Three-Way Correspondence Analysis in R

by Rosaria Lombardo, Michel van de Velden, and Eric J. Beh

Abstract Three-way correspondence analysis is a suitable multivariate method for visualising the association in three-way categorical data, modelling the global dependence, or reducing dimensionality. This paper provides a description of an R package for performing three-way correspondence analysis: CA3variants. The functions in this package allow the analyst to perform several variations of this analysis, depending on the research question being posed and/or the properties underlying the data. Users can opt for the classical (symmetrical) approach or the non-symmetric variant - the latter is particularly useful if one of the three categorical variables is treated as a response variable. In addition, to perform the necessary three-way decompositions, a Tucker3 and a trivariate moment decomposition (using orthogonal polynomials) can be utilized. The Tucker3 method of decomposition can be used when one or more of the categorical variables is nominal while for ordinal variables the trivariate moment decomposition can be used. The package also provides a function that can be used to choose the model dimensionality.

1 Introduction

In many applications, one encounters problems where detecting and describing the association between three categorical variables is of interest. For example, one may wish to analyse animal counts stratified by species-by-site-by-time, treatment success stratified by cure-by-therapy-by-hospital, customer satisfaction-by-service's quality-by-country, or two interacting genes in expression under the genotypes of another gene. One method specifically designed for analysing such data is three-way correspondence analysis (Carlier and Kroonenberg, 1996). For this method of analysis, a three-way contingency table is decomposed in such a way that the maximum amount of association is reflected in a low-dimensional display. Depending on the underlying data, and the research questions being asked, there are various ways to quantify and decompose the association in the table, generate a visual display of the association and calculate the accompanying numerical summaries. Hence, several variants of three-way correspondence analysis exist. Common among all the variants that we describe below is the emphasis that is placed on data exploration through the visualization of the associations.

There exists a sizable body of literature that examines the various theoretical properties and extensions of three-way correspondence analysis. For example, Kroonenberg (1989), Carlier and Kroonenberg (1996), Kroonenberg (2008, Chap. 17), Beh and Lombardo (2014, Chap. 11) and Lombardo et al. (2021) discuss a wide range of issues concerned with this technique. However, there also appears to be only a few applications that use these techniques (Carlier and Kroonenberg, 1998; van Herk and van de Velden, 2007; Lombardo et al., 2019). One reason for the lack of applications could be the absence of R software packages to perform three-way correspondence analysis.

In this paper, we introduce **CA3variants**, a comprehensive R package that allows researchers to apply variants of three-way correspondence analysis. In Section 2.2, we introduce the notation that we adopt as well as two key measures of association - Pearson's three-way phi-squared statistic and Marcotorchino's three-way index. These measures lie at the core of the three-way correspondence analysis variants that we describe below. In Section 2.3 we present three methods for decomposing a three-way contingency table, with a particular focus on the appropriateness of the different variants. In Section 2.4 we show how the two association measures above, can be partitioned in bivariate and trivariate association terms, and how can be used to define variants of three-way correspondence analysis, and we consider specific issues concerned with the visualization and selection of the dimensionality of the three-way correspondence analysis solution. In Section 2.5, we briefly review the software that is currently available for three-way analyses. In Section 2.6, we introduce our three-way correspondence analysis package, **CA3variants**, and illustrate its features and application through some illustrative examples. Some final comments are left for Section 2.7.

2 Measures of three-way association

Three-way correspondence analysis provides a numerical and graphical summary of how categories and variables are related to one another. Rather than only considering the bivariate associations between pairs of variables, three-way correspondence analysis also considers the trivariate associations (Lombardo et al., 2021).

When performing three way correspondence analysis, the dependence structure of the three categorical variables that are cross-classified to form a contingency table is analysed by considering

an appropriate measure of association. This measure can then be partitioned to reveal more detail about the nature of the association that exists between the variables. Two measures of association are implemented in the **CA3variants** package: Pearson's phi-squared statistic and Marcotorchino's index. Pearson's three-way statistic is appropriate when studying deviations from three-way independence and when the variables are symmetrically associated, while Marcotorchino's three-way index is a more suitable choice when the variables are not symmetrically associated. Depending on the choice of the measure of association used, the appropriately scaled three-way table can be decomposed into (low-dimensional) components for each of the variables. Before discussing these three-way decomposition methods, we first introduce the notation used throughout this paper. We then provide a brief description of Pearson's phi-squared statistic and Marcotorchino's index.

2.1 Notation

Suppose we have data from a sample of n subjects on three categorical variables. Such data can be represented by a three-way contingency table consisting of I rows, J columns and K tubes, where each cell value represents the count within an intersection of the levels of each of the three variables.

Denote $\underline{\mathbf{N}}$ to be the contingency table of order $I \times J \times K$ belonging to the space $\Re^{I\times J\times K}$, subscripted by i, j and k for $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, whose (i, j, k)th term is n_{ijk}, while $\underline{\mathbf{P}}$ is the table of joint relative frequencies of $\underline{\mathbf{N}}$ whose (i, j, k)th term is $p_{ijk} = n_{ijk}/n$, such that $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk} = 1$. Define $p_{i\bullet\bullet} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}$, $p_{\bullet j\bullet} = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{ijk}$, $p_{\bullet\bullet k} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ijk}$, $p_{ij\bullet} = \sum_{k=1}^{K} p_{ijk}$, $p_{i\bullet k} = \sum_{j=1}^{J} p_{ijk}$ and $p_{\bullet jk} = \sum_{i=1}^{I} p_{ijk}$ to be the univariate and bivariate marginal relative frequencies of the three-way contingency table. In addition, define $\mathbf{I}_I$ to be the identity matrix of order $I \times I$ in the space $\Re^I$, and let $\mathbf{D}_I$, $\mathbf{D}_J$, $\mathbf{D}_K$ be the diagonal matrices containing the univariate marginal relative frequencies in $\Re^I$, $\Re^J$ and $\Re^K$ whose general term is $p_{i\bullet\bullet}$, $p_{\bullet j\bullet}$ and $p_{\bullet\bullet k}$, respectively.

2.2 Pearson's three-way statistic

When the association between the categorical variables of a three-way contingency table, $\underline{\mathbf{N}}$, is considered to be symmetric, we can analyse the strength of this association using Pearson's three-way phi-squared statistic

$$\begin{aligned}\Phi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{ijk} - p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}} - 1\right)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\pi_{P_{ijk}}\right)^2.\end{aligned} \tag{1}$$

The symmetric nature of this measure implies that the three variables are all treated as predictor variables. That is, none are deemed to be dependent on the outcome of any other variable being studied. It can be shown that, under the independence assumption, Φ^2 can be partitioned as

$$\begin{aligned}\Phi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet\bullet}p_{\bullet j\bullet}\left(\frac{p_{ij\bullet} - p_{i\bullet\bullet}p_{\bullet j\bullet}}{p_{i\bullet\bullet}p_{\bullet j\bullet}}\right)^2 + \sum_{i=1}^{I}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet\bullet k}\left(\frac{p_{i\bullet k} - p_{i\bullet\bullet}p_{\bullet\bullet k}}{p_{i\bullet\bullet}p_{\bullet\bullet k}}\right)^2 \\ &+ \sum_{j=1}^{J}\sum_{k=1}^{K} p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{\bullet jk} - p_{\bullet j\bullet}p_{\bullet\bullet k}}{p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{ijk} -_{\alpha} p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^2,\end{aligned} \tag{2}$$

where

$$_{\alpha}\hat{p}_{ijk} = \hat{p}_{ij\bullet}\hat{p}_{\bullet\bullet k} + \hat{p}_{i\bullet k}\hat{p}_{\bullet j\bullet} + \hat{p}_{\bullet jk}\hat{p}_{i\bullet\bullet} - 2\hat{p}_{i\bullet\bullet}\hat{p}_{\bullet j\bullet}\hat{p}_{\bullet\bullet k}. \tag{3}$$

For further details see Carlier and Kroonenberg (1996) and Lombardo et al. (2020). Briefly, we get

$$\Phi^2 = \Phi^2_{IJ} + \Phi^2_{IK} + \Phi^2_{JK} + \Phi^2_{IJK}. \tag{4}$$

Observe that this partition also concerns Pearson's chi-squared statistic, X^2, (Lancaster, 1951; Lombardo et al., 2020) obtained by multiplying each of the terms of phi-squared in equation (4) by the sample size, n. Indeed, Pearson's chi-squared statistic is well established for testing association

between variables in contingency tables. Hence, deviations from three-way independence can be orthogonally partitioned into three deviations from independence (for each of the two-way tables formed by summing over each variable of the three-way contingency table) and a three-way association term, as it will be shown in Section 2.6.1. This partition has been extensively discussed by Carlier and Kroonenberg (1996) and more recently by Kroonenberg (2008, Chap. 17), Loisel and Takane (2016) and Lombardo et al. (2020).

2.3 Marcotorchino's three-way index

If the three categorical variables are non-symmetrically associated, or if one is interested in exploring an non-symmetric association between the variables, a more appropriate measure is the three-way Marcotorchino index. This index is defined by

$$\tau_M = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{ijk}}{p_{\bullet j\bullet}p_{\bullet\bullet k}} - p_{i\bullet\bullet}\right)^2}{1-\sum_{i=1}^{I} p_{i\bullet\bullet}^2}. \tag{5}$$

See, for example, Marcotorchino (1984a,b), Lombardo et al. (1996), Beh et al. (2007), Beh and Lombardo (2014, Section 11.4.2) and Beh and Lombardo (2021b, Section 7.5). Since the denominator of equation (5) is independent on the cell values of $\underline{\mathbf{N}}$, the numerator of the Marcotorchino index suffices as a measure of association when performing three-way correspondence analysis. This numerator measures the absolute increase in predictability of the response variable, given the predictor variables (Marcotorchino, 1985; Lombardo et al., 1996). Like Pearson's three-way phi-squared statistic, Marcortorchino's index is based on deviations from the three-way independence model. Without loss of generality, assume that the row variable is considered to be dependent on the column and tube variables. In doing so, the numerator of equation (5), which we shall simply refer to as Marcotorchino's $\tau_{M_{num}}$ statistic, is equal to

$$\begin{aligned}\tau_{M_{num}} &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{ijk}}{p_{\bullet j\bullet}p_{\bullet\bullet k}} - p_{i\bullet\bullet}\right)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\pi_{M_{ijk}}\right)^2.\end{aligned} \tag{6}$$

As in the symmetric case, an additive orthogonal partition of $\tau_{M_{num}}$ exists and is given by

$$\begin{aligned}\tau_{M_{num}} = &\sum_{i=1}^{I}\sum_{j=1}^{J} p_{\bullet j\bullet}\left(\frac{p_{ij\bullet}}{p_{\bullet j\bullet}} - p_{i\bullet\bullet}\right)^2 + \sum_{i=1}^{I}\sum_{k=1}^{K} p_{\bullet\bullet k}\left(\frac{p_{i\bullet k}}{p_{\bullet\bullet k}} - p_{i\bullet\bullet}\right)^2 \\ &+\frac{1}{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{\bullet jk} - p_{\bullet j\bullet}p_{\bullet\bullet k}}{p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^2 \\ &+\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{p_{ijk} -_{\alpha} p_{ijk}}{p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^2,\end{aligned} \tag{7}$$

where

$${}_{\alpha}p_{ijk} = \hat{p}_{ij\bullet}\hat{p}_{\bullet\bullet k} + \hat{p}_{i\bullet k}\hat{p}_{\bullet j\bullet} + \frac{\hat{p}_{\bullet jk}}{I} - \hat{p}_{i\bullet\bullet}\hat{p}_{\bullet j\bullet}\hat{p}_{\bullet\bullet k} - \hat{p}_{\bullet j\bullet}\frac{\hat{p}_{\bullet\bullet k}}{I}.$$

The partition of $\tau_{M_{num}}$ may be more simply expressed as

$$\tau_{M_{num}} = \tau_{IJ} + \tau_{IK} + \tau_{JK} + \tau_{IJK}. \tag{8}$$

Hence, like Pearson's three-way phi-squared statistic, $\tau_{M_{num}}$ (and hence the total predictability measure τ_M) is partitioned into four additive terms. The first three of these terms reflect the two-way associations and the fourth term reflects the three-way association. The first two bivariate terms of equation (8) are equal to the numerators of the Goodman-Kruskal indices (Goodman and Kruskal, 1954) between the response (row) variable and each of the two predictor (column and tube) variables, respectively. These terms are also equal to the inertias of the marginal two-way tables in classical two-way non-symmetric correspondence analysis (Lauro and D'Ambra, 1984; D'Ambra and Lauro, 1989; Kroonenberg and Lombardo, 1999; Takane and Jung, 2008). The third bivariate term of (8) is (up

to the constant $1/I$) equal to Φ^2_{JK}, which is Pearson's phi-squared statistic for the $J \times K$ contingency table formed by aggregating over the row categories. This term can be seen as a measure of the symmetric association between the two predictor variables. Finally, the last term of equation (8) is a measure of the trivariate association between the variables. Beh et al. (2007) showed that the test statistic associated with Marcotorchino's three way index is the generalization of the C-statistic (Light and Margolin, 1971), referred to here as the C_M-statistic, and is defined by

$$C_M = (n-1)(I-1)\tau_M \sim \chi^2_{\alpha,df}. \tag{9}$$

Therefore, for both Pearson's three-way chi-squared statistic and Marcotorchino's three-way τ_M statistic, under the null hypothesis of complete independence, each term of the partition is a chi-squared random variable. For further details see Light and Margolin (1971), Beh et al. (2007), Beh and Lombardo (2014, Section 11.5.2) and Beh and Lombardo (2021b, Section 7.5.2).

