Statistics and Computing (2025) 35:50
https://doi.org/10.1007/s11222-025-10579-w

ORIGINAL PAPER

Multilevel latent class models for cross-classified categorical data: model definition and estimation through stochastic EM

S. Columbu[1] · N. Piras[1] · J. K. Vermunt[2]

Received: 26 June 2024 / Accepted: 24 January 2025 / Published online: 13 February 2025

Abstract

We present an extension of the multilevel latent class model for dealing with multilevel cross-classified categorical data. Cross-classified data structures arise when observations are simultaneously nested within two or more groups, for example, children nested within both schools and neighborhoods. More specifically, we propose extending the standard hierarchical latent class model, which contains mixture components at two levels, say for children and schools, by including a separate set of mixture components for each of the higher-level crossed classifications, say for schools and neighborhoods. Because of the complex dependency structure arising from the cross-classified nature of the data, it is no longer possible to obtain maximum likelihood estimates of the model parameters, for example, using the EM algorithm. As a solution to the estimation problem, we propose an approximate estimation approach using a stochastic version of the EM algorithm. The performance of this approach, which resembles Gibbs sampling, was investigated through a set of simulation studies. Moreover, the application of the new model is illustrated using an Italian dataset on the quality of university experience at degree programme level, with degree programmes nested in both universities and fields of study.

Keywords Latent class · Cross-classified · Stochastic EM · Gibbs sampling · Multilevel

1 Introduction

The popularity of latent class (LC) analysis (Lazarsfeld 1950; Hagenaars and McCutcheon 2002) as a powerful and flexible statistical tool for model-based clustering with categorical data is still increasing, among others in applied fields such as the social sciences, machine learning, and information sciences. Various extensions of the standard LC model have been introduced to handle more complex data structures, for instance when the statistical units are grouped within one or more higher-level observations; that is, when there is a hierarchical nesting of observations at different levels, for example, children nested within schools or patients nested within hospitals. An approach proposed by Vermunt (see Vermunt 2003, 2008, 2004) involves having mixture components at each level (thus for both schools and children), implying that a discrete latent variable is introduced at the lower and the higher level of the hierarchy. For parameter estimation a special implementation of the E step of the EM algorithm can be used which is referred to the upward-downward algorithm.

However, sometimes the nesting consists of multiple higher levels which are not hierarchically linked, but instead cross-classified, for example, children could be considered nested within both schools and neighborhoods. In such situations, there is the need to introduce two sets of higher-level mixture components, with associated latent variables, one for each of the two cross-classified levels. The idea of considering separate latent variables for the mixture at each level of the hierarchy, resembles what is commonly done in mixed effects models, in which continuous latent variables that vary between clusters (random-effects) are introduced to capture the unobserved heterogeneity (Skrondal and Rabe-Hesketh 2004; Goldstein 2010). While either mixed effects than LC approaches assume the presence of underlying unobserved variables, their fundamental distinction is that in mixed

✉ S. Columbu
silvia.columbu@unica.it

N. Piras
nicola.piras97@unica.it

J. K. Vermunt
J.K.Vermunt@tilburguniversity.edu

[1] Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy

[2] Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands

effects models the latent variables are continuous and fully parametric, typically a Gaussian distribution is assumed, whereas the latent variables introduced to account for the multilevel structure in mixture models are discrete and no parametric assumptions about their distribution is formulated (Bartolucci et al. 2022).

In this paper we propose an approach for latent class analysis of multilevel cross-classified data that extends the one defined in Vermunt (2003). Given the intractability of the derived likelihood, the standard EM algorithm can no longer be applied. We therefore propose using a stochastic version of the EM algorithm that can handle the hierarchy of units but also their double cross-classification. The idea behind the estimation approach is borrowed from the co-clustering literature and in particular from latent block models (Holland et al. 1983; Keribin et al. 2015; Biernacki et al. 2023). Co-clustering approaches deal with the simultaneous clustering of rows and columns of data matrices, and show strong resemblance with the cross-classified structure of our model, in the sense that each observational unit is simultaneously nested within a combination of cross-classified units.

Stochastic approaches for approximate maximum likelihood estimation have been proved to be able to handle non-tractable M steps in complex likelihood frames, as well as to be very useful to overcome initialization issues usually encountered in the application of standard EM. However, a well-known complication occurring when applying such stochastic approaches with mixtures models is label switching, which can be even more prominent in complex settings such as those discussed in this paper. We present possible solutions to label switching when discussing the results of simulation studies.

The aim of the model we propose is to cluster both the lower-level units and the higher-level cross-classified units. More specifically, we simultaneously obtain a clustering of the lower-level units nested within cross-classified units and separate clusterings for each of the two cross-classified units. Therefore, the resulting mixture model formulation contains three discrete latent variables, one for each block of units in the data structure. This is in contrast to the standard multilevel LC model, which contains two discrete latent variables.

The paper is organized as follows. In Section 2 we present the multilevel cross-classified latent class model (MCCLC) formulation for categorical data, and in Section 3 we describe the estimation methods adopted. Section 4 is devoted to the investigation of the method by means of simulations studies under different conditions. Section 5 shows an application to real data on Italian University quality. Finally, Section 6 is a discussion section with concluding remarks and perspectives on future extensions.

2 Model formulation

For model formulation we extend the notation introduced in Vermunt (2003) to handle the double grouping of observations. We use the index j to refer to a first-level unit belonging to a unique combination of two second-level units indexed by k and q, where $k = 1, \ldots, K$ and $q = 1, \ldots, Q$, implying the number of second-level units equals K and Q, respectively. The number of first-level units within each (k, q) combination is indicated by n_{kq}. If $\boldsymbol{y}$ is the data matrix composed of I variables, then y_{ijkq} denotes the value of variable i ($i = 1, \ldots, I$) of first level unit j belonging to the cross-classified (CC) group level units k and q. A particular category of variable i is denoted by m_i, its number of categories by M_i, and a possible answer pattern by $\mathbf{m}$. As explained in detail below, our model contains three discrete membership variables, one at level-1 and two at level-2, which are referred to as X_{jkq}, W_k, and Z_q, with ℓ, h and r indicating a particular latent class and L, H and R the number of latent classes.

The proposed LC model for CC multilevel data consists of separate mixture distributions for each level of the hierarchical structure, which can be expressed through two separate equations. Because the first-level observations belong to a combined grouping (k, q), the mixture model should take into account the joint belonging to the corresponding CC latent classes, which, in the complete data form, yields the following model formulation:

$$
\begin{aligned}
&P(\mathbf{Y}_{kq}, W_k = h, Z_q = r) \\
&\quad = P(W_k = h, Z_q = r) \\
&\quad P(\mathbf{Y}_{kq} | W_k = h, Z_q = r) \\
&\quad = P(W_k = h, Z_q = r) \\
&\quad \prod_{j=1}^{n_{kq}} P(\mathbf{Y}_{jkq} = \mathbf{m} | W_k = h, Z_q = r) \\
&\quad = P(W_k = h) P(Z_q = r) \\
&\quad \prod_{j=1}^{n_{kq}} P(\mathbf{Y}_{jkq} = \mathbf{m} | W_k = h, Z_q = r).
\end{aligned}
\tag{1}
$$

For the first level, the model structure is similar to that of a standard mixture model (see among others McLachlan and Peel 2004), where instead mixing proportions and data distribution are conditional to their belonging to a specific combination of latent classes along the multilevel structure; that is,

$$
\begin{aligned}
&P(\mathbf{Y}_{jkq} = \mathbf{m}|W_k = h, Z_q = r) \\
&= \sum_{\ell=1}^{L} P(X_{jkq} = \ell|W_k = h, Z_q = r) \\
&P(\boldsymbol{Y}_{jkq} = \boldsymbol{m}|X_{jkq} = \ell) \\
&= \sum_{\ell=1}^{L} P(X_{jkq} = \ell|W_k = h, Z_q = r) \\
&\prod_{i=1}^{I} P(Y_{ijkq} = m_i|X_{jkq} = \ell)
\end{aligned} \tag{2}
$$

with

$$
P(Y_{ijkq} = m_i|X_{jkq} = \ell) = \prod_{m_i=1}^{M_i} (\pi_{y_i=m_i|\ell})^{y_{ijkq}^{m_i}}
$$

where $y_{ijkq}^{m_i} = 1$ if the observation takes value m_i and 0 otherwise, and $\pi_{y_i=m_i|\ell}$ are the probability parameters in the multinomial distribution. The vector of the whole parameters to be estimated is $\boldsymbol{\theta} = \{P(X_{jkq} = \ell|W_k = h, Z_q = r), P(W_k = h), P(Z_q = r), P(Y_{ijkq} = m_i|X_{jkq} = \ell)\}$.
In the above model definition, we used the following (conditional) independence assumptions:

a) *constraint*: the parameters defining the conditional distributions for the response variables are independent from CC level latent classes (W_k and Z_q) once conditioned on first level latent classes (X_{jkq}):

$$
\begin{aligned}
P(Y_{ijkq} &= m_i|X_{jkq} = \ell) \\
&= P(Y_{ijkq} = m_i|X_{jkq} = \ell, W_k = h, Z_q = r);
\end{aligned}
$$

b) *cross-classified independence*: membership variables at the second cross-classified level are assumed to be independent:

$$
P(W_k = h, Z_q = r) = P(W_k = h)P(Z_q = r);
$$

c) *local independence*: response variables are independent from each other once conditioned on first level latent classes (X_{jkq}):

$$
P(\boldsymbol{Y}_{jkq} = \boldsymbol{m}|X_{jkq} = \ell) = \prod_{i=1}^{I} P(Y_{ijkq} = m_i|X_{jkq} = \ell);
$$

d) *group level conditional independence*: the level-1 observations within group (k, q) are independent from each other once conditioned on the CC latent classes W_k and Z_q:

$$
\begin{aligned}
&P(\boldsymbol{Y}_{kq}|W_k = h, Z_q = r) \\
&= \prod_{j=1}^{n_{kq}} P(\boldsymbol{Y}_{jkq} = \boldsymbol{m}|W_k = h, Z_q = r).
\end{aligned}
$$

The assumption b) of independence between the two latent variables W and Z at level-2 could be in principle relaxed, however, this would result in a complication in the estimation process. Such independence assumption makes sense because we aim at keeping the separation of the two cross-classified groups. In case of dependency, the model could be reformulated as a hierarchical one with a single group level given by $Z \times W$, similarly to what is proposed for cross classified mixed models with continuous random effects when an interaction term is present.

