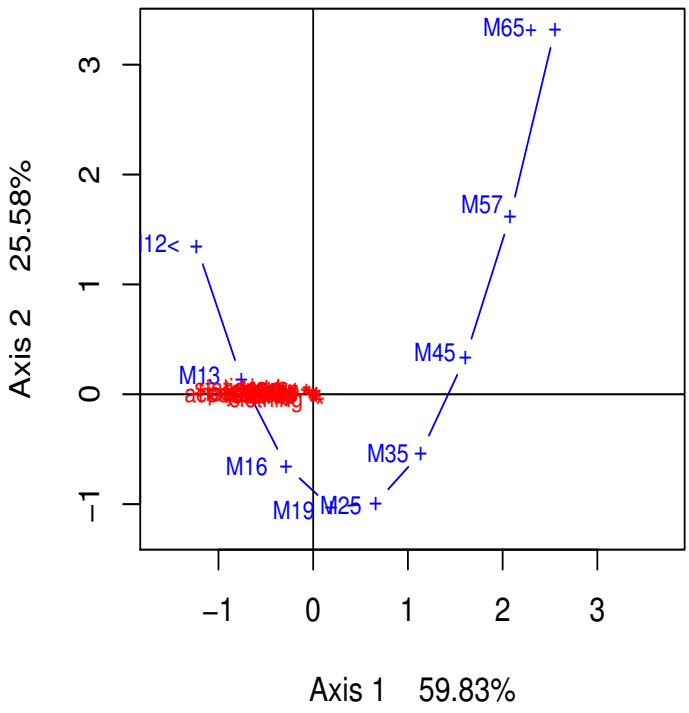

Figure 3: 95% confidence ellipses in the row isometric polynomial biplot of singly ordered NSCA of the shoplifting data: *Stolen goods* and *Age*.

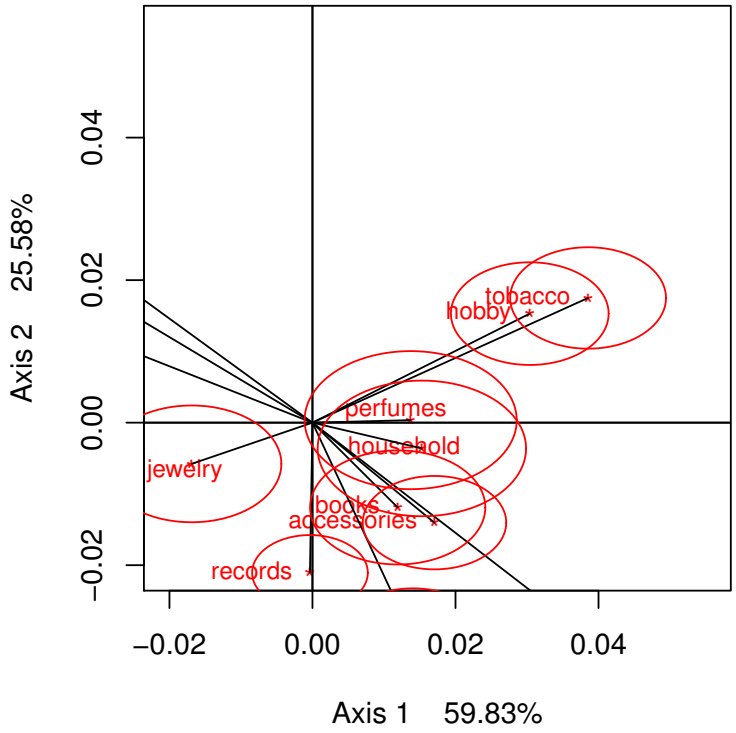

Figure 4: A zoomed view of the origin of the row-isometric polynomial biplot given by Figure 3

a p-value of 0.297. If we now consider the age of the males in the sample, and specify `M = 8` when constructing confidence ellipses and calculating p-values, see the last column of the table titled `Ellipse axes,Area,p-values of columns`, all age groups are useful predictors of the items that are stolen.

