Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A dimensionally reduced finite mixture model for multilevel data

Daniela G. Calò [*], Cinzia Viroli

Department of Statistics, University of Bologna, via Belle Arti, 41, Bologna, Italy

ARTICLE INFO

Article history:
Received 9 July 2009
Available online 23 July 2010

AMS subject classifications:
primary 62H30
secondary 62H25

Keywords:
Cluster analysis
Factor mixture model
Dimension reduction
EM-algorithm
Multilevel latent class analysis

ABSTRACT

Recently, different mixture models have been proposed for multilevel data, generally requiring the local independence assumption. In this work, this assumption is relaxed by allowing each mixture component at the lower level of the hierarchical structure to be modeled according to a multivariate Gaussian distribution with a non-diagonal covariance matrix. For high-dimensional problems, this solution can lead to highly parameterized models. In this proposal, the trade-off between model parsimony and flexibility is governed by assuming a latent factor generative model.

1. Introduction

Multilevel data are collected in different applied fields of research, such as sociology, behavioral science and medicine. Data from students nested within schools or patients hospitalized in different primary care centers are typical examples of two-level data: pupils and patients are referred to as lower level units, while being schools and hospitals the respective higher level units. A two-level multivariate data set, where multiple responses are recorded for each lower level unit, can also be seen as a three-level data set by interpreting multivariate responses as nested univariate responses.

Different models have been proposed for describing multilevel data (see, for example, [17,18]). Recently, various types of latent class [9] and finite mixture models [13] for data sets having a multilevel structure have been proposed (see, for example, [20,21]). In particular, Vermunt [19,22] and Vermunt and Magidson [24] have introduced a hierarchical extension of these two model-based clustering methods, with the aim of clustering multilevel data on the basis of a set of categorical or continuous response variables. As in standard model-based clustering, the main goal of their proposal is to build a meaningful cluster model for the lower level units. However, in order to take the nested data structure into account, certain model parameters are allowed to differ randomly across higher level units. A particular variant—that is dealt with in the present paper—is based on the introduction of a discrete random effect that is assumed to capture higher level variation in lower level class membership probabilities. Since this amounts to assume that higher level units belong to a given number of latent classes, the resulting multilevel model contains a separate finite mixture distribution at each level of nesting, thus yielding a simultaneous model-based clustering of units at both the levels of the hierarchy.

The above-mentioned multilevel mixture model is presented in [22] by making the additional assumption that the observed responses are conditionally independent of one another given the class membership of the generic lower level unit (which is sometimes referred to as the *local independence assumption*). If the observed variables are modeled as a finite Gaussian mixture, this amounts to assume all the component covariance matrices to be diagonal. By removing such a

[*] Corresponding author.
E-mail address: danielagiovanna.calo@unibo.it (D.G. Calò).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.07.003

constraint, more flexible models can be defined. On the other hand, a fully unconstrained parameterization of the component covariance matrices involves a quadratic increasing number of parameters as dimensionality increases. When dealing with high-dimensional data, this choice could lead to unreliable estimates or to over-fitting problems.

An alternative method for dealing with local dependencies can be devised by interpreting a two-level multivariate data set as a three-level data set, with response variables representing the lowest level units in the hierarchy. It consists in introducing one (or more) continuous latent variable(s) in the lowest level of hierarchy (see [23], Section 3.3).

In the present paper, a specific mixture model belonging to this latter methodological framework is proposed that weakens the local independence assumption while governing the trade-off between model flexibility and parsimony. It represents the multilevel variant of the *Heteroscedastic Factor Mixture Model* (HFMM), which has been introduced by Montanari and Viroli [14] as an effective method for dimensionally reduced model-based clustering.

The basic idea is to assume that the p response variables are generated by q latent factors (with $q < p$), which are jointly distributed as a finite Gaussian mixture. Under these conditions, it can be proved that the observed variables are modeled as a finite Gaussian mixture as well, with parameters depending on the parameters of the factor distribution. In so doing, the number of model parameters is controlled via the dimensionality of the reduced latent space; in addition, the latent variables can provide useful information for better understanding the investigated phenomenon.