The assumption of local independence reduces the multivariate complexity of the model as only univariate distribution functions are involved. This assumption could be relaxed via the introduction of multivariate distributions when dealing with responses of the same nature.

The model, together with independence assumptions, can be represented as a diagram with a tree structure as in Fig. 1. The upper nodes are the discrete membership variables at the higher cross-classified level. Then for each combination of higher level units we have the n_{kq} lower level latent classes, the observed responses will be nested within the lower level latent classes.

3 Parameter estimation through stochastic EM

The estimation of the vector $\boldsymbol{\theta}$ of model parameters requires the maximization of the observed likelihood of the model in the form:

$$
\begin{aligned}
L(\boldsymbol{\theta}; \boldsymbol{y}) = &\sum_{h_1=1}^{H} \sum_{h_2=1}^{H} \\
&\cdots \sum_{h_K=1}^{H} \sum_{r_1=1}^{R} \sum_{r_2=1}^{R} \cdots \sum_{r_Q=1}^{R} \\
&\prod_{k=1}^{K} \prod_{q=1}^{Q} P(\boldsymbol{Y}_{kq}) \\
= &\sum_{h_1=1}^{H} \sum_{h_2=1}^{H} \cdots \sum_{h_K=1}^{H} \sum_{r_1=1}^{R} \sum_{r_2=1}^{R} \\
&\cdots \sum_{r_Q=1}^{R} \prod_{k=1}^{K} P(W_k = h_k) \prod_{q=1}^{Q} P(Z_q = r_q) \\
&\prod_{j=1}^{n_{kq}} \left[\sum_{\ell=1}^{L} P(X_{jkq} = \ell|W_k = h_k, Z_q = r_q) \right. \\
&\left. P(\boldsymbol{Y}_{jkq} = \boldsymbol{m}|X_{jkq} = \ell) \right].
\end{aligned} \tag{3}
$$

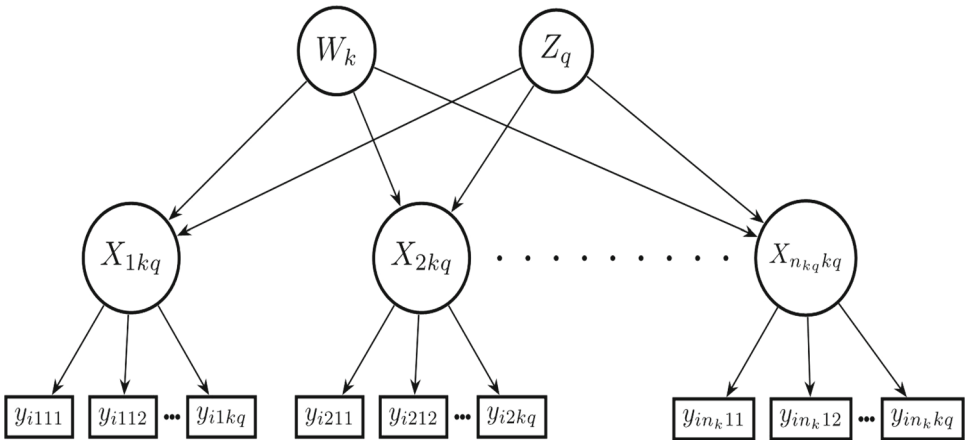

Fig. 1 Diagram associated to a MCCLC model. Latent variables are expressed at each level of the multilevel structure

The presence of a double missing data structure at higher level, with W_k and Z_q unobserved, causes that the likelihood cannot factorize to the product of the mixing probabilities as for standard LC and multilevel LC models. In fact, it is required a marginalization over all possible label configurations involving the sum of $H^K R^Q$ terms, which causes that the computation of the likelihood (and its logarithm) gets rapidly not tractable.

The standard approach for the estimation of model parameters in latent class analysis and mixture modelling consists in performing maximum likelihood estimation through the implementation of Expectation Maximization (EM)-type algorithms (Dempster et al. 1977; McLachlan and Krishnan 2008), applied to the complete data log-likelihood ($L_c(\boldsymbol{\theta})$). The E step of the algorithm involves the computation of the expectation of the log-likelihood, which in our cross-classified model takes the form:

$$
\begin{aligned}
&E(\log L_c(\boldsymbol{x}, \boldsymbol{w}, z, \boldsymbol{\theta})) \\
&\quad = \sum_{k=1}^{K}\sum_{h=1}^{H} P(W_k = h|\boldsymbol{y}_{kq}) \log \pi_h \\
&\qquad + \sum_{q=1}^{Q}\sum_{r=1}^{R} P(Z_q = r|\boldsymbol{y}_{kq}) \log \pi_r \\
&\qquad + \sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{h=1}^{H}\sum_{r=1}^{R}\sum_{\ell=1}^{L}\sum_{j=1}^{n_{kq}} \\
&P(W_k = h, Z_q = r, X_{jkq} = \ell|\boldsymbol{y}_{kq}) \log \pi_{\ell|hr} \\
&\qquad + \sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{h=1}^{H}\sum_{r=1}^{R}\sum_{\ell=1}^{L}\sum_{j=1}^{n_{kq}} \\
&P(W_k = h, Z_q = r, X_{jkq} = \ell|\boldsymbol{y}_{kq}) \sum_{i=1}^{I} \log \pi_{y_i = m_i|\ell}, \qquad (4)
\end{aligned}
$$

where $\pi_h = P(W_k = h)$, $\pi_r = P(Z_q = r)$, $\pi_{\ell|hr} = P(X_{jkq} = \ell|W_k = h, Z_q = r)$ and $\pi_{y_i = m_i|\ell} = P(Y_{ijkq} = m_i|X_{jkq} = \ell)$ and $\boldsymbol{\theta} = \{\pi_{\ell|hr}, \pi_h, \pi_r, \pi_{y_i = m_i|\ell}\}$. The E step involves the computation of the joint conditional posterior probabilities $P(W_k = h, Z_q = r, X_{jkq} = \ell|\boldsymbol{y}_{kq})$ given the current estimates of model parameters, which consist of $n_{kq} + 2$ latent variables for each group (k, q). Since these terms cannot be factorized due to the specific dependency structure of the CC higher level labels conditionally to the observations, standard EM algorithms cannot be directly applied here. Therefore, following the literature on latent block modeling (see Keribin et al. 2015), we propose using a stochastic version of the EM algorithm which involves the inclusion of a Gibbs sampling scheme between the E and the M step; that is, the SEM-Gibbs algorithm. The stochastic step consists in the consecutive sampling from the univariate marginal posterior distributions of a latent variable conditionally on the sampled values of the other latent variables, which reduces the computational burden. In the M step the parameters $\boldsymbol{\theta}$ are updated to maximize the complete log-likelihood. Stochastic versions of the EM (see Celeux and Diebolt 1986; Celeux et al. 1996) do not increase the log-likelihood at each iteration, but instead generate an irreducible Markov chain whose unique stationary distribution concentrates around the maximum of the likelihood. The implementation of a M step based on sampled labels avoids the need of analytical computation of the maximum of the expected log-likelihood, allowing to overcome intractable situations. It has also been proved that, given the fluctuation of the chain generated, such algorithms are less sensitive to initial values, thus reducing the risk of getting stuck in local maxima. We propose two versions of the stochastic algorithm: i) a full SEM-Gibbs, in which we sample the latent class memberships at both higher and lower level; ii) a hybryd SEM-Gibbs, in which the stochastic part is applied only at level-2, while at level-1 a standard E step is performed. This second version is possible due to the fact that, as in the hierarchical multilevel case (Vermunt 2008), we can consider the following partial factorization

$$
\begin{aligned}
&P(W_k = h, Z_q = r, X_{jkq} = \ell | \boldsymbol{y}_{kq}) \\
&\quad = P(W_k = h, Z_q = r | \boldsymbol{y}_{kq}) \\
&\quad P(X_{jkq} = \ell | \boldsymbol{y}_{kq}, W_k = h, Z_q = r) \\
&\quad = P(W_k = h, Z_q = r | \boldsymbol{y}_{kq}) \\
&\quad P(X_{jkq} = \ell | \boldsymbol{y}_{jkq}, W_k = h, Z_q = r),
\end{aligned}
$$

where we have applied the conditional independence of subjects (j) and their class membership given the memberships of the cross-classified groups.

