Advances in Data Analysis and Classification (2024) 18:381–407
https://doi.org/10.1007/s11634-023-00539-5

REGULAR ARTICLE

Composite likelihood methods for parsimonious model-based clustering of mixed-type data

Monia Ranalli[1] · **Roberto Rocci**[1]

Received: 17 February 2021 / Revised: 10 March 2023 / Accepted: 13 March 2023 /
Published online: 9 April 2023
© The Author(s) 2023

Abstract
In this paper, we propose twelve parsimonious models for clustering mixed-type (ordinal and continuous) data. The dependence among the different types of variables is modeled by assuming that ordinal and continuous data follow a multivariate finite mixture of Gaussians, where the ordinal variables are a discretization of some continuous variates of the mixture. The general class of parsimonious models is based on a factor decomposition of the component-specific covariance matrices. Parameter estimation is carried out using a EM-type algorithm based on composite likelihood. The proposal is evaluated through a simulation study and an application to real data.

Keywords Mixture models · Factor analyzers · Composite Likelihood · EM algorithm · Mixed-type data

Mathematics Subject Classification 62-07 · 62H25 · 62H30

1 Introduction

In several fields, such as genetics, economics, engineering, social sciences and many others, data often present a complex structure where variables are measured on different scales: some are continuous, some others ordinal. If the goal of the analysis is to find subgroups in the population, all information should be properly used. In other words, all the variables should concur symmetrically, i.e. in the same way, to the estimation of the groups. However, the literature has been for the most part developed for continuous variables. In this framework, several clustering methods exist, mainly divided into

✉ Monia Ranalli
monia.ranalli@uniroma1.it

Roberto Rocci
roberto.rocci@uniroma1.it

1 Sapienza University of Rome, Piazzale Aldo Moro 5, Rome, Italy

distance-based, such as k-means, and model-based. Under the model-based approach, the finite Gaussian mixture models are the most commonly used (Hennig et al. 2015) for clustering continuous data. They are intensively used in many fields and with different classification purposes (e.g. unsupervised, semi-supervised and supervised). Their success is mainly due to their simplicity to be fitted and interpreted. According to a clustering point of view, they provide a coherent strategy for classifying data accounting for uncertainties through probabilities. Each mixture component can be interpreted as a sub-population, i.e. cluster. The same framework cannot be directly applied to ordinal data. The challenge to model ordinal data is mainly due to the lack of metric properties. For this reason, among practitioners, it is still common to analyze ordinal data following a naive approach whereby their nature is ignored. Ranks are treated as interval-scaled, and thus clustering techniques developed for continuous data are applied. However, the estimates are biased and the clustering structure may be wrong (see e.g. Dolan 1994; DiStefano 2002; Rhemtulla et al. 2012 in the SEM framework; see e.g. Ranalli and Rocci 2016, 2017 in the clustering framework). It follows that ordinal variables should be modeled properly. This can be achieved adopting the underlying variable approach (URV, Jöreskog 1990; Lee et al. 1990; Muthén 1984) where the ordinal variables are assumed to be generated by thresholding some latent continuous variables. This approach allows us to cluster mixed-type data (continuous and ordinal variables) satisfying two main requirements: dealing with ordinal data properly and modeling dependences between ordinal and continuous variables. Both continuous and ordinal variables follow a heteroscedastic Gaussian mixture model, by assuming that the ordinal variables are some variates of the mixture only partially observed through a discretization (see e.g. Ranalli and Rocci 2017; Everitt 1988).

Adopting mixture models for mixed-type data, two main closely related issues should be faced with when the dimensionality of the data increases: the number of parameters increases polynomially; a large number of ordinal variables makes the full maximum likelihood estimation infeasible.

To solve the first issue, the model should be more parsimonious in terms of number of parameters to estimate. At this aim, appropriate reparameterizations need to be assumed for the covariance matrices. In literature, there exists a general class of parsimonious mixture models for continuous data by imposing a factor decomposition on component-specific covariance matrices. The loadings and variances of error terms of the factor model may be constrained to be equal or unequal across mixture components (McNicholas and Murphy 2008; McLachlan et al. 2003; Ghahramani and Hinton 1996). More precisely, Ghahramani and Hinton (1996) constrains the (variance) error term equal across groups, McLachlan et al. (2003) imposes no constrains, and McNicholas and Murphy (2008) use eight models with varying constraints on the loadings and/or (variance) error terms. This means that even if the number of variables P is high, it is still possible to estimate the component-specific covariance matrices with few latent factors $K(K << P)$.

In our proposal, we define a general class of parsimonious mixture models for mixed-type data by introducing several possible parsimonious reparameterizations for the covariance matrices starting from the idea of McNicholas and Murphy (2008) formulated only for continuous data. In particular, we introduce twelve models; eight are defined constrained; four are defined semi-constrained, since they are more flexible.

In the first class of models, we consider different constraints on the loadings and/or (variance) error terms. In the second one, the latent factors in each clusters are the same but with different variances.

As regard the second issue, we note that the maximum likelihood estimation is rather complex. Indeed the presence of ordinal variables requires the computation of many high dimensional integrals, whose evaluation is computationally demanding as the number of ordinal variables increases. The problem is usually solved by substituting the likelihood function with a surrogate function. More precisely, we replace the full likelihood with the composite likelihood (Lindsay 1988), defined as the product of m-dimensional marginals or conditional events. The composite likelihood methods are flexible ways to create consistent estimators, which inherit the main desirable properties of the maximum likelihood estimators: under some regularity conditions (Molenberghs and Verbeke 2005), asymptotically unbiased and normally distributed with the variance given by the inverse of the Godambe Information (Lindsay 1988; Varin et al. 2011). Moreover, they have some varying degrees of robustness (Xu and Reid 2011), they are fully efficient and identical to the full maximum likelihood estimators in exponential families under a certain closure property (Mardia et al. 2009). In general efficiency is not easy to achieve and it is strictly linked to the design issue, but in all cases much more efficient in terms of computational complexity. In the current work, a composite likelihood approach is adopted for model estimation. The surrogate function is built as the product of all possible marginals of two ordinal and all continuous variables. However, as long as sparsity is not a problem and computations are feasible, it is possible to use a higher m, including more ordinal variables. The computation of parameter estimates is carried out through an EM-type algorithm based on the complete-data composite log-likelihood.

The remainder of the paper is organised as follows. Section 2 introduces the general model. Section 3 describes the estimation procedure and some issues about classification, model selection and identifiability. A theoretical comparison with the most related models some is presented in Sect. 4. While the results of a simulation study are presented in Sect. 5. A real data analysis is conducted in Sect. 6 and some concluding remarks are pointed out in Sect. 7. The models presented in this work have been implemented in MatLab code, which may be found online at https://github.com/moniar412/parsFMMmixdata.

2 Model

Let $\mathbf{y}^{\bar{O}} = [y_1, \ldots, y_{P-O}]$ and $\mathbf{x} = [x_{P-O+1}, \ldots, x_P]$ be $\bar{O} = P - O$ continuous variables and O ordinal variables, respectively. The associated categories for each ordinal variable are denoted by $c_i = 1, \ldots, C_i$ with $i = \bar{O} + 1, \ldots, P$.

Following the underlying response variable approach, observed variables $\mathbf{x}$ are considered as a discretization of continuous latent variables $\mathbf{y}^O = [y_{\bar{O}+1}, \ldots, y_P]$. The latent relationship between $\mathbf{x}$ and $\mathbf{y}^O$ is explained by a threshold model defined as follows,

$$\gamma^{(i)}_{c_i-1} \le y_i < \gamma^{(i)}_{c_i} \Leftrightarrow x_i = c_i,$$

Table 1 The covariance structure of latent parsimonious Gaussian mixture models with constrained or unconstrained factor loadings matrix

Model ID	$\boldsymbol{\Lambda}_g$	$\boldsymbol{\Psi}_g$	Isotropic	$\boldsymbol{\Sigma}_g$
CCC	C	C	C	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \psi\mathbf{I}_P$
CCU	C	C	U	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$
CUC	C	U	C	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \psi_g\mathbf{I}_P$
CUU	C	U	U	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g$
UCC	U	C	C	$\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi\mathbf{I}_P$
UCU	U	C	U	$\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}$
UUC	U	U	C	$\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi_g\mathbf{I}_P$
UUU	U	U	U	$\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$

Table 2 The covariance structure of latent parsimonious Gaussian mixture models with a semiconstrained factor loadings matrix

Model ID	$\boldsymbol{\Lambda}_g$	$\boldsymbol{\Psi}_g$	Isotropic	$\boldsymbol{\Sigma}_g$
(SC)CC	SC	C	C	$\boldsymbol{\Lambda}\mathbf{L}_g\boldsymbol{\Lambda}' + \psi\mathbf{I}_P$
(SC)CU	SC	C	U	$\boldsymbol{\Lambda}\mathbf{L}_g\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$
(SC)UC	SC	U	C	$\boldsymbol{\Lambda}\mathbf{L}_g\boldsymbol{\Lambda}' + \psi_g\mathbf{I}_P$
(SC)UU	SC	U	U	$\boldsymbol{\Lambda}\mathbf{L}_g\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g$

where $-\infty = \gamma_0^{(i)} < \gamma_1^{(i)} < \ldots < \gamma_{C_i-1}^{(i)} < \gamma_{C_i}^{(i)} = +\infty$ are the thresholds defining the C_i categories.

according to our proposal $\mathbf{y} = [\mathbf{y}^{\bar{O}}, \mathbf{y}^{O}]$ follows a finite mixture of factor analyzers (McNicholas and Murphy 2008; McLachlan et al. 2003; Ghahramani and Hinton 1996)

$$f(\mathbf{y}) = \sum_{g=1}^{G} p_g \phi(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g)$$

where ϕ is the multivariate normal density, $\boldsymbol{\Lambda}_g$ is the $P \times K$ matrix of factor loadings, and $\boldsymbol{\Psi}_g$ is the diagonal matrix of uniqueness. The latter can be assumed of the isotropic form $\psi_g\mathbf{I}$, leading to the probabilistic principal component analysis model (Tipping et al. 1999). Each term may be constrained to be equal or unequal across mixture components. The result of imposing, or not, such constraints generates the family of the eight parsimonious Gaussian mixture models (PGMMs), described in Table 1 introduced by McNicholas and Murphy (2008) in the context of continuous data. Each member of this family of models has a number of covariance parameters that is linear in data dimensionality. By assuming a common covariance structure, an even more parsimonious model can be used.

