British Journal of Mathematical and Statistical Psychology (2014)

The British Psychological Society

www.wileyonlinelibrary.com

A latent class distance association model for cross-classified data with a categorical response variable

José Fernando Vera[1]*, Mark de Rooij[2] and Willem J. Heiser[2]

[1]University of Granada, Spain
[2]Leiden University, The Netherlands

In this paper we propose a latent class distance association model for clustering in the predictor space of large contingency tables with a categorical response variable. The rows of such a table are characterized as profiles of a set of explanatory variables, while the columns represent a single outcome variable. In many cases such tables are sparse, with many zero entries, which makes traditional models problematic. By clustering the row profiles into a few specific classes and representing these together with the categories of the response variable in a low-dimensional Euclidean space using a distance association model, a parsimonious prediction model can be obtained. A generalized EM algorithm is proposed to estimate the model parameters and the adjusted Bayesian information criterion statistic is employed to test the number of mixture components and the dimensionality of the representation. An empirical example highlighting the advantages of the new approach and comparing it with traditional approaches is presented.

1. Introduction

Most data analysis problems in psychology, and in social and behavioural sciences in general, involve categorical variables. Furthermore, many problems involve the prediction of a single response variable from a set of explanatory variables; that is the traditional regression or classification problem. In this paper both the response and the explanatory variables are categorical. Any combination of the categories of the explanatory variables is called a *profile*. In the most general case, the data consist of a profile by response contingency table. A special case is when the profiles represent a single categorical explanatory variable, in which case we have a simple two-way contingency table. When the profiles represent many explanatory variables, the contingency table becomes large and often sparse: that is, many entries of this contingency table will show zero frequencies.

Let us consider as a running example some study data concerning the 2006 Dutch parliamentary election. A total of 2,176 survey respondents provided data on their membership of various organizations and how they voted. Although many more political parties took part in the election, we will confine our analysis to the seven largest parties: the Christian Democratic Party (CDA, 594 votes in the sample), the Labour Party (PvdA, 461), the Conservative Liberals (VVD, 312), the Green Left Party (GL, 108), the Socialist

*Correspondence should be addressed to José Fernando Vera, Department of Statistics and Operational Research, Faculty of Sciences, University of Granada, 18071 Granada, Spain (e-mail: jfvera@ugr.es).

DOI:10.1111/bmsp.12038

Party (SP, 379), the Christian Union (CU, 93), the Party for Freedom (PVV, 109), and all other votes (a heterogeneous group of political parties, 120) are grouped as Other.

The participants were interviewed about their membership of nine different types of organizations. The following questions were asked:

E Are you a member of an environmental or nature organization?
H Are you a member of an organization active in Third World, human rights or peace issues?
T Are you a member of a trade union?
P Are you a member of a professional organization?
W Are you a member of an employers' organization?
N Are you a member of a neighbourhood or residents' organization or social club?
M Are you a member of a musical or cultural organization?
S Are you a member of a sports organization?
C Are you a member of a church or ideological organization?

These nine types of membership resulted in 512 different membership profiles, of which only 253 different membership patterns occur in our data set. This leaves us with a fairly sparse table of size 253×8 with a total of 1,336 zero entries.

An excerpt from the data can be found in Table 1, where the rows present different profiles as defined by the first nine columns. A '1' denotes membership of the specific organization, while a '2' represents non-membership. As an example, the first row represents the subjects who are a member of E, H, P, W, N, M, S, and C, but not of T. The last eight columns provide frequencies of voting for the political parties of those subjects; that is, of the two subjects with profile 1 (first row), one voted for CDA and one for VVD. There are no persons with this profile of memberships voting for the other political parties. As can be seen from the first seven rows, this contingency table is very sparse, having many zeros and small frequencies.

A traditional approach employed in the analysis of such data is to use a multinomial regression model (also known as a multinomial baseline category logit model: Agresti, 2013, Chapter 8) or a log-linear model. In a multinomial regression model the response variable (political vote) is linked to the set of explanatory variables (the memberships).

Table 1. First seven rows of the data. The first nine columns provide information on the memberships of types of organization as specified in the text, a '1' denoting membership, a '2' non-membership. The last eight columns provide frequencies of votes. So, of the subjects with the profile as described in row 1, one person voted for CDA, and one for VVD. No subjects with this profile voted for any of the other parties

E	H	T	P	W	N	M	S	C	CDA	PvdA	VVD	GL	SP	CU	PVV	Other
1	1	2	1	1	1	1	1	1	1	0	1	0	0	0	0	0
2	1	2	1	1	1	1	1	1	0	0	0	1	0	0	0	0
2	2	1	2	1	1	1	1	1	0	0	0	0	0	0	1	0
2	2	2	2	1	1	1	1	1	0	0	1	0	0	0	0	0
1	1	1	1	2	1	1	1	1	3	1	0	0	0	0	0	0
2	2	1	1	2	1	1	1	1	1	1	0	0	0	0	0	0
2	2	2	1	2	1	1	1	1	3	0	1	0	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

A major difficulty with the analysis of contingency tables like the one above is that many entries in the contingency table are zero, rendering a subset of models inappropriate and goodness-of-fit criteria ill-defined. That is, when a contingency table is sparse, including empty cells, estimated odds ratios based on these empty cells are either zero, infinity, or undefined. The implication is that the standard log-linear model for such a table results in parameters that are infinite. To overcome this problem a model with fewer parameters for the association can be fitted. One of the earliest examples of such a reduced model is the RC(*M*) association model (Goodman, 1985; Wickens, 1989, Chapter 11). In this model the association between two categorical variables is defined in terms of a low-dimensional inner product relation. By reducing the dimensionality, the model is not as vulnerable to sparse data.

For non-sparse data, the RC(*M*) association model reduces the number of parameters for modelling the association as compared to the log-linear model. Such a reduction in parameters often results in more stable models, especially when the number of categories of the two variables is large. Another advantage of the RC(*M*) model is that there exists a graphical representation that can aid interpretation. De Rooij & Heiser (2005), however, show that the standard graphical representation of the RC(*M*) model, which is in terms of inner products, is prone to misinterpretations and proposed a distance representation: the distance association (DA) model. The standard graphical representation of the RC(*M*) association model is often interpreted by a distance rule, while this has no basis in the model. In fact, the relationship between row and column points in the RC(*M*) model should be interpreted in terms of an inner product: take the product of the distance of the two points from the origin and multiply by the cosine of the angle. De Rooij & Heiser (2005) present several examples where a distance interpretation of an inner product relation fails. De Rooij and Heiser go on to propose a distance representation: the DA model. The DA model and the RC(*M*) association model are equivalent (De Rooij, 2007 2008; De Rooij & Heiser, 2005), that is, the two models have the same fit to any data set, and parameters from one model can be transformed into the parameters of the other.

The graphical interpretation of the DA model is based on a distance rule, leading to less confusion. In terms of distances, one of the main aims of DA models for a cross-classified data set is to represent the I row category elements $V = \{v_1, \ldots, v_I\}$ and the J column category elements $O = \{o_1, \ldots, o_J\}$, in a Euclidean space of low dimension M, with the distances between points inversely describing the relationship between the categories of the two sets (De Rooij & Heiser, 2005). Thus, a relatively large frequency, indicating a strong association, will correspond to a small distance between the points representing the corresponding row and column categories, and conversely a relatively small frequency will be related to a large distance.

For the type of data described above, although the DA model can be estimated, the final result may be difficult to interpret because of the proliferation of row profiles. That is, a graphical representation with 253 row points becomes very cluttered and the distinct points are hard to distinguish. One way to proceed is to check whether this number can be reduced. In other contexts, simultaneous dimension reduction and clustering have been proposed. Vichi and Kiers (2001), for example, show how to use *k*-means within principal component analysis. Such a clustering reduces the number of objects in the graphical display and therefore makes the final result better interpretable. In a least squares unfolding framework, classification and representation methods have already been developed to enhance the interpretation of the solution and/or to obtain an adequate fit of the model when the number of elements is large (Vera *et al.*, 2013). In a probabilistic

context, mixture distribution formulations are the natural way to proceed. A dual latent class unfolding model for continuous rating data, with the assumption that within each homogeneous group or latent class the data are independently and normally distributed, was developed by Vera, Macías, and Heiser (2009b) (see also Vera, Macías, & Angulo, 2009; and Vera, Macías, & Heiser, 2009a). In a general cross-classified framework, this approach is of particularly interest when the number of row (column) categories is large, or when the data are sparse, as is the case in the present paper. Clustering the row profiles may lead to a solution that is better interpretable, with a relatively small number of parameters.

Clustering techniques for contingency tables were proposed earlier (Goodman, 1981; Kateri & Iliopoulos, 2003), but these pertain to simple two-variable contingency tables. Moreover, they are all based on within-cluster homogeneity (Goodman, 1981) and are often coupled with the RC(M) association model (Goodman, 1985). The present approach uses a different cluster criterion. The latent class model collapses rows that belong to the same row distribution (this is the homogeneity concept here), and thus equally distributed row profiles are represented at the same location in the Euclidean space.

In this paper we propose a latent class distance association (LCDA) model for profile by response data. The model allows us to cluster the profiles while simultaneously representing the classes by the categories of the response variable in a low-dimensional Euclidean space using the DA model. The model can also be applied to any two-mode cross-classified data without requiring us to distinguish between predictor and response categories. A model selection strategy based on the Bayesian information criterion (BIC) statistic is used to test the number of latent classes and the dimensionality of the representation.

This paper is organized as follows. The next section presents the LCDA model. Section 3 describes maximum likelihood estimation using a generalized EM algorithm. Starting values and model selection are also discussed in this section. Section 4 starts with a series of simulated studies to verify the properties of the algorithm and the model selection procedure. The Dutch parliamentary election data are then analysed using the LCDA model, and the results are compared with those obtained with the standard DA model and with multinomial logistic regression. The paper concludes with a general discussion.

2. The LCDA model

Consider a partition $\mathcal{P}(\mathbf{F})$ of the profiles (row space) of an $I \times J$ contingency table $\mathbf{F} = (f_{ij})$ into T latent classes.

