Statistics and Computing (2025) 35:177
https://doi.org/10.1007/s11222-025-10697-5

ORIGINAL PAPER

Hierarchical Mixtures of Latent Trait Analyzers with concomitant variables for multivariate binary data

Dalila Failli[1] · **Maria Francesca Marino**[2] · **Bruno Arpino**[3]

Received: 9 January 2025 / Accepted: 24 July 2025

Abstract

Analyzing three-way data is challenging due to complex dependencies between observations, which must be accounted for to ensure reliable results. We focus on hierarchical, multivariate, binary data organized in a three-way data structure, where rows correspond to first-level units, columns to variables, and layers to second-level units within which the first-level units are nested. In this framework, model-based clustering methods can be effectively employed for dimensionality reduction purposes, facilitating a clear understanding of the phenomenon under investigation. In this work, we propose a novel modeling tool for a hierarchical clustering of first- and second-level units. We extend the Mixture of Latent Trait Analyzers (MLTA) with concomitant variables by letting prior component probabilities depend also on second-level-specific random effects. Parameter estimation is performed by means of a double EM algorithm based on a variational approximation of the model log-likelihood function, along with a nonparametric maximum likelihood estimation of the second-level-specific random effect distribution. This latter approach allows to estimate a discrete distribution which directly provides a clustering of second-level units. Within (conditional on) each of such clusters, first-level units are partitioned thanks to the MLTA specification. The proposal is applied to data from the European Social Survey to partition countries (second-level units) according to the baseline attitude of their residents (first-level units) toward digital technologies (variables). Within these clusters, residents are partitioned on the basis of their attitude toward specific digital skills. The influence of socio-economic factors on the identification of digitalization profiles is also taken into consideration via a concomitant variable approach.

Keywords Model-based clustering · Latent variables · Finite mixtures · EM algorithm · Variational inference · NPML

1 Introduction

Data analysis in many sectors often has to deal with nested or hierarchical data structures, e.g., students belonging to the same schools, employees working in the same companies, and individuals living in the same regions. In these cases, data come in the form of three-way datasets, where rows correspond to first-level units, columns to variables, and layers to second-level units. The dependence structure characterizing such data may be rather complex and unobserved heterogeneity between layers must be taken into consideration to ensure reliable results from the analyses.

In this framework, finite mixture models represent an interesting tool of analysis. In detail, the multilevel latent class model introduced by Vermunt (2003) is suitable in the presence of categorical multivariate data having a multilevel structure. Here, first-level units are partitioned into clusters by means of a discrete latent variable, whose distribution is allowed to depend on a second-level latent variable (random effect) that accounts for the multilevel data structure. The second-level latent variable may be assumed to be continuous and the corresponding distribution may be parametrically specified. A more flexible alternative consists of leaving the distribution of the second-level latent variable

✉ Dalila Failli
dalila.failli@unipg.it

Maria Francesca Marino
mariafrancesca.marino@unifi.it

Bruno Arpino
bruno.arpino@unipd.it

[1] Department of Political Science, University of Perugia, Perugia, Italy

[2] Department of Statistics, Computer Science, Applications "Giuseppe Parenti", University of Florence, Florence, Italy

[3] Department of Statistical Sciences and Department of Philosophy, Sociology, Education and Applied Psychology, University of Padua, Padova, Italy

Published online: 28 August 2025

unspecified and estimating it from the data. This is done via a nonparametric maximum likelihood approach (NPML; see Laird 1978; Lindsay 1983a; Aitkin 1996, 1999). According to this approach, it is possible to identify latent clusters at each level of the hierarchy, while also avoiding stringent parametric assumptions and heavy computations (Vermunt and Van Dijk 2001). See also Vermunt and Magidson (2005). Vermunt (2007) further extends this model to deal with continuous multivariate responses organized in the form of general three-way structures.

In this paper, we build on the Mixture of Latent Trait Analyzers (MLTA) with concomitant variables proposed by Failli et al. (2024) as an extension of the original work by Gollini and Murphy (2014). This model allows to deal with binary data by identifying clusters of units sharing unobserved characteristics via a finite mixture specification. A multidimensional, continuous, latent variable (trait) included in the model also allows to capture the possible residual heterogeneity between units in a given cluster in the way they respond to the different items. In addition, thanks to the concomitant variable specification, the model allows measuring the effect that covariates have on cluster formation.

In the spirit of Vermunt (2003) and Vermunt and Magidson (2005), we extend such a proposal in a multilevel perspective by modeling prior component probabilities as a function of both concomitant variables and second-level-specific random effects. We leave the distribution of such random effects unspecified and estimate it from the data via the NPML approach. The resulting discrete distribution is totally free of stringent parametric assumptions and, thanks to such a discreteness, we obtain a clustering of second-level units as a by-product. Within each of these clusters (in the following referred to as "blocks"), the MLTA specification still allows to partition first-level units. Overall, the proposal defines a Hierarchical Mixture of Latent Trait Analyzers (HMLTA).

When compared to the alternatives available in the literature, the main advantage of the proposed approach is that it allows to account for the possible residual dependence (heterogeneity) within first-level units belonging to the same cluster using the continuous latent trait of the model. This means that, unlike competitors, it is based on a less stringent form of local independence that, in many instances, may lead to the identification of an excessive number of clusters (see Gollini and Murphy 2014 and Failli et al. 2024).

Model parameter estimates are obtained within a maximum likelihood framework by indirectly maximizing the log-likelihood function using an Expectation-Maximization (EM) algorithm (Dempster et al. 1977). Following the approach by Vermunt (2003) and Vermunt and Magidson (2005), an upward-downward strategy is employed to reduce the computational complexity of the estimation process. However, the current setting requires modifications to handle the multi-dimensional integrals over latent traits that do not have closed-form solutions. To approximate these integrals, Gauss-Hermite quadrature methods (Abramowitz and Stegun 1964; Pinheiro and Bates 1995) may be employed, although the computational burden increases substantially as the dimensionality of the latent trait grows. As a solution, we extend the double EM algorithm originally proposed by Gollini and Murphy (2014) and subsequently extended by Failli et al. (2024). This is based on a variational approximation of the model log-likelihood (Jaakkola and Jordan 1997; Tipping 1998). Overall, such an approach is computationally fast and straightforward to implement, making it particularly useful for the analysis of large datasets.

A large-scale simulation study is conducted to test the performance of the proposal in terms of parameters and cluster recovery, also compared to the alternative approaches available in the literature. We consider various simulation settings with different sizes for the three-way data structure and a varying number of first- and second-level partitions. In addition, the computational effort required to obtain estimates and the model selection performances are evaluated.

The model is applied to data from the European Social Survey (ESS Round 10 2020). Specifically, focusing on information on digital skills collected in the last round of the ESS, we consider a three-way binary data matrix entailing the mastery of various digital technologies (variables) by residents (first-level units) in different countries (second-level units). The proposed model allows to perform a hierarchical clustering of respondents based on their digital skills, and, on the top of that, a clustering of countries according to the baseline attitudes to digital technology of their residents. Furthermore, the model enables examining the influence of socio-economic and demographic factors on individuals' digitalization profiles, as well as taking into account the potential residual heterogeneity of individuals' responses on the digital skills' variables.

The paper is organized as follows. In Section 2, we introduce the assumptions underlying the original specification of the MLTA model, while Section 3 describes the proposed extension. Section 4 outlines the double EM algorithm for maximum likelihood estimation of model parameters, while Section 5 describes the procedure for standard errors' computation, model selection, and clustering. Section 6 shows the results of the simulation study. In Section 7, we describe the ESS data and present the application of the proposed model, while Section 8 contains concluding remarks and points to further extensions of the proposal.

2 MLTA for binary data

Let Y_{ik} denote the k-th binary random variable and y_{ik} the corresponding observed value for the i-th unit, with $i = 1, \ldots, N$, and $k = 1, \ldots, R$. The MLTA model, introduced by Gollini and Murphy (2014) and later extended by Gollini (2020), integrates latent class and latent trait analysis, assuming that a set of N units can be partitioned into G distinct classes (or clusters) and that the propensity of each unit to "positively" respond to the R variables also depends on a multidimensional continuous latent trait. In detail, we start by considering a discrete, unit-specific, latent variable of the type $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})' \overset{iid}{\sim}$ Multinomial $(1, (\eta_{i1}, \ldots, \eta_{iG})')$ whose generic element is

$$z_{ig} = \begin{cases} 1 & \text{if unit } i \text{ belongs to cluster } g, \\ 0 & \text{otherwise.} \end{cases}$$

The parameter η_{ig} denotes the probability for unit i to belong to cluster g, with $g = 1, \ldots, G$, under the constraints that $\sum_{g=1}^{G} \eta_{ig} = 1$ and $\eta_{ig} > 0$, $g = 1, \ldots, G$. Gollini and Murphy (2014) assume these parameters to be constant across units (i.e., $\eta_{ig} = \eta_g \; \forall i = 1, \ldots, N$), while Failli et al. (2024) model them as a function of unit-specific concomitant variables.

Second, let $\boldsymbol{u}_i$ be a D-dimensional continuous latent trait distributed according to a D-variate Gaussian density, with zero mean vector and identity covariance matrix, i.e., $\boldsymbol{u}_i \sim \mathcal{N}_D(\boldsymbol{0}, \boldsymbol{I})$. This is assumed to capture the residual variability of units within each cluster in the way they respond to the different variables. The MLTA assumes that, conditional on $z_{ig} = 1$ and $\boldsymbol{u}_i$, variables contained in the vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iR})'$ are independent each other and follow a Bernoulli distribution with parameter $\pi_{gk}(\boldsymbol{u}_i), k = 1, \ldots, R$, modeled through the following logistic function:

$$\begin{aligned} \pi_{gk}(\boldsymbol{u}_i) &= \Pr(Y_{ik} = 1 \mid \boldsymbol{u}_i, z_{ig} = 1) \\ &= \frac{1}{1 + \exp\{-(b_{gk} + \boldsymbol{w}'_{gk}\boldsymbol{u}_i)\}}. \end{aligned}$$

Here, the model intercept b_{gk} represents the baseline tendency of units in the g-th cluster to "positively" respond to the k-th variable. More specifically, a positive (respectively, negative) value for this parameter indicates a tendency toward (respectively, against) a "positive" response to the k-th variable (item) for units in the g-th cluster. Furthermore, slopes $\boldsymbol{w}_{gk}$ associated with the latent variable $\boldsymbol{u}_i$ are meant to capture the influence of the latent trait on the probability of a "positive" response to variable k from units belonging to the g-th cluster. Therefore, statistically significant estimates of $\boldsymbol{w}_{gk}$ indicate residual association between variables, as well as the presence of heterogeneity within clusters with respect to the baseline level dictated by b_{gk}. A simpler and more parsimonious version of the model is obtained by assuming $\boldsymbol{w}_{gk}$ to be constant across clusters (i.e., $\boldsymbol{w}_{gk} = \boldsymbol{w}_k \; \forall g = 1, \ldots, G$), meaning that the latent trait has the same effect in all clusters.

3 Hierarchical MLTA

In this section, we extend the MLTA model described in Section 2 in a multilevel perspective with the aim of obtaining a hierarchical clustering of first- and second-level units in a three-way binary data structure. In detail, Section 3.1 introduces the model, while Section 3.2 discusses model identifiability.

3.1 Model assumptions

Let Y_{hik} be the k-th binary variable associated to the i-th first-level unit nested within the h-th second-level unit, with $h = 1, \ldots, H$, $i = 1, \ldots, n_h$, and $k = 1, \ldots, R$. We assume that first-level units belong to homogeneous clusters, represented by the latent variable $\boldsymbol{z}_{hi} = (z_{hi1}, \ldots, z_{hiG})' \overset{iid}{\sim}$ Multinomial $(1, (\eta_{hi1}, \ldots, \eta_{hiG})')$, whose generic element is

$$z_{hig} = \begin{cases} 1 & \text{if first-level unit } i \text{ within second-level unit } h \\ & \text{belongs to cluster } g, \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand, η_{hig} denotes the prior probability for the i-th first-level unit within second-level unit h to belong to cluster g. Furthermore, as standard in the MLTA framework, we assume the existence of a D-dimensional, standard Gaussian, latent trait associated with each first-level unit in a given second-level unit, i.e., $\boldsymbol{u}_{hi} \sim \mathcal{N}_D(\boldsymbol{0}, \boldsymbol{I})$.