3.1 Full SEM-Gibbs

SE step

After initialization of $\pi_h, \pi_r, \pi_{\ell|hr}, \pi_{y_i=m_i|\ell}$ iterate the following sampling steps

1) Draw $\boldsymbol{w}^{(t)}$ from a Multinomial distribution with probabilities

$$
\begin{aligned}
&P(W_k = h | \boldsymbol{y}_k, \boldsymbol{z}^{(t-1)}) \\
&\quad = \frac{\pi_h P(\boldsymbol{Y}_k = \boldsymbol{m}_k | \boldsymbol{z}^{(t-1)}, W_k = h)}{P(\boldsymbol{Y}_k = \boldsymbol{m}_k | \boldsymbol{z}^{(t-1)})}, \\
&P(\boldsymbol{Y}_k = \boldsymbol{m}_k | \boldsymbol{z}, W_k = h) \\
&\quad = \prod_{q_k=1}^{Q_K} \prod_{r=1}^{R} \left[\prod_{j=1}^{n_{kq}} P(\boldsymbol{Y}_{jkq} = \boldsymbol{m} | W_k = h, Z_q = r) \right]^{z_q^r};
\end{aligned}
$$

2) Draw $\boldsymbol{z}^{(t)}$ from a Multinomial distribution with probabilities

$$
\begin{aligned}
&P(Z_q = r | \boldsymbol{y}_q, \boldsymbol{w}^{(t)}) \\
&\quad = \frac{\pi_r P(\boldsymbol{Y}_q = \boldsymbol{m}_q | \boldsymbol{w}^{(t)}, Z_q = r)}{P(\boldsymbol{Y}_q = \boldsymbol{m}_q | \boldsymbol{w}^{(t)})}, \\
&P(\boldsymbol{Y}_q = \boldsymbol{m}_q | \boldsymbol{w}, Z_q = r) \\
&\quad = \prod_{k_q=1}^{K_Q} \prod_{h=1}^{H} \left[\prod_{j=1}^{n_{kq}} P(\boldsymbol{Y}_{jkq} = \boldsymbol{m} | W_k = h, Z_q = r) \right]^{w_k^h};
\end{aligned}
$$

3) Draw $\boldsymbol{x}^{(t)}$ from a Multinomial distribution with probabilities

$$
\begin{aligned}
&P(X_{jkq} = \ell | \boldsymbol{y}_{jkq}, \boldsymbol{w}^{(t)}, \boldsymbol{z}^{(t)}) \\
&\quad = \frac{\left[\pi_{\ell|hr} P(\boldsymbol{Y}_{jkq} = \boldsymbol{m} | X_{jkq} = \ell) \right]^{w_{jk}^h z_{jq}^r}}{P(\boldsymbol{Y}_{jkq} = \boldsymbol{m})},
\end{aligned}
$$

where w_k^h, z_q^r, w_{jk}^h and z_{jq}^r are all binary indicators of units' membership at different levels, in particular w_{jk}^h, z_{jq}^r are the expansion of higher level latent class indicators over the first level units j.

M step: update the parameters $\pi_h, \pi_r, \pi_{\ell|hr}, \pi_{y_i=m_i|\ell}$ using the information from the previous step

$$
\begin{aligned}
&\pi_h = \frac{\sum_{k=1}^{K} w_k^{h(t)}}{K}, \quad \pi_r = \frac{\sum_{q=1}^{Q} z_q^{r(t)}}{Q}, \\
&\pi_{\ell|hr} = \frac{\sum_{j=1}^{n} w_{jk}^{h(t)} z_{jq}^{r(t)} x_{jkq}^{\ell(t)}}{\sum_{j=1}^{n} w_{jk}^{h(t)} z_{jq}^{r(t)}}, \quad \pi_{y_i=m_i|\ell} = \frac{\sum_{j=1}^{n} x_{jkq}^{\ell(t)} y_{ijkq}^{m_i}}{\sum_{j=1}^{n} x_{jkq}^{\ell(t)}}.
\end{aligned}
$$

where x_{jkq}^{ℓ} is a binary indicator of units' membership at the lower level.

3.2 Hybrid SEM-Gibbs

SE step

1) After initialization of $\pi_h, \pi_r, \pi_{\ell|hr}, \pi_{y_i=m_i|\ell}$ run a Gibbs sampler

1.1) Draw $\boldsymbol{w}^{(t)}$ from a Multinomial distribution with probabilities

$$
\begin{aligned}
&P(W_k = h | \boldsymbol{y}_k, \boldsymbol{z}^{(t-1)}) = \frac{\pi_h P(\boldsymbol{Y}_k = \boldsymbol{m}_k | \boldsymbol{z}^{(t-1)}, W_k = h)}{P(\boldsymbol{Y}_k = \boldsymbol{m}_k | \boldsymbol{z}^{(t-1)})}, \\
&P(\boldsymbol{Y}_k = \boldsymbol{m}_k | \boldsymbol{z}, W_k = h) \\
&\quad = \prod_{q_k=1}^{Q_K} \prod_{r=1}^{R} \left[\prod_{j=1}^{n_{kq}} P(\boldsymbol{Y}_{jkq} = \boldsymbol{m} | W_k = h, Z_q = r) \right]^{z_q^r};
\end{aligned}
$$

1.2) Draw $\boldsymbol{z}^{(t)}$ from a Multinomial distribution with probabilities

$$
\begin{aligned}
&P(Z_q = r | \boldsymbol{y}_q, \boldsymbol{w}^{(t)}) = \frac{\pi_r P(\boldsymbol{Y}_q = \boldsymbol{m}_q | \boldsymbol{w}^{(t)}, Z_q = r)}{P(\boldsymbol{Y}_q = \boldsymbol{m}_q | \boldsymbol{w}^{(t)})}, \\
&P(\boldsymbol{Y}_q = \boldsymbol{m}_q | \boldsymbol{w}, Z_q = r) \\
&\quad = \prod_{k_q=1}^{K_Q} \prod_{h=1}^{H} \left[\prod_{j=1}^{n_{kq}} P(\boldsymbol{Y}_{jkq} = \boldsymbol{m} | W_k = h, Z_q = r) \right]^{w_k^h};
\end{aligned}
$$

2) Compute

$$
\begin{aligned}
&P(X_{jkq} = \ell | \boldsymbol{y}_{jkq}, \boldsymbol{w}^{(t)}, \boldsymbol{z}^{(t)}) \\
&\quad = \frac{\left[\pi_{\ell|hr} P(\boldsymbol{Y}_{jkq} = \boldsymbol{m} | X_{jkq} = \ell) \right]^{w_{jk}^h z_{jq}^r}}{P(\boldsymbol{Y}_{jkq} = \boldsymbol{m})},
\end{aligned}
$$

M step: update the parameters $\pi_h, \pi_r, \pi_{\ell|hr}, \pi_{y_i=m_i|\ell}$ using the information from the previous step

$$
\pi_h = \frac{\sum_{k=1}^{K} w_k^{h(t)}}{K}, \quad \pi_r = \frac{\sum_{q=1}^{Q} z_q^{r(t)}}{Q},
$$

$$\pi_{\ell|hr} = \frac{\sum_{j=1}^{n} w_{jk}^{h(t)} z_{jq}^{r(t)} P(X_{jkq} = \ell | \boldsymbol{y}_{jkq}, \boldsymbol{w}^{(t)}, \boldsymbol{z}^{(t)})}{\sum_{j=1}^{n} w_{jk}^{h(t)} z_{jq}^{r(t)}},$$

$$\pi_{y_i=m_i|\ell} = \frac{\sum_{j=1}^{n} P(X_{jkq} = \ell | \boldsymbol{y}_{jkq}, \boldsymbol{w}^{(t)}, \boldsymbol{z}^{(t)}) y_{ijkq}}{\sum_{j=1}^{n} P(X_{jkq} = \ell | \boldsymbol{y}_{jkq}, \boldsymbol{w}^{(t)}, \boldsymbol{z}^{(t)})}.$$

3.3 Estimation conditions

Final Parameters Estimation. Final estimates $\hat{\boldsymbol{\theta}}$ of $\{\pi_{\ell|hr}, \pi_h, \pi_r, \pi_{y_i=m_i|\ell}\}$ are calculated as the mean over the total number of iterations, burn-in period excluded. Numerical experiments have proved that it is sufficient to consider a single iteration over the Gibbs sampler for each SE step.

Classification step. A sample of $(\boldsymbol{w}, \boldsymbol{z}, \mathbf{x})$ is generated by several SE steps with $\boldsymbol{\theta}$ fixed to $\hat{\boldsymbol{\theta}}$. The final classification $(\hat{\boldsymbol{w}}, \hat{z}, \hat{\boldsymbol{x}})$ is estimated by the mode of their sampling distribution. In the hybrid SEM-Gibbs version for $(\hat{\boldsymbol{x}})$ the mode of the posterior probability is taken.

Initialization. $\boldsymbol{w}$ and $\boldsymbol{z}$ are randomly initialized with the multinomial distribution with probabilities $(\frac{1}{H}, \ldots, \frac{1}{H})$ and $(\frac{1}{R}, \ldots, \frac{1}{R})$ respectively. The level-1 mixing proportions $\pi_{\ell|hr}$ are initialized to $(\frac{1}{L}, \ldots, \frac{1}{L})$ for all $h = 1, \ldots, H$ and $r = 1, \ldots, R$, and the multinomial response probabilities $\pi_{y_i=m_i|\ell}$ to the maximum likelihood estimates obtained with a standard LC model. It should be noted that we tested various initialization strategies in several simulation settings. However, the approach here presented turned out to be the most appropriate to prevent local maxima.

Identifiability. While the intractability of the likelihood makes it hard or even impossible to come up with necessary and sufficient conditions for (local) identifiability, we can still say something about parameter identification. For the lower-level, standard results on the identification of LC models apply, such as that the Jacobian should be full rank. As shown by Vermunt (2005) and Bennink et al. (2016), similar conditions apply to the higher level of multilevel LC models. Basically, the second-level parameters are identified if the number of lower-level units within the higher-level unit is large enough given the number of higher-level latent classes. Our conjecture is that the higher-level parameters of the cross-classified model are identified if the separate models for the two nestings are identified. Note that in typical applications, the number of lower-level units in each of the two groupings will be large enough, and identification is then not an issue. Unlike the hierarchical version of the model, identifiability assessment based on the Jacobian cannot be applied as the likelihood is not available.

4 Simulation study

A simulation study was conducted to evaluate the performance of the proposed model in combination with the stochastic EM methods when applied to categorical response data. Various scenarios were examined, (see Sect. 4.1) which differed among others in terms of the level-1 and level-2 class separation. The design of numerical experiments is based on an entropy based R^2 measure that allows to quantify the degree of class separation and determine how well latent class memberships can be predicted using the posterior class membership probabilities (see Magidson 1981 and Vermunt and Magidson 2016). Its value ranges between 0 and 1, where 0 indicates no separation and 1 perfect separation.