It is assumed that a profile 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. The rows in $\mathbf{F}$ are arranged by permuting them in accordance with the sequence in the index sets of the latent classes; thus, in terms of the frequency table $\mathbf{F}$, the situation can be described assuming a row block shaped partition $\mathcal{P}(\mathbf{F})$ of the rectangular matrix $\mathbf{F}$ into T blocks $\mathbf{F}_t$ of r_t elements $\mathbf{f}_i = (f_{i1}, \ldots, f_{iJ})'$, with $\mathbf{f}_i \in \mathbf{F}_t$. Hence, each row vector of $\mathbf{F}$ belongs to one and only one of the T subsets $\mathbf{F}_t$, but we do not known in advance which latent block a particular row belongs to. The unconditional probability that any row element $\mathbf{f}_i$ belongs to latent class $\mathbf{F}_t$ is denoted by γ_t, with $0 \leq \gamma_t \leq 1$ and

$$\sum_{t=1}^{T} \gamma_t = 1. \tag{1}$$

For the LCDA model it is assumed that it is not the profiles by themselves but the cluster centres that are represented by points $\mathbf{x}_t$ in Euclidean space of dimension M, whose coordinates are gathered in a $T \times M$ configuration matrix $\mathbf{X}$. The categories of the response variable are represented by points $\mathbf{y}_j$ gathered in the rows of the $J \times M$ configuration matrix $\mathbf{Y}$. Thus, in the general multiplicative form, the expected frequency of row i and column j, with $\mathbf{f}_i \in \mathbf{F}_t$, is given by the expected frequency μ_{tj} of cluster t and column j, which can be written as

$$\mu_{tj} = \mu\alpha_t\beta_j \exp(-d_{tj}^2), \tag{2}$$

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

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

3. The algorithm

We utilize the well-known equivalence of the multinomial and Poisson distribution (Agresti, 2013, pp. 8, 361; Birch, 1963) to derive a generalized EM algorithm. The equivalence of the product multinomial and the Poisson distributions for the LCDA model is shown in Appendix A.

In a standard Poisson sampling model, counts are considered as independent random variables (see Agresti, 2013), and the probability $h_t(\cdot)$ for the data of a row element $\mathbf{f}_i \in \mathbf{F}_t$ is given by

$$h_t(\mathbf{f}_i|\mathbf{x}_t, \mathbf{Y}, \mu, \alpha_t, \boldsymbol{\beta}) = \prod_{j=1}^{J} \frac{\mu_{tj}^{f_{ij}}}{f_{ij}!} \exp(-\mu_{tj}), \tag{3}$$

where μ_{tj} is given by (2) and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$. Because we do not know in advance which latent class a row belongs to, the probability density function (p.d.f.) ($g(\cdot)$) of the random variable $\mathbf{f}_i$ becomes a finite mixture of Poisson densities given by (3), which can be expressed as

$$g(\mathbf{f}_i|\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{t=1}^{T} \gamma_t h_t(\mathbf{f}_i|\mathbf{x}_t, \mathbf{Y}, \mu, \alpha_t, \boldsymbol{\beta}), \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)'$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)'$. So the log-likelihood function to be maximized subject to (1) can be written as

$$\log L = \sum_{i=1}^{I} \log \sum_{t=1}^{T} \gamma_t h_t(\mathbf{f}_i|\mathbf{x}_t, \mathbf{Y}, \mu, \alpha_t, \boldsymbol{\beta}). \tag{5}$$

Given the maximum likelihood estimators $\hat{\mathbf{X}}$, $\hat{\mathbf{Y}}$, $\hat{\mu}$, $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\gamma}}$, the posterior probability that an element $\mathbf{f}_i$ belongs to latent class $\mathbf{F}_t$ is calculated by means of Bayes' theorem as follows:

$$\pi_{it}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \frac{\hat{\gamma}_t h_t(\mathbf{f}_i|\hat{\mathbf{x}}_t, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\alpha}_t, \hat{\boldsymbol{\beta}})}{g(\mathbf{f}_i|\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})}. \tag{6}$$

Hence, from the maximum likelihood estimators, an element $\mathbf{f}_i$ will be assigned to the class that it is most likely to belong to given these posterior probabilities, and for the parameter estimation, the EM algorithm (Dempster, Laird, & Rubin, 1977) can be employed.

3.1. The EM algorithm

As usual in the EM algorithm formulation, the following mixture component indicator variables are introduced:

$$z_{it} = \begin{cases} 1, & \text{if} \quad \mathbf{f}_i \in \mathbf{F}_t, \\ 0, & \text{otherwise.} \end{cases}$$

Let us define the column vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{iT})'$, and the $I \times T$ matrix $\mathbf{Z}$ written in terms of its row vectors as $(\mathbf{z}_1, \ldots, \mathbf{z}_I)'$. It will be assumed that the $\mathbf{z}_i$ are independently and identically multinomially distributed variables with probabilities $\boldsymbol{\gamma}$ such that

$$\sum_{t=1}^{T} z_{it} = 1 \quad \text{and} \quad \sum_{i=1}^{I}\sum_{t=1}^{T} z_{it} = I.$$

Then, given the indicator nature of z_{it}, the p.d.f. of $\mathbf{f}_i$, given $\mathbf{z}_i$, can be written as

$$\Psi(\mathbf{f}_i|\mathbf{z}_i, \mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{t=1}^{T} h_t(\mathbf{f}_i|\mathbf{x}_t, \mathbf{Y}, \mu, \alpha_t, \boldsymbol{\beta})^{z_{it}}, \tag{7}$$

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

$$p(\mathbf{z}_i|\boldsymbol{\gamma}) = \prod_{t=1}^{T} \gamma_t^{z_{it}}. \tag{8}$$

Using (7) and (8), the complete p.d.f. of $\mathbf{f}_i$ and $\mathbf{z}_i$ can be written as

$$\begin{aligned} \Phi(\mathbf{f}_i, \mathbf{z}_i|\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \Psi(\mathbf{f}_i|\mathbf{z}_i, \mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\mathbf{z}_i|\boldsymbol{\gamma}) \\ &= \prod_{t=1}^{T} (\gamma_t h_t(\mathbf{f}_i|\mathbf{z}_i, \mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}))^{z_{it}}, \end{aligned} \tag{9}$$

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

$$\log L(\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{F}, \mathbf{Z}) = \sum_{i=1}^{I}\sum_{t=1}^{T} z_{it} \log \gamma_t + \sum_{i=1}^{I}\sum_{t=1}^{T} z_{it} \log h_t(\mathbf{f}_i|\mathbf{x}_t, \mathbf{Y}, \mu, \alpha_t, \boldsymbol{\beta}). \tag{10}$$

In practice, the indicators in $\mathbf{Z}$ are non-observed variables and the EM algorithm estimates them by means of their expected value (E-step). Then, the remaining parameters are

estimated bymaximizing (10), given $\mathbf{F}$ and the previously estimated values of $\mathbf{Z}$ (M-step). The two steps are repeated in an alternating iterative process, which can be stopped when a convergence criterion is reached.

3.1.1. E-step

The EM algorithm starts with the initial parameter estimation process (see Section 3.2), denoting the initial parameter values by $\hat{\mathbf{Z}}^{(0)}, \hat{\mathbf{X}}^{(0)}, \hat{\mathbf{Y}}^{(0)}, \hat{\mu}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\boldsymbol{\beta}}^{(0)}$, and $\hat{\boldsymbol{\gamma}}^{(0)}$. Then the expectation of $\mathbf{Z}$ in the sth iteration, given $\mathbf{F}$ and previous estimated parameter values $\Theta^{(s-1)} = \{\hat{\mathbf{X}}^{(s-1)}, \hat{\mathbf{Y}}^{(s-1)}, \hat{\mu}^{(s-1)}, \hat{\boldsymbol{\alpha}}^{(s-1)}, \hat{\boldsymbol{\beta}}^{(s-1)}, \hat{\boldsymbol{\gamma}}^{(s-1)}\}$, can be determined due to the linearity of $\log L$ on z_{it} as

$$\begin{aligned}\mathcal{Q}(\Theta, \Theta^{(s-1)}) = &\sum_{i=1}^{I}\sum_{t=1}^{T} E[z_{it}|\mathbf{F}, \Theta^{(s-1)}] \log \gamma_t \\ &+ \sum_{i=1}^{I}\sum_{t=1}^{T} E[z_{it}|\mathbf{F}, \Theta^{(s-1)}] \log h_t(\mathbf{f}_i|\mathbf{x}_t, \mathbf{Y}, \mu, \alpha_t, \boldsymbol{\beta}),\end{aligned} \tag{11}$$

where $E[z_{it}|\mathbf{F}, \Theta^{(s-1)}]$ denotes the expectation of z_{ij} in the sth iteration. Because the non-observed z_{it} are Bernoulli distributed variables, $E[z_{it}|\mathbf{F}, \Theta^{(s-1)}]$ represents the probability that $\mathbf{f}_i$ belongs to $\mathbf{F}_t$, after the partition of $\mathbf{F}$ is known and considering the previous estimated parameter values $\Theta^{(s-1)}$. Thus, $E[z_{it}|\mathbf{F}, \Theta^{(s-1)}] = \pi_{it}(\Theta^{(s-1)})$ and

$$\hat{z}_{it}^{(s)} = \hat{\pi}_{it}(\Theta^{(s-1)}), \tag{12}$$

from which in the M-step the unobserved values of $\mathbf{Z}$ are substituted by these posterior probabilities.

3.1.2. M-step

In the M-step, (10), or equivalently (11), is maximized with respect to parameters $\mathbf{X}$, $\mathbf{Y}$, μ, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, under previously estimated values $\hat{z}_{it}^{(s)}$. It can be easily shown that the expression for the estimator of γ_t at the sth iteration is given by

$$\hat{\gamma}_t^{(s)} = \frac{1}{I}\sum_{i=1}^{I} \hat{z}_{it}^{(s)}. \tag{13}$$

The estimation of the remaining parameters can be carried out by maximizing (11) under previously estimated values of $\hat{z}_{it}^{(s)}$ and $\hat{\gamma}_t^{(s)}$, or equivalently, by maximizing

$$q(\mathbf{X}, \mathbf{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}|\hat{\mathbf{Z}}^{(s)}) = \sum_{i=1}^{I}\sum_{t=1}^{T} \hat{z}_{it}^{(s)} \log h_t(\mathbf{f}_i|\mathbf{x}_t, \mathbf{Y}, \mu, \alpha_t, \boldsymbol{\beta}). \tag{14}$$

Writing $f_{tj} = \sum_{i=1}^{I} \hat{z}_{it} f_{ij}$, we have

$$f_{..} = \sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj}, \quad f_{t.} = \sum_{j=1}^{J} f_{tj}, \quad f_{.j} = \sum_{t=1}^{T} f_{tj},$$

and letting $\lambda = \log \mu$, $\lambda_t^R = \log \alpha_t$, and $\lambda_j^C = \log \beta_j$, we find (see Appendix B) that

$$q(\mathbf{X},\mathbf{Y},\mu,\boldsymbol{\alpha},\boldsymbol{\beta}|\hat{\mathbf{Z}}^{(s)}) = f_{..}\lambda + \sum_{t=1}^{T} f_{t.}\lambda_t^R + \sum_{j=1}^{J} f_{.j}\lambda_j^C - \mathrm{tr}\mathbf{X}'\mathbf{D}_R\mathbf{X} - \mathrm{tr}\mathbf{Y}'\mathbf{D}_C\mathbf{Y} + 2\mathrm{tr}\mathbf{X}'\mathbf{F}_T\mathbf{Y} - \sum_{t=1}^{T}\sum_{j=1}^{J} I\gamma_t \exp(\lambda + \lambda_t^R + \lambda_j^C - \mathbf{x}_t'\mathbf{x}_t - \mathbf{y}_j'\mathbf{y}_j + 2\mathbf{x}_t'\mathbf{y}_j), \tag{15}$$

where $\mathbf{D}_R = \mathrm{diag}(f_{1.} \ldots f_{T.})$ denotes a $T \times T$ diagonal matrix, $\mathbf{D}_C = \mathrm{diag}(f_{.1} \ldots f_{.J})$ a $J \times J$ diagonal matrix, and $\mathbf{F}_T = (f_{tj})$ the $T \times J$ block matrix of $\mathcal{P}(\mathbf{F})$.

