PSYCHOMETRIKA—VOL. 82, NO. 4, 1007–1034
DECEMBER 2017
https://doi.org/10.1007/s11336-017-9578-5

A MODEL-BASED APPROACH TO SIMULTANEOUS CLUSTERING AND DIMENSIONAL REDUCTION OF ORDINAL DATA

MONIA RANALLI

THE PENNSYLVANIA STATE UNIVERSITY

ROBERTO ROCCI

UNIVERSITY OF TOR VERGATA

The literature on clustering for continuous data is rich and wide; differently, that one developed for categorical data is still limited. In some cases, the clustering problem is made more difficult by the presence of noise variables/dimensions that do not contain information about the clustering structure and could mask it. The aim of this paper is to propose a model for simultaneous clustering and dimensionality reduction of ordered categorical data able to detect the discriminative dimensions discarding the noise ones. Following the underlying response variable approach, the observed variables are considered as a discretization of underlying first-order latent continuous variables distributed as a Gaussian mixture. To recognize discriminative and noise dimensions, these variables are considered to be linear combinations of two independent sets of second-order latent variables where only one contains the information about the cluster structure while the other one contains noise dimensions. The model specification involves multidimensional integrals that make the maximum likelihood estimation cumbersome and in some cases infeasible. To overcome this issue, the parameter estimation is carried out through an EM-like algorithm maximizing a composite log-likelihood based on low-dimensional margins. Examples of application of the proposal on real and simulated data are performed to show the effectiveness of the proposal.

Key words: mixture models, reduction ordinal data, composite likelihood.

1. Introduction

Cluster analysis aims at partitioning the data into meaningful groups which should differ considerably from each other. The literature on clustering for continuous data is rich and wide; differently, that one developed for categorical data is still limited. Only in the last decades there has been an increasing interest in clustering categorical data, although they are encountered in many fields, such as in behavioral, social and health sciences. These variables are frequently of ordinal type, measuring attitudes, abilities or opinions, and practitioners often apply on their ranks models and techniques developed for continuous data. Several authors have shown how this procedure can give biased estimates and is definitely less efficient than a proper modelization that is able to take into account the ordinal nature of the data (e.g., Ranalli & Rocci, 2016a). Such models mainly adopt two approaches developed mainly in the factor analysis framework: Item Response Theory (IRT) (see e.g., Bock & Moustaki, 2007; Bartholomew et al., 2011), and the Underlying Response Variable (URV) (see e.g., Jöreskog 1990; Lee et al., 1990; Muthén 1984). In the former, the probabilities of the categories are assumed to be analytic functions of some latent variables having a particular cluster structure. The best known model is latent class analysis (LCA; Goodman, 1974) where the latent variable is nominal. Examples where the latent variables are continuous are found in Cagnone and Viroli (2012), McParland et al. (2014), Gollini and Murphy (2014). In the URV approach, the ordinal variables are seen as a

Correspondence should be made to Monia Ranalli, Department of Statistics, The Pennsylvania State University, State College, PA, USA. Email: mxr459@psu.edu

discretization of continuous latent variables jointly distributed as a finite mixture; examples are: Everitt (1988), Lubke and Neale (2008), Ranalli and Rocci (2016a). In both approaches, the use of latent continuous variables makes the estimation rather complex because it requires the computation of many high-dimensional integrals. The problem is usually solved by approximating the likelihood function. Indeed, several lines of research propose different approximations, but they share the same idea: replacing the full likelihood with a surrogate that is easier to maximize and make inference about model parameters. In this regard we mention some useful surrogate functions, such as the variational likelihood (Gollini & Murphy, 2014; Yang et al., 2014) or the pairwise likelihood (Ranalli & Rocci, 2016a) to cluster categorical or ordinal data. Besides this, other approaches based on simulating the hidden variables exist.

In some cases, the clustering problem is made more difficult by the presence of dimensions (named noise) that are uninformative for recovering the groups and could obscure the true cluster structure. Different approaches exist in the literature to identify discriminative dimensions that emphasize group separability and give a representation of the cluster structure discarding the irrelevant and redundant noise dimensions. Several techniques for simultaneous clustering and dimensionality reduction (SCR) have been proposed in a non-model-based framework for quantitative (e.g., Vichi & Kiers 2001; Rocci et al., 2011) or categorical data (e.g., Buuren & Heiser, 1989; Hwang et al., 2006). There are also approaches based on a family of mixture models which fits the data into a common discriminative subspace (see e.g., Kumar & Andreou, 1998; Bouveyron & Brunet, 2012b). The key idea is to assume a common latent subspace to all groups that is the most discriminative one. This allows to project the data into a lower dimensional space preserving the clustering characteristics in order to improve visualization and interpretation of the underlying structure of the data. The model can be formulated as a finite mixture of Gaussians with a particular set of constraints on the parameters.

The aim of this paper is to propose a model for SCR on ordered categorical data. Following the URV approach, the observed variables are considered as a discretization of underlying first-order latent continuous variables that are linear combinations of two independent sets of second-order latent variables where only one contains the information about the cluster structure, defining a discriminative subspace, while the other one contains noise dimensions. Technically, the variables in the first set are distributed as a finite mixture of Gaussians, while in the second set as a multivariate normal. When in the dataset there are noise variables, then they tend to coincide with the set of second-order noise latent variables. If they are not present, then the model could still be able to identify a reduced set of second-order discriminative latent dimensions. This allows us to reduce the number of parameters and identify the main features of the clustering structure.

The model specification involves multidimensional integrals that make the maximum likelihood estimation rather cumbersome and in some cases infeasible. To overcome this issue, the model is estimated within the EM framework maximizing a composite likelihood based on m-dimensional marginals. In the current work, we present the model estimation considering $m = 2$, i.e., a pairwise likelihood, that is the product of all possible likelihoods based on the bivariate marginals, as proposed in Ranalli and Rocci (2016a) and used to estimate models for ordinal data in different contexts (such as in Jöreskog & Moustaki 2001; Leon, 2005; Leon & Carrigre 2007; Katsikatsou et al., 2012, Katsikatsou & Moustaki, 2016). However, as long as sparsity is not a problem and computations are feasible, it is possible to use a higher m, as shown in the simulation study and real data applications. Under some regularity conditions (Lindsay, 1988; Molenberghs & Verbeke, 2005), the estimators are consistent, asymptotically unbiased and normally distributed. In general, they are less efficient than the full maximum likelihood estimators, even if in many cases the loss in efficiency is very small or almost null (Lindsay, 1988; Varin et al. 2011), but much more efficient in terms of computational complexity.

The plan of the paper is the following: in the second section, we present the general model; in Section 3 we describe how to take into account noise dimensions and/or variables; then the

pairwise algorithm used to estimate the model parameters is presented in Section 4. Sections 5, 6 and 7 deal with model identifiability issue, model selection problem, and comparison with related models, respectively. In Section 8 a simulation study has been conducted to investigate the behavior of the proposed methodology, while in Section 9 two applications to real data are illustrated. In the last section, some remarks are pointed out.

2. The General Model

Let $x_1, x_2, \ldots, x_P$ be ordinal variables and $c_i = 1, \ldots, C_i$ the associated number of categories for $i = 1, 2, \ldots, P$. There are $R = \prod_{i=1}^{P} C_i$ possible response patterns of the form $\mathbf{x}_r = (x_1 = c_1, x_2 = c_2, \ldots, x_P = c_P)$. Let $\mathbf{y}$ be the heteroscedastic latent Gaussian mixture $f(\mathbf{y}) = \sum_{g=1}^{G} p_g \phi\left(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$, where the p_g's are the mixing weights, such that $p_g > 0$ and $\sum_{g=1}^{G} p_g = 1$, and $\phi\left(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$ is the density of a P-variate normal distribution with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. Under the URV approach, the ordinal variables are considered as a discretization of $\mathbf{y}$, i.e., generated by thresholding $\mathbf{y}$, as follows

$$\gamma_{c_i-1}^{(i)} \leq y_i < \gamma_{c_i}^{(i)} \Leftrightarrow x_i = c_i,$$

where $-\infty = \gamma_0^{(i)} < \gamma_1^{(i)} < \cdots < \gamma_{C_i-1}^{(i)} < \gamma_{C_i}^{(i)} = +\infty$ are the thresholds defining the C_i categories. Let us set $\boldsymbol{\psi} = \left\{p_1, \ldots, p_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G, \boldsymbol{\gamma}\right\} \in \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is the parameter space. The probability of response pattern $\mathbf{x}_r$ is given by

$$\begin{aligned} Pr(x_1 = c_1, x_2 = c_2, \ldots, x_P = c_P; \boldsymbol{\psi}) &= \sum_{g=1}^{G} p_g \int_{\gamma_{c_1-1}^{(1)}}^{\gamma_{c_1}^{(1)}} \cdots \int_{\gamma_{c_P-1}^{(P)}}^{\gamma_{c_P}^{(P)}} \phi(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \mathrm{d}\mathbf{y} \\ &= \sum_{g=1}^{G} p_g \pi_r(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\gamma}), \end{aligned}$$

where $\pi_r\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\gamma}\right)$ is the probability of response pattern $\mathbf{x}_r$ in the cluster g and p_g is the probability of belonging to group g. Let u_{nr} be a dummy variable which assumes value 1 to indicate the response pattern occurred for observation n, and 0 otherwise. Thus, for a random i.i.d. sample of size N, the contribution of the nth observation to the log-likelihood is

$$\ell(\boldsymbol{\psi}; \mathbf{x}_n) = \sum_{r=1}^{R} u_{nr} \log \left[\sum_{g=1}^{G} p_g \pi_r \left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\gamma}\right) \right];$$

by summing over n, the log-likelihood can be written as

$$\ell(\boldsymbol{\psi}; \mathbf{x}) = \sum_{r=1}^{R} n_r \log \left[\sum_{g=1}^{G} p_g \pi_r \left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\gamma}\right) \right], \tag{1}$$

where n_r is the observed sample frequency of response pattern r and $\sum_{r=1}^{R} n_r = N$.

3. How to Detect the Presence of Noise Dimensions

Sometimes noisy dimensions are present in the data. These are dimensions that do not contain information about the cluster structure and could mask the true classes. It means that there exists a proper discriminative subspace, with a dimension less than the number of variables, where the clusters lie. In order to identify the discriminative subspace, in the previously described model, it is assumed that there is a second-order set of P latent variables $\tilde{\mathbf{y}}$, which in turn is formed of two independent subsets of variables. In the first one, there are Q (with $Q \leq P$) variables that have some clustering information, while in the second set there are $\bar{Q} = P - Q$ noise variables defining the so-called noise dimensions. Thus, it is assumed that only the first Q elements of $\tilde{\mathbf{y}}$ carry any class discrimination information defining the so-called discriminative subspace. Technically, the Q informative elements of $\tilde{\mathbf{y}}$ are assumed to be distributed as a mixture of Gaussians with class conditional means and variances equal to $E(\tilde{\mathbf{y}}^{Q} \mid g) = \boldsymbol{\eta}_g$ and $\text{Cov}(\tilde{\mathbf{y}}^{Q} \mid g) = \boldsymbol{\Omega}_g$, respectively. The $P - Q$ noisy elements do not contain information about the cluster structure, it follows that they are independent of $\tilde{\mathbf{y}}^{Q}$ and their distribution does not vary from one class to another. In particular we assume that $E(\tilde{\mathbf{y}}^{\bar{Q}} \mid g) = \boldsymbol{\eta}_0$ and $\text{Cov}(\tilde{\mathbf{y}}^{\bar{Q}} \mid g) = \boldsymbol{\Omega}_0$. The link between the two orders of latent variables $\tilde{\mathbf{y}}$ and $\mathbf{y}$ is given by a non-singular $P \times P$ matrix $\mathbf{A}$, as $\mathbf{y} = \mathbf{A}\tilde{\mathbf{y}}$. This means requiring a particular structure on the mean vectors and covariance matrices of $\mathbf{y}$. The assumption of multivariate normality in each component provides a convenient way of specifying the parameter structure. For each component g, the mean vector and the covariance matrix present the following structures,

$$\boldsymbol{\mu}_g = E(\mathbf{y} \mid g) = \mathbf{A}E(\tilde{\mathbf{y}} \mid g) = \mathbf{A}\begin{bmatrix} \eta_{g,1} \\ \vdots \\ \eta_{g,Q} \\ \eta_{0,1} \\ \vdots \\ \eta_{0,P-Q} \end{bmatrix} = \mathbf{A}\begin{bmatrix} \boldsymbol{\eta}_g \\ \boldsymbol{\eta}_0 \end{bmatrix}$$

and

$$\boldsymbol{\Sigma}_g = \text{Cov}(\mathbf{y} \mid g) = \mathbf{A}\text{Cov}(\tilde{\mathbf{y}} \mid g)\mathbf{A}' = \mathbf{A}\begin{bmatrix} \boldsymbol{\Omega}_g & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_0 \end{bmatrix}\mathbf{A}'.$$