3 Decomposing three-way tables

The choice of which measure of association to use should be made based on the data at hand and the research question under investigation. Depending on the choice, an appropriately scaled matrix can be constructed. Three-way correspondence analysis can then be performed and involves fitting a model to the data. In particular, low-dimensional component matrices as well as a core matrix that links the different components, are fitted to the data in such a way that the sum-of-squares of the deviations between the low-dimensional approximation and the original table is as small as possible.

Several decomposition models have been proposed in the literature for three-way contingency tables. In the **CA3variants** package three types of decomposition are implemented. They are the Tucker3 model (Tucker, 1963; Kroonenberg, 1983, 2008; Kiers et al., 1992) for when all three variables are nominal, the trivariate moment decomposition (Lombardo et al., 2016b, 2021) for when all three variables are ordinal, and a hybrid decomposition for a mix of nominal and ordinal categorical variables (Lombardo and Beh, 2017). In the following subsections, we briefly review these decomposition methods and how they apply to the different variants of three-way correspondence analysis.

3.1 Tucker3 decomposition for three-way tables

For the Tucker3 decomposition, a three-way matrix $\underline{\mathbf{X}}$ with elements x_{ijk} is decomposed such that

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}a_{ip}b_{jq}c_{kr} + e_{ijk},$$

where P, Q and R ($P \leq I$, $Q \leq J$, $R \leq K$) are the fixed number of the components corresponding to the row, column and tube variables, respectively. The a_{ip}, b_{jp} and c_{kp} values are elements of the column matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, respectively, and give component loadings for the row, column and tube variables, while g_{pqr} is an element of the $P \times Q \times R$ core array. The term e_{ijk} is the error of approximation. By "flattening" the three-way matrix $\underline{\mathbf{X}}$ – for example, by concatenating the K tubes of $\underline{\mathbf{X}}$ – we can write the Tucker3 decomposition in matrix form by

$$\text{Tucker3}(\mathbf{X}) = \mathbf{AG}\left(\mathbf{B}^T \otimes \mathbf{C}^T\right) + \mathbf{E}, \tag{10}$$

where $\mathbf{X}$ and $\mathbf{G}$ are, respectively, the $I \times JK$ matrix of (flattened) data values and the $P \times QR$ matrix of core elements.

The solution to $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{G}$ is obtained by minimizing the sum-of-squares of the elements of $\mathbf{E}$ (matrix of the errors of approximation) using an alternating least-squares algorithm. The general framework of the algorithm that **CA3variants** uses is based on the Tuckals3 alternating least squares algorithm discusssed by Kroonenberg and Leeuw (1980) and Kroonenberg (1983, 1994).

Symmetric three-way correspondence analysis

For symmetric three-way correspondence analysis, the elements of Pearson's three-way phi-squared statistic are decomposed using a Tucker3 decomposition. In particular, the Tucker3 decomposition is

applied to the appropriately scaled three-way array $\underline{\mathbf{\Pi}_P}$ with elements

$$\pi_{P_{ijk}} = \frac{p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}} - 1\,, \tag{11}$$

where the component matrices, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are constrained to be orthonormal with respect to the diagonal matrices of univariate marginal relative frequencies such that

$$\mathbf{A}^T\mathbf{D}_I\mathbf{A} = \mathbf{I}_P\,, \quad \mathbf{B}^T\mathbf{D}_J\mathbf{B} = \mathbf{I}_Q\,, \quad \text{and} \quad \mathbf{C}^T\mathbf{D}_K\mathbf{C} = \mathbf{I}_R\,. \tag{12}$$

Note that the weighted sum-of-squares of the elements of $\pi_{P_{ijk}}$ is equal to Pearson's three-way phi-squared statistic; see equation (1). In other words, the symmetric variant of three-way correspondence analysis amounts to minimizing the weighted squared differences between the standardized deviations of independence in the three-way table with the approximated values using the Tucker3 model. That is:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\pi_{P_{ijk}} - \widehat{\pi}_{P_{ijk}}\right)^2\,,$$

is minimized where, for some value of P, Q and R

$$\widehat{\pi}_{P_{ijk}} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}a_{ip}b_{jq}c_{kr}\,.$$

The constraints of equation (12) are similar to the constraints used in the traditional approach to simple (two-way) correspondence analysis. Consequently, symmetric three-way correspondence analysis can be seen as a direct extension of the traditional two-way correspondence analysis approach. For more details see, for example, Carlier and Kroonenberg (1996).

Non-symmetric three-way correspondence analysis

For non-symmetric three-way correspondence analysis, one variable needs to be selected as the response variable. In the following discussion we choose, without loss of generality, the first (row) variable to serve as the response variable. When performing non-symmetric three-way correspondence analysis, we use the Tucker3 decomposition to decompose Marcotorchino's three-way $\tau_{M_{num}}$ statistic defined by equation (6). Let $\underline{\mathbf{\Pi}_M}$ represents the three-way matrix with elements

$$\pi_{M_{ijk}} = \frac{p_{ijk}}{p_{\bullet j\bullet}p_{\bullet\bullet k}} - p_{i\bullet\bullet}\,. \tag{13}$$

Non-symmetric three-way correspondence analysis is then performed by applying the Tucker3 decomposition to $\underline{\mathbf{\Pi}_M}$ where the components contained in the row, column and tube matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, are constrained to be orthornormal with respect to the weight matrices $\mathbf{I}_I$, $\mathbf{D}_J$, $\mathbf{D}_K$. That is,

$$\mathbf{A}^T\mathbf{A} = \mathbf{I}_P\,, \quad \mathbf{B}^T\mathbf{D}_J\mathbf{B} = \mathbf{I}_Q\,, \quad \text{and} \quad \mathbf{C}^T\mathbf{D}_K\mathbf{C} = \mathbf{I}_R\,.$$

Note that, for the decomposition of $\underline{\mathbf{\Pi}_M}$, these constraints ensure that the weighted quadratic norm of the low-dimensional approximation $\underline{\widehat{\mathbf{\Pi}}_M}$, can be written as

$$\|\underline{\widehat{\mathbf{\Pi}}_M}\|^2 = \tau_{M_{num}} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}^2\,.$$

3.2 Trivariate moment decomposition

Rather than considering component matrices as the Tucker3 decomposition does, the trivariate moment decomposition is based on column matrices consisting of orthogonal polynomials. The decomposition was first proposed by Beh (1998b, Chap. 7) and has since been described by, for example, Beh and Davy (1998), Lombardo et al. (2016b) and Lombardo et al. (2021, eq. 10), as an alternative method of three-way decomposition. It is particularly useful when a variable consists of ordered categories, either increasing or decreasing. The decomposition can be applied to either $\underline{\mathbf{\Pi}_P}$ or $\underline{\mathbf{\Pi}_M}$ and allows the researcher to incorporate the ordinality by replacing the Tucker3 components with the orthogonal polynomials for the ordinal variable. These polynomials are typically generated using the three-term recurrence formulae of Emerson (1968) who demonstrated their computational efficiency when compared with

the Gram-Schmidt orthogonalization process. Refer to Beh (1997, 1998a,b) and Beh and Lombardo (2021a) for a definition and properties of these polynomials when performing correspondence analysis.

To form the polynomial basis space we generate as many orthogonal polynomials as there are ordered categories. The matrix of row, column and tube orthogonal polynomials is denoted by $\boldsymbol{\mathcal{A}} = \{\alpha_{iu}\}$, (for $i = 1, \ldots, I$ and $u = 0, \ldots, I-1$), $\boldsymbol{\mathcal{B}} = \{\beta_{jv}\}$ (for $j = 1, \ldots, J$ and $v = 0, \ldots, J-1$) and $\boldsymbol{\mathcal{C}} = \{\gamma_{kw}\}$ (for $k = 1, \ldots, K$ and $w = 0, \ldots, K-1$), respectively. Like the Tucker3 components, when a symmetric variant of three-way correspondence analysis is performed, the row polynomials are orthogonal with respect to the marginal relative frequencies $p_{i\bullet\bullet}$, while the column and tube polynomials are orthogonal with respect to $p_{\bullet j\bullet}$ and $p_{\bullet\bullet k}$, respectively. In general, the first polynomial that is computed for each ordered variable of the three-way table represents the *zeroth-order* polynomial and is equal to 1 when in its normalized state. The second polynomial is the *first-order* polynomial and reflects the variation in the linearity of the categories. The third polynomial is the *second-order* orthogonal polynomial and reflects the variation in the dispersion of the categories. Higher-order polynomials represent higher-order moments of the ordered categories. These polynomials have been used extensively in the correspondence analysis literature. For more information, see, for example, Beh (1997, 1998a), Beh and Lombardo (2014, p. 94), Lombardo et al. (2016a) and Beh and Lombardo (2021b, Chap. 4).

When using the trivariate moment decomposition for symmetric and non-symmetric three-way correspondence analysis, the decomposition of the arrays $\underline{\boldsymbol{\Pi}_P}$ and $\underline{\boldsymbol{\Pi}_M}$ is defined by replacing the matrices of components **A**, **B** and **C** (see equation (10)) with their orthogonal polynomial equivalents. In particular, for the non-symmetric case and given the different row weights, we consider $\boldsymbol{\alpha}_u^* = p_{i\bullet\bullet}^{1/2}\boldsymbol{\alpha}_u$, $\boldsymbol{\beta}_v$ and $\boldsymbol{\gamma}_w$ such that

$$\pi_{M_{ijk}} = \sum_{u=0}^{U}\sum_{v=0}^{V}\sum_{w=0}^{W} \tilde{z}_{uvw}\alpha_{iu}^*\beta_{jv}\gamma_{kw}\,. \tag{14}$$

For the decomposition given by equation (14), the row polynomials are weighted such that $\sum_{i=1}^{I}\alpha_{iu}^{*2} = 1$ while the column and tube polynomials are weighted so that $\sum_{j=1}^{J} p_{\bullet j\bullet}\beta_{jv}^2 = 1$ and $\sum_{k=1}^{K} p_{\bullet\bullet k}\gamma_{kw}^2 = 1$, respectively. Note that the indices u, v, w are from 0 to U, V and W (where $U \leq I-1$, $V \leq J-1$, $W \leq K-1$), respectively, and correspond to the orders of the polynomials. The $\tilde{z}_{uvw}$ value in equation (14) is analogous to the core element g_{pqr} in the nominal case and is therefore referred to as the *polynomial core element* and is defined by

$$\tilde{z}_{uvw} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \pi_{M_{ijk}} p_{\bullet j\bullet} p_{\bullet\bullet k}\alpha_{iu}^*\beta_{jv}\gamma_{kw}\,,$$

and is of order (u, v, w). Such a term has also been referred to as a *generalized correlation*. See, for example, Rayner and Beh (2009), Beh and Lombardo (2014, Chap. 6) and Beh and Lombardo (2021b, Chap. 5). Observe that, unlike the Tucker3 decomposition given by equation (10), the trivariate moment decomposition has a closed form that justifies the absence of the error of approximation in equation (14).

3.3 Hybrid decomposition for nominal and ordinal variables

The hybrid decomposition involves computing Tucker3 components for the nominal variables, and orthogonal polynomials for the ordinal variables (Lombardo and Beh, 2017; Lombardo et al., 2021). Generally for the analysis of three-way contingency tables, we distinguish the following two cases: 1) there are two ordinal variables and one nominal variable, and 2) there are two nominal variables and only one ordinal variable. Suppose we consider the case where we have a three-way contingency table in which the row and column variables are ordinal and the tube variable is nominal. Then the *hybrid decomposition*, for case 1, involves calculating the polynomials for the row and column variables and the Tucker3 components for the nominal tube variable. When the row variable is treated as a response variable, three-way non-symmetric correspondence analysis can be performed using the hybrid decomposition of $\pi_{M_{ijk}}$ such that

$$\begin{aligned}\pi_{M_{ijk}} &= \widehat{\pi}_{M_{ijk}} + e_{ijk} \\ &= \sum_{u=0}^{U}\sum_{v=0}^{V}\sum_{r=1}^{R} z_{uvr}\alpha_{iu}^*\beta_{jv}c_{kr} + e_{ijk}\,.\end{aligned} \tag{15}$$

Here $\boldsymbol{\alpha}_u^*$ and $\boldsymbol{\beta}_v$ are the uth order row and vth order column polynomials, respectively, while $\boldsymbol{c}_r$ is the rth tube (Tucker3) component. The value of z_{uvr} in equation (15) is defined by

$$z_{uvr} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \pi_{M_{ijk}} p_{\bullet j\bullet} p_{\bullet\bullet k} \alpha_{iu}^* \beta_{jv} c_{kw} \,,$$

and is referred to as the *hybrid core element* of order (u, v, r). While the number of orthogonal polynomials for the rows and columns should always be equal to the number of categories that define the variable (see Section 2.3.2), the number of Tucker3 components for the tube variable can be smaller ($R \leq K$). A complete orthogonal decomposition is always used when all the three variables are ordered, as it is for equation (14), but is seldom used in practice when the variables are not all ordered. Like the Tucker3 decomposition (see equation (10)) and unlike the trivariate moment decomposition (see equation (14)), the hybrid decomposition given by equation (15) includes the error of approximation, e_{ijk}, because the decomposition no longer has a closed form solution because of the presence of the Tucker3 components.