Thus, in the sth iteration and under the estimated values of $\hat{\gamma}^{(s)}$ and $\hat{\mathbf{Z}}^{(s)}$, a weighted generalization of the De Rooij & Heiser (2005) two-mode estimation procedure can be employed for the estimation of $\hat{\mathbf{X}}^{(s)}$, $\hat{\mathbf{Y}}^{(s)}$, $\hat{\mu}^{(s)}$, $\hat{\boldsymbol{\alpha}}^{(s)}$, and $\hat{\boldsymbol{\beta}}^{(s)}$ (see Appendix B).

In summary, the steps in the estimation process can be described as follows:

1. Start with the initial parameter values $\Theta^{(0)}$ and calculate $\widehat{z}_{it}^{(0)} = \pi_{it}(\boldsymbol{\Theta}^{(0)})$.
2. Maximize (11) with respect to Θ to give $\widehat{\Theta} = \{\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\gamma}\}$.
3. With the last estimation of the parameter values, return to 1, and iterate 1 and 2 to achieve convergence.

3.2. *Initial parameter estimation and identification*

The EM algorithm starts with certain initial values that must be specified in advance. Different starting strategies and stopping rules can lead to widely differing estimates (see Seidel, Mosler, & Alker, 2000). Therefore, the EM algorithm needs to be run repeatedly using different sets of starting values. The proposed EM algorithm alternates iteratively between the E- and M-steps, until no significant improvement is found in the likelihood function.

The initial estimation for unobserved data $\mathbf{Z}^{(0)}$ is carried out from a random or any given classification of the I row elements in T groups under the computational restriction that $r_t \geq 1$, $t = 1, \ldots, T$, to avoid the presence of zero columns in $\mathbf{Z}$. The initial values for $\hat{\gamma}^{(0)}$ are given by (13), and $\mathbf{F}_T^{(0)} = \mathbf{Z}^{(0)'}\mathbf{F}$ is calculated. Then the initial parameter estimation is carried out by maximizing (15).

As shown by De Rooij & Heiser (2005), indeterminacies are present in the parameter estimates and an identified solution can be obtained by writing all parameters as a function of singular values and singular vectors. First determine $\mathbf{G}_T^{(0)} = \log(\mathbf{F}_T)$; denoting by $\overline{\overline{g}}$ the global mean of the entries of $\mathbf{G}_T^{(0)}$, and by $\overline{g}_t$ and $\overline{g}_j$ the marginal means for the tth row and for the jth column of $\mathbf{G}_T^{(0)}$ respectively, define $\tilde{\lambda} = \overline{\overline{g}}$, $\tilde{\lambda}_t^R = \overline{g}_t - \overline{\overline{g}}$, $\tilde{\lambda}_j^C = \overline{g}_j - \overline{\overline{g}}$, and $\Delta = \mathbf{G}_T^{(0)} - \tilde{\lambda} - \tilde{\lambda}_t^R - \tilde{\lambda}_j^C$. From the singular value decomposition of $\Delta = \mathbf{U}\boldsymbol{\gamma}\boldsymbol{\Lambda}'$, 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, writing $d_{x,t} = \sum_m x_{tm}^2$ and $d_{y,j} = \sum_m y_{jm}^2$, identified parameters are obtained by

$$\dot{\lambda}_t^R = \tilde{\lambda}_t^R + d_{x,t} - \log(I\gamma_t), \tag{16}$$

$$\dot{\lambda}_j^C = \tilde{\lambda}_j^C + d_{y,j}, \tag{17}$$

$$\lambda = \tilde{\lambda} + \frac{1}{T}\sum_{t=1}^{T} \dot{\lambda}_t^R + \frac{1}{J}\sum_{j=1}^{J} \dot{\lambda}_j^C, \tag{18}$$

$$\lambda_t^R = \dot{\lambda}_t^R - \frac{1}{T}\sum_{t=1}^{T} \dot{\lambda}_t^R, \tag{19}$$

$$\lambda_j^C = \dot{\lambda}_j^C - \frac{1}{J}\sum_{j=1}^{J} \dot{\lambda}_j^C. \tag{20}$$

The singular value decomposition is unique and is characterized by $M(M + 2)$ constraints, and the mean of the values of $\tilde{\lambda}_t^R, t = 1, \ldots, T$ and of $\tilde{\lambda}_j^C, j = 1, \ldots, J$, is equal to zero. Then, after identified parameters are obtained, the model is characterized by $2 + M(M + 2)$ further constraints.

3.3. Optimal partition and model selection

After the maximum likelihood parameter estimators of $\hat{\mathbf{X}}$, $\hat{\mathbf{Y}}$, $\hat{\mu}$, $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, $\hat{\gamma}$, have been found, identified parameters are obtained as shown for the initial solution but now considering $\mathbf{G}_T = (\log(\hat{\mu}_{tj}))$, and $\dot{\lambda}_t^R = \tilde{\lambda}_t^R + d_{x,t}$ in (16). For each row of $\mathbf{F}$, the final posterior probabilities that the profile $\mathbf{f}_i$, $i = 1, \ldots, I$, belongs to $\mathbf{F}_t$, $t = 1, \ldots, T$, written in vector form $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i1}, \ldots, \hat{\pi}_{iT})$, are obtained by (6). Then, the optimal row partition is given by $\hat{\mathbf{Z}}$, defined as

$$\hat{z}_{it} = \begin{cases} 1, & \text{for} \quad t = \arg\max(\hat{\boldsymbol{\pi}}_i), \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

This rule, known as the Bayes rule, may not be uniquely defined for a profile $\mathbf{f}_i$ if the maximum of the posterior probabilities is achieved with respect to more than one latent class. In this situation, the profile can be assigned arbitrarily to one of the classes for which the corresponding posterior probabilities are equal to the maximum value (see McLachlan & Peel, 2000).

For the purpose of parameter estimation, the values of T and M are assumed to be known. Nevertheless, in most practical situations the investigator will have to determine the number of clusters in the data, as well as the dimension for the representation. One of the principal goals of latent class models is to determine the number of mixture components, T. However, it is well known that regularity conditions do not hold for the likelihood ratio test when comparing mixtures with different numbers of component distributions, and several procedures have been proposed to resolve this question (see McLachlan & Peel, 2000). The bootstrap approach (Hope, 1968) is a widely employed alternative procedure. Although this method has been employed in the context of latent class models for unfolding or multidimensional scaling (see Vera *et al.*, 2009b, for further details), it adds a lot to CPU time.

For model selection purposes, the BIC (Schwarz, 1978) is employed. Rissanen (1986, 1989) derived this same criterion in a model selection context, from the perspective of coding information theory (see McLachlan & Peel, 2000, Section 6.9.3, for further details).

The proposed model also provides the possibility of determining the dimensionality of the unfolding representation using information criteria. To outperform the BIC in this context, we have included the sample size adjustment suggested by Rissanen (1978), where the number of rows I is adjusted by $(I + 2)/24$ (Yang &Yang, 2007). Under the adjusted BIC, the expression

$$\text{BIC}^* = -2\log L + l\log h$$

is used, where $h = (I + 2)/24$ and $l = TJ + (T - 1)$ for the unconstrained model. When geometrical constraints are imposed, the number of unknown parameters is given by

$$\begin{aligned} l =& 1 + (T-1) + (J-1) + (T+J)M - M(M+2) + (T-1) \\ =& 2T + J + (T + J - M - 2)M - 2 \end{aligned}$$

for the Poisson sampling model and by $l - T$ for product multinomial sampling.

Therefore, the number of latent classes is indicated as corresponding to the lower value of the BIC* statistics, when the proposed procedure is applied for a range of values of T without imposing the geometrical constraints, whence the expected frequencies are given by

$$\hat{\mu}_{tj} = \frac{\sum_{i=1}^{I} \hat{z}_{it} f_{ij}}{\sum_{i=1}^{I} \hat{z}_{it}}. \tag{22}$$

Given a number of latent classes for the row category elements, the BIC* criterion can be employed to establish the dimension of the distance association representation, since regularity conditions do not hold for the likelihood ratio test when models with different numbers of dimensions are compared (Takane, Van der Heijden, & Browne, 2003). Hence, under the previously selected values of T and by imposing geometrical constraints, the dimensionality corresponding to the lower BIC* value when the procedure is run in several dimensions can be selected as the best representation model.

3.4. Overview and implementation

An overview of the algorithm is provided in Figure 1, where the different optimization and identification steps are put in the appropriate order. In this overview the different estimation steps are explicated by referring to the appropriate equation in the paper.

The proposed procedure was implemented in Matlab,[1] and the best solution in 100 replications was chosen as the final solution for the LCDA algorithm, due to the well-known problem of local minima with the EM algorithm. The convergence criterion used is a difference in subsequent log-likelihood values of less than 10^{-8}. Additional stopping criteria are a maximum of $1000(I + J)$ iterations for the EM estimation procedure and of 10,000 for the Newton–Raphson algorithm in the M-step. For none of our applications or simulations (see Section 4) were the stopping criteria reached. In all cases the algorithm stopped because the difference in subsequent log-likelihood values was smaller than the threshold.

---

[1] The program and data are available upon request.