Finally, one important step is to identify the observed variables that could be considered as noise. Intuitively x_p is a noise variable if y_p is well explained by $\tilde{\mathbf{y}}^{\bar{Q}}$. This information is included in the correlation matrix between the first- and second-order latent variables, $\mathbf{y}$ and $\tilde{\mathbf{y}}$, respectively. The matrix $\mathbf{A}$ plays a central role in estimating the correlation between $\mathbf{y}$ and $\tilde{\mathbf{y}}$, whose covariance matrix is given by $\mathbf{A}\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}}$; we remark that $\text{Cov}(\tilde{\mathbf{y}}) = \boldsymbol{\Sigma}_{\tilde{\mathbf{y}}}$ accounts for both the within and the between variance of the mixture. The variables $\mathbf{y}$ that are most correlated with variables $\tilde{\mathbf{y}}^{\bar{Q}}$ are identified as noise. However, it is worth noticing that, exploiting the independence between $\tilde{\mathbf{y}}^{Q}$ and $\tilde{\mathbf{y}}^{\bar{Q}}$, it is possible to compute proportions of each variable's variance that can be explained by the noise factors, and by one's complement, the proportions of each variable's variance that can be explained by the discriminative factors. They are very helpful in identifying the noise variables/dimensions when the correlations between $\mathbf{y}$ and $\tilde{\mathbf{y}}$ are not high.

4. Pairwise EM Algorithm

In the previous section, we have introduced the main contribution of the paper, that is the methodology aimed at reducing and clustering ordinal data simultaneously. In this section, we deal

with the model parameter estimation. It could be carried out through a full maximum likelihood estimation, but this becomes infeasible as the number of observed variables increases due to the presence of multidimensional integrals. Indeed, multidimensional integrals should be evaluated for each response pattern in the sample at several points of the parameter space. This is computationally demanding with a very low number of variables P and makes the model estimation prohibitive with P greater than 5. To overcome this computational issue, we suggest an alternative estimation procedure based on m-dimensional marginals, i.e., low-dimensional margins. As suggested in Ranalli and Rocci (2016a), the model is estimated within the expectation–maximization (EM) framework maximizing a pairwise likelihood, i.e., $m = 2$. It is a robust estimation method and its estimators have been proven to be consistent, asymptotically unbiased and normally distributed, under some regularity conditions (Lindsay, 1988; Varin et al., 2011; Molenberghs & Verbeke, 2005). In general they are less efficient than the full maximum likelihood estimators, or estimators obtained with a higher m, but in many cases the loss in efficiency is very small or almost null (Lindsay, 1988; Mardia et al., 2009). In the sequel we refer to the case $m = 2$, but, as long as sparsity is not a problem and computations are feasible, the framework can be easily extended to higher m.

The pairwise log-likelihood can be constructed as follows. Let $u^{(ij)}_{nc_ic_j}$ be a dummy variable which assumes value 1 to indicate the response pattern occurred for observation n, considering the pair of variables (x_i, x_j) with $i = 1, \ldots, P-1$ and $j = i+1, \ldots, P$, and 0 otherwise. Then, for a random i.i.d. sample of size N, the contribution of the nth observation to the pairwise log-likelihood is given by

$$p\ell(\boldsymbol{\psi}; \mathbf{x}_n) = \sum_{i=1}^{P-1} \sum_{j=i+1}^{P} \sum_{c_i=1}^{C_i} \sum_{c_j=1}^{C_j} u^{(ij)}_{nc_ic_j} \log \left[\sum_{g=1}^{G} p_g \pi^{(ij)}_{r_{c_ic_j}} (\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\gamma}) \right],$$

by summing over n the pairwise log-likelihood can be written as

$$p\ell(\boldsymbol{\psi}; \mathbf{x}) = \sum_{i=1}^{P-1} \sum_{j=i+1}^{P} \sum_{c_i=1}^{C_i} \sum_{c_j=1}^{C_j} n^{(ij)}_{c_ic_j} \log \left[\sum_{g=1}^{G} p_g \pi^{(ij)}_{c_ic_j} (\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\gamma}) \right], \tag{2}$$

where now, after the reparameterization, the set of models parameters is $\boldsymbol{\psi} = \{p_1, \ldots, p_G, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_G, \boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_G, \mathbf{A}, \boldsymbol{\gamma}\}$, $n^{(ij)}_{c_ic_j}$ is the observed frequency of a response in category c_i and c_j for variables x_i and x_j respectively, while $\pi^{(ij)}_{c_ic_j}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\gamma})$ is the corresponding probability obtained by integrating the (i, j) bivariate marginal of the normal distribution with parameters $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ between the given thresholds. Let $\mathbf{Z}$ denote the group membership matrix of order $\left(\sum_{i=1}^{P-1} \sum_{j=i+1}^{P} C_i \times C_j\right) \times G$, where $z^{(ij)}_{c_ic_j;g} = 1$ if the cell (c_i, c_j) belongs to component g and $z^{(ij)}_{c_ic_j;g} = 0$ otherwise, for $g = 1, \ldots, G$. The complete pairwise log-likelihood is defined as the pairwise log-likelihood assuming the component for each observation is known and it is given by

$$p\ell_c(\boldsymbol{\psi}; \mathbf{Z}, \mathbf{x}) = \sum_{i=1}^{P-1} \sum_{j=i+1}^{P} \sum_{c_i=1}^{C_i} \sum_{c_j=1}^{C_j} \sum_{g=1}^{G} n^{(ij)}_{c_ic_j} z^{(ij)}_{c_ic_j;g} \left[\log \left(\pi^{(ij)}_{c_ic_j} (\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\gamma}) \right) + \log \left(p_g \right) \right].$$

The E-step requires the computation of the expected value of the complete-data pairwise log-likelihood given the current estimates of the model parameters $\hat{\boldsymbol{\psi}}$. This is given by

$$Q\left(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}}\right) = E_{\hat{\boldsymbol{\psi}}}\left[p\ell_c(\boldsymbol{\psi}; \mathbf{Z}, \mathbf{x}|\mathbf{x})\right] = \sum_{i=1}^{P-1}\sum_{j=i+1}^{P}\sum_{c_i=1}^{C_i}\sum_{c_j=1}^{C_j}\sum_{g=1}^{G} n_{c_i c_j}^{(ij)}\hat{z}_{c_i c_j;g}^{(ij)}\left[\log\left(\pi_{c_i c_j}^{(ij)}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\gamma})\right) + \log\left(p_g\right)\right], \quad (3)$$

where

$$\hat{z}_{c_i c_j;g}^{(ij)} = E_{\hat{\boldsymbol{\psi}}}\left[Z_{c_i c_j;g}^{(ij)} = 1|x_i = c_i, x_j = c_j\right] = Pr_{\hat{\boldsymbol{\psi}}}\left[Z_{c_i c_j;g}^{(ij)} = 1|x_i = c_i, x_j = c_j\right].$$

In the M-step, we maximize the complete pairwise log-likelihood function subject to some identifiability constraints that will be specified in the sequel. Looking at the expected value in (3), the maximization can be decomposed in two parts: the former corresponds to the component parameters $\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$ and thresholds $\boldsymbol{\gamma}$, the second one to the mixture weights p_g's. The first part of the M-step has not a closed form; hence, to obtain the updates, its maximization has carried out through a quasi Newton-Raphson optimization method implemented by using an optimization routine called "fmincon" in Matlab (2013). On the other hand, the update of component weight $\hat{p}_g$ has a closed form and they are easily carried out as follows,

$$\hat{p}_g = \frac{\sum_{i<j}\sum_{c_i=1}^{C_i}\sum_{c_j=1}^{C_j} n_{c_i c_j}^{(ij)}\hat{z}_{c_i c_j;g}^{(ij)(t)}}{N},$$

with $g = 1, \ldots, G$. In order to ensure the positive-definiteness of the covariance matrices we compute them through their Cholesky decomposition. It means that the objective function is maximized with respect to $\mathbf{T}_g$ rather than $\boldsymbol{\Omega}_g$ where the $\mathbf{T}_g$'s are upper triangular matrices such that $\mathbf{T}_g\mathbf{T}_g' = \boldsymbol{\Omega}_g$ for $g = 0, 1, \ldots, G$. Finally, the threshold parameters do not change over the components, but each component is characterized by a different set of parameters; now, standardizing each component by making a change of variable, i.e., $w_i = (y_i - \mu_g^{(i)})/\sigma_g^{(ii)}$, we obtain new integration limits changing over the components. These are defined as

$$\tau_{c_{i;g}}^{(i)} = \frac{\gamma_{c_i}^{(i)} - \mu_g^{(i)}}{\sigma_g^{(ii)}}.$$

This allows to compute the probability of a response in category c_i and c_j for variables x_i and x_j, respectively, in (3) as

$$\begin{aligned}\pi_{c_i c_j}^{(ij)}(\mathbf{0}, \mathbf{R}_g, \boldsymbol{\tau}_{\cdot,g}^{(i)}, \boldsymbol{\tau}_{\cdot,g}^{(j)}) &= \int_{\tau_{c_i-1,g}^{(i)}}^{\tau_{c_i,g}^{(i)}}\int_{\tau_{c_j-1,g}^{(j)}}^{\tau_{c_j,g}^{(j)}} \phi\left(w_i, w_j; \rho_{ij}^{(g)}\right)\mathrm{d}w_i\mathrm{d}w_j \\ &= \Phi_2(\tau_{c_i,g}^{(i)}, \tau_{c_j,g}^{(j)}; \rho_{ij}^{(g)}) - \Phi_2(\tau_{c_i,g}^{(i)}, \tau_{c_j-1,g}^{(j)}; \rho_{ij}^{(g)}) \\ &\quad -\Phi_2(\tau_{c_i-1,g}^{(i)}, \tau_{c_j,g}^{(j)}; \rho_{ij}^{(g)}) + \Phi_2(\tau_{c_i-1,g}^{(i)}, \tau_{c_j-1,g}^{(j)}; \rho_{ij}^{(g)}), \end{aligned} \quad (4)$$

where $\Phi_2(a, b; \rho)$ is the bivariate cumulative standard normal distribution with correlation ρ evaluated at the thresholds a and b and $\phi(u, v; \rho)$ is the corresponding density. As regards the classification, in the context of mixture models estimated through a full likelihood, an observation is assigned to the component with the maximum a posteriori probability (MAP criterion), that is the component with the maximum scaled fit (scaled by the corresponding mixing weight). However, when we adopt a pairwise (or, more in general, composite likelihood) approach, this is not possible anymore, since we do not have the joint density for each observation. To solve the problem, there are at least two different solutions. According to the first one, it is possible to reconstruct the joint probabilities through the joint density: the composite likelihood estimates are plugged into the full likelihood. It follows that the classification of the observations can be easily based on the MAP criterion, as it happens in the context of standard mixture models. On the other hand, in the same fashion, for each observation it is possible to evaluate the scaled pairwise-fit (or composite fit) of each component and assign the observation to the component corresponding to the maximum fit (CMAP criterion). In the first case, it is true that there are still multidimensional integrals, but they have not to be evaluated many times (as it is needed in the estimation), but only once. However, CMAP is more efficient computationally, with competitive performance, as shown in Ranalli and Rocci (2017). For these reasons, in this work, the classification is based on the CMAP criterion.

5. Model Identifiability

The subject of this section is model identifiability. In the sequel, we investigate this aspect starting from some general thoughts about mixture models and their composite estimation, then we specialize them with the regards to the specific features of our proposal.