4 Three-way correspondence analysis variants

Combining the two measures of three-way association described in Section 2.2 with the three methods for decomposing three-way tables outlined in Section 2.3 gives four variants of three-way correspondence analysis:

- Symmetric three-way correspondence analysis: this analysis is based on the partition of Pearson's three-way phi-squared statistic and the Tucker3 decomposition of $\underline{\mathbf{\Pi}_P}$. It executes three-way correspondence analysis by treating all variables symmetrically and corresponds to the analysis described by Carlier and Kroonenberg (1996).
- Non-symmetric three-way correspondence analysis: this corresponds to partitioning Marcotorchino's three-way statistic and applies a Tucker3 decomposition to $\underline{\mathbf{\Pi}_M}$. In this analysis, one of the three variables is treated as a response variable and the other two are treated as predictor variables (Lombardo et al., 1996).
- Ordered symmetric three-way correspondence analysis: for this analysis, either the trivariate moment decomposition (if all variables are ordinal) or the hybrid decomposition (if one or two of the three variables are ordinal) is applied to $\underline{\mathbf{\Pi}_P}$ leading to the partition of Pearson's three-way phi-squared statistic (Lombardo et al., 2021).
- Ordered non-symmetric three-way correspondence analysis: this analysis is based on the trivariate moment decomposition (if all variables are ordinal) or the hybrid decomposition (if one or two of the three variables are ordinal) of $\underline{\mathbf{\Pi}_M}$ and leads to the partition of Marcotorchino's three-way index. Hence, in this analysis one variable is treated as the response variable and the other two are treated as predictor variables.

These four variants of three-way correspondence analysis are incorporated into the **CA3variants** package.

4.1 Visualizing three-way correspondence analysis solutions

Like the traditional approach to two-way correspondence analysis, visualization in three-way correspondence analysis is an important feature and helps to provide a descriptive analysis of the data. To visually display the (symmetric or non-symmetric) association that exists among the variables we consider the *interactive biplot* (Carlier and Kroonenberg, 1996), also called as *nested biplot* by Kroonenberg (2008, p. 441). In the interactive biplot, the categories of one variable, referred to as a *reference variable*, are jointly visualized with all pair-wise combinations of the categories of the other two variables. Hence, depending on the choice of reference variable, we can distinguish three different *interactive coordinates*: row-column, row-tube and column-tube interactive coordinates.

To see how this works, and why the resulting visualizations are indeed biplots, note that all four three-way correspondence analysis variants described in Section 2.4 yield three sets of "coordinates" (one for each variable), as well as an array of core elements that describe the strength of association between these values. Differences between variants can be described in terms of the different measures of association under consideration (i.e., Pearson's three-way phi-squared statistic or Marcotorchino's index), the orthogonalization constraints adopted, or the type of decomposition (i.e., Tucker3, trivariate moment decomposition or hybrid decomposition) used.

Recall that the general form of the Tucker3 decomposition is given by equation (10). As we described above, this matrix formulation is based on a "flattened" version of the three-way matrices

that involves the concatenation of the categories of a variable. In fact, the concatenation of a variable that leads to the $P \times Q \times R$ approximation can be seen in the following three ways

$$\begin{aligned} \mathbf{X}_{JK,I} &= (\mathbf{B} \otimes \mathbf{C})\, \mathbf{G}_1 \mathbf{A}^T \\ \mathbf{X}_{IK,J} &= (\mathbf{A} \otimes \mathbf{C})\, \mathbf{G}_2 \mathbf{B}^T \\ \mathbf{X}_{IJ,K} &= (\mathbf{A} \otimes \mathbf{B})\, \mathbf{G}_3 \mathbf{C}^T . \end{aligned} \tag{16}$$

Note that these arrangements have no influence on the approximated values of $\underline{\mathbf{X}}$. The subscripted and flattened $\mathbf{G}$'s indicate that, although their elements are the same, the organization differs between them. Each of the formulations in equation (16) constitutes a biplot. To show this, suppose we consider the decomposition of $\mathbf{X}_{JK,I}$. Then the rows of $(\mathbf{B} \otimes \mathbf{C})\, \mathbf{G}_1$ are the principal coordinates of the pairwise combinations of columns and tubes categories of $\underline{\mathbf{X}}$. Hence, plotting these jointly with the row standard coordinates contained in the rows of $\mathbf{A}$ provides the analyst with a biplot interpretation of the association. For an extensive discussion of biplot interpretations in the context of correspondence analysis see, for example, Greenacre (2010) and Gower et al. (2011, Chapters 7 and 8). For a more general treatment of data visualizations in dimension reduction methods, see Gower et al. (2014).

For each approximation in equation (16), the interactive coordinates can be expressed in either their standard or principal form (whose features are the same of those derived for biplots in the classical approach to correspondence analysis) and so leads to two types of *interactive biplots*:

- For the first type of interactive biplot, we can factorize each equation in such a way that the categories for the non-interactive variable are displayed using standard coordinates so that they are orthonormal with respect to the appropriate metric. Therefore, observing the combination of categories from the other two variables (which constitutes the "interactive" structure of the categories) are defined using principal coordinates (Kroonenberg, 2008, p. 273). Algebraically, this choice simply means that the interactive coordinates are a form of principal coordinates. They are calculated from the Kronecker product of two component matrices (for example, $\mathbf{B}$ and $\mathbf{C}$) multiplied by the appropriate $\mathbf{G}$ matrix (for example, $\mathbf{G}_1$). When displaying the standard coordinates of the non-interactive variable, the points are often displayed as a projection from the origin to their position defined by their standard coordinate.
- For the second type of interactive biplot, the $\mathbf{G}$ matrix is applied to the non-interactive variable. Hence, the categories for this variable are displayed in terms of their principal coordinates while the coordinates corresponding to the combination of categories from the other two variables (i.e., the interactive coordinates) are depicted as standard coordinates (Lombardo et al., 2021).

4.2 Selecting the number of components

The three-way decompositions described in Section 2.3 require a chosen number of components (P, Q and R) for each of the variables of $\underline{\mathbf{N}}$. A common approach is to consider various solutions for the components, resulting in different values of dimensionality (i.e., values of P, Q and R) and then inspect their appropriateness using a goodness-of-fit (or a lack-of-fit) measure with respect to the degrees-of-freedom of the approximation obtained from these solutions. By increasing the number of components the model becomes more complex but the goodness-of-fit of the model improves. Hence, by considering a goodness- (or lack-) of-fit measure for different model complexities, the trade-off between model fit and model complexity can be assessed.

Unfortunately, there is no "best" way to determine the optimal trade-off between model fit and model complexity. Often, the choice of what dimensionality to select is made by visually inspecting a plot of the goodness-of-fit against the degrees-of-freedom of the model. One such plot is a scree-like plot and selecting the desired dimensionality is made by using a variety of strategies including simply looking for an "elbow". One may also select the dimensionality by observing where the "elbow" lies in the lower boundary of the convex hull (Kroonenberg and Oort, 2003; Murakami and Kroonenberg, 2003; Kroonenberg, 2008). Scree-like plots can also be considered by using a measure of goodness- (or lack-) of-fit on the y-axis and the degrees of freedom (or the number of free parameters) on the x-axis. In this case, the analyst selects a model on or close to the "elbow" near the upper boundary of the convex hull (Timmerman and Kiers, 2000; Ceulemans and Kiers, 2006).

To aid in the visual detection of an "elbow" in the convex hull, Ceulemans and Kiers (2006) introduce the *st*-criterion which looks at the smallest angle on the convex hull and allows one to choose a model on the higher (lower) boundary of the convex hull, with the best balance of goodness-(or lack-) of-fit and df (or free parameters). Given the goodness-fit-value, f, and the model complexity-value, df, the st criterion for a model of dimensionality l can be written as

$$st(l) = \left(\frac{f(l) - f(l-1)}{df(l) - df(l-1)} \right) / \left(\frac{f(l+1) - f(l)}{df(l+1) - df(l)} \right) . \tag{17}$$

The number of models to consider when constructing the convex hull depends on the choice of dimensionality, l, made, since there are as many models available to consider as there are combinations of the three dimensions.

In addition to evaluating the goodness-of-fit for the different models, it may also be insightful to asses how stable models of certain dimensionalities are. This can be done by first applying re-sampling procedures to the three-way tables and then considering the resulting convex hulls. Several ways to facilitate such an assessment have been implemented in the **CA3variants** package. More details of the relevant functions and options can be found in Section 2.6.

5 Related software

Currently, there are no packages available in R devoted to three-way correspondence analysis. However, the R packages **PTAk** (Leibovici, 2010), **ThreeWay** (Giordano et al., 2014), **rTensor** (Li et al., 2018), **multiway** (Eilers, 2019), **psych** (Revelle, 2018), **tensorA** (Statnikov, 2018), **mvoutlier** (Zhou, 2019) and **irlba** (Hoffman, 2017) can be used to perform several different three-way decompositions, including the Tucker3 decomposition. An overview of the areas of data analysis that these packages cover is summarised in Table 1.

A complete three-way methods program is also available in Pieter Kroonenberg's Fortran package 3WayPack and includes functionality to perform a multi-way correspondence analysis; see `http://three-mode.leidenuniv.nl/` of the *The Three-Mode Company*. Similarly, an extensive collection of three-way methods and decomposition tools are available for MATLAB through the N-Way Toolbox (Bro, 2020). However, while these packages can be used to calculate solutions for the three-way correspondence analysis variants based on the Tucker3 decomposition, doing so requires some non-trivial data preparation and output processing steps.

Table 1: R packages for three-way data analysis. CA3: symmetric three-way correspondence analysis; NSCA3: non-symmetric three-way correspondence analysis; OCA3: ordered symmetric three-way correspondence analysis; ONSCA3: ordered non-symmetric three-way correspondence analysis; PCA3: three-way principal component analysis

	Three-way Data Analysis				
package	CA3	NSCA3	OCA3	ONSCA3	PCA3
CA3variants	x	x	x	x	
ThreeWay					x
PTAk	x				x
rTensor					x
multiway					x
psych					x
tensorA					x
mvoutlier					x
irlba					x

The R package **CA3variants** provides a straightforward way to perform the different variants of three-way correspondence analysis described above on a three-way contingency table. Moreover, in addition to the Tucker3 variants of three-way correspondence analysis, the package also allows for the application of trivariate moment decomposition and hybrid decomposition methods, suitable when variable categories are ordered.

6 CA3variants: Package description and examples

In this section, we introduce the main functions, arguments and options available in the **CA3variants** package. These functions are `tunelocal()` and `CA3variants()`.

The `tunelocal()` function can be used to determine an appropriate number of dimensions in the approximation of $\underline{\mathbf{\Pi}_P}$ or $\underline{\mathbf{\Pi}_M}$, while the function `CA3variants()` can be used to perform all four methods described in Section 2.4. Some similarities and differences of these four methods are summarized in Table 2.

The `CA3variants()` and `tunelocal()` functions return S3 objects from which the `plot()`, `print()`

and summary() functions are available. Note that both functions require the input arguments Xdata, ca3type, resp and norder. Respectively, these arguments specify the three-way data, the type of analysis being performed (which can be chosen from those outlined in Section 2.4), the response variable (in the case of a non-symmetric variant) and the number of ordinal variables (when an ordered variant is performed). Xdata can be a three-way table, or an $(n \times 3)$ data matrix where the rows represent the n observations/objects and the 3 columns correspond to three categorical variables, i.e. the row, column and tube variables (the levels/categories of each variable are given by integer numbers).

The tunelocal() function can help the user to choose an appropriate number of dimensions for any variant of three-way correspondence analysis. A list detailing the fit of all of the models considered can be obtained using print(tune.out); here tune.out is the output object produced using the tunelocal() function. This function considers the decompositions of the original data for all triplets of dimensions. The stability of the fit of the solutions for different dimensionalities can also be assessed by adding arguments related to the implementation of three resampling schemes (when 'boots = TRUE' and 'nboots = 100'). The available schemes are a non-parametric bootstrap resampling method or a parametric bootstrap method using one of two distributions (multinomial or Poisson). The parametric bootstrap can be considered as a *simple* parametric bootstrap ('boottype = "bootpsimple"') when the row, column and tube marginals are fixed to equal those of the original three-way table. Alternatively, it can be performed using a stratified parametric bootstrap method ('boottype = "bootpstrat"') where the row and column marginals are fixed for each tube (for $k = 1, ..., K$) of the original three-way table.

