Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

A Latent Block Distance-Association Model for Profile by Profile Cross-Classified Categorical Data

J. Fernando Vera & Mark De Rooij

To cite this article: J. Fernando Vera & Mark De Rooij (2019): A Latent Block Distance-Association Model for Profile by Profile Cross-Classified Categorical Data, Multivariate Behavioral Research, DOI: 10.1080/00273171.2019.1634995

To link to this article: https://doi.org/10.1080/00273171.2019.1634995

MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2019.1634995

A Latent Block Distance-Association Model for Profile by Profile Cross-Classified Categorical Data

J. Fernando Vera[a] and Mark de Rooij[b]

[a]Department of Statistics and O.R. Faculty of Sciences, University of Granada; [b]Methodology and Statistics Unit, Institute of Psychology, Leiden University

ABSTRACT

Distance association models constitute a useful tool for the analysis and graphical representation of cross-classified data in which distances between points inversely describe the association between two categorical variables. When the number of cells is large and the data counts result in sparse tables, the combination of clustering and representation reduces the number of parameters to be estimated and facilitates interpretation. In this article, a latent block distance-association model is proposed to apply block clustering to the outcomes of two categorical variables while the cluster centers are represented in a low dimensional space in terms of a distance-association model. This model is particularly useful for contingency tables in which both the rows and the columns are characterized as profiles of sets of response variables. The parameters are estimated under a Poisson sampling scheme using a generalized EM algorithm. The performance of the model is tested in a Monte Carlo experiment, and an empirical data set is analyzed to illustrate the model.

KEYWORDS

Distance-association model; clustering; latent class analysis; mixture distribution; EM algorithm; BIC

Introduction

Data are often collected with multiple explanatory variables and multiple response variables. For the analysis of such data, multivariate regression models may be fitted providing an insight into the relationship between the explanatory variables and the response variables. When categorical response variables are considered, multivariate logistic regression models are usually appropriate. Examples include logistic regression models estimated using a generalized estimating equation (Zeger, Liang, & Albert, 1988) approach, or the multivariate logistic distance model (Worku & De Rooij, 2018) that considers a dimensional structure of the response variables.

On the other hand, instead of this kind of variable-oriented approach, a person-oriented approach to data analysis might be preferred (Bergman & Magnusson, 1997), focusing on the personal profiles of the variables. When both the explanatory and the response variables are categorical, we might be interested in examining how the first set of profiles is related to the second. Addressing this question often involves the analysis of large contingency tables with explanatory profiles in the rows and response profiles in the columns. A disadvantage of these large contingency tables is that they are sparse even with large samples, i.e., many cells are not present in the data set.

As an example of such a situation, we consider data in which the row profiles are based on gender and on five personality variables, while the column profiles are cross-classifications of five mental disorders. This data set has been analyzed before, taking a variable-oriented approach (Spinhoven, De Rooij, Heiser, Penninx, & Smit, 2009). In this article, we take a person-oriented approach, in which the personality variables are Neuroticism, Extraversion, Openness to experience, Agreeableness, and Conscientiousness, each categorized as Low, Medium or High. Therefore, there are $2 \times 3^5 = 486$ different row profiles. The columns contain the following mental disorders: Major Depressive Disorder, Dysthymia, Generalized Anxiety Disorder, Social Phobia, and Panic Disorder. The subjects are diagnosed as being with or without the disorder, which produces $2^5 = 32$ different profiles of mental disorders. The resulting contingency table, therefore, is large (with dimensions 486 by 32). The sample is composed of 2,938 subjects, scattered throughout the

CONTACT J. Fernando Vera jfvera@ugr.es Department of Statistics and O.R. Faculty of Sciences, University of Granada C/ Fuente Nueva s/n, 18071 Granada, Spain.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hmbr.

contingency table; thus of the 15,552 cells 13,956 are empty, and only 1,596 are nonempty. Such sparse contingency tables are problematic for many standard analysis techniques. Loglinear analysis, for example, completely fails for these data simply because of the large number of zero cells.

One solution to this problem may be to use clustering techniques. Clustering in combination with modeling the relationships between the categorical response variables is a widely employed procedure in data analysis. In general, most classical two-mode clustering methods are designed to attain homogeneous row-by-column clusters, while other methods are designed to find partitions based on the optimization of within-block interactions for a single quantitative dependent variable (Schepers, Bock, & Van Mechelen, 2017). For contingency tables, various procedures have been proposed to combine categories in terms of a particular homogeneity criterion (Goodman, 1981; Govaert & Nadif, 2014; Kateri & Iliopoulos, 2003), or seeking to maximize a measure of dependence (Bock, 2003; Govaert, 1995). Latent block clustering methods have been proposed using a Poisson model, for example in information retrieval (Li & Zha, 2006), or for sequencing data (Witten, 2011), among others. With the aim of reducing the number of parameters and at the same time to facilitate the interpretation, clustering and representation methods have been proposed in different areas for different data sets (see, e.g., Kim, Choi, & Hwang, 2017; Vera, Macías, & Heiser, 2009a, Vera, Macías, & Heiser, 2009b; Vera, Macías, & Heiser, 2013).

Association patterns among categorical variables have traditionally been studied by log-linear and association models (Agresti, 2013). However, the application of log-linear models to large contingency tables usually produces a large number of parameters. Models with fewer parameters for the association have been proposed for nonsparse data such as the RC(M) association model (Goodman, 1985) or the distance association (DA) model (De Rooij & Heiser, 2005). These models are equivalent, except that the latter is based on distances and so is easier to interpret (De Rooij, 2007, 2008). In DA models, row and column categories for cross-classified data are represented in a Euclidean space of low dimensionality such that the distances between points inversely describe the association between the categories of the two sets. The presence of zero entries in a contingency table means that the estimated odds ratios will be either zero, infinity, or undefined. Therefore, when standard log-linear models are used for sparse tables in which the number of cells is large relative to the sample size, estimation problems are often experienced (Vera, de Rooij, & Heiser, 2014).

As noted by the latter authors, for nonsparse tables involving profiles, the DA model can be estimated but the association plot may be difficult to interpret due to the presence of a large number of points (profiles). For sparse tables such as the one presented above, the DA model fails due to the excessive number of zero cells. For cross-classified data where the row categories correspond to the profiles, Vera et al. (2014) propose a latent class distance association model (LCDA) that reduces the number of parameters, which makes the model suitable for sparse tables. This model clusters the row profiles into a small number of classes and represents them with the column categories in an association plot. When the column categories also represent profiles of a set of variables, it is advisable, additionally, to cluster the column profiles into a small number of classes. In this case, the number of parameters can be further reduced by simultaneously collapsing the rows and columns of the table.

In the context of the mental disorder example mentioned above, clustering the column profiles is also of substantive interest. A major issue in mental disorder studies is that of understanding comorbidity, i.e., when an individual presents multiple mental disorders. By clustering profiles, we can obtain classes in which certain disorders occur concurrently, i.e., classes that represent patterns of comorbidity.

In this article, we address the problem of block clustering and the simultaneous representation of association between row and column clusters in a contingency table. A latent block distance association model (LBDA) is formulated that simultaneously partitions the rows and the columns of a contingency table, while the between-cluster associations are represented in a low dimensional space in terms of Euclidean distances. In the LBDA model, odds are defined in terms of the block-related main effects and of the distances, while odds ratios are defined only in terms of the distances. This model can be viewed as an alternative to traditional models for representing associations between two categorical variables, one that reduces the number of parameters to be estimated and facilitates interpretation.

The rest of this article is organized as follows. In the next section, the LBDA model is formulated and estimated by maximum likelihood using a generalized expectation-maximization (GEM) algorithm. A model selection procedure using the BIC statistic and a guideline for model interpretation are then discussed in Model selection section and Model interpretation

section, respectively. Experimental results are shown in Experimental results section. The performance of the LBDA model is considered in a Monte Carlo experiment, first using simulated nonsparse data sets and then focusing on the influence of sampling zeros in recovering the structure of clustered contingency tables. The above-described personality profile data set is then analyzed and, in the final section, the results obtained are discussed.

Latent block distance association model

The model

Consider an $I \times J$ contingency table $\mathbf{F} = (f_{ij})$ that collects the counts of combinations of row and column categories, which can represent profiles of variables. Let us consider a partition $\mathcal{P}_I(\mathbf{F})$ of the row categories into T latent classes $\mathbf{F}_t^R$ with n_t rows each, $n_1 + \cdots + n_T = I$, and a partition $\mathcal{P}_J(\mathbf{F})$ of the column categories into K latent classes $\mathbf{F}_k^C$ with n_k columns each, $n_1 + \cdots + n_K = J$. It is assumed that a category belongs to one and only one subset of its corresponding partition, and that we do not know in advance which latent class a particular element belongs to. Without loss of generality we can assume that both rows and columns in $\mathbf{F}$ are arranged by permuting them in accordance with the sequence in the index sets of the latent classes.

In terms of the frequency table the situation is equivalent to having a block shaped partition $\mathcal{P}(\mathbf{F})$ of the rectangular matrix $\mathbf{F}$ into $T \times K$ blocks $\mathbf{F}_{tk}$ of $n_{tk} = n_t n_k$ frequencies f_{ij}, corresponding to the entries simultaneously present in the row vectors $\mathbf{f}_i^R = (f_{i1}, \ldots, f_{iJ})'$, with $\mathbf{f}_i^R \in \mathbf{F}_t^R$, and in the column vectors $\mathbf{f}_j^C = (f_{1j}, \ldots, f_{Ij})'$, with $\mathbf{f}_j^C \in \mathbf{F}_k^C$. The unconditional probability that any row vector $\mathbf{f}_i^R$ belongs to latent class $\mathbf{F}_t^R$ is denoted by γ_t^R, with $0 \leq \gamma_t^R \leq 1$, and that any column vector $\mathbf{f}_j^C$ belongs to latent class $\mathbf{F}_k^C$ is denoted by γ_k^C, with $0 \leq \gamma_k^C \leq 1$. It is assumed that the unconditional probability that any frequency f_{ij} belongs to a latent block $\mathbf{F}_{tk}$ is given by $\gamma_{tk} = \gamma_t^R \gamma_k^C$. Thus

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_{tk} = \sum_{t=1}^{T}\gamma_t^R = \sum_{k=1}^{K}\gamma_k^C = 1. \tag{1}$$

The aim of the latent block distance association model is to represent, not the row and column categories (or profiles) themselves, but the corresponding cluster centers by points in a Euclidean space of low dimension. Thus, let us define the $T \times M$ matrix $\mathbf{X}$ and the $K \times M$ matrix $\mathbf{Y}$, whose row vectors $\mathbf{x}_t$, $t = 1, \ldots, T$ and $\mathbf{y}_k, k = 1, \ldots, K$ are the coordinates of the centers of the T clusters for the rows and the K clusters for the columns respectively in dimension M. Under the general multiplicative form in the distance association model, the expected frequency μ_{tk} of any $f_{ij} \in \mathbf{F}_{tk}$ is assumed to be given by

$$\mu_{tk} = \mu\alpha_t\beta_k \exp\left(-d_{tk}^2\right), \tag{2}$$

where μ is the overall scale parameter, α_t is the latent row-class effect parameter, β_k is the latent column-class effect parameter and $d_{tk}^2 = d^2(\mathbf{x}_t, \mathbf{y}_k)$ is the squared Euclidean distance given by

$$d^2(\mathbf{x}_t, \mathbf{y}_k) = \sum_{m=1}^{M}(x_{tm} - y_{km})^2.$$

Equation (2) represents a distance association model in which associations between row and column clusters are inversely related to the squared distances between the corresponding points in the estimated configuration. This model extends the LCDA model of Vera et al. (2014), by further considering that it is not the column modalities itself, but their simultaneously estimated cluster centers, that are of interest.