Conditional on $\boldsymbol{u}_{hi}$ and $z_{hig} = 1$, the elements in the vector $\boldsymbol{Y}_{hi} = (Y_{hi1}, \ldots, Y_{hiR})'$ are assumed to be independent each other, so that the following factorization holds

$$f(\boldsymbol{y}_{hi} \mid \boldsymbol{u}_{hi}, z_{hig} = 1) = \prod_{k=1}^{R} f(y_{hik} \mid \boldsymbol{u}_{hi}, z_{hig} = 1).$$

Here, $f(y_{hik} \mid \boldsymbol{u}_{hi}, z_{hig} = 1)$ denotes the Bernoulli probability mass function, with success probability $\pi_{gk}(\boldsymbol{u}_{hi})$ modeled via the following logistic function:

$$\begin{aligned} \pi_{gk}(\boldsymbol{u}_{hi}) &= \Pr(Y_{hik} = 1 \mid \boldsymbol{u}_{hi}, z_{hig} = 1) \\ &= \frac{1}{1 + \exp\{-(b_{gk} + \boldsymbol{w}'_{gk}\boldsymbol{u}_{hi})\}}. \end{aligned} \tag{1}$$

As typical of the MLTA specification, the continuous latent trait is meant to capture the potential residual heterogeneity within clusters of first-level units. This means that, conditional on the membership of first-level unit i (nested within the second-level unit h) to the g-th cluster, the probability of a "positive" response to variable k is specified by a latent trait model with parameters b_{gk} and $\boldsymbol{w}_{gk}$, with $g = 1, \ldots, G$, and $k = 1, \ldots, R$. The meaning and the interpretation of such parameters are the same as those detailed in Section 2.

To account for the effect that concomitant variables may have on the clustering of first-level units and also provide a clustering of second-level units, we parametrically model prior component probabilities, η_{hig}, by extending the concomitant variable approach introduced by Failli et al. (2024) for the MLTA specification. Specifically, we start by considering a latent class regression model (Agresti 2002) for η_{hig}, which is assumed to depend upon (i) a $(J+1)$-dimensional vector of cluster-specific parameters, $\boldsymbol{\beta}_g = (\beta_{0g}, \beta_{1g}, \ldots, \beta_{Jg})'$, measuring the impact of concomitant variables, $\boldsymbol{x}_{hi} = (1, x_{hi1}, \ldots, x_{hiJ})'$, on cluster formation; (ii) a second-level-specific random effect, ϵ_h, that is meant to capture sources of unobserved heterogeneity determining dependence between first-level units nested within the h-th second-level unit. This is also in line with the multilevel latent class specification introduced by Vermunt (2003). Overall, the following conditional model is considered:

$$\eta_{hig} = \Pr(z_{hig} = 1 \mid \epsilon_h, \boldsymbol{x}_{hi}) = \frac{\exp\{\boldsymbol{x}'_{hi}\boldsymbol{\beta}_g + \epsilon_h\}}{1 + \sum_{g'=2}^{G} \exp\{\boldsymbol{x}'_{hi}\boldsymbol{\beta}_{g'} + \epsilon_h\}}, \quad g = 2, \ldots, G. \tag{2}$$

A standard assumption for the random effect distribution is that of Gaussianity (Laird and Ware 1982). Other approaches rely on the use of mixtures of Gaussian distributions (Magder and Zeger 1996; Verbeke and Lesaffre 1996) or semi-parametric, smooth, distributions (e.g., Zhang and Davidian 2001; Ghidey, Lesaffre and Eilers 2004). Here, we decide to follow a different route, avoiding any unverifiable assumptions on the distribution of ϵ_h (the mixing distribution). To this aim, we adopt a NPML approach (Laird 1978; Lindsay 1983a; Aitkin 1996, 1999) which is based on leaving the mixing distribution unspecified and estimating it from the observed data in a maximum likelihood framework.

According to the NPML theory, as long as the likelihood function is bounded, it is maximized by a discrete mixing distribution defined over a finite number of support points, say $\{\gamma_1, \ldots, \gamma_Q\}$, with corresponding weights, say $\{\rho_1, \ldots, \rho_Q\}$, such that $\rho_q = \Pr(\epsilon_h = \gamma_q)$, with $\sum_{q=1}^{Q} \rho_q = 1$ and $\rho_q > 0$, for all $q = 1, \ldots, Q$; see Lindsay (1983a) and Lindsay (1983b). Within this framework, Equation (2) is modified as:

$$\eta_{higq} = \Pr(z_{hig} = 1 \mid \epsilon_h = \gamma_q, \boldsymbol{x}_{hi}) = \frac{\exp\{\boldsymbol{x}'_{hi}\boldsymbol{\beta}_g + \gamma_q\}}{1 + \sum_{g'=2}^{G} \exp\{\boldsymbol{x}'_{hi}\boldsymbol{\beta}_{g'} + \gamma_q\}}, \quad g = 2, \ldots, G. \tag{3}$$

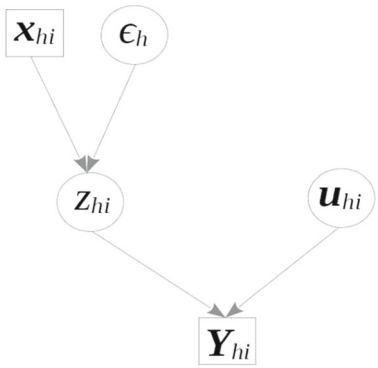

Fig. 1 Path diagram for the proposed HMLTA model.

The estimated discrete mixing distribution can be interpreted as representing homogeneous subsets of second-level units: each of them belongs to one out of Q unobserved clusters. In the following, we refer to such clusters as "blocks" to avoid confusion. Overall, the proposed parametrization allows for a hierarchical clustering of first- and second-level units. The latter are partitioned into blocks thanks to the discrete nature of the estimated distribution for ϵ_h; within each block, first-level units are partitioned into clusters via the discrete latent variable z_{hi} characterizing the MLTA specification.

The modeling assumptions detailed so far are summarized in the path diagram reported in Figure 1. From here, it is evident that the outcomes of interest in $\boldsymbol{Y}_{hi}$ are influenced by both the discrete latent variable z_{hi} and the continuous latent trait $\boldsymbol{u}_{hi}$. In turn, z_{hi} is influenced by observed covariates in $\boldsymbol{x}_{hi}$ and the discrete, second-level-specific, random effect ϵ_h.

3.2 Model identifiability

An important aspect to discuss for the specification of the proposed model is the identifiability of parameters. First, starting from the D-variate standard Gaussian distribution assumed for the latent trait $\boldsymbol{u}_{hi}$, it is relevant to highlight that the zero-mean assumption allows us to fix the location, while the identity covariance matrix allows to fix the scale of the latent trait (Skrondal and Rabe-Hesketh 2004).

Second, as typically happens in latent trait models, the slope parameters $\boldsymbol{w}_{gk}$ are only identifiable up to a rotation of the factors. As for this point, let us denote by $\boldsymbol{W}_g$, $g = 1, \ldots, G$, the $R \times D$ matrix formed by stacking by rows the $\boldsymbol{w}_{gk}$ vectors, for $k = 1, \ldots, R$. We may opt for imposing

constraints on these matrices to allow for rotation invariance. In this respect, we may fix the elements in the upper triangular of $\boldsymbol{W}_g$ to zero and constrain the diagonal elements to be positive (Niku et al. 2017). Alternatively, we may leave these matrices free to vary and allow for rotation: the flexibility of the model is not restricted and the latent variable components, $\boldsymbol{w}'_{gk}\boldsymbol{u}_{hi}$, allow to describe the correlation between variables that is not captured by the finite mixture specification via the residual $R \times R$ covariance matrix $\boldsymbol{\Lambda}_g = \boldsymbol{W}_g\boldsymbol{W}'_g$ (on the linear predictor scale) for $g = 1, \ldots, G$.

Last, choosing the NPML estimator for the second-level specific random effect requires additional constraints to ensure parameter identifiability. In detail, the intercepts $\beta_{0g} \in \boldsymbol{\beta}_g$, $g = 1, \ldots, G$, appearing in Equation (3) need to be set equal to 0. Alternatively, we may consider a block- and cluster-specific intercept $\gamma^*_{qg} = \gamma_q + \beta_{0g}$, representing the baseline propensity for first-level units to belong to the g-th cluster ($g = 1, \ldots, G$), conditional on the corresponding second-level unit being in block $q = 1, \ldots, Q$.

4 Maximum likelihood estimation

Let $\boldsymbol{\theta}$ be the vector of all free model parameters. Given the assumptions described in the previous section, the model log-likelihood function can be written as:

$$\ell(\boldsymbol{\theta}) = \sum_{h=1}^{H} \log \left(\sum_{q=1}^{Q} \rho_q \prod_{i=1}^{n_h} \sum_{g=1}^{G} \eta_{higq} \times \right. \\ \left. \times \int f(\boldsymbol{y}_{hi} \mid \boldsymbol{u}_{hi}, z_{hig} = 1) f(\boldsymbol{u}_{hi}) d\boldsymbol{u}_{hi} \right), \quad (4)$$

which clearly corresponds to the log-likelihood function of a hierarchical mixture model (Vermunt 2003; Vermunt and Magidson 2005).

As it is typical of latent variable models, maximum likelihood estimates of model parameters can be obtained via an indirect approach based on the use of an EM algorithm (Dempster et al. 1977). As in this context we need to deal with intractable integrals, to avoid the use of heavy approximation techniques, such as the Gaussian quadrature (Abramowitz and Stegun 1964; Pinheiro and Bates 1995), we may rely on a variational approach, as suggested by Jaakkola and Jordan (1997). A detailed description of the estimation procedure is provided in the following. To start, we describe the variational approach (Section 4.1); then, in Section 4.2, we detail the EM algorithm used for the indirect maximization of the likelihood function.

4.1 Variational approximation of the likelihood function

Let us start by the model log-likelihood function detailed in Equation (4). As evident, it requires the solution of a set of multi-dimensional integrals allowing to marginalize out the latent traits from the joint distribution of the responses and the latent variables in the model. As a solution, we may rely on a variational approximation.

The goal is to approximate the joint conditional distribution of the responses in $\boldsymbol{y}_{hi}$, $f(\boldsymbol{y}_{hi} \mid \boldsymbol{u}_{hi}, z_{hig} = 1)$, with a lower bound that is conjugate to the Gaussian density postulated for $\boldsymbol{u}_{hi}$, i.e., $f(\boldsymbol{u}_{hi})$. Since the logistic function defining $f(\boldsymbol{y}_{hi} \mid \boldsymbol{u}_{hi}, z_{hig} = 1)$ is convex, any of its tangents lies below it and is part of the family of lower-bounds. As a result, $f(\boldsymbol{y}_{hi} \mid \boldsymbol{u}_{hi}, z_{hig} = 1)$ can be approximated by the following function

$$\tilde{f}(\boldsymbol{y}_{hi} \mid \boldsymbol{u}_{hi}, z_{hig} = 1, \boldsymbol{\xi}_{hig}) = \prod_{k=1}^{R} \sigma(\xi_{higk}) \exp\left\{ \frac{A_{higk} - \xi_{higk}}{2} + \right. \\ \left. + \lambda(\xi_{higk})(A^2_{higk} - \xi^2_{higk}) \right\},$$

where $\boldsymbol{\xi}_{hig} = (\xi_{hig1}, \ldots, \xi_{higR})'$ are auxiliary variational parameters, with $\xi_{higk} \neq 0$, for all $k = 1, \ldots, R$. Furthermore,

$$A_{higk} = (2y_{hik} - 1)(b_{gk} + \boldsymbol{w}'_{gk}\boldsymbol{u}_{hi}),$$
$$\sigma(\xi_{higk}) = (1 + e^{-\xi_{higk}})^{-1},$$
$$\lambda(\xi_{higk}) = \frac{1/2 - \sigma(\xi_{higk})}{2\xi_{higk}}.$$

After some little algebra, the logarithm of the integral in Equation (4)

$$\mathcal{L}_{hig} = \log \int f(\boldsymbol{y}_{hi} \mid \boldsymbol{u}_{hi}, z_{hig} = 1) f(\boldsymbol{u}_{hi}) d\boldsymbol{u}_{hi}$$

can be approximated by

$$\mathcal{L}_{hig}(\boldsymbol{\xi}) = \log \int \tilde{f}(\boldsymbol{y}_{hi} \mid \boldsymbol{u}_{hi}, z_{hig} = 1, \boldsymbol{\xi}_{hig}) f(\boldsymbol{u}_{hi}) d\boldsymbol{u}_{hi} \\ = \log \int \prod_{k=1}^{R} \sigma(\xi_{higk}) \exp\left\{ \left(y_{hik} - \frac{1}{2} \right) b_{gk} + \right. \\ - \frac{\xi_{higk}}{2} + \lambda(\xi_{higk}) b^2_{gk} - \lambda(\xi_{higk}) \xi^2_{higk} + \\ \left. + \frac{\boldsymbol{\mu}'_{hig} \boldsymbol{\Sigma}^{-1}_{hig} \boldsymbol{\mu}_{hig}}{2} \right\} \mid \boldsymbol{\Sigma}_{hig} \mid^{1/2} \times \\ \times (2\pi)^{-D/2} \mid \boldsymbol{\Sigma}_{hig} \mid^{-1/2} \times \\ \times \exp\left\{ -\frac{1}{2} (\boldsymbol{u}_{hi} - \boldsymbol{\mu}_{hig}) \boldsymbol{\Sigma}^{-1}_{hig} (\boldsymbol{u}_{hi} - \boldsymbol{\mu}_{hig})' \right\} d\boldsymbol{u}_{hi}. \quad (5)$$