Differently from tunelocal(), another important argument of CA3variants() is dims. The argument dims defines the dimensionality of the solution which can be driven by first using tunelocal(). The available variants for ca3type are:

- ca3type = "CA3" for symmetric three-way correspondence analysis. This option is appropriate when all variables are assumed, or known, to be nominal and symmetrically associated. This is also the default analysis that is performed.
- ca3type = "NSCA3" for non-symmetric three-way correspondence analysis. This option is appropriate when one of the variables is defined as the response variable which can be chosen by specifying resp = "row" (the default choice), resp = "column" or resp = "tube". All three variables are treated as being nominal.
- ca3type = "OCA3" for three-way ordered symmetric correspondence analysis. This option is appropriate when at least one of the three variables consists of ordered categories.
- ca3type = "ONSCA3" for three-way ordered non-symmetric correspondence analysis. This option is appropriate when at least one of three variables consists of ordered categories and one of the variables is defined as the response variable. The analyst can specify the response variable in the same way that the response variable is defined for non-symmetric three-way correspondence analysis (see ca3type = "NSCA3").

Method	Variables	Association	Decomposition method
ca3type = "CA3"	nominal	symmetric	Tucker3
ca3type = "NSCA3"	nominal	non-symmetric	Tucker3
ca3type = "OCA3"	ordinal	symmetric	Trivariate moment
ca3type = "ONSCA3"	ordinal	non-symmetric	Trivariate moment
ca3type = "OCA3"	one or two variables are ordinal	symmetric	Hybrid
ca3type = "ONSCA3"	one or two variables are ordinal	non-symmetric	Hybrid

Table 2: Similarities and differences of three-way correspondence analysis methods in CA3variants().

Finally, the package contains four example data sets that can be used to test and benchmark the different methods, all with varying features and variable structures. They are: happy - a $4 \times 6 \times 4$ contingency table with $n = 40323$ - (Davis, 1977), happyNL - a $4 \times 5 \times 4$ contingency table with $n = 1669$ - (from the European Social Survey of 2016, http://www.europeansocialsurvey.org/), museum - a 253×3 data matrix with $n = 253$ - (from a 2019 survery promoted by the University "Luigi Vanvitelli", Italy), and ratrank - a $9 \times 9 \times 5$ contingency table with $n = 44568$ - (van Herk and van de Velden, 2007). In Section 2.6.2 we illustrate the package by performing a NSCA3 on the data set happyNL, while Sections 2.6.1 and 2.6.3 consider two symmetric analyses of the ratrank data set (a nominal three-way correspondence analysis and a hybrid three-way analysis).

6.1 Three-way symmetric correspondence analysis: Ranking and rating data

The dataset `ratrank` is one of the four datasets included in the **CA3variants** package. It is a data array of size $9 \times 9 \times 5$ that is formed from the cross-classification of the *Rating* (row), *Ranking* (column), and *Country* (tube) variables, and was analyzed by van Herk and van de Velden (2007).

Participants from five European countries were asked to rate and rank the same nine values taken from the list of values (LOV) described by Kahle (1983). For each of these European countries, a contingency table was constructed with counts of co-occurrences of rating numbers and rankings. The ranking task required participants to provide a strict ranking of the items. In the rating task, participants are asked to provide ratings (on a 9 point scale) to the same items. It gives the participants the freedom to rank the items in any way they desire, however, it is also open to response tendencies. Such tendencies can be referred to as response styles. For example, some individuals may be more inclined to use extreme ratings (lowest or highest) where others only use middle ratings to express their preferences. The observed correspondence between the ratings and rankings could then be used to inspect response tendencies and, in this study, to relate such tendencies to nationalities. For more details on the data and the theory underlying the response tendencies, we refer to the van Herk and van de Velden (2007). Our objective here is to illustrate the application of the **CA3variants** package by reproducing some of the results published in their paper. After downloading and installing **CA3variants** from the Comprehensive R Archive Network (CRAN), we load the package:

```
library("CA3variants")
```

Dimensionality of the solution

We use the `tunelocal()` function to determine an appropriate triplet of dimensions:

```
tune.ca3.out <- tunelocal(ratrank, ca3type = "CA3")
print(tune.ca3.out)
plot(tune.ca3.out)
```

The function `tunelocal()` yields an object containing goodness-of-fit measures, model complexity and, when `boots = TRUE`, the bootstrap samples used. However, using `print(tune.ca3.out)` we show the following numerical results for the models on the boundary:

```
#> # Convex hull (upper bound)

#> # Selected model(s):
#>           complexity      fit
#> c(2, 2, 1)          1 17726.27


#> All models on upper bound:
#>            complexity      fit        st
#> c(1, 6, 1)          0 14265.56        NA
#> c(2, 2, 1)          1 17726.27 16.097440
#> c(3, 3, 1)          4 18371.23  3.116993
#> c(3, 3, 2)         12 18923.00  1.233506
#> c(3, 4, 2)         17 19202.58  1.846595
#> c(3, 4, 3)         28 19535.67  1.650064
#> c(4, 4, 3)         39 19737.53  1.291062
#> c(4, 4, 4)         54 19950.73  1.538958
#> c(5, 5, 4)         88 20264.76  1.368834
#> c(6, 6, 4)        130 20548.15  1.497295
#> c(7, 7, 4)        180 20773.47        NA
```

The numerical and graphical output of `tune.ca3.out` show that an appropriate triplet of dimensions is (2, 2, 1). Note that Figure 1 is generated when using `plot(tune.ca3.out)`. While cluttered, as the result of the large number of triplets that can be considered in the analysis of `ratrank`, Figure 1 also shows that an appropriate dimensionality of the solution is (2, 2, 1). Each point in Figure 1 corresponds to a combination of dimensions. The y-axis gives the goodness-of-fit measure for each model which we use as the criterion for choosing the most appropriate dimensionality for the solution. More complex models that involve higher dimensionalities (or, equivalently, higher degrees of freedom) have a better fit. The red line outlines the convex hull where models on this line are superior to higher dimensional options with a similar fit. For example, in Figure 1, for the models that lie below the red line there is typically an alternative, less complex, model that achieves the same fit. Alternatively,

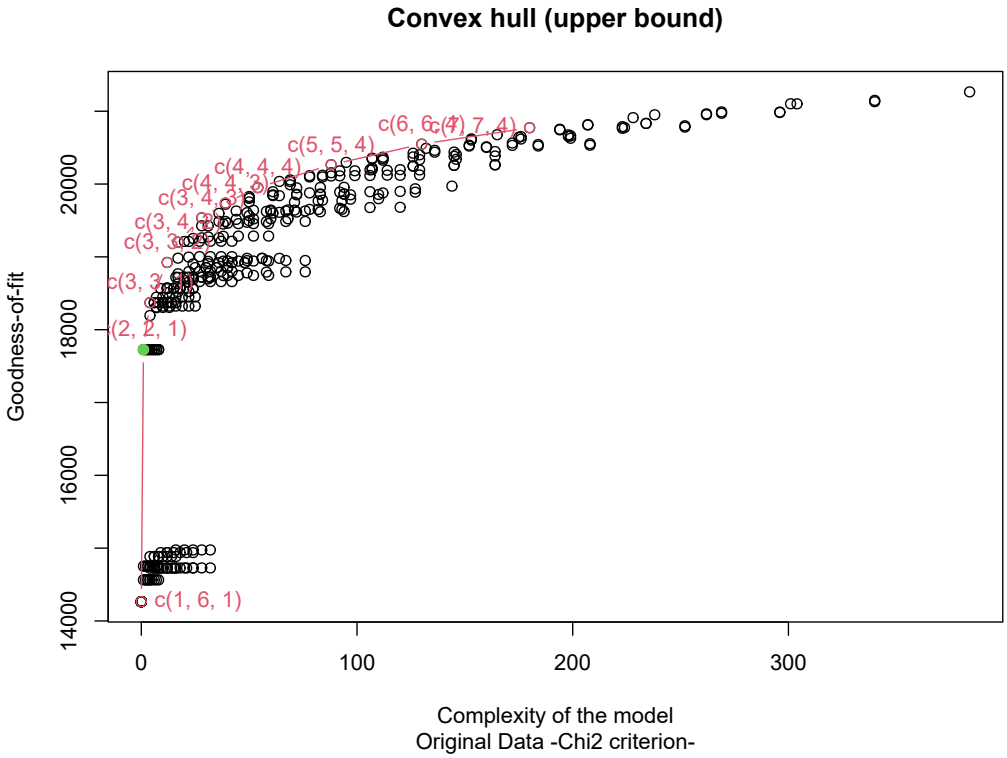

Figure 1: Model fit versus complexity for three-way nominal CA of the `ratrank` data.

there is also an equally complex model having a better fit. In three-way correspondence analysis, like other dimension reduction methods, users can factor in subjective criteria (such as interpretability) when selecting the dimensionality of a model. Here, in accordance with Ceulemans and Kiers (2006), we follow the *st* criterion and select the model, marked green in Figure 1, with two dimensions for the row (*Rating*) and column (*Ranking*) variables and one for tube (*Country*) variable.

Numerical summary of the association

By following the analysis described in van Herk and van de Velden (2007) we perform a symmetric three-way correspondence analysis, the default method, on the `ratrank` data. Following the output of the tunelocal function and in accordance with van Herk and van de Velden (2007), we specify two dimensions for the row (*Rating*) and column (*Ranking*) categories, and a single dimension for the tube (*Country*) categories. This can be achieved by:

```
ca3.out <- CA3variants(ratrank, dims = c(2, 2, 1))
```

The `print()` function returns several key measures of association that are included in this output. These include the percentage of explained inertia along each dimension, the partition of Pearson's three-way chi-squared and phi-squared statistics into four terms (see equation (4)), the corresponding degrees of freedom, the p-value, and the relative sizes of each term of the partition (allowing for comparisons between chi-squared values from different, asymptotic chi-squared distributions):

```
print(ca3.out)

#> # Percentage contributions of the components to the total inertia for column-tube
biplots

#>     p1     p2
#> 67.008 16.347

#> # Percentage contributions of the components to the total inertia for row-tube
biplots
```

```
#>     q1     q2
#> 67.008 16.347

#> # Percentage contributions of the components to the total inertia for row-column
biplots

#>     r1
#> 83.355

#> #  Index partition

#>                Term-IJ Term-IK Term-JK Term-IJK Term-total
#> Chi-squared  18359.272 589.605 254.629 2062.404  21265.910
#> Phi-squared      0.412   0.013   0.006    0.046      0.477
#> % of Inertia    86.332   2.773   1.197    9.698    100.000
#> df              64.000  32.000  32.000  256.000    384.000
#> p-value          0.000   0.000   0.000    0.000      0.000
#> X2/df          286.864  18.425   7.957    8.056     55.380
```

This output shows that the Pearson chi-squared statistic of `ratrank` is 21265.91 and, with a p-value that is less than 0.0001, there is a statistically significant association between at least two of the variables of the data set. Further insight into the nature of the association can be obtained from the terms of the partition of the overall chi-squared. The output shows that the most dominant source of association exists between the *Rating* and *Ranking* variables (`Term-IJ`), and contributes to 18359.27, or 86.33%, of the total association among the three variables. The association between the *Rating-Country* variables (`Term-IK`) and *Ranking-Country* variables (`Term-JK`) accounts for relatively little in comparison (2.77% and 1.20%, respectively), but are still statistically significant sources of association. The association among all three variables (`Term-IJK`) contributes to the remainder (or nearly 10%) of the association between the variables. Further information about the nature of the association can be obtained visually by performing a correspondence analysis.

Visual summary of the association

To reproduce the results from van Herk and van de Velden (2007), we consider here the row-tube (*Rating - Country*) interactive biplot, so that the interactive row-tube points are plotted using principal coordinates. This biplot is given by Figure 2 and is produced from the command:

```
plot(ca3.out, biptype = "row-tube", addlines = F)
```

By default, the `plot()` function uses a straight line from the origin to each standard coordinate to depict the non-interactive variable. However, with so many points in Figure 2, adding projection lines for each of the nine *Ranking* categories leads to a cluttered plot. Hence, and in accordance with van Herk and van de Velden (2007), we use `addlines = F` to remove the lines. Furthermore, to control the size of the points and their labels, the `plot()` function uses two arguments `size1` and `size2` (for the points and labels, respectively); by default `size1 = 1` and `size2 = 3`.

Finally, to avoid any further clutter of points close to the origin, a scaling argument can be used that helps to reveal important features of the association without impacting the approximation. The default for this scaling argument, which was applied here, is set such that the average sum of squares for the two sets of points is the same, and is thus in accordance with the recommendations given by Gower et al. (2010) and van de Velden et al. (2017). This default can be overwritten by specifying a value for the `scaleplot` argument in the `plot()` function. Note that, except for this scaling, the biplot given by Figure 2 is identical to Figure 1 in van Herk and van de Velden (2007).