We start by noticing that composite likelihood estimation methods provide good estimators as long as the model is identified, although their estimators are less efficient than the maximum likelihood estimators. If the composite likelihood is rich enough to include all the information about the parameters, the model is identified Mardia et al. (2009). It follows that adopting a low-dimensional margins approach, we should make sure that the model is identified even by considering only all the m-variate marginal distributions. The main point is to see if the m-variate marginals are able to capture and identify the true cluster structure generating the data. This aspect plays a key-role. Indeed, by marginalizing we are losing information about the true cluster structure and there could exist situations where the latter is lost with $m < P$. In other words, an identified model could be not identified looking only at all the m-variate marginal distributions. To clarify this crucial point, we investigate this aspect through an illustrative example for continuous data.

The first row Figure 1 displays two different cluster structures, both generated from a trivariate homogeneous Gaussian mixture model with four components equally weighted. The only difference is given by the centroids of the clusters. The last three rows represent all the bivariate marginals generated by the cluster structures A and B on the left and on the right, respectively. These result to be the same: in other words, looking at the marginals, it is not possible to identify the true cluster structure. The same bivariate marginals correspond to two different configurations of four clusters. The true cluster structure is identified, and thus well captured with $m = 3$, and it cannot be identified with $m = 2$.

The non-identifiability is due to the perfect overlapping of the centroids of the clusters on the marginals, in addition to the same covariance matrix and mixture weights. In practice, these conditions are strict and very unlikely because they are based on several equalities of parameters. In the sequel, by relaxing these equalities over parameters in the bivariate margins, we show that

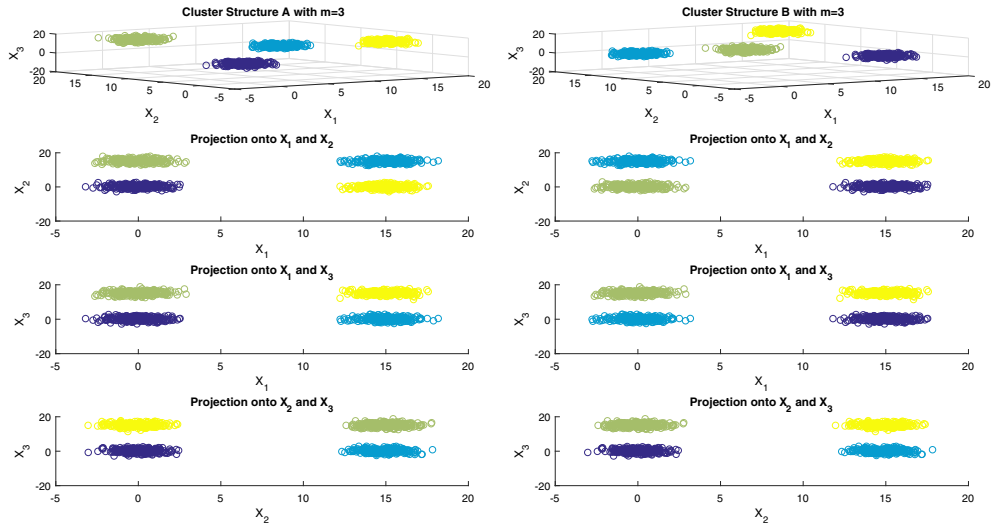

FIGURE 1.
Simulated example where true cluster structure can be captured only with $m = 3$. Two different cluster structures (*first row*) lead to the same bivariate marginals (*last three rows*).

the true cluster structure is identified looking at the m-variate marginal distributions under some sufficient conditions.

In this respect, we tried to give an answer to the identifiably in the case of continuous data and we refer to the pairwise case, $m = 2$, but what follows can be easily extended to a higher m. Before to state the sufficient conditions needed to identify Gaussian mixture models from the bivariate marginals, we should set up the background. At this aim, let us first recall formally what is a finite mixture in our context and when it is considered identified.

Let $\mathcal{F} = \left\{ f(\mathbf{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Omega}, \mathbf{x} \in \Re^P \right\}$ be the class of density functions from which mixture are to be formed. Let define the class of finite mixtures on $\mathcal{F}$ with appropriate class of density functions, $\mathcal{H}_F$, as

$$\mathcal{H}_{\mathcal{F}} = \{h(\mathbf{x}; \boldsymbol{\psi}) : h(\mathbf{x}; \boldsymbol{\psi}) = \sum_{g=1}^{G} p_g f(\mathbf{x}; \boldsymbol{\theta}_g), \sum_{g=1}^{G} p_g = 1, \forall G = 1, 2, \ldots,$$
$$p_g > 0, f(\cdot; \boldsymbol{\theta}_g) \in \mathcal{F}, \forall g = 1, \ldots, G\};$$

where $\boldsymbol{\psi} = \{p_1, \ldots, p_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G\}$.

To avoid trivial ambiguities, in the sequel, we will always assume that the components of a mixture are all distinct.

Definition. $\mathcal{H}_{\mathcal{F}}$ is said to be identified if for any two members $h(\mathbf{x}; \boldsymbol{\psi}) = \sum_{g=1}^{G} p_g f(\mathbf{x}; \boldsymbol{\theta}_g)$ and $h(\mathbf{x}; \tilde{\boldsymbol{\psi}}) = \sum_{g=1}^{\tilde{G}} \tilde{p}_g f(\mathbf{x}; \tilde{\boldsymbol{\theta}}_g)$, the equality $h(\mathbf{x}; \boldsymbol{\psi}) \equiv h(\mathbf{x}; \tilde{\boldsymbol{\psi}})$ implies $G = \tilde{G}$ and the summations can be reordered such that $p_g = \tilde{p}_g$ and $f(\mathbf{x}; \boldsymbol{\theta}_g) = f(\mathbf{x}; \tilde{\boldsymbol{\theta}}_g)$ for $g = 1, \ldots, G$.

Finally, we need to recall an additional well-known result due to Yakowitz and Spragins (1968) before discussing model identifiability under our proposal.

Theorem. (Yakowitz & Spragins, 1968). *The class $\mathcal{H}_{\mathcal{G}^{\mathcal{P}}}$, where $\mathcal{G}^{\mathcal{P}}$ is the class of P-dimensional Gaussian densities, is identified.*

Once we have defined the setting and recalled the conditions needed to have a mixture model considered identified, we can proceed further. First of all, identifying the true cluster structure

within a pairwise framework means identifying the correct number of mixture components by looking only at the bivariate marginals. So, first, we need to say when this is possible for the continuous case.

Proposition 1. *Let* $h(\boldsymbol{x};\boldsymbol{\psi}) = \sum_{g=1}^{G} p_g\phi(\boldsymbol{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ *be an element of* $\mathcal{H}_{\mathcal{GP}}$ *and* $h_{ij}(x_i,x_j;\boldsymbol{\psi}^{(ij)})$ *be its bivariate marginal density with respect to* x_i *and* x_j*, if for any* $g \neq h$ $(g,h=1,\ldots,G)$ *we have*

$$(\mu_{i,g},\mu_{j,g},\sigma_{ii,g},\sigma_{ij,g},\sigma_{jj,g}) \neq (\mu_{i,h},\mu_{j,h},\sigma_{ii,h},\sigma_{ij,h},\sigma_{jj,h}) \tag{5}$$

then $h_{ij}(x_i,x_j;\boldsymbol{\psi}^{(ij)})$ *belongs to* $\mathcal{H}_{\mathcal{G}^2}$ *and it has* G *components.*

Proof. By definition $h_{ij}(x_i,x_j;\boldsymbol{\psi}^{(ij)}) = \int h(\mathbf{x};\boldsymbol{\psi})d\mathbf{x}_{-ij} = \sum_{g=1}^{G} p_g \int \phi(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)d\mathbf{x}_{-ij} = \sum_{g=1}^{G} p_g\phi(x_i,x_j;\mu_{i,g},\mu_{j,g},\sigma_{ii,g},\sigma_{ij,g},\sigma_{jj,g})$. This shows that the marginal is always a finite mixture of at most G components. From (5), we deduce that the components of the marginal mixture are all distinct, this implies that the number of components cannot be less than G being $\mathcal{H}_{\mathcal{GP}}$ identified.

From the proof of Proposition 1 is clear that the number of clusters that you can see on the bivariate margin (x_i,x_j) is always less than or equal to the true one. The inequality is strict if (5) is false for at least one pair of components. That is, if at least two components have the same bivariate margin (x_i,x_j). □

On the basis of the previous result, we can establish some sufficient conditions to identify model parameters from the bivariate marginals.

Proposition 2. *Let* $h(\boldsymbol{x};\boldsymbol{\psi}) = \sum_{g=1}^{G} p_g\phi(\boldsymbol{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ *and* $h(\boldsymbol{x};\tilde{\boldsymbol{\psi}}) = \sum_{g=1}^{\tilde{G}} p_g\phi(\boldsymbol{x};\tilde{\boldsymbol{\mu}}_g,\tilde{\boldsymbol{\Sigma}}_g)$ *be two members of* $\mathcal{H}_{\mathcal{GP}}$*. If they are such that*

$$p_1 > p_2 > \cdots > p_G, \tilde{p}_1 > \tilde{p}_2 > \cdots > \tilde{p}_G \tag{6}$$

and for every $i \neq j = 1,2,\ldots,P$

$$h_{ij}(x_i,x_j;\boldsymbol{\psi}^{(ij)}) = h_{ij}(x_i,x_j;\tilde{\boldsymbol{\psi}}^{(ij)}) \tag{7}$$

$$(\mu_{i,g},\mu_{j,g},\sigma_{ii,g},\sigma_{ij,g},\sigma_{jj,g}) \neq (\mu_{i,h},\mu_{j,h},\sigma_{ii,h},\sigma_{ij,h},\sigma_{jj,h}),\ g \neq h = 1,2,\ldots,G \tag{8}$$

$$(\tilde{\mu}_{i,g},\tilde{\mu}_{j,g},\tilde{\sigma}_{ii,g},\tilde{\sigma}_{ij,g},\tilde{\sigma}_{jj,g}) \neq (\tilde{\mu}_{i,h},\tilde{\mu}_{j,h},\tilde{\sigma}_{ii,h},\tilde{\sigma}_{ij,h},\tilde{\sigma}_{jj,h}),\ g \neq h = 1,2,\ldots,\tilde{G}, \tag{9}$$

then $h(\boldsymbol{x};\boldsymbol{\psi}) = h(\boldsymbol{x};\tilde{\boldsymbol{\psi}})$.

Proof. Equation (8) is such that, by proposition 1, we can deduce that the bivariate marginals of $h(\mathbf{x};\boldsymbol{\psi})$ have G components. Similarly, from (9), we deduce that $\tilde{G}$ is the number of components of the bivariate marginals of $h(\mathbf{x};\tilde{\boldsymbol{\psi}})$. Equality (7) implies, by the theorem, $G = \tilde{G}$. By using inequalities (6) along with (7) and the theorem, we derive $p_g = \tilde{p}_g$ for $g = 1,2,\ldots,G$. It implies that not only the mixing weights are equal, but also the labeling order of the components is the same. By using once again (7) and the theorem, we can say that for every $i \neq j = 1,\ldots,P$ and $g = 1,\ldots,G$ we have

$$(\mu_{i,g},\mu_{j,g},\sigma_{ii,g},\sigma_{ij,g},\sigma_{jj,g}) = (\tilde{\mu}_{i,g},\tilde{\mu}_{j,g},\tilde{\sigma}_{ii,g},\tilde{\sigma}_{ij,g},\tilde{\sigma}_{jj,g})$$

i.e.,

$$(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) = (\tilde{\boldsymbol{\mu}}_g,\tilde{\boldsymbol{\Sigma}}_g)$$

and this completes the proof. □

The Proposition 2 says that if the bivariate marginals have at least one different parameter and the mixing weights are different, then the cluster structure captured through the marginals is well captured, and thus identified in the sense specified by the proposition. The conditions to identify the true cluster structure are not necessary. Thus, they do not result to be sharp and they became even less sharp as m increases. Besides the investigation of the necessary conditions, it would also be interesting to generalize Proposition 2 to different cases: without assuming (9), when (8) is true only for some bivariate marginals, and when the mixture is discretized.

Finally, it is worth pointing out some further comments raised by a referee. If one is in the near-non-identifiable situation (such as when there is much loss of information in the marginalization), then the variance of the composite likelihood estimator might be much larger than that for estimator from the full likelihood. It follows that the efficiency also depends on the mixture components being easily identified from low-dimensional margins. This means that the efficiency of the composite estimators depends not only on (5), but also on the closeness of the pairs of vectors in (5) for all (i, j) and (g, h).