In the last part of the above equation, it is possible to recognize the kernel of a multivariate Gaussian density with covariance matrix and mean vector given by

$$\boldsymbol{\Sigma}_{hig} = \left[\boldsymbol{I} - 2\sum_{k=1}^{R}\lambda(\xi_{higk})\boldsymbol{w}'_{gk}\boldsymbol{w}_{gk}\right]^{-1}, \tag{6}$$

and

$$\boldsymbol{\mu}_{hig} = \boldsymbol{\Sigma}_{hig}\left[\sum_{k=1}^{R} y_{hik} - \frac{1}{2} + 2\lambda(\xi_{higk})b_{gk}\right]\boldsymbol{w}_{gk}, \tag{7}$$

respectively. This allows for the integration of the latent trait $\boldsymbol{u}_{hi}$ from the integral in Equation (5) and leads to

$$\begin{aligned}\mathcal{L}_{hig}(\boldsymbol{\xi}) = \log\Bigg(&\prod_{k=1}^{R}\sigma(\xi_{higk}) + \\ &+ \exp\left\{\left(y_{hik} - \frac{1}{2}\right)b_{gk} - \frac{\xi_{higk}}{2} + \lambda(\xi_{higk})b_{gk}^2 + \right. \\ &\left. - \lambda(\xi_{higk})\xi_{higk}^2 + \frac{\boldsymbol{\mu}'_{hig}\boldsymbol{\Sigma}_{hig}^{-1}\boldsymbol{\mu}_{hig}}{2} + \frac{\log\mid\boldsymbol{\Sigma}_{hig}\mid}{2}\right\}\Bigg).\end{aligned}$$

In turn, the log-likelihood function in Equation (4) is approximated as

$$\ell(\boldsymbol{\theta}, \boldsymbol{\xi}) = \sum_{h=1}^{H}\log\left(\sum_{q=1}^{Q}\rho_q\prod_{i=1}^{n_h}\sum_{g=1}^{G}\eta_{higq}\tilde{f}(\boldsymbol{y}_{hi}\mid z_{hig}=1,\boldsymbol{\xi}_{hig})\right),$$

where $\tilde{f}(\boldsymbol{y}_{hi}\mid z_{hig}=1,\boldsymbol{\xi}_{hig}) = \exp\{\mathcal{L}_{hig}(\boldsymbol{\xi})\}$.

4.2 EM algorithm

Once the integral is approximated, the log-likelihood function can be maximized in an indirect manner via an EM algorithm (Dempster et al. 1977). This starts from the definition of the following complete data log-likelihood function

$$\begin{aligned}\ell_c(\boldsymbol{\theta}) = &\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{g=1}^{G} z_{hig}\log f(\boldsymbol{y}_{hi}\mid\boldsymbol{u}_{hi}, z_{hig}=1) + \\ &+ \sum_{h=1}^{H}\sum_{i=1}^{n_h}\log f(\boldsymbol{u}_{hi}) + \\ &+ \sum_{h=1}^{H}\sum_{q=1}^{Q}\sum_{i=1}^{n_h}\sum_{g=1}^{G} v_{hq}\, z_{hig}\, \log\eta_{higq} + \\ &+ \sum_{h=1}^{H}\sum_{q=1}^{Q} v_{hq}\log\rho_q,\end{aligned} \tag{8}$$

where v_{hq} is an indicator variable being equal to 1 if second-level unit h belongs to block q and 0 otherwise; that is, $v_{hq} = \mathbb{I}(\epsilon_h = \gamma_q)$.

At the generic iteration of the algorithm, the E-step consists of computing the expectation of the complete data log-likelihood in Equation (8), conditional on the observed data and the current parameter estimates $\boldsymbol{\theta}^{(t)}$. That is, we need to compute

$$\begin{aligned}Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(t)}) = &\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{g=1}^{G}\int\hat{z}_{hig}f(\boldsymbol{u}_{hi}\mid z_{hig}=1,\boldsymbol{y}_{hi})\times \\ &\times\log f(\boldsymbol{y}_{hi}\mid\boldsymbol{u}_{hi}, z_{hig}=1)d\boldsymbol{u}_{hi} + \\ &+ \sum_{h=1}^{H}\sum_{i=1}^{n_h}\int f(\boldsymbol{u}_{hi}\mid\boldsymbol{y}_{hi})\log f(\boldsymbol{u}_{hi})d\boldsymbol{u}_{hi} + \\ &+ \sum_{h=1}^{H}\sum_{q=1}^{Q}\sum_{i=1}^{n_h}\sum_{g=1}^{G}\hat{v}_{hq}\hat{z}_{hig|q}\log\eta_{higq} + \\ &+ \sum_{h=1}^{H}\sum_{q=1}^{Q}\hat{v}_{hq}\log\rho_q.\end{aligned} \tag{9}$$

Here, $\hat{z}_{hig}$ and $\hat{v}_{hq}$ denote the posterior expectation of z_{hig} and v_{hq}, respectively, while $\hat{z}_{hig|q}$ is the posterior expectation of z_{hig}, conditional on $v_{hq}=1$. By doing some little algebra, these quantities can be computed as

$$\begin{aligned}\hat{z}_{hig} &= \frac{\sum_{q=1}^{Q}\rho_q\,\eta_{higq}\int f(\boldsymbol{y}_{hi}\mid\boldsymbol{u}_{hi}, z_{hig}=1)f(\boldsymbol{u}_{hi})d\boldsymbol{u}_{hi}}{\sum_{q=1}^{Q}\rho_q\sum_{g=1}^{G}\eta_{higq}\int f(\boldsymbol{y}_{hi}\mid\boldsymbol{u}_{hi}, z_{hig}=1)f(\boldsymbol{u}_{hi})d\boldsymbol{u}_{hi}} \\ &\approx \frac{\sum_{q=1}^{Q}\rho_q\,\eta_{higq}\,\tilde{f}(\boldsymbol{y}_{hi}\mid z_{hig}=1,\boldsymbol{\xi}_{hig})}{\sum_{q=1}^{Q}\rho_q\sum_{g=1}^{G}\eta_{higq}\,\tilde{f}(\boldsymbol{y}_{hi}\mid z_{hig}=1,\boldsymbol{\xi}_{hig})},\end{aligned}$$

$$\begin{aligned}\hat{v}_{hq} &= \frac{\rho_q\prod_{i=1}^{n_h}\sum_{g=1}^{G}\eta_{higq}\int f(\boldsymbol{y}_{hi}\mid\boldsymbol{u}_{hi}, z_{hig}=1)f(\boldsymbol{u}_{hi})d\boldsymbol{u}_{hi}}{\sum_{q=1}^{Q}\rho_q\prod_{i=1}^{n_h}\sum_{g=1}^{G}\eta_{higq}\int f(\boldsymbol{y}_{hi}\mid\boldsymbol{u}_{hi}, z_{hig}=1)f(\boldsymbol{u}_{hi})d\boldsymbol{u}_{hi}} \\ &\approx \frac{\rho_q\prod_{i=1}^{n_h}\sum_{g=1}^{G}\eta_{higq}\,\tilde{f}(\boldsymbol{y}_{hi}\mid z_{hig}=1,\boldsymbol{\xi}_{hig})}{\sum_{q=1}^{Q}\rho_q\prod_{i=1}^{n_h}\sum_{g=1}^{G}\eta_{higq}\,\tilde{f}(\boldsymbol{y}_{hi}\mid z_{hig}=1,\boldsymbol{\xi}_{hig})},\end{aligned}$$

and

$$\begin{aligned}\hat{z}_{hig|q} &= \frac{\eta_{higq}\int f(\boldsymbol{y}_{hi}\mid\boldsymbol{u}_{hi}, z_{hig}=1)f(\boldsymbol{u}_{hi})d\boldsymbol{u}_{hi}}{\sum_{g=1}^{G}\eta_{higq}\int f(\boldsymbol{y}_{hi}\mid\boldsymbol{u}_{hi}, z_{hig}=1)f(\boldsymbol{u}_{hi})d\boldsymbol{u}_{hi}} \\ &\approx \frac{\eta_{higq}\,\tilde{f}(\boldsymbol{y}_{hi}\mid z_{hig}=1,\boldsymbol{\xi}_{hig})}{\sum_{g=1}^{G}\eta_{higq}\,\tilde{f}(\boldsymbol{y}_{hi}\mid z_{hig}=1,\boldsymbol{\xi}_{hig})}.\end{aligned}$$

To derive the above quantities, an upward-downward approach is employed (Vermunt 2003, Vermunt and Magidson 2005). Specifically, the algorithm first integrates out the latent variables from the lower to the higher levels, and then computes the marginal posterior probabilities in the reverse direction, i.e., from higher to lower levels. This strategy allows the

computational cost, in terms of both memory and processing time, to increase linearly with the number of lower-level observations, in contrast to the exponential growth typical of a standard EM algorithm.

The M-step of the algorithm consists of updating the model parameter estimates by maximizing the expected complete data log-likelihood function in Equation (9) with respect to $\boldsymbol{\theta}$. To this end, the following steps are carried out.

1. The multinomial logit coefficients $\boldsymbol{\beta}_g$ and γ_q are estimated via a Newton-Raphson algorithm, by maximizing the weighted likelihood of a multinomial logit model, with weights equal to $\hat{v}_{hq} \times \hat{z}_{hig|q}$. The prior probabilities η_{higq} are updated accordingly. Moreover, the prior block probabilities ρ_q are updated as:

$$\hat{\rho}_q = \frac{\sum_{h=1}^{H} \hat{v}_{hq}}{H}.$$

2. A second EM algorithm, nested within the first, is required to update the posterior mean vectors and covariance matrices $\boldsymbol{\mu}_{hig}$ and $\boldsymbol{\Sigma}_{hig}$, as well as the variational parameters ξ_{higk}, and the latent trait parameters, b_{gk} and $\boldsymbol{w}_{gk}$. These are obtained as described in the following.

   (i) In the second-level E-step, we update the covariance matrices $\boldsymbol{\Sigma}_{hig}$ and the mean vectors $\boldsymbol{\mu}_{hig}$ according to Equations (6) and (7), conditional on the observations $\boldsymbol{y}_{hi}$, the posterior probabilities $\hat{z}_{hig}$, and the variational parameters $\hat{\boldsymbol{\xi}}_{hig}$.

   (ii) In the second-level M-step, the variational parameters ξ_{higk} are updated as

$$\hat{\xi}^2_{higk} = b^2_{gk} + 2b_{gk}\boldsymbol{w}'_{gk}\boldsymbol{\mu}_{hig} + \boldsymbol{w}'_{gk}(\boldsymbol{\Sigma}_{hig} + \boldsymbol{\mu}_{hig}\boldsymbol{\mu}'_{hig})\boldsymbol{w}_{gk},$$

   while for the latent trait parameters in the vector $\hat{\boldsymbol{\zeta}}_{gk} = (\hat{\boldsymbol{w}}'_{gk}, \hat{b}_{gk})'$, we have the following updating rule:

$$\hat{\boldsymbol{\zeta}}_{gk} = -\left[2\sum_{h=1}^{H}\sum_{i=1}^{n_h} \hat{z}_{hig}\lambda(\hat{\xi}_{higk})\mathbb{E}[\hat{\boldsymbol{u}}_{hi}\hat{\boldsymbol{u}}'_{hi}]_g\right]^{-1} \times \times \left[\sum_{h=1}^{H}\sum_{i=1}^{n_h} \hat{z}_{hig}\left(y_{hik} - \frac{1}{2}\right)\hat{\boldsymbol{\alpha}}_{hig}\right].$$

   Here, we have

$$\hat{\boldsymbol{\alpha}}_{hig} = (\hat{\boldsymbol{\mu}}'_{hig}, 1)'$$

   and

$$\mathbb{E}[\hat{\boldsymbol{u}}_{hi}\hat{\boldsymbol{u}}'_{hi}]_g = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{hig} + \hat{\boldsymbol{\mu}}_{hig}\hat{\boldsymbol{\mu}}'_{hig} & \hat{\boldsymbol{\mu}}_{hig} \\ \hat{\boldsymbol{\mu}}'_{hig} & 1 \end{bmatrix}.$$

The lower bound of the log-likelihood function, $\ell(\boldsymbol{\theta}, \boldsymbol{\xi})$, is updated accordingly.