The GEM algorithm

Following the usual mixture approach to the estimation problem, the probability for the frequency $f_{ij} \in \mathbf{F}_{tk}$ in a standard Poisson sampling model is expressed as

$$h_{tk}\left(f_{ij}|\mathbf{x}_t, \mathbf{y}_k, \mu, \alpha_t, \beta_k\right) = \frac{\mu_{tk}^{f_{ij}}}{f_{ij}!}\exp\left(-\mu_{tk}\right), \tag{3}$$

where μ_{tk} is given by (2). Because in this context it is not known in advance which latent class a frequency belongs to, the probability density function (p.d.f.) of the random variable f_{ij} becomes a finite mixture of Poisson densities, i.e.,

$$g\left(f_{ij}|\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}\right) = \sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_{tk}h_{tk}\left(f_{ij}|\mathbf{x}_t, \mathbf{y}_k, \mu, \alpha_t, \beta_k\right), \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ and $\boldsymbol{\Gamma} = (\gamma_{tk})$ is the $T \times K$ matrix of unconditional probabilities. Therefore, the log-likelihood function to be maximized subject to (1) can be written as

$$\log L = \sum_{i=1}^{I}\sum_{j=1}^{J}\log\sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_{tk}h_{tk}\left(f_{ij}|\mathbf{x}_t, \mathbf{y}_k, \mu, \alpha_t, \beta_k\right). \tag{5}$$

For the formulation of the mixture problem in the EM (Dempster, Laird, & Rubin, 1977) framework the following mixture component indicator variables are introduced,

$$z_{ij,tk} = \begin{cases} 1, & \text{if } f_{ij} \in \mathbf{F}_{tk}, \\ 0, & \text{otherwise.} \end{cases}$$

As usual, let us define the vector $\mathbf{z}_{ij} = (z_{ij,11}, \ldots, z_{ij,TK})'$, and the $IJ \times TK$ matrix $\mathbf{Z}$ written by their row vectors $\mathbf{z}_{ij}$. It will be assumed that the $\mathbf{z}_{ij}$ are observations of independently and identically distributed multinomial variables, with probabilities given by the entries of the matrix $\boldsymbol{\Gamma}$ such that

$$\sum_{t=1}^{T}\sum_{k=1}^{K} z_{ij,tk} = 1.$$

The p.d.f. of f_{ij}, given $\mathbf{z}_{ij}$, can be written as,

$$\Psi(f_{ij}|\mathbf{z}_{ij}, \mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{t=1}^{T}\prod_{k=1}^{K} h_{tk}(f_{ij}|\mathbf{x}_t, \mathbf{y}_k, \mu, \alpha_t, \beta_k)^{z_{ij,tk}} \tag{6}$$

and the unconditional p.d.f. of $\mathbf{z}_{ij}$ is expressed as,

$$p(\mathbf{z}_{ij}|\boldsymbol{\Gamma}) = \prod_{t=1}^{T}\prod_{k=1}^{K} \gamma_{tk}^{z_{ij,tk}}. \tag{7}$$

Using (6) and (7), the complete p.d.f. of f_{ij} and $\mathbf{z}_{ij}$ can be written as

$$\begin{aligned} \Phi(f_{ij}, \mathbf{z}_{ij}|\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}) &= \Psi(f_{ij}|\mathbf{z}_{ij}, \mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\mathbf{z}_{ij}|\boldsymbol{\Gamma}) \\ &= \prod_{t=1}^{T}\prod_{k=1}^{K} \left(\gamma_{tk} h_{tk}(f_{ij}|\mathbf{X}, \mathbf{Y}, \mu, \alpha_t, \beta_k)\right)^{z_{ij,tk}}, \end{aligned} \tag{8}$$

and the log-likelihood of the complete data $\mathbf{F}$ and $\mathbf{Z}$ can be expressed as

$$\begin{aligned} \log L(\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}|\mathbf{F}, \mathbf{Z}) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{ij,tk} \log \gamma_{tk} \\ &+ \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{ij,tk} \log h_{tk} \\ &\times (f_{ij}|\mathbf{x}_t, \mathbf{y}_k, \mu, \alpha_t, \beta_k). \end{aligned} \tag{9}$$

In general, the direct estimation of parameters for mixture models is a difficult task, for which the usual iterative Expectation-Maximization (EM) algorithm, or a generalized version (GEM) of this is usually employed (Dempster et al., 1977). This procedure estimates the unobserved values of $\boldsymbol{Z}$ by means of their expected value when the remaining parameters of the model, $\boldsymbol{\Theta}$, are known from a previous iteration, using the full conditional probability, $p(\boldsymbol{Z}|\hat{\boldsymbol{\Theta}}, \mathbf{F})$ (see, e.g., McLachlan and Peel (2000) for an extensive description). From these estimated values, the model parameters are then re-estimated in a M-step. The algorithm alternates between these two steps, while at each iteration the complete log-likelihood never decreases, and the process usually concludes when a certain convergence criterion is achieved.

Various estimation procedures have been proposed to apply the maximum likelihood approach to the latent block model when full conditional probabilities are difficult to obtain (see, e.g., Govaert & Nadif, 2014). In this respect, Neal and Hinton (1998) described a non-standard perspective of the EM algorithm for which at the E-step, the unknown value of $\boldsymbol{Z}$ can be viewed as representing a distribution of values, while the M-step performs maximum likelihood estimation for the joint data using these values. The same authors also observed an association between parameter estimation via the EM algorithm and a lower bound of the complete log-likelihood obtained when any probability distribution is considered for the unobserved values $\boldsymbol{Z}$. Hence, when the full conditional distribution $p(\boldsymbol{Z}|\hat{\boldsymbol{\Theta}}, \mathbf{F})$ cannot be computed, a variational approach of the EM algorithm (Jordan, Ghahramani, Jaakkola, & Saul, 1999) may be employed. Subsequently, Govaert and Nadif (2005) introduced a general variational EM algorithm for the maximum likelihood estimation of latent block models to solve the problem when difficulties arise from the dependence structure of the variables.

In this model, a GEM algorithm is employed for the parameter estimation. The algorithm was implemented in MatLab[1] and the best solution from 100 random starts was chosen as the final solution. The convergence criterion used is that the difference in subsequent log-likelihood values is less than 10^{-8} (see Appendix C for an algorithm overview). It can be shown (see Appendix A) that the proposed GEM algorithm and the variational EM approach of Govaert and Nadif (2014) results in an equivalent estimation procedure in this LBDA model.

E-step

Let $\boldsymbol{\Theta} = \{\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}\}$ be the vector of parameters of the model. In the first iteration initial values for the parameters of the model $\hat{\boldsymbol{\Theta}}^{(0)}$ are set. In general, in the sth-iteration, the conditional expectation of the log-likelihood ($\log L$) given $\mathbf{F}$, and previous estimated parameters values $\hat{\boldsymbol{\Theta}}^{(s=1)}$, can be determined from the linearity of $\log L$ on $z_{ij,tk}$ as,

[1]The program and data are available upon request

$$\mathcal{Q}\left(\Theta|\hat{\Theta}^{(s-1)}\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} E\left[z_{ij,tk}|\mathbf{F},\hat{\Theta}^{(s-1)}\right]\log\gamma_{tk} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} E\left[z_{ij,tk}|\mathbf{F},\hat{\Theta}^{(s-1)}\right] \times\log h_{tk}\left(f_{ij}|\mathbf{x}_t,\mathbf{y}_k,\mu,\alpha_t,\beta_k\right), \tag{10}$$

where $E[z_{ij,tk}|\mathbf{F},\Theta^{(s-1)}]$ denotes the expectation of $z_{ij,tk}$ in the s-th iteration.

As $\mathcal{P}(\mathbf{F})$ is a block-shaped partition of $\mathbf{F}$, we introduce row and column indicator variables such that, $z_{ij,tk} = z^R_{it}z^C_{jk}$, where $z^R_{it} = 1$ if $\mathbf{f}^R_i \in \mathbf{F}^R_t$, and $z^C_{jk} = 1$ if $\mathbf{f}^C_j \in \mathbf{F}^C_k$, and zero otherwise. Therefore it follows that,

$$\sum_{t=1}^{T} z^R_{it} = \sum_{j=1}^{J} z^C_{jk} = 1, \sum_{i=1}^{I}\sum_{t=1}^{T} z^R_{it} = I, \text{and} \sum_{j=1}^{J}\sum_{k=1}^{K} z^C_{jk} = J.$$

Let us define the $I\times T$ indicator matrix $\mathbf{Z}^R$ in terms of its row vectors $\boldsymbol{z}^R_i = (z^R_{i1},\ldots z^R_{iT})'$ for the $\mathcal{P}_I(\mathbf{F})$ partition, and the $J\times K$ indicator matrix $\mathbf{Z}^C$ in terms of its row vectors $\boldsymbol{z}^C_j = (z^C_{j1},\ldots z^C_{jK})'$ for the $\mathcal{P}_J(\mathbf{F})$ partition. Then, $\mathbf{Z} = \mathbf{Z}^R \otimes \mathbf{Z}^C$, where $\otimes$ denotes the usual Kronecker product of two matrices.