In conclusions in the near-non-identifiable situation—that is the cluster structure is well identified by the joint density, but it is weakly identified by the bivariate marginals—it is recommended to use a higher m. This leads to increase the efficiency of the composite estimators and to improve the model identifiability.

However, when the mixture is assumed to be latent and discretized, as in our proposal, the previous results hold with some cautions. In fact, as always it happens in the URV approach, the discretization could mask the original features of the underlying model. In most cases, we can expect to have identifiable models.

What said above regards sufficient conditions. Now, we deal with the necessary condition needed to identify the SCR model. In this respect, we investigate the maximum number of parameters that can be estimated. As said in Ranalli and Rocci (2016a), the pairwise likelihood is obtained by the product of all bivariate marginal likelihood contributions and thus the maximum number of estimable parameters is equal to the number of non redundant parameters involved in the bivariate marginals. This equals the number of parameters of a log-linear model with only two factor interaction terms. As a consequence, given a $C_1 \times C_2 \times \cdots \times C_P$ contingency table, a necessary condition for the identifiability of a model is that the number of parameters is at most

$$\sum_{i=1}^{P}(C_i - 1) + \sum_{i=1}^{P-1}\sum_{j=i+1}^{P}(C_i - 1)(C_j - 1). \tag{10}$$

Such number can be computed as follows. Under the URV approach, the means and the variances of the first-order latent variables are fixed to 0 and 1, respectively, because they are not identified. In Ranalli and Rocci (2016a), the authors set the means and the variances of the reference component to 0 and 1, respectively. This constraint identifies uniquely the mixture components (ignoring the label switching problem), as well described in Millsap and Yun-Tein (2004). This is sufficient to estimate both thresholds and component parameters if all the observed variables have three categories at least and when groups are known. Given the particular structure of the mean vectors and covariance matrices, it is preferable to adopt an alternative, but equivalent, parametrization. This is analogous to that one used by Jöreskog and Sörbom (1996); it consists in setting the first two thresholds to 0 and 1, respectively. This means that there is a one-to-one correspondence between the two sets of parameters. If there is a binary variable, then the variance of the corresponding latent variable is set equal to 1 (while its mean should be still kept free).

Finally, we note that the model has the same rotational freedom that characterizes the classical factor analysis model. In other words, so far we have said how many parameters can be

estimated but different parametrizations still can exist. This can be shown by noting that, writing $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2]$, our model can be seen as the sum of two factor analysis models, i.e.,

$$\mathbf{y} = \mathbf{A}\tilde{\mathbf{y}} = \mathbf{A}_1\tilde{\mathbf{y}}^{Q} + \mathbf{A}_2\tilde{\mathbf{y}}^{\bar{Q}}.$$

The first one is related to the discriminating variables with loadings matrix $\mathbf{A}_1$, the other one is related to the noise variables with loading matrix $\mathbf{A}_2$. Let $\mathbf{T}_1$ and $\mathbf{T}_2$ be two $Q \times Q$ and $\bar{Q} \times \bar{Q}$ non-singular matrices, we can equivalently rewrite our model as

$$\begin{aligned}\mathbf{y} &= \mathbf{A}_1\mathbf{T}_1\mathbf{T}_1^{-1}\tilde{\mathbf{y}}^{Q} + \mathbf{A}_2\mathbf{T}_2\mathbf{T}_2^{-1}\tilde{\mathbf{y}}^{\bar{Q}} \\ &= \mathbf{A}_1^{*}\tilde{\mathbf{y}}^{*Q} + \mathbf{A}_2^{*}\tilde{\mathbf{y}}^{*\bar{Q}}.\end{aligned}$$

It follows that only the subspaces generated by the columns of $\mathbf{A}_1$ and $\mathbf{A}_2$ are identified. In order to estimate such subspaces, we impose some constraints on the model parameters, in complete analogy with what is usually done in the factor analysis model. In this way, we select a particular solution, one which is convenient to find, and leave the experimenter to apply whatever rotation he thinks desirable, as suggested by Lawley and Maxwell (1962). In particular, we require a spherical distribution for the second-order noise factors, i.e., $\mathbf{\Omega}_0 = \mathbf{I}$, and informative factors in the first cluster, i.e., $\mathbf{\Omega}_1 = \mathbf{I}$. Such constraints still allow a rotational freedom by orthonormal matrices. This can be eliminated by requiring a "lower" triangular form for the two loading matrices. For example, if $P = 5$ and $Q = 3$, $\mathbf{A}_1$ is of order 5×3 and it looks like

$$\mathbf{A}_1 = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \\ a_{51} & a_{52} & a_{53} \end{bmatrix};$$

while $\mathbf{A}_2$ is of order 5×2 and it looks like

$$\mathbf{A}_2 = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \\ a_{51} & a_{52} \end{bmatrix}.$$

In general, $\mathbf{A}_1$ and $\mathbf{A}_2$ have a lower triangular matrix in the first Q and $(P - Q)$ rows, respectively. Of course, after the estimation, the parameter matrices can be rotated to enhance the interpretation.

In conclusion, the number of parameters needed to estimate the model with Q variables carrying classification power, $\bar{Q}$ noise variables and G components is given by

$$\begin{aligned}&\underbrace{G-1}_{p_g} + \underbrace{PQ - Q(Q-1)/2}_{\mathbf{A}_1} + \underbrace{P\bar{Q} - \bar{Q}(\bar{Q}-1)/2}_{\mathbf{A}_2} \\ &+ \underbrace{(G-1)Q(Q+1)/2}_{\mathbf{\Omega}_g} + \underbrace{GQ}_{\boldsymbol{\eta}_g} + \underbrace{\bar{Q}}_{\boldsymbol{\eta}_0} + \underbrace{\sum_{i=1}^{P} C_i - 3P}_{thresholds}.\end{aligned}$$

This should be less or equal to the maximum number of parameters needed to saturate a log-linear model with two factor interaction terms in (10).

As a further investigation of the identifiability issue, we prompt to assess it empirically. First of all, it is always possible to initialize the pairwise algorithm from different starting points. If different parameter estimates lead to the same pairwise log-likelihood value, then the model is not identified. Further, at the end of the estimation step, it is possible to perform a test for the goodness of fit (comparing the empirical probabilities with the estimated ones—that are generated through a large sample with the parameter estimates carried out by the pairwise EM algorithm). If there is agreement between the empirical and the estimated probabilities, then the pairwise likelihood was enough to capture the cluster structure. On the other hand, if there is poor agreement, then the pairwise likelihood may be replaced with a three- or more-wise likelihood. As the latent class analysis captures the true cluster structure, if the local independence assumption holds in the data, similarly, the pairwise likelihood ($m = 2$) is an efficient estimation method, if the data show the true cluster structure in the bivariate margins.

6. Model Selection

In the estimation procedure, we assume that both the number of mixture components and the number of noisy factors are fixed. In practice, they are frequently unknown and thus, they must be estimated through the data. The best model is chosen by minimizing the C-BIC, introduced by Gao and Song (2010).

$$\text{C-BIC} = -2p\ell(\hat{\boldsymbol{\psi}}; \mathbf{x}) + \text{tr}\left(\hat{\mathbf{H}}^{-1}\hat{\mathbf{V}}\right)\log N. \tag{11}$$

where $\mathbf{H}$ is the sensitivity matrix, $\mathbf{H} = E(-\nabla^2 p\ell(\boldsymbol{\psi}; \mathbf{x}))$ while $\mathbf{V}$ is the variability matrix (the covariance matrix of the score vector), $\mathbf{V} = \text{Cov}(\nabla p\ell(\boldsymbol{\psi}; \mathbf{x}))$. The C-BIC contains the BIC as a particular case. In fact, when the full likelihood is used then $\mathbf{H}$ and $\mathbf{V}$ coincide and the penalty term becomes equal to the number of parameters multiplied by the logarithm of the sample size. Sample estimates of $\mathbf{H}$ and $\mathbf{V}$ are

$$\hat{\mathbf{H}} = -\frac{1}{N}\sum_{n=1}^{N}\nabla^2 p\ell(\hat{\boldsymbol{\psi}}; \mathbf{x}_n)$$

and

$$\hat{\mathbf{V}} = \frac{1}{N}\sum_{n=1}^{N}(\nabla p\ell(\hat{\boldsymbol{\psi}}; \mathbf{x}_n)\nabla p\ell(\hat{\boldsymbol{\psi}}; \mathbf{x}_n))'.$$

A simulation study testing its performance in a context of mixture models has been provided in Ranalli and Rocci (2016a, 2016b). In the current work, in order to obtain the empirical estimates of the sensitivity and variability matrices, we used the same numerical approximation technique described there.

7. Related Models

The aim of our proposal is mainly exploratory. We propose a way to reduce the data dimensionality by identifying latent factors that are able (informative) or unable (noise) to explain the clustering structure underlying the data. Such construction mainly allows us to identify the factors that explain the between variability in terms of different class conditional means, variances and

covariances; in short, it allows us to perform a simultaneous clustering and reduction (SCR). The model can also be used for variable selection and/or parsimonious modeling purposes. It follows that our model can also be compared to models/methods following one of the two aforementioned purposes.

In particular, within the first purpose, Raftery et al. (2006) formulates the problem of variable selection, for continuous data, as a model comparison problem using the BIC. Here, the variables are partitioned into two exclusive subsets representing the relevant, or discriminative, and the irrelevant, or noise, variables, respectively. Maugis et al. (2009) extend this approach, while Witten and Tibshirani (2010) propose to perform the variable selection by using a lasso-penalty. Many other authors have extended the aforementioned works or proposed different approaches but almost exclusively on continuous data. In the context of categorical data, there are only few proposals. We mention Dean and Raftery (2010) and White et al. (2014) who extend the work of Raftery et al. (2006) to the latent class model. In summary, variable selection has the main aim to select the useful variables for classification in a iterative-fashion. They are heuristic methods whose selection processes are usually computationally demanding. Furthermore, discarding the irrelevant variables, it means that only noise variables may exist—it is not assumed the existence of noisy dimensions. Our model can be used to identify noise variables; if they exist, they are identified by being closely related with the noise factors. One advantage of our approach is that, by looking at the relations between the first- and second-order factors, we can understand how much a variable is informative or not for the classification.

Within the second purpose, parsimonious modeling, the idea is to define models able to capture the clustering structure by using a reduced set of parameters. One of the earliest *parsimonious* proposal is given in Celeux and Govaert (1995), where a mixture of Gaussians is considered and made parsimonious by imposing some equality constraints between components on some elements of the spectral decompositions of the class conditional covariance matrices. Another *parsimonious* proposal is given by the mixture of factor analyzers (MFA). The MFA model differs from the factor analysis model in having different local factor models, one for each component/cluster. The MFA was originally proposed by Ghahramani and Hinton (1997) and Hinton et al. (1997) and extended to the case of mixtures of t-distribution by McLachlan et al. (2007). Later, a general framework for the MFA model was proposed by McNicholas and Murphy (2008). Furthermore, we point the reader to see also Tipping and Bishop (1999) and Bishop (1998) who considered the related model of mixtures of principal component analyzers for the same purpose. Further references may be found in chapter 8 of Mclachlan and Peel (2000) and in a recent review on model-based clustering of high-dimensional data (Bouveyron & Brunet, 2012a). As regards categorical data, we find few analogous proposals (see e.g., Gollini & Murphy, 2014, McParland et al., 2014; Marbac et al., 2014a, Marbac et al., 2014b).

All the aforementioned proposals use a FA or PCA approach to reparameterize the covariance matrices. They do not aim at identifying the informative dimensions—indeed they are just parsimonious model-based clustering techniques. Their dimensionality reduction is only local and within the components. Differently, our model can be used to cluster observations by taking into account the presence of global informative/noise dimensions. One advantage is that we can pursue both clustering and identification of informative variables (reduction) by using a parsimonious structure on the means and covariance matrices of the mixture model.

8. Simulation Study

To evaluate the empirical behavior of the proposal, a simulation study has been conducted in four different scenarios. Our model has been estimated with $m = 2$ and, for sensitivity purposes, $m = 3$. It has been compared with the unreduced model, estimated with $m = 2$ (Pairwise

TABLE 1.
True values of the latent mixture model and thresholds under different scenarios.