The different steps of the estimation algorithm are iterated until the following stopping rule is satisfied:

$$|| \ \ell(\hat{\boldsymbol{\theta}}^{(t+1)}, \hat{\boldsymbol{\xi}}^{(t+1)}) - \ell(\hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{\xi}}^{(t)}) \ ||< \delta,$$

where t denotes the iteration index and $\delta = 0.1^6$ is the chosen tolerance level.

Since the variational approach considers a lower bound of the true log-likelihood function, a refined evaluation of the log-likelihood is obtained in the final step using Gauss-Hermite quadrature (Abramowitz and Stegun 1964; Pinheiro and Bates 1995), as suggested by Gollini and Murphy (2014).

5 Standard errors, model selection, and clustering

To evaluate uncertainty associated with the estimates obtained from the variational EM algorithm, we employ a nonparametric bootstrap approach (Efron 1979) within each second-level unit, to preserve the multilevel structure of the data (Goldstein 2011). Given the three-way data matrix $\boldsymbol{Y}$, the nonparametric bootstrap consists of repeatedly drawing with replacement a sample having the same size of $\boldsymbol{Y}$, and then estimating the model parameters. In detail, for each bootstrap replicate, n_h rows are drawn with repetition from the $(n_h \times R)$-dimensional matrix $\boldsymbol{Y}_h$, for $h = 1, \ldots, H$, so that each first-level unit can appear several times within each second-level unit. The model is estimated for fixed G, D, and Q.

Let $\hat{\boldsymbol{\theta}}_{(s)}$ denote the vector of estimates obtained from the s-th bootstrap sample. By paying attention to the label switching issue via a sorting of model parameters at each iteration, bootstrap standard errors correspond to the squared root of the diagonal elements of the following matrix:

$$\hat{V}(\hat{\boldsymbol{\theta}}) = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\boldsymbol{\theta}}_{(s)} - \hat{\boldsymbol{\theta}}_{(.)}\right)\left(\hat{\boldsymbol{\theta}}_{(s)} - \hat{\boldsymbol{\theta}}_{(.)}\right)',$$

with $\hat{\boldsymbol{\theta}}_{(.)}$ being the empirical mean vector $\hat{\boldsymbol{\theta}}_{(.)} = \frac{1}{S}\sum_{s=1}^{S}\hat{\boldsymbol{\theta}}_{(s)}$.

Regarding model selection, the HMLTA specification is estimated for different values of G, D, and Q; the optimal combination $[G, D, Q]$ is selected by means of standard penalized information criteria. Specifically, the simulation study discussed in Section 6.4 shows that a good model selection strategy is based on considering the Bayesian Information Criterion (BIC; Schwarz 1978)

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}}) + \nu \log(N).$$

Here, the penalization term considers the total number of free model parameters, i.e., $\nu = (G \times M) + (G \times M \times D - D \times (D-1)/2) + J \times (G-1) + Q + (Q-1)$, as well as the overall number of first-level units in the sample, i.e., $N = \sum_{h=1}^{H} n_h$. Alternative strategies may also be considered. In detail, the optimal model may be selected according to the Akaike Information Criterion (AIC; Akaike 1974)

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}) + 2\nu,$$

or the Integrated Classification Likelihood (ICL; Biernacki et al. 2000)

$$\text{ICL} = \text{BIC} - \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{g=1}^{G}\hat{z}_{hig}\log(\hat{z}_{hig}).$$

In this latter case, the optimal model is the one corresponding to the maximum ICL.

Note that in the framework of multilevel latent class models, Lukočienė et al. (2010) suggest instead the use of a three-step strategy to select the optimal number of clusters of first- and second-level units. Considering the good performance of the BIC, we do not further investigate this latter approach.

To conclude, it is important to emphasize that when the model selection strategy outlined above results in an HMLTA specification with $D = 0$, the model simplifies to a multilevel latent class model, with second-level-specific (non-parametric) random effects, as introduced by Vermunt (2003). This choice suggests that the simpler model introduced in the literature by Vermunt (2003) is enough to account for dependencies in the observed data.

Last, as regards clustering, this is obtained when convergence of the estimation algorithm is achieved. Specifically, first- and second-level units are hierarchically assigned to blocks and clusters, respectively, on the basis of a maximum a posteriori (MAP) rule applied to the approximate posterior probabilities obtained from the variational EM algorithm.

6 Simulation study

To evaluate the ability of the proposed approach to correctly estimate the model parameters and classify first- and second-level units, we conduct a simulation study, as described below. Results are based on $B = 100$ replicates, each considering 100 random starting values.

6.1 Simulation setup

We consider 24 different scenarios based on a different number of first-level units ($N = 500, 1000, 2000$), variables ($R = 7, 14$), clusters ($G = 3, 4$), and blocks ($Q = 2, 3$), while the size of the latent trait and the number of second-level units are kept constant and equal to $D = 1$ and $H = 20$, respectively. Note that a uni-dimensional latent trait is considered for convenience. Furthermore, all second-level units group the same number of first-level units. The membership of first-level units to clusters is defined in terms of a single concomitant variable, i.e., $\boldsymbol{x}_{hi} = (1, x_{hi1})'$, with x_{hi1} being drawn from a Gaussian distribution with mean and variance both equal to 1. As detailed in Section 3.2, for identifiability reasons, we always constraint the model intercepts to be equal to zero; that is, $\beta_{0g} = 0$, for each $g = 1, \ldots, G$, both when $G = 3$ and $G = 4$. The slope parameter associated with x_{hi1}, β_{1g}, is set instead equal to -0.4, -0.9, and -1.4, when $g = 2$, 3, and 4, respectively. Clearly, the latter value (-1.4) is considered only for those scenarios in which $G = 4$. Further, we set $\boldsymbol{\gamma} = (0, -2)'$ and $\boldsymbol{\rho} = (0.5, 0.5)'$ when Q=2, and $\boldsymbol{\gamma} = (0, -1, -2)'$ and $\boldsymbol{\rho} = (0.2, 0.3, 0.5)'$ when Q=3.

As for the observed layer of the model, we simulate the baseline tendency parameter $\boldsymbol{b}_g = (b_{g1}, \ldots, b_{gR})'$ from multivariate Gaussian distributions with mean vectors $-3\boldsymbol{I}_R$, $0\boldsymbol{I}_R$, $3\boldsymbol{I}_R$, and $6\boldsymbol{I}_R$, for $g = 1, 2, 3$, and 4, respectively. As before, this latter value is considered only when $G = 4$. The covariance matrix of the $\boldsymbol{b}_g$ parameters is set equal to $2.5\boldsymbol{I}_R$ in all scenarios. As regards the slopes $\boldsymbol{w}_{gk}$, we assume they are constant across clusters (i.e., $\boldsymbol{w}_{gk} = \boldsymbol{w}_k, \ \forall\ g = 1, \ldots, G$), meaning that the latent trait has the same effect in all clusters of first-level units. These are simulated from a Gaussian distribution with mean equal to -1 and variance equal to 1.

6.2 Simulation results: clustering recovery

The ability of the proposal to correctly classify first- and second-level units is evaluated via the Adjusted Rand Index (ARI; Rand 1971). This measures the agreement between the true partitions and the estimated ones. Its expected value is zero in the case of random partitioning, while it is 1 in the case of perfect agreement between the two partitions. Results of the simulation study are shown in Table 1. Looking at this table, we observe that, as N and R grows larger, the classification improves for both first- and second-level units. As expected, worse results are obtained in terms of first-level units partitioning as G increases. In detail, when $G = 4$, a larger number of first-level units and variables is needed to obtain good clustering performance. Similar conclusions can also be drawn for the partitioning of second-level units: generally, the larger Q, the worse the performance.

6.3 Simulation results: model parameters

Table 2 shows the mean squared error (MSE) values across samples for the slope parameters $\tilde{\boldsymbol{\beta}} = (\beta_{11}, \ldots, \beta_{1G})'$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_Q)'$, and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_Q)'$ by letting the

Table 1 ARI mean (median) across samples for first- and second-level units' partitions, with varying N, R, G, and Q.

	First-level units			
	G=3		G=4	
R=7	Q=2	Q=3	Q=2	Q=3
N=500	0.791 (0.788)	0.781 (0.782)	0.689 (0.671)	0.681 (0.695)
N=1000	0.793 (0.781)	0.786 (0.783)	0.689 (0.703)	0.684 (0.683)
N=2000	0.796 (0.785)	0.790 (0.786)	0.693 (0.691)	0.685 (0.682)
	G=3		G=4	
R=14	Q=2	Q=3	Q=2	Q=3
N=500	0.843 (0.838)	0.841 (0.841)	0.838 (0.847)	0.835 (0.830)
N=1000	0.844 (0.841)	0.842 (0.841)	0.839 (0.838)	0.836 (0.834)
N=2000	0.846 (0.847)	0.844 (0.838)	0.841 (0.836)	0.836 (0.838)
	Second-level units			
	G=3		G=4	
R=7	Q=2	Q=3	Q=2	Q=3
N=500	0.960 (1.000)	0.556 (0.573)	0.852 (0.900)	0.518 (0.518)
N=1000	0.980 (1.000)	0.673 (0.757)	0.960 (1.000)	0.627 (0.647)
N=2000	1.000 (1.000)	0.915 (0.883)	1.000 (1.000)	0.896 (1.000)
	G=3		G=4	
R=14	Q=2	Q=3	Q=2	Q=3
N=500	0.971 (1.000)	0.570 (0.560)	0.886 (1.000)	0.553 (0.581)
N=1000	1.000 (1.000)	0.706 (0.813)	0.980 (1.000)	0.675 (0.715)
N=2000	1.000 (1.000)	0.921 (0.962)	1.000 (1.000)	0.906 (0.955)

number of first-level units and partitions vary. As known, the MSE is computed as the mean of squared differences between the estimates and the true value of model parameters. Therefore, the smaller its value, the closer the estimates to the true parameters. The results indicate that increasing the data size improves the model's ability to recover the true data-generating process. However, as expected, MSE values increase as the number of partitions increases. In detail, for G = 4 and Q=3, a larger number of first-level units and variables is needed to obtain good results in terms of parameters' recovery. We expect that by further increasing the number of first-level units and variables, estimates improve.

6.4 Simulation results: model selection

In this section, we evaluate the performance of the BIC index in correctly identifying the true number of clusters and blocks, as well as the size of the latent trait. In detail, we simulate $B = 50$ datasets made of $H = 20$ second-level units, $N = 500$ first-level units, and $R = 7$ binary variables, from an HMLTA model based on $G = 3$ clusters, $Q = 2$ blocks, and a uni-dimensional latent trait ($D = 1$). Parameter values are set exactly as described in Section 6.1 for the setting based on the same number of clusters and blocks. For each dataset, we estimate an HMLTA model with a varying number of clusters ($G = 3, 4$), blocks ($Q = 1, 2, 3$), and latent trait size ($D = 1, 2$), and retain the optimal model by selecting the combination $[G, D, Q]$ providing the minimum BIC index. Results reported in Table 3 show that the correct model (i.e., the one with $G = 3$, $Q = 2$, and $D = 1$) is selected 98% of the time, thus suggesting very good performances of this criterion.

Furthermore, when comparing the suggested model selection approach based on the BIC index with strategies based on the AIC or ICL, we observe that BIC and ICL agree 94% of the time. In contrast, the AIC index generally favors more complex models and performs consistently worse than the alternatives.

6.5 A comparison with the multilevel latent class model

In this section, we compare the proposed HMLTA model with the multilevel LC model by Vermunt (2003). Note that, as highlighted before, this latter coincides with an HMLTA specification with $D = 0$. We consider two different scenarios by varying the number of first-level units ($N = 500, 20'000$) and clusters, ($G = 3, 4$), while the

Table 2 MSE values across samples for $\tilde{\boldsymbol{\beta}}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\rho}$, with varying N, R, G, and Q.