With respect to the proposal of McNicholas and Murphy (2008), we decided to add some extra flexibility maintaining a certain degree of parsimony. We introduce four new models, see Table 2, that are in between the first and the last four models of Table 1 in terms of flexibility. This is achieved by assuming that the matrix of factor loadings can be written in the form $\boldsymbol{\Lambda}\mathbf{L}_g$, where $\mathbf{L}_g$ is a positive definite diagonal matrix of factor saliences. The interpretation is the following one: the latent factors in each clusters are the same but with different variances recorded by the matrices $\mathbf{L}_g$. This is a particular form of factorial invariance firstly introduced by Cattell (1944) and then developed by several authors in the context of three-way analysis where the same variables are measured on the same subjects in different occasions (Carroll and Chang 1970; Harshman et al. 1970). It has been also extended, and successfully applied, to the case of multi-group factor analysis, where the same variables are observed on different groups of observations (see Stegeman and Lam 2016 and references there in).

A nice feature of the semi-constrained models in Table 2 is that, under mild conditions, the factors are unique. In other terms, it is not possible to rotate the factors as in the classical factor analysis model. This property can be shown by using the following result found by Kruskal (1977). Let us denote by k-rank($\mathbf{Z}$) the so-called k-rank of a matrix $\mathbf{Z}$. It is defined as the largest number k such that every subset of k columns of $\mathbf{Z}$ is linearly independent. Moreover, let ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$) and ($\mathbf{A}_T$, $\mathbf{B}_T$, $\mathbf{C}_T$) be two triplets of matrices with K columns such that

$$\mathbf{A}\text{diag}(\mathbf{c}_g)\mathbf{B}' = \mathbf{A}_T\text{diag}(\mathbf{c}_{Tg})\mathbf{B}'_T \tag{1}$$

with $g = 1, 2, \ldots, G$, where, diag($\mathbf{d}$) is the diagonal matrix having the elements of vector $\mathbf{d}$ on the main diagonal, $\mathbf{c}_g[\mathbf{c}_{Tg}]$ is g-th row of the $G \times K$ matrix $\mathbf{C}[\mathbf{C}_T]$. Kruskal (1977) has shown that if

$$k\text{-rank}(\mathbf{A}) + k\text{-rank}(\mathbf{B}) + k\text{-rank}(\mathbf{C}) \geq 2K + 2 \tag{2}$$

then there exists a permutation matrix $\mathbf{P}$ and three diagonal matrices $\mathbf{D}_A$, $\mathbf{D}_B$ and $\mathbf{D}_C$, for which $\mathbf{D}_A\mathbf{D}_B\mathbf{D}_C = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix, such that

$$\mathbf{A}_T = \mathbf{APD}_A, \mathbf{B}_T = \mathbf{BPD}_B, \mathbf{C}_T = \mathbf{CPD}_C. \tag{3}$$

In words, if (2) holds then the solution ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$) is unique up to scaling and a simultaneous column permutation. Although Kruskal's condition has been extended by some other authors (see Giordani et al. 2020 for an overview), what follows is based on such a condition because practitioners mainly refer to it in their applications. In our case, let us suppose that the part of the covariance matrices due to the common factors has not a unique representation and it is possible to write

$$\boldsymbol{\Lambda}\mathbf{L}_g\boldsymbol{\Lambda}' = \boldsymbol{\Lambda}_T\mathbf{L}_{Tg}\boldsymbol{\Lambda}'_T \tag{4}$$

with $g = 1, 2, \ldots, G$. Indicating with $\mathbf{M}$ the $G \times K$ matrix having the diagonal of $\mathbf{L}_g$ as the g-th row, from the Kruskal's results we deduce that if

$$2 \cdot k\text{-rank}(\boldsymbol{\Lambda}) + k\text{-rank}(\mathbf{M}) \geq 2K + 2 \tag{5}$$

then $\boldsymbol{\Lambda}[\mathbf{M}]$ differs from $\boldsymbol{\Lambda}_T[\mathbf{M}_T]$ only for the scaling and/or position of the columns. It is very important to note that the inequality (5) is satisfied if $\boldsymbol{\Lambda}$ and $\mathbf{M}$ are of full column rank, as usual in practical applications.

For a random i.i.d. sample of size N, the log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \left[\sum_{g=1}^{G} p_g \phi(\mathbf{y}_n^{\bar{O}}; \boldsymbol{\mu}_g^{\bar{O}}, \boldsymbol{\Sigma}_g^{\bar{O}\bar{O}}) \pi(\mathbf{x}_n; \boldsymbol{\mu}_{ng}^{O|\bar{O}}, \boldsymbol{\Sigma}_g^{O|\bar{O}}, \boldsymbol{\gamma}) \right]$$

where

$$\boldsymbol{\mu}_g^{\bar{O}} = \mathrm{E}[\mathbf{y}^{\bar{O}} \mid g], \, \boldsymbol{\Sigma}_g^{\bar{O}\bar{O}} = \mathrm{V}(\mathbf{y}^{\bar{O}} \mid g),$$
$$\boldsymbol{\mu}_{ng}^{O|\bar{O}} = \mathrm{E}[\mathbf{y}^{O} \mid \mathbf{y}_n^{\bar{O}}, g], \, \boldsymbol{\Sigma}_g^{O|\bar{O}} = \mathrm{V}(\mathbf{y}^{O} \mid \mathbf{y}_n^{\bar{O}}, g)$$

and

$$\pi(\mathbf{x}_n; \boldsymbol{\mu}_{ng}^{O|\bar{O}}, \boldsymbol{\Sigma}_g^{O|\bar{O}}, \boldsymbol{\gamma}) = \int_{\gamma_{c_1-1}^{(\bar{O}+1)}}^{\gamma_{c_1}^{(\bar{O}+1)}} \cdots \int_{\gamma_{c_P-1}^{(P)}}^{\gamma_{c_P}^{(P)}} \phi(\mathbf{y}^{Q}; \boldsymbol{\mu}_{ng}^{O|\bar{O}}, \Sigma_g^{O|\bar{O}}) d\mathbf{y}^{O}$$

is the probability of response pattern $\mathbf{x}_n$ in the g-th component mixture with mean and covariance matrix conditioned on the continuous variables. As said before, the covariance matrices could have different structures according to the specific parsimonious model chosen (see Tables 1 and 2). This likelihood causes non trivial computational problems due to the presence of multidimensional integrals. In the next section, we are going to solve this problem through the use of a composite likelihood.

3 Estimation

As suggested in Ranalli and Rocci (2017) and references therein, a composite likelihood approach could be adopted. It allows us to simplify the problem by replacing the full likelihood with a surrogate function based on m-dimensional marginals. It is a robust estimation method and its estimators have been proven to be consistent, asymptotically unbiased and normally distributed, under some mild regularity conditions (Lindsay 1988; Varin et al. 2011; Molenberghs and 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 based on $O(O-1)/2$ marginal distributions each of them composed of two ordinal variables and $\bar{O}$ continuous variables. The composite log-likelihood will be the sum of $O(O-1)/2$ sub-log-likelihoods, one for each marginal distribution. In formulas

$$c\ell(\boldsymbol{\theta}) = \sum_{i=1}^{O-1} \sum_{j=i+1}^{O} \sum_{n=1}^{N} \log \left[\sum_{g=1}^{G} p_g \phi(\mathbf{y}_n^{\bar{O}}; \boldsymbol{\mu}_g^{\bar{O}}, \boldsymbol{\Sigma}_g^{\bar{O}\bar{O}}) \pi(\mathbf{x}_n^{ij}; \boldsymbol{\mu}_{ng}^{ij|\bar{O}}, \boldsymbol{\Sigma}_g^{ij|\bar{O}}, \boldsymbol{\gamma}^{ij}) \right],$$