Because the unobserved $z_{ij,tk}$ are Bernoulli distributed variables, $E[z_{ij,tk}|\mathbf{F},\Theta^{(s-1)}] = P[z_{ij,tk}=1|\mathbf{F},\hat{\Theta}^{(s-1)}]$. Since $z_{ij,tk} = z^R_{it}z^C_{jk}$, and assuming (conditional) independence, $E[z_{ij,tk}|\mathbf{F},\Theta^{(s-1)}] = E[z^R_{it}|\mathbf{F},\Theta^{(s-1)}]$ $E[z^C_{jk}|\mathbf{F},\Theta^{(s-1)}]$, where $E[z^R_{it}|\mathbf{F},\Theta^{(s-1)}] = \pi^R_{it}(\Theta^{(s-1)})$ is the posterior probability that f^R_i belongs to $\mathbf{F}^R_t$, and $E[z^C_{jk}|\mathbf{F},\Theta^{(s-1)}] = \pi^C_{jk}(\Theta^{(s-1)})$ is the posterior probability that f^C_j belongs to $\mathbf{F}^C_k$. Therefore,

$$\hat{z}^{(s)}_{ij,tk} = \hat{\pi}^R_{it}\left(\Theta^{(s-1)}\right)\hat{\pi}^C_{jk}\left(\Theta^{(s-1)}\right), \tag{11}$$

and the unobserved values of $\mathbf{Z}$ are calculated by these products of posterior probabilities that are estimated using the Bayes theorem (see Appendix A).

M-step

This step requires the optimization of (10) with respect to parameters $\mathbf{X}$, $\mathbf{Y}$, μ, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$, under previously estimated values $\hat{z}^{(s)}_{ij,tk}$. First, the estimation of the unconditional probabilities under (1), is obtained by maximizing

$$\log\mathcal{L}^* = \log\mathrm{L} - \tau_R\left(\sum_{t=1}^{T}\gamma^R_t - 1\right) - \tau_C\left(\sum_{k=1}^{K}\gamma^C_k - 1\right), \tag{12}$$

where τ_R and τ_C are Lagrange multipliers. It can easily be shown that the expressions for the estimators of γ^R_t and of γ^C_k in the s-th iteration are given by

$$\hat{\gamma}^{R^{(s)}}_t = \frac{1}{I}\sum_{i=1}^{I}\hat{z}^{(s)}_{it} \text{ and } \hat{\gamma}^{C^{(s)}}_k = \frac{1}{J}\sum_{j=1}^{J}\hat{z}^{(s)}_{jk}, \tag{13}$$

and therefore, $\hat{\gamma}^{(s)}_{tk} = \hat{\gamma}^{R^{(s)}}_t\hat{\gamma}^{C^{(s)}}_k$.

The remaining parameters of the model are estimated by maximizing (10) under previously estimated values of $\hat{\mathbf{Z}}^{(s)}$, which is equivalent to maximizing

$$q\left(\mathbf{X},\mathbf{Y},\mu,\boldsymbol{\alpha},\boldsymbol{\beta}|\hat{\mathbf{Z}}^{(s)}\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K}\hat{z}^{(s)}_{ij,tk}\log h_{tk} \times\left(f_{ij}|\mathbf{x}_t,\mathbf{y}_k,\mu,\alpha_t,\beta_k\right). \tag{14}$$

Details of the parameter estimation at this step using the Newton-Raphson procedure are shown in Appendix B.

Model selection

In the estimation procedure the number of clusters for the rows T and for the columns K is assumed to be known, as well as the number of dimensions M for the configuration. Since these values are unknown in many practical situations, a model selection procedure is needed, and the BIC (Schwarz, 1978) criterion is an appropriate alternative in this framework (see McLachlan & Peel, 2000). The sample size adjustment suggested by Rissanen (1978) for the BIC will be used here, for which the number of cells IJ is adjusted by $(IJ+2)/24$ (Vera et al., 2014).

The adjusted criterion is defined by $BIC^* = -2\log L + l\log h$, where $h = (IJ+2)/24$, and l is the number of parameters to be estimated. Without geometrical constraints μ_{tk} is the average frequency in a corresponding block representing TK parameters, which together with the prior probability parameters makes $l = TK + (T-1) + (K-1)$. When geometrical constraints are imposed, there are $(1+T+K)$ parameters for the marginal-class effects, $(T+K)M$ for both configurations and $(T-1)+(K-1)$ for the prior probabilities. Nevertheless, after identification (see Appendix C) the number for the marginal-class effects is reduced to $1+(T-1)+(K-1)$. Moreover, the singular value decomposition introduces $M(M+2)$ constraints. Therefore, in this situation the number of parameters to be estimated is $l = 1+2(T-1)+2(K-1)+M(T+K-M-2) = (M+2)(T+K-M)-3$. Since the number of parameters when geometrical constraints are imposed should not be greater than when they are not considered, $(T+K)M-M(M+$

$2) \leq (T-1)(K-1)$, and the maximum number of dimensions $M \leq \min(T, K)-1$.

The usual procedure to determine the number of clusters is to calculate, without considering geometrical constraints, the value of the BIC^* statistic, for a range of predetermined values of T and K; therefore, the optimal combination is the one related to the lowest BIC^* value. The dimensionality is then determined by again minimizing the BIC^* statistic for different values of M, with fixed values of T and K.

Model interpretation

Once the most appropriate model has been selected, row and column profile classes can be interpreted in terms of the original variables. To this end we use the size of the classes, in terms of the number of profiles and of the number of elements. After obtaining the parameters, the $\boldsymbol{Z}^R$ are estimated such that $z_{it}^R = 1, t = \text{argmax}_t(\hat{\pi}_{it}^R)$, and zero otherwise, and equivalently for the $\boldsymbol{Z}^C$ (see Appendix D). By this approach, we know exactly which profiles are assigned to each class (and so its attributes are identified), and how many participants are related to each profile (and therefore how many participants present the variable attributes that constitute this profile). Hence, for each class, we have the total number of participants in each modality of each variable that shape the profiles in the class, as well as the number of profiles and the number of participants in the class. Then for every class we determine the conditional probability of a category of the original variables given the class membership.

To interpret the associations we use an association plot in which the row and column classes are represented. The odds of a column class k against a column class k' for a given row class t, are given by

$$\log\left(\frac{\mu_{tk}}{\mu_{tk'}}\right) = \log(\beta_k) - \log(\beta_{k'}) - d_{tk}^2 + d_{tk'}^2. \qquad (15)$$

The odds are a function of the main effect parameters and the distances. In the latter respect, the odds are in favor of the closest category. Concerning the main effects the odds are in favor of the category with the largest β value. For a detailed discussion of the interplay between the β's and distances see Takane (1998) and De Rooij (2009). The odds ratio is another useful tool, which can be defined in terms of squared distances (see De Rooij & Heiser, 2005)

$$\frac{\mu_{tk} \times \mu_{t'k'}}{\mu_{tk'} \times \mu_{t'k}} = \exp\left(-d_{tk}^2 - d_{t'k'}^2 + d_{tk'}^2 + d_{t'k}^2\right). \qquad (16)$$

Experimental results

Monte carlo experiments

Simulation study I

To test the performance of the model, twenty twelve-point matrices were simulated in two dimensions representing the cluster centers for the LBDA model. From each set of twelve points, $T = 7$ row and $K = 5$ column cluster centers were selected, such that the intermixedness index score (Busing, Groenen, & Heiser, 2005) was below 0.05. The distances d_{tk} between the row and column cluster centers were then calculated. Three conditions were distinguished, in which each cluster had an average of 20, 40 or 60 elements, by means of n_t and n_k rounded random draws of the normal distribution with mean values of 20, 40 and 60, and standard deviations of 5, 10 and 15, respectively. Thus, data matrices with average sizes ranging from 140 by 100 to 420 by 300 were created, in which each table presented seven row clusters and five column clusters.

With these elements, the expected frequencies for sixty contingency tables were obtained on the basis of the multiplicative model (Equation (2)) taking $\mu_{tk} = 100\exp(-d_{tk}^2)$, with d_{tk} as the distances between the cluster centers and equating the remaining main effect parameters to unity in order to make the distances monotonically related to the joint probabilities and thus suitable for comparison (see Vera et al., 2014, Takane, 1998). For each table, identified parameter values for the model were obtained from these expected frequencies, following the procedure described in Appendix C. These parameter values then constituted the reference values for comparison. The observed frequencies were obtained by reference to the Poisson distribution. Each contingency table thus generated was analyzed, without imposing geometrical constraints, to determine the number of clusters (four to ten clusters for the rows and two to eight clusters for the columns); for all matrices the lower BIC^* value was in agreement with the correct number of clusters.

Each of the contingency tables generated was then analyzed subject to geometrical constraints for $T = 7$ and $K = 5$ clusters in two, three and four dimensions, and the lower BIC^* value was always obtained in two dimensions. The estimated parameter values in two dimensions were compared with the true ones in terms of the row and column classifications recovered, and of the configuration arising from Procrustes analysis (Cliff, 1966). In all datasets, the original partition was correctly recovered. The normalized Procrustes

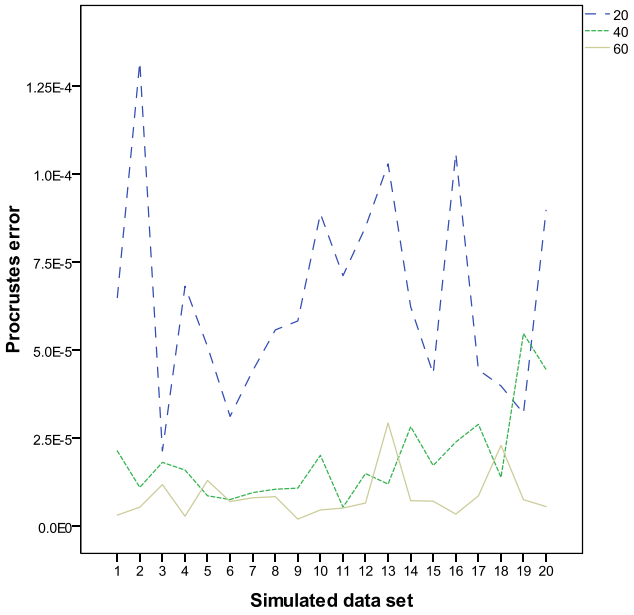

Figure 1. Normalized Procrustes sum of squared errors between simulated and recovered configurations, for the 20 contingency tables in the three cluster size conditions.

sum of squared errors produced average (near zero) values of 0.0000645, 0.0000188, and 0.0000084, for cluster sizes of 20, 40 and 60, respectively, thus reflecting a good match between the estimated and the true configurations for all the data sets simulated, as shown in Figure 1.