R=7		G=3		G=4	
		Q=2	Q=3	Q=2	Q=3
N=500	$\tilde{\boldsymbol{\beta}}$	0.01 0.02	0.01 0.02	0.08 0.05 0.03	0.12 0.03 0.03
	$\boldsymbol{\gamma}$	0.00 0.07	0.00 0.09 0.41	0.00 0.12	0.00 0.24 0.11
	$\boldsymbol{\rho}$	0.01 0.01	0.01 0.01 0.00	0.01 0.01	0.01 0.01 0.00
N=1000	$\tilde{\boldsymbol{\beta}}$	0.01 0.02	0.01 0.02	0.05 0.05 0.02	0.05 0.02 0.02
	$\boldsymbol{\gamma}$	0.00 0.04	0.00 0.08 0.10	0.00 0.08	0.00 0.22 0.08
	$\boldsymbol{\rho}$	0.01 0.01	0.01 0.01 0.00	0.01 0.01	0.01 0.01 0.00
N=2000	$\tilde{\boldsymbol{\beta}}$	0.01 0.00	0.01 0.01	0.03 0.01 0.01	0.02 0.01 0.02
	$\boldsymbol{\gamma}$	0.00 0.04	0.00 0.06 0.09	0.00 0.05	0.00 0.07 0.02
	$\boldsymbol{\rho}$	0.01 0.01	0.01 0.01 0.00	0.01 0.01	0.01 0.01 0.00
R=14		G=3		G=4	
		Q=2	Q=3	Q=2	Q=3
N=500	$\tilde{\boldsymbol{\beta}}$	0.01 0.02	0.01 0.02	0.03 0.03 0.03	0.04 0.01 0.02
	$\boldsymbol{\gamma}$	0.00 0.02	0.00 0.06 0.06	0.00 0.07	0.00 0.23 0.18
	$\boldsymbol{\rho}$	0.01 0.01	0.01 0.01 0.00	0.01 0.01	0.01 0.01 0.00
N=1000	$\tilde{\boldsymbol{\beta}}$	0.01 0.01	0.01 0.01	0.02 0.01 0.01	0.02 0.02 0.02
	$\boldsymbol{\gamma}$	0.00 0.02	0.00 0.06 0.02	0.00 0.07	0.00 0.21 0.01
	$\boldsymbol{\rho}$	0.01 0.01	0.01 0.01 0.00	0.01 0.01	0.01 0.01 0.00
N=2000	$\tilde{\boldsymbol{\beta}}$	0.01 0.00	0.01 0.00	0.01 0.01 0.01	0.01 0.01 0.02
	$\boldsymbol{\gamma}$	0.00 0.01	0.00 0.05 0.02	0.00 0.03	0.00 0.06 0.02
	$\boldsymbol{\rho}$	0.01 0.01	0.01 0.01 0.00	0.01 0.01	0.01 0.01 0.00

Table 3 Proportion of times the lowest BIC value occurred for varying G, Q, and D. The true model is highlighted in bold.

		G			G			G	
$Q=1$		3	4	$Q=2$	3	4	$Q=3$	3	4
D	1	0.01	0.00		**0.98**	0.01		0.00	0.00
	2	0.00	0.00		0.00	0.00		0.00	0.00

Table 4 Average ARI for first- and second-level units partitions, respectively, run time (in minutes), number of iterations (t) and log-likelihood (LL) across samples for the proposed HMLTA and the alternative multilevel LC model with varying N and G.

		$G=3$							
		HMLTA				Multilevel LC			
		ARI	Time	t	LL	ARI	Time	t	LL
N	500	0.79 0.96	3.95	204	-1642.25	0.62 0.89	2.36	168	-1826.08
	20000	0.80 1.00	180.53	245	-66144.97	0.62 1.00	128.10	186	-72546.29
		$G=4$							
		HMLTA				Multilevel LC			
		ARI	Time	t	LL	ARI	Time	t	LL
N	500	0.69 0.85	4.54	238	-1278.08	0.61 0.77	2.80	196	-1467.32
	20000	0.70 1.00	183.49	245	-51333.72	0.61 0.97	167.10	237	-58969.64

number of second-level units, variables, latent trait size, and blocks is kept constant and set to $H = 20$, $R = 7$, $D = 1$, and $Q = 2$, respectively. Model parameters for generating the data are fixed as described in Section 6.1. For each simulation scenario, we estimate both the proposed model and the multilevel LC model. In both cases, parameter initialization is based on a single random start. The aim is that of comparing the models in terms of clustering (measured by the ARI), computational runtime (in minutes), number of iterations, and log-likelihood value achieved at convergence of the EM algorithm. Results reported in Table 4 show that the proposed HMLTA outperforms the multilevel LC model in terms of clustering performance, as the ARI value obtained when fitting the former is higher than that of the latter in all scenarios. This is, clearly, an expected result. Furthermore, the proposed model has slightly higher computational times due to the larger number of iterations required to achieve convergence compared to the multilevel LC model. This is due to the presence of the multidimensional latent trait that requires a variational approximation for parameter estimation. However, the HMLTA exhibits a more gradual increase in the computational time and number of iterations before convergence with increasing N and G compared to the multilevel LC model. Overall, results show that the HMLTA is computationally efficient even on large-scale datasets ($N = 20'000$), thus making the proposal worth to be considered also in large-scale problems.

7 Application

In this section, we analyze data from the ESS using the proposed HMLTA with concomitant variables. The ESS is an academic cross-national survey whose main objective is to analyze the evolution of attitudes toward different topics throughout Europe. The survey involves subjects aged 15 years and older, resident in private households across several countries, regardless of their nationality, citizenship, language, or social status. The ESS includes a wide range of questions on social and political attitudes and behaviors, employment, education, health, well-being, social values, and trust in institutions. Each round includes both a core questionnaire that remains mostly invariant across time, and rotating modules. We focus on data from the module on digital social contacts in work and family life (Abendroth et al. 2023) introduced for the first time in the last available round of the survey (ESS Round 10 2020). Here, respondents are asked to provide information on the mastery of various digital skills. We report in Table 5 the list of such skills with the corresponding description.

Due to the impact of the COVID-19 pandemic, numerous countries faced restrictions on conducting face-to-face interviews during ESS round 10. As an alternative, these countries adopted a self-completion approach, which includes both web-based and paper surveys. Due to different data collection

Table 5 Digital skills' description.

Digital skill	Description
Internet	How often the respondent uses the Internet on computers, tablets, smartphones or other devices, whether for work or personal use
Preference settings	How familiar the respondent is with preference settings
Advanced search	How familiar the respondent is with advanced search
PDFs	How familiar the respondent is with PDFs
Video calls with child aged 12+	How often the respondent talks to child aged 12 or more on a screen
Video calls with parent	How often the respondent talks to parent on a screen
Video calls with manager	How often the respondent talks to the line manager on a screen
Video calls with colleagues	How often the respondent talks to colleagues on a screen
Messages with child aged 12+	How often the respondent communicates with child aged 12+ via text, email or messaging apps
Messages with parent	How often the respondent communicates with parent via text, email or messaging apps
Messages with manager	How often the respondent communicates with the line manager via text, email or messaging apps
Messages with colleagues	How often the respondent communicates with colleagues via text, email or messaging apps
Online political posts	During the last 12 months the respondent posted or shared anything about politics online, for example on blogs, via email or on social media such as Facebook or Twitter

methods and variations in questionnaire structure between self-completion and standard face-to-face interviews, comparability issues can arise (ESS 2022). Consequently, our analysis focuses exclusively on the subset of 21 countries that conducted data collection through traditional face-to-face interviews: Belgium, Bulgaria, Croatia, Czech Republic, Estonia, Finland, France, Hungary, Iceland, Ireland, Italy, Lithuania, Montenegro, Netherlands, North Macedonia, Norway, Portugal, Slovakia, Slovenia, Switzerland, and United Kingdom. In addition, given that digital skills are generally very high among young people, we focus on digitalization among older adults. Thus, we select only individuals aged 50 and older among the ESS respondents. All subsequent analyses are conducted on this subset.

When applying the proposed model specification, we aim at performing a joint, hierarchical, partitioning of European countries into blocks and, among such blocks, identifying clusters of residents sharing similar digital skills. In addition, we aim at taking into account how residents' socio-economic and demographic features influence the propensity of being digitally skilled. In Section 7.1, we provide a description of the data at hand, while Section 7.2 presents the results of the analysis.

7.1 Data description

As detailed above, we focus on ESS data that involve the mastery of various digital technologies by older residents in different countries. In this framework, countries identify second-level units, residents identify first-level units, and digital skills the variables of interest.

To simplify the data structure and ease the interpretation of results from subsequent analyses, we dichotomize the data on digital abilities. In the corresponding binary three-way data matrix, a 1 is reported if an individual states to use the Internet "a few times a week", "most days" or "every day", or to be "somewhat familiar", "very familiar", or "completely familiar" with preference settings, advanced search, and PDFs, or to video call or send messages to children aged 12 or over, parents, managers, and colleagues at least once in the last 12 months ("often in a day", "once a day", "often in a week", "often in a month", "once a month", or "less often"), or to have posted or shared anything about politics online in the last 12 months.

We report in Table 6 the distribution of residents who master the above technologies across countries and overall. From the last row of Table 6, we can see that 67% of the respondents frequently use the Internet, while this percentage reduces to 50% when focusing on preference settings, advanced search, and PDFs use. The percentage of those who frequently make video calls is 35%, while about 51% of older residents send messages. Lastly, only 11% of respondents share online political posts. As regards the comparison between countries, it can be seen that, in general, the percentage of respondents who own the different digital skills is equal to or greater than 50% in Belgium, Czechia, Fin-

land, Iceland, Netherlands, Norway, Switzerland, and UK. This percentage is around 40% in France, Ireland, Italy, and Slovakia, while it is lower than 40% in Bulgaria, Croatia, Estonia, Hungary, Lithuania, Montenegro, North Macedonia, Portugal, and Slovenia. The digital gap between countries is particularly evident in terms of Internet use, while it narrows when considering advanced skills such as the use of messages, video calls, and the sharing of online political posts.

Selecting only respondents with no missing records, we obtain a three-way binary data matrix made of 13'174 first-level units (residents aged 50 and above) nested within 21 second-level units (countries), and 7 variables (digital skills).

We also consider relevant information included in the ESS on socio-demographic and economic features of respondents: age, self-rated overall health, being hampered in daily activities, being born in the country of current residence, gender, educational level, legal marital status, partner's education, income, having/having had children aged 12 or over in the household (coded as "Children in the household"), and respondents' main employment activity. For the sake of interpretability, these variables are all transformed into a suitable number of dummy variables or into variables with a smaller number of categories. Table 7 reports the distribution of such transformed variables, from which it is evident that approximately 50% of respondents are between 50 and 65 years of age, in good health, not hampered in daily activities, of female gender, with a partner, a medium/high educational level and income, with children aged 12 or over, and retired. Also, about 92% of older respondents were born in the country of residence and about 25% of them have a partner with a medium educational level.

7.2 Analysis

In this section, we analyze the digital divide in Europe using the proposed modeling approach. We aim at performing a hierarchical clustering of countries and residents, also taking into account how socio-economic and demographic features influence the propensity of being digitally skilled.

For the sake of interpretability, we estimate the proposed model letting $\boldsymbol{w}_{gk}$ be constant across clusters ($\boldsymbol{w}_{gk} = \boldsymbol{w}_k$). The model is estimated for varying G, D, and Q. The optimal model is selected according to the BIC index. To prevent the estimation algorithm from remaining trapped in local maxima, a multi-start strategy based on 10 random starts is considered. As regards the concomitant variables affecting the latent layer of the model, the categorical variables described in Table 7 are considered.

7.2.1 Choosing the number of blocks and clusters

We start by reporting in Table 8 the BIC values for varying G, Q, and D. The optimal specification (i.e., the one providing the smallest BIC value) corresponds to $G = 4$ clusters of residents, $Q = 2$ blocks of countries, and $D = 1$ for the latent trait size.

7.2.2 Describing the blocks of countries

The two blocks of countries identified by the proposed model are represented in Figure 2a. As it is evident, the model clusters Belgium, Czechia, Estonia, Finland, France, Ireland, Iceland, Netherlands, Norway, and Switzerland in the first block, while Bulgaria, Croatia, Hungary, Italy, Lithuania, Montenegro, North Macedonia, Portugal, Slovenia, Slovakia, and United Kingdom are assigned to the second block. This partition clearly reflects the estimated prior block probabilities ($\hat{\rho}_1 = 0.48$ and $\hat{\rho}_2 = 0.52$), even though these are updated on the basis of the observed data when computing the approximate posterior block probabilities.