Read data: Read data matrix $\boldsymbol{F}_{I\times J}$

Initialize parameters: $T, ndim$

Determine: $Z^{(0)}$, and $F_t = {Z^{(0)}}^t F$

Estimate: $\gamma^{(0)}$ (using 13), and $\mathbf{X}^{(0)}$, $\mathbf{Y}^{(0)}$, $\mu^{(0)}$, $\boldsymbol{\alpha}^{(0)}$, $\boldsymbol{\beta}^{(0)}$ (see Appendix B)

Calculate: $LL(1) = \log L(Z^{(0)}, \mathbf{X}^{(0)}, \mathbf{Y}^{(0)}, \mu^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)})$ (see 5)

Repeat (Iter=Iter+1)

E-step

$\hat{z}_{it} = \pi_{it}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ using (6)

M-step

$\boldsymbol{F}_t = \hat{\boldsymbol{Z}}^t \boldsymbol{F}$, $\hat{\gamma}_t = 1/I \sum_i \hat{z}_{it}$. Initialize $\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\lambda}, \hat{\boldsymbol{\lambda}}^R, \hat{\boldsymbol{\lambda}}^C$ (see Section 3.2)

Calculate $LL_Z(1) = \log L(\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\lambda}, \hat{\boldsymbol{\lambda}}^R, \hat{\boldsymbol{\lambda}}^C | Z)$ (see 14)

Repeat (It=It+1)

Calculate λ by (B10), λ_t by (B11), λ_j by (B12)

For m=1 **to** ndim

Calculate x_{tm} by (B13), y_{jm} by (B14)

End (For)

Identify $\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\lambda}, \hat{\boldsymbol{\lambda}}^R, \hat{\boldsymbol{\lambda}}^C$ (see Section 3.2)

Evaluate LL_Z(It).

Until $(LL(It) - LL(It-1) > 10E^{-6})$

Evaluate LL(Iter)

Until $(LL(Iter) - LL(Iter-1) > 10E^{-6})$ or $(Iter = Iter_{\max})$

Figure 1. Pseudocode for the LCDA algorithm.

4. Application

4.1. Artificial data

We first present a series of simulation studies in which the properties of our model and estimation procedure are investigated. Taking into account the equivalence between the Poisson and product multinomial sampling schemes in this context (see Appendix A), the performance of the model is first tested for general clustered non-sparse frequency data and afterwards for sparse data sets. The performance of the model is also tested for known clustered configurations, in an experimental design where dimensionality, cluster dispersion/overlap and cluster size are considered as factors. Afterwards we present an application to the Dutch parliamentary election studies and a comparison to standard DA models and the multinomial logistic regression for this data set.

Table 2. Results for the $T = 5$ model, when categories are scaled in one, two and in three dimensions for the artificial data set

No. of dimensions	No. of parameters	Log-likelihood	BIC*
1	58	−54,492.91	109,143.92
2	89	−34,945.87	70,113.70
3	122	−34,930.62	70,140.99

4.1.1. Simulation I

This methodology was first applied to an artificial data set in which each profile consists of a single category. A rectangular matrix of artificial frequency data was generated, after locating in a plane 21 points regularly distributed according to a 3×7 equally spaced grid representing the column categories, and with five cluster centres regularly distributed on it representing the row cluster centres. The Euclidean distances between the coordinates of the cluster centres and the points in the grid were calculated. Then, for each distance d_{tj}, $t = 1, \ldots, 5, j = 1, \ldots, 21$, a sample of 100 frequency values was generated from a Poisson distribution with parameters $\mu_{tj} = 100 \exp(-d_{tj}^2)$, according to (3), considering the remaining effect parameter values equal to unity to make the distances monotonically related to the joint probabilities for comparison purposes (see De Rooij & Heiser, 2005; Takane, 1998).

The 500×21 matrix of frequencies was first analysed without imposing geometrical constraints, in order to determine the number of clusters. The lowest value for the BIC* statistic (70,173.43) was found for the value of $T = 5$, as was to be expected. Following the model selection procedure, the geometrical constraint (2) was considered in the $T = 5$ latent class model. As shown in Table 2, two dimensions were selected to represent the cluster centres, which corresponds to the lowest value of the BIC* statistics.

For the $T = 5$ model scaled in two dimensions, the artificial row structure of 100 elements per cluster is recovered, and therefore equal mixing proportion values of $\hat{\gamma}_t = 0.2$ are obtained, which is in accordance with the model. The estimated maximum posterior probability that a row element belongs to its corresponding cluster is $\hat{\pi}_{it} = 1$ in all situations. The estimated row effect parameter values for the latent class effect are located in the interval $\hat{\alpha}_t \in (0.82, 1.21), t = 1, \ldots, 5$, while the column effects are in the interval $\hat{\beta}_j \in (0.90, 1.09), j = 1, \ldots, 21$; the estimated value for the overall scale parameter is about $\hat{\mu} = 113.1$, all of which is again in accordance with the model simulated. To best determine the effectiveness in recovering the underlying structure, twenty 500×21 frequency matrices were generated following the procedure described above. In terms of CPU time, a value of 272 s per run was found on average. The original data structure was effectively recovered by the LCDA model in all data sets after a Procrustes analysis (Cliff, 1966) was conducted to eliminate the rotational indeterminacy (see Figure 2), with a mean Procrustes value of 0.328.

4.1.2. Simulation II

The performance of the algorithm was also analysed for sparse data sets by setting equal to zero a percentage of randomly selected frequency entries, previously given by the above simulation process. Thus, 20 frequency tables for each density value of 0.75, 0.55, 0.35, and 0.25 were considered, that is, tables with $100(1 - \textit{density}) = 25$, 45, 65, and 75% zero entries. The corresponding given mean Procrustes values were 0.0357, 0.0622,

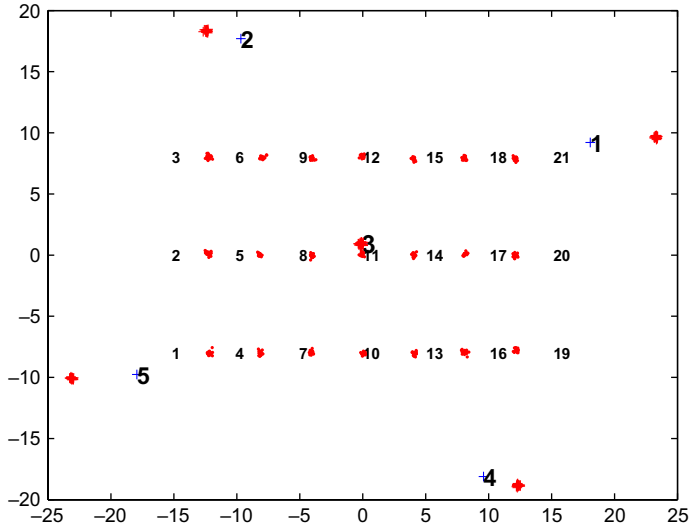

Figure 2. Representation of the true (numbered) and recovered solution (blue crosses and red circles) for the 20 non-sparse Poisson artificial data sets. The large-sized numbers are labels for the cluster centres.

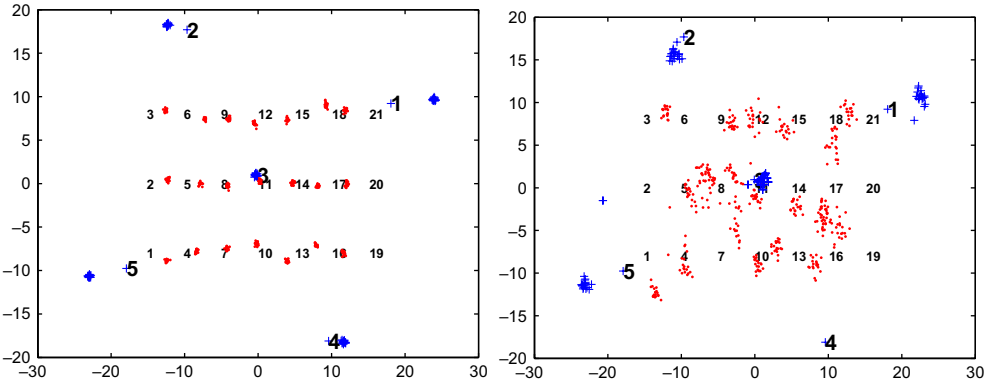

Figure 3. Representation of the true (numbers) and recovered solution (blue crosses and red points) for 20 artificial sparse data sets. The results are shown for tables with densities of 0.75 (left) and 0.35 (right).

0.1824, and 0.6467, when comparing the given configurations with the original one. Figure 3 shows the given configurations corresponding to the 0.75 (left) and the 0.35 (right) density values, corresponding to the lowest and highest acceptable Procrustes values, respectively. As can be seen, the LCDA model recovers the original configuration well, up to fairly sparse tables with decreasing performance for very low densities (about 70% zero entries), as could be expected.

4.1.3. Simulation III

To further analyse the performance of the proposed LCDA model, a Monte Carlo experiment was carried out on the basis of recovering the original cluster center structure

(and thus the original distances) of a configuration of clustered synthetic data sets. Different factors were considered for the sampling scheme such as the cluster size (equal or unequal), the cluster overlap (in terms of the overall dispersion), and the number of dimensions (two or three) for the configuration.

Two configurations of nine points $\mathbf{y}_j$ randomly distributed in two and in three dimensions were considered for the column categories, and with five points $\mathbf{x}_t$ regularly distributed on it representing the row cluster centres. For each configuration, the Euclidean distances d_{tj} between the coordinates of the cluster centres and the points representing column categories were calculated, constituting the parameter values of the interaction term at (2).

First, for each point representing a row cluster center $\mathbf{x}_t$, a sample of 100 clustered points $\mathbf{x}_{i_t}$, $i_t = 1, \ldots, 100$, was generated from a normal distribution of mean $\mathbf{x}_t$ and diagonal covariance matrix given by a value $\mathbf{I}\sigma_t$ that is selected such that the sum of the within-cluster points dispersion represents a previously fixed percentage of the overall dispersion. Percentage values of 20, 40, 60, and 80 were considered, and for each data set the overall dispersion is calculated as the trace of the covariance matrix for the complete data set, while the within-cluster dispersion is given as the trace of the covariance matrix for the corresponding points in the cluster. Also, samples of different cluster sizes were analysed by considering one cluster representing 50% of the total number of rows of $\mathbf{F}$, while the remaining cluster sizes were fixes to 15%, except for one fixed at 5% (see Figure 4).