From the simulation process, after identification, the overall scale parameter μ showed averaged values of 85.23, 82.34, and 84.09 for cluster sizes of 20, 40 and 60, respectively. The largest differences between true and estimated values for μ in terms of absolute residuals were 0.5124, 0.6378, and 0.2575, respectively, for these cluster sizes. In terms of the marginal effects, after identification, the original values in the 20 simulated tables and for the three cluster sizes gave the following results: for α_t in the intervals $(0.5215, 2.2449)$, $(0.4227, 1.8333)$ and $(0.5046, 2.0201)$, and for β_k in the intervals $(0.5233, 2.0540), (0.4214, 2.3968)$ and $(0.4924, 2.2352)$, respectively. The largest differences between the true and estimated values, in terms of absolute residuals, were 0.0133, 0.0106 and 0.0058 for the row effects and 0.0115, 0.0121, and 0.0093 for the column effects, in all the datasets, for the respective cluster sizes. Thus, in all datasets, the estimated parameter values were close to their true values, and closer still for the larger cluster sizes.

Simulation study II

In this simulation experiment we investigate the influence of random zero entries on the stability of the cluster structure in terms of the LBDA model. It is important to note the difference between structural-zero entries, which would be expected in a profile-by-profile cross-classification table, and sampling-zero entries, which are zero entries with expected values that are not required to be zero (see, e.g., Baker, Clarke, & Lane, 1985). In general, equating to zero a number of entries in a block clustered table may originate a new partition in a different cluster structure. This is a well-known problem for sparse tables, and does not affect the DA model alone.

Table 1. Results for artificially sparse tables.

Density	%ENC	%WRC	$\bar{p}$	SD
60	45	25	0.00017	0.00020
70	50	25	0.00018	0.00031
80	70	60	0.03512	0.11332
90	85	80	0.04712	0.13035

For each density value the percentage of recovery classifications in the same number of clusters as in the original nonsparse datasets (%ENC), and the percentage of well recovered classifications (%WRC) are shown. In terms of comparable configurations, the average Procrustes error (p) and standard deviation (SD) values are shown.

Taking the previous comment into account, eighty clustered contingency tables were obtained following the above-described procedure, for $T = 5$ and $K = 3$ clusters, each with a cluster size of 100. Density values of 0.60, 0.70, 0.80 and 0.90, respectively, were considered in each set of twenty simulated tables, by equating to zero a percentage of randomly selected cells given by $(1-\text{density}) * 100$. These sparse contingency tables were analyzed with the LBDA model, labeling the elements in each estimated cluster as in the original nonsparse table. The labels of the elements in each cluster were then tabulated, and corresponding row and column clusters were labeled with the most frequently observed label. Tables in which the number of clusters was lower than in the original ones (i.e., different clusters with equal label value) were not considered for the Procrustes analysis.

For the twenty data sets analyzed in each density, Table 1 shows the percentage of recovered partitions in the same number of clusters (%ENC) as in the original nonsparse datasets, as well as the percentage of classifications that, as well as presenting the same number of clusters, also matched in terms of cluster memberships (%WRC). The mean Procrustes error (p) and related standard deviation (SD) for comparable configurations after identification are also shown. Few block-shaped partitions were recovered in the same number of clusters for fairly sparse tables, since the estimated row and/or column partitions produced a different number of clusters than for the corresponding nonsparse data set. Nevertheless, the Procrustes values were low in all comparable

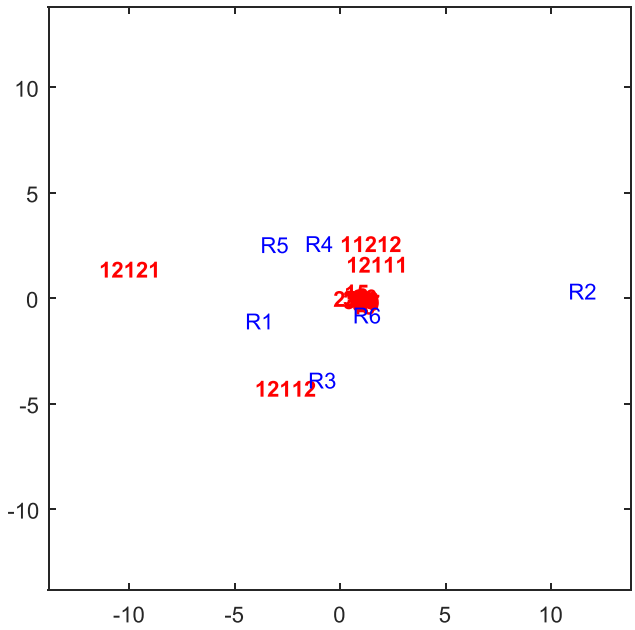

Figure 2. Association plot for the six row-clusters and 32 columns in the LCDA model.

partitions, which indicates that in spite of the presence of zeros the model represents associations quite well, and therefore preserves the odds ratio in datasets. Thus, for comparable configurations, the LBDA model is shown to perform well.

Empirical data

LCDA model

The empirical data presented in the introduction are now revisited. First, the personality data set was analyzed with the LCDA model of Vera et al. (2014) in a two-step procedure for a Poisson sampling scheme. In the first step, the LCDA model was run to cluster the rows and represent associations between row clusters and columns categories. The model with $T = 6$ clusters in two dimensions was related to the lower BIC^* value. The 6×32 configuration shown in Figure 2 is difficult to interpret because most of the response categories are close together in the origin. Moreover, the resulting contingency table in this model remains sparse (density of 0.7657). In the second step, the LCDA model was applied to cluster the columns of the complete data set, and the lower BIC^* value was again obtained for $K = 6$ clusters (now in the columns) and in two dimensions. The resulting 434×6 configuration was largely uninterpretable.

LBDA model

The models with 2 to 20 classes for the rows, i.e., the profiles based on gender and personality, and 2 to 5 classes for the columns, i.e., the mental disorder profiles, were investigated. First, the number of classes was determined without geometric constraints, and then the dimensionality was determined.

Table 2. BIC^* values for the personality dataset and each combination of row and column classes (2 to 20 classes for the rows and 2 to 5 classes for the columns).

	2	3	4	5
2	13,877.55	13,799.75	13,797.91	13,814.78
3	13,671.20	13,640.84	13,635.88	13,657.20
4	13,632.01	13,578.70	13,582.57	13,615.06
5	13,613.06	13,561.00	13,580.14	13,616.61
6	13,599.14	13,568.14	13,600.42	13,644.54
7	13,611.65	13,590.34	13,629.20	13,678.40
8	13,629.23	13,611.84	13,657.99	13,714.71
9	13,644.01	13,636.13	13,688.38	13,751.49
10	13,664.14	13,659.00	13,719.76	13,788.48
11	13,680.01	13,683.69	13,750.09	13,826.05
12	13,699.20	13,708.58	13,781.96	13,864.09
13	13,715.17	13,732.96	13,813.02	13,901.33
14	13,735.13	13,757.89	13,845.31	13,939.43
15	13,754.46	13,782.92	13,876.86	13,977.53
16	13,772.79	13,808.33	13,907.95	14,015.67
17	13,791.72	13,833.20	13,940.28	14,053.75
18	13,810.41	13,859.37	13,971.50	14,091.72
19	13,828.28	13,884.54	14,003.78	14,129.98
20	13,848.24	13,910.00	14,035.22	14,168.18

Table 3. Some characteristics of the five row-classes and the three column-classes.

Row classes			Column classes		
Class	r_t	n_t	Class	r_k	n_k
R1	267	893	C1	1	1266
R2	81	544	C2	7	1058
R3	16	318	C3	24	614
R4	67	961			
R5	3	222			

r_t and r_k give the number of membership profiles in class R_t or C_k, n_t and n_k give the number of participants in the class.

Table 2 shows the BIC^* values obtained, where the model with $T = 5$ and $K = 3$ is related to the lower value. The BIC^* values for $M = 1$, 2 were 13,543.96 and 13,561.35, respectively, which corroborates the one dimensional solution. These values are smaller than those corresponding to the solutions derived by the LCDA model, as might be expected. Specifically, the BIC^* value of 14,410.09 was obtained for the $T = 6$, $K = 32$ model, where $\boldsymbol{Z}^C$ is the identity matrix on dimension J, and a corresponding value of 27,937.09 was obtained for the $T = 434$, $K = 6$ model, where $\boldsymbol{Z}^R$ is the identity matrix on dimension I.

Therefore we choose the model with five row classes and three column classes in one dimension. Table 3 shows the number of profiles and number of participants in each class. Thus, the first row class contains 267 different profiles, with a total of 893

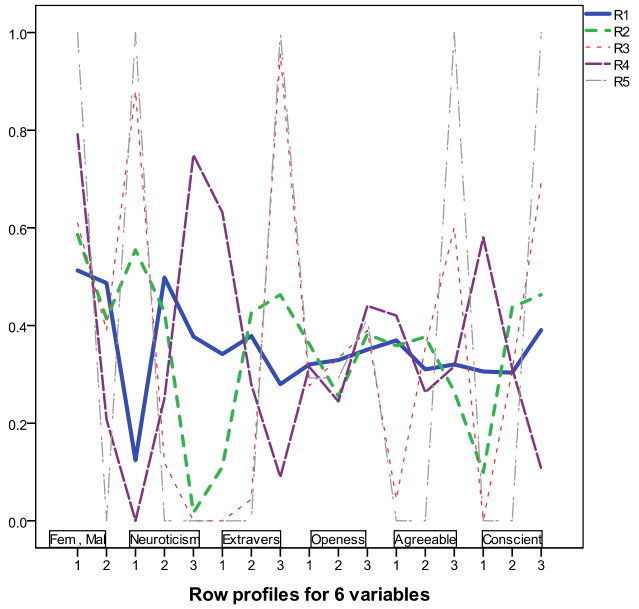

Figure 3. Representation of the probability of gender (1: female, 2: male), and of answering 1: low, 2: medium, and 3: high in personality variables, for each of the five row-classes, given the class membership.