For a better understanding of these blocks, we look at the Digital Economy and Society Index (DESI; European Commission 2022), which measures the progress in digitalization of European member states. The distribution of this index across countries is reported in Figure 2b. Here, the darker the color, the higher the index. Comparing the graphical representations in Figures 2a and 2b, it is evident the association between the identified blocks and the DESI levels. This is not obvious, given that DESI is a general measure of countries' progress in terms of digitalization and is a composite index across four dimensions. The first one (human capital) measures users' skills in the general population, i.e., it is not limited to older people, while the other dimensions (connectivity, integration of digital technology, and digital public services) measure factors, such as broadband coverage, that may facilitate the use and learning of digital technologies. More in detail, when looking at the distribution of the DESI across blocks, we observe that its first, second, and third quartiles are equal to 0.52, 0.62, and 0.65, respectively, in the first block. On the other hand, countries in the second block all have a DESI below 0.57; the first, the second, and the third quartiles are here equal to 0.44, 0.48, and 0.53, respectively, thus suggesting that this block clusters the least digitalized countries in Europe.

In addition, we expect that countries characterized by higher digital skills are also those in better economic conditions, as well as with more educated residents. In this regard, we report in Figure 2c and 2d the distribution of the GDP at market prices (euro per capita; Eurostat 2022a) and the percentage of people aged 15-74 years with tertiary education (Eurostat 2022b) across countries, respectively. These figures confirm our beliefs, with countries classified in the first block (i.e., those with a higher digital level) corresponding to those with higher GDP values and higher percentage of highly educated individuals (see, e.g., Finland, France, Ireland, Norway and Switzerland). On the contrary, lower GDP

Table 6 Distribution of digital skills across countries and overall after dichotomization (% values). Data from residents aged 50 and above.

Country	Internet use	Preference settings	Advanced search	PDFs	Video calls	Messages	Political posts	Overall
Belgium	80	56	64	56	34	57	14	52
Bulgaria	50	32	33	26	30	29	7	30
Croatia	52	46	38	34	24	38	5	34
Czechia	71	42	45	44	64	77	8	50
Estonia	65	38	42	47	24	38	9	38
Finland	84	72	68	59	41	75	11	59
France	75	45	53	56	28	49	12	45
Hungary	48	42	43	33	31	41	7	35
Ireland	73	41	46	49	53	70	10	49
Iceland	92	37	61	59	37	61	23	53
Italy	61	42	42	40	35	51	12	40
Lithuania	58	35	40	35	35	42	10	36
Montenegro	62	38	37	34	48	47	7	39
Netherlands	93	73	80	77	38	72	16	64
North Macedonia	42	24	23	14	42	37	8	27
Norway	95	72	73	79	44	76	23	66
Portugal	49	32	28	27	25	31	6	28
Slovenia	59	43	44	41	18	43	10	37
Slovakia	54	45	49	50	40	53	9	43
Switzerland	84	72	76	70	32	60	13	58
UK	83	54	56	63	43	65	21	55
Overall	67	47	49	47	35	51	11	44

Table 7 Distribution of demographic and socio-economic variables after transformation. Pooled sample of residents aged 50 and above from all countries.

Variable	Category	%	Variable	Category	%
Age	$50 \vdash 65$	49	Health	Bad	12
	$65 \vdash 75$	31		Fair	35
	75+	20		Good	53
Hampered in daily activities	Yes	37	Born in country	Yes	92
	No	63		No	8
Gender	Male	46	Marital status	With partner	58
	Female	54		Without partner	42
Education	Low	27	Household's total net income	Low	37
	Medium	39		Medium	33
	High	35		High	29
Children in the household	No children	41	Main activity	Employed	37
	Children < 12	5		Unemployed	13
	Children ≥ 12	54		Retired	50
Partner's education	Low	15			
	Medium	25			
	High	21			
	Not applicable	39			

Table 8 BIC for varying G, Q, and D. The selected model is highlighted in bold.

		G				
$Q = 1$		1	2	3	4	5
D	1	93634.88	88459.42	86241.20	87033.70	93409.04
	2	91400.51	88557.90	85391.04	88472.25	99293.92
	3	91447.96	88539.75	86425.04	91238.00	90875.28
$Q = 2$						
D	1	93634.88	84126.13	82783.09	**81848.99**	82239.38
	2	91400.51	84135.27	83029.45	82009.49	81925.86
	3	91447.96	84341.12	82945.94	81995.58	82904.78
$Q = 3$						
D	1	93634.88	83808.40	82803.18	81897.24	83575.87
	2	91400.51	83929.99	82839.41	83425.37	81948.35
	3	91447.95	84082.13	82974.73	82204.79	81858.22

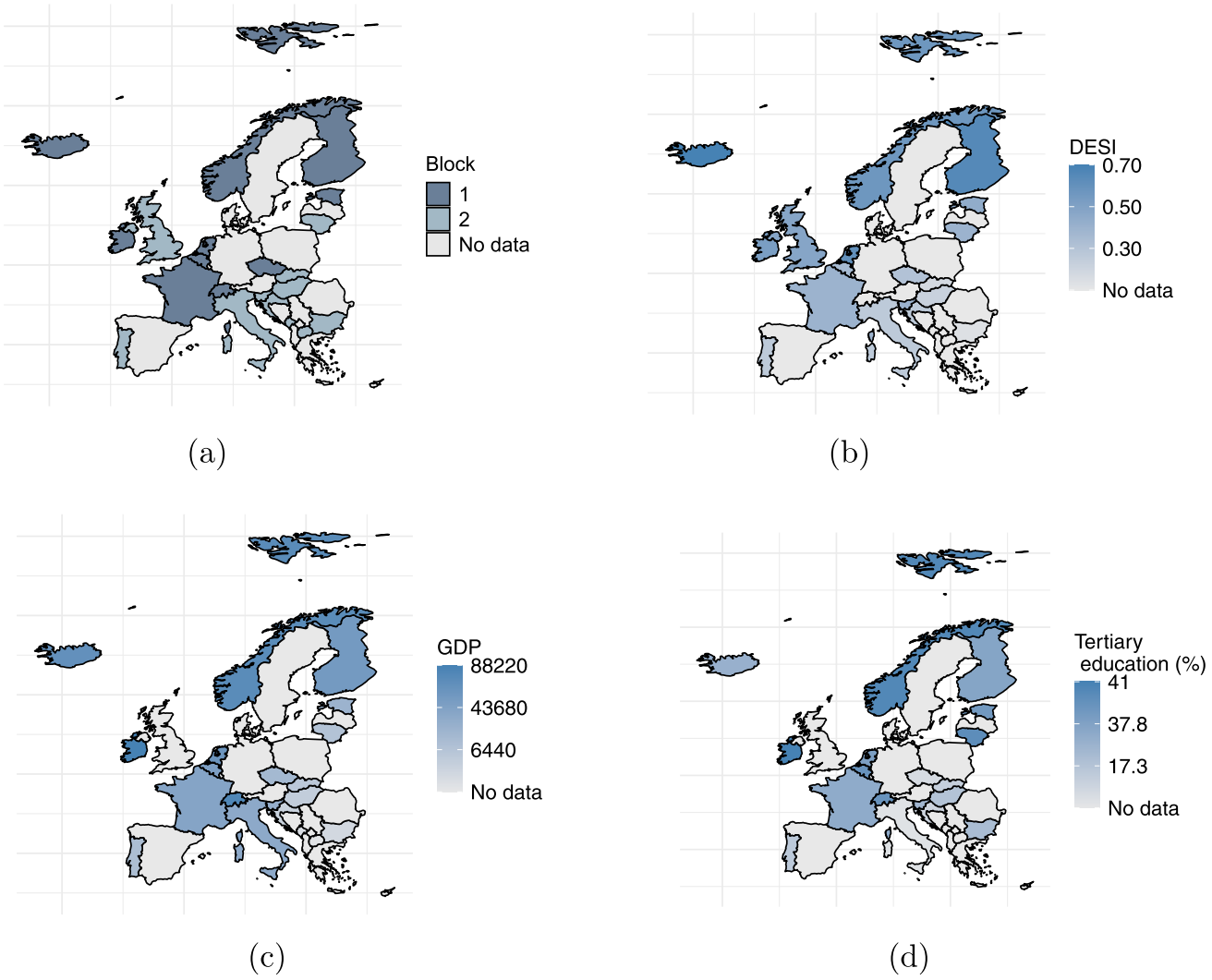

Fig. 2 Map comparison for the countries considered in the analysis. The different shades identify the two different blocks of countries (a), the different DESI values (b), the GDP values at market prices in euro per capita (c), and the % of people aged 15-74 with a tertiary educational level (d) across countries.

Table 9 Proportion of residents assigned to each cluster ($g = 1, 2, 3, 4$) across country blocks ($q = 1, 2$).

		g=1	g=2	$g = 3$	g=4
q	1	0.06	0.11	0.22	0.61
	2	0.28	0.16	0.20	0.36

values and a lower prevalence of highly educated residents are observed for those countries classified in the second block (the least digitalized one).

7.2.3 Describing the clusters of residents

Table 9 shows the proportion of residents assigned to the different clusters across country blocks. As we can see from this table, 61% of residents are assigned to cluster 4 within the block of countries with a higher digitalization level ($q = 1$). On the other hand, only 6% of residents are classified in cluster 1 within this block of countries. In contrast, within lower digitalized countries (q=2), residents are more equally split across the four clusters.

To interpret residents' clusters, we report in Table 10 the estimated b_{gk} and w_k parameters, with the corresponding 95% bootstrap confidence intervals for the optimal model specification. In this context, b_{gk} represents the baseline tendency of individuals in cluster g to master the k-th digital technology, while w_k captures the influence of the resident-specific latent trait on the probability of mastering the k-th technology. Looking at Table 10, we observe that the baseline tendency coefficients in the first cluster, i.e., $\boldsymbol{b}_1 = (b_{11}, \ldots, b_{1R})'$, are all negative and statistically different from zero. This means that residents in this cluster do have a negative propensity toward all digital skills considered in the analysis. A similar result is observed for residents in cluster 2, where the only positive and significant coefficient in the vector $\boldsymbol{b}_2 = (b_{21}, \ldots, b_{2R})'$ is the one related to message exchanges. Cluster 3 differs from cluster 2 in that residents in this cluster use the Internet frequently in addition to message exchanges. Lastly, cluster 4 is characterized by an intensive use of all digital technologies considered here, with the only exception of video calls and sharing of online political posts.

Regarding the influence coefficients w_k reported in the last column of Table 10, these have to be interpreted in terms of residual heterogeneity between residents (allocated in any of the clusters) in the way they master the k-th digital technology with respect to the cluster-specific baseline level dictated by b_{gk}. By looking at the table, we observe that these parameters are all positive and significantly different from 0, with the exception of that related to advanced search ability. Focusing the attention on their magnitude, we may observe that heterogeneity is particularly relevant for video calls and messages.

To better interpret the main characteristics of the identified clusters of residents, Figure 3 shows the distribution of the predicted probabilities of owning each digital skill across clusters. More specifically, for each g, the figure shows the box-plot of the predicted probabilities $\hat{\pi}_{gk}(\hat{\boldsymbol{u}}_{hi})$ obtained from Equation (1) by replacing $\boldsymbol{u}_{hi}$ with the corresponding posterior mean $\hat{\boldsymbol{u}}_{hi}$, computed as:

$$\hat{\boldsymbol{u}}_{hi} = \mathbb{E}[\boldsymbol{u}_{hi} \mid \boldsymbol{y}_{hi}] = \int_{\mathbb{R}^D} \boldsymbol{u}_{hi} f(\boldsymbol{u}_{hi} \mid \boldsymbol{y}_{hi}) d\boldsymbol{u}_{hi}.$$

The plot also reports the average predicted probabilities across residents (red bullet). From this figure, it is evident that, for the first cluster, predicted probabilities are zero or very close to zero, with the exception of Internet use for which the mean is 0.10. For the second cluster, the average probability of mastering the different digital technologies is high only when looking at video calls and messages (around 0.50); in this cluster, the average predicted probability of frequent Internet use increases to 0.20. From this result, we can speculate that individuals in this cluster use the Internet rather rarely but, when they do, it is mostly for video calls and messaging. In the third cluster, all average probabilities tend to be higher than in the first and second cluster (with the only exception of video calls and messages); in particular, the average probability of frequently using the Internet reaches 0.89. Lastly, the fourth cluster is characterized by a high probability of mastering each of the technologies considered in this analysis, with the exception of sharing online political posts, for which the average predicted probability is the highest across all clusters (0.22), though.

Based on these results, we can attribute the following labels to the clusters and summarize their main characteristics as follows:

(1) "Digital Laggards":
   (i) very low percentages of digital tool usage and digital skills;
   (ii) limited engagement with various digital technologies.

(2) "Communication Technologies Users":
   (i) higher percentages for video calls and text messaging compared to "Digital Laggards";
   (ii) low digital skills.