Thus, for each of the 2 (dimension) $\times$ 4 (dispersion) $\times$ 2 (cluster size) combinations of sampling factors, 20 configurations were generated, and the Euclidean distances between the corresponding configurations of rows and columns, $d_{i_t j}$, for $i_t = 1, \ldots, n_t$, $t = 1, \ldots, T$, were calculated. Then the frequency values were generated as $f_{ij} = 100 \exp(-d_{tj}^2)$, for $\mathbf{f}_i \in \mathbf{F}_t$, and the 320 data sets were analysed by an LCDA model under a Poisson sampling scheme, taking into account the random nature of the point generation process.

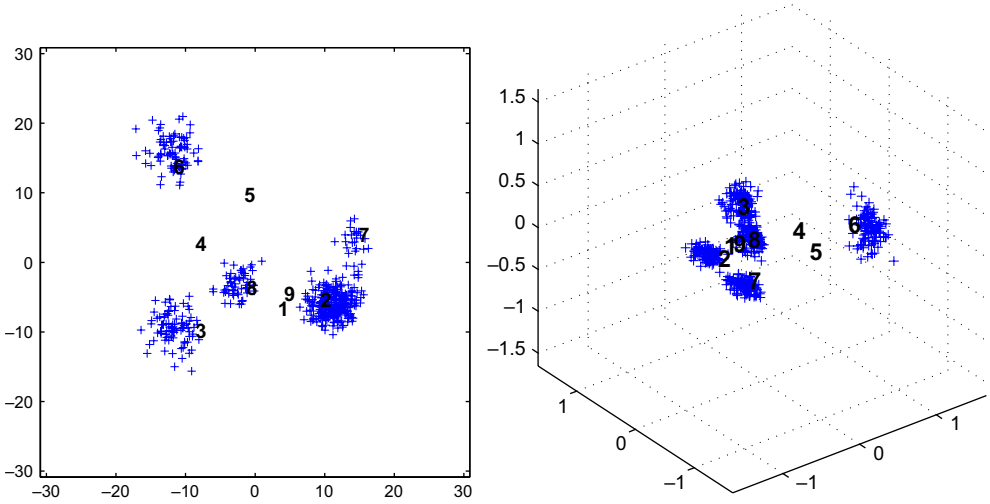

Figure 4. Sample of 500 simulated clustered data from a multivariate normal distribution in two dimensions for different cluster sizes (left) and in three dimensions for equal cluster sizes (right), with cluster dispersion corresponding to 20% of the overall dispersion.

The well-known labelling problem for Monte Carlo experiments in mixture distributions was solved by assigning the labels to the given cluster centres in connection with the nearest column point. The given results were compared with the original configuration using Procrustes analysis (see Figure 5). Table 3 shows the given results in terms of the mean value of Procrustes index, which measures the accuracy of the given configuration, and the mean CPU time, for each combination of factors. Although the nature of the simulated data sets deviates somewhat from the hypothesis of the proposed model, very low values of the Procrustes index were found in all situations, which indicates that the given configuration matched very well to the corresponding original configuration in all situations. Although the aim here is not to recover the configuration of the points in itself, but the cluster centres, the Procrustes value indicates the classification of points was in general well recovered. The performance of the LCDA procedure is improved when minor overlap exists between clusters, as in general can be expected when mixture models are used for clustering. In terms of CPU time, the algorithm is efficient in all situations.

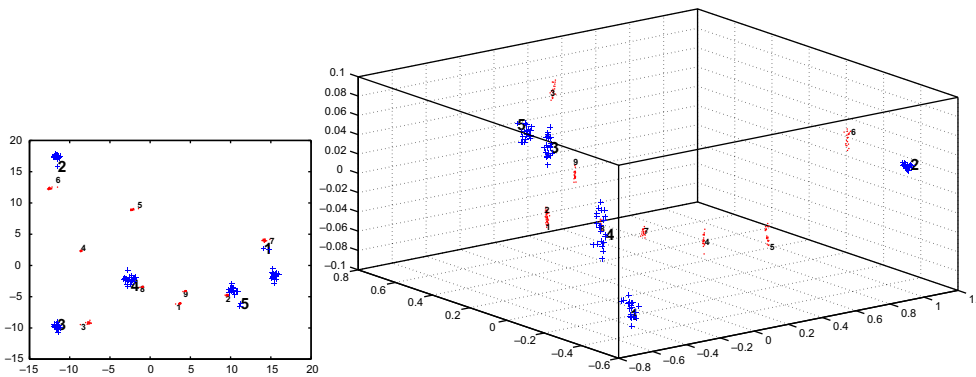

Figure 5. Procrustes configuration for 20 samples of 500 simulated clustered data from a multivariate normal distribution in two dimensions for different cluster sizes (left) and in three dimensions for equal cluster sizes (right), with cluster dispersion corresponding to 60% of the overall dispersion.

Table 3. Mean Procrustes values $\overline{p}$ and CPU time $t(s)$ for the normal simulated data sets with equal and different cluster sizes in two and three dimensions. Rows represent different cluster dispersions (overlap) in terms of percentage of variability with respect to the overall dispersion for each generated data set

	Dim 2				Dim 3			
	Equal size		Unequal size		Equal size		Unequal size	
Dispersion (%)	$\overline{p}$	$t(s)$	$\overline{p}$	$t(s)$	$\overline{p}$	$t(s)$	$\overline{p}$	$t(s)$
20	.0021	12.1	.0022	15.3	.0001	21.6	.0002	21.8
40	.0025	12.4	.0087	22.6	.0006	12.8	.0006	20.1
60	.0211	16.1	.0184	18.7	.0015	15.9	.0017	21.8
80	.0377	17.2	.0239	21.1	.0029	19.2	.0068	27.25

In summary, from the Monte Carlo experiment it follows that under the hypothesis of the model, the proposed LCDA procedure well recovered the clustering structure, even for fairly sparse data sets. For synthetically clustered data sets, both the original configuration and the clustering structure were also well recovered, in all tested combinations of dimensionality, cluster overlap and size.

4.2. Dutch parliamentary election studies

We now analyse the data introduced in Section 1. In Section 4.2.1 we show the results of our new approach, the LCDA model. Then we show the results obtained using a standard DA model (Section 4.2.2) and a multinomial regression model (Section 4.2.3).

4.2.1. LCDA model

The proposed model was applied to the 2006 Dutch parliamentary election studies. To determine the appropriate number of latent classes, we tested the model up to 20 classes without considering geometrical constraints, and seven classes were selected. Minor differences were found between the values for the BIC^* statistics in one, two and three dimensions, and the $T = 7$ class solution in two dimensions ($BIC^* = 2{,}657.25$) was selected, corresponding to a model of 33 parameters. Table 4 presents the characteristics of the seven classes. It can be seen that class 3 (485 participants) only has two membership patterns (no membership of any organization, and membership of only sports organizations), whereas class 2 (378 participants) has 181 different membership patterns.

Table 5 shows the mean profiles for each of the classes. From this table we can derive the following. Class 1 has a very high probability (greater than .9) of being a member of a church or an ideological organization (C). Class 2 has similar probabilities for membership

Table 4. Some characteristics of the seven classes. r_t is the number of membership profiles in class t, n_t the number of participants in the class

Class	r_t	n_t
1	10	212
2	181	378
3	2	485
4	28	204
5	3	209
6	22	322
7	7	366

Table 5. Conditional proportions by cluster for each type of organization

	E	H	T	P	W	N	M	S	C
1	0.3113	0.1887	0.1509	0.0802	0	0.2689	0.2123	0.3066	0.9198
2	0.5529	0.5344	0.4656	0.5423	0.1772	0.5026	0.3942	0.5556	0.4762
3	0	0	0	0	0	0	0	0.3299	0
4	0.4510	0.2892	0.1912	0.1618	0.0294	0.5735	0.5833	0.4853	0.8676
5	0	0	0	0	0	0.2249	0	0.2249	1
6	0.5870	0.3509	0.2826	0.1149	0	0.3230	0.2609	0.4938	0
7	0.3470	0.0929	0.2760	0	0	0.2842	0	0.4344	0

of all types of organizations, except employers' organizations (W). Class 3 is characterized by membership only of a sports club (S) at most. Class 4 has a very high probability (greater than .8) of church membership, and similar probabilities of belonging to a music or cultural organization (M), a neighbourhood organization (N), a sports organization, and an environmental or nature organization (E). Participants in class 5 are all members of the church or an ideological organization, have small probabilities for being a member of a sports club or neighborhood organization, and are not a member of all other organizational types. Class 6 has a moderate probability of being a member of an environmental or nature organization, as well as of belonging to a sports organization, and no probability of being a member of an employers' organization, or a church or ideological organization, and some probability of being member of the other organizations. Finally, class 7 is characterized by not belonging to a professional organization (P), employers' organization, music or cultural organization, or a church or ideological organization, but having a moderate probability of being a member of a sports organization and an environmental or nature organization, and also some probability of being a member of a neighbourhood organization and a trade union (T).

The solution with the seven latent classes is shown in Figure 6. In interpreting this, the odds and odds ratios are of some importance. The odds of a response j against a response j' for a given class t are given by

$$\log\left(\frac{\mu_{tj}}{\mu_{tj'}}\right) = \log(\beta_j) - \log(\beta_{j'}) - d_{tj}^2 + d_{tj'}^2.$$

The odds are a function of both the main effect parameters and the distances. Concerning the distances, the odds are in favour of the closest category. Concerning

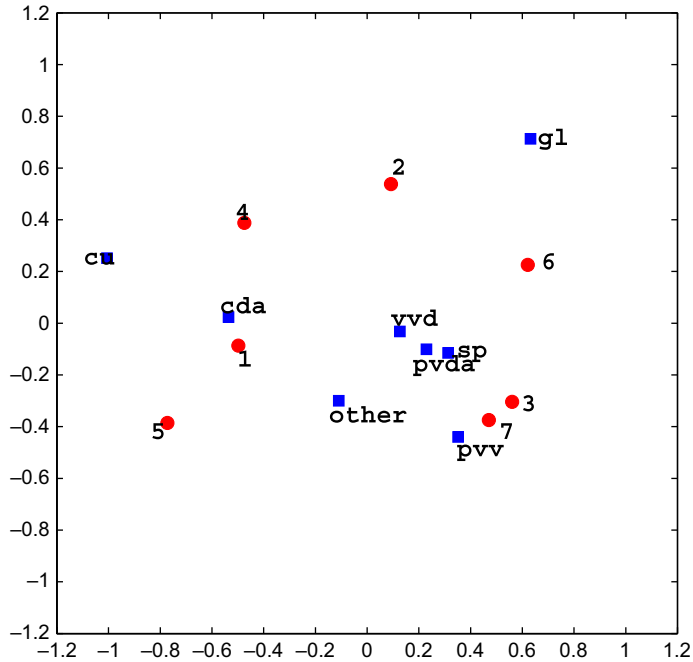

Figure 6. Solution of the LCDA model for membership and vote data.