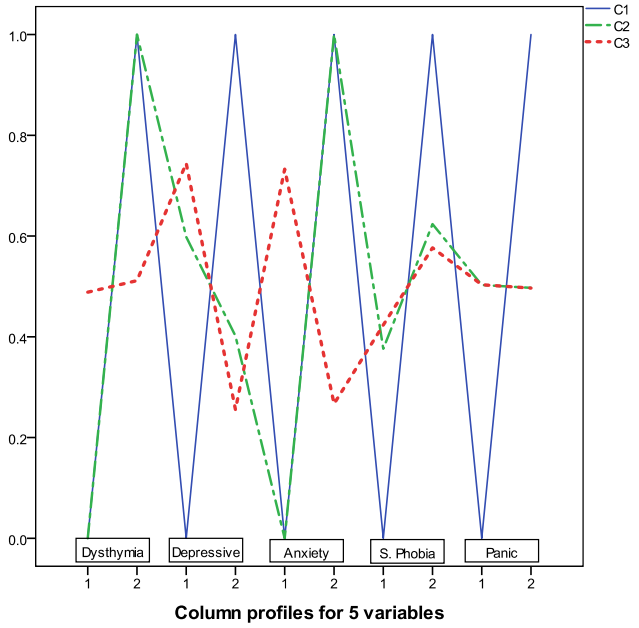

Figure 4. Probabilities of presenting (1) or not presenting (2) each personality disorder, given the class membership.

participants, while the first column class represents a single profile with 1,266 participants.

For the rows (Figure 3), class R1 consists of 267 profiles and 893 subjects. Gender is distributed evenly (51% female) and the personality variables are distributed likewise. Class R2 consists of 81 profiles and 544 subjects. This class has a slight female majority (59%), scores low or medium on neuroticism, medium or high on extraversion and conscientiousness and evenly on agreeableness and openness. Class R3 consists of 16 profiles and 318 subjects. Gender (61% female) scores low on neuroticism, high on extraversion and conscientiousness, and medium or high on agreeableness. Class R4 consists of 67 profiles and 961 subjects. The class is mainly female (79%) scores high on neuroticism and low on extraversion and conscientiousness. Class R5 consists of three profiles and 222 subjects. This class consists exclusively of female subjects (100%), scores low on neuroticism and high on extraversion, agreeableness, and conscientiousness.

There is a striking similarity between these latent classes and those obtained by Spinhoven, De Rooij, Heiser, Smit, and Penninx (2012). Class R4 corresponds to what Spinhoven et al. called the High Overcontrollers, class R1 to the Low Overcontrollers, class R3 to the Medium Resilients, and class R5 to the High Resilients.

For the columns (Figure 4), the first class (*C1*) consists of one profile with 1,266 subjects. This is the profile with no disorders, i.e., the healthy profile. The second class category (*C2*) category has seven profiles and 1,058 subjects. It has zero probability for dysthymia and generalized anxiety disorder, and evenly probability of presenting or not a major depressive disorder, social phobia, and panic disorder. The third category (*C3*) consists of 24 profiles with 614 subjects. This class has a high probability of presenting each disorder, i.e., this is the comorbid class.

The log of the β for column class C1 is $\log(\beta_1) = 2.93$, for column class C2 it is $\log(\beta_2) = -0.51$, and for column class C3 it is $\log(\beta_3) = -2.42$. The first class is very dominant in terms of the main effect, while the third class has a very low value. This strongly influences the odds (see Equation (15)); thus for every row class the odds are in favor of column class C1, and never in favor of class C3.

Figure 5 gives the one-dimensional configuration, showing that the subjects in row classes R4 and *R1* are the most vulnerable to comborbid disorders (column class C3), whereas the subjects in row class R5 are probably healthy. In more detail, the odds ratios (see Equation (16)) are only based on the distances, which can be read from Figure 5. The odds of patients with a row profile in class R4 being diagnosed with a profile of column class C3 rather than one in column class C2 are 1.5 greater than those for row class R1, 4.4 times greater than those for row class R2, 6.0 times greater than those for row class R3, and 8.7 times greater than those for row class R5.

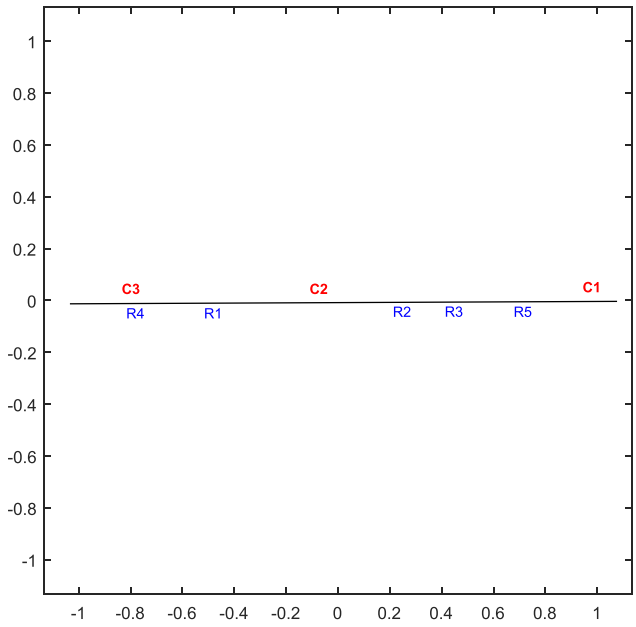

Figure 5. Association plot for row (*R*) and column (*C*) classes in one dimension.

Similarly, the odds of patients with a row profile in class R5 being diagnosed with the healthy profile (*C1*) instead of the profile in column class C2 are 23.2 times greater than those for row class R4, 11.8 times greater than those for row class R1, 2.6 times greater than those for row class R2, and 1.7 times greater than those for row class R3. Hence, the odds ratio can be understood in terms of the distances, and the row classes in the neighborhood of column classes have a relatively high odds ratio for that class compared to any other class.

Conclusion

In this article we propose a latent-block distance association model (LBDA) to analyze the relationships between categorical variables. The model allows us to block cluster the outcomes of two categorical variables while simultaneously representing the row and column classes in a low-dimensional Euclidean space. The estimated distances between cluster centers inversely describe the association between the partitioned variables.

It is assumed that the data are independent counts related to two categorical response variables, or in general, two different sets of response variables. In this latter general framework, any combination of the categories of the variables in a set is called a profile, and the data consist of a profile-by-profile contingency table. In this situation, contingency tables are usually large, and often present many zero entries, i.e., the contingency table is sparse. When there are zero entries in a contingency table, the estimated odds ratios are either zero, infinity, or undefined, and standard methods for categorical data analysis with sparse tables may encounter estimation problems.

In this study, the GEM algorithm is employed for parameter estimation, using a Poisson sampling scheme. This model can be viewed as an extension of the LCDA model (Vera et al., 2014), which facilitates the representation of associations for tables with large numbers of column modalities, possibly with many zero entries. We propose the Bayesian information criterion as a useful means of determining the number of latent classes for the rows and for the columns, as well as the dimensionality of the representation.

The well-known problem of the local optimum in the likelihood equation, as well as a slow rate of convergence and the dependence on appropriate initial values of the GEM algorithm (McLachlan & Peel, 2000; Shireman, Steinley, & Brusco, 2016) are also experienced with our algorithm. Therefore, the performance of other co-clustering estimation procedures within the DA framework, together with other sampling schemes, constitute interesting areas for future research. Another topic that remains to be studied is that of another, closely-related model, in which the association parameters are clustered (either for the rows or columns separately or together as in the present article) but the main effect parameters are not.

Article Information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant ECO2013-48413-R from the Spanish Ministry of Economy and Competitiveness (co-financed by FEDER) and Grant RTI2018-099723-B-I00 from the Spanish Ministry of Science, Innovation and Universities (co-financed by FEDER).

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and

conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors thank the Action Editor, Douglas Steinley and two anonymous reviewers for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions or the funding agencies is not intended and should not be inferred.

ORCID

J. Fernando Vera http://orcid.org/0000-0002-6499-7132
Mark de Rooij http://orcid.org/0000-0001-7308-6210

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). New York: Wiley.

Baker, R. J., Clarke, M. R. B., & Lane, P. W. (1985). Zero entries in contingency tables. *Computational Statistics and Data Analysis*, *3*, 33–45. doi:10.1016/0167-9473(85)90056-8

Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology*, *9*(2), 291–319. doi:10.1017/S095457949700206X

Becker, M. P. (1990). Maximum likelihood estimation of the RC(m) association model. *Applied Statistics*, *39*(1), 152–167. doi:10.2307/2347833

Bock, H.-H. (2003). Two-way clustering for contingency tables: Maximizing a dependence measure. In M. Schader, W. Gaul, & M. Vichi (Eds.), *Between data science and applied data analysis. Studies in classification, data analysis, and knowledge organization* (pp. 143–154). Berlin: Springer. doi:10.1007/978-3-642-18991-3_17

Busing, F. M. T. A., Groenen, P. J. F., & Heiser, W. J. (2005). Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, *70*(1), 71–98. doi:10.1007/s11336-001-0908-1

Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika*, *31*(1), 33–42. doi:10.1007/BF02289455

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38. doi:10.1111/j.2517-6161.1977.tb01600.x

De Rooij, M. (2007). The distance perspective of generalized biadditive models: Scalings and transformations. *Journal of Computational and Graphical Statistics*, *16*(1), 210–227. doi:10.1198/106186007X180101

De Rooij, M. (2008). The analysis of change, Newton's law of gravity, and Association models. *Journal of the Royal Statistical Society, Series A*, *171*, 137–157. doi:10.1111/j.1467-985x.2007.00498.x

De Rooij, M. (2009). Ideal point discriminant analysis revisited with an emphasis on visualization. *Psychometrika*, *74*(2), 317–330. doi:10.1007/s11336-008-9105-9

De Rooij, M., & Heiser, W. J. (2005). Graphical representations and odds ratios in a distance association model for the analysis of cross-classified data. *Psychometrika*, *70*(1), 99–122. doi:10.1007/s11336-000-0848-1

Goodman, L. A. (1981). Criteria for determining whether certain categories in a cross-classification table should be combined, with special reference to occupational categories in an occupational mobility table. *American Journal of Sociology*, *87*(3), 612–650. doi:10.1086/227498

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetric models for contingency tables with or without missing entries. *The Annals of Statistics*, *13*(1), 10–69. doi:10.1214/aos/1176346576