In particular, as done in Lukočiene et al. (2010) for the hierarchical latent class model, we provide a new definition of R^2 entropy for each level of the CC data structure.

The R^2_{Entropy} is defined separately for lower and higher latent level classes. For level-1 membership we have:

$$\mathrm{R}^2_{\text{Entropy},\ell ow} = 1 - \frac{\sum_{j=1}^{n} \sum_{\ell=1}^{L} -P(X_{jkq} = \ell | \boldsymbol{y}_{jkq}) \log(P(X_{jkq} = \ell | \boldsymbol{y}_{jkq}))}{n \sum_{\ell=1}^{L} -P(X_{jkq} = \ell) \log(P(X_{jkq} = \ell))}$$

with

$$P(X_{jkq} = \ell) = \sum_{h=1}^{H} \sum_{r=1}^{R} \pi_{\ell|hr} \pi_h \pi_r$$

and

$$P(X_{jkq} = \ell | \boldsymbol{y}_{jkq}) = \frac{\prod_{i=1}^{I} P(Y_{ijkq} = m_i | X_{jkq} = \ell) P(X_{jkq} = \ell)}{\sum_{\ell=1}^{L} \prod_{i=1}^{I} P(Y_{ijkq} = m_i | X_{jkq} = \ell) P(X_{jkq} = \ell)}.$$

For level-2 we can compute a cross-classified entropy, but also a separate measure for each of the two sets of latent classes. In particular, for the total cross-classified entropy we take:

$$R^2_{\text{Entropy},CC} = 1 - \frac{\sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{h=1}^{H} \sum_{r=1}^{R} -P(W_k = h, Z_q = r | \boldsymbol{y}_{kq}) \log(P(W_k = h, Z_q = r | \boldsymbol{y}_{kq}))}{C \sum_{h=1}^{H} \sum_{r=1}^{R} -P(W_k = h, Z_q = r) \log(P(W_k = h, W_q = r))}$$

with C the number of KQ cross-classified units in the data. We also define separate entropy measures for W and Z

$$\mathrm{R}^2_{\text{Entropy},W_k} = 1 - \frac{\sum_{h=1}^{H} \sum_{k=1}^{K} -P(W_k = h | \boldsymbol{y}_k) \log(P(W_k = h | \boldsymbol{y}_k))}{K \sum_{h=1}^{H} -P(W_k = h) \log(P(W_k = h))}$$

and

$$\mathrm{R}^2_{\mathrm{Entropy},Z_q} = 1 - \frac{\sum_{r=1}^{R}\sum_{q=1}^{Q} -P(Z_q = r|\boldsymbol{y}_q)\log(P(Z_q = r|\boldsymbol{y}_q))}{Q\sum_{r=1}^{R} -P(Z_q = r)\log(P(Z_q = r))}.$$

We compute all the R^2 entropy measures described above using the true values of parameters $\boldsymbol{\theta}$ to obtain an a priori degree of class separation for the data we simulate. In particular, the joint probability $P(W_k = h, Z_q = r|\boldsymbol{y}_{kq})$ is estimated as the proportion of co-occurrences along the Gibbs chain obtained when iterating the SE step with true values of parameters $\boldsymbol{\theta}$.

4.1 Design

The factors varied in the simulation study were the number of level-1 and level-2 classes, the number of observations within each level-2 unit, the number of level-2 units, and the mixing proportions. Both algorithms were applied with the datasets simulated.

The simulation settings take 6 categorical variables, two binary, two with 3 modalities and two with 4 modalities. All simulations were set in order to have a medium separation at level-1 ($R^2_{\mathrm{Entropy},\ell ow} \approx 0.65$), fixing the number of level-1 latent classes to 4 ($L = 4$). The probability parameters for the two binary variables were $\pi_{y_1=1|\ell} = (0.2, 0.1, 0.7, 0.9)$, and $\pi_{y_2=1|\ell} = (0.3, 0.2, 0.8, 0.7)$. For the two variables of 3 modalities we set

$$\pi_{y_3|\ell} = \begin{pmatrix} 0.7 & 0.80 & 0.1 & 0.15 \\ 0.1 & 0.05 & 0.7 & 0.65 \\ 0.2 & 0.15 & 0.2 & 0.20 \end{pmatrix}, \quad \pi_{y_4|\ell} = \begin{pmatrix} 0.80 & 0.1 & 0.7 & 0.2 \\ 0.05 & 0.7 & 0.1 & 0.7 \\ 0.15 & 0.2 & 0.2 & 0.1 \end{pmatrix}.$$

The last two variables have probability parameters

$$\pi_{y_5|\ell} = \begin{pmatrix} 0.65 & 0.10 & 0.6 & 0.2 \\ 0.10 & 0.65 & 0.1 & 0.6 \\ 0.10 & 0.10 & 0.2 & 0.1 \\ 0.15 & 0.15 & 0.1 & 0.1 \end{pmatrix}, \quad \pi_{y_6|\ell} = \begin{pmatrix} 0.75 & 0.20 & 0.7 & 0.10 \\ 0.05 & 0.60 & 0.1 & 0.70 \\ 0.05 & 0.05 & 0.1 & 0.15 \\ 0.15 & 0.15 & 0.1 & 0.05 \end{pmatrix}.$$

Scenario 1 In the first scenario we focused on level-1 characteristics by assuming an almost perfect class separation at level-2 (near 1 in terms of $R^2_{\mathrm{Entropy},CC}$). More specifically, we assumed two latent classes at each of the two cross-classified higher levels ($H = 2$, $R = 2$), with mixing proportions $\pi_h = (0.625, 0.375)$ and $\pi_r = (0.25, 0.75)$, and sample sizes $K = 24$ and $Q = 36$. At level-2, data were generated according to a stratified sampling, constraining the level-2 mixing proportions to their nominal value. The matrix of lower-level mixing proportions $\pi_{\ell|hr}$ was set to

$$\pi_{\ell|hr} = \begin{pmatrix} 0.4 & 0.15 & 0.05 & 0.30 \\ 0.3 & 0.05 & 0.55 & 0.05 \\ 0.2 & 0.35 & 0.15 & 0.50 \\ 0.1 & 0.45 & 0.25 & 0.15 \end{pmatrix}.$$

Table 1 Average error in the estimate of mixing conditional proportions $\hat{\pi}_{\ell|11}$ at level-1. Scenario 1, true values $\pi_{\ell|11} = (0.4, 0.3, 0.2, 0.1)$. These values are shown for all the combinations of algorithm version (full/hybrid) and sample size ($n_{kq} = 13/n_{kq} = 24$) in the rows and for the four lower level latent classes (ℓ) in the columns

$\hat{\pi}_{\ell\|11}$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
full, $n_{kq} = 13$	−0.0049	0.0012	0.0033	0.0004
hybrid, $n_{kq} = 13$	0.0006	0.0006	−0.0013	0.0001
full, $n_{kq} = 24$	−0.0003	0.0006	−0.0014	0.0011
hybrid, $n_{kq} = 24$	−0.0012	−0.0003	0.0011	0.0003

At level-1 we varied the number of units considering two conditions:

1. $n_{kq} = 13$ with a total of $n = 11232$
2. $n_{kq} = 24$ with a total of $n = 20736$.

Scenario 2 In the second scenario we lowered the separation at level-2 to around 0.85 in terms of total $R^2_{\mathrm{Entropy},CC}$, with separate $R^2_{\mathrm{Entropy},W_k} \approx 0.66$, and $R^2_{\mathrm{Entropy},Z_q} \approx 1$. This was achieved by increasing the number of latent classes at this level, taking $H = 3$ and $R = 2$. The mixing proportions set were $\pi_h = (0.20, 0.30, 0.50)$ and $\pi_r = (0.40, 0.60)$ and the sample sizes were $K = 50$ and $Q = 12$. At level-1 the total number of units was $n = 3000$, with $n_{kq} = 5$ for each combined level-2 unit. The matrix of level-1 mixing proportions $\pi_{\ell|hr}$ was set to

$$\pi_{\ell|hr} = \begin{pmatrix} 0.40 & 0.15 & 0.15 & 0.30 & 0.30 & 0.05 \\ 0.30 & 0.10 & 0.35 & 0.10 & 0.35 & 0.25 \\ 0.20 & 0.35 & 0.20 & 0.40 & 0.15 & 0.40 \\ 0.10 & 0.40 & 0.30 & 0.20 & 0.20 & 0.30 \end{pmatrix}.$$

In the data generation process, given the small number of units at level-2, we explored two different simulation schemes:

i. random sampling of memberships of level-2 units with probabilities π_h and π_r;
ii. stratified sampling of memberships of level-2 units constraining the proportion of units to their nominal probabilities of π_h and π_r.

All computations were implemented in C++, and performed with the aid of the Rcpp R package (Eddelbuettel et al. 2024). Experimental observations proved that the algorithm implemented is scalable; that is, the running time increases almost linearly with increasing total sample size n. The C++ code, together with the R code to generate data and perform simulations, is made available on a Github repository at the link https://github.com/NicolaPiras97/MCCLC.