Figure 2 shows that the highest value rank ("rank9") generally receives the highest possible rating ("9") across all five countires. However, for the second highest value rank ("rank8") the ratings tend to vary from 4 to 8, showing some heterogeneity in how "rank8" is perceived in terms of the *Rating* categories. For the lowest valued rank ("rank1"), we see a clear association with the lowest rating ("1"). However, this level of rating is also often linked to items that received a rank of "2". Moreover, for the items that receive a rank of "1" up to "7", we see that individuals tend to assign to them a rating of between "1" and "3" (inclusive). Finally, each of the ratings appear rather homogeneous across all five countries. However, with ratings from Germany being consistently furthest from the origin, and those from the United Kingdom being closest to the origin, these *Country* categories provide, relatively speaking, the strongest and weakest (respectively), contribution to the association. See van Herk and van de Velden (2007) for a more in-depth analysis and explanation of this analysis.

6.2 Three-way non-symmetric correspondence analysis: Happiness data

Since 1977, the study of the relationship between happiness, household characteristics and education using data obtained from social survey data has received a great deal of attention. For an analysis of this data set, see, for example, Davis (1977), Clogg (1982), Beh and Davy (1998) and Kroonenberg (2008, Chap.17). Davis' data set Davis (1977) examines the association between happiness, number of siblings and years of schooling completed of 1517 individuals and is included in **CA3variants** as `happy`. Kroonenberg (2008, Chap.17) studied Davis' data by performing a symmetric three-way correspondence analysis.

Following on from those studies mentioned above, we analyze a three-way contingency table, obtained from the 2016 European Social Survey (http://www.europeansocialsurvey.org/). It involves a sample of 1669 respondents from the Netherlands and investigates the association between their reported level of *Happiness*, the level of *Education* and the number of people in their *Household*. As an illustration of one of the three-way variants implemented in the **CA3variants** package, we now turn our attention to performing the non-symmetric variant of three-way correspondence analysis.

Defining the variables

To assess the level of *Happiness* of a respondent, people were asked to reply to the question:

> *"Taking all things together, how happy would you say you are?"*

Responses were made on a scale from 1 ("extremely unhappy") to 10 ("extremely happy"). After observing the distribution of counts in the data, we re-coded these scales into the following four

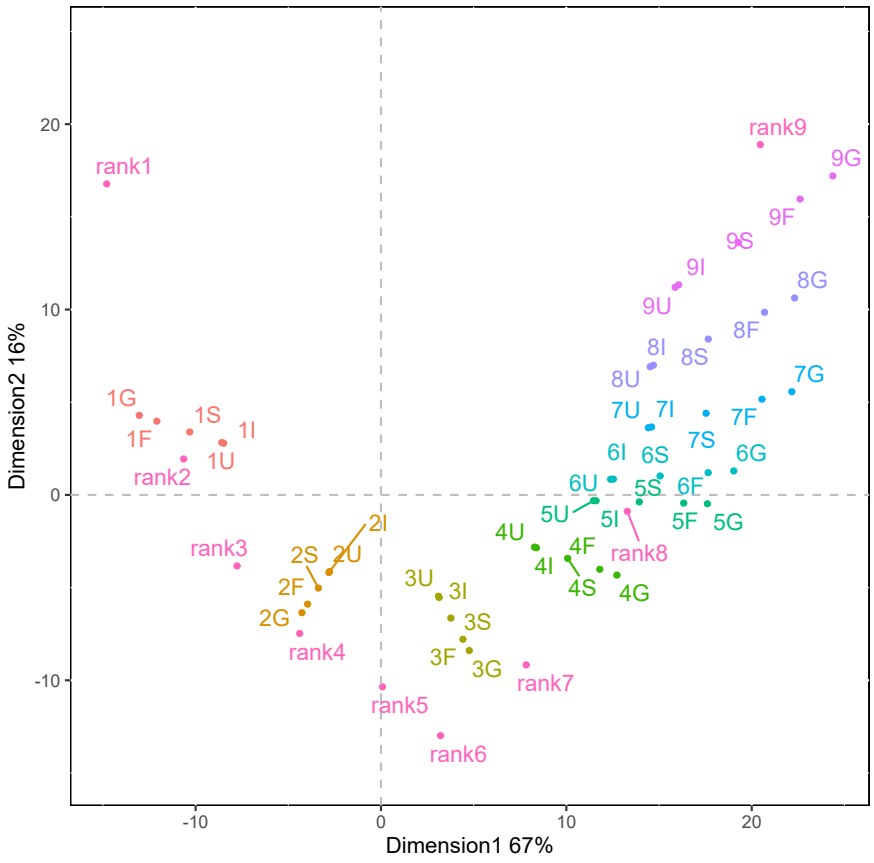

Figure 2: Interactive row-tube biplot for `ratrank`. The column points (*Ranking* categories) are depicted using standard coordinates and are labeled as "rank1" to "rank9". The interactive row-tube points (*Rating-Country* categories) are depicted using principal coordinates and are labelled by the rating number (1 to 9) followed by the first letter of the country: France (F), Germany (G), Italy (I), Spain (S) and the United Kingdom (U).

categories of *Happiness*: "low" (for ratings < 6), "middle" (for ratings between $6 - 7$), "high" (for a rating equal to 8), and "very-high" (for ratings > 8).

The *Education* variable is defined using four categories. These are "Less than lower secondary education" (coded "ED1"), "Lower secondary education completed" (coded "ED2"), "Upper secondary education completed" (coded "ED3") and "Post-secondary and/or tertiary education completed" (coded "ED45").

Finally, for the *Household* variable, the respondents were asked to reply to the question:

> *"Including yourself, how many people - including children - live here regularly as members of this household?"*

The four categories from this question were defined as follows: a one person household is coded "HS1", a two person household is coded "HS2", a three person household is coded "HS3", a four person household is coded "HS4", a five person household is coded "HS5' and a household containing more than five people is coded ">HS5".

The cross-classification of the *Happiness*, *Education* and *Household* variables forms a three-way contingency table which has been included in the package with the object name `happyNL`.

For our analysis of this contingency table, we consider the non-symmetric three-way correspondence analysis variant with the row variable (*Happiness*) treated as the response variable, and the column (*Education*) and tube (*Household*) variables defined as the predictor variables.

Dimensionality of the solution

Before performing a three-way NSCA on `happyNL` we first need to determine the dimensionality of the solution. This can be done by comparing the fit and complexity of models of different dimensionality using the `tunelocal()` function. For this example, we consider decompositions applied to 100 resampled data tables (using the parametric bootstrap; the default), and calculate, for each triplet of dimensions, the mean goodness of fit over the bootstrap samples. Note that, by doing so, the overall number of estimated models equals $I \times J \times K \times nboots = 80 \times 100 = 8000$. All resampled data tables are collected in the object named 'XG' of the output of the `tunelocal()` function:

```
tune.nsca3.out <- tunelocal(happyNL, ca3type = "NSCA3", resp = "row", boots = T)
plot(tune.nsca3.out)
```

The resulting plot is given as Figure 3. Each point in this plot corresponds to a combination of dimensions. The y-axis gives the goodness-of-fit measure for each model. More complex models, that is those involving higher dimensionalities have a better fit. The red line denotes the convex hull where models on this line are superior to higher dimensional options with a similar fit. For example, in Figure 3, for the models which are below the red line there is an alternative, less complex, model achieving the same (or similar) fit, or there is an equally complex model having a better fit. In three-way correspondence analysis, as with other dimension reduction techniques, users can factor in subjective criteria such as interpretability when selecting the dimensionality of a model. Here, in accordance with Ceulemans and Kiers (2006), we follow the *st* criterion and select the model, marked green in Figure 3, with two dimensions for the row (*Education*) column (*Household*) and tube (*Happiness*) variables.

Using `print(tune.nsca3.out)` we obtain the numerical results (i.e. fit) for the models that lie along the red line in Figure 3. The output from this command is:

```
#> # Note that when boots = T,  the data samples generated
#> # are given in the object named 'XG'

#> # Results for choosing the optimal model dimension

#> # Convex hull (upper bound)

#> # Selected model(s):
#>           complexity      fit
#> c(2, 2, 2)          4 187.6015


#> # All models on upper bound:
#>           complexity      fit       st
#> c(1, 1, 4)          0 111.6055       NA
#> c(1, 2, 2)          1 145.6444 2.433839
```

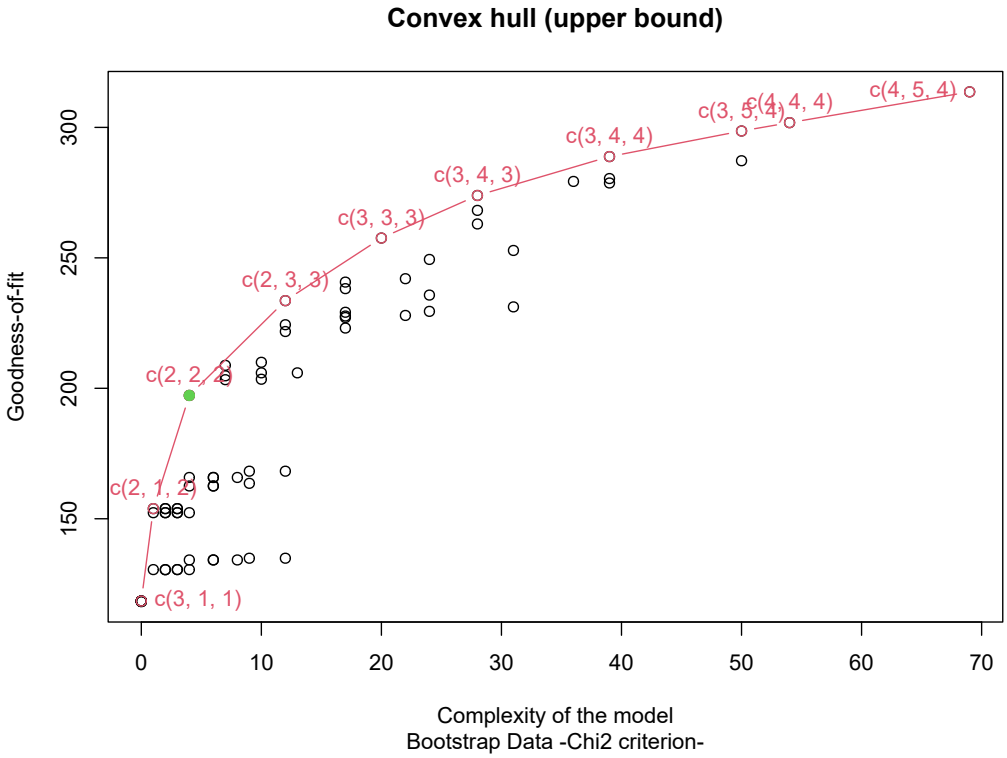

Figure 3: Model fit versus complexity for three-way NSCA of the happiness bootstrapped data.

```
#> c(2, 2, 2)          4 187.6015 2.965930
#> c(2, 3, 3)         12 225.3251 1.766711
#> c(3, 3, 3)         20 246.6776 1.382302
#> c(3, 4, 3)         28 262.1246 1.246718
#> c(3, 4, 4)         39 279.1611 1.734889
#> c(3, 5, 4)         50 288.9810 1.190973
#> c(4, 5, 4)         69 303.2229       NA
```

Numerical summary of the association

The `CA3variants()` function can be used to perform a three-way non-symmetric correspondence analysis on `happyNL` by specifying the arguments of the function so that they define the data table, the dimensionality of the solution, the type of analysis and the response variable. Here, using the suggested dimensions, the analysis is performed so that:

```
nsca3.out <- CA3variants(happyNL, ca3type = "NSCA3", resp = "row",
dims = c(2, 2, 2))
```

The numerical output from this analysis is obtained using `print(nsca3.out)`:

```
print(nsca3.out)

#> # Percentage contributions of the components to the total inertia for pred biplots

#>    p1     p2
#> 47.407 22.696

#> #  Index partition

#>                Term-IJ Term-IK Term-JK Term-IJK Term-total
#> Tau Numerator    0.014   0.005   0.008    0.006      0.032
#> Tau              0.021   0.007   0.012    0.008      0.047
#> % of Inertia    43.162  14.719  24.749   17.371    100.000
```

```
#> CM-Statistic    102.583  34.981  58.819   41.285    237.669
#> df               12.000   9.000  12.000   36.000     69.000
#> p-value           0.000   0.000   0.000    0.251      0.000
#> CM-Statistic/df   8.549   3.887   4.902    1.147      3.444
```

By reducing the dimensionality of the solution to (2, 2, 2), the first set of values (`p1` and `p2`) are the percentages of the total association which is explained by the two axes of a biplot; we will speak more on these two values shortly. The summary of values that follow `Index partition` gives the four terms of the partition of the Marcotorchino index, its numerator and its associated test statistic, C_M-statistic (see Section 2.2.3). Note that the last column, labeled `Term-total` corresponds to the three-way index being partitioned. Consequently, the seven rows of this output summarize the elements of each term of this partition, including their p-value and their relative sizes (allowing for comparisons between C_M-statistic values from different, asymptotic chi-squared distributions).