The remainder of the paper is organized as follows. After introducing some preliminaries in the next section, Section 3 provides a short description of the hierarchical finite mixture model for nested data developed in [24,22]. Section 4 presents the proposed clustering model and describes the restrictions needed for its identification. In Section 5, maximum likelihood estimation of model parameters via the EM algorithm is developed. Section 6 presents the classification performance of the proposed model on artificial two-level multivariate data sets with varying higher level clustering structures; moreover, a real data example is illustrated where the proposed model provides a meaningful and relatively more parsimonious model-based clustering.

2. Preliminaries

The multilevel models described in the following two sections deal with two-level p-variate data. However, three-level data notation and terminology will be used in their formulation since multivariate responses are conceived as nested univariate responses (which yields an extra level of nesting). For example, multivariate data observed on students nested within schools can be viewed in a hierarchical structure where third-level units are schools, second-level units are students and first-level units are the univariate responses nested within each student.

First-level units are referred to by the index h, ranging from 1 to p, second-level units are denoted by the index i, and third-level units are labeled by the index j, with $j = 1, \ldots, J$. The number of second-level units within third-level unit j is denoted by n_j, where $\sum_{j=1}^{J} n_j = n$.

Let $y_{hij} \in \mathbf{R}$ denote the value of the hth observed variable for second-level unit i within third-level unit j. Notation $\mathbf{y}_{ij} = [y_{hij}]_{h=1,\ldots,p}$ is used to refer to the p-dimensional vector of responses for case i within third-level unit j.

3. Hierarchical finite mixture model

The model proposed in [24,22] allows a simultaneous model-based clustering of third-level units into L latent classes and second-level units into K latent classes. Both L and K are assumed to be fixed but unknown. In the following, the two corresponding latent class variables will be represented as allocation vectors: namely, $\mathbf{s}_j = [s_{jl}]_{l=1,\ldots,L}$ and $\mathbf{r}_{ij} = [r_{ijk}]_{k=1,\ldots,K}$ for third-level and second-level units, respectively. More specifically, with reference to a (second- or third-level) unit, the generic component of an allocation vector takes value 1 in correspondence of the class the unit belongs to, and 0 otherwise.

The model consists of the following two parts:

$$f(\mathbf{y}_j) = \sum_{l=1}^{L} w_l \prod_{i=1}^{n_j} f(\mathbf{y}_{ij}|s_{j1} = 0, \ldots, s_{jl} = 1, \ldots, s_{jL} = 0) \tag{1}$$

$$f(\mathbf{y}_{ij}|s_{j1} = 0, \ldots, s_{jl} = 1, \ldots, s_{jL} = 0) = \sum_{k=1}^{K} \pi_{k|l} f(\mathbf{y}_{ij}|r_{ij1} = 0, \ldots, r_{ijk} = 1, \ldots, r_{ijK} = 0), \tag{2}$$

where $\mathbf{s}_j \sim \mathcal{B}(1; w_1, \ldots, w_l, \ldots, w_L)$ and $\mathbf{r}_{ij}|(s_{j1} = 0, \ldots, s_{jl} = 1, \ldots, s_{jL} = 0) \sim \mathcal{B}(1; \pi_{1|l}, \ldots, \pi_{k|l}, \ldots, \pi_{K|l})$, with $\mathcal{B}(1; \ldots)$ denoting the multinomial distribution with number of trials equal to 1.

In Eq. (1), third-level units are assumed to belong to one of L latent classes with prior probabilities equal to w_l (where $\sum_{l=1}^{L} w_l = 1$), and observations within third-level unit j are assumed to be mutually independent given the class membership of unit j.

In Eq. (2), the conditional density $f(\mathbf{y}_{ij}|\mathbf{s}_j)$ is modeled as a finite mixture: second-level units within third-level unit j are assumed to belong to one of K latent classes with prior probabilities equal to $\pi_{k|l}$ (where $\sum_{k=1}^{K} \pi_{k|l} = 1$). These latter probabilities depend on the class membership of unit j, whereas the component densities are assumed to not differ across third-level classes.