Table 6. Squared distances between classes and political parties. The smallest distance for each class is shown in bold; distances that are also relatively small are shown in italic. The final row gives the estimated log βs

	CDA	PvdA	VVD	GL	SP	CU	PVV	Other
1	**0.0136**	0.5290	0.3931	1.9153	0.6579	0.3719	0.8465	*0.1960*
2	0.6589	*0.4271*	*0.3261*	**0.3214**	*0.4742*	1.2877	1.0235	0.7423
3	1.3101	*0.1511*	0.2624	1.0400	*0.0974*	2.7621	**0.0623**	0.4498
4	**0.1366**	0.7348	0.5382	1.3287	0.8726	*0.3009*	1.3685	0.6062
5	**0.2237**	1.0837	0.9329	3.1783	1.2504	*0.4608*	1.2653	*0.4459*
6	1.3792	*0.2604*	*0.3109*	*0.2378*	**0.2109**	2.6467	0.5160	0.8103
7	1.1715	*0.1330*	*0.2355*	1.2093	*0.0925*	2.5707	**0.0184**	0.3424
log (β)	1.1390	0.5048	0.1249	-0.4421	0.3066	-0.1431	-0.8134	-0.6767

the main effects, the odds are in favour of the category with the largest β value. For a detailed examination of the interplay of the βs and distances, see Takane (1998) and De Rooij (2009). The squared distances and estimated log βs are given in Table 6. The log βs show that the odds are generally in favour of CDA, while the value for Other is very small.

Figure 6 and Table 6 show that the classes with high probabilities for membership of a church (1, 4, and 5) are close to the Christian Democrats (CDA) and the Christian Union (CU). Class 6 is a left-oriented class close to the Socialist Party (SP), Green Left (GL), and the Labour Party (PvdA). Classes 3 and 7 are close to the Party for Freedom (PVV) but also relatively close to the PvdA and the SP. Class 2 is closest to the GL, but also relatively close to PvdA, CDA, Conservative Liberals (VVD), SP, and other, (i.e. this is a relatively heterogeneous class).

The odds ratio can be defined in terms of squared distances (De Rooij & Heiser, 2005):

$$\frac{\mu_{tj} \times \mu_{t'j'}}{\mu_{tj'} \times \mu_{t'j}} = \exp\left(-d_{tj}^2 - d_{t'j'}^2 + d_{tj'}^2 + d_{t'j}^2\right). \tag{23}$$

Using the distances reported in Table 6 and this formula, the following conclusions can be drawn:

- The odds that participants from class 1 choose CDA instead of PvdA are

$$\exp(-0.0136 - 0.1330 + 0.5290 + 1.1715) = \exp(1.5539) = 4.7298$$

times the odds that participants from class 7 choose CDA instead of PvdA.

- The odds that participants from class 6 choose GL instead of PVV are

$$\exp(-0.2378 - 0.0184 + 0.5160 + 1.2093) = \exp(1.4691) = 4.3453$$

times these odds for participants from class 7.

- The odds that participants from class 2 choose GL instead of VVD are

$$\exp(-0.3214 - 0.3109 + 0.3261 + 0.2378) = \exp(-0.0684) = 0.9339$$

times these odds for participants from class 6.

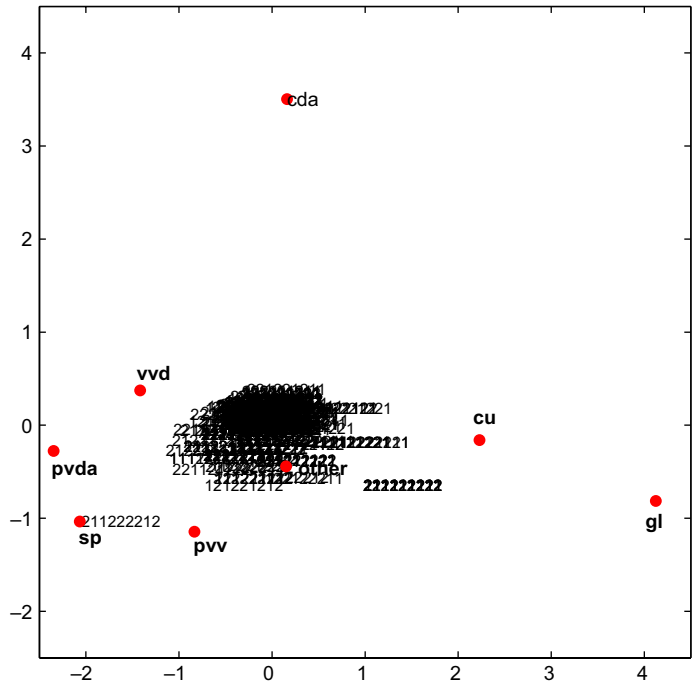

Figure 7. Result of the distance association model. Profiles are given in the form 112112222 where a '1' indicates membership and '2' non-membership. The order of the numbers corresponds to EHTPWNMSC.

4.2.2. Distance association model

By way of comparison we also give the result of a two-dimensional DA model (Figure 7). It can be seen that this graph is largely uninterpretable due to the large number of profiles. The number of free parameters for this model is 774; the BIC* statistic equals 5,575.477.

4.2.3. Multinomial regression model

In general, a researcher would also be interested in interaction effects between the predictor variables. With nine dichotomous predictor variables many models are possible, for example $(9 \times 8)/2 = 36$ different two-variable interactions can be included, and more interactions are possible when three- or more-variable interactions are also taken into account. An approach often taken is to use backward selection. For the Dutch parliamentary election studies, however, such an approach fails due to the excessive number of zero frequencies in the data. That is, we started with a hierarchical model using up to five-variable interactions between the predictor variables, but this analysis failed – parameter estimates cannot be computed. We then deleted the five-variable interactions and started with four-variable interactions, but this, too, failed. Removing the four-variable interactions again failed. Even the backward elimination scheme starting with a model with only pairwise associations failed.

The reason why these models fail is that the odds ratios (or ratios of odds ratios for two-way and higher interaction effects) are undefined. Consider a 2×2 frequency table. An estimate of the odds ratio for this table is obtained by multiplying the diagonal elements

Table 7. Regression coefficients of the multinomial logistic regression model

	CDA	PvdA	VVD	GL	SP	CU	PVV
Int	1.0060	1.4080	0.8537	−0.5899	1.2656	−1.5852	0.3588
C	1.0010*	−0.7320*	−0.6354*	−1.2224*	−1.0089*	2.0959*	−1.0703*
M	0.6460*	0.3160	0.7823*	0.3951	0.4684	0.6353	−0.2631
W	1.3820	1.1140	2.1154*	1.1802	0.7867	−18.2656†	1.4597
T	0.1350	0.8030*	−0.5535	0.4597	0.4412	0.2299	0.1250
E	−0.5300	−0.5560*	−0.2462	0.4124	−0.3658	−0.6272*	−0.5553
H	−0.2120	0.3060	−0.5034	0.9066*	0.3137	0.8952	−1.0161
P	0.1450	−0.3640	0.3385	0.0087	−0.4574	−0.4014	0.1181
S	0.3740	0.2020	0.6883*	0.3814	0.2658	−0.3935	0.1531

*Regression coefficient with *p*-values smaller than .05. †Estimate for which the standard error cannot be computed.

and dividing by the product of off-diagonal elements. Now, if a diagonal cell equals zero and an off-diagonal cell equals zero the odds ratio is undefined. In our large sparse data set this occurs for many 2×2 subtables. Since the parameters of the multinomial logistic regression are directly based on these undefined odds ratios the model is not estimable.

In the backward elimination scheme starting with only the main effects, the variable *N* is removed while the others must be included in the model. Some parameter values of this final model are very large due to complete separation. Complete separation occurs in a 2×2 table if either a diagonal cell or an off-diagonal cell equals zero. The log of the odds ratio in that case is either infinity or minus infinity. The estimates are given in Table 7. It can be seen that membership of a church or religious organization influences all aspects, raising the odds for CDA and CU, and lowering all other odds. There is one very large estimate, for which the standard error cannot be computed. SPSS warns that validity of this model is uncertain. Therefore, we did not proceed with the interpretation of all coefficients.

This final multinomial response model has 63 parameters (as shown in Table 7), whereas the LCDA model has 33.

5. Conclusion

In this paper we develop a latent class model for profile by response cross-classified data. The proposed model allows us to find the most appropriate classification for the profiles (combinations of the categories of the explanatory variables) while simultaneously representing the classes and column categories in a low-dimensional space by means of a DA model. In a cross-classified framework, this approach is particularly suitable when the number of profiles is large, and hence model selection procedures are cumbersome, when a very large number of parameters must be estimated, or when there are difficulties with sparse data sets containing too many zero entries.

As in all finite-mixture models, the proposed procedure allows the possibility of determining the number of classes using several procedures, among which the adjusted BIC* statistic seems to be an efficient criterion for also selecting the appropriate dimensionality for the model.

The model lets us analyse cross-classified data without distinguishing between predictor and response categories. Li and Zha (2006) also proposed a two-way Poisson

mixture model for cross-classified data, but their model is somewhat tangential to ours. In their model, each column class is characterized by a mixture model with a row clustering structure. The distribution of the column vectors within each class is modelled by a mixture of multivariate Poisson distributions for which a clustering structure is also imposed in the rows, leading to a two-way mixture model for each class. Also, this model does not include the dimension reduction provided by the DA model.

The EM algorithm is a fundamental tool for fitting mixture models by maximum likelihood, particularly in the proposed latent class model. As mentioned in Section 3.2, different starting values for the EM algorithm can lead to different estimates. Convergence of the EM algorithm is slow and the situation will be exacerbated by a poor choice of initial parameter values. Furthermore, in several situations, the sequence of estimates generated by the EM algorithm may diverge if the initial parameter values chosen are too close to the boundary (see McLachlan & Peel, 2000). Another common problem with mixture models is that the likelihood equation will usually have multiple local maxima. Therefore, the performance of optimization heuristics to escape from local optima is an important issue, and this is currently being investigated by the authors.