Conclusion

There are many freely downloadable programs/code available for performing classical correspondence analysis. For example, the R code of Nenadic and Greenacre (2007) and De Leeuw and Mair (2009a) may be considered for performing simple and joint correspondence analysis. However, the **CAvariants** package provides variants of correspondence analysis which are not offered by other correspondence analysis R packages on the CRAN. To the best of these authors' knowledge, **CAvariants** is the only package available that provides the user with the option of performing six variants of two-way correspondence analysis and, in particular, ordered symmetrical and non symmetrical correspondence analysis variants. Indeed, symmetrical correspondence analysis for ordered variables was implemented in SPLUS by Beh (2004b) and has been easily adapted for R.

Subsequent versions of the function may allow for more flexibility by giving the user more tools to assess the reliability of graphical results. These may include bootstrap confidence regions to complement the algebraic regions developed by these authors, or three-dimensional polynomial biplots. While Beh and Lombardo (2014) and Lombardo et al. (2016) describe the theoretical aspects of these variants of correspondence analysis for two-way contingency tables in detail, they also provide fragments of R code to undertake the relevant calculations. However, this paper has described the CAvariants package by demonstrating the applicability of one variant and providing new insight into the development of elliptical regions for ordered variants of correspondence analysis.

Bibliography

G. Alberti. CAinterprTools: An R package to help interpreting correspondence analysis results. *SoftwareX*, 1-2:26–31, 2015. [p1, 2]

J. Baxter, M and M. Cool, H E. Correspondence analysis in R for archaeologists: An educational account. *Archeologia e Calcolatori*, 21:211–228, 2010. [p2]

E. Beh, J. Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical Journal*, 39:589–613, 1997. [p4]

E. Beh, J. Partitioning pearson's chi-squared statistic for singly ordered two-way contingency tables. *The Australian and New Zealand Journal of Statistics*, 43:327–333, 2001. [p4]

E. Beh, J. Simple correspondence analysis: A bibliographic review. *International Statistical Review*, 72: 257–284, 2004a. [p1]

E. Beh, J. S-PLUS code for ordinal correspondence analysis. *Computational Statistics*, 19:593–612, 2004b. [p18]

E. Beh, J. Simple correspondence analysis of nominal-ordinal contingency tables. *Journal of Applied Mathematics and Computer Sciences*, 8:1–12, 2008. [p4]

E. Beh, J. Elliptical confidence regions for simple correspondence analysis. *Journal of Statistical Planning and Inference*, 140:2582–2588, 2010. [p5, 7]

E. Beh, J and R. Lombardo. *Correspondence Analysis: Theory, Practice and New Strategies*. John Wiley & Sons, 2014. [p1, 2, 4, 5, 6, 7, 18]

E. Beh, J and R. Lombardo. Confidence regions and p-values for classical and non-symmetric correspondence analysis. *Communications in Statistics, Theory and Methods*, 44:95–114, 2015. [p6, 7]

J. Benzécri, P. *Analyse des Données*. Dunod, Paris, 1973. [p1]

J. Best, D and J. Rayner, W C. Nonparametric analysis for doubly ordered two-way contingency tables. *Biometrics*, 52:1153–1156, 1996. [p4]

D. Chessel, A. Dufour, B, and J. Thioulouse. The ade4 package – i: One-table methods. *R News*, 4:5–10, 2004. [p2]

G. Clavel, J, S. Nishisato, and A. Pita. Dual scaling analysis of multiple choice data. https://cran.r-project.org/web/packages/dualScale, 2014. [p1]

L. D'Ambra and N. Lauro, C. Non-symmetrical correspondence analysis for three-way contingency table. In S. B. R. Coppi, editor, *Multiway Data Analysis*, pages 301–315. Elsevier, Amsterdam, 1989. [p3]

J. De Leeuw. Correspondence analysis in R. www.cuddyvalley.org/psychoR/ca, 2006. [p1]

J. De Leeuw and P. Mair. Simple and canonical correspondence analysis using the R package Anacor. *UCLA Statistics Preprint Series*, 31:1–18, 2009a. [p1, 18]

J. De Leeuw and P. Mair. Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software*, 31(4):1–20, 2009b. [p1]

S. Dray and A. Dufour, B. The ade package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22 (4):20 pages, 2007. [p2]

P. Emerson, L. Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics*, 24:696–701, 1968. [p4]

J. Gower, S. Lubbe, and N. le Roux. *Understanding Biplots*. John Wiley & Sons, Chichester, 2011. [p7]

M. Greenacre. *Theory and Application of Correspondence Analysis*. London Academic Press, London, 1984. [p1, 5]

A. Israëls. *Eigenvalue Techniques for Qualitative Data*. DSWO Press, Leiden, 1987. [p9]

B. Kostov, M. Bécue-Bertaut, and F. Husson. Correspondence analysis on generalised aggregated lexical tables (CA-GALT) in the FactoMineR package. *R Journal*, 7/1:109–117, 2015. [p1]