The data set `happyNL` has a C_M-statistic of 237.669. Its small p-value (< 0.0001, df = 69) confirms that there is very strong evidence to conclude that the *Household* and *Education* variables are statistically significant predictors of *Happiness*. By partitioning the C_M-statistic associated with the Marcotorchino index, we can examine the sources of non-symmetric association that exists in the three-way table. We see that all the bivariate association terms are statistically significant, but not the trivariate association term (p-value < 0.251, df = 36) which assesses the increase in predictability of *Happiness* given the number of people in a *Household* and the highest level of *Education* of the participants.

Visual summary of the association

While the trivariate term from the partition of the C_M-statistic is not statistically significant, we shall nonetheless visually explore how people's level of *Happiness* is influenced by the number of people in their *Household* and their highest level of *Education*. This shall be done by generating an interactive biplot with the interaction of each combination of categories of the predictor variables depicted using principal coordinates and, therefore, setting `biptype = "pred"`. Note that there is indeed an "interaction" (via a symmetric association) between the two predictor variables since the `Term-JK` p-value is less than < 0.0001. The categories of the response variable are depicted in the biplot using standard coordinates when `biptype = "pred"`.

Applying the `plot()` function to the **CA3variants** object can be used to generate different biplots. A description of some of all available plotting arguments can be found in Table 3. However, when a non-symmetric variant is applied to the **CA3variants** object, a suitable interactive biplot that portrays the non-symmetric association can be obtained using the command:

```
plot(nsca3.out, biptype = "pred")
```

which produces the interactive biplot of Figure 4. Figure 4 displays straight lines from the origin to each standard coordinate to depict the non-interactive variable for the four levels of the *Happiness* variable. Such lines are convenient for visualizing how the interactive points relate to the non-interactive points. This is because the proximity of the points from the origin reflect deviations from independence.

In Figure 4, we see that the first dimension accounts for 47% (rounded to the nearest integer) of the association between the variables while the second dimension accounts for 23%. Thus, Figure 4 captures approximately 70% of the association between the three categorical variables (when treated non-symmetrically) of `happyNL`. These two percentages are also included as `p1` and `p2`, respectively, from the numerical summaries included in `print(nsca3.out)`. Since Figure 4 provides a good visual summary of the non-symmetric association of the variables of `happyNL`, we now turn our attention to describing the nature of this association. The left side of Figure 4 shows a group of points corresponding to HS1 (a single person household) combined with all levels of education. It shows that respondents tend to exhibit lower levels of happiness when they live alone, regardless of education level. Due to the non-symmetric nature of the association we can also infer that for these single households, the groups with lower levels of education (HS1ED1 and HS1ED2) lead (or help predict) a low, or middle, level of happiness. For those with a higher education, Figure 4 also suggests that having a higher level of education does not necessarily lead to (or help to predict) a very-high happiness level. Furthermore, respondents in a two person household (HS2) tend to be very happy (HS2 is a good predictor of very-high levels of happiness), especially for those with a lower level of education (HS2ED1 and HS2ED2). The interactive biplot shows that those with higher levels of education in a two person household are still more associated with a very-high level of happiness (HS2ED3 and HS2ED45) but less than those with less of an education.

For the large households (HS4 and >HS5), we observe that the effect of education level on happiness appears to be stronger. That is, for these larger households, respondents with a higher (ED45) or a middle-high (ED3) level of education tend to be more happy (high and very-high) than people with a

Arguments	Description
`Xout`	The output of `CA3variants()`.
`biptype`	Specifies the type of interactive biplot being produced. When `ca3type = "CA3"` or `= "OCA3"` there are six options: `biptype = "row"`, `"column"`, `"tube"`, `"row-column"`, `"row-tube"` and `"column-tube"`. Each option refers to what is depicted using principal coordinates. For instance, `"row"` specifies that the row points are depicted using principal coordinates and, consequently, the interactive column-tube points are depicted using standard coordinates. When `ca3type = "NSCA3"` or `"ONSCA3"`, there are only two biplot options: `biptype = "resp"` or `"pred"`. The option `"resp"` specifies that the response categories are depicted using principal coordinates, while the option `"pred"` indicates that the interactive predictor points are in principal coordinates.
`scaleplot`	A biplot scaling argument used to avoid spatial cluttering by pulling points away from the origin. See the description of the "gamma scaling" in Gower et al. (2011, Section 2.3.1). By default, `scaleplot` is the overall average of the sum-of-squares of the two sets of coordinates (principal and standard ones), so that the average sum-of-squares for the two sets of points is the same (van de Velden et al., 2017).
`addlines`	Specifies whether the points in standard coordinates are represented using axes. By default, `addlines = TRUE`.

Table 3: Summary of important plotting options available in `plot.CA3variants()`. For all options use `?plot.CA3variants`

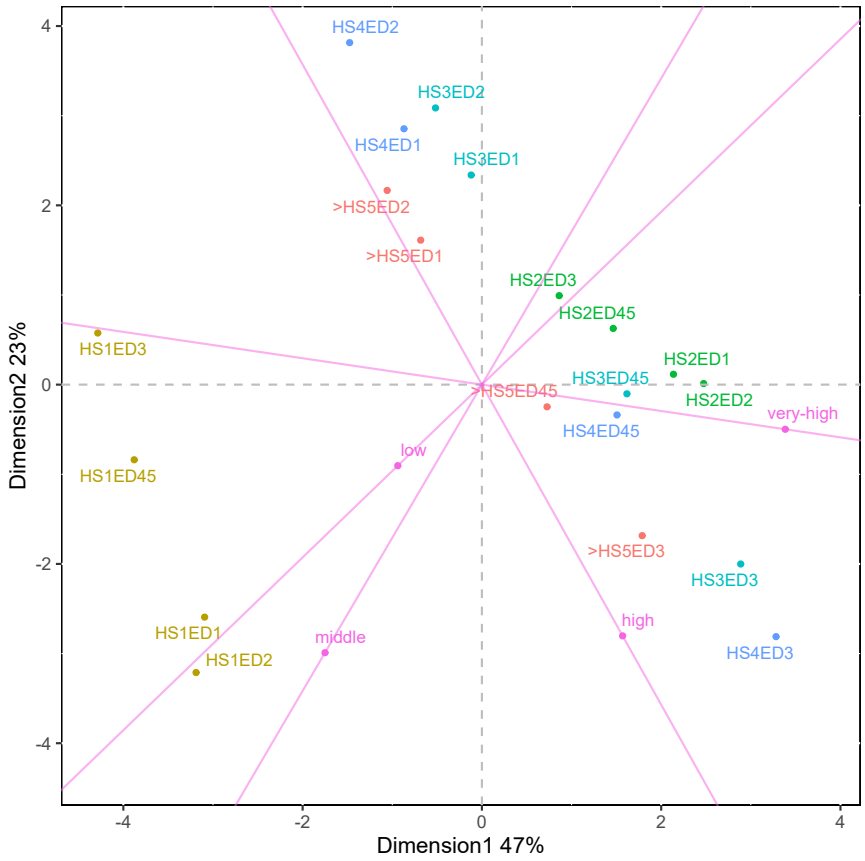

Figure 4: Interactive biplot from the NSCA3 of `happyNL` with *Happiness* and *Education* the interactive variables

lower level of education (ED1 and ED2). Respondents that live in large households and have a low level of education (HS4ED2, HS4ED1, >HS5ED2 and >HS5ED1) are not highly happy individuals. Indeed, these interative points are on the opposite side of Figure 4 to the high level of happiness.

6.3 Ordered three-way correspondence analysis: Ranking and rating data

In the analysis of `ratrank` in Section 2.6.1 we treated the *Ranking* and *Rating* variables as nominal when they are, in fact, ordinal variables. Using the `CA3variants` package we can incorporate this ordinality in the decomposition. In particular, we perform the analysis by treating *Country* as a nominal variable and *Ranking* and *Rating* as ordinal by using the hybrid decomposition described in Section 2.3.3.

Dimensionality of the solution

Before we construct a low-dimensional display of the association between the ordinal variables (*Rating* and *Ranking*) and the nominal variable (*Country*), we determine the appropriate dimension of the solution. The $9 \times 9 \times 5$ data set `ratrank` has a $8 \times 8 \times 4$ sized matrix of hybrid core elements that reflect the trivariate sources of association between the three variables. Not all these sources are important for describing the analysis, or are even practically relevant. In most practical cases the linear and quadratic sources of association are sufficient and provide a meaningful description of the association. We use the `tunelocal()` function to determine the appropriate number of hybrid core elements to define the dimensionality of the solution. When using the `tunelocal()` function for analysing ordinal variables, one needs to specify the number of them; this is done by setting the argument `norder = 2`. The numerical and graphical summaries from using this function are obtained using the commands:

```
tune.oca3.out <- tunelocal(ratrank, ca3type = "OCA3", norder = 2)
print(tune.oca3.out)
```

The numerical and graphical output of a similar form to those seen in the previous examples. The visual and numerical outputs (not given here) show that the highest order hybrid core element is of order $(2, 2, 1)$ (i.e. the quadratic-by-quadratic-by-first order component association) so that all terms, up to the (2, 2, 1) term, together account for most of the association between the three variables.

Numerical summary of the association

We perform OCA3 on `ratrank` using the dimensionalities as suggested by the output of the `tunelocal()` function. Hence, we confine our attention to sources of association no higher than the quadratic-by-quadratic-by-first hybrid core so that:

```
oca3.out <- CA3variants(ratrank, ca3type = "OCA3", dims = c(2, 2, 1), norder = 2)
```

We note that from such an analysis, only four of the 256 hybrid core elements are required to account for most of the association that exists between the variables. We can gain more insight into the structure of the association by inspecting the core elements from the hybrid decomposition. When using the function `summary()`, the elements of the core and squared core arrays (Lombardo et al., 2021), respectively, can be obtained:

```
summary(oca3.out)

#> Core table
#> , , r1

#>       q1     q2
#> p1 -0.527 -0.143
#> p2  0.198 -0.246

#> Squared core table
#> , , r1

#>      q1    q2
#> p1 0.278 0.020
#> p2 0.039 0.061

#> Explained inertia (reduced dimensions)
#> [1] 0.398

#> Total inertia (complete dimensions)
#> [1] 0.477
```

```
#> Proportion of explained inertia (when reducing dimensions)
#> [1] 0.834
```

Note that by confining the solution to include terms no higher than (2, 2, 1), the sum of squares of these four squared core elements is 83.4% of the association that exists between the three variables of the contingency table.

The four terms from this output are all adequately described using the linear and quadratic polynomials for *Rating* and *Ranking* and just one Tucker3 component for *Country*, and are:

- the linear-by-linear polynomial component term (`0.278`) which describes the association between the ordered variables in terms of any differences that exist in the linearity of each ordered set of categories that form the *Rating* and *Ranking* variables,
- the linear-by-quadratic polynomial component term (`0.020`) which describes the association between the ordered variables in terms of any linear differences in the *Rating* variable and dispersion differences in the *Ranking* variable,
- the quadratic-by-linear polynomial component (`0.039`) which describes the association between the ordered variables in terms of any dispersion differences in the *Rating* variable and any linear differences that exist in the *Ranking* variable, and
- the quadratic-by-quadratic polynomial component (`0.061`) which describes the association between the ordered variables in terms of any dispersion differences that exist in the *Rating* and *Ranking* variables.

Using the `print()` function we obtain the percentage contributions of the components to the total inertia for different biplots, the overall decomposition information, as well as the partitionings of the four terms of the Pearson three-way chi-squared statistic into their polynomial components. For example, suppose we focus on the pair-wise association between the *Rating* and *Ranking* variables. The partial output corresponding to the row and column (linear and non-linear) components of the `Chi2-IJ` term can be shown:

```
print(oca3.out)
#> # ...
#> # Partition of the Term-IJ using polynomials

#>          Term-IJ-poly %inertia df p-value
#> poly-row1    12701.275   69.182  8       0
#> poly-row2     3882.996   21.150  8       0
#> poly-row3      833.867    4.542  8       0
#> poly-row4      346.473    1.887  8       0
#> poly-row5      177.262    0.966  8       0
#> poly-row6      167.495    0.912  8       0
#> poly-row7      122.977    0.670  8       0
#> poly-row8      126.927    0.691  8       0
#> Chi2-IJ      18359.272  100.000 64       0
#> poly-col1    13819.761   75.274  8       0
#> poly-col2     3177.816   17.309  8       0
#> poly-col3      605.183    3.296  8       0
#> poly-col4      189.747    1.034  8       0
#> poly-col5      149.127    0.812  8       0
#> poly-col6      160.638    0.875  8       0
#> poly-col7      124.849    0.680  8       0
#> poly-col8      132.150    0.720  8       0
#> Chi2-IJ      18359.272  100.000 64       0
#> # ...
```

This output shows that all components are statistically significant (with p-values smaller than 0.0001). It also shows that the variation in the *Rating* variable (row variable) is dominated by the difference in the linearity of its categories - the linear component accounts for $100 \times 12701.28/18359.27 = 69.18\%$ of the variation in this variable. The linear component also accounts for the largest source of variation in the *Ranking* variable (column variable), contributing to $100 \times 13819.76/18359.27 = 75.27\%$ of the variables' variation. Thus, if we were to confine our attention to just exploring further the association between the *Rating* and *Ranking* variables by generating a visual summary of the association this can be done using the correspondence analysis approach introduced in Beh (1997) and described by Beh and Lombardo (2014, Chap. 6) and Beh and Lombardo (2021b, Chap. 4). Since both variables are dominated by differences in the linearity of their categories, such an analysis will produce a correspondence plot that is dominated more by the first axis than any of the other axis in the optimal plot.