Table 2 Average error in the estimate of $\hat{\pi}_{y_1=1|\ell}$ probability parameters. Scenario 1, true values $\pi_{\ell|11} = (0.2, 0.1, 0.7, 0.9)$. These values are shown for all the combinations of algorithm version (full/hybrid) and sample size ($n_{kq} = 13/n_{kq} = 24$) in the rows and for the four lower level latent classes (ℓ) in the columns

$\hat{\pi}_{y_1=1\|\ell}$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
full, $n_{kq} = 13$	−0.00001	0.00007	0.00117	0.00038
full, $n_{kq} = 24$	0.00032	0.00244	−0.00056	−0.00027
hybrid, $n_{kq} = 13$	−0.03206	−0.00118	−0.01248	0.04209
hybrid, $n_{kq} = 24$	0.00003	−0.00061	−0.00093	0.00005

4.2 Results

Scenario 1 In the first scenario we evaluated the finite estimates' performances of both the full and the hybrid SEM-Gibbs algorithm using 100 simulated datasets. The required number of iterations was determined by inspecting the evolution of the simulated chains. Convergence was achieved with 50 burn-in iterations and another 100 iterations to obtain the parameter estimates. Figure 4 shows the chain evolution of a mixing conditional proportion parameter at level-1 and a probability parameter for one of the datasets simulated when $n_{kq} = 13$. As can be seen, with both versions of the algorithm, the chain reaches sufficient stability after few iterations. It is important to observe that the chains arising from the hybrid version of the Gibbs have no-fluctuation along the iterations, which is a consequence of the closed form estimation of lower-level parameters. The computing time was comparable for the hybrid and full version of the algorithm. In particular, in setting 1. with $n_{kq} = 13$, estimation on 100 datasets took 19 and 21 min for the hybrid and full version, respectively, while in setting 2. it took 37 and 39 min.

To evaluate the bias of the estimates we looked at the average deviation from the true values across replications (see Tables 1 and 2) and the sampling distribution of this deviation (see Figs. 2 and 3). We observe that, at level-1 of the structure, the performance is good in all scenarios considered for both versions of the SEM-Gibbs algorithm. Similar results are observed for the other parameter estimates.

The classification performance at level-1 was investigated by computing percentage of correctly classified units across simulations. For the condition with $n_{kq} = 13$ combined with the application of the full SEM-Gibbs algorithm, we observed an average of 84% classification accuracy of level-1 units across level-2 latent classes, with a minimum of 81% for $h = 1$ and $r = 2$ and a maximum of 87% for $h = 2$ and $r = 1$. Instead, the classification at the CC levels is perfect by construction, as explained in the simulation design.

Scenario 2 In the second scenario we have extensively investigated the performance of the hybrid SEM-Gibbs approach using 100 simulated datasets, and focused on higher-level parameters of our CC LC model. After 200 burn-in iterations, we performed another 600 iterations, and thinned the chain (we keep one iteration every 3) to reduce autocorrelations. The sample size used to obtain the final estimates is then 200. Considering all the chains generated, we computed an average integrated autocorrelation time of 2.3, and an average effective sample size of 100. The computing time on 100 simulated datasets was about 25 min. Average errors for $\hat{\pi}_h$ are available in Table 3 and error distributions are summarized in Fig. 5. We observe similar average errors for π_h under both data generation schemes, the random sampling of class labels (i.) and the stratified (and thus constrained) sampling (ii.). In the estimates of π_r, which correspond to level-2 latent classes with a higher separation and a small number of units, the constrained generative scheme produced no-error, as could be expected. When looking at classification accuracy under random generating sampling scheme (i.), we observed perfect classification across the $R = 2$ level-2 latent classes associated to the latent variable Z, while the percentage of well classified units across the $H = 3$ latent classes associated to the latent variable W was 89%.

In this scenario we often encountered label switching at level-2. This was overcome by imposing an ordering constraint ex post for the final estimation of level-2 mixing-proportions π_h and π_r (see Marin et al. 2005). An example of chain evolution for all levels of the structure is available in Fig. 6. The scenarios investigated guaranteed a convergence to local maxima. However, with lower degree of separation the problem of encountering spurious solutions is more likely, and the implementation of alternative initialization strategies may help in avoiding them.

5 Application

We applied our method to the analysis of data coming from a survey on the profile of Italian university graduates. Data are publicly available through the Almalaurea Interuniversity Consortium, a system which provides a continuous monitoring of 78 Italian universities. Their surveys supply an overview of graduates characteristics, giving access to infor-

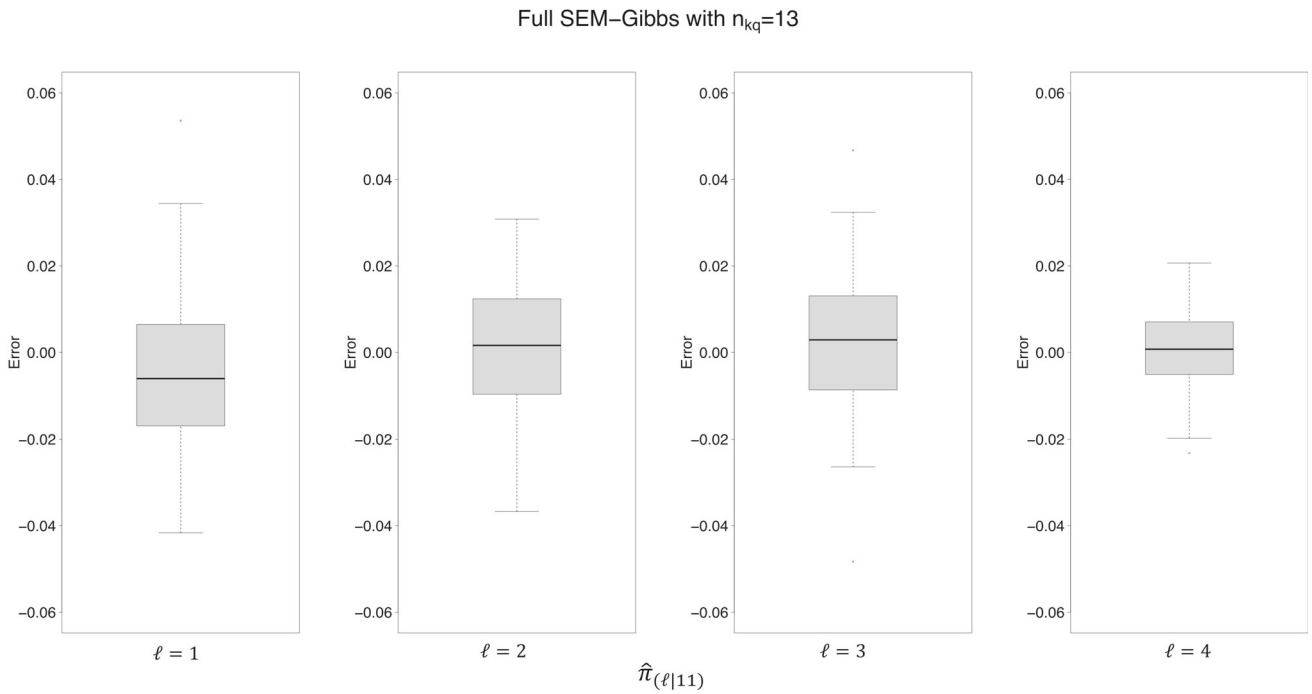

Fig. 2 Error distribution of the estimates of level-1 conditional mixing proportion $\hat{\pi}_{\ell|11}$ over 100 simulations. Estimates were obtained through the application of the full SEM-Gibbs algorithm when $n_{kq} = 13$. True values $\pi_{\ell|11} = (0.4, 0.3, 0.2, 0.1)$

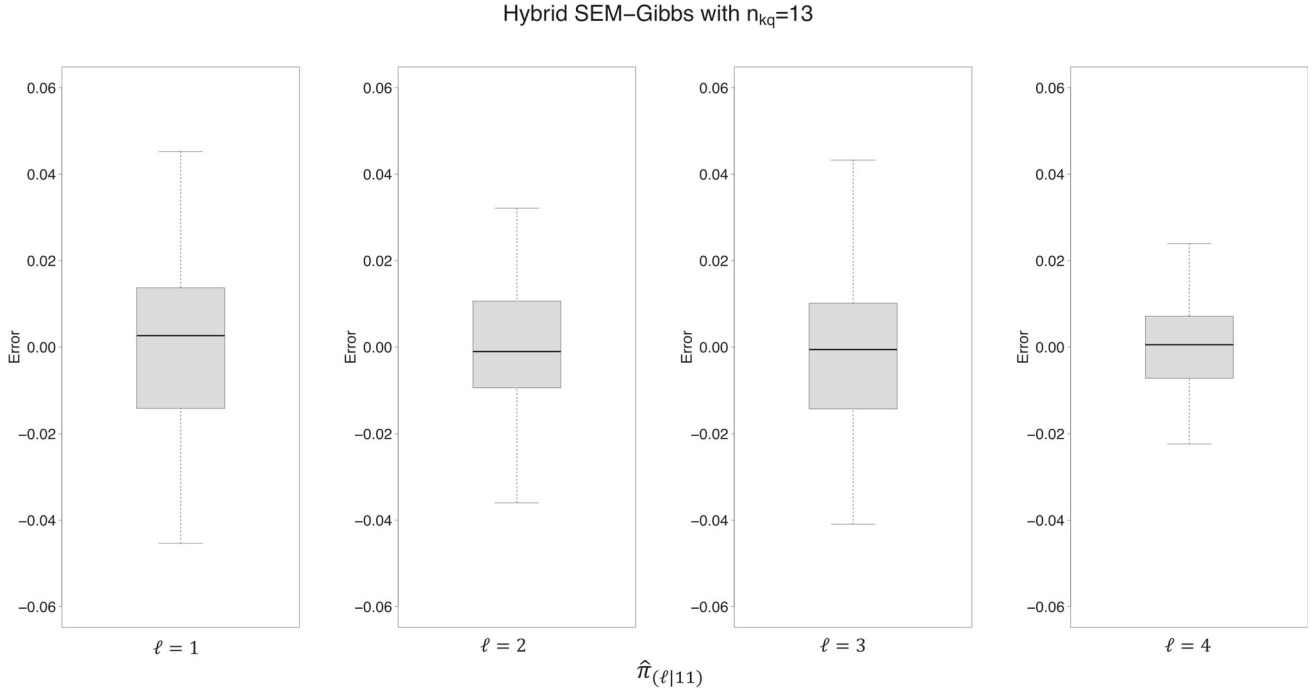

Fig. 3 Error distribution of the estimates of level-1 conditional mixing proportion $\hat{\pi}_{\ell|11}$ over 100 simulations. Estimates were obtained through the application of the hybrid SEM-Gibbs algorithm when $n_{kq} = 13$. True values $\pi_{\ell|11} = (0.4, 0.3, 0.2, 0.1)$