(3) "Web Navigators":
   (i) similar to "Communication Technologies Users" but with a significant portion frequently using the Internet;
   (ii) moderate digital skills.

(4) "Digital Proficients":

Table 10 Estimates for the baseline tendency parameters b_{gk} and influence parameters w_k and 95% bootstrap confidence intervals.

	$\boldsymbol{b}_1$	$\boldsymbol{b}_2$	$\boldsymbol{b}_3$	$\boldsymbol{b}_4$	$\boldsymbol{w}$
Internet	-2.33 (-2.63; -2.04)	-1.67 (-3.2; -0.31)	1.94 (1.58; 2.30)	3.24 (3.11; 3.38)	0.12 (0.09; 0.16)
Pref. settings	-5.31 (-6.11; -4.50)	-4.46 (-5.50; -3.42)	-1.40 (-1.69; -1.10)	2.44 (2.31; 2.56)	0.04 (0.01; 0.07)
Adv. search	-6.20 (-7.30; -5.10)	-4.86 (-5.78; -3.93)	-1.21 (-1.54; -0.88)	3.29 (3.07; 3.50)	0.01 (-0.02; 0.03)
PDFs	-6.56 (-7.91; -5.22)	-5.28 (-5.98; -4.58)	-1.05 (-1.31; -0.79)	2.15 (2.06; 2.25)	0.05 (0.03; 0.07)
Video calls	-13.66 (-15.75; -11.57)	0.19 (-0.02; 0.40)	-0.76 (-0.93; -0.59)	-0.57 (-0.67; -0.48)	1.09 (1.04; 1.13)
Messages	-13.14 (-16.67; -9.61)	0.60 (0.41; 0.78)	0.20 (0.04; 0.36)	0.52 (0.34; 0.71)	1.10 (1.05; 1.15)
Political posts	-7.76 (-8.67; -6.85)	-5.60 (-6.21; -4.98)	-2.17 (-2.35; -1.99)	-1.42 (-1.48; -1.35)	0.14 (0.10; 0.19)

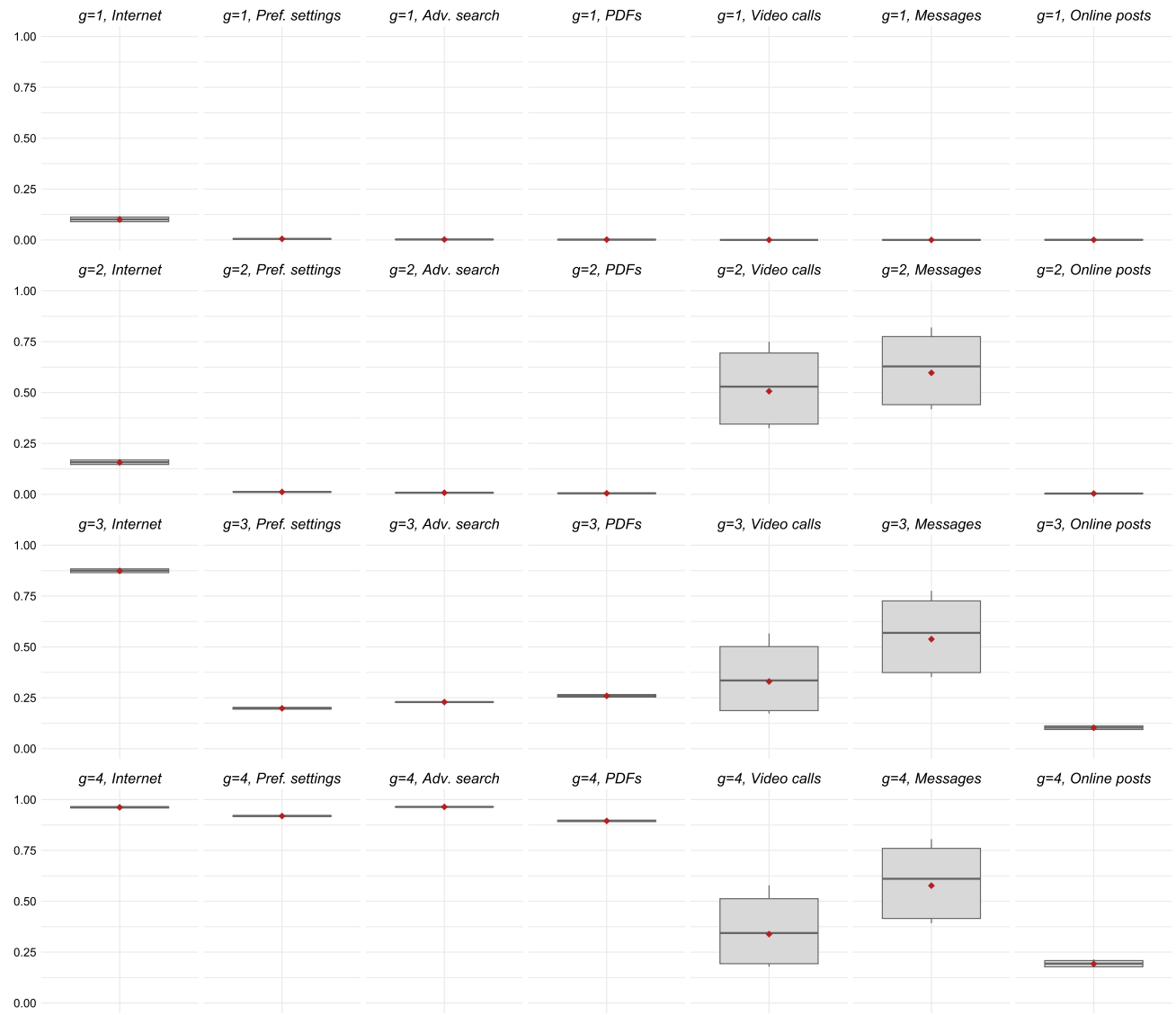

Fig. 3 Distribution of the predicted probabilities of mastering each digital technology for residents belonging to each cluster. Mean values are highlighted in red.

(i) highest percentages among all clusters of digital skills and use of digital technologies;
(ii) engaged across a wide range of digital tools and technologies.

Note that the identified clusters differ not only with respect to the general level of digital proficiency of residents, but also with respect to different types of digital tools used.

7.2.4 Understanding covariates' influence on cluster formation

Table 11 shows estimates and 95% bootstrap confidence intervals of the $\boldsymbol{\beta}_g$ parameters measuring the effect that demographic and socio-economic variables have on residents' partitioning. In particular, these parameters measure the impact of concomitant variables on the probability that residents in a given country belong to cluster $g = 2, 3, 4$ with respect to the first cluster, the "Digital Laggards". From this table, we can see that the probability of belonging to the "Communication Technologies Users" (cluster 2), as opposed to the "Digital Laggards" cluster, and compared to the respective reference categories, is significantly higher for residents between 65 and 75 years old, in fair health conditions, not hampered in daily activities, with a medium educational level and a low/medium educated partner, medium/low income, with young children or no children at all, retired or unemployed. The probability of belonging to the "Web Navigators" (cluster 3), is higher for individuals who are between 65 and 75 years old, in fair/good health conditions, not hampered in daily activities, males, with medium income and educational level and a low/medium educated partner, with young children or no children at all, and unemployed. On the other hand, the probability of belonging to cluster 3 with respect to cluster 1 is lower for those over 75 years of age, with a low level of education, low income, and retired. Lastly, the probability of belonging to the "Digital Proficients" (cluster 4) is higher for those who are in fair/good health conditions, not hampered in daily activities, born in the country of residence, males, in a relationship, with a low/medium educated partner, a medium income level, with young children or no children at all, and unemployed. In contrast, this probability is lower for those who are over 75 years of age, have a low/medium level of education, have a low income, and are retired.

As far as the vector of block-specific intercepts $\boldsymbol{\gamma}_q^* = (\gamma_{1q}^*, \gamma_{2q}^*, \gamma_{3q}^*)'$ is concerned, its estimate is $\hat{\boldsymbol{\gamma}}_1^* = (-0.76, 0.49, 1.41)'$ for $q = 1$, and $\hat{\boldsymbol{\gamma}}_2^* = (-2.36, -1.11, -0.19)'$ for $q = 2$. These values are coherent with the findings discussed in Section 7.2.2 regarding the DESI. That is, thanks to these parameters, we are able to capture the latent digitalization level measured by such an index.

Lastly, we report in Table 12 the predicted probabilities of belonging to the different clusters of residents for some attributes of interest and for the two blocks of countries identified by the model. Younger residents (i.e., younger than 65 years of age) display the highest probability of belonging to the cluster of the "Digital Proficients" (g=4) in both blocks of countries. They are also more likely than older individuals to belong to the "Web Navigators" cluster (g=3), but only in the block of most digitalized countries ($q = 2$). Within the block of least digitalized countries, we find a considerably higher probability of belonging to the cluster of "Communication Technologies Users" ($g = 2$) for the oldest individuals (i.e., those aged 75 and more). Interestingly, the probability of belonging to the "Digital Laggards" (g=1) cluster displays very different patterns in the two blocks of countries, while, within the block of highly digitalized countries, residents aged 65 and above report considerably higher probabilities of belonging to the digitally excluded cluster than younger individuals. Very small differences across age classes are observed in the first block of countries. Education and income show similar patterns of effects on the probability of belonging to the four clusters of residents within both blocks of countries. In fact, individuals with the highest level of education or income display substantially higher probabilities of belonging to the "Digital Proficients" cluster ($g = 4$) compared with individuals with lower educational or income levels, besides the block the country of residence belongs to. Within the most digitalized countries ($q = 1$), low and medium education or income are associated with higher probabilities of being in the "Digital Laggards" cluster ($g = 1$) compared with individuals with high education or income. This is not found within the least digitalized block of countries ($q = 2$) where, instead, low and medium education or income lead to the highest probabilities of belonging to the "Communication Technologies Users" ($g = 2$). In both blocks of countries, individuals with better health conditions ("fair" or "good") have higher probabilities of being "Digital Proficients" ($g = 4$) than their counterparts who self-report a "bad" health status. Lastly, those with young children are (slightly) more likely to belong to the "Digital Proficients" cluster ($g = 4$) and less likely to be in the "Communication Technologies Users" ($g = 2$) than childless individuals, in both blocks of countries.

8 Conclusions

The proposed model extends the Mixture of Latent Trait Analyzers with concomitant variables and provides a hierarchical clustering of three-way binary data. This is achieved by considering a hierarchical mixture of latent trait models. In detail, first-level units are partitioned via an MLTA specification. Here, latent class probabilities are modeled as a

Table 11 Estimated $\boldsymbol{\beta}_g$ parameters and 95% bootstrap confidence intervals. Reference categories are in brackets.

Variable	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_3$	$\boldsymbol{\beta}_4$
Age 65-75 (50-65)	0.47 (0.36; 0.58)	0.28 (0.16; 0.40)	0.10 (-0.05; 0.25)
Age 75+ (50-65)	0.28 (-0.12; 0.68)	-0.32 (-0.63; -0.02)	-0.86 (-1.17; -0.54)
Health fair (bad)	0.12 (0.02; 0.22)	0.18 (0.12; 0.24)	0.47 (0.35; 0.59)
Health good (bad)	0.26 (-0.10; 0.62)	0.64 (0.56; 0.72)	1.41 (1.26; 1.57)
Not hampered	0.28 (0.09; 0.47)	0.16 (0.11; 0.22)	0.26 (0.21; 0.32)
Born in country	-0.01 (-0.06; 0.04)	0.01 (-0.05; 0.07)	0.26 (0.17; 0.35)
Male	0.04 (-0.01; 0.09)	0.21 (0.13; 0.30)	0.28 (0.23; 0.33)
Educ. low (high)	0.08 (-0.25; 0.40)	-0.43 (-0.54; -0.32)	-1.91 (-2.18: -1.64)
Educ. medium (high)	0.28 (0.06; 0.49)	0.28 (0.10; 0.45)	-0.62 (-0.74; -0.49)
With partner	0.16 (-0.01; 0.33)	0.27 (0.19; 0.36)	0.36 (0.24; 0.47)
Partner educ. low (high)	0.28 (0.09; 0.46)	0.28 (0.22; 0.33)	0.28 (0.22; 0.33)
Partner educ. med. (high)	0.27 (0.04; 0.52)	0.28 (0.17; 0.39)	0.28 (0.14; 0.41)
Income low (high)	0.28 (0.08; 0.47)	-0.10 (-0.16; -0.05)	-0.76 (-0.87; -0.65)
Income medium (high)	0.28 (0.23; 0.33)	0.28 (0.14; 0.41)	0.07 (0.01; 0.12)
Children <12 (≥ 12)	2.37 (2.29; 2.44)	1.16 (0.83; 1.49)	1.48 (1.22; 1.75)
No children (≥ 12)	1.74 (1.68; 1.80)	0.75 (0.50; 1.00)	1.04 (0.88; 1.20)
Retired (employed)	0.28 (0.03; 0.53)	-0.39 (-0.44; -0.34)	-0.60 (-0.80; -0.40)
Unemployed (employed)	0.67 (0.61; 0.72)	0.28 (0.18; 0.37)	0.28 (0.13; 0.42)

Table 12 Predicted probabilities of belonging to each cluster for some covariates of interest and for the two blocks of countries.