	Common parameters	
Means of noise variables	$\boldsymbol{\eta}_0 = [0, 0, 0]'$	
Thresholds	for each variable: [0, 1, 2, 3]	
Component weights	$p_1 = 0.3$	$p_2 = 0.7$
	High separation. Independent noise variables	
	$\boldsymbol{\eta}_1 = [-2.0, 4.0]'$	$\boldsymbol{\eta}_2 = [2.5, 0.5]'$
$\mathbf{A} = \begin{bmatrix} \sqrt{1.5} & 0 & 0 \\ 0 & \sqrt{1.5} & 0 \\ 0 & 0 & \sqrt{1.5} \end{bmatrix}$	$\boldsymbol{\Omega}_1 = \begin{bmatrix} 0.8 & 0 \\ 0 & 0.8 \end{bmatrix}$	$\boldsymbol{\Omega}_2 = \begin{bmatrix} 1.25 & 0.75 \\ 0.75 & 1.25 \end{bmatrix}$
	Low separation. Independent noise variables	
	$\boldsymbol{\eta}_1 = [-0.5, 3.5]'$	$\boldsymbol{\eta}_2 = [2.5, 0.5]'$
$\mathbf{A} = \begin{bmatrix} \sqrt{1.5} & 0 & 0 \\ 0 & \sqrt{1.5} & 0 \\ 0 & 0 & \sqrt{1.5} \end{bmatrix}$	$\boldsymbol{\Omega}_1 = \begin{bmatrix} 0.8 & 0 \\ 0 & 0.8 \end{bmatrix}$	$\boldsymbol{\Omega}_2 = \begin{bmatrix} 2.3 & 1.3 \\ 1.3 & 2.3 \end{bmatrix}$
	High Separation. Correlated noise variables	
	$\boldsymbol{\eta}_1 = [-2.0, 4.0]'$	$\boldsymbol{\eta}_2 = [2.5, 0.5]'$
$\mathbf{A} = \begin{bmatrix} 1.2247 & 0 & 0 \\ 0.6124 & 1.0607 & 0 \\ 0.6124 & 0.3536 & 1 \end{bmatrix}$	$\boldsymbol{\Omega}_1 = \begin{bmatrix} 0.8 & 0.5 \\ 0.5 & 0.8 \end{bmatrix}$	$\boldsymbol{\Omega}_2 = \begin{bmatrix} 1.25 & 0.75 \\ 0.75 & 1.25 \end{bmatrix}$

The data were generated according the structure assumed by the SCR model with $G = 2$.

Clustering, PC) as proposed by Ranalli and Rocci (2016a), to assess what is the effect of ignoring the noise dimensions. It has also been compared with the LCA, estimated with $m = P$ (full likelihood), i.e., with the standard approach to cluster categorical data. In the first scenario, we simulated 250 samples from a latent two-component Gaussian mixture. By its discretization, we generated five ordinal variables with five categories, where only two ($Q = 2$) variables carry group discrimination information, while the others are noise variables (see Table 1). In the second scenario, we explored the behavior of the methods when the number of variables increases. We simulated 250 samples from a latent three-component Gaussian mixture. By its discretization, we generated eight ordinal variables with five categories, where only three ($Q = 3$) variables carry group discrimination information, while the others are noise variables (see Table 2). Under both scenarios, we have analyzed two different experimental factors: the sample size ($N = 1000, 5000$) and the degree of separation between clusters (high, low). The matrix $\mathbf{A}$ has been chosen diagonal; it follows that there is a one-to-one correspondence between the second-order and the observed variables. Hence, $P - Q$ observed variables are noise and not correlated. This is coherent with the main LCA assumption, i.e., the local independence. In the third scenario, to assess its effect when the noise variables are correlated, we modified the structure of $\mathbf{A}$ in a non-diagonal form as shown in Tables 1 and 2, when there is a high degree of separation, that is the most favorable condition. In the fourth scenario, we assessed the model selection performance of C-BIC. We simulated 200 samples from a latent two-component Gaussian mixture model with high degree of discrimination. We generated five ordinal variables with five categories, and we assumed that there are $Q = 1, 2, 3, 4$ noise variables. Under these conditions, we have fitted different models: our proposal with $Q = 1, 2, 3, 4$ and the *standard* clustering model proposed by Ranalli and Rocci (2016a). In each case we selected the best model according to C-BIC.

Throughout the simulation study, the clustering performance has been evaluated through the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) between the true partition matrix and that

TABLE 2.
True values of the latent mixture model and thresholds under different scenarios.

	Common parameters		
Means of noise variables	$\boldsymbol{\eta}_0 = [0, 0, 0, 0, 0]'$		
Thresholds	for each noise variable:[0, 1, 2, 3]		
Component weights	$p_1 = 0.2$	$p_2 = 0.4$	$p_3 = 0.4$
	High Separation. Independent noise variables		
	$\boldsymbol{\eta}_1 = [-4.0, 5.0, 3.5]'$	$\boldsymbol{\eta}_2 = [2.0, -2.0, -1.0]'$	$\boldsymbol{\eta}_3 = [5.0, 3.0, -3.0]'$
$\mathbf{A} = \begin{bmatrix} \sqrt{1.5} & 0 & 0 & 0 & 0 \\ 0 & \sqrt{1.5} & 0 & 0 & 0 \\ 0 & 0 & \sqrt{1.5} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{1.5} & 0 \\ 0 & 0 & 0 & 0 & \sqrt{1.5} \end{bmatrix}$	$\boldsymbol{\Omega}_1 = \begin{bmatrix} 0.8 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.8 \end{bmatrix}$	$\boldsymbol{\Omega}_2 = \begin{bmatrix} 1.25 & 0.75 & 0.75 \\ 0.75 & 1.25 & 0.75 \\ 0.75 & 0.75 & 1.25 \end{bmatrix}$	$\boldsymbol{\Omega}_3 = \begin{bmatrix} 1.5 & -0.5 & -0.5 \\ -0.5 & 1.5 & -0.5 \\ -0.5 & -0.5 & 1.5 \end{bmatrix}$
Thresholds	for each of the Q variables: [0, 1, 1.5, 3.5]		
	Low separation. Independent noise variables		
	$\boldsymbol{\eta}_1 = [-2.0, 5.0, 2.5]'$	$\boldsymbol{\eta}_2 = [0.5, -1.0, -1.0]'$	$\boldsymbol{\eta}_3 = [5.0, 3.0, -3.0]'$
$\mathbf{A} = \begin{bmatrix} \sqrt{1.5} & 0 & 0 & 0 & 0 \\ 0 & \sqrt{1.5} & 0 & 0 & 0 \\ 0 & 0 & \sqrt{1.5} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{1.5} & 0 \\ 0 & 0 & 0 & 0 & \sqrt{1.5} \end{bmatrix}$	$\boldsymbol{\Omega}_1 = \begin{bmatrix} 0.8 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.8 \end{bmatrix}$	$\boldsymbol{\Omega}_2 = \begin{bmatrix} 1.25 & 0.75 & 0.75 \\ 0.75 & 1.25 & 0.75 \\ 0.75 & 0.75 & 1.25 \end{bmatrix}$	$\boldsymbol{\Omega}_3 = \begin{bmatrix} 1.5 & -0.5 & -0.5 \\ -0.5 & 1.5 & -0.5 \\ -0.5 & -0.5 & 1.5 \end{bmatrix}$
Thresholds	for each of the Q variables: [0, 1, 1.5, 3.5]		
	High separation. Correlated noise variables		
	$\boldsymbol{\eta}_1 = [-2.0, 3.5, 2.0]'$	$\boldsymbol{\eta}_2 = [2.0, -2.0, -1.0]'$	$\boldsymbol{\eta}_3 = [4.0, 2.5, -3.0]'$
$\boldsymbol{\Omega}_0 = \begin{bmatrix} 1.2247 & 0 & 0 & 0 & 0 \\ 0.5307 & 1.1038 & 0 & 0 & 0 \\ 0.5307 & 0.3337 & 1.0521 & 0 & 0 \\ 0.5307 & 0.3337 & 0.2442 & 0.7398 & 0 \\ 0.5307 & 0.3337 & 0.2442 & 0.2667 & 0.9880 \end{bmatrix}$	$\boldsymbol{\Omega}_1 = \begin{bmatrix} 0.8 & 0.5 & 0.5 \\ 0.5 & 0.8 & 0.5 \\ 0.5 & 0.5 & 0.8 \end{bmatrix}$	$\boldsymbol{\Omega}_2 = \begin{bmatrix} 1.25 & 0.75 & 0.75 \\ 0.75 & 1.25 & 0.75 \\ 0.75 & 0.75 & 1.25 \end{bmatrix}$	$\boldsymbol{\Omega}_3 = \begin{bmatrix} 1.5 & -0.5 & -0.5 \\ -0.5 & 1.5 & -0.5 \\ -0.5 & -0.5 & 1.5 \end{bmatrix}$
Thresholds	for each of the Q variables: [0, 1, 2, 4]		

The data were generated according the structure assumed by the SCR model with $G = 3$.

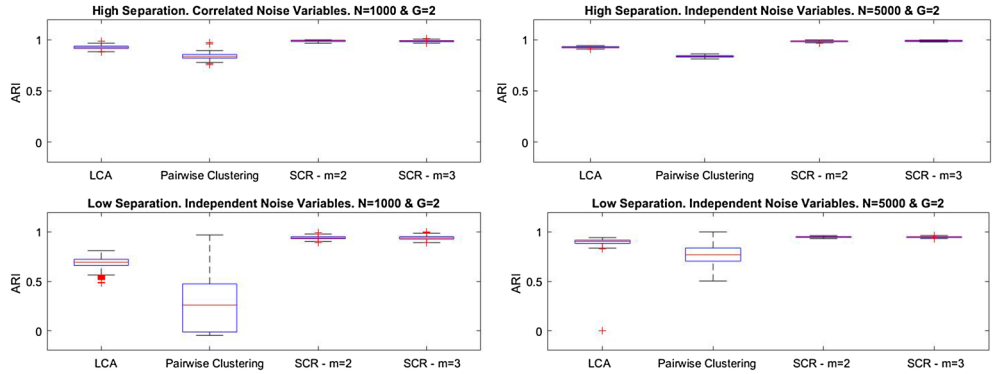

FIGURE 2.
Box-plots of ARI for the posterior probabilities. Data generated from a two-component latent mixture; 5 ordinal variables with 5 categories; 3 of them are noise variables. $N = 1000, 5000$. High/Low separation. Independent noise variables.

estimated. The index has expected value zero for independent clusterings and maximum value 1 for identical clusterings. Finally, for the SCR model and the latent mixture model for ordinal data, the parameter estimates were carried out through a pairwise or triple-wise EM algorithm, that has been initialized using *rational* starting points. In detail, we first fitted a Gaussian mixture model, treating the ranks as continuous. Then, we rearranged its output to initialize the pairwise or three-wise EM algorithm. The algorithms were stopped when the increase in the asymptotic estimated log-likelihood between two consecutive steps was less than 10^{-2}. As regards the latent class analysis, we used the R package poLCA (Linzer & Lewis, 2011) on the same simulated samples. Due to the structure of the function design, we could not use the partition of a Gaussian mixture model as initialization of the algorithm. Thus, to initialize the algorithm, 20 different starting points have been considered.

8.1. *First Scenario:* $P = 5$ *and* $G = 2$

All simulation results are reported in "appendix". Figure 2 shows the distributions of the adjusted rand index in the four conditions of the first scenario. On the left side, the sample size is equal to 1000, while on the right one is equal to 5000; in the first row, the Q variables have a different degree of discrimination, varying from high to low. The composite estimators, estimated with $m = 2$ and $m = 3$, show consistency: as N increases we obtain better classification performance. Furthermore, the clustering performance becomes poorer as the Q variables have less classification power. Comparing the four fitted models, we observe that SCRs estimated with $m = 2$ or $m = 3$ behave similarly. Both outperform PC and LCA in all conditions. The gap in performance depends on the specific condition, as expected it seems to increase when the degree of discriminative power is low.