Govaert, G. (1995). Simultaneous clustering of rows and columns. *Control and Cybernetics*, *24*(4), 437–458. http://control.ibspan.waw.pl:3000/contents/export?filename=1995-4-06_govaert.pdf

Govaert, G., & Nadif, M. (2005). An EM algorithm for Block Mixture Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27* (4), 643–647. doi:10.1109/TPAMI.2005.69

Govaert, G., & Nadif, M. (2014). *Co-clustering: Models, algorithms and applications*. Hoboken, NJ: Wiley. doi:10.1002/9781118649480

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). *Machine Learning*, *37*(2), 183–533. doi:10.1023/A:1007665907178

Kateri, M., & Iliopoulos, G. (2003). On collapsing categories in two-way contingency tables. *Statistics*, *37*, 443–455. doi:10.1080/0233188031000123780

Kim, S., Choi, J. Y., & Hwang, H. (2017). Two-way regularized fuzzy clustering of multiple correspondence analysis. *Multivariabe Behavioral Research*, *52* (1), 31–46. doi:10.1080/00273171.2016.1246996

Li, J., & Zha, H. (2006). Two-way Poisson mixture models for simultaneous classification and word clustering. *Computational Statistics and Data Analysis*, *50*(1), 163–180. doi:10.1016/j.csda.2004.07.013

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models. Wiley series in probability and statistics*. New York: Wiley. doi:10.1002/0471721182

Neal, R., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 143–154). Cambridge, MA: MIT Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471. doi:10.1016/0005-1098(78)90005-5

Schepers, J., Bock, H.-H., & Van Mechelen, I. (2017). Maximal interaction two-mode clustering. *Journal of Classification*, *34*(1), 49–75. doi:10.1007/s00357-017-9226-x

Shireman, E. M., Steinley, D., & Brusco, M. J. (2016). Local optima in mixture modeling. *Multivariate Behavioral Research*, *51*(4), 466–481. doi:10.1080/00273171.2016.1160359

Schwarz, G. (1978). Estimating the dimensions of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Spinhoven, P., De Rooij, M., Heiser, W. J., Penninx, B., & Smit, J. (2009). The role of personality in comorbidity

among anxiety and depressive disorders in primary care and specialty care: A cross sectional analysis. *General Hospital Psychiatry*, *31*(5), 470–477. doi:10.1016/j.genhosppsych.2009.05.002

Spinhoven, P., De Rooij, M., Heiser, W. J., Smit, J., & Penninx, B. (2012). Personality and changes in comorbidity patterns among anxiety and depressive disorders. *Journal of Abnormal Psychology*, *121*(4), 874–884. doi:10.1037/a0028234

Takane, Y. (1998). Visualization in ideal point discriminant analysis. In J. Blasius & M.J. Greenacre (Eds.), *Visualization of categorical data* (pp. 441–459). New York: Academic Press. doi:10.1016/b978-012299045-8/50034-6

Vera, J. F., Macías, R., & Heiser, W. J. (2009a). A latent class multidimensional scaling model for two-way one-mode continuous rating dissimilarity data. *Psychometrika*, *74*(2), 297–315. doi:10.1007/s11336-008-9104-x

Vera, J. F., Macías, R., & Heiser, W. J. (2009b). A dual latent class unfolding model for two-way two-mode preference rating data. *Computational Statistics and Data Analysis*, *53*(8), 3231–3244. doi:10.1016/j.csda.2008.07.019

Vera, J. F., Macías, R., & Heiser, W. J. (2013). Cluster differences unfolding for two-way two-mode preference rating data. *Journal of Classification*, *30*(3), 370–396. doi:10.1007/s00357-017-9226-x

Vera, J. F., de Rooij, M., & Heiser, W. J. (2014). A latent class distance association model for cross-classified data with a categorical response variable. *British Journal of Mathematical and Statistical Psychology*, *67*(3), 514–540. doi:10.1111/bmsp.12038

Worku, H. M., & de Rooij, M. (2018). A multivariate logistic distance model for the analysis of multiple binary responses. *Journal of Classification*, *35* (1), 124–146. doi:10.1007/s00357-018-9251-4

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *The Annals of Applied Statistics*, *5* (4), 2493–2518. doi:10.1214/11-AOAS493

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, *44* (4), 1049–1060. doi:10.2307/2531734

Appendix A

Estimation of posterior probabilities

Assume $\hat{\Theta} = \{\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Gamma}}\}$ is known (e.g., from the previous iteration step). At the E-step, the expected value of the log-likelihood (10) is given in terms of $E[z_{ij,tk}|\mathbf{F}, \hat{\Theta}] = P[z_{ij,tk} = 1|\mathbf{F}, \hat{\Theta}]$.

Assuming independence between the row and column indicator variables of $\mathbf{F}$, since $z_{ij,tk} = z^R_{it} z^C_{jk}$, we have that $E[z_{ij,tk}|\mathbf{F}, \hat{\Theta}] = E[z^R_{it}|\mathbf{F}, \hat{\Theta}]E[z^C_{jk}|\mathbf{F}, \hat{\Theta}]$, that is, $P[z^R_{it} = 1, z^C_{jk} = 1|\mathbf{F}, \hat{\Theta}] = P[z^R_{it} = 1|\mathbf{F}, \hat{\Theta}]P[z^C_{jk} = 1|\mathbf{F}, \hat{\Theta}]$. In this model the conditional probabilities $P[z^R_{it} = 1|z^c_{jk} = 1, \mathbf{F}, \hat{\Theta}] = P[z^R_{it} = 1|\mathbf{F}, \hat{\Theta}]$, and $P[z^C_{jk} = 1|z^R_{it} = 1, \mathbf{F}, \hat{\Theta}] = P[z^C_{jk} = 1|\mathbf{F}, \hat{\Theta}]$, taking into account the independence of z^R_{it} and z^C_{jk} for each pair (i, j) (see, e.g., Agresti, 2013, Section 2.3.4).

Then, denote by $f_{ik} = \sum_{j=1}^{J} z^C_{jk} f_{ij\cdot}$, $i = 1, ..., I, k = 1..., K$, where $\hat{\mathbf{Z}}^C$ is a known classification of the columns of $\mathbf{F}$. If $f_{ij} \in \mathbf{F}_{tk}$, the probability that $\mathbf{f}^R_i \in \mathbf{F}^R_t$ can be expressed as

$$\begin{aligned} h^R_t(\mathbf{f}^R_i|\mathbf{x}_t, \mathbf{Y}, \mu, \alpha_t, \boldsymbol{\beta}) &= \prod_{k=1}^{K}\prod_{j=1}^{J}\left(\frac{\mu^{f_{ij}}_{tk}}{f_{ij}!}\exp(-\mu_{tk})\right)^{z^C_{jk}} \\ &= \prod_{k=1}^{K}\frac{\mu^{f_{ik}}_{tk}\exp\left(-\sum_{j=1}^{J} z^C_{jk}\mu_{tk}\right)}{\prod_{j=1}^{J}(f_{ij}!)^{z^C_{jk}}} \\ &= \prod_{k=1}^{K}\frac{\left(\sum_{j=1}^{J} z^C_{jk}\mu_{tk}\right)^{f_{ik}}\exp\left(-\sum_{j=1}^{J} z^C_{jk}\mu_{tk}\right)}{\left(\sum_{j=1}^{J} z^K_{jk}\right)^{f_{ik}}\prod_{j=1}^{J}(f_{ij}!)^{z^K_{jk}}} \\ &= \prod_{k=1}^{K}\frac{\dot{\mu}^{f_{ik}}_{tk}\exp(-\dot{\mu}_{tk})}{\left(\sum_{j=1}^{J} z^C_{jk}\right)^{f_{ik}}\prod_{j=1}^{J}(f_{ij}!)^{z^C_{jk}}}, \end{aligned} \tag{A.1}$$

which except for a constant term, is a product of Poisson distributions of parameter $\dot{\mu}_{tk} = \sum_{j=1}^{J} z^C_{jk}\mu_{tk}$. Then, since it is unknown in advance to which block a row belongs, the p.d.f of the random variable $\mathbf{f}^R_i$ is a mixture distribution given by

$$g^R(\mathbf{f}^R_i|\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}) = \sum_{t=1}^{T}\gamma_t h^R_t(\mathbf{f}^R_i|\mathbf{x}_t, \mathbf{Y}, \mu, \alpha_t, \boldsymbol{\beta}), \tag{A.2}$$

and therefore, the posterior probabilities are expressed as

$$\pi^R_{it}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Gamma}}) = \frac{\hat{\gamma}_t h^R_t\left(\mathbf{f}^R_i|\hat{\mathbf{x}}_t, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\alpha}_t, \hat{\boldsymbol{\beta}}\right)}{g^R\left(\mathbf{f}^R_i|\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Gamma}}\right)}. \tag{A.3}$$

Equivalently, given $\hat{\mathbf{Z}}^R$, if $\mathbf{f}_{ij} \in \mathbf{F}_{tk}$, the posterior probability that $\mathbf{f}^C_j$ belongs to $\mathbf{F}^C_k$ is expressed as

$$\pi^C_{jk}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Gamma}}) = \frac{\hat{\gamma}_k h^C_k\left(\mathbf{f}^C_j|\hat{\mathbf{X}}, \hat{\mathbf{y}}_k, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\beta}_k\right)}{g^C\left(\mathbf{f}^C_k|\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Gamma}}\right)}, \tag{A.4}$$

where denoting by $f_{tj} = \sum_{i=1}^{R} z_{it} f_{ij}$, and $\ddot{\mu}_{tk} = \sum_{i=1}^{I} z^R_{it}\mu_{tk}$, the p.d.f. of the variable $\mathbf{f}^C_j$ is given by

$$g^C\left(\mathbf{f}^C_j|\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}\right) = \sum_{k=1}^{K}\gamma_k h^C_k\left(\mathbf{f}^C_j|\mathbf{X}, \mathbf{y}_k, \mu, \boldsymbol{\alpha}, \beta_k\right), \tag{A.5}$$

with

$$\begin{aligned} h^C_k\left(\mathbf{f}^C_j|\mathbf{X}, \mathbf{y}_k, \mu, \boldsymbol{\alpha}, \beta_k\right) &= \prod_{t=1}^{T}\prod_{i=1}^{I}\left(\frac{\mu^{f_{ij}}_{tk}}{f_{ij}!}\exp(\mu_{tk})\right)^{z^R_{it}} \\ &= \prod_{t=1}^{T}\frac{\ddot{\mu}^{f_{ij}}_{tk}\exp(-\ddot{\mu}_{tk})}{\left(\sum_{i=1}^{I} z^R_{it}\right)^{f_{tj}}\prod_{i=1}^{I}(f_{ij}!)^{z^R_{it}}}. \end{aligned} \tag{A.6}$$

Therefore, at the E-step, $E[z_{ij,tk}|\mathbf{F}, \hat{\Theta}] = \pi^R_{it}(\hat{\Theta})\pi^C_{jk}(\hat{\Theta})$.