Finally, in the model defined by (1) and (2) the parameters $\{\pi_{k|l}, k = 1, \dots, K\}$ are allowed to differ randomly across third-level units, according to the discrete random effect described by $\mathbf{s}_j \sim \mathcal{B}(1; w_1, \dots, w_l, \dots, w_L)$.

In the case of finite mixtures of Gaussian distributions, Eq. (2) takes the following form:

$$f(\mathbf{y}_{ij}|s_{j1} = 0, \dots, s_{jl} = 1, \dots, s_{jL} = 0) = \sum_{k=1}^{K} \pi_{k|l} \phi_k^{(p)}(\mathbf{y}_{ij}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3}$$

where $\phi_k^{(p)}$ denotes the p-dimensional Gaussian density with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. Under the local independence assumption, the responses are taken to be mutually independent given $\mathbf{r}_{ij}$, then Eq. (3) simplifies to

$$f(\mathbf{y}_{ij}|s_{j1} = 0, \dots, s_{jl} = 1, \dots, s_{jL} = 0) = \sum_{k=1}^{K} \pi_{k|l} \prod_{h=1}^{p} \phi_k^{(1)}(y_{hij}; \mu_{hk}, \sigma_{hk}^2). \tag{4}$$

Maximum likelihood estimates of the parameters of model (1)–(2) can be obtained by an adapted EM algorithm [6] that has been implemented in the `Latent GOLD` software package [25]. Then, the units of both levels can be partitioned into clusters according to the criterion of the maximum estimated posterior class membership probability.

4. The proposed model

The proposed model includes the same two parts described in Eqs. (1) and (2), but contains a further part that serves to connect the p responses (i.e. first-level units) observed on the same second-level unit. This connection is described through a generative factor model involving q latent variables (with $q < p$). Therefore, the model involves not only the two latent class variables $\mathbf{s}$ and $\mathbf{r}$ introduced in Section 3, but also a further q-dimensional continuous latent variable linked to the first-level of the hierarchy.

More specifically, the additional assumption we introduce is that the p observed variables contained in vector $\mathbf{y}$ are generated by the following factor model:

$$\mathbf{y}_{ij} = \boldsymbol{\gamma} + \boldsymbol{\Lambda}\mathbf{z}_{ij} + \mathbf{u}_{ij}, \tag{5}$$

where $\mathbf{z}$ denotes the q-dimensional vector of factors, $\boldsymbol{\Lambda}_{p\times q}$ is the factor loading matrix and $\mathbf{u}$ represents a p-dimensional Gaussian term with zero mean and diagonal covariance matrix $\boldsymbol{\Psi}_{p\times p}$. In Eq. (7), notation $\mathbf{y}_{ij}$, $\mathbf{z}_{ij}$, and $\mathbf{u}_{ij}$ denotes the realizations of the correspondent random variables for the ith second-level unit within jth third-level unit.

Conditionally to third-level class, the factor distribution is modeled as a mixture of K Gaussian densities through the allocation variable $\mathbf{r}_{ij}|\mathbf{s}_j$ defined in Section 3:

$$f(\mathbf{z}_{ij}|s_{j1} = 0, \dots, s_{jl} = 1, \dots, s_{jL} = 0) = \sum_{k=1}^{K} \pi_{k|l} \phi^{(q)}(\mathbf{z}_{ij}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{6}$$

It can be shown that fitting the model in (6) amounts to fitting the following K-component Gaussian mixture model in the original p-dimensional space:

$$f(\mathbf{y}_{ij}|s_{j1} = 0, \dots, s_{jl} = 1, \dots, s_{jL} = 0) = \sum_{k=1}^{K} \pi_{k|l} \phi^{(p)}(\mathbf{y}_{ij}; \boldsymbol{\gamma} + \boldsymbol{\Lambda}\boldsymbol{\mu}_k, \boldsymbol{\Lambda}\boldsymbol{\Sigma}_k\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}), \tag{7}$$

where the parameters of the K Gaussian components share the same $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ matrices and depend on the parameters of the factor distribution, which is defined on a q-dimensional space.