8.2. *Second Scenario:* $P = 8$ *and* $G = 3$

Figure 3 shows the distributions of the adjusted rand index in the four conditions of the second scenario. On the left side, the sample size is equal to 1000, while on the right one is equal to 5000; in the first row the Q variables have a different degree of discrimination, varying from high to low. In this case we have three groups and a higher number of variables. The considerations made in the previous scenario are still valid. In summary, margin order choice does not affect significantly the results when the model parameters are identified. The composite estimators, estimated with $m = 2$ and $m = 3$, show consistency and our proposal outperform the alternative clustering models in

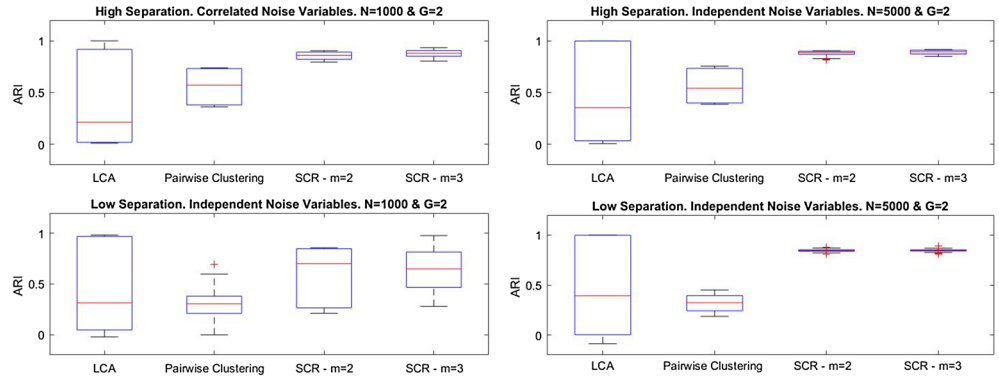

FIGURE 3.
Box-plots of ARI for the posterior probabilities. Data generated from a three-component latent mixture; 8 ordinal variables with 5 categories; 5 of them are noise variables. $N = 1000, 5000$. High/Low separation. Independent noise variables.

all conditions. The gap depends on the specific condition design. When we are departing from local independence assumption (in the first component), the difference in clustering performances between SCR and LCA is more evident. However, in all conditions, the models aimed at capturing the clustering structure show poorer performances compare to scenarios with $G = 2$. Possible explanation of the behavior of LCA is the following one. When $G = 3$, according to the simulation design, the local independence assumption does not hold in two out of three clusters. On the other hand, PC still accounts for the dependencies within each cluster, although it does not detect the presence of noisy variables/dimensions.

8.3. Third Scenario: Correlated Noise Variables

In this third scenario, 250 samples were generated from a two- and three-component Gaussian mixture with high degree of separation. In the first case, there were five ordinal variables with five categories where three of them were noise variables. In the second case, there were eight ordinal variables with five categories, where five of them were noise variables. In both cases the noise variables were correlated; in other words, the matrix $\mathbf{A}$ was a non-diagonal matrix. Figure 4 shows the distribution of ARI in the four conditions. On the left side the sample size is equal to 1000, while on the right one is equal to 5000; the first row refers to the two-component mixture model ($G = 2$), while the second row refers to the three-component mixture model ($G = 3$). Besides the considerations made in the previous scenarios, it is interesting to investigate the effect of the correlation between noise variables. LCA assumes the local independence of the variables; comparing this scenario with the previous ones, we notice a worsening in the clustering performances of LCA, while all the other methods have similar performances. Indeed, PC accounts for dependencies within each cluster and dependencies between clusters, although is does not detect the presence of noise variables/dimensions, while SCR accounts for within and between dependencies and presence of noise variables/dimensions. In conclusion, the assumption of correlated noise variables has a significant effect only on the LCA's clustering performances (Figure 5).

8.4. Fourth Scenario: Model Selection

This last scenario aims at assessing the model selection power of C-BIC. We simulated 200 samples from a latent two-component Gaussian mixture with high degree of separation and $N = 1000$. We generated five ordinal variables with five categories, and we assumed that there

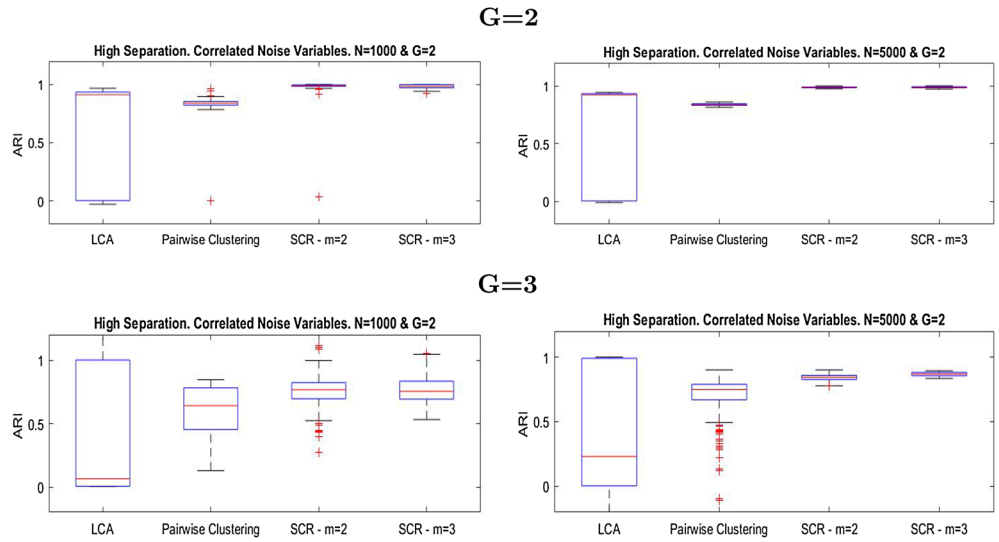

FIGURE 4.
Box-plots of ARI for the posterior probabilities. 250 samples generated from $G = 2$ and $G = 3$, with correlated noise variables, and $N = 1000, 5000$. $G = 2$: two-component latent mixture, 5 ordinal variables with 5 categories, 3 of them are noise variables. $G = 3$: three-component latent mixture, 8 ordinal variables with 5 categories, 5 of them are noise variables.

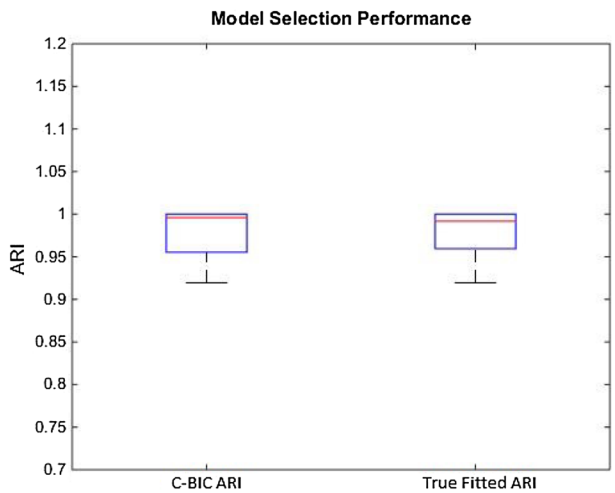

FIGURE 5.
Box-plots of ARI for the best model chosen through C-BIC compared to the ARI of the true model. Data generated from a two-component latent mixture; 5 ordinal variables with 5 categories; three of them are noise variables. High degree of discriminative power. $N = 1000$. 50 samples have been generated with $Q = 1, 2, 3, 4$. For each of the 200 samples 5 different models have been fitted.

are $Q = 1, 2, 3, 4$ noise variables. Under these conditions, we have fitted different models: our proposal with $Q = 1, 2, 3, 4$ estimated with $m = 2$ and the PC model estimated with $m = 2$ proposed by Ranalli and Rocci (2016a). In each case we selected the best model according C-BIC and we computed the corresponding ARI.

The main summary statistics of true fitted ARI, that is the ARI corresponding to the true fitted model (i.e., we fit the model with the true Q), are very close to those of the ARI obtained by

TABLE 3.
Simulation results: ARI for the best model chosen through C-BIC compared to the ARI of the true model.

	Mean	St.Dev	$q = 0.025$	$q = 0.25$	$q = 0.5$	$q = 0.75$	$q = 0.975$
C-BIC ARI	0.9674	0.0796	0.9431	0.9694	0.9959	1.0000	1.0000
True Fitted ARI	0.9797	0.0247	0.9488	0.9796	0.9918	1.0000	1.0000

Data generated from a two-component latent mixture; 5 ordinal variables with 5 categories; three of them are noise variables. High degree of separation and independent noise variables. $N = 1000$. 50 samples have been generated with $Q = 1, 2, 3, 4$. For each of the 200 samples, 5 different models have been fitted.

selecting the best model according to C-BIC over $Q = 1, 2, 3, 4$. For example, the distribution of the true fitted ARI has mean 0.98, while the distribution of the ARI chosen through C-BIC has mean 0.97. This means that C-BIC selects the right model in most cases; thus, the true cluster structure is recovered satisfactorily (Table 3).

9. Application to Real Data

In this section, our proposal is applied to two different real datasets, that are well known in the literature. The main aim of these applications is to show the effectiveness of the proposal in identifying potential noise dimensions/variables. As said throughout the paper, the proposal consists of two main contributions: methodologically we have introduced a model to classify and reduce ordinal data simultaneously, and computationally we have suggested the use of a composite likelihood based on low-dimensional margins as an efficient estimation method. In the simulation study we have investigated both aspects. In the applications to real data, we mainly focus on the first one. Furthermore, we conduct a sensitivity analysis to investigate the model identifiability more in depth; in other words, to assess whether $m = 2$ is enough to capture the cluster structure. Indeed, if the model is weakly identified, by increasing m, the estimates may change. At this aim, we compared the output obtained with $m = 2$ with the full likelihood. As said throughout the paper, it is not always possible to estimate the model with the full likelihood. If the number of variables is too large for full likelihood, we advice to conduct sensitivity analyses with trivariate margins or four-variate margins, i.e., $m = 3$ or $m = 4$.

9.1. ISSP Dataset

In this section, we apply the model to a multi-way table taken from the International Social Survey Programme (ISSP) on environment in 1993. It is available in the R package ca (Nenadic & Greenacre, 2007), and it has been previously analyzed by Greenacre (2007). There are seven variables, three of which are demographic (gender, age and education). These were not included in the analysis. The other variables correspond to four questions reported in Table 4. The possible answers to each question are: (1) strongly agree, (2) somewhat agree, (3) neither agree nor disagree, (4) somewhat disagree, (5) strongly disagree.

We initialized the pairwise EM algorithm considering 100 different random starting points. We run 16 different scenarios, considering $G = 1, 2, 3, 4$ and $Q = 1, 2, 3, 4$ estimated with $m = 2$ (pairwise likelihood) and $m = 4$ (full likelihood). In both cases, the best fitted model is given by the combination $Q = 3$ and $G = 2$. It is the model minimizing the C-BIC (when the model is estimated with $m = 2$) and BIC (when the model is estimated with $m = 4$) (Table 5).

For measuring the agreement between the classifications provided by $m = 2$ and $m = 4$, ARI is used. Its value is equal to 0.7815. The two best fitted models lead to the same conclusions; for this reason, we only report those obtained through the pairwise likelihood.

TABLE 4.
The ISSP survey data set.

Questions
X_1: We believe too often in science, and not enough in feelings and faith
X_2: Overall, modern science does more harm than good
X_3: Any change humans cause in nature, no matter how scientific, is likely to make things worse
X_4: Modern science will solve our environmental problems with little change to our way of life

TABLE 5.
Model choice according to C-BIC and BIC for pairwise likelihood approach ($m = 2$) and full likelihood approach ($m = 4$), respectively.

	C-BIC: $m = 2$				BIC: $m = 4$			
	$G = 1$	$G = 2$	$G = 3$	$G = 4$	$G = 1$	$G = 2$	$G = 3$	$G = 4$
$Q = 1$	37228	38003	34875	34887	35735	31070	31107	31130
$Q = 2$	34850	39094	35223	34925	33815	31069	31076	31167
$Q = 3$	37111	**33267**	34925	34968	37450	**31036**	31054	31161
$Q = 4$	38957	38468	38092	40876	32569	31172	31246	31287

Bold values indicate the best fitted model.