Variational EM approach

The same estimators are found if the variational EM approximation is employed. As shown by Govaert and Nadif (2014), the variational EM criterion replaces the maximization of the log-likelihood by that of the fuzzy criterion,

$$\begin{aligned} G\left(\tilde{\mathbf{\Pi}}^R, \tilde{\mathbf{\Pi}}^C, \mathbf{\Theta}\right) = & \sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{\pi}_{it}^R \log\left(\gamma_t\right) + \sum_{j=1}^{J}\sum_{k=1}^{K}\tilde{\pi}_{jk}^C \log\left(\gamma_k\right) \\ & + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K}\tilde{\pi}_{it}^R\tilde{\pi}_{jk}^C \log\left(h_{tk}\left(f_{ij}|\mathbf{x}_t, \mathbf{y}_k, \mu, \alpha_t, \beta_k\right)\right) \\ & + H\left(\tilde{\mathbf{\Pi}}^R\right) + H\left(\tilde{\mathbf{\Pi}}^C\right), \end{aligned} \tag{A.7}$$

where $H(\tilde{\mathbf{\Pi}}^R) = -\sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{\pi}_{it}^R \log\left(\tilde{\pi}_{it}^R\right)$ and $H(\tilde{\mathbf{\Pi}}^C) = -\sum_{j=1}^{J}\sum_{k=1}^{K}\tilde{\pi}_{jk}^C \log\left(\tilde{\pi}_{jk}^C\right)$ are the entropy of distributions $\tilde{\mathbf{\Pi}}^R$ and $\tilde{\mathbf{\Pi}}^C$ respectively. The maximization of this function is given in an alternating estimation procedure as in the GEM algorithm. In one step, G is maximized in terms of $\tilde{\mathbf{\Pi}}^R$ and $\tilde{\mathbf{\Pi}}^C$, for fixed values of $\mathbf{\Theta}$, and in a second step, G is maximized in terms of $\mathbf{\Theta}$ for fixed values of $\tilde{\mathbf{\Pi}}^R$ and $\tilde{\mathbf{\Pi}}^C$, in a M-step that is equivalent to that of the traditional GEM algorithm. These fuzzy classification matrices are estimated by means of the following alternating estimation procedure, in which $\tilde{\mathbf{\Pi}}^R$ is estimated for fixed values of $\tilde{\mathbf{\Pi}}^C$, and vice versa. To estimate $\tilde{\mathbf{\Pi}}^R$, given $\hat{\tilde{\mathbf{\Pi}}}^C$, the only item that must be maximized is

$$\begin{aligned} \tilde{G}\left(\tilde{\mathbf{\Pi}}^R|F, \hat{\mathbf{\Theta}}, \hat{\tilde{\mathbf{\Pi}}}^C\right) = & \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K}\tilde{\pi}_{it}^R\hat{\tilde{\pi}}_{jk}^C \log\left(h_{tk}\left(f_{ij}|\hat{\theta}_{tk}\right)\right) \\ & + \sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{\pi}_{it}^R \log\left(\hat{\gamma}_t\right) - \sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{\pi}_{it}^R \log\left(\tilde{\pi}_{it}^R\right) \\ & - \sum_{i=1}^{I}\tau_i\left(\sum_{t=1}^{T}\tilde{\pi}_{it}^R - 1\right) = \sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{\pi}_{it}^R\left(a_{it} - \log\tilde{\pi}_{it}^R\right) \\ & - \sum_{i=1}^{I}\tau_i\left(\sum_{t=1}^{T}\tilde{\pi}_{it}^R - 1\right) \end{aligned} \tag{A.8}$$

where $\hat{\theta}_{tk} = (\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_k, \hat{\mu}, \hat{\alpha}_t, \hat{\beta}_k)$, $a_{it} = \sum_{j=1}^{J}\sum_{k=1}^{K}\hat{\tilde{\pi}}_{jk}^C \log\left(h_{tk}(f_{ij}|\hat{\theta}_{tk})\right) + \log\left(\hat{\gamma}_t\right)$, and τ_i are the Lagrange multipliers related to the constraints $\sum_{t=1}^{T}\tilde{\pi}_{it}^R = 1$, $\forall i$. It can be easily shown that

$$\begin{aligned} \hat{\tilde{\pi}}_{it}^R = \frac{e^{a_{it}}}{\sum_{t=1}^{T}e^{a_{it}}} &= \frac{\exp\left(\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{\tilde{\pi}}_{jk}^C \log\left(h_{tk}\left(f_{ij}|\hat{\theta}_{tk}\right)\right) + \log\left(\hat{\gamma}_t\right)\right)}{\sum_{t=1}^{T}\exp\left(\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{\tilde{\pi}}_{jk}^C \log\left(h_{tk}\left(f_{ij}|\hat{\theta}_{tk}\right)\right) + \log\left(\hat{\gamma}_t\right)\right)} \\ &= \frac{\hat{\gamma}_t\prod_{j=1}^{J}\prod_{k=1}^{K}h_{tk}\left(f_{ij}|\hat{\theta}_{tk}\right)^{\hat{\tilde{\pi}}_{jk}^C}}{\sum_{t=1}^{T}\hat{\gamma}_t\prod_{j=1}^{J}\prod_{k=1}^{K}h_{tk}\left(f_{ij}|\hat{\theta}_{tk}\right)^{\hat{\tilde{\pi}}_{jk}^C}} \end{aligned}$$

$$= \frac{\hat{\gamma}_t h_t^R\left(\mathbf{f}_i^R|\hat{\mathbf{x}}_t, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\alpha}_t, \hat{\boldsymbol{\beta}}\right)}{g^R\left(\mathbf{f}_i^R|\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{\Gamma}}\right)} = \hat{\pi}_{it}^R(\mathbf{\Theta}). \tag{A.9}$$

A similar result can be obtained for $\hat{\tilde{\pi}}_{jk}^C$.

Appendix B

Parameter estimation at the M-step using the Newton–Raphson procedure

Denoting by $\tilde{f}_{tk} = \sum_{i=1}^{I}\sum_{j=1}^{J}\hat{z}_{ij,tk}f_{ij}$, the entries of the $T \times K$ matrix $\tilde{\mathbf{F}}_{TK} = \mathbf{Z}^{R'}\mathbf{F}\mathbf{Z}^C$, we define

$$\tilde{f}_{..} = \sum_{t=1}^{T}\sum_{k=1}^{K}\tilde{f}_{tk}, \tilde{f}_{t.} = \sum_{k=1}^{K}\tilde{f}_{tk}, \tilde{f}_{.k} = \sum_{t=1}^{T}\tilde{f}_{tk}.$$

and $\lambda = \log\mu$, $\lambda_t^R = \log\alpha_t$, and $\lambda_k^C = \log\beta_k$.

We wish to maximize

$$\begin{aligned} & q\left(\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}|\hat{\mathbf{Z}}^{(s)}\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K}\hat{z}_{ij,tk}^{(s)}\left[f_{ij}\log\left(\mu_{tk}\right) - \mu_{tk}\right] \\ & = \lambda\sum_{t=1}^{T}\sum_{k=1}^{K}\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{z}_{ij,tk}^{(s)}f_{ij}\right) + \sum_{t=1}^{T}\lambda_t^R\sum_{k=1}^{K}\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{z}_{ij,tk}^{(s)}f_{ij}\right) \\ & + \sum_{k=1}^{K}\lambda_k^C\sum_{t=1}^{T}\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{z}_{ij,tk}^{(s)}f_{ij}\right) - \sum_{t=1}^{T}\sum_{k=1}^{K}d_{tk}^2(\mathbf{x}_t, \mathbf{y}_k)\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{z}_{ij,tk}^{(s)}f_{ij}\right) \\ & - \sum_{t=1}^{T}\sum_{k=1}^{K}\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{z}_{ij,tk}^{(s)}\right)\exp\left(\lambda + \lambda_t^R + \lambda_k^C - d_{tk}^2(\mathbf{x}_t, \mathbf{y}_k)\right). \end{aligned} \tag{B.1}$$

Taking into account that

$$\begin{aligned} \sum_{t=1}^{T}\sum_{k=1}^{K}\tilde{f}_{tk}d_{tk}^2(\mathbf{x}_t, \mathbf{y}_k) &= \sum_{t=1}^{T}\tilde{f}_{t.}\mathbf{x}_t'\mathbf{x}_t + \sum_{k=1}^{K}\tilde{f}_{.k}\mathbf{y}_k'\mathbf{y}_k \\ & -2\sum_{t=1}^{T}\sum_{k=1}^{K}\tilde{f}_{tk}\mathbf{x}_t'\mathbf{y}_k, \end{aligned} \tag{B.2}$$

then, (B.1) can be written as,

$$\begin{aligned} q\left(\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}|\hat{\mathbf{Z}}^{(s)}\right) &= \tilde{f}_{..}\lambda + \sum_{t=1}^{T}\tilde{f}_{t.}\lambda_t^R + \sum_{k=1}^{K}\tilde{f}_{.j}\lambda_k^C \\ & -\mathrm{tr}\mathbf{X}'\mathbf{D}_R\mathbf{X} - \mathrm{tr}\mathbf{Y}'\mathbf{D}_C\mathbf{Y} + 2\mathrm{tr}\mathbf{X}'\tilde{\mathbf{F}}_{TK}\mathbf{Y} \\ & -\sum_{t=1}^{T}\sum_{k=1}^{K}M_{tk} \end{aligned} \tag{B.3}$$

where $\mathbf{D}_R = \mathrm{diag}(\tilde{f}_{1.}, ..., \tilde{f}_{T.})$ and $\mathbf{D}_C = \mathrm{diag}(\tilde{f}_{.1}, ..., \tilde{f}_{.K})$ are diagonal matrices, and

$$M_{tk} = IJ\hat{\gamma}_{tk}\mu_{tk} = IJ\hat{\gamma}_{tk}\exp\left(\lambda + \lambda_t^R + \lambda_k^C - \mathbf{x}_t'\mathbf{x}_t - \mathbf{y}_k'\mathbf{y}_k + 2\mathbf{x}_t'\mathbf{y}_k\right).$$

The above function is similar to that of the DA model (De Rooij & Heiser, 2005), except for the last term of (B.3), which is a weighted sum that depends on previously estimated values for the $T \times K$ unconditional probabilities. Compared to the LCDA model (Vera et al., 2014), this term now depends on the classification of the rows and of the columns of $\mathbf{F}$. Therefore, in the s-th iteration at the M-step,

a weighted generalization of the procedure followed in Vera et al. (2014) is employed for parameter estimation.