Acknowledgements

Part of this research was conducted while J. Fernando Vera was a guest researcher at the Division of Methodology and Statistics, Institute of Psychology, Faculty of Social Sciences, at Leiden University. This research was also conducted while M. de Rooij was sponsored by the Netherlands Organization for Scientific Research (NWO), Innovational Grant, 452-06-002. The data utilized in this publication were originally collected for the Dutch Parliamentary Election Studies 2006 by Statistics Netherlands and by Kees Aarts, Henk van der Kolk, Martin Rosema and Martha Brinkman on behalf of the Foundation for Electoral Research in the Netherlands (Stichting Kiezersonderzoek Nederland, SKON). These studies have been made possible by grants from Statistics Netherlands, the Netherlands Organization for Scientific Research (NWO), the Ministry of the Interior and Kingdom Relations (BZK), the Ministry of Health, Welfare and Sports (VWS), the Social and Cultural Planning Office (SCP), and the Department of Political Science and Research Methodology, University of Twente. The original collectors of the data bear no responsibility for the analyses or interpretations published here. The data are distributed by NIWI/Steinmetz Archive, Amsterdam, by the ICPSR, Ann Arbor, Michigan, USA, and the Zentral Archiv, Cologne, Germany.

References

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

Becker, M. P. (1990). Maximum likelihood estimation of the RC(M) association model. *Applied Statistics*, *39*, 152–167.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B*, *25*, 220–233.

Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika*, *31*, 33–42.

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

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.

De Rooij, M. (2009). Ideal point discriminant analysis revisited with an emphasis on visualization. *Psychometrika*, *74*, 317–330.

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*, 99–122.

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.

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*, 612–650.

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. *Annals of Statistics*, *13*, 10–69.

Hope, A. C. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society, Series B*, *30*, 582–598.

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

Li, J., & Zha, H. (2006). Two-way Poisson mixture models for simultaneous classification and word clustering. *Computational Statistics and Data Analysis*, *50*, 163–180.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

Rissanen, J. (1986). Stochastic complexity. *Annals of Statistics*, *14*, 1080–1100.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. World Scientific series in computer science, v. 15. Singapore/Teaneck, NJ: World Scientific.

Schwarz, G. (1978). Estimation the dimensions of a model. *Annals of Statistics*, *6*, 461–464.

Seidel, W., Mosler, K., & Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, *52*, 481–487.

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.

Takane, Y., Van der Heijden, P. G. M., & Browne, M. (2003). On likelihood ratio tests for dimension selection. In T. Higuchi, Y. Iba, & M. Ishiguro (Eds.), *Proceedings of Science of Modeling: The 30th Anniversary of the Information Criterion (AIC)* (pp. 348–349). Report on Research and Education 17. Tokyo, Japan: Institute of Statistical Mathematics.

Vera, J. F., Macías, R., & Angulo, J. M. (2009). A latent class MDS model with spatial constraints for non-stationary spatial covariance estimation. *Stochastic Environmental Research and Risk Assessment*, *23*(6), 769–779.

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.

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.

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.

Vichi, M., & Kiers, H. A. L. (2001). Factorial *k*-means analysis for two-way data. *Computational Statistics and Data Analysis*, *37*, 49–64.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Yang, C.-C., & Yang, C.-C. (2007). Separating latent classes by information criteria. *Journal of Classification*, *24*, 183–203.

Received 19 December 2012; revised version received 23 December 2013

Appendix A: Connection between the LCD models from Poisson and product multinomial sampling

In a Poisson sampling model, counts are considered as independent random variables, and the total count $N = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij}$ is random rather than fixed. In general, conditional on the total count N, cell frequencies are no longer independent, and assuming that samples in different rows are independent, product multinomial sampling arises. Thus, if we start from a Poisson model considering the row totals as fixed, the conditional distribution is also a product multinomial distribution (see Agresti, 2013, for further details). Nevertheless, conditional on a row i, multinomial sampling arises directly by assuming the observations on $\mathbf{f}_i$ are independent samples, each having probability distribution $\{\pi_{1|i}, \ldots, \pi_{J|i}\}$, where, considering the usual decomposition of the expected frequency in the general multiplicative form,

$$\pi_{j|i} = \frac{\beta_j \exp\left(-d_{ij}^2\right)}{\sum_k \beta_k \exp\left(-d_{ik}^2\right)}.$$

In a product multinomial sampling model, the probability for the data of a row element $\mathbf{f}_i \in \mathbf{F}_t$ is given by

$$h_t(\mathbf{f}_i|\mathbf{x}_t, \mathbf{Y}, \boldsymbol{\beta}) = \frac{f_{i.}!}{\prod_j f_{ij}!}\prod_{j=1}^{J}\left(\frac{\beta_j \exp(-d_{tj}^2)}{\sum_k \beta_k \exp(-d_{tk}^2)}\right)^{f_{ij}}, \tag{A1}$$

where $f_{i.} = \sum_j f_{ij}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$. The p.d.f. of the random variable $\mathbf{f}_i$ is a finite mixture of product multinomial densities given by (A1), which can be expressed as

$$g(\mathbf{f}_i|\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{t=1}^{T} \gamma_t h_t(\mathbf{f}_i|\mathbf{x}_t, \mathbf{Y}, \boldsymbol{\beta}), \tag{A2}$$

where $\boldsymbol{\gamma} = (\gamma_1 \ldots, \gamma_T)'$. So the log-likelihood function to be maximized subject to (1) can be written as

$$\log L = \sum_{i=1}^{I} \log \sum_{t=1}^{T} \gamma_t h_t(\mathbf{f}_i|\mathbf{x}_t, \mathbf{Y}, \boldsymbol{\beta}). \tag{A3}$$

Considering the EM algorithm for the parameter estimation, at the M-step, the value of $\hat{\gamma}_t$ is estimated from $\hat{\mathbf{Z}}$ and the remaining parameters are estimated by maximizing

$$q_{PM}(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}|\hat{\mathbf{Z}}) = \sum_{i=1}^{I}\sum_{t=1}^{T} \hat{z}_{it} \log h_t(\mathbf{f}_i|\mathbf{x}_t, \mathbf{Y}, \boldsymbol{\beta}). \tag{A4}$$

Then, taking logarithms in (A1) and writing $f_{tj} = \sum_i \hat{z}_{it} f_{ij}$, expression (A4) can be written as

$$q_{PM}(\mathbf{X},\mathbf{Y},\boldsymbol{\beta}|\hat{\mathbf{Z}}) = \sum_{i} \log(f_{i.}!) - \sum_{i=1}^{I}\sum_{j=1}^{J} \log(f_{ij}!) + \sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj}\log(\beta_j) - \sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj} d_{tj}^2 - \sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj}\log\left(\sum_{k=1}^{J}\beta_k \exp(-d_{tk}^2)\right). \tag{A5}$$

In the M-step of the EM algorithm in a Poisson sampling scheme, after substituting (2) into (3), the parameters are estimated by maximizing (14), which, given $\hat{\mathbf{Z}}$, can be written as

$$\begin{aligned} q_P(\mathbf{X},\mathbf{Y},\mu,\boldsymbol{\alpha},\boldsymbol{\beta}|\hat{\mathbf{Z}}) = {} & \sum_{i=1}^{I}\sum_{t=1}^{T}\hat{z}_{it}\sum_{j=1}^{J}\log(\mu\alpha_t) + \sum_{i=1}^{I}\sum_{t=1}^{T}\hat{z}_{it}\sum_{j=1}^{J} f_{ij}\log(\beta_j) \\ & - \sum_{i=1}^{I}\sum_{t=1}^{T}\hat{z}_{it}\sum_{j=1}^{J} f_{ij} d_{tj}^2 - \sum_{i=1}^{I}\sum_{t=1}^{T}\hat{z}_{it}\mu\alpha_t\sum_{j=1}^{J}\beta_j\exp(-d_{tj}^2) \\ & - \sum_{i=1}^{I}\sum_{t=1}^{T}\hat{z}_{it}\sum_{j=1}^{J}\log(f_{ij}!). \end{aligned} \tag{A6}$$

The stationary equations for μ and α_t are given by

$$\begin{aligned} \frac{\partial q_P}{\partial \mu} &= \sum_{i=1}^{I}\sum_{t=1}^{T}\hat{z}_{it}\sum_{j=1}^{J} f_{ij} - \mu\sum_{i=1}^{I}\sum_{t=1}^{T}\hat{z}_{it}\alpha_t\sum_{j=1}^{J}\beta_j\exp(-d_{tj}^2) = 0, \\ \frac{\partial q_P}{\partial \alpha_t} &= \sum_{i=1}^{I}\hat{z}_{it}\sum_{j=1}^{J} f_{ij} - \alpha_t\mu\sum_{i=1}^{I}\hat{z}_{it}\sum_{j=1}^{J}\beta_j\exp(-d_{tj}^2) = 0, \end{aligned} \tag{A7}$$

whence

$$\mu\alpha_t = \frac{\sum_{j=1}^{J} f_{tj}}{\sum_{i=1}^{I}\hat{z}_{it}\sum_{j=1}^{J}\beta_j\exp(-d_{tj}^2)}. \tag{A8}$$

Thus, substituting (A8) into (A6), the remaining parameters are given by maximizing the conditional log-likelihood given by

$$\begin{aligned} q_P(\mathbf{X},\mathbf{Y},\boldsymbol{\beta}|\mu,\boldsymbol{\alpha},\hat{\mathbf{Z}}) = {} & \sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj}\log\left(\sum_{i=1}^{I}\hat{z}_{it}\right) + \sum_{t=1}^{T}\sum_{=1}^{J} f_{tj}\log(f_{t.}) - f_{..} \\ & - \sum_{i=1}^{I}\sum_{j=1}^{J}\log(f_{ij}!) + \sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj}\log(\beta_j) - \sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj} d_{tj}^2 \\ & - \sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj}\log\left(\sum_{k=1}^{J}\beta_k\exp(-d_{tk}^2)\right), \end{aligned} \tag{A9}$$

where $f_{..} = \sum_t \sum_j f_{tj}$. Thus, writing $f_{t.} = \sum_j f_{tj}$ and $f_{i.} = \sum_j f_{ij}$,

$$q_{PM}(\mathbf{X},\mathbf{Y},\boldsymbol{\beta}|\hat{\mathbf{Z}}) = q_P(\mathbf{X},\mathbf{Y},\boldsymbol{\beta}|\mu,\boldsymbol{\alpha},\hat{\mathbf{Z}}) - \sum_i \log(f_{i.}!) - \sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj}\log\left(\sum_{i=1}^{I}\hat{z}_{it}\right) + \sum_{t=1}^{T}\sum_{=1}^{J} f_{tj}\log(f_{t.}) - f_{..}. \tag{A10}$$