The component weights for the groups are equal to 0.53 and 0.47, respectively. The matrix **A** is given by

$$\mathbf{A} = \begin{bmatrix} 0.3256 & 0 & 0 & -0.7696 \\ -0.2942 & 1.1804 & 0 & -0.2975 \\ -0.1449 & 0.2120 & 0.9410 & -0.1779 \\ 0.6821 & -0.1544 & -0.3561 & 0.4231 \end{bmatrix},$$

while the correlation between the first- and second-order variables (by rows and by columns, respectively) is

$$\begin{array}{c} \\ \mathbf{y} \end{array} \begin{array}{c} \tilde{\mathbf{y}} \\ \left[\begin{array}{ccc|c} 0.9258 & 0.5422 & 0.2135 & 0.4377 \\ 0.4350 & 0.9421 & 0.5070 & 0.4466 \\ 0.3094 & 0.4952 & 0.8756 & 0.5150 \\ \hline -0.2758 & -0.2592 & -0.4968 & 0.8082 \end{array}\right] \end{array}.$$

It follows that the noisy dimension is mainly given by the fourth question.

These conclusions agree with some empirical evidences; indeed, we computed some association measures between all the possible pairs of variables, summarized in the following table. Indeed, if there exist noise variables, then we should have empirical evidence of independence between those variables and the remaining ones (Table 6).

It follows that the association between the variables $\{X_1, X_2, X_3\}$ and X_4 is very weak: the association between X_1 and X_2 is stronger than that between X_1 & X_3, or X_2 & X_3. We may conclude that the identification of noise dimensions is empirically justified.

TABLE 6.
Empirical Evidence on the presence of a noise dimension between X_1, X_2, X_3 and X_4 by pairs.

Bivariate marginals	Polychoric corr.	ϕ-coefficient	Cramer's V	Goodman–Kruskal γ (s.e.) [CI 95%]
X_1 & X_2	0.421	0.488	0.244	0.402 (0.0304) [0.335, 0.470]
X_1 & X_3	0.400	0.441	0.22	0.400 (0.035) [0.330, 0.467]
X_2 & X_3	0.488	0.545	0.273	0.474 (0.032) [0.411, 0.537]
X_1 & X_4	0.034	0.229	0.115	0.039 (0.039) [−0.037, 0.116]
X_2 & X_4	0.005	0.291	0.146	0.026 (0.040) [−0.052 0.105]
X_3 & X_4	0.072	0.393	0.197	−0.073 (0.041) [−0.154 0.007]

TABLE 7.
Three-way cross-classification of U.S. sample according to their reported happiness, years of schooling and number of siblings.

Year of school completed	Number of siblings				
	0–1	2–3	4–5	6–7	8+
Not too happy					
<12	15	34	36	22	61
12	31	60	46	25	26
13–16	35	45	30	13	8
17+	18	14	3	3	4
Pretty happy					
<12	17	53	70	67	79
12	60	96	45	40	31
13–16	63	74	39	24	7
17+	15	15	9	2	1
Very happy					
<12	7	20	23	16	36
12	5	12	11	12	7
13–16	5	10	4	4	3
17+	1	2	9	0	1

Looking at the posterior probabilities of the response patterns (although they are not included in the paper), it seems that the two groups cluster individuals based on the score assigned to the questions—high score on the first two questions (bad feeling toward science) and low score on the third one. As a consequence, the two groups can be interpreted as degree of belief in science (or faith, conversely)—strong vs. weak.

9.2. General Social Survey Dataset

To illustrate how the model can be used we apply it to a set of data taken from the General Social Survey and displayed in Table 7. This is a well-known dataset in educational field, analyzed by Goodman and Clogg (1984) and re-analyzed recently by Giordan and Diana (2011) and Ranalli and Rocci (2016a). It is a three-way cross-classification table of 1517 people on three ordinal variables: $X_1 =$ years of completed schooling (4 categories), $X_2 =$ number of siblings (5 categories) and $X_3 =$ happiness (3 categories).

TABLE 8.
Model choice according to C-BIC and BIC for pairwise likelihood approach ($m = 2$) and full likelihood approach ($m = 3$), respectively.

	C-BIC: $m = 2$			BIC: $m = 3$		
	$G = 1$	$G = 2$	$G = 3$	$G = 1$	$G = 2$	$G = 3$
$Q = 1$	24717	**22848**	22890	12950	**12610**	12770
$Q = 2$	23151	22881	22891	13193	12634	12831
$Q = 3$	22937	22896	22972	12294	12367	12463

Bold values indicate the best fitted model.

We initialized the pairwise EM algorithm considering 100 different random starting points. We run 9 different scenarios varying both the number of clusters $G = 1, 2, 3$ and the number of variables with classification power $Q = 1, 2, 3$, estimated with $m = 2$ (pairwise likelihood) and $m = 3$ (full likelihood). All models with G greater than 3 cannot be identified under a pairwise likelihood approach. In both cases, the best fitted model is given by the combination Q=1 and G=2. It is the model minimizing the C-BIC (when the model is estimated with $m = 2$) and BIC (when the model is estimated with $m = 3$) (Table 8).
Once again, ARI is used to measure the agreement between the classifications provided by $m = 2$ and $m = 3$. Its value is equal to 0.8603. The two best fitted models lead to the same conclusions; for this reason, we only report those obtained with $m = 2$.

The component weights equal to 0.29 and 0.71, respectively. There is a clear classification between the two groups as the number of completed years of schooling increases. Moreover, it is interesting to note that years of completed schooling is the only variable with discriminative power, and, as expected, the posterior probabilities do not change substantially over the levels of happiness or the number of siblings.
The matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} 1.0542 & 0.0682 & 0 \\ -0.4806 & 1.0406 & -0.2546 \\ -0.0944 & 0.1320 & 0.6111 \end{bmatrix},$$

while the correlation between the first- and second-order variables (by rows and by columns, respectively) is

$$\begin{array}{c} \\ \mathbf{y} \end{array} \begin{array}{c} \tilde{\mathbf{y}} \\ \left[\begin{array}{c|cc} 0.9987 & 0.0509 & 0.0000 \\ \hline -0.4951 & 0.8439 & -0.2065 \\ -0.1884 & 0.2074 & 0.9600 \end{array}\right] \end{array}.$$

This leads to some straightforward conclusions: to detect the noisy variables/dimensions, we should look at the highest correlation on the last two columns ($\tilde{y}_2, \tilde{y}_3$). The most correlated variables are y_2 and y_3 with correlations equal to 0.84 and 0.96, respectively.

If there exists noise variables, then we have empirical evidence of independence between happiness and years of schooling, and happiness and number of siblings. However, in this real example, it is better to speak about noisy dimensions. Looking at the correlation between the first- and second-order variables, we notice that also the second variable contributes to the classification (its correlation with $\tilde{y}_1$ is equal to -0.50, it means that 25% of its variability is explained by the

TABLE 9.
Empirical Evidence on the presence of noise dimensions between years of schooling (X_1), number of siblings (X_2) and happiness (X_3) by pairs.

Bivariate marginals	Polychoric corr.	ϕ-coefficient	Cramer's V	Goodman–Kruskal γ (s.e.) [CI 95%]
X_1 & X_2	−0.425	0.394	0.227	−0.425 (0.025) [−0.474, −0.377]
X_1 & X_3	−0.161	0.165	0.116	−0.169 (0.036) [−0.24, −0.099]
X_2 & X_3	0.073	0.13	0.092	0.07 (0.033) [0.006, 0.135]

TABLE 10.
Model choice according to C-BIC using the pairwise likelihood approach (m 2) when an extra noisy ordinal variable is included into the original dataset.

	$G = 1$	$G = 2$	$G = 3$
$Q = 1$	44407	**44133**	44151
$Q = 2$	44719	44182	44166
$Q = 3$	44423	44162	44186
$Q = 4$	44809	44219	44313

Pairwise likelihood approach.
Bold value indicates the best fitted model.

informative dimension). These conclusions agree with some empirical evidences; indeed, we computed some association measures between all the possible pairs of variables, summarized in the following table (Table 9). There is still association between X_1 and X_2 (the polychoric correlation is −0.43), but the most discriminative variable is X_1.

It follows that the association between X_1 and X_2 is stronger than that between X_1 and X_3, or X_2 and X_3. Thus, the fact that our proposal identifies noise dimensions is empirically justified.

Furthermore, in order to test the right behavior of our proposal, in the original dataset, we included a noisy ordinal variable with three categories obtained by thresholding a standard normal variable (Table 10).

As expected, the best fitted model is that one minimizing C-BIC, that is the model with $G = 2$ and $Q = 1$ with a C-BIC value of 44133.

The correlation between the first- and second-order variables (by rows and by columns, respectively) is

$$\begin{array}{c} \\ \mathbf{y} \end{array} \begin{array}{c} \tilde{\mathbf{y}} \\ \left[\begin{array}{c|ccc} 0.9986 & -0.1726 & -0.0259 & -0.0096 \\ \hline -0.4800 & 0.9286 & 0.0286 & 0.0019 \\ -0.2157 & 0.0349 & 0.9816 & 0.0038 \\ -0.0580 & -0.0119 & 0.0260 & 0.9985 \end{array}\right] \end{array}.$$

This leads to some straightforward conclusions: to detect the noise variable we should look at the highest correlation on the last three columns ($\tilde{y}_2$, $\tilde{y}_3$, $\tilde{y}_4$). The most correlated variables are y_2, y_3, y_4 with correlations equal to 0.93, 0.98 and 1, respectively.

Finally, as a side note, we recall that the M-step is carried out through an optimization routine in Matlab. It requires some user-specified input—such as the number of functions evaluated or the number of iterations—that influence both the speed of convergence and the quality of estimates.

It seems that a constrained optimization requires a lower number of functions evaluated. In order to compare the models, the same number of functions has been considered for the competitive models. A further investigation on how to chose the number of functions evaluated taking into account both the sample size and the complexity of the problem (such as the number of variables) is needed.

10. Discussion and Concluding Remarks

In this paper, we proposed a model that reduces the data dimensionality by identifying latent factors that are able (informative) or unable (noise) to explain the clustering structure underlying the data. This allows to identify the factors that explain the between variability in terms of different class conditional means, variances and covariances; in other words, it allows us to perform a simultaneous clustering and reduction (SCR). However, the model can also be used for variable selection and/or parsimonious modeling purposes. If there are noise factors that are highly correlated with some of the observed variables, the model can be used for a variable selection purpose. If there is no noise factor, but only noisy dimensions, it reduces to a more parsimonious model-based clustering for ordinal data (compared to the existing proposals in the literature). Whatever the purpose is (apart from the independence case), the full likelihood always involves multidimensional integrals that cannot be computed in a closed form. For this reason, the parameter estimation is carried out through the maximization of an easier surrogate function, that is based on a low-dimensional margins ($m = 2$ or higher). After exploring the effectiveness of the proposal through a large-scale simulation study, applications to real datasets have been analyzed. To validate the proposal, a further experiment was conducted: an ordinal noise variable was added to the original General Social Survey dataset. In all cases, the best fitted model has been chosen by minimizing the information criterion C-BIC.

Even if the proposal seems to be promising, there are some open issues.

In this paper, we focused on the URV approach. Although the development of the IRT approach is different, it may result in a higher computational efficiency; but, it is still not clear how to take into account the noise variables/dimensions. However, our model can be reparametrized (Takane & Leeuw, 1987) into the IRT approach with probit link.

In addition, model identifiability needs further investigation, especially when the model may result to be weekly identified. In this paper, to study it theoretically, we started with the continuous case by stating some sufficient conditions needed to identify the true cluster structure within a pairwise framework. However, when the continuous mixture is assumed to be latent and discretized, the discretization could mask the original features of the underlying model. But, it is always possible to assess model identifiability empirically, as said at the end of Section 5.

Acknowledgments

We would like to thank the three reviewers for their helpful comments and suggestions. We believe that the comments have identified important areas which required improvement.

Appendix

See Tables 11 and 12.

TABLE 11.
Simulation results: ARI for the posterior probabilities.