where $\pi(\mathbf{x}_n^{ij}; \boldsymbol{\mu}_{ng}^{ij|\bar{O}}, \boldsymbol{\Sigma}_g^{ij|\bar{O}}, \boldsymbol{\gamma}^{ij})$ is the conditional probability of response pattern $\mathbf{x}_n^{ij}$, i.e. the response pattern $\mathbf{x}_n$ restricted to only the variables i and j, given all the $\bar{O}$ continuous variables, i.e. $\mathbf{Y}^{\bar{O}} = \mathbf{y}_n^{\bar{O}}$; while $\boldsymbol{\gamma}^{ij}$ is the set of thresholds for variables i and j. This conditional probability is obtained by integrating the density of a bivariate normal distribution with parameters $(\boldsymbol{\mu}_{ng}^{ij|\bar{O}}, \boldsymbol{\Sigma}_g^{ij|\bar{O}})$ between the corresponding threshold parameters contained in γ^{ij}. The computation of parameter estimates is carried out using simultaneously a standard EM algorithm on each sub-likelihood having the same set of parameters. We start by writing the complete composite log-likelihood $c\ell_c(\boldsymbol{\theta})$ by introducing the group membership matrix $\mathbf{z}^{ij}$ indicating if the observation n belongs to mixture component g in the sub-likelihood corresponding to the marginal distribution of the ordinal variables i and j and all the continuous variables, as follows

$$c\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{O-1} \sum_{j=i+1}^{O} \sum_{n=1}^{N} \sum_{g=1}^{G} z_{ng}^{ij} \log \left[p_g \phi(\mathbf{y}_n^{\bar{O}}; \boldsymbol{\mu}_g^{\bar{O}}, \boldsymbol{\Sigma}_g^{\bar{O}\bar{O}}) \pi(\mathbf{x}_n^{ij}; \boldsymbol{\mu}_{ng}^{ij|\bar{O}}, \boldsymbol{\Sigma}_g^{ij|\bar{O}}, \boldsymbol{\gamma}^{ij}) \right],$$

The E-step requires the computation of the expected value of the complete-data composite log-likelihood given the current estimates of the model parameters. This is given by

$$Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(r-1)}) = \mathrm{E}_{\boldsymbol{\theta}^{(r-1)}} \left[c\ell_c(\boldsymbol{\theta}; \mathbf{y}^{\bar{O}}, \mathbf{x}, \mathbf{z} \mid \mathbf{y}^{\bar{O}}, \mathbf{x}) \right].$$

At the r-th iteration, the E-step consists of updating the group membership matrix $\mathbf{z}^{ij}$ of order $N \times G$ as

$$\hat{z}_{ng}^{ij} = \frac{p_g \phi(\mathbf{y}_n^{\bar{O}}; \boldsymbol{\mu}_g^{\bar{O}}, \boldsymbol{\Sigma}_g^{\bar{O}\bar{O}}) \pi(\mathbf{x}_n^{ij}; \boldsymbol{\mu}_{ng}^{ij|\bar{O}}, \boldsymbol{\Sigma}_g^{ij|\bar{O}}, \boldsymbol{\gamma}^{ij})}{\sum_{h=1}^{G} p_h \phi(\mathbf{y}_n^{\bar{O}}; \boldsymbol{\mu}_h^{\bar{O}}, \boldsymbol{\Sigma}_h^{\bar{O}\bar{O}}) \pi(\mathbf{x}_n^{ij}; \boldsymbol{\mu}_{nh}^{ij|\bar{O}}, \boldsymbol{\Sigma}_h^{ij|\bar{O}}, \boldsymbol{\gamma}^{ij})},$$

for $i = 1, \ldots, Q-1$, $j = i+1, \ldots, Q$. Then, given the E-step, the M-step is performed in blocks. First, at iteration r, the mixing weights are updated by averaging the group membership matrices, then the complete composite log-likelihood is maximized with respect to the other parameters. Since the parameter estimates of the mixture components do not have a closed form, we use an optimization routine to obtain all the parameter estimates (apart from $p_1, \ldots, p_G$). More precisely we use an optimization routine ("fmincon") in Matlab based on a quasi-Newton approximation (for more details see MATLAB (2013)). Any other optimization routines can be used: in any case, the complete composite log-likelihood needs to be coded, such that each block of marginals is weighted by the corresponding group membership matrix updated in the E-step. Given the parameter estimates, the E-step can be performed once again. The E and M steps are repeated until convergence is reached. We halted the estimation process and assumed convergence to the maximum when the relative difference between two consecutive composite log-likelihood values is less than 10^{-5}.

3.1 Classification

As regards the classification, each observation is assigned to the component with the maximum fit according to CMAP criterion (Ranalli and Rocci 2017). In a context of standard mixture models, the classification of the observations is usually based on the MAP criterion. This means that the observation is assigned to the component corresponding to the maximum fit. However, since the composite likelihood is constructed as the product of $O(O-1)/2$ sub-likelihoods, following the same principle, the fit of each observation is obtained by multiplying the corresponding $O(O-1)/2$ fits

$$s_{gn}(\boldsymbol{\theta}) = \prod_{i=1}^{O-1} \prod_{j=i+1}^{O} \left[p_g \phi(\mathbf{y}_n^{\bar{O}}; \boldsymbol{\mu}_g^{\bar{O}}, \boldsymbol{\Sigma}_g^{\bar{O}\bar{O}}) \pi(\mathbf{x}_n^{ij}; \boldsymbol{\mu}_{ng}^{ij|\bar{O}}, \boldsymbol{\Sigma}_g^{ij|\bar{O}}, \boldsymbol{\gamma}^{ij}) \right],$$

In order to express the fit in terms of degree of membership, the fit of each observation is normalized (i.e. it varies between 0 and 1), that is

$$\max_g \frac{s_{gn}}{\sum_{h=1}^{g} s_{gh}}.$$

3.2 Model selection

In the estimation procedure, we assume that the number of mixture components and the structure of covariance matrices are fixed. In practice, they are often unknown and thus, they have to be selected through the data. A criterion to select the best model could be the so-called composite BIC (Gao and Song 2010). However, its use requires the computation of the gradient of the contribution of each observation to the composite log-likelihood, see (Ranalli and Rocci 2017) for details. This makes its use rather cumbersome or infeasible when the dimensionality of the data increases. For this reason, in this work, the best model is chosen by selecting the one minimizing the additive BIC, that is the sum of BICs computed for each sub-likelihood. We refer to the additive BIC as aBIC. The idea is quite simple. Each BIC should obtain a minimum on the true model as well as their sum. The only problem in this reasoning is the fact that the BICs are computed by using the composite likelihood estimates instead of the full likelihood ones. However, if the sample size is large enough they should not be very different.

3.3 Identifiability

A further important point of the proposed model, that is worth to be discussed, is parameter identifiability. To estimate both thresholds and component parameters if all the ordinal variables have three categories at least, we set the first two thresholds to 0 and 1, respectively. This identification constraint allows us to identify uniquely means and variances of the latent variates of the mixture components (ignoring the label switching problem), as well described in Millsap and Yun-Tein (2004). This parameterization is

equivalent to that one used by Jöreskog and Sörbom (1996), where the means and the variances of the latent variables (of the first component, in the mixture framework) are set to 0 and 1, respectively. There is a one-to-one correspondence between the two sets of parameters. If there are binary variables, then the unique threshold should be set to zero while their variances should be set equal to 1 (while their means should be still kept free). However, this is a necessary condition, but it is not sufficient. Within a full maximum likelihood approach, it is well known that a sufficient condition for local identifiability is given by the non singularity of the information matrix; while a necessary condition is that the number of parameters must be less than or equal to the number of canonical parameters. Such conditions should be modified when model parameters are estimated by maximizing a composite likelihood. The sufficient condition should be reformulated by investigating the Godambe information matrix, that is, the analogous of the information matrix in composite likelihood estimation. However, as far as we know, such modification has not been formally investigate yet.
About the necessary condition, we note that in the composite likelihood only some marginal distributions are involved. It implies that we have to count the number of canonical parameters only by considering the ones involved in such marginals. As an example, if there are only ordinal variables the number of canonical 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. In particular, given a $C_{\bar{O}+1} \times C_{\bar{O}+2} \times \ldots \times C_O$ contingency table such number is

$$\sum_{i=\bar{O}+1}^{O} (C_i - 1) + \sum_{i=\bar{O}+1}^{O-1} \sum_{j=i+1}^{O} (C_i - 1)(C_j - 1).$$

However, heuristically, we are always able to see if a model is not identified, that is when the same maximized likelihood (or composite likelihood) is obtained with different parameter estimates.
The factorial reparameterisation of a component-specific covariance is not uniquely identified for models in Table 1. Indeed, we note that it has the same rotational freedom that characterizes the classical factor analysis model. Only the subspaces generated by the columns of $\boldsymbol{\Lambda}_g$ 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 lower triangular form in the first K rows of the loading matrix. Of course, after the estimation the parameter matrices can be rotated to enhance the interpretation. In Sect. 2, we have shown that such rotational freedom disappears for models in Table 2. However, in this case the columns of the matrix of factor loadings $\boldsymbol{\Lambda}$ can be arbitrarly rescaled by adjusting the matrices $\mathbf{L}_g$ accordingly. We remove such ambiguity by setting $\mathbf{L}_1 = \mathbf{I}$. It is important to say that rules for the identifiability of a factor analysis model (Shapiro 1985), like the so-called Ledermann bound (Ledermann 1937) for the number of factors

$$K \leq P + (1 - \sqrt{8P + 1})/2,$$

hold for the unconstrained models, while, probably, they could be relaxed for the constrained ones. Finally, we mention that in factor analysis there is the possibility that the estimate of the variance of the error term for a variable, the so-called uniqueness, is exactly zero. Such possibility is named an Heywood case and considered, in the factor analysis field, an improper solution because it corresponds to assume that one of the common factors coincides with one of the variables. In our experience, we did not encounter Heywood cases, however, if this would happen we suggest to introduce some constraints, e.g. $\boldsymbol{\Psi} = \psi_g \mathbf{I}$, or eliminating some variables causing the Heywood case (see Farooq 2022 and references there in).