It is worth noting that, in analogy to (2), class-conditional densities in (7) do not differ across third-level classes, whereas class membership probabilities $\pi_{k|l}$ depend on the class membership of third-level unit j. By comparing Eq. (7) to Eq. (2), it can be seen that the proposed model can provide a proper balance between model flexibility and parsimony in high-dimensional problems. In fact, local dependencies between observed variables are dealt with, but class-specific covariance matrices are defined on a dimensionally reduced space.

By making the assumptions in (5) and (6), common factor analysis and latent class analysis are combined. In fact, second-level latent classes have to be inferred from the data jointly with the underlying factor structure. This combination has been recently presented by Montanari and Viroli [14], as an heteroscedastic variant of the *Factor Mixture Model* [10], in which factor covariance matrices in (6) are held equal across classes.

The model presented in this section represents a multilevel extension of the *heteroscedastic factor mixture model* (HFMM), where a discrete random effect is assumed to affect class membership probabilities. A multilevel variant of the *factor mixture model* has also been explored by Allua [2], which however uses a continuous latent variable at the third-level of the nesting. For a review on modeling multilevel data by including multiple latent variables, see Asparouhov and Muthen [3].

4.1. Identifiability conditions

As is well known, the classical factor model requires specific restrictions to be identified. In particular, given an invertible matrix $\mathbf{A}$, the factor model (5) and the transformed factor model $\mathbf{y} = \boldsymbol{\Lambda}\mathbf{A}^{-1}\mathbf{A}\mathbf{z} + \mathbf{u} = \boldsymbol{\Lambda}^*\mathbf{z}^* + \mathbf{u}$ are indistinguishable.

In order to avoid this ambiguity, we impose the standard assumption that the factors have zero mean and identity covariance matrix:

$$E(\mathbf{z}) = \sum_{l=1}^{L} w_l \sum_{k=1}^{K} \pi_{k|l} \boldsymbol{\mu}_k = \mathbf{0}, \tag{8}$$

$$V(\mathbf{z}) = \sum_{l=1}^{L} w_l \sum_{k=1}^{K} \pi_{k|l} \left(\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top \right) - \left(\sum_{l=1}^{L} w_l \sum_{k=1}^{K} \pi_{k|l} \boldsymbol{\mu}_k \right) \left(\sum_{l=1}^{L} w_l \sum_{k=1}^{K} \pi_{k|l} \boldsymbol{\mu}_k \right)^\top = \mathbf{I}_q. \tag{9}$$

Despite restrictions in (8) and (9), model (5) is still invariant under orthogonal transformations. Hence, we further impose the constraint that $\boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}$ is diagonal with elements in the decreasing order (see, for instance, [12]). This latter condition introduces $q(q-1)/2$ restrictions on $\boldsymbol{\Lambda}$ to be uniquely defined.

As a consequence of the above-mentioned conditions, the set of free parameters of the proposed model consists of the $p \times q - q(q-1)/2$ factor loadings in $\boldsymbol{\Lambda}$, the p diagonal elements of $\boldsymbol{\Psi}$, the $(K-1) \times q$ elements of the component means, the $(K-1) \times q(q+1)/2$ elements of the full component covariance matrices and finally the $L(K-1)$ mixing proportions $\pi_{k|l}$ and the $L-1$ weights w_l.

5. Inference

Given the notation in Section 4, we define $\mathbf{y}_j$ as the full vector of responses for third-level unit j. The likelihood of the proposed model is

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{J} L_j(\boldsymbol{\theta}; \mathbf{y}_j), \tag{10}$$

where $\boldsymbol{\theta}$ collectively denotes the whole set of parameters and $L_j(\boldsymbol{\theta}; \mathbf{y}_j)$ is the marginal likelihood contribution for third-level unit j. Since observations within unit j are assumed to be mutually independent given the class membership of unit j, this latter function is defined as

$$L_j(\boldsymbol{\theta}; \mathbf{y}_j) = \sum_{l=1}^{L} w_l \prod_{i=1}^{n_j} f(\mathbf{y}_{ij} | s_{j1} = 0, \ldots, s_{jl} = 1, \ldots, s_{jL} = 0), \tag{11}$$

where $f(\mathbf{y}_{ij} | s_{j1} = 0, \ldots, s_{jl} = 1, \ldots, s_{jL} = 0)$ is given in Eq. (7).