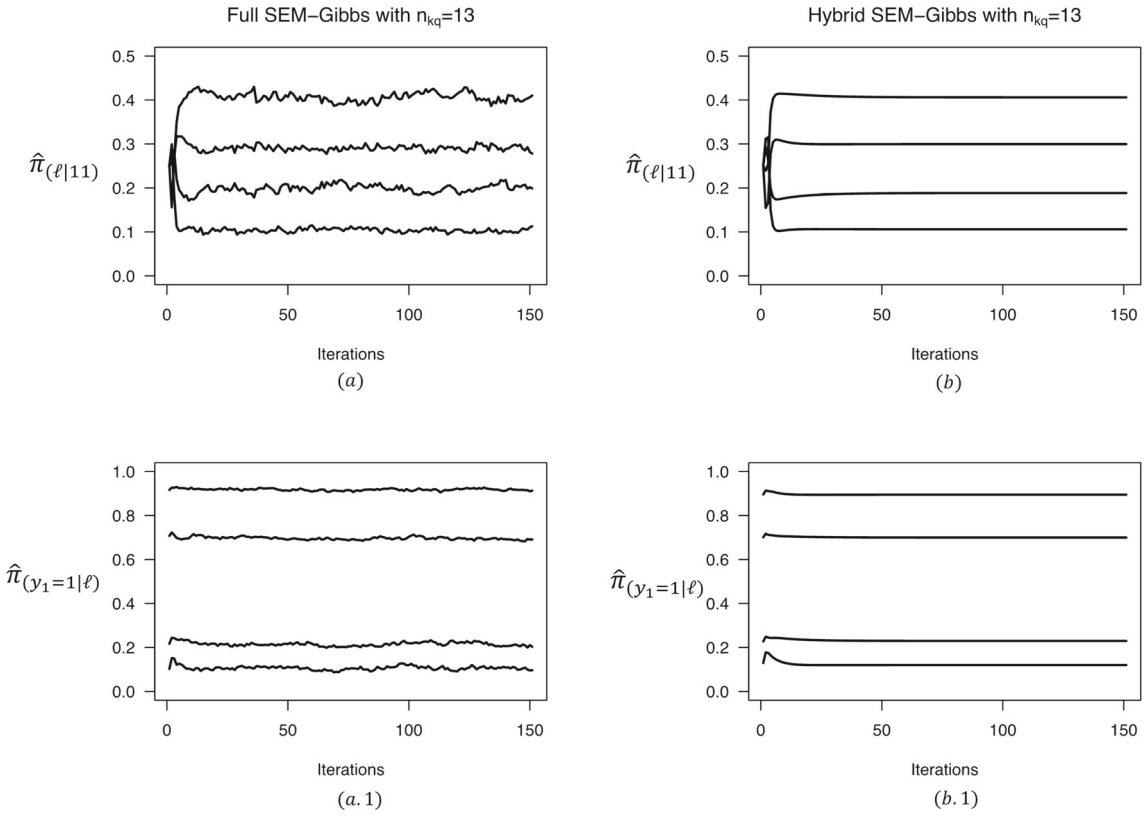

Fig. 4 Evolution of the SEM-Gibbs chains over 150 iterations. Panel (**a**) and (**a.1**) refer to the application of the full version of the estimation algorithm, while panels (**b**) and (**b.1**) to the hybrid one. True values of parameters are $\pi_{\ell|11} = (0.4, 0.3, 0.2, 0.1)$ and $\pi_{y_1=1|\ell} = (0.2, 0.1, 0.7, 0.9)$

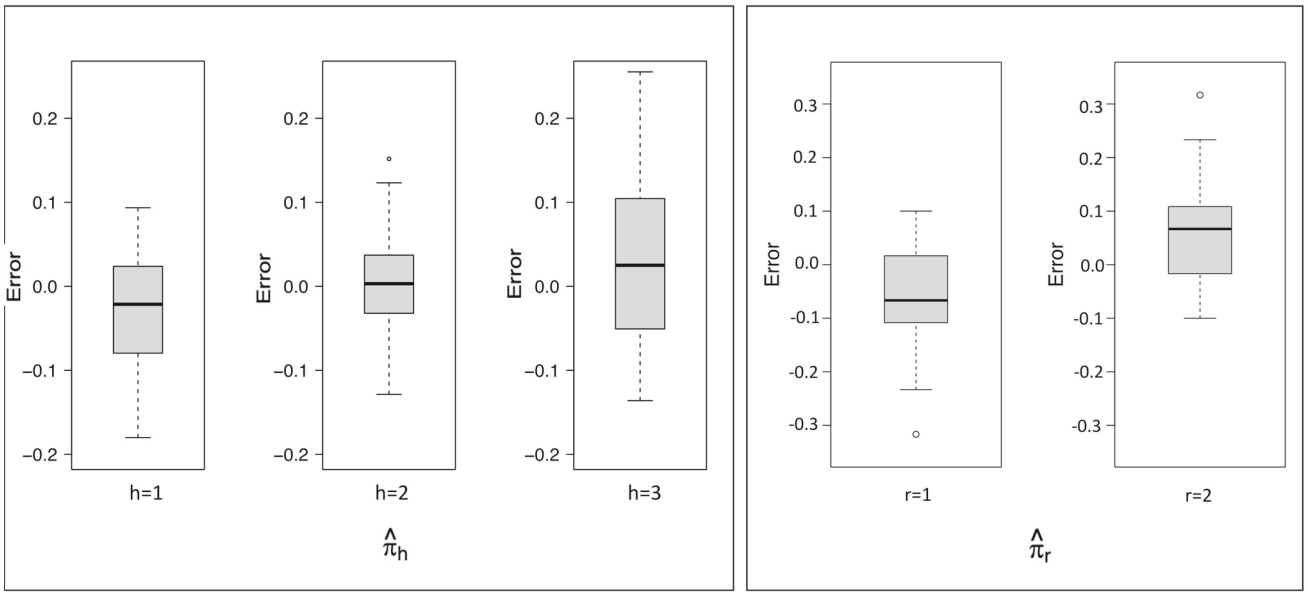

Fig. 5 Error distribution of level-2 mixing proportions ($\hat{\pi}_h$ and $\hat{\pi}_r$) in scenario 2 under generative scheme (i)

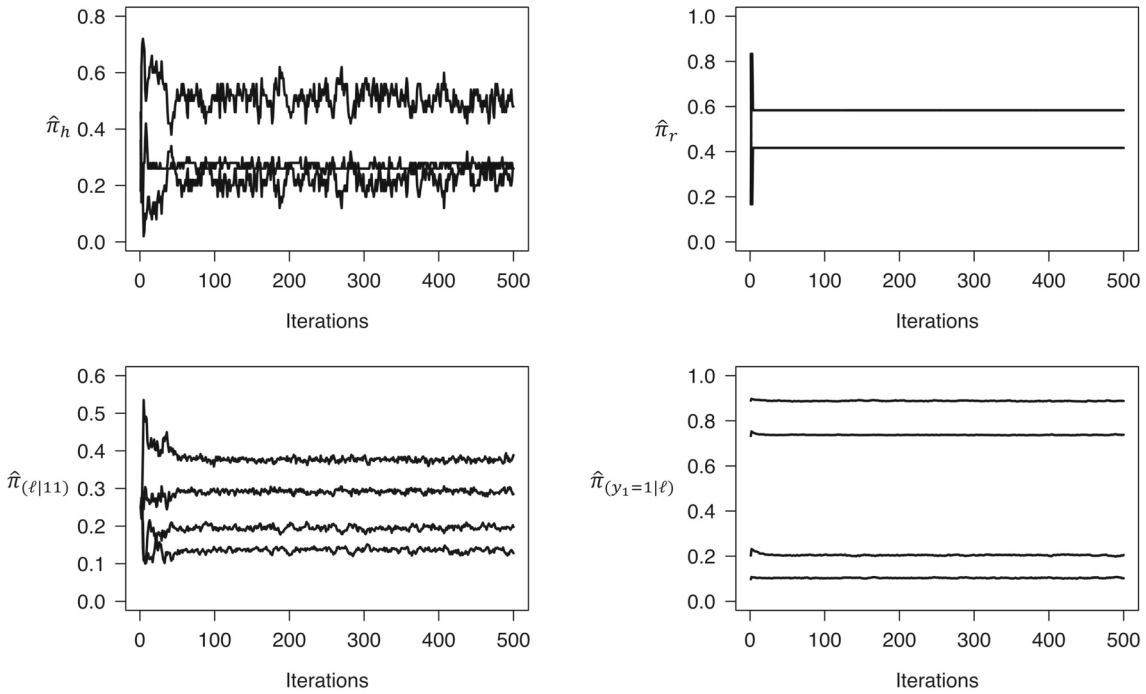

Fig. 6 Evolution of the SEM-Gibbs chains at all levels of the cross-classified data structure trend in scenario 2, under generative scheme (i). True values of parameters are $\pi_h = (0.2, 0.3, 0.5)$, $\pi_r = (0.4, 0.6)$, $\pi_{\ell|11} = (0.4, 0.3, 0.2, 0.1)$ and $\pi_{y_1=1|\ell} = (0.2, 0.1, 0.7, 0.9)$

Table 3 Average error in the estimate of mixing proportions of level-2 classes, $\hat{\pi}_h$ and $\hat{\pi}_r$. These results refer to the second scenario for both random (i.) and constrained (ii.) sampling of memberships of level-2 units. True values $\pi_h = (0.2, 0.3, 0.5)$, $\pi_r = (0.4, 0.6)$

$\hat{\pi}_h$	$h = 1$	$h = 2$	$h = 3$
(i.)	−0.0339	−0.0001	0.0339
(ii.)	−0.0178	−0.0004	0.0183
$\hat{\pi}_r$	$r = 1$	$r = 2$	
(i.)	−0.0414	0.0414	-
(ii.)	0	0	-

mation on their university achievements, the experiences they have gained during their studies and the evaluation of the studies they have completed. In this application we used seven indicators on the quality of the university experience as perceived by bachelor degree graduates. The information is available at different aggregation levels. We treated the degree programme as the first level which is simultaneously nested within a combination of the level-2 units university and field of study. We selected 38 Italian bachelor programmes in 10 fields of study according to the "International Standard and Classification of Education: Field of education and training" (ISCED-F 2013, see UNESCO 2014). The ISCED-F 2013 is an international classification system, implemented since 2016, that considers a classification of education programmes on the basis of the similarities between disciplinary contents.