4 Related models

The present proposal can be considered an extension of the work (Ranalli and Rocci 2017) where mixed type data, with ordinal and continuous variables, is used to cluster a sample of observations. The model is a finite mixture of Gaussians were some variates, the ones corresponding to the ordinal variables, are observed only trough a discretization. The parameters are estimated by maximizing a composite likelihood built on three blocks. The first is given by all the continuous variables, the second by all the bivariate marginals obtained considering pairs of ordinal variables, the third by the marginals obtained considering one ordinal variables and all the continuous ones. In this paper we refined this scheme eliminating the first two blocks and extending the third including two ordinal variables and all the continuous in each sub-likelihood. This modification allowed us to improve computational efficiency without worsening the quality of the estimates. Another improvement over (Ranalli and Rocci 2017) is the introduction of several possible parsimonious reparameterizations for the covariance matrices starting from the idea of McNicholas and Murphy (2008) formulated only for continuous data. A similar approach has been also adopted by Mcparland and Gormley (2015). It is a model based clustering procedure for data of mixed type based on latent variables. The latters, following a mixture of Gaussian distributions, generates the observed data of mixed type: continuous, ordinal, binary or nominal. It employs a parsimonious diagonal covariance structure for the latent variables, leading to six clustering models that varying in complexity. Each model can be estimated by using the package `clustMD` available in R. The main differences with our approach are that the thresholds parameters are estimated in a separate step using the single variables, even nominal variables are considered but it is essentially based on the local independence assumption, i.e. the variables are independent conditionally to the groups. As a side note, it is necessary to caution the reader on the presence of a further model in the R package `clustMD`, called BD model, although there is no theoretical explanation about the assumptions underlying the corresponding data generation process. Due to the lack of information about the model and the method/algorithm used for the parameter estimation, we decided to exclude it from the main analysis. In the following simulation study, we focus only on the first six parsimonious models included in the R package `clustMD`. However, as explicitly requested by an anonymous reviewer, we have also considered the BD model as possible competitor of our proposal in the supplementary material, although given the lack of its description, we have not been able to make reasonable comments on results.

The mixture of factor analyzers model Ghahramani and Hinton (1996) has been extended to the mixed type data by McParland et *al* (Mcparland et al. 2014, 2017). Compared to our proposal there are some differences: the authors estimate the model using a Bayesian approach; furthermore, they consider even variables measured on a nominal scale but constrain the diagonal matrix of uniqueness, $\boldsymbol{\Psi}$ in our notation, to be equal to the identity matrix in each component.

Finally, it is important to say that there are approaches where the variables does not play a symmetric role. For example, this happens in Murphy and Murphy (2020) were the model for each component is a regression and the distribution of some continuous variables is formulated conditionally to some covariates that could be categorical. A similar example is the proposal of Ingrassia et al. (2015) where the regressions are univariate and the covariates are random with a joint distribution built on the hypothesis of local independence.

5 Simulation study

In this section, we illustrate and discuss the results of a large simulation study aimed at assessing the effectiveness of the maximum composite likelihood estimator under different settings in terms of sample size, number of components and factors, number of variables and categories. The composite estimator has been also compared to the full likelihood one, in terms of precision and computational time. A further comparison has been done in terms of the Adjusted Rand index (Hubert and Arabie 1985) by considering the naive approach, i.e. our model treating all variables as they were continuous, and the six parsimonious models of `clustMD` (Mcparland and Gormley 2015). Furthermore, as explicitly requested by an anonymous reviewer, we have also considered the BD model as possible competitor of our proposal in the supplementary material. Finally, we tested the effectivness of the aBIC in finding the correct number of components.

The experiments are conducted generating the data from the $(SC)UU$ model with eight variables, four continuous and four ordinal with four categories. In some experiments the number of variables and categories has been increased to 15 (of which 10 are ordinal) and 10, respectively. For $G = 2$, we specify the mixture weights by $p_g = [0.30, 0.70]$, while the group-specific mean vector by

$$\boldsymbol{\mu}_1 = [-0.5, 0.5, 1, 1, -1, 2, -2, -1],$$
$$\boldsymbol{\mu}_2 = [1.5, 1.5, 0, 0, 1, 0, 2, 0].$$

For $G = 3$, we specify the mixture weights by $p_g = [0.25, 0.35, 0.40]$ and we added the further group-specific mean vector,

$$\boldsymbol{\mu}_3 = [-0.5, -0.5, -1, -1, 0, -2, 0, -1].$$

At last, $\boldsymbol{\Lambda}$ is randomly drawn from a uniform distribution on the interval $[-1, 1]$, the diagonal elements of $\mathbf{L}_g$ are randomly drawn from a uniform distribution on the interval $[0, 2]$, and a reasonable level of error was added by generating the diagonal elements of $\boldsymbol{\Psi}_g$ from a uniform in $[0, 1]$. The thresholds for the ordinal variables are $[0, 1, 2]$ when the categories are four. When the categories are five or ten, the thresholds for the ordinal variables are $[0, 1, 1.5, 2.5]$ and $[0, 1, 1.5, 1.833, 2.166, 2.499, 2.832, 3.165, 3.498]$, respectively.

Maximum composite likelihood estimates are computed by following the EM-like algorithm previously described. We halted the estimation process and assumed convergence to the maximum when the relative difference between two consecutive composite log-likelihood values is less than 10^{-5}. To initialize the model parameters, we worked out the output of the Gaussian mixture model where we treated all variables as they were continuous. The initial values for the thresholds have been computed as follows: for each variable, we have considered the empirical relative frequency of each category and then we have minimized the quadratic difference between this frequency and the corresponding probability of the mixture. As regards the factor loadings, starting from the specific-component covariance matrix (the output of the Gaussian mixture model) we estimate a factor analysis model. Then we rotate the obtained loading matrix in order to obtain a lower triangular form for the square sub-matrix given by the first K columns. The error variances are obtained as the difference between the main diagonal of the within covariance matrix of the Gaussian mixture model output and $\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g'$. We averaged $\boldsymbol{\Lambda}_g$ to get $\boldsymbol{\Lambda}$. Finally $\mathbf{L}_g$ is set to 1. Of course, this inizialitazion can be adapted properly to accommodate the other cases listed in Tables 1 and 2. The choice of initial values influences the speed of convergence of the algorithm and its ability to reach the global maximum. We suggest to use a rational start because, in our experience, a purely random initialization is generally extremely worse than our rational start in terms of local optima and computational time. However, our rational start does not guarantee to reach the global optimum and further studies are needed to improve it.

We analyzed the performance of an estimator by computing for each sample the Euclidean squared distance between the estimates and the true values for different set of parameters: group-specific means ($\boldsymbol{\mu}_c$, $\boldsymbol{\mu}_o$), thresholds ($\boldsymbol{\gamma}$), mixture weights ($\mathbf{p}$), factor loadings ($\boldsymbol{\Lambda}$), saliences ($\mathbf{L}$), uniqueness ($\boldsymbol{\Psi}$). These indexes serve to evaluate the accuracy of the estimators. In order to identify uniquely the sign of the columns of the loading matrices, in the simulation study, we impose the first non-zero element in each column to be positive. The performance of recovering the cluster structure is measured by the Adjusted Rand Index (ARI) (Hubert and Arabie 1985), which is a measure of agreement between the estimated and the true cluster memberships. It takes its maximum (value one) in case of perfect agreement.
in what follows, we describe the different experiments that compose our simulation study.

5.1 Efficiency of the composite estimator

In this experiment we assess the performances of the composite likelihood estimator over different scenarios obtained combining three factors: number of observations ($N = 500, 1000$), number of groups ($G = 2, 3$) and number of latent factors ($K = 2, 4$). For each of the scenarios, we generated 250 samples and compute the aforementioned measures of performance. The results are depicted in Table 3.
Overall, we note a normal behaviour of the composite estimator. The sample size N and the model complexity influence the imprecision of the estimates and the goodness of classification. In particular, when N increases the imprecision decreases while the ARI increases, i.e. the estimator performs better. On the contrary, when the model complexity increases, in terms of the number of components G and/or latent factors K, the estimator efficiency decreases.

5.2 Comparison between the composite and the full likelihood estimators

In this experiment we compare the composite likelihood with the full likelihood approach. This allows us to evaluate both the statistical efficiency (the full likelihood is more precise) and the computational efficiency (the composite likelihood is much faster). We generated 250 samples of size $N = 500$, with $G = 3$ components and $K = 4$ latent factors. Table 4 displays the results in terms of imprecision indexes and computational time.

In terms of error in parameter estimation, the full likelihood is slightly more efficient. However, it is more than 4 times in median and 11 times in mean slower than the composite likelihood. It follows that the loss in statistical efficiency of the composite likelihood is well paid in terms of computational time.
The experiment has been extended by considering on the same samples even the naive approach, where the ordinal variables are considered as continuous, and six models of the `clustMD` package. Being different models with different parameters, the performances have been compared only in terms of goodness of recovery by using the ARI. The results are depicted in Table 5.