Under the hierarchical structure of the model, the maximum likelihood estimation problem can be solved by means of the EM algorithm.

5.1. The complete likelihood

In order to derive the complete data log-likelihood, it is worth noting that the proposed model involves three different latent variables devoted to different tasks. As already described in Section 4, the latent variable $\mathbf{z}$ accomplishes the dimension reduction step, whereas variables $\mathbf{r}$ and $\mathbf{s}$ perform the classification of second- and third-level units, respectively.

The complete likelihood function

$$L_c(\boldsymbol{\theta}) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} f(\mathbf{y}_{ij}, \mathbf{z}_{ij}, \mathbf{r}_{ij}, \mathbf{s}_j; \boldsymbol{\theta}), \tag{12}$$

can be written as follows:

$$\begin{aligned} L_c(\boldsymbol{\theta}) &= \prod_{j=1}^{J} \prod_{i=1}^{n_j} f(\mathbf{y}_{ij} | \mathbf{z}_{ij}, \mathbf{r}_{ij}, \mathbf{s}_j; \boldsymbol{\theta}) f(\mathbf{z}_{ij} | \mathbf{r}_{ij}, \mathbf{s}_j; \boldsymbol{\theta}) f(\mathbf{r}_{ij} | \mathbf{s}_j; \boldsymbol{\theta}) f(\mathbf{s}_j; \boldsymbol{\theta}) \\ &= \prod_{j=1}^{J} \prod_{i=1}^{n_j} f(\mathbf{y}_{ij} | \mathbf{z}_{ij}; \boldsymbol{\theta}) f(\mathbf{z}_{ij} | \mathbf{r}_{ij}; \boldsymbol{\theta}) f(\mathbf{r}_{ij} | \mathbf{s}_j; \boldsymbol{\theta}) f(\mathbf{s}_j; \boldsymbol{\theta}). \end{aligned} \tag{13}$$

The four terms involved in the previous expression are defined as

$$f(\mathbf{y}_{ij} | \mathbf{z}_{ij}; \boldsymbol{\theta}) = \phi^{(p)}(\boldsymbol{\gamma} + \boldsymbol{\Lambda} \mathbf{z}_{ij}, \boldsymbol{\Psi})$$
$$f(\mathbf{z}_{ij} | \mathbf{r}_{ij}; \boldsymbol{\theta}) = \phi^{(q)}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

according to (5) and (6) respectively, and $f(\mathbf{r}_{ij} | \mathbf{s}_j; \boldsymbol{\theta}) = \prod_{k=1}^{K} \pi_{k|l}^{r_{ijk}}, f(\mathbf{s}_j; \boldsymbol{\theta}) = \prod_{l=1}^{L} w_l^{s_{jl}}$.

Remark 1. Function $f(\mathbf{z}_{ij} | \mathbf{r}_{ij}, \mathbf{s}_j; \boldsymbol{\theta})$ in (13) simplifies to $f(\mathbf{z}_{ij} | \mathbf{r}_{ij}; \boldsymbol{\theta})$ because the parameters defining the class-specific conditional distributions for the factors do not vary across third-level classes, as can be seen in Eq. (6).

Remark 2. In expression (13), $f(\mathbf{y}_{ij}|\mathbf{z}_{ij}, \mathbf{r}_{ij}, \mathbf{s}_j; \boldsymbol{\theta})$ can be replaced by $f(\mathbf{y}_{ij}|\mathbf{z}_{ij}; \boldsymbol{\theta})$ for the following reasons: as shown in Eq. (7), the class-conditional distributions for the observed variables do not depend on variable **s**, in direct analogy with Remark 1; moreover, **y** depends on **r** only through **z**, meaning that information provided by variable **r** is redundant when conditioning on variable **z**.

Remark 3. In the following, without loss of generality, it will be assumed that the p observed variables in vector $\mathbf{y}_{ij}$ are mean centered, so that $\boldsymbol{\gamma} = \mathbf{0}$.