Table 4 Variables description

Variable	Description	Frequencies		
		1	2	3
y_1: SATISF-PROF	1: full satisfaction; 2: weak satisfaction or insatisfaction	1079	39	
y_2: SATISFCOL	1: full satisfaction; 2: weak satisfaction or insatisfaction	976	313	
y_3: SATISFCLASSROOMS	1: always adequate; 2: often adequate; 3: rarely/never adequate	155	876	258
y_4: SATISFPCLAB	1: adequate number, 2: inadequate number	636	653	
y_5: SATISFLIB	1: positive; 2: quite positive; 3: negative	303	971	15
y_6: SATISFSTUDYROOMS	1: adequate; 2: inadequate	858	431	
y_7: ADEQSTUDY	1: full positive; 2: positive; 3: negative	132	1099	58

In Frequencies column, for each variable and for each category, is reported the absolute frequency of each modality

We used the survey of bachelor students graduated in 2017 (see Almalaurea 2018). After some preliminary data cleaning, removing degree programmes with a small number of students for which information was not accessible, our dataset consisted of 1289 degree programmes nested within 72 universities and 10 ISCED fields of study. The seven indicators used are: (1) their satisfaction with respect to human relationships with professors (SATISF-PROF), (2) their satisfaction with respect to relationships with other students (SATISFCOL), (3) evaluation of classrooms (SATISFCLASSROOMS), (4) evaluation of computer science labs (SATISFPCLAB), (5) evaluation of library services and accessibility (SATISFLIB), (6) evaluation of spaces available where to study (SATISFSTUDYROOMS), (7) evaluation of the adequateness of study workload relatively to the duration of the degree programme (ADEQSTUDY). While the original variables were expressed as percentage of preferences for each of the possible multiple choices associated to each item, we used a categorized version of them by associating to each degree programme the most frequent option. Categories with a frequency smaller than 5 over 1289 units were combined with the closest one. A description of final variables is summarized in Table 4.

When it comes to University enrollment, students make their choice after careful evaluation of multiple aspects that include previous experience reported from other undergraduates, and University rankings. Official rankings usually do not take into account the high heterogeneity in the educational offer within and between universities, and are mainly based on aggregated information at university level. The application of the MCCLC model to the indicators of perceived quality, aims to produce a classification of degree programmes that can be differentiated with respect to the institutional vocation. The model provides a clustering of degree programmes within the combination of universities and fields of study, but also two separate clusterings of universities and fields of study.

We fit a model with 3 classes at the degree programme level (level-1), and 2 classes for each of the second cross-classified levels (University and Field of study), with a total of four combined level-2 classes. The number of classes at lower level has been selected considering the Bayesian Information Criterion (BIC) in the estimate of a standard latent class model, ignoring the nesting structure. The BIC values obtained letting vary the number of level-1 L classes are in Table 5, the number of classes suggested is $L = 3$. The choice of the number of level-2 CC classes requires the computation of suitable selection measures, that involve the double nesting structure. Indeed, the application of standard information criteria cannot be directly applied as, when the cross-classification is considered the likelihood value is not available in closed form. Possible candidate measures should involve the finite approximations of the likelihood.

Table 5 BIC values for different choices of L, number of level-1 latent classes

	$L = 2$	$L = 3$	$L = 4$
BIC	9745.02	9670.86	9745.02

This aspect needs a full discussion that will be covered in a specific work.

Given the number of L, H and R classes, a labelling has been assigned on the basis of $\hat{\pi}_{y_i=m_i|\ell}$ estimates for the seven indicators and their aggregate values at level-2. At degree programme level we have found a class of Low, one of Medium and one of High satisfaction, whereas, taking the combination of University and field of study we have the ordered classes of Low, Medium, Medium-High and High satisfaction. A graphical synthesis of results obtained can be given through a bivariate barplot reporting the proportion of degree programmes classified within the combination of university and Field of study classes, see Fig. 7. The proportions depicted in Fig. 7 are obtained using the final classification as explained in Sect. 3.3. Parameter estimates are summarized in Tables 6 and 7, where respectively mixing proportions and density parameters, corresponding to conditional probabilities for each categorical variable, are reported. The values obtained are in agreement with the final classification, in fact the class of Medium satisfaction at University-Fields of study level is on average composed of degree programmes classified as Medium. The cross-classified class of Low satisfaction of degree programmes with Low satisfaction and so on.

The benefit of our approach compared to the hierarchical latent class model, is that we are able to provide a classification in terms of quality perceived either of both universities and fields of study, obtaining at the same time a classification of degree programmes within each combination ($H \cdot R = 4$) of universities and fields of study clusters. The application presented has some limitations, the data used are a categorized version of numerical ones resulting in modalities which are under represented, and moreover the ordinal nature of the categories has been ignored in model formulation. These limitations can be easily overcome as the proposed modeling framework can be adapted to other distributional assumptions (see for instance D'Elia and Piccolo 2005; Jacques and Biernacki 2018). Albeit some weaknesses in the nature of the data, the application presented is suitable to provide a complete illustration of the potentials of the method proposed.

6 Conclusions and discussion

This paper proposed an extension of LC analysis which can be used to handle multilevel cross-classified data structures.

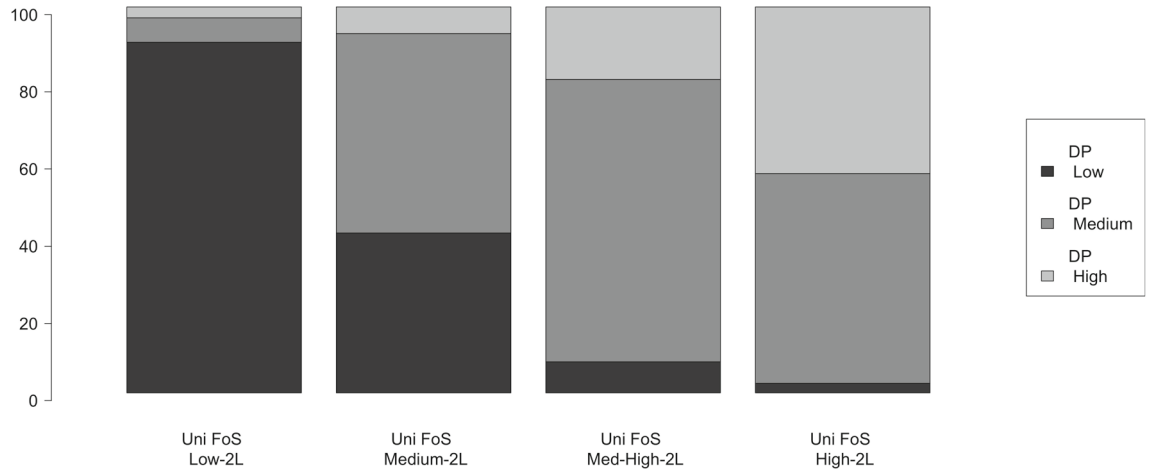

Fig. 7 Classification of degree programmes (DP) within the combination of university (Uni) and field of study (FoS) classes

Table 6 Class probabilities estimated at each level of the structure. Labels to classes have been assigned based on parameter estimates for the seven indicators

	$\ell = 1$ (Highest)	$\ell = 2$ (Medium)	$\ell = 3$ (Low)
$h = 1, r = 1$ (Medium)	0.11	0.48	0.41
$h = 1, r = 2$ (Low)	0.03	0.09	0.88
$h = 2, r = 1$ (High)	0.43	0.53	0.04
$h = 2, r = 2$ (Medium-High)	0.24	0.64	0.12
	$h = 1$	$h = 2$	
	0.21	0.79	
	$r = 1$	$r = 2$	
	0.10	0.90	

Table 7 Estimates of conditional probabilities for the 7 categorical indicators

	$\ell = 1$	$\ell = 2$	$\ell = 3$
$\hat{\pi}(y_1 = 1\|\ell)$	0.1721	0.0004	0.0041
$\hat{\pi}(y_2 = 1\|\ell)$	0.8525	0.6888	0.8115
$\hat{\pi}(y_3 = 1\|\ell)$	0.4963	0.0309	0.0108
$\hat{\pi}(y_3 = 2\|\ell)$	0.4828	0.8963	0.4321
$\hat{\pi}(y_4 = 1\|\ell)$	0.8955	0.5578	0.0895
$\hat{\pi}(y_5 = 1\|\ell)$	0.6759	0.1741	0.0282
$\hat{\pi}(y_5 = 2\|\ell)$	0.3164	0.8256	0.9370
$\hat{\pi}(y_6 = 1\|\ell)$	0.9217	0.7728	0.2898
$\hat{\pi}(y_7 = 1\|\ell)$	0.3709	0.0253	0.0480
$\hat{\pi}(y_7 = 2\|\ell)$	0.6098	0.9477	0.8562

The illustrative real data application involved the clustering of Italian universities and fields of study based on the clustering of degree programmes.