In terms of recovering the clustering structure, the full likelihood is the best one, even if the composite likelihood shows similar results (0.80 vs. 0.79 in terms of means). The ARI for the naive case is 0.6 in mean. The six parsimonious models show ARI values inferior to the others, between 0.58 and 0.60. This is probably due to the fact they assume a local independence that in this case is false.

5.3 The effect of a higher number of categories

In this experiment we evaluate the effect of a higher number of categories for the ordinal variables. This allows us to evaluate how the error in parameter estimation changes with different number of categories. This means that some categories could have zero or very low frequencies, but this is not a problem for the algorithm. Furthermore we compare the clustering performance, in terms of ARI, of the proposal with the naive case (i.e. our model treating all variables as they were continuous) and six parsimonious models

Table 3 Quartiles, mean and standard deviation (in brackets) of the imprecision indexes for the parameter estimates, and the Adjusted Rand index (ARI) obtained by the EM-like algorithm over 250 samples generated in different settings from the model $(SC)UU$

	$G = 2$ and $K = 2$							
	$N = 500$				$N = 1000$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
$\boldsymbol{\mu}_c$	0.0026	0.0042	0.0069	0.0435 (0.4582)	0.0014	0.0023	0.0036	0.0288 (0.3674)
$\boldsymbol{\mu}_o$	0.0046	0.0071	0.0116	0.0246 (0.1927)	0.0024	0.0036	0.0058	0.0249 (0.2768)
$\boldsymbol{\gamma}$	0.0067	0.0145	0.0241	0.0911 (0.8503)	0.0038	0.0074	0.0133	0.0166 (0.0457)
$\mathbf{p}$	0.0000	0.0001	0.0005	0.0004 (0.0006)	0.0000	0.0001	0.0003	0.0003 (0.0014)
$\boldsymbol{\Lambda}$	0.3828	0.5913	0.8956	0.6318 (0.3866)	0.3915	0.5999	0.8224	0.6141 (0.3697)
$\mathbf{L}$	0.0346	0.2188	0.4843	0.3404 (0.4737)	0.0216	0.2901	0.5112	0.3883 (0.5054)
$\boldsymbol{\Psi}$	0.0054	0.0082	0.0120	0.0123 (0.0152)	0.0033	0.0048	0.0080	0.0086 (0.0158)
ARI	0.9918	1.0000	1.0000	0.9871 (0.0693)	0.9918	1.0000	1.0000	0.9887 (0.0645)
	$G = 2$ and $K = 4$							
	$N = 500$				$N = 1000$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
$\boldsymbol{\mu}_c$	0.0043	0.0068	0.0102	0.1132 (0.7209)	0.0023	0.0034	0.0055	0.0463 (0.4697)
$\boldsymbol{\mu}_o$	0.0070	0.0113	0.0165	0.0916 (0.5711)	0.0036	0.0053	0.0073	0.0207 (0.1591)
$\boldsymbol{\gamma}$	0.0073	0.0120	0.0205	0.0170 (0.0189)	0.0039	0.0069	0.0115	0.0097 (0.0162)
$\mathbf{p}$	0.0000	0.0002	0.0006	0.0010 (0.0041)	0.0000	0.0001	0.0003	0.0002 (0.0005)
$\boldsymbol{\Lambda}$	0.5009	0.6703	0.8064	0.6619 (0.2018)	0.4817	0.6273	0.7772	0.6445 (0.2197)
$\mathbf{L}$	0.2121	0.3525	0.5159	0.4484 (0.3721)	0.1985	0.3473	0.5783	0.4435 (0.3597)
$\boldsymbol{\Psi}$	0.0159	0.0264	0.0417	0.0357 (0.0351)	0.0112	0.0198	0.0344	0.0293 (0.0309)
ARI	0.9598	0.9837	1.0000	0.9492 (0.1352)	0.9632	0.9877	0.9959	0.9674 (0.0851)

Table 3 continued

	$G = 3$ and $K = 2$							
	$N = 500$				$N = 1000$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
$\boldsymbol{\mu}_c$	0.0043	0.0076	0.0168	0.3999 (0.8532)	0.0023	0.0039	0.0065	0.1904 (0.6676)
$\boldsymbol{\mu}_o$	0.0131	0.0227	0.6699	0.5072 (0.9089)	0.0059	0.0120	0.0559	0.3193 (0.8463)
$\boldsymbol{\gamma}$	0.0084	0.0189	0.0278	0.0215 (0.0192)	0.0049	0.0085	0.0161	0.0161 (0.0275)
$\mathbf{p}$	0.0001	0.0002	0.0005	0.0006 (0.0015)	0.0001	0.0001	0.0003	0.0003 (0.0007)
$\boldsymbol{\Lambda}$	0.4258	0.5968	0.8003	0.6029 (0.3126)	0.3626	0.5817	0.8474	0.6304 (0.4115)
$\mathbf{L}$	0.1770	0.3299	0.7482	0.4991 (0.4240)	0.1098	0.2848	0.5499	0.4006 (0.3898)
$\boldsymbol{\Psi}$	0.0124	0.0191	0.0510	0.0425 (0.0498)	0.0063	0.0099	0.0241	0.0272 (0.0387)
ARI	0.8273	0.8926	0.9383	0.8523 (0.1363)	0.9400	0.9654	0.9839	0.9347 (0.1115)
	$G = 3$ and $K = 4$							
	$N = 500$				$N = 1000$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
$\boldsymbol{\mu}_c$	0.0089	0.0155	0.4670	0.4449 (0.7739)	0.0036	0.0064	0.0107	0.2597 (0.6184)
$\boldsymbol{\mu}_o$	0.0171	0.0270	0.9474	0.5748 (0.9731)	0.0077	0.0122	0.0279	0.4021 (0.9426)
$\boldsymbol{\gamma}$	0.0090	0.0177	0.0281	0.0206 (0.0152)	0.0053	0.0096	0.0165	0.0144 (0.0316)
$\mathbf{p}$	0.0001	0.0003	0.0007	0.0006 (0.0010)	0.0001	0.0002	0.0003	0.0003 (0.0005)
$\boldsymbol{\Lambda}$	0.4913	0.6322	0.7808	0.6398 (0.2214)	0.4773	0.6257	0.7658	0.6373 (0.2292)
$\mathbf{L}$	0.3202	0.4880	0.6507	0.5213 (0.2848)	0.2859	0.4687	0.6811	0.5062 (0.3303)
$\boldsymbol{\Psi}$	0.0265	0.0388	0.0711	0.0574 (0.0482)	0.0142	0.0266	0.0505	0.0397 (0.0360)
ARI	0.8288	0.8943	0.9425	0.8613 (0.1281)	0.8710	0.9159	0.9428	0.8914 (0.0848)

Table 4 Quartiles, mean and standard deviation (in brackets) of the imprecision indexes for the parameter estimates obtained by maximizing the composite likelihood and the full likelihood over 250 samples generated from the model $(SC)UU$ with $G = 3$, $K = 4$, $N = 500$

	Full likelihood case			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
$\boldsymbol{\mu}_c$	0.0199	0.0411	0.6961	0.5193 (0.8298)
$\boldsymbol{\mu}_o$	0.0342	0.0726	0.5093	0.3537 (0.4880)
$\boldsymbol{\gamma}$	0.0126	0.0245	0.0505	0.0358 (0.0352)
$\mathbf{p}$	0.0001	0.0003	0.0010	0.0008 (0.0014)
$\boldsymbol{\Lambda}$	0.4029	0.5050	0.6103	0.5166 (0.1655)
$\mathbf{L}$	0.2723	0.3382	0.4099	0.3459 (0.1133)
$\boldsymbol{\Psi}$	0.0824	0.1014	0.1230	0.1041 (0.0284)
	Composite likelihood case			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
$\boldsymbol{\mu}_c$	0.0092	0.0177	0.8184	0.5717 (1.0149)
$\boldsymbol{\mu}_o$	0.0155	0.0306	1.0790	0.5936 (0.9846)
$\boldsymbol{\gamma}$	0.0068	0.0157	0.0267	0.0224 (0.0228)
$\mathbf{p}$	0.0001	0.0003	0.0007	0.0008 (0.0016)
$\boldsymbol{\Lambda}$	0.4720	0.6055	0.7713	0.6229 (0.2138)
$\mathbf{L}$	0.3015	0.4771	0.6981	0.5314 (0.2966)
$\boldsymbol{\Psi}$	0.0338	0.0508	0.0945	0.0686 (0.0458)
Computational Time Ratios				
$\frac{\text{Full}}{\text{Composite}}$	2.2040	4.3822	7.3894	11.4821 (25.7926)

The last row contains the same statistics for the distribution of the ratios of the computational times

Table 5 Quartiles, mean and standard deviation (in brackets) of the ARI for the estimated partition obtained by different estimators over 250 samples generated from the model $(SC)UU$ with $G = 3$, $K = 4$ and $N = 500$

	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
Full	0.7450	0.8907	0.9416	0.8048 (0.1906)
Composite	0.6933	0.8876	0.9397	0.7945 (0.2021)
Naive	0.4194	0.5108	0.8935	0.6103 (0.2421)
EII	0.4540	0.5846	0.7233	0.5892 (0.1814)
VII	0.4512	0.5781	0.7290	0.5839 (0.1832)
EEI	0.4384	0.6238	0.7666	0.6031 (0.2105)
VEI	0.4314	0.6089	0.7528	0.5925 (0.2140)
EVI	0.4245	0.5981	0.7636	0.5996 (0.2128)
VVI	0.4233	0.5919	0.7547	0.5944 (0.2117)

available in the package `clustMD`. We considered the following design: number of observations ($N = 500$), number of groups ($G = 2, 3$) and number of latent factors ($K = 4$). The number of variables is 8 of which 4 are ordinal. For the ordinal variables we considered two cases: 5 or 10 categories for each ordinal variable. The other model parameters are generated as described previously under the model $(SC)UU$. We generated 250 samples.