P. Kroonenberg, M and R. Lombardo. Nonsymmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research Journal*, 34:367–397, 1999. [p3]

N. Lauro, C and L. D'Ambra. L'analyse non symmetrique des correspondances. In E. Diday, editor, *Data Analysis and Informatics III*, pages 433–446. Elsevier, Amsterdam, 1984. [p3]

S. Lé, J. Josse, and F. Husson. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008. [p1]

L. Lebart, A. Morineau, and K. Warwick, W. *Multivariate Descriptive Statistical Analysis*. John Wiley & Sons, New-York, USA, 1984. [p1, 5, 7]

G. Leibovici, D. Spatio-temporal multiway decomposition using principal tensor analysis on k-modes: the R package PTAk. *Journal of Statistical Software*, 34 (10):34 pages, 2010. [p2]

G. Leibovici, D. Principal tensor analysis on k modes. http://c3s2i.free.fr/, 2015. [p2]

R. Light, J and B. Margolin, H. An analysis of variance for categorical data. *Journal of the American Statistical Association*, 66:534–544, 1971. [p4]

M. Linting, J. Meulman, J, P. Groenen, F J, and A. Van der Kooij, J. Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap. *Psychological Methods*, 12 (3):359–379, 2007. [p5]

R. Lombardo and T. Ringrose, J. Bootstrap confidence regions in non-symmetrical correspondence analysis. *Electronic Journal of Applied Statistical Analysis*, 5:413–417, 2012. [p5, 7]

R. Lombardo, E. Beh, J, and P. Kroonenberg, M. Modelling trends in ordered correspondence analysis using orthogonal polynomials. *Psychometrika*, 81:325–349, 2016. [p2, 3, 4, 5, 10, 18]

M. Markus, T. *Bootstrap Confidence Regions in Non-Linear Multivariate Analysis*. DSWO Press, 1994. [p5]

F. Murtagh. *Correspondence Analysis and Data Coding with Java and R*. Boca Raton, FL: Chapman & Hall/CRC, 2005. [p1, 2]

O. Nenadic and M. Greenacre. Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20:1–13, 2007. [p1, 18]

B. Nieto Librero, A, P. Willems, and P. Galindo, Villardon. Canonical non-symmetrical correspondence analysis in R (1th edition, 2015). http://search.r-project.org/library/cncaGUI/html/cncaGUI-package.html, 2015. [p2]

S. Nishisato. *Multidimensional Nonlinear Descriptive Analysis*. Taylor & Francis Group, LLC, 2007. [p1, 3]

J. Oksanen, G. Blanchet, and K. *et al.* Ordination methods, diversity analysis and other functions for community and vegetation ecologists. http://cran.r-project.org,http://vegan.r-forge.r-project.org/, 2013. [p2]

J. Rayner, W C and E. Beh, J. Towards a better understanding of correlation. *Statistica Neerlandica*, 63: 324–333, 2009. [p4]

T. Ringrose, J. Bootstrap confidence regions for correspondence analysis. *Journal of Statistical Computation and Simulation.*, 83:1397–1413, 2012. [p1, 5]

B. Ripley, D, W. Venables, N, D. Bates, M, and *et al.* Functions and datasets to support venables and ripley, modern applied statistics with S (4th edition, 2002). http://cran.r-project.org/web/packages/MASS/index.html, 2002. [p1]

J. Thioulouse, D. Chessel, S. Dolédec, and M. Olivier, J. ade-4: a multivariate analysis and graphical display software. *Statistics and Computing*, 7:75–83, 1997. [p2]

W. Venables, N and B. Ripley, D. *Modern Applied Statistics with SPLUS*. 3rd edn. Springer, 1999. [p1]

Rosaria Lombardo
Department of Economics, Second University of Naples
via Gran Priorato di Malta, Capua 81043
Italy
rosaria.lombardo@unina2.it

Eric J. Beh
School of Mathematical & Physical Sciences, University of Newcastle
University Drive, Callaghan, NSW, 2308 Australia
eric.beh@newcastle.edu.au