	Mean	St.Dev	$q = 0.025$	$q = 0.25$	$q = 0.5$	$q = 0.75$	$q = 0.975$
High degree of separation. Independent noise variables							
$N = 1000$							
LCA	0.9271	0.0165	0.9117	0.9204	0.9281	0.9353	0.9401
Pairwise C	0.8390	0.0266	0.8135	0.8278	0.8378	0.8501	0.8615
SCR: $m = 2$	0.9895	0.0089	0.9795	0.9869	0.9918	0.9959	1.0000
SCR: $m = 3$	0.9875	0.0081	0.9801	0.9841	0.9872	0.9909	0.9955
$N = 5000$							
LCA	0.9281	0.0071	0.9210	0.9246	0.9283	0.9316	0.9349
Pairwise C	0.8387	0.0105	0.8285	0.8351	0.8382	0.8430	0.8490
SCR: $m = 2$	0.9855	0.0062	0.9796	0.9828	0.9849	0.9877	0.9918
SCR: $m = 3$	0.9893	0.0062	0.9828	0.9845	0.9886	0.9936	0.9967
Low degree of separation. Independent noise variables							
$N = 1000$							
LCA	0.6803	0.0632	0.6282	0.6722	0.6929	0.7139	0.7343
Pairwise C	0.2868	0.2857	−0.0465	0.0747	0.2606	0.4190	0.6057
SCR: $m = 2$	0.9410	0.0156	0.9275	0.9343	0.9403	0.9476	0.9550
SCR: $m = 3$	0.9403	0.0175	0.9227	0.9332	0.9393	0.9463	0.9564
$N = 5000$							
LCA	0.8963	0.0614	0.8758	0.8949	0.9039	0.9133	0.9215
Pairwise C	0.7703	0.0976	0.6757	0.7280	0.7675	0.8105	0.8718
SCR: $m = 2$	0.9482	0.0056	0.9432	0.9457	0.9484	0.9506	0.9537
SCR: $m = 3$	0.9475	0.0059	0.9412	0.9453	0.9478	0.9500	0.9534
High degree of separation. Correlated noise variables							
$N = 1000$							
LCA	0.5615	0.4545	−0.0036	0.0130	0.9125	0.9275	0.9397
Pairwise C	0.8368	0.0589	0.8173	0.8289	0.8396	0.8488	0.8631
SCR: $m = 2$	0.9866	0.0613	0.9834	0.9877	0.9918	0.9959	1.0000
SCR: $m = 3$	0.9825	0.0161	0.9654	0.9757	0.9844	0.9943	1.0000
$N = 5000$							
LCA	0.5951	0.4478	−0.0004	0.0137	0.9223	0.9299	0.9341
Pairwise C	0.8379	0.0090	0.8295	0.8327	0.8384	0.8431	0.8464
SCR: $m = 2$	0.9876	0.0064	0.9820	0.9837	0.9869	0.9894	0.9949
SCR: $m = 3$	0.9879	0.0064	0.9818	0.9850	0.9883	0.9909	0.9946

Data generated from a two-component latent mixture; 5 ordinal variables with 5 categories; three of them are noisy variables. High/low separation. Independent/correlated noise variables. $N = 1000$, 5000 and $R = 250$ samples.

TABLE 12.
Simulation results: ARI for the posterior probabilities.

	Adjusted Rand Index						
	Mean	St.Dev	$q = 0.025$	$q = 0.25$	$q = 0.5$	$q = 0.75$	$q = 0.975$
High degree of separation. Independent noise variables							
$N = 1000$							
LCA	0.4109	0.4140	0.0123	0.0333	0.2132	0.6877	0.9865
Pairwise C	0.5578	0.1571	0.3671	0.4064	0.5727	0.7090	0.7360
SCR: $m = 2$	0.8573	0.0365	0.8091	0.8377	0.8618	0.8839	0.8978
SCR: $m = 3$	0.8772	0.0340	0.8391	0.8595	0.8796	0.8985	0.9153
$N = 5000$							
LCA	0.4585	0.3934	0.0252	0.2592	0.3538	0.5750	1.0000
Pairwise C	0.5647	0.1535	0.3892	0.4190	0.5419	0.7067	0.7510
SCR: $m = 2$	0.8820	0.0229	0.8570	0.8795	0.8888	0.8981	0.9022
SCR: $m = 3$	0.8911	0.0206	0.8666	0.8808	0.8958	0.9066	0.9127
Low degree of separation. Independent noise variables							
$N = 1000$							
LCA	0.3133	0.4469	0.0140	0.0236	0.0317	0.0463	0.9982
Pairwise C	0.2995	0.1250	0.1806	0.2464	0.3052	0.3604	0.4172
SCR: $m = 2$	0.5826	0.2694	0.2177	0.3675	0.6998	0.8278	0.8560
SCR: $m = 3$	0.6415	0.1942	0.4095	0.5229	0.6494	0.7573	0.8689
$N = 5000$							
LCA	0.4622	0.4538	0.0031	0.0047	0.3928	0.8800	0.9998
Pairwise C	0.3181	0.0807	0.2242	0.2620	0.3230	0.3772	0.4118
SCR: $m = 2$	0.8451	0.0103	0.8352	0.8405	0.8448	0.8500	0.8548
SCR: $m = 3$	0.8467	0.0108	0.8380	0.8444	0.8475	0.8509	0.8551
High degree of separation. Correlated noise variables							
$N = 1000$							
LCA	0.3889	0.4413	0.0032	0.0090	0.0645	0.6664	1.0000
Pairwise C	0.5906	0.2138	0.3195	0.5210	0.6406	0.7464	0.8064
SCR: $m = 2$	0.7558	0.1200	0.6662	0.7214	0.7664	0.7949	0.8482
SCR: $m = 3$	0.7646	0.1057	0.6647	0.7115	0.7537	0.7978	0.8690
$N = 5000$							
LCA	0.3842	0.4496	0.0043	0.0046	0.2314	0.5928	0.9989
Pairwise C	0.6997	0.1454	0.6107	0.7021	0.7501	0.7755	0.8008
SCR: $m = 2$	0.8175	0.1103	0.8116	0.8280	0.8441	0.8567	0.8675
SCR: $m = 3$	0.8192	0.1005	0.7209	0.7732	0.8244	0.8593	0.9202

Data generated from a three-component latent mixture; 8 ordinal variables with 5 categories; 5 of them are noisy variables. High/low separation. Independent/correlated noise variables. $N = 1000$, 5000 and $R = 250$ samples.

References

Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd ed.). Wiley Series in Probability and Statistics. Wiley.

Bishop, C. M. (1998). Latent variable models. In *Learning in graphical models*. Springer Netherlands (pp. 371–403).

Bock, D., & Moustaki, I. (2007). *Handbook of statistics on psychonometrics, chap. Item response theory in a general framework*. Amsterdam: Elsevier.

Bouveyron, C., & Brunet, C. (2012a). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, *71*, 52–78.

Bouveyron, C., & Brunet, C. (2012b). Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, *22*(1), 301–324.

Cagnone, S., & Viroli, C. (2012). A factor mixture analysis model for multivariate binary data. *Statistical Modelling*, *12*, 257–277.
Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, *28*(5), 781–793.
Dean, N., & Raftery, A. E. (2010). Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, *62*(1), 11–35.
de Leon, A. R. (2005). Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & Probability Letters*, *75*(1), 49–57.
de Leon, A. R., & Carrigre, K. C. (2007). General mixed-data model: Extension of general location and grouped continuous models. *Canadian Journal of Statistics*, *35*(4), 533–548.
Everitt, B. (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, *6*(5), 305–309.
Gao, X., & Song, P. X. K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, *105*(492), 1531–1540.
Ghahramani, Z., & Hinton, G. E. (1997). The EM algorithm for mixtures of factor analyzers. Technical Report, University of Toronto.
Giordan, M., & Diana, G. (2011). A clustering method for categorical ordinal data. *Communications in Statistics: Theory and Methods*, *40*(7), 1315–1334.
Gollini, I., & Murphy, T. B. (2014). Mixture of latent trait analyzers for model-based clustering of categorical data. *Statistics and Computing*, *24*, 569–588.
Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.
Goodman, L. A., & Clogg, C. C. (1984). *The analysis of cross-classified data having ordered categories*. Cambridge, MA: Harvard University Press.
Greenacre, M. (2007). *Correspondence analysis in practice*. London: CRC Press.
Hinton, G. E., Dayan, P., & Revow, M. (1997). Modeling the manifolds of images of handwritten digits. *IEEE Transactions on Neural Networks*, *8*(1), 65–74.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.
Hwang, H., Montréal, H., Dillon, W., & Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika*, *71*(1), 161–171.
Jöreskog, K. G. (1990). New developments in lisrel: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, *24*(4), 387–404.
Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis for ordinal variables: A comparison of three approaches. *Multivariate Behavioural Research*, *36*, 347–387.
Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software.
Katsikatsou, M., & Moustaki, I. (2016). Pairwise likelihood ratio tests and model selection criteria for structural equation models with ordinal variables. *Psychometrika*, *81*(4), 1046–1068.
Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, *56*(12), 4243–4258.
Kumar, N., & Andreou, A. G. (1998). Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, *26*(4), 283–297.
Lawley, D. N., & Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *12*(3), 209–229.
Lee, S. Y., Poon, W. Y., & Bentler, P. (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics & Probability Letters*, *9*(1), 91–97.
Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, *80*, 221–239.
Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent. *Journal of Statistical Software*, *42*(10), 1–29.
Lubke, G., & Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*, *43*(4), 592–620.
Marbac, M., Biernacki, C., & Vandewalle, V. (2014a). Model-based clustering for conditionally correlated categorical data. ArXiv preprint arXiv:1401.5684.
Marbac, M., Biernacki, C., & Vandewalle, V. (2014b). Finite mixture model of conditional dependencies modes to cluster categorical data. ArXiv preprint arXiv:1402.5103.
Mardia, K. V., Kent, J. T., Hughes, G., & Taylor, C. C. (2009). Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika*, *96*(4), 975–982.
MATLAB. (2013). *User's guide, R2013b*. MathWorks.
Maugis, C., Celeux, G., & Martin-Magniette, M. L. (2009). Variable selection for clustering with gaussian mixture models. *Biometrics*, *65*(3), 701–709.
McLachlan, G., Bean, R. W., & Ben-Tovim, J. L. (2007). Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. *Computational Statistics & Data Analysis*, *51*, 5327–5338.
Mclachlan, G., & Peel, D. (2000). *Finite mixture models* (1st ed.). Wiley Series in Probability and Statistics. Wiley.
McNicholas, P., & Murphy, T. (2008). Parsimonious gaussian mixture models. *Statistics and Computing*, *18*(3), 285–296.
McParland, D., Gormley, I., Clark, S., McCormick, T., Kabudula, C., & Collinson, M. (2014). Clustering south african households based on their asset status using latent variable models. *The Annals of Applied Statistics*, *8*(2), 747–776.
Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer Series in Statistics Series. Springer, Incorporated New York.
Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.
Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The CA package. *Journal of Statistical Software, 20*(3), 1–13. http://www.jstatsoft.org.
Raftery, A. E., Dean, N., & Graduate, N. D. I. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, *101*, 168–178.
Ranalli, M., & Rocci, R. (2016a). Mixture models for ordinal data: A pairwise likelihood approach. *Statistics and Computing*, *26*(1), 529–547.
Ranalli, M., & Rocci, R. (2016b). Standard and novel model selection criteria in the pairwise likelihood estimation of a mixture model for ordinal data. In A. F. X. Wilhelm & H. A. Kestler (Eds.), *Studies in classification, data analysis, and knowledge organization. Analysis of large and complex data* (pp. 53–68).
Ranalli, M., & Rocci, R. (2017). Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis*, *110*, 87–102.
Rocci, R., Gattone, S. A., & Vichi, M. (2011). A new dimension reduction method: Factor discriminant k-means. *Journal of Classification*, *28*(2), 210–226.
Takane, Y., & Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.
Tipping, M., & Bishop, C. (1999). Mixtures of probabilistic principal component analyzers. *Neural Computation*, *11*(2), 443–482.
Van Buuren, S., & Heiser, W. J. (1989). Clustering objects into k groups under optimal scaling of variables. *Psychometrika*, *54*(4), 699–706.
Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*(1), 1–41.
Vichi, M., & Kiers, H. A. (2001). Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, *37*(1), 49–64.
White, A., Wyse, J., & Murphy, T. B. (2014). Bayesian variable selection for latent class analysis using a collapsed Gibbs sampler. ArXiv preprint arXiv:1402.6928.
Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, *105*, 490.
Yakowitz, S. J., & Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *39*(1), 209–214.
Yang, T., Browne, R. P., & McNicholas, P. D. (2014). Model based clustering of high-dimensional binary data. ArXiv preprint arXiv:1404.3174.

Manuscript Received: 9 APR 2015
Final Version Received: 22 MAY 2017
Published Online Date: 6 SEP 2017