Table 6 Quartiles, mean and standard deviation (in brackets) of the imprecision indexes for the parameter estimates obtained by maximizing the composite likelihood and the ARI for the estimated partition obtained by different estimators over 250 samples generated from the model $(SC)UU$ with $G = 2, 3$, $K = 4$, $N = 500$, ordinal variables with 5 or 10 categories

	5 categories							
	$G = 2$				$G = 3$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
$\boldsymbol{\mu}_c$	0.0044	0.0071	0.0129	0.0672 (0.4778)	0.0093	0.0175	0.5245	0.5177 (0.9924)
$\boldsymbol{\mu}_o$	0.0074	0.0122	0.0177	0.0552 (0.4253)	0.0180	0.0319	0.7657	0.5212 (0.9669)
$\boldsymbol{\gamma}$	0.0097	0.0154	0.0285	0.0749 (0.4174)	0.0118	0.0212	0.0366	0.0896 (0.5714)
$\mathbf{p}$	0.0001	0.0002	0.0007	0.0011 (0.0038)	0.0001	0.0004	0.0008	0.0009 (0.0018)
$\boldsymbol{\Lambda}$	0.4386	0.5846	0.7690	0.6097 (0.2201)	0.4269	0.5931	0.7457	0.5950 (0.2252)
$\mathbf{L}$	0.2186	0.3504	0.5347	0.4330 (0.3455)	0.2961	0.4557	0.6800	0.5159 (0.2960)
$\boldsymbol{\Psi}$	0.0166	0.0286	0.0649	0.0470 (0.0435)	0.0245	0.0438	0.0961	0.0653 (0.0535)

	5 categories - ARI							
	$G = 2$				$G = 3$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
Composite	0.8327	0.9428	0.9834	0.8642 (0.1998)	0.7621	0.8591	0.9285	0.7949 (0.1892)
Naive	0.5431	0.8881	0.9362	0.7053 (0.3151)	0.4019	0.4806	0.7551	0.5583 (0.2375)
EII	0.7709	0.8888	0.9692	0.8435 (0.1579)	0.4761	0.6000	0.7357	0.6020 (0.1757)
VII	0.7583	0.8890	0.9673	0.8337 (0.1728)	0.4641	0.5865	0.7229	0.5963 (0.17749)
EEI	0.7148	0.8814	0.9753	0.7836 (0.2579)	0.4179	0.5976	0.7594	0.5955 (0.2104)
VEI	0.6515	0.8816	0.9754	0.7714 (0.2651)	0.4323	0.6064	0.7636	0.5944 (0.2161)
EVI	0.6688	0.8934	0.9754	0.7763 (0.2602)	0.4289	0.5860	0.7553	0.5911 (0.2124)
VVI	0.5356	0.8724	0.9754	0.7440 (0.2818)	0.4237	0.5888	0.7569	0.5874 (0.2126)

Table 6 continued

	10 categories							
	$G = 2$				$G = 3$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
$\boldsymbol{\mu}_c$	0.0043	0.0077	0.0154	0.2419 (0.7407)	0.0092	0.0176	0.3887	0.5065 (0.9507)
$\boldsymbol{\mu}_o$	0.0077	0.0127	0.0262	0.2304 (0.5579)	0.0168	0.0287	0.9542	0.4849 (0.7551)
$\boldsymbol{\gamma}$	0.0231	0.0409	0.8116	0.4711 (0.8236)	0.0268	0.0498	1.0771	0.6374 (0.9596)
$\mathbf{p}$	0.0000	0.0003	0.0010	0.0037 (0.0087)	0.0001	0.0004	0.0009	0.0012 (0.0033)
$\boldsymbol{\Lambda}$	0.4635	0.6237	0.7541	0.6181(0.2062)	0.4403	0.6002	0.7870	0.6214 (0.2362)
$\mathbf{L}$	0.1831	0.3265	0.4838	0.3838 (0.2992)	0.2978	0.4355	0.5973	0.4834 (0.2734)
$\boldsymbol{\Psi}$	0.0175	0.0255	0.0573	0.0458 (0.0434)	0.0220	0.0446	0.0878	0.0598 (0.0469)

	10 categories - ARI							
	$G = 2$				$G = 3$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
Composite	0.9196	0.9754	1.0000	0.8706 (0.2560)	0.7280	0.8480	0.9224	0.7788 (0.1977)
Naive	0.1168	0.6546	0.9837	0.5531 (0.4190)	0.2458	0.4294	0.6472	0.4611 (0.2486)
EII	0.7625	0.8889	0.9754	0.8433 (0.1572)	0.4051	0.5531	0.6613	0.5605 (0.1660)
VII	0.7499	0.9043	0.9751	0.8356 (0.1742)	0.4223	0.5441	0.7089	0.5528 (0.1833)
EEI	0.7122	0.8872	0.9754	0.7860 (0.2601)	0.3972	0.5768	0.7429	0.5629 (0.2038)
VEI	0.6151	0.8945	0.9755	0.7712 (0.2732)	0.3623	0.5872	0.7422	0.5567 (0.2211)
EVI	0.6716	0.8921	0.9757	0.7777 (0.2637)	0.3767	0.5212	0.7170	0.5455 (0.2099)
VVI	0.5269	0.8749	0.9757	0.7357 (0.2914)	0.3585	0.5461	0.7243	0.5399 (0.2195)

Table 6 displays the results in terms of imprecision indexes and ARI values. As regards the performances of the composite likelihood estimator, overall we noted that when the model complexity increases in terms of the number of components G or higher number of categories, the estimator efficiency decreases. In terms of recovering the clustering structure, we noted that in some cases the algorithm of six models available in the R package `clustMD` was not able to provide a solution. Looking at the ARI values our proposal is still the best one (mean values varying between 0.78 and 0.87), the naive case results to be the worst (mean values varying between 0.46 and 0.71). The six parsimonious models show lower ARI values; the mean values are between 0.54 and 0.84.

5.4 The effect of larger number of ordinal variables

Finally, in the last experiment we evaluate the effect of higher number of ordinal variables. This allows to evaluate how the sparsity in the data could affect the parameter estimation and the classification performances. This means that some profiles could have zero or very low frequencies, but this is not a problem for the algorithm. Furthermore we compare the clustering performance, in terms of ARI, of the proposal with the naive case (i.e. our model treating all variables as they were continuous) and six parsimonious models available in the package `clustMD`. We considered the following design: number of observations ($N = 1000$), number of groups ($G = 2, 3$) and number of latent factors ($K = 4$). The number of variables is 15 of which 10 are ordinal. We considered 5 categories for each ordinal variable.
For $G = 2$, we specify the group-specific mean vector by

$$\begin{aligned}\boldsymbol{\mu}_1 &= [0.5, 0.5, 1, 1, 0.5, 0.5, 1, 1, 1, 1, -1, 2, -2, -1, -1],\\ \boldsymbol{\mu}_2 &= [1.5, 1.5, 0, 0, 1.5, 1.5, 0, 0, 0, 0, 1, 0, 2, 0, 0].\end{aligned}$$

For $G = 3$, we added the further group-specific mean vector,

$$\boldsymbol{\mu}_3 = [0.5, -0.5, -1, -1, 0.5, -0.5, -1, -1, -1, -1, 0, -2, 0, -1, -1].$$

The other model parameters are generated as described previously under the model $(SC)UU$. We generated 250 samples.

Table 7 displays the results in terms of imprecision indexes and ARI values. As regards the performances of the composite likelihood estimator, we do not observe particular issues. Once again when the complexity of the model increases in terms of number of components G, the estimator efficiency decreases, as expected. In terms of recovering the clustering structure, we noted that is many cases the algorithm of six models available in the R package `clustMD` was not able to provide a solution, especially when $G = 3$. This is may be due to the presence of zero frequency categories. Also in this case, we noted that as the number of groups increases the ARI values decrease, even if, less compared to the previous cases. This may be explained by a larger sample size ($N = 1000$). More in details, looking at the ARI values, our proposal is still the best one- the mean values are equal to 0.97 and 0.79 for $G = 2$

Table 7 Quartiles, mean and standard deviation (in brackets) of the imprecision indexes for the parameter estimates obtained by maximizing the composite likelihood and the ARI for the estimated partition obtained by different estimators over 250 samples generated from the model $(SC)UU$ with $G = 2, 3$, $K = 4$, $N = 1000$, 15 variables where 10 are ordinal