To solve the estimation problem, we proposed two variants of a stochastic EM algorithm. The method presented is an extension of the LC model Vermunt (2003) proposed for nested structures and integrates estimation procedures applied in the co-clustering context with latent-block models (see Biernacki et al. 2023). These algorithms were tested in simulation studies in which the generative schemes were defined to give different levels of data separation both at the lower and the higher level of the multilevel structure. Simulation studies showed that the proposed stochastic EM algorithms provide good tools for handling the rather complex structure of cross-classified data, characterised by dependencies across and within levels.

Though the new method was presented for categorical responses, it can be easily extended to consider other data types, including mixed categorical and continuous responses. Future developments involve the investigation of model selection procedures, and the definition of criteria to verify the algorithm convergence. For what concerns model selection, it should be noted that Information Criteria based on penalization of the log-likelihood cannot be directly applied because it is not possible to compute the log-likelihood value. However, it may be possible to use results from co-clustering on information criteria based on asymptotic approximations (see Keribin et al. 2015). Moreover, Columbu et al. (2024) presented preliminary results obtained with approximations of the log-likelihood value using output from the stochastic EM algorithm. Another issue to be addressed involves the tailoring of the model selection strategy to the multilevel

data structure, which involves a simultaneous decision on the number of latent classes at multiple levels. For multilevel nested LC models Lukočiene et al. (2010) presented a three-step model selection procedure using different definitions of Information Criteria for each level of the hierarchical structure. A similar stepwise approach may be developed for the cross-classified version of the multilevel model.

A limitation of the proposed new method is that, as common for mixture models (Wu 1983), the convergence to global maximum of the likelihood cannot be guaranteed. In particular, with lower separation and small number of units, we have observed higher rates of convergence to local maxima. In such situations multiple stochastic EM runs must be performed to ensure convergence, possibly with multiple starting points strategy (see among others Biernacki et al. 2003; Maruotti and Punzo 2021). An alternative approach that is worth investigating is regularization using full Bayesian estimation, where non-informative Dirichlet priors for mixing proportions and distribution parameters can be specified. In the co-clustering context (Biernacki et al. 2023) it has been observed that such an approach prevents spurious solutions and empty clusters, and we believe that similar benefits could be encountered in the multilevel cross-classified model.

We did not investigate the estimation performances of the cross-classified LC model in the presence of missing data. Recently, there has been an increasing interest in dealing with this aspect in the frame of model-based clustering or co-clustering. In recent contributions (see Selosse et al. 2020; Sportisse et al. 2024) this issue is overcome taking advantage of the typical augmented EM formulation and including missing-data patterns in the definition of the complete data matrix. We believe that similar strategies could be also integrated in the estimation procedure proposed for the cross-classified multilevel LC.

Another possible extension of this work involves the inclusion of covariates. In fact, when applying LC models, researchers are often interested in relating the class membership to possible explanatory variables. The literature (see among others Vermunt 2010; Di Mari et al. 2023) suggests that in these situations the most appropriate estimation approach is to consider stepwise estimators in which the latent class model for the observed responses is implemented separately from the estimation of the regression models for the latent variables given the covariates. We think that an analogous strategy should be investigated for the cross-classified LC model, especially because the implementation of a one-step approach would further increase the computational complexity.

Acknowledgements Nicola Piras and Silvia Columbu gratefully acknowledge the support to their research by project "GraphNet: models, computation and estimation in networks and graph analysis" funded by D.M. 737/2021 - Linea d'intervento Iniziative di ricerca interdisciplinare su temi di rilievo trasversale per il PNR (grant number F25F21002720001). Silvia Columbu also acknowledges the support of the projects "FIATLUCS - Favorire l'Inclusione e l'Accessibilità per i Trasferimenti Locali Urbani a Cagliari e Sobborghi" funded by the PNRR RAISE Liguria, Spoke 01, (grant number F23C24000240006) and "MAPS" funded by Fondazione di Sardegna (grant number F73C23001550007).

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

References

Almalaurea: 2018 Reports on Graduates' Profile and Occupational Condition. www.almalaurea.it

Biernacki, C., Celeux, G., Govaert, G.: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Comput. Statist. Data Anal. **41**(3–4), 561–575 (2003)

Bennink, M., Croon, M.A., Kroon, B.E.A.: Micro-macro multilevel latent class models with multiple discrete individual-level variables. Adv. Data Anal. Classif. **10**, 139–154 (2016)

Biernacki, C., Jacques, J., Keribin, C.: A survey on model-based co-clustering: high dimension and estimation challenges. J. Classif. **40**, 332–381 (2023)

Bartolucci, F., Pandolfi, S., Pennoni, F.: Discrete latent variable models. Ann. Rev. Statist. **5**, 425–452 (2022)

Celeux, G., Chauveau, D., Diebolt, J.: Stochastic versions of the EM algorithm: an experimental study in the mixture case. J. Stat. Comput. Simul. **55**(4), 287–314 (1996)

Celeux, G., Diebolt, J.: L'algorithme sem: un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. Revue de statistique appliquée **34**(4), 35–52 (1986)

Columbu, S., Piras, N., Vermunt, J.K.: Log-likelihood approximation in stochastic em for multilevel latent class models. In: Plaia, A., Egidi, L., Abbruzzo, A. (eds.) Proceedings of the Statistics and Data Science 2024 Conference - New Perspectives on Statistics and Data Science, Palermo (2024)

Dempster, A.P., Laird, N.M., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser. B (Methodol.) **39**(1), 1–38 (1977)

Di Mari, R., Bakk, Z., Oser, J., Kuha, J.: A two-step estimator for multilevel latent class analysis with covariates. Psychometrika **88**(4), 1144–1170 (2023)

D'Elia, A., Piccolo, D.: A mixture model for preference data analysis. Comput. Stat. Data Anal. **49**(3), 917–934 (2005)

Eddelbuettel, D., Francois, R., Allaire, J., Ushey, K., Kou, Q., Russell, N., Ucar, I., Bates, D., Chambers, J.: Rcpp: Seamless R and C++ Integration. (2024). R package version 1.0.13. https://CRAN.Rproject.org/package=Rcpp

Goldstein, H.: Multilevel Statistical Models. Wiley Series in Probability and Statistics, London (2010)
Holland, P.W., Blackmond Laskey, K., Leinhardt, S.: Stochastic blockmodels: first steps. Soc. Netw. **5**(2), 109–137 (1983)
Hagenaars, J.A., McCutcheon, A.L.: Applied Latent Class Analysis. Cambridge University Press, Cambridge (2002)
Jacques, J., Biernacki, C.: Model-based co-clustering for ordinal data. Comput. Stat. Data Anal. **123**, 101–115 (2018)
Keribin, C., Brault, V., Celeux, G., Govaert, G.: Estimation and selection for the latent block model on categorical data. Stat. Comput. **25**(6), 1201–1216 (2015)
Lazarsfeld, P.F.: The logical and mathematical foundation of latent structure analysis. Studies in Social Psychology in World War II Vol. IV : Measurement and Prediction, 362–412 (1950)
Lukočiene, O., Varriale, R., Vermunt, J.K.: The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. Sociol. Methodol. **40**(1), 247–283 (2010)
Magidson, J.: Qualitative variance, entropy, and correlation ratios for nominal dependent variables. Soc. Sci. Res. **10**, 177–194 (1981)
McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. Wiley Series in Probability and Statistics, Wiley, New Jersey (2008)
Marin, J.M., Mengersen, K., Robert, C.P.: Bayesian modelling and inference on mixtures of distributions. In: Dey, D.K. (ed.) Handbook of Statistics, vol. 25, pp. 459–507. Elsevier, Amsterdam (2005)
McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley Series in Probability and Statistics, Wiley, UK (2004)
Maruotti, A., Punzo, A.: Initialization of hidden Markov and semi-Markov models: a critical evaluation of several strategies. Int. Stat. Rev. **89**(3), 447–480 (2021)
Selosse, M., Jacques, J., Biernacki, C.: Model-based co-clustering for mixed type data. Comput. Stat. Data Anal. **144**, 106866 (2020)
Sportisse, A., Marbac, M., Laporte, F., Celeux, G., Boyer, C., Josse, J., Biernacki, C.: Model-based clustering with missing not at random data. Statist. Comput. **34**, 135 (2024)
Skrondal, A., Rabe-Hesketh, S.: Generalized Latent Variable Modelling: Multilevel Longitudinal and Structural Equation Models. Chapman and Hall/CRC, Boca Raton (2004)
UNESCO: Isced fields of education and international standard classification of education 2011. Technical report, UNESCO Institute for Statistics, Montréal (2014)
Vermunt, J.K.: Multilevel latent class models. Sociol. Methodol. **33**, 213–239 (2003)
Vermunt, J.K.: An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. Stat. Neerl. **58**, 220–233 (2004)
Vermunt, J.K.: Mixed-effects logistic regression models for indirectly observed discrete outcome variables. Multivar. Behav. Res. **40**(3), 281–301 (2005)
Vermunt, J.K.: Latent class and finite mixture models for multilevel data sets. Stat. Methods Med. Res. **17**, 33–51 (2008)
Vermunt, J.K.: Latent class modeling with covariates: Two improved three-step approaches. Polit. Anal. **18**, 450–469 (2010)
Vermunt, J.K., Magidson, J.: Upgrade Manual for Latent GOLD 5.1: Basic, Advanced, and Syntax. Statistical Innovations Inc., Belmont Massachusetts (2016)
Wu, C.J.: On the convergence properties of the EM algorithm. Ann. Stat. **11**, 95–103 (1983)