Appendix B: A Newton–Raphson algorithm for maximum likelihood estimation

Given a classification $\hat{\mathbf{Z}}$ of the rows of $\mathbf{F}$, the unconditional probability for each latent class $\hat{\gamma}_t$ is given by (13). Then, the overall log-likelihood (10) can be maximized with respect to parameters $\mathbf{X}$, $\mathbf{Y}$, μ, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by maximizing (15), which, except for a constant term, can be expressed as

$$\begin{aligned} q(\mathbf{X},\mathbf{Y},\mu,\boldsymbol{\alpha},\boldsymbol{\beta}|\hat{\mathbf{Z}}^{(s)}) &= \sum_{i=1}^{I}\sum_{t=1}^{T}\hat{z}_{it}^{(s)}\sum_{j=1}^{J}\left[f_{ij}\log(\mu_{tj}) - \mu_{tj}\right] \\ &= \sum_{t=1}^{T}\sum_{j=1}^{J}\left(\sum_{i=1}^{I}\hat{z}_{it}^{(s)}f_{ij}\right)\log(\mu) + \sum_{t=1}^{T}\sum_{j=1}^{J}\left(\sum_{i=1}^{I}\hat{z}_{it}^{(s)}f_{ij}\right)\log(\alpha_t) \\ &\quad + \sum_{t=1}^{T}\sum_{j=1}^{J}\left(\sum_{i=1}^{I}\hat{z}_{it}^{(s)}f_{ij}\right)\log(\beta_j) - \sum_{t=1}^{T}\sum_{j=1}^{J}\left(\sum_{i=1}^{I}\hat{z}_{it}^{(s)}f_{ij}\right)d_{tj}^2(\mathbf{x}_t,\mathbf{y}_j) \\ &\quad - \sum_{t=1}^{T}\sum_{j=1}^{J}\left(\sum_{i=1}^{I}\hat{z}_{tj}^{(s)}\right)\exp\Big(\log(\mu) + \log(\alpha_t) + \log(\beta_j) - d_{tj}^2(\mathbf{x}_t,\mathbf{y}_j)\Big). \end{aligned} \tag{B1}$$

Writing $\lambda = \log\,\mu$, $\lambda_t^R = \log\alpha_t$, $\lambda_j^C = \log\beta_j$, and $f_{tj} = \sum_{i=1}^{I}\hat{z}_{it}f_{ij}$,

$$f_{..} = \sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj}, \quad f_{t.} = \sum_{j=1}^{J} f_{tj}, \quad f_{.j} = \sum_{t=1}^{T} f_{tj},$$

and taking into account that

$$\begin{aligned} \sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj}d_{tj}^2(\mathbf{x}_t,\mathbf{y}_j) &= \sum_{t=1}^{T}\left(\sum_{j=1}^{J} f_{tj}\right)\mathbf{x}_t'\mathbf{x}_t + \sum_{j=1}^{J}\left(\sum_{t=1}^{T} f_{tj}\right)\mathbf{y}_j'\mathbf{y}_j \\ &\quad - 2\sum_{t=1}^{T}\sum_{j=1}^{J} f_{tj}\mathbf{x}_t'\mathbf{y}_j, \end{aligned} \tag{B2}$$

then (B1) can be written as

$$
\begin{aligned}
q(\mathbf{X},\mathbf{Y},\mu,\boldsymbol{\alpha},\boldsymbol{\beta}|\hat{\mathbf{Z}}^{(s)}) =& f_{..}\lambda + \sum_{t=1}^{T} f_{t.}\lambda_t^R + \sum_{j=1}^{J} f_{.j}\lambda_j^C \\
&- \mathrm{tr}\mathbf{X}'\mathbf{D}_R\mathbf{X} - \mathrm{tr}\mathbf{Y}'\mathbf{D}_C\mathbf{Y} + 2tr\mathbf{X}'\mathbf{F}_T\mathbf{Y} \\
&- \sum_{t=1}^{T}\sum_{j=1}^{J} I\gamma_t \exp(\lambda + \lambda_t^R + \lambda_j^C - \mathbf{x}_t'\mathbf{x}_t - \mathbf{y}_j'\mathbf{y}_j + 2\mathbf{x}_t'\mathbf{y}_j),
\end{aligned} \tag{B3}
$$

where $\mathbf{D}_R = \mathrm{diag}(f_{1.} \ldots, f_{T.})$ denotes a $T \times T$ diagonal matrix, $\mathbf{D}_C = \mathrm{diag}(f_{.1} \ldots, f_{.J})$ a $J \times J$ diagonal matrix, and $\mathbf{F}_T = (f_{tj})$ the $T \times J$ block matrix of $\mathcal{P}(\mathbf{F})$.

To maximize (B3), an iterative procedure based on the usual Newton–Raphson theory can be employed. Thus, writing

$$
M_{tj} = I\hat{\gamma}_t\mu_{tj} = I\hat{\gamma}_t \exp(\lambda + \lambda_t^R + \lambda_j^C - \mathbf{x}_t'\mathbf{x}_t - \mathbf{y}_j'\mathbf{y}_j + 2\mathbf{x}_t'\mathbf{y}_j),
$$

(B3) can be written as

$$
\begin{aligned}
q(\mathbf{X},\mathbf{Y},\mu,\boldsymbol{\alpha},\boldsymbol{\beta}|\hat{\mathbf{Z}}) =& f_{..}\lambda + \sum_{t=1}^{T} f_{t.}\lambda_t^R + \sum_{j=1}^{J} f_{.j}\lambda_j^C \\
&- \mathrm{tr}\mathbf{X}'\mathbf{D}_R\mathbf{X} - \mathrm{tr}\mathbf{Y}'\mathbf{D}_C\mathbf{Y} + 2\mathrm{tr}\mathbf{X}'\mathbf{F}_T\mathbf{Y} \\
&- \sum_{t=1}^{T}\sum_{j=1}^{J} M_{tj}.
\end{aligned} \tag{B4}
$$

Taking the derivative of q with respect to λ, and writing $M_{t.} = \sum_{j=1}^{J} M_{tj}, M_{.j} = \sum_{t=1}^{T}$, and $M_{..} = \sum_{t=1}^{T}\sum_{j=1}^{J} M_{tj}$, we obtain

$$
\frac{\partial q}{\partial \lambda} = f_{..} - M_{..}(\lambda) = 0, \quad \frac{\partial^2 q}{\partial^2 \lambda} = -M_{..}(\lambda), \tag{B5}
$$

$$
\frac{\partial q}{\partial \lambda_t^R} = f_{t.} - M_{t.}(\lambda_t^R) = 0, \quad \frac{\partial^2 q}{\partial^2 \lambda_t^R} = -M_{t.}(\lambda_t^R), \tag{B6}
$$

$$
\frac{\partial q}{\partial \lambda_j^C} = f_{t.} - M_{j}(\lambda_j^C) = 0, \quad \frac{\partial^2 q}{\partial^2 \lambda_j^C} = -M_{j}(\lambda_j^C), \tag{B7}
$$

$$\begin{aligned}\frac{\partial q}{\partial \mathbf{x}_t} &= -\frac{\partial}{\partial \mathbf{x}_t}\mathrm{tr}\mathbf{X}'\mathbf{D}_R\mathbf{X} + 2\frac{\partial}{\partial \mathbf{x}_t}\mathrm{tr}\mathbf{X}'\mathbf{F}_T\mathbf{Y} - \frac{\partial}{\partial \mathbf{x}_t}\mathrm{tr}\mathbf{Y}'\mathbf{D}_C\mathbf{Y} \\ &= -2f_{t.}\mathbf{x}_t + 2\sum_{j=1}^{J} f_{tj}\mathbf{y}_j - 2\sum_{j=1}^{J} I\gamma_t\mu_{tj}(\mathbf{y}_j - \mathbf{x}_t) \\ &= 2\sum_{j=1}^{J}(M_{tj} - f_{tj})(\mathbf{x}_t - \mathbf{y}_j) = 0, \\ \frac{\partial^2 q}{\partial^2 \mathbf{x}_t} &= -4\sum_{j=1}^{J} M_{tj}(\mathbf{x}_t - \mathbf{y}_j)^2 + 2\sum_{j=1}^{J}(M_{tj} - f_{tj}),\end{aligned} \tag{B8}$$

$$\begin{aligned}\frac{\partial q}{\partial \mathbf{y}_j} &= 2\sum_{t=1}^{T}(M_{tj} - f_{tj})(\mathbf{y}_j - \mathbf{x}_t) = 0, \\ \frac{\partial^2 q}{\partial^2 \mathbf{y}_j} &= -4\sum_{t=1}^{T} M_{tj}(\mathbf{y}_j - \mathbf{x}_t)^2 + 2\sum_{t=1}^{T}(M_{tj} - f_{tj}).\end{aligned} \tag{B9}$$

Then the updates in the $(s+1)$th iteration, according to the Newton–Raphson theory, are given by

$$\lambda^{(s+1)} = \lambda^{(s)} + \frac{f_{..} - M_{..}(\lambda^{(s)})}{M_{..}(\lambda^{(s)})}, \tag{B10}$$

$$\lambda_t^{R(s+1)} = \lambda_t^{R(s)} + \frac{f_{t.} - M_{t.}(\lambda_t^{R(s)})}{M_{t.}(\lambda_t^{R(s)})}, \tag{B11}$$

$$\lambda_j^{C(s+1)} = \lambda_j^{C(s)} + \frac{f_{.j} - M_{.j}(\lambda_j^{C(s)})}{M_{.j}(\lambda_c^{C(s)})}, \tag{B12}$$

$$x_{tm}^{(s+1)} = x_{tm}^{(s)} + \frac{2\sum_{j=1}^{J}(f_{tj} - M_{tj})(x_{tm}^{(s)} - y_{jm}^{(s)})}{2\sum_{j=1}^{J}(f_{tj} - M_{tj}) - 4\sum_{j=1}^{J} M_{tj}(x_{tm}^{(s)} - y_{jm}^{(s)})^2}, \tag{B13}$$

$$y_{jm}^{(s+1)} = y_{jm}^{(s)} + \frac{2\sum_{t=1}^{T}(f_{tj} - M_{tj})(y_{jm}^{(s)} - x_{tm}^{(s)})}{2\sum_{t=1}^{T}(M_{tj} - f_{tj}) - 4\sum_{t=1}^{T} M_{tj}(y_{jm}^{(s)} - x_{tm}^{(s)})^2}. \tag{B14}$$

The likelihood function has many local maxima, and initial estimates for the iterative procedure can be given using Becker's (1990) procedure as described in Section 3.2.