Using the Newton-Raphson method the parameter values are estimated in an iterative procedure. The updates in the $(s+1)$th iteration are as follows:

$$\lambda^{(s+1)} = \lambda^{(s)} + \frac{\tilde{f}_{..} - M_{..}(\lambda^{(s)})}{M_{..}(\lambda^{(s)})} \tag{B.4}$$

$$\lambda_t^{T(s+1)} = \lambda_t^{T(s)} + \frac{\tilde{f}_{t.} - M_{t.}\left(\lambda_t^{T(s)}\right)}{M_{t.}\left(\lambda_t^{T(s)}\right)} \tag{B.5}$$

$$\lambda_k^{C(s+1)} = \lambda_k^{C(s)} + \frac{\tilde{f}_{.k} - M_{.k}\left(\lambda_k^{C(s)}\right)}{M_{.k}\left(\lambda_k^{C(s)}\right)} \tag{B.6}$$

$$x_{tm}^{(s+1)} = x_{tm}^{(s)} + \frac{2\sum_{k=1}^{K}\left(\tilde{f}_{tk} - M_{tk}\right)\left(x_{tm}^{(s)} - y_{km}^{(s)}\right)}{2\sum_{k=1}^{K}\left(\tilde{f}_{tk} - M_{tk}\right) - 4\sum_{k=1}^{K} M_{tk}\left(x_{tm}^{(s)} - y_{km}^{(s)}\right)^2} \tag{B.7}$$

$$y_{km}^{(s+1)} = y_{km}^{(s)} + \frac{2\sum_{t=1}^{T}\left(\tilde{f}_{tk} - M_{tk}\right)\left(y_{km}^{(s)} - x_{tm}^{(s)}\right)}{2\sum_{t=1}^{T}\left(M_{tj} - \tilde{f}_{tc}\right) - 4\sum_{t=1}^{T} M_{tk}\left(y_{km}^{(s)} - x_{tm}^{(s)}\right)^2}. \tag{B.8}$$

where $M_{t.} = \sum_{k=1}^{K} M_{tk}, M_{.k} = \sum_{t=1}^{T} M_{tk}$, and $M_{..} = \sum_{t=1}^{T}\sum_{k=1}^{K} Mtk$.

Initial estimates for the iterative procedure can be given using Becker's (1990) procedure as described in Appendix C.

Appendix C

Initial estimates and identification

Besides the classification of the rows, $\hat{\boldsymbol{Z}}^R$ and of the columns $\hat{\boldsymbol{Z}}^C$, maximum likelihood estimators for $\hat{\Theta} = (\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\gamma})$ are obtained at the end of the GEM procedure for given values of T, K and M. Parameter estimates in distance association models suffer from indeterminacies. To obtain an identified solution the parameters are expressed as a function of singular values and singular vectors, since the singular value decomposition is unique and is characterized by $M(M+2)$ constraints.

Denoting by $\hat{\boldsymbol{\mu}}_{TK} = (\hat{\mu}_{tk})$ the matrix of estimated expected frequencies (2), let us denote by $\mathbf{G}_{TK} = (g_{tk})$, the matrix of entries $g_{tk} = \log(\hat{\mu}_{tk})$. Then, we take $\bar{\bar{g}}$ as the global mean of the entries of $\mathbf{G}_{TK}$, and $\bar{g}_t$ and $\bar{g}_k$ as the marginal means for the t-th row and for the k-th column of $\mathbf{G}_{TK}$ respectively. Then define $\tilde{\lambda} = \bar{\bar{g}}$, $\tilde{\lambda}_t^R = \bar{g}_t - \bar{\bar{g}}$, $\tilde{\lambda}_{c_k}^C = \bar{g}_k - \bar{\bar{g}}$, and $\boldsymbol{\Delta}$ the matrix of entries $\delta_{tk} = g_{tk} - \tilde{\lambda} - \tilde{\lambda}_t^R - \tilde{\lambda}_k^C$. From the singular value decomposition of $\boldsymbol{\Delta} = \mathbf{U}\boldsymbol{\Gamma}\boldsymbol{\Lambda}'$, it follows that $\mathbf{X}\sqrt{2} = \mathbf{U}\boldsymbol{\Gamma}^{1/2}$ and $\mathbf{Y}\sqrt{2} = \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Lambda}'$, and denoting by $d_{x,t} = \sum_m x_{tm}^2$, and $d_{y,k} = \sum_m y_{km}^2$, identified parameters are obtained;

$$\dot{\lambda}_t^R = \tilde{\lambda}_t^R + d_{x,t} - \log(I\gamma_t) \tag{C.1}$$

$$\dot{\lambda}_k^C = \tilde{\lambda}_k^C + d_{y,k} - \log(J\gamma_k) \tag{C.2}$$

$$\lambda = \tilde{\lambda} + \frac{1}{T}\sum_{t=1}^{T}\dot{\lambda}_t^R + \frac{1}{K}\sum_{k=1}^{K}\dot{\lambda}_k^C \tag{C.3}$$

$$\lambda_t^R = \dot{\lambda}_t^R - \frac{1}{T}\sum_{t=1}^{R}\dot{\lambda}_t^R \tag{C.4}$$

$$\lambda_k^C = \dot{\lambda}_k^C - \frac{1}{K}\sum_{k=1}^{K}\dot{\lambda}_k^C. \tag{C.5}$$

The mean of the values of $\tilde{\lambda}_t^R, t = 1, \ldots, T$ and of $\tilde{\lambda}_k^C, k = 1, \ldots, K$ is equal to zero, and $g_{tk} = \lambda + (\lambda_t^R + \log(I\gamma_t^R)) + (\lambda_t^C + \log(J\gamma_k^C) - d_{tk}^2(\boldsymbol{x}_t, \boldsymbol{y}_k)$. Then, after the identification step the model is characterized by $2 + M(M+2)$ further constraints. This procedure can also be used to determine the initial solution at the M-step in the GEM algorithm using $\mathbf{G}_{TK} = \log(\tilde{\mathbf{F}}_{TK})$, where $\tilde{\mathbf{F}}_{TK} = \mathbf{Z}^{R'}\mathbf{F}\mathbf{Z}^C$ are the values associated with the given classifications at the E-step.

Appendix D

Algorithm flow

From a computational standpoint and to speed up the convergence of the GEM procedure, a multicycle GEM algorithm is employed for parameter estimation. First, an E-step related to the classification of the rows, given the previous classification for the column is calculated, followed by a partial update of the M-step. Then a similar procedure for the columns is followed using previously updated parameter values, the M-step is performed for the estimated classifications, and the final log-likelihood is evaluated. Because of the well-known problem of local minima of the EM algorithm, the latter is usually applied for a number of random starts or from a known optimal initial solution. Here, only nonempty initial classifications of the I row elements into T groups and of the J column elements into K groups are considered. In summary, the steps in the estimation process are:

1. An initial classification $\hat{\boldsymbol{Z}}^{(0)}$ is given and initial parameter values $\Theta^{(0)}$ are calculated maximizing (B.3). Then the parameters are corrected in terms of identification.
2. At the sth step and from previously estimated values of $\hat{\boldsymbol{Z}}^{(s-1)}$ and $\hat{\Theta}^{(s-1)}$, a multicycle estimation procedure is employed. First, the expected value $\hat{\boldsymbol{Z}}^{R^{(s)}} = E[z_{it}^R|\mathbf{F}, \Theta^{(s-1)}]$ is calculated. Then, (B.3)) is maximized with respect to Θ for the values of $\hat{\boldsymbol{Z}}^{R^{(s)}}$ and $\hat{\boldsymbol{Z}}^{C^{(s-1)}}$ to obtain a partial update of Θ at the sth iteration, denoted by $\hat{\Theta}^{(st)}$. Now, $\hat{\boldsymbol{Z}}^{C^{(s)}} = E[z_{jk}^C|\mathbf{F}, \Theta^{(st)}]$ is obtained for the given value of $\hat{\boldsymbol{Z}}^{R^{(s)}}$, and the final value for $\hat{\Theta}^{(s)}$ is obtained using $\hat{\boldsymbol{Z}}^{R^{(s)}}$ and $\hat{\boldsymbol{Z}}^{C^{(s)}}$ by again maximizing (B.3). Then, the parameters are corrected in terms of the identification purpose.
3. The above alternating step is repeated in an iterative cycle until the convergence is achieved, usually when two consecutive values of the log-likelihood (B.3) do not change more than a small, previously determined, value usually 10^{-8}.

For identified values for $\hat{\boldsymbol{\Theta}}$, the final posterior probabilities $\hat{\boldsymbol{\pi}}_i^R = (\hat{\pi}_{i1}^R, ..., \hat{\pi}_{iT}^R)'$ that $\mathbf{f}_i^R \in \mathbf{F}_t^R, i = 1, ..., I, t = 1, ..., T$, and $\hat{\boldsymbol{\pi}}_j^C = (\hat{\pi}_{j1}^C, ..., \hat{\pi}_{jK}^C)'$ that $\mathbf{f}_j^C \in \mathbf{F}_k^C, j = 1, ..., J, k = 1, ..., K$, are obtained by (A.3) and (A.4), respectively. Then, at the end of the iterative procedure the optimal block-shaped partition is given by the well-known Bayes (optimal) rule defined by $\hat{z}_{it,jk} = 1$ if $t = \text{argmax}(\hat{\boldsymbol{\pi}}_i^R)$ and $k = \text{argmax}(\hat{\boldsymbol{\pi}}_j^C)$, and zero otherwise. When the maximum of posterior probabilities are related to more than one latent class the corresponding row and/or column category can be assigned arbitrarily to one of the classes for which the corresponding posterior probabilities are equal to the maximum value (McLachlan & Peel, 2000).