		Concomitant variable													
		Age			Health			Education			Income			Children	
		<65	$65 \vdash 75$	75+	Bad	Fair	Good	Low	Med.	High	Low	Med.	High	<12	No
q=1	g=1	0.11	0.13	0.19	0.38	0.14	0.08	0.25	0.16	0.11	0.19	0.13	0.15	0.16	0.05
	g=2	0.24	0.35	0.43	0.34	0.25	0.16	0.43	0.35	0.19	0.40	0.31	0.21	0.25	0.51
	g=3	0.31	0.28	0.24	0.26	0.26	0.24	0.26	0.35	0.30	0.27	0.31	0.30	0.27	0.19
	g=4	0.34	0.24	0.14	0.02	0.35	0.52	0.06	0.14	0.40	0.14	0.25	0.34	0.32	0.25
q=2	g=1	0.02	0.42	0.55	0.35	0.44	0.29	0.62	0.49	0.06	0.53	0.44	0.02	0.19	0.21
	g=2	0.26	0.23	0.24	0.33	0.16	0.13	0.22	0.21	0.17	0.23	0.20	0.28	0.26	0.42
	g=3	0.34	0.19	0.13	0.18	0.17	0.18	0.13	0.21	0.33	0.16	0.20	0.32	0.27	0.16
	g=4	0.38	0.16	0.08	0.14	0.23	0.40	0.03	0.09	0.44	0.08	0.16	0.38	0.28	0.21

function of both concomitant variables and second-level random effects. The distribution of these latter is left unspecified and is directly estimated from the data via a NPML approach. This leads to the estimation of a discrete distribution that is totally free of stringent parametric assumptions and provides a clustering of second-level units. Overall, we obtain a partitioning of second-level units into blocks and, in each of such blocks, a partitioning of first-level units into clusters. A double EM algorithm based on a variational approximation of the model log-likelihood function is used to derive model parameter estimates. A simulation study assesses the effectiveness of the proposal in correctly identifying model parameters and providing a good classification of both first- and second-level units. The good performance of the model selection criterion and the computational efficiency of the estimation method are also demonstrated.

The proposed model is applied to the analysis of digital skills among older people in different countries, with the goal of identifying clusters of respondents with similar profiles of digital skills, also taking into account the multilevel structure of the data (i.e., residents nested within countries) and the influence of demographic features on clustering. The model identifies four different clusters of residents who differ not only for their overall level of digital skills, but also with respect to the specific digital tools used and skills they declare to have. Furthermore, the country-specific, discrete, random effect estimated via the NPML approach allows countries to be classified in terms of the baseline attitude to digital technologies of their residents. Thus, countries are clustered in

two blocks with different average levels of digital skills of their residents. The results show that the digitalization profiles are strongly influenced by age, income and educational level, as well as by health and by the presence of children in the household.

Both the simulation study and the real data analysis have been carried out using the open-source R software (Pinheiro and Bates 1995). All the codes and data used in this paper are available on the first author's GitHub page: https://github.com/dfailli.

It is worth noting that while NPML provides a flexible approach for modeling unobserved heterogeneity without requiring strong parametric assumptions, it has some potential limitations. First, the method is sensitive to the choice of the number of support points in the mixing distribution. Too few points may lead to an oversimplified model that fails capturing the underlying variability, while too many may result in numerical instability. Second, model identifiability may be problematic since different combinations of support points and weights can yield similar likelihood values. These limitations highlight the importance of model diagnostics and careful interpretation of results.

From a methodological point of view, the proposed approach can be extended to properly handle response variables with more than two categories, as well as in contexts of more than two levels of clustering; concomitant variables at each level may also be considered for a deeper understanding of the phenomenon of interest. Furthermore, dealing with missing data poses a number of challenges because of the potential bias in parameter estimates due to incomplete information. In our analysis, we assume that the data are missing at random (MAR), meaning that the probability that responses are missing depends on the set of observed measures, but it is unrelated to unobserved values that, in principle, should have been obtained (Rubin 1976). In this respect, the proposed model can be extended to deal with missing not at random (MNAR) data. To generalize the results obtained to the entire population, a sensitivity analysis comparing the models for MAR and MNAR data is required. Lastly, in the multilevel LC model framework, a number of goodness-of-fit statistics are available in the literature to assess the adequacy of model assumptions (Nagelkerke et al. 2015). A further line of research could involve the extension of such measures to the present framework.

Author Contributions Dalila Failli led all parts of the work. Bruno Arpino and Maria Francesca Marino have equally contributed to this work.

Funding Open access funding provided by Università degli Studi di Perugia within the CRUI-CARE Agreement. The work of Dalila Failli has been developed with the support of the Italian Ministry of University and Research under grant "FSE REACT-EU 2022-2024" and the European Union – Next Generation EU, Mission 4 Component 1 CUP J33C22002910001 – Project INCLUDE-ALL: social INCLUsion and Digital indicators Estimation At Local Level. Maria Francesca Marino gratefully acknowledge financial support by the Italian Ministry of University and Research under the PRIN 2022 Call (D.D. 104/2022) — Project title "Latent variable models and dimensionality reduction methods for complex data" – Project No. 20224CRB9E, CUP B53C24006320006, as well as the Department of Excellence project 2023-2027 — Project title "ReDS 'Rethinking Data Science" — Department of Statistics, Computer Science, Applications — University of Florence. The work of Bruno Arpino has been developed with the support of the Italian Ministry of University and Research under the project "Socio-demographic determinants of Information and Communication Technologies (ICT) use among older people in Italy (ICTAGE)".

Availability of data and materials The data that support the findings of this study are openly available at https://ess-search.nsd.no/en/study/172ac431-2a06-41df-9dab-c1fd8f3877e7

Declarations

Disclosure statement The authors report there are no competing interests to declare.

Competing interests The authors declare no competing interests.

References

Abendroth, A., Lükemann, L., Hargittai, E., Billari, F., Treas, J., Van Der Lippe, T.: Digital Social Contacts in Work and Family Life Topline results From Round 10 of the European Social Survey, ESS Topline Results Series, Issue 12 (2023). https://europeansocialsurvey.org/sites/default/files/2023-10/TL012_Digital-Social-Contacts-English.pdf

Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th edn. Dover, New York (1964)

Agresti, A.: Categorical data analysis. Wiley, New York (2002)

Aitkin, M.: A general maximum likelihood analysis of overdispersion in generalized linear models. Stat. Comput. **6**, 251–262 (1996)

Aitkin, M.: A general maximum likelihood analysis of variance components in generalized linear models. Biometrics **55**(1), 117–28 (1999)

Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**, 716–723 (1974)

Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. **22**(7), 719–725 (2000)

Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. J. R. Stat. Soc. Series B Stat. Methodol. **39**(1) 1–38 (1977)

Efron, B.: Bootstrap Methods: Another Look at the Jackknife. Ann. Stat. 7(1), 1–26 (1979)
ESS Round 10: European Social Survey Round 10 Data. Data file edition 3.1. Sikt -Norwegian Agency for Shared Services in Education and Research, Norway - Data Archive and distributor of ESS data for ESS ERIC. (2020). https://doi.org/10.21338/NSD-ESS10-2020.
ESS (2022). Note on ESS Round 10 data releases December 2022 (2022). https://www.europeansocialsurvey.org/sites/default/files/2023-06/ESS10_note_for_data_users_0.pdf. Accessed on 2 december 2023
European Commission. Digital Economy and Society Index (DESI) reports (2022)
Eurostat. Gross domestic product at market prices https://ec.europa.eu/eurostat/databrowser/view/tec00001/default/table?lang=en
Eurostat. Labour Force Survey (LFS) - Employment rate by sex, age group, and educational attainment. Retrieved from https://ec.europa.eu/eurostat/databrowser/view/edat_lfs_9911__custom_16060562/default/table?lang=en
Failli, D., Marino, M. F., Martella, F.: Finite mixtures of latent trait analyzers with concomitant variables for bipartite networks: an analysis of COVID-19 data. Multivar. Behav. Res. **59**(4), 801–817 (2024) https://doi.org/10.1080/00273171.2024.2335391
Ghidey, W., Lesaffre, E., Eilers, P.: Smooth random effects distribution in a linear mixed model. Biometrics **60**(4), 945–953 (2004)
Goldstein, H.: Bootstrapping in multilevel models.: In: Hox, J. J., Roberts, J. K. (eds.) Handbook for advanced multilevel analysis, pp. 163–171. Routledge/Taylor & Francis Group (2011)
Gollini, I., Murphy, T.B.: Mixture of latent trait analyzers for model-based clustering of categorical data. Stat. Comput. **24**(4), 569–588 (2014)
Gollini, I.: A mixture model approach for clustering bipartite networks. Challenges in Social Network Research Volume in the Lecture Notes in Social Networks (LNSN - Series of Springer) (2020)
Jaakkola, T. S., Jordan, M. I.: Bayesian logistic regression: A variational approach. In Madigan, D., Smyth, P. (Eds.) Proceedings of the 1997 Conference on Artificial Intelligence and Statistics. Ft. Lauderdale, FL (1997)
Laird, N.: Nonparametric maximum likelihood estimation of a mixing distribution. J. Am. Stat. Assoc. **73**, 805–811 (1978)
Laird, N.M., Ware, J.H.: Random-Effects Models for Longitudinal Data. Biometrics 38(4), 963–974 (1982). https://doi.org/10.2307/2529876
Lindsay, B.G.: The geometry of mixture likelihoods: a general theory. Ann. Stat. **11**, 86–94 (1983)
Lindsay, B.G.: The geometry of mixture likelihoods, Part II: the Exponential Family. Ann. Stat. **11**, 783–792 (1983)
Lukočienė, 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). https://doi.org/10.1111/j.1467-9531.2010.01231.x
Magder, L.S., Zeger, S.L.: A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. J. Am. Stat. Assoc. **91**(435), 1141–1151 (1996)
Nagelkerke, E., Oberski, D.L., Vermunt, J.K.: Goodness-of-fit of Multilevel Latent Class Models for Categorical Data. Sociol. Methodol. **46**(1), 252–282 (2015). https://doi.org/10.1177/0081175015581379
Niku, J., Warton, D.I., Hui, F.K.C., Taskinen, S.: Generalized linear latent variable models for multivariate count and biomass data in ecology. J. Agric. Biol. Environ. Stat.22, 498–522 (2017)
R Core Team.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2024). https://www.R-project.org/
Rand, W.M.: Objective Criteria for the Evaluation of Clustering Methods. J. Am. Stat. Assoc. **66**, 846–850 (1971)
Rubin, D.B.: Inference and Missing Data. Biometrika **63**(3), 581–592 (1976)
Schwarz, G.: Estimating the Dimension of a Model. Ann. Stat. **6**, 461–464 (1978)
Skrondal, A. Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models (1st ed.)* Chapman and Hall/CRC
Tipping, M. E. (1998). Probabilistic visualisation of high-dimensional binary data. In Proceedings of the 11th International Conference on Neural Information Processing Systems (NIPS'98).(pp. 592–598). MIT Press
Verbeke, G., Lesaffre, E.: A linear mixed-effects model with heterogeneity in the random-effects population. J. Am. Stat. Assoc. **91**(433), 217–221 (1996)
Vermunt, J.K., Van Dijk, L.: A Nonparametric Random-Coefficients Approach: The Latent Class Regression Model. Multilevel Modelling Newsletter **13**, 6–13 (2001)
Vermunt, J.K.: Multilevel latent class models. Sociol. Methodol. **33**, 213–239 (2003)
Vermunt, J.K., Magidson, J.: Hierarchical mixture models for nested data structures. In: Weihs, C., Gaul, W. (eds.) Classification: The Ubiquitous Challenge, pp. 176–183. Springer (2005)
Vermunt, J.K.: A hierarchical mixture model for clustering three-way data sets. Computational Statistics & Data Analysis **51**(11), 5368–5376 (2007). https://doi.org/10.1016/j.csda.2006.08.005
Zhang, D., Davidian, M.: Linear mixed models with flexible distributions of random effects for longitudinal data. Biometrics **57**(3), 795–802 (2001)