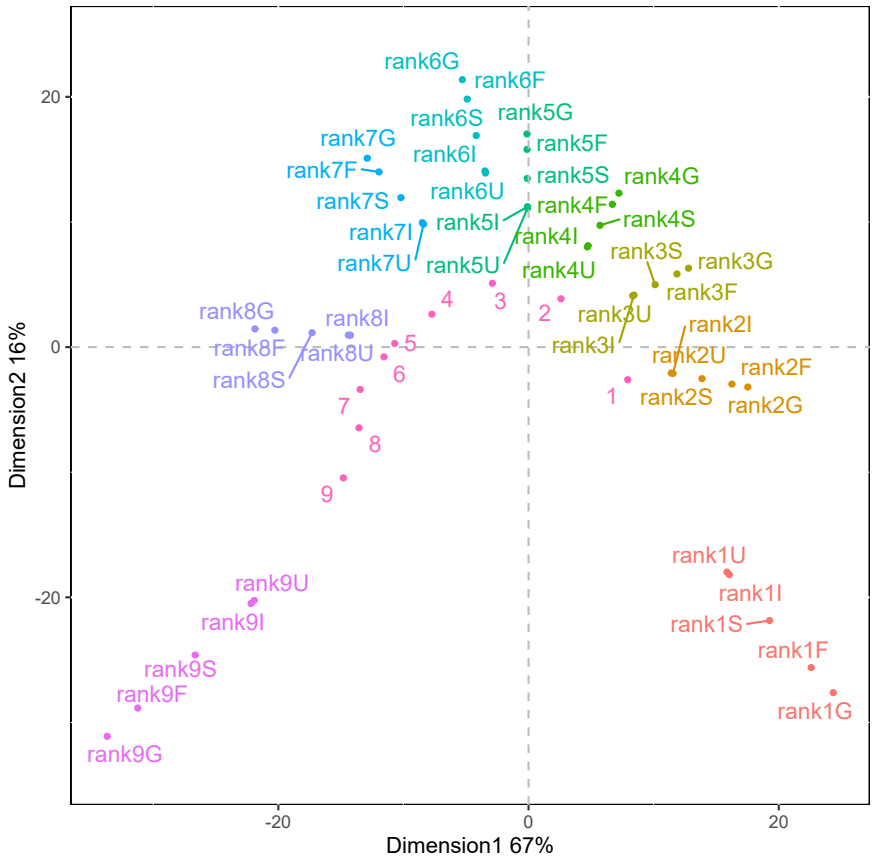

Figure 5: The row biplot from the classical three-way correspondence analysis of `ratrank`.

Visual summary of the association

Since the three-way association term is statistically significant ($X^2 = 21265.91$, p-value < 0.0001), we can examine the nature of this association term more closely. Visually summarizing this three-way association can be done by considering the coordinate systems that generate the biplots described in Section 2.4.1. Recall that in Section 2.6.1, we performed the classical approach to three-way correspondence analysis and visualized the results using the row-tube (interactive) biplot. In doing so, the *Ranking-Country* association – which is comparatively weak (contributing to 1.2% of the association) but is statistically significant (p-value < 0.0001) – is depicted using standard coordinates while the *Rating* categories are depicted using principal coordinates that are akin to $\mathbf{X}_{IK,J}$ in (16).

To highlight differences between the classical and ordered three-way correspondence analysis, we construct the row biplot of Figure 5 using the command:

```
plot(ca3.out, biptype = "row",  addlines = F,  scaleplot = 15)
```

note that `ca3.out` is the output from the classical analysis performed in Section 2.6.1. When *Rating* and *Ranking* are treated as ordinal variables, Figure 6 gives the row biplot that can be obtained from the command:

```
plot(oca3.out, biptype = "row",  scaleplot = 15)
```

where the value for `scaleplot = 15` was chosen by trial and error to ensure a reasonable separation of the points in the biplot, without afffecting the approximation of the association between the variables. While Figure 5 and Figure 6 both give parabolic configurations of the points, these configurations are quite different since the former treats the variables as nominal and uses the components from a Tucker3 decomposition while the latter is constructed using orthogonal polynomials for the ordered row and column (*Ratings and Rankings*) variables and a Tucker3 component for the tube (*Country*) variable. Observe that the parabolic shape of *Rating* in Figure 5 is more pronounced than the parabolic configuration of *Rating* in Figure 6.

In addition to the visual differences between the two configurations of points, there are also some features that make Figures 5 and 6 distinct. Indeed, the first axis of Figure 6 is constructed using the

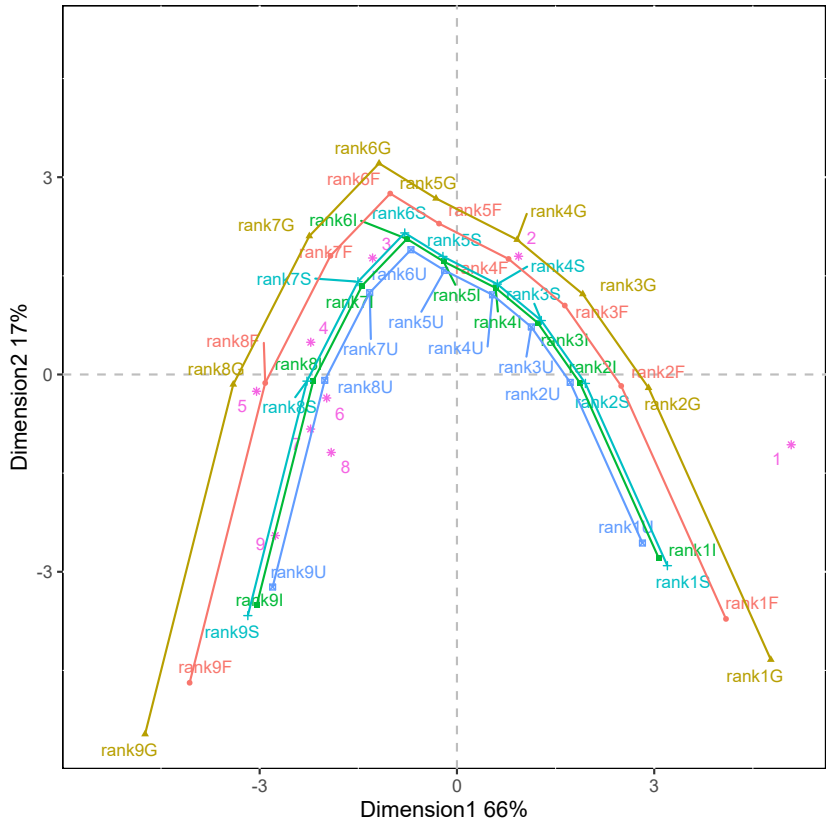

Figure 6: The row biplot from the ordered (hybrid) three-way correspondence analysis of `ratrank`.

linear orthogonal polynomial while the second axis is constructed using the quadratic orthogonal polynomial. The linear and dispersion components contribute to 66% and 17%, respectively, of the total inertia of the data; these percentages can be obtained from the output of the `print(oca3.out)` command.

When considering all variables as nominal, as was done in our analysis in Section 2.6.1, the ratings appear closely associated with the rankings across the five countries. However, treating the two variables (*Rating* and *Ranking*) as ordinal provides additional information on some aspects of the variable distribution (mean and variability). For example, Figure 6 shows that the configuration of the *Rating* categories along the first (linear polynomial) axis is different to the configuration along the first axis of Figure 5. This is because the variation of the *Rating* variabe is dominated more by differences in the linearity of its categories than by its dispersion differences. This dominant linear component affects the variable association and is captured by the configuration of points in Figure 6.

7 Conclusion

The **CA3variants** package described in this paper is, to the best of our knowledge, the only package that allows practitioners and researchers to directly perform four variants of three-way correspondence analysis, including the classical three-way correspondence analysis (Carlier and Kroonenberg, 1996), the non-symmetric variant and the two ordered versions of three-way correspondence analysis (Lombardo et al., 2021). Subsequent versions of the package may allow for additional flexibility by providing the user more tools to numerically and visually explore the association structure between categorical variables. These include, but are not confined to, the decomposition of the generalised Cressie-Read family of divergence statistics (Pardo, 1996). Indeed, Pearson's statistic is one of many measures of symmetric association that can be considered. Alternatives include the Freeman-Tukey statistic, log-likelihood ratio statistic, Neyman's chi-squared statistic, and the Cressie-Read statistic, which were originally developed to study two variables (Cressie and Read, 1984; Beh and Lombardo, 2023). These measures are all special cases of the Cressie-Read family of divergence statistics and have been adapted for three-way and multi-way contingency tables (Pardo, 1996; Pardo and Pardo, 2003;

Lombardo and Beh, 2022). Thus, this family of statistics may be incorporated into the **CA3variants** package, thereby providing the user with greater flexibility for the choice of symmetric association they wish to consider. Furthermore, next version of the package might consider the construction of confidence regions that determine those categories (and interactions) that provide a statistically significant contribution to the association between the variables (Beh, 2010; Ringrose, 1996, 2012). Numerical summaries that accompany such regions, including p-values (Beh and Lombardo, 2015) can certainly be incorporated and would provide similar functionality that is available in the **CAvariants** package used for the correspondence analysis of two cross-classified categorical variables (Lombardo and Beh, 2016).

References

E. J. Beh. Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical Journal*, 39:589 – 613, 1997. URL https://doi.org/10.1002/bimj.4710390507. [p242, 256]

E. J. Beh. A comparative study of scores for correspondence analysis with ordered categories. *Biometrical Journal*, 40:413 – 429, 1998a. URL https://doi.org/10.1002/(SICI)1521-4036(199808)40:4<413::AID-BIMJ413>3.0.CO;2-V. [p242]

E. J. Beh. *Correspondence Analysis using Orthogonal Polynomials*. Unpublished PhD Thesis, University of Wollongong, Australia, 1998b. [p241, 242]

E. J. Beh. Elliptical confidence regions for simple correspondence analysis. *Journal of Statistical Planning and Inference*, 140:2582 – 2588, 2010. URL http://dx.doi.org/10.1016/j.jspi.2010.03.018. [p259]

E. J. Beh and P. J. Davy. Partitioning Pearson's chi-squared statistic for a completely ordered three-way contingency table. *The Australian and New Zealand Journal of Statistics*, 40:465 – 477, 1998. URL https://doi.org/10.1111/1467-842X.00050. [p241, 250]

E. J. Beh and R. Lombardo. *Correspondence Analysis, Theory, Practice and New Strategies*. John Wiley & Sons, Chichester, UK, 2014. [p237, 239, 240, 242, 256]

E. J. Beh and R. Lombardo. Confidence regions and approximate p-values for classical and non symmetric correspondence analysis. *Communications in Statistics - Theory and Methods*, 44:95 – 114, 2015. URL https://doi.org/10.1080/03610926.2013.768665. [p259]

E. J. Beh and R. Lombardo. Features of the polynomial biplot for ordered contingency tables. *Journal of Computational and Graphical Statistics*, 31:403 – 412, 2021a. URL https://doi.org/10.1080/10618600.2021.1990773. [p242]

E. J. Beh and R. Lombardo. *An Introduction to Correspondence Analysis*. John Wiley & Sons, Chichester, UK, 2021b. [p239, 240, 242, 256]

E. J. Beh and R. Lombardo. Correspondence analysis using the cressie–read family of divergence statistics. *International Statistical Review*, 2023. URL https://doi.org/10.1111/insr.12541. [p258]

E. J. Beh, B. Simonetti, and L. D'Ambra. Partitioning a non-symmetric measure of association for three-way contingency tables. *Journal of Multivariate Analysis*, 98:1391 – 1411, 2007. URL https://doi.org/10.1016/j.jmva.2007.01.011. [p239, 240]

R. Bro. The N-Way Toolbox. 2020. URL https://www.mathworks.com/matlabcentral/fileexchange/1088-the-n-way-toolbox. [p245]

A. Carlier and P. M. Kroonenberg. Decompositions and biplots in three-way correspondence analysis. *Psychometrika*, 61:355 – 373, 1996. URL https://doi.org/10.1007/BF02294344. [p237, 238, 239, 241, 243, 258]