	Large Dataset							
	$G = 2$				$G = 3$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
$\boldsymbol{\mu}_c$	0.0033	0.0050	0.0087	0.1211 (0.6715)	0.0060	0.0113	0.1743	0.3349 (0.7321)
$\boldsymbol{\mu}_o$	0.0066	0.0099	0.0144	0.0368 (0.1292)	0.0262	0.0450	0.2506	0.2099 (0.3145)
$\boldsymbol{\gamma}$	0.0136	0.0222	0.0315	0.0250 (0.0164)	0.0186	0.0262	0.0374	0.0288 (0.0151)
$\mathbf{p}$	0.0000	0.0001	0.0003	0.0008 (0.0039)	0.0001	0.0002	0.0005	0.0005 (0.0009)
$\boldsymbol{\Lambda}$	0.4186	0.5123	0.6154	0.5269 (0.1488)	0.4146	0.5365	0.6806	0.5420 (0.1674)
$\mathbf{L}$	0.1975	0.2875	0.4464	0.3320 (0.1858)	0.2204	0.2948	0.4002	0.3197 (0.1429)
$\boldsymbol{\Psi}$	0.0429	0.0577	0.0774	0.0624 (0.0263)	0.0796	0.1007	0.1206	0.1015 (0.0280)

	ARI							
	$G = 2$				$G = 3$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
Composite	0.9724	0.9917	1.0000	0.9749 (0.0422)	0.7488	0.8542	0.9206	0.7893 (0.1997)
Naive	0.0784	1.0000	1.0000	0.7137 (0.4433)	0.4074	0.4834	0.8187	0.5346 (0.3200)
EII	0.7902	0.9388	0.9835	0.8476 (0.2018)	0.4854	0.6176	0.7558	0.6369 (0.1744)
VII	0.7547	0.9258	0.9807	0.8377 (0.2038)	0.5359	0.6342	0.7920	0.6645 (0.1745)
EEI	0.6265	0.9307	0.9837	0.7650 (0.3047)	0.4733	0.6490	0.8324	0.6570 (0.2021)
VEI	0.3718	0.9193	0.9876	0.7253 (0.3369)	0.4789	0.5753	0.8664	0.6500 (0.2178)
EVI	0.5109	0.9200	0.9848	0.7493 (0.3141)	0.4464	0.7467	0.8672	– (–)
VVI	0.3241	0.9085	0.9837	0.6945 (0.3516)	0.4357	0.8087	0.8733	– (–)

Table 8 Percentage distribution of the number of components chosen by aBIC over 250 samples generated from the model $(SC)UU$ with $G = 3$, $K = 2$ and $N = 1000$

Criterion	$G = 2$	$G = 3$	$G = 4$
aBIC	1.20%	82.80%	16.00%

and $G = 3$, respectively. The naive model results to be the worst when $G = 3$ (mean value is 0.53), while the other six parsimonious models show lower ARI values in mean, varying between 0.69 and 0.85 when $G = 2$ and between 0.64 and 0.66 when $G = 3$.

5.5 A simulation study for the model selection

To complete our large simulation study, we tested the effectivness of the aBIC in finding the correct number of components.
We considered the following design: number of observations ($N = 1000$), number of groups ($G = 3$) and number of latent factors ($K = 2$). The number of variables is 8 of which 4 are ordinal. The ordinal variables have 5 categories. The other model parameters are generated as described previously under the model $(SC)UU$. We generated 250 samples. For each sample we fitted the true model, keeping fixed the number of latent factors. Further works are needed to calibrate the penalization term for the additive BIC.

Looking at the results of 8, the aBIC is able to choose the right number for components most of the times (82.80%). In conclusion, even if the idea of measuring the complexity of the model by counting the number of parameters involved in each sub-likelihood may seem simplistic, in this context it gives good results. This does not exclude that further improvements are possible and needed.

6 Real data analysis

We apply the proposal to a set of data taken from the survey carried out by the Italian National Statistical Institute (ISTAT) in 2015 on academic graduates' vocational integration by interviewing a sample of graduates who attained the university degree four years before. The aim of the survey is to detect graduates' employment conditions about four years after graduation. We select the following variables: final grade for high school, age at graduation, final grade for the MSc degree, seven variables to detect job satisfaction, four variables to detect the propensity to move abroad, monthly income, length of study (in years), and gap between graduation and job (in months). We focus only on students of Master degree in Economics and Statistics. Furthermore we restricted the analysis to the observations with non-missing values. The final dataset is composed of 1033 students and 16 variables (12 ordinal variables and 4 continuous variables). The seven variables to detect job satisfaction have 11 categories, while the four variables to detect the propensity to move abroad have 4 categories. We fitted twelve models to the data for $G = 2, 3, 4$ and $K = 1, 2, 3$, and computed the additive BIC values for each model. The values are shown in Fig. 1.

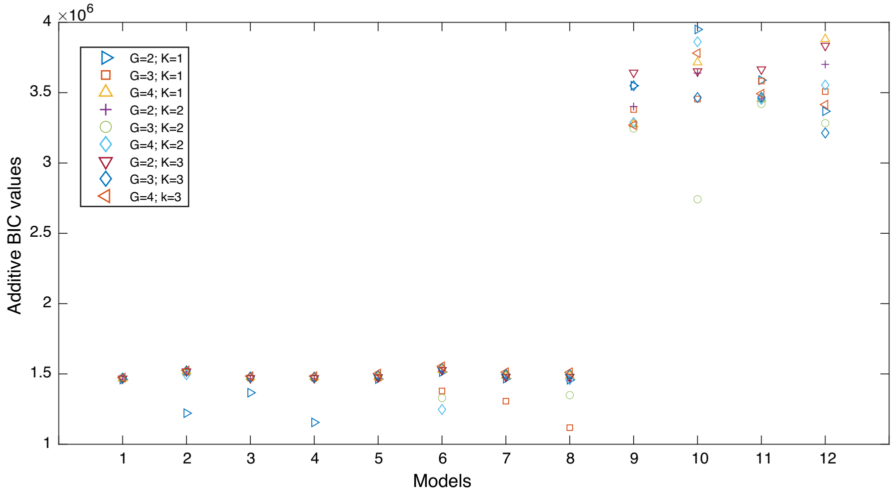

Fig. 1 Model selection according to additive BIC with $G = 2, 3, 4$ and $K = 1, 2, 3$

The best model is that one minimizing the additive BIC. The model with the lowest additive BIC value was a three component mixture with the UUU covariance structure with a single common factor. The factors are changing over the groups, as well as the error terms for each variable. The Table 9 reports the empirical medians and proportions within each group.

Looking at Table 9, it is possible to note that the first two groups are very similar in terms of final grade, satisfaction with the current job, and salary, although median values and proportions are slightly different in terms of willingness to move abroad. On the other hand, in the third group the final mark, the satisfaction with the current job and the monthly income are the lowest, making this group quite different in terms of academic graduates' vocational integration. More in detail, the first group is the smallest (18.48%) and it is composed by the youngest graduates. They got the highest median value as final mark for the degree (110); they are very happy and satisfied with their current work; they are also willing to move abroad only for better qualification (93.10%) or higher salary (69.83%); their monthly median income is 1750 euros, and finally, 50% of graduates started to work just two months after their graduation. The second group is the largest (55.83%), composed mainly by graduates that got 108 as median final mark for the degree. Furthermore, they are quite satisfied with their current job; they want also move abroad for better qualification (70.07%) or higher salary (70.07%), although the percentages are lower than those in the first group; 50% of graduates found job at most three months after graduation and their monthly median income is 1700 euros. The last group is composed by of 26% of graduates. Half of them got 106.5 or less as final median mark; they are not so satisfied with their current job, mainly in terms of long term (43.83%), salary (43.21%) and career perspectives (40.12%). However, the proportions of graduates that want to move abroad is very similar to the second group; 50% of graduates found their job at most three months after the degree, but their monthly median income is the lowest, that is 1300 euros.

Table 9 Empirical medians and proportions with each group based on the classification given by the best model, i.e. UUU with $G = 3$ and $K = 1$

Median or $\hat{\pi}$ for	Group 1 (18.48%)	Group 2 (55.83%)	Group 3 (25.69%)
Grade for High School	88	90	86
Age at Graduation [> 25]	73.28%	76.95	88.89%
Grade for Degree	110	108	106.5
Happy for Duties [6, 10]	98.28%	93.64%	72.22%
Happy for Long Term [6, 10]	93.97%	87.02%	43.83%
Happy for Independence/Responsibilities [6, 10]	100.00%	93.91%	79.01%
Happy for Knowledge from University [6, 10]	87.93%	74.44%	56.17%
Happy for Salary [6, 10]	90.51%	80.66%	43.21%
Happy for Career [6, 10]	96.55%	82.91%	40.12%
Happy for Professional growth [6, 10]	95.69%	89.93%	63.58%
Moving abroad for Qualification (Much or Enough)	93.10%	70.07%	75.31%
Moving abroad for Salary (Much or Enough)	69.83%	70.07%	70.99%
Moving abroad for Opportunity (Much or Enough)	50.00%	50.59%	55.56%
Moving abroad for Personal reasons (Much or Enough)	25.00%	39.34%	36.42%
Total Monthly Income	1750	1700	1300
Gap between Degree and Job	2	3	3

Table 10 Correlations between the variables and the single common factor of each group

Variables	Group 1	Group 2	Group 3
Grade for High School	0.81	0.16	0.76
Age at Graduation [> 25]	0.79	0.64	0.73
Grade for Degree	0.71	−0.75	0.65
Happy for Duties [6, 10]	0.89	0.89	0.79
Happy for Long Term [6, 10]	0.56	0.94	0.74
Happy for Independence/Responsibilities [6, 10]	0.57	0.86	0.76
Happy for Knowledge from University [6, 10]	0.88	0.78	0.77
Happy for Salary [6, 10]	0.72	0.73	0.50
Happy for Career [6, 10]	0.74	0.74	0.79
Happy for Professional growth [6, 10]	0.83	0.86	0.64
Moving abroad for Qualification (Much or Enough)	0.47	0.58	0.74
Moving abroad for Salary (Much or Enough)	0.88	0.77	0.86
Moving abroad for Opportunity (Much or Enough)	0.74	0.88	0.70
Moving abroad for Personal reasons (Much or Enough)	0.81	0.96	0.76
Total Monthly Income	0.57	0.84	0.52
Gap between Degree and Job	0.69	0.97	0.65

In each group we have a single factor that is, being unique, a sort of *overall*. Looking at the correlations between the variables and the latent factors, we note that, as expected, they are all positive (see Table 10). The only exception is the grade for degree in the second group that is negative. It implies that this variable in the second group has a negative correlation with the other variables while they are positive in the other groups. In practice, in the second group a higher grade is associated with a lower degree of happiness and willingness to move. A possible explanation could be that in this group the expectations of the students are strictly related with the grade for degree: high grade implies high expectations that are rarely satisfied; low grade implies low expectations that are easier to satisfy.