A. Carlier and P. M. Kroonenberg. The case of the French cantons: An application of three-way correspondence analysis. In J. Blasius and M. Greenacre, editors, *Visualization of Categorical Data*, pages 253 – 275. Academic Press, San Diego, 1998. [p237]

E. Ceulemans and H. A. L. Kiers. Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical & Statistical Psychology*, 59:133 – 150, 2006. URL https://doi.org/10.1348/000711005X64817. [p244, 248, 251]

C. C. Clogg. Some models for the analysis of association in multiway cross-classifications having ordered categories. *Journal American Statistical Association*, 77:803 – 815, 1982. URL https://doi.org/10.2307/2287311. [p250]

N. A. C. Cressie and T. R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, 46:440 – 464, 1984. URL http://www.jstor.org/stable/2345686. [p258]

L. D'Ambra and N. C. Lauro. Non-symmetrical correspondence analysis for three-way contingency table. In R. Coppi and S. Bolasco, editors, *Multiway Data Analysis*, pages 301 – 315. Elsevier, Amsterdam, 1989. [p239]

J. A. Davis. Codebook for the 1977 General Social Survey, 1977. National Opinion Research Centre, Chicago. [p246, 250]

P. H. C. Eilers. **multiway**: Analysis of multi-way arrays, 2019. URL https://CRAN.R-project.org/package=multiway.Rpackageversion1.0-6. [p245]

P. Emerson. Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics*, 24:696 – 701, 1968. URL https://doi.org/10.2307/2528328. [p241]

P. Giordano, H. A. Kiers, and M. A. D. Ferraro. Three-way component analysis using the R package threeway. *Journal of Statistical Software*, 57:1 – 23, 2014. URL 10.18637/jss.v057.i07. [p245]

L. A. Goodman and W. H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49:732 – 764, 1954. URL https://doi.org/10.2307/2281536. [p239]

J. C. Gower, P. J. F. Groenen, and M. van de Velden. Area biplots. *Journal of Computational and Graphical Statistics*, 19:46 – 61, 2010. URL https://www.jstor.org/stable/25651299. [p249]

J. C. Gower, S. Lubbe, and N. le Roux. *Understanding Biplots*. Wiley, Chichester, 2011. [p244, 254]

J. C. Gower, P. J. F. Groenen, M. van de Velden, and K. Vines. Better perceptual maps: Introducing explanatory icons to facilitate interpretation. *Food Quality and Preference*, 36:61 – 69, 2014. URL https://doi.org/10.1016/j.foodqual.2014.01.004. [p244]

M. Greenacre. *Biplots in Practice*. Fundación BBVA, Barcelona, 2010. [p244]

M. J. Hoffman. **irlba**: Fast truncated singular value decomposition and principal components analysis for large dense and sparse matrices, 2017. URL https://CRAN.R-project.org/package=irlba.Rpackageversion2.3.5.1. [p245]

L. R. Kahle. *Social Values and Social Change: Adaptation to Life in America*. Praeger, New York, 1983. [p247]

H. A. L. Kiers, P. M. Kroonenberg, and J. M. F. T. Berge. An efficient algorithm for TUCKALS3 on data with large numbers of observation units. *Psychometrika*, 57:415 – 422, 1992. URL https://doi.org/10.1007/BF02295429. [p240]

P. Kroonenberg and R. Lombardo. Nonsymmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research Journal*, 34:367 – 397, 1999. URL https://doi.org/10.1207/S15327906MBR3403_4. [p239]

P. Kroonenberg and F. J. Oort. Three-mode analysis of multi-mode covariance matrices. *British Journal of Mathematical and Statistical Psychology*, 56:305 – 336, 2003. URL https://doi.org/10.1348/000711003770480066. [p244]

P. M. Kroonenberg. *Three Mode Principal Component Analysis*. DSWO Press, Leiden, 1983. [p240]

P. M. Kroonenberg. Singular value decompositions of interactions in three-way contigency tables. In R. Coppi and S. Bolasco, editors, *Multiway Data Analysis*, pages 169 – 184. Elsevier, Amsterdam, 1989. [p237]

P. M. Kroonenberg. The TUCKALS line: A suite of programs for three-way data analysis. *Computational Statistics and Data Analysis*, 18:73 – 96, 1994. URL https://doi.org/10.1016/0167-9473(94)90133-3. [p240]

P. M. Kroonenberg. *Applied Multiway Data Analysis*. John Wiley & Sons, Hoboken, NJ, 2008. [p237, 239, 240, 243, 244, 250]

P. M. Kroonenberg and J. D. Leeuw. Principal component analysis of three mode data by means of alternating least squares algorithms. *Psychometrika*, 45:69 – 97, 1980. URL https://doi.org/10.1007/BF02293599. [p240]

H. O. Lancaster. Complex contingency tables treated by the partition of the chi-square. *Journal of Royal Statistical Society, Series B*, 13:242 – 249, 1951. URL https://www.jstor.org/stable/2984066. [p238]

N. C. Lauro and L. D'Ambra. L'analyse non symétrique des correspondances. In E. Diday and et al, editors, *Data Analysis and Informatics III*, pages 433 – 446. Elsevier, Amsterdam, 1984. [p239]

D. Leibovici. Spatio-temporal multiway data decomposition using principal tensor analysis on k-modes: The r package **PTAk**. *Journal of Statistical Software*, 34(10):34 pages, 2010. URL 10.18637/jss.v034.i10. [p245]

J. Li, J. Bien, and M. T. Wells. **rTensor**: An R package for multidimensional array (tensor) unfolding, multiplication, and decomposition. *Journal of Statistical Software*, 87(10):31 pages, 2018. URL DOI:10.18637/jss.v087.i10. [p245]

R. J. Light and H. B. Margolin. An analysis of variance for categorical data. *Journal of the American Statistical Association*, 66:534 – 544, 1971. URL https://doi.org/10.2307/2283520. [p240]

S. Loisel and Y. Takane. Partitions of Pearson's chi-square statistic for frequency tables: A comprehensive account. *Computational Statistics*, 31:1429 – 1452, 2016. URL https://doi.org/10.1007/s00180-015-0619-1. [p239]

R. Lombardo and E. J. Beh. Variants of simple correspondence analysis. *The R Journal*, 8/2:167 – 184, 2016. [p259]

R. Lombardo and E. J. Beh. Three–way correspondence analysis for ordinal–nominal variables. In A. Petrucci and R. Verde, editors, *SIS 2017 Statistics and Data Science: New Challenges, New Generations, 28–30 June 2017, Florence (Italy) Proceedings of the Conference of the Italian Statistical Society*, pages 613 – 620. Firenze Press, 2017. [p240, 242]

R. Lombardo and E. J. Beh. Partitioning the Cressie-Read divergence statistic for three-way contingency tables: a study on environmental sustainability data. In R. Lombardo, I. Camminatiello, and V. Simonacci, editors, *IES 2022 Innovation & Society 5.0: Statistical and Economic Methodologies for Quality Assessment. Book of Short Papers*, pages 491–497. PKE Press, 2022. [p259]

R. Lombardo, A. Carlier, and L. D'Ambra. Nonsymmetric correspondence analysis for three-way contingency tables. *Methodologica*, 4:59 – 80, 1996. [p239, 243]

R. Lombardo, E. J. Beh, and P. M. Kroonenberg. Modelling trends in ordered correspondence analysis using orthogonal polynomials. *Psychometrika*, 81:325 – 349, 2016a. URL https://doi.org/10.1007/s11336-015-9448-y. [p242]

R. Lombardo, P. Kroonenberg, and E. J. Beh. Modelling trends in ordered three-way non-symmetrical correspondence analysis. In M. Pratesi and C. Perna, editors, *Proceedings of the 48th Scientific Meeting of the Italian Statistical Society, June 8-10, 2016*, page 14 pages. Springer, 2016b. [p240, 241]

R. Lombardo, E. J. Beh, and L. Guerrero. Analysis of three-way non-symmetrical association of food concepts in cross-cultural marketing. *Quality & Quality*, 53:2323 – 2337, 2019. URL https://doi.org/10.1007/s11135-018-0733-6. [p237]

R. Lombardo, Y. Takane, and E. J. Beh. Familywise decompositions of Pearson's chi-square statistic in the analysis of contingency tables. *Advances in Data Analysis and Classification*, 14 (3):629 – 649, 2020. URL https://doi.org/10.1007/s11634-019-00374-7. [p238, 239]

R. Lombardo, E. J. Beh, and P. M. Kroonenberg. Symmetrical and non-symmetrical variants of three-way correspondence analysis for ordered variables. *Statistical Science*, 36 (4):542 – 561, 2021. URL https://doi.org/10.1214/20-STS814. [p237, 240, 241, 242, 243, 244, 255, 258]

M. Marcotorchino. Utilisation des comparaisons par paires en statistique des contingencies: Partie i. Etude du Centre Scientifique, IBM, France, No F 069, 1984a. [p239]

M. Marcotorchino. Utilisation des comparaisons par paires en statistique des contingencies: Partie ii. Etude du Centre Scientifique, IBM, France, No F 071, 1984b. [p239]

M. Marcotorchino. Utilisation des comparaisons par paires en statistique des contingencies: Partie iii. Etude du Centre Scientifique, IBM, France, No F 081, 1985. [p239]

T. Murakami and P. M. Kroonenberg. Three-mode models and individual differences in semantic differential data. *Multivariate Behavioral Research*, 38:247 – 283, 2003. URL https://doi.org/10.1207/S15327906MBR3802_5. [p244]

L. Pardo and M. C. Pardo. Minimum power-divergence estimator in three-way contingency tables. *Journal of Statistical Computation and Simulation*, 73:819 – 831, 2003. URL https://doi.org/10.1080/0094965031000097782. [p258]

M. C. Pardo. An empirical investigation of cressie and read tests for the hypothesis of independence in three-way contingency tables. *Kybernetika*, 32:175 – 183, 1996. URL http://hdl.handle.net/10338.dmlcz/124180. [p258]

J. C. W. Rayner and E. J. Beh. Towards a better understanding of correlation. *Statistica Neerlandica*, 63:324 – 333, 2009. URL https://doi.org/10.1111/j.1467-9574.2009.00425.x. [p242]

W. Revelle. **psych**: Procedures for psychological, psychometric, and personality research, 2018. URL https://CRAN.R-project.org/package=psych. [p245]

T. J. Ringrose. Alternative confidence regions for canonical variate analysis. *Biometrika*, 83:575 – 587, 1996. URL https://doi.org/10.1093/biomet/83.3.575. [p259]

T. J. Ringrose. Bootstrap confidence regions for correspondence analysis. *Journal of Statistical Computation and Simulation*, 82:1397 – 1413, 2012. URL https://doi.org/10.1080/00949655.2011.579968. [p259]

A. Statnikov. **tensorA**: Algebra for tensors, 2018. URL https://CRAN.R-project.org/package=tensorA.Rpackageversion0.36.2. [p245]

Y. Takane and S. Jung. Regularized partial and/or constrained redundancy analysis. *Psychometrika*, 73:671 – 690, 2008. URL https://doi.org/10.1007/s11336-008-9067-y. [p239]

M. Timmerman and H. A. L. Kiers. Three-mode principal component analysis: Choosing the numbers of components and sensitivity to local optima. *British Journal of Mathematical and Statistical Psychology*, 53:1 – 16, 2000. URL https://doi.org/10.1348/000711000159132. [p244]

L. R. Tucker. Implications of factor analysis of three-way matrices for measurement of change. In C. W. Harris, editor, *Problems in Measuring Change*, pages 122 – 137. University of Wisconsin Press, 1963. [p240]

M. van de Velden, A. I. D'Enza, and F. Palumbo. Cluster correspondence analysis. *Psychometrika*, 82:158 – 185, 2017. URL https://doi.org/10.1007/s11336-016-9514-0. [p249, 254]

H. van Herk and M. van de Velden. Insight into the relative merits of rating and ranking in a cross-national context using three-way correspondence analysis. *Food Quality and Preference*, 18:1096 – 1105, 2007. URL https://doi.org/10.1016/j.foodqual.2007.05.006. [p237, 246, 247, 248, 249]

X. Zhou. **mvoutlier**: Multivariate outlier detection, 2019. URL https://CRAN.R-project.org/package=mvoutlier.Rpackageversion2.1.1. [p245]

Rosaria Lombardo
Department of Economics, University of Campania "Luigi Vanvitelli"
Capua (CE), Italy
rosaria.lombardo@unicampania.it

Michel van de Velden
Econometric Institute, Erasmus University
Rotterdam, The Netherlands
vandevelden@ese.eur.nl

Eric J. Beh
National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong
Wollongong, Australia
and
Centre for Multi-Dimensional Data Visualisation (MuViSU)
Stellenbosch University
Stellenbosch, South Africa
ericb@uow.edu.au