Furthermore we fit the data with six flexible models available in the R package `clustMD`. However we noted some issues: if some categories have zero frequencies or there are many categories, the algorithms are not able to provide a solution. To overcome this issue, we merged the categories such that all the categories have non-zero frequencies and the number of categories is reduced. The solutions are not directly comparable. For this reason we did not report the results.

7 Discussion

In this paper, we have introduced a general class of parsimonious Gaussian mixture models for clustering ordinal and continuous variables. It includes known parameterizations proposed for continuous data as mixture of factor analysers and mixture of probabilistic principal component analysis as special cases. In order to increase the flexibility, we also introduced a new parameterization introduced in the context of multiway and multigroup data. The main advantage is that the number of covariance parameters grows linearly with the number of variables, rather than polynomially. This feature, along with the maximum composite likelihood estimation, makes the application of such models possible even for high dimensional mixed type data. The effectiveness of the proposal has been tested through a simulation study. Additionally, this class of models appears to be particularly good at modelling situations where some of the variables are highly correlated within the groups, as expected in high dimensional data. The application to the university graduates' employment conditions about four years after graduation in Italy indicates that the model gives excellent clustering performance. The clusters found using the models showed greater ability to capture the group structure, by defining the main features of the graduates in each group.

Finally, we summarize some limits and possible extensions of our proposal. First of all, we note that counting and nominal variables are not considered. Their introduction would be quite easy under a local independence assumption, i.e. assuming that the observed variables are independent within the components. However, this assumption is in contrast with the spirit of this work where the dependencies among the variables within a component are modelled trough a factor analysis model. Such a modelling is not trivial; for example some authors argue that in some cases factor analysis is not applicable to nominal data (Revuelta et al. 2019), and probably would lead to a different way to build the composite likelihood. All in all, further studies are needed. Another direction of extension of the present work is to introduce other ways to constrain the parameters of the factor analysis model across the groups. As an example, McNicholas and Murphy (2010) propose to rewrite the diagonal matrix of the uniqueness $\boldsymbol{\Psi}_g$ as the product of a scalar ω_g and a diagonal matrix $\boldsymbol{\Delta}_g$, having the determinant equal to 1. New parsimonious parameterizations are obtained constraining the first or the second factor of the product to be the same across the components. In our model, such constraints could be applied to the uniqueness matrices and to the saliences matrices $\mathbf{L}_g$. Other models could be simply obtained by allowing the number of factors to be cluster specific. However, this does not mean that finding the best fitting model based on information criteria is so trivial. Indeed it may result to be time consuming, since we need to simultaneously choose $G, K_1, \ldots, K_G$, and the covariance parameterisation. To overcome this issue it could be useful adopting a penalization approach. As an example, by introducing a lasso term, as explored in a different context (see e.g. Khalili and Chen (2007); Chen and Khalili (2008)), it may be possible to define a regularization path for the model selection. This would allow us to fit a lower number of models and to choose the best one based on a pre-specified information criterion. This potential solution could be developed in a future work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11634-023-00539-5.

Acknowledgements We would like to thank the three reviewers and the editor for their helpful comments and suggestions. We believe that they helped us to improve the quality of the manuscript.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Carroll JD, Chang JJ (1970) Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. Psychometrika 35(3):283–319

Cattell RB (1944) "Parallel proportional profiles" and other principles for determining the choice of factors by rotation. Psychometrika 9(4):267–283

Chen J, Khalili A (2008) Order selection in finite mixture models with a nonsmooth penalty. J Am Stat Assoc 103(484):1674–1683

DiStefano C (2002) The impact of categorization with confirmatory factor analysis. Struct Equ Model 9(3):327–346

Dolan CV (1994) Factor analysis of variables with 2, 3, 5 and 7 response categories: a comparison of categorical variable estimators using simulated data. Br J Math Stat Psychol 47(2):309–326

Everitt B (1988) A finite mixture model for the clustering of mixed-mode data. Stat Probab Lett 6(5):305–309

Farooq R (2022) Heywood cases: possible causes and solutions. Int J Data Anal Tech Strateg 14:79–88

Gao X, Song PXK (2010) Composite likelihood Bayesian information criteria for model selection in high-dimensional data. J Am Stat Assoc 105(492):1531–1540
Ghahramani Z, Hinton GE (1996) The em algorithm for mixtures of factor analyzers
Giordani P, Rocci R, Bove G (2020) Factor uniqueness of the structural parafac model. Psychometrika 85(3):555–574
Harshman RA, et al (1970) Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis
Hennig C, Meila M, Murtagh F, Rocci R (2015) Handbook of cluster analysis. CRC Press
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218
Ingrassia S, Punzo A, Vittadini G, Minotti S (2015) The generalized linear mixed cluster-weighted model. J Classif 32:85–113
Jöreskog KG (1990) New developments in lisrel: analysis of ordinal variables using polychoric correlations and weighted least squares. Qual Quant 24(4):387–404
Jöreskog K G, Sörbom D (1996) LISREL 8: User's Reference Guide. Scientific Software
Khalili A, Chen J (2007) Variable selection in finite mixture of regression models. J Am Stat Assoc 102(479):1025–1038
Kruskal JB (1977) Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. Linear Algebra Appl 18(2):95–138
Lawley DN, Maxwell AE (1962) Factor analysis as a statistical method. J R Stat Soc Ser D (Stat) 12(3):209–229
Ledermann W (1937) On the rank of the reduced correlation matrix in multiple-factor analysis. Psychometrika 2:85–93
Lee SY, Poon WY, Bentler P (1990) Full maximum likelihood analysis of structural equation models with polytomous variables. Stat Probab Lett 9(1):91–97
Lindsay B (1988) Composite likelihood methods. Contemp Math 80:221–239
Mardia KV, Kent JT, Hughes G, Taylor CC (2009) Maximum likelihood estimation using composite likelihoods for closed exponential families. Biometrika 96(4):975–982
MATLAB. User's Guide, R2013b. MathWorks (2013)
McLachlan G, Peel D, Bean R (2003) Modelling high-dimensional data by mixtures of factor analyzers. Comput Stat Data Anal 41(3):379–388
McNicholas P, Murphy T (2008) Parsimonious gaussian mixture models. Stat Comput 18(3):285–296
McNicholas PD, Murphy TB (2010) Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics 26(21):2705–2712
Mcparland D, Gormley I (2015) Model based clustering for mixed data: clustmd. Adv Data Anal Classif 10:1
Mcparland D, Gormley I, McCormick T, Clark S, Kabudula C, Collinson M (2014) Clustering south African households based on their asset status using latent variable models. Ann Appl Stat 8:1
Mcparland D, Phillips C, Brennan L, Roche H, Gormley I (2017) Clustering high-dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data. Stat Med 36:1
Millsap RE, Yun-Tein J (2004) Assessing factorial invariance in ordered-categorical measures. Multivar Behav Res 39(3):479–515
Molenberghs G, Verbeke G (2005) Models for discrete longitudinal data. Springer Series in Statistics Series. Springer, New York
Murphy K, Murphy TB (2020) Gaussian parsimonious clustering models with covariates and a noise component. Adv Data Anal Classif 14:293–325
Muthén B (1984) A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika 49(1):115–132
Ranalli M, Rocci R (2016) Mixture models for ordinal data: a pairwise likelihood approach. Stat Comput 1:1–19
Ranalli M, Rocci R (2017) Mixture models for mixed-type data through a composite likelihood approach. Comput Stat Data Anal 110(C):87–102
Revuelta J, Maydeu Olivares A, Ximénez C (2019) Factor analysis for nominal (first choice) data. Struct Equ Modeling 27(5):781–797
Rhemtulla M, Brosseau-Liard PÉ, Savalei V (2012) When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. Psychol Methods 17(3):354

Shapiro A (1985) Identifiability of factor analysis: some results and open problems. Linear Algebra Appl 70:1–7
Stegeman A, Lam TTT (2016) Multi-set factor analysis by means of parafac2. Br J Math Stat Psychol 69:1–19
Tipping M, House SG, Street G, Nh CC (1999) Probabilistic visualisation of high-dimensional binary data
Varin C, Reid N, Firth D (2011) An overview of composite likelihood methods. Stat Sin 21(1):1–41
Xu X, Reid N (2011) On the robustness of maximum composite likelihood estimate. J Stat Plan Inference 141(9):3047–3054

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.