Given the factorization in (13), the conditional expected value of the complete log-likelihood function evaluated at a given set of parameters, $\boldsymbol{\theta}'$, has the following form:

$$\begin{aligned} E_{\mathbf{z},\mathbf{r},\mathbf{s}|\mathbf{y},\boldsymbol{\theta}'}[\log L_c(\boldsymbol{\theta})] &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\int f(\mathbf{z}_{ij}|\mathbf{y}_j; \boldsymbol{\theta}')\log f(\mathbf{y}_{ij}|\mathbf{z}_{ij}; \boldsymbol{\theta})\mathrm{d}\mathbf{z}_{ij} \\ &+ \sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{k=1}^{K}\int f(\mathbf{z}_{ij}, r_{ijk}=1|\mathbf{y}_j; \boldsymbol{\theta}')\log f(\mathbf{z}_{ij}|r_{ijk}=1; \boldsymbol{\theta})\mathrm{d}\mathbf{z}_{ij} \\ &+ \sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{l=1}^{L}\sum_{k=1}^{K} f(r_{ijk}=1, s_{jl}=1|\mathbf{y}_j; \boldsymbol{\theta}')\log f(r_{ijk}=1|s_{jl}=1; \boldsymbol{\theta}) \\ &+ \sum_{j=1}^{J}\sum_{l=1}^{L} f(s_{jl}=1|\mathbf{y}_j; \boldsymbol{\theta}')\log f(s_{jl}=1; \boldsymbol{\theta}), \end{aligned} \tag{14}$$

where $f(\mathbf{z}_{ij}, r_{ijk}=1|\mathbf{y}_j; \boldsymbol{\theta}') = f(r_{ijk}=1|\mathbf{y}_j; \boldsymbol{\theta}')f(\mathbf{z}_{ij}|r_{ijk}=1, \mathbf{y}_j; \boldsymbol{\theta}')$ and $f(r_{ijk}=1, s_{jl}=1|\mathbf{y}_j; \boldsymbol{\theta}') = f(s_{jl}=1|\mathbf{y}_j; \boldsymbol{\theta}')f(r_{ijk}=1|s_{jl}=1, \mathbf{y}_j; \boldsymbol{\theta}')$. Since each term on the right-hand side of the conditional expectation depends on a different set of parameters, all the cross-derivatives are null and the maximization of the four terms in (14) can be carried out separately.

5.2. E-step

Calculation of the E-step involves computing three sets of posterior probabilities, namely $f(r_{ijk}=1|s_{jl}=1, \mathbf{y}_j)$, $f(s_{jl}=1|\mathbf{y}_j)$ and $f(r_{ijk}=1|\mathbf{y}_j)$, and the first and second moments of the conditional densities $f(\mathbf{z}_{ij}|r_{ijk}=1, \mathbf{y}_j)$ and $f(\mathbf{z}_{ij}|\mathbf{y}_j)$.

The first two probabilities are computed by exploiting the fact

$$f(r_{ijk}=1|s_{jl}=1, \mathbf{y}_j) = f(r_{ijk}=1|s_{jl}=1, \mathbf{y}_{ij}), \tag{15}$$

which is due to the assumption that subjects belonging to a group are independent of one another given the group class membership. Therefore,

$$\begin{aligned} f(r_{ijk}=1|s_{jl}=1, \mathbf{y}_{ij}) &= \frac{f(r_{ijk}=1|s_{jl}=1)f(\mathbf{y}_{ij}|r_{ijk}=1, s_{jl}=1)}{\sum_{k=1}^{K} f(r_{ijk}=1|s_{jl}=1)f(\mathbf{y}_{ij}|r_{ijk}=1, s_{jl}=1)} \\ &= \frac{\pi_{k|l}\phi^{(p)}(\mathbf{y}_{ij}; \boldsymbol{\Lambda}\boldsymbol{\mu}_k, \boldsymbol{\Lambda}\boldsymbol{\Sigma}_k\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi})}{\sum_{k=1}^{K}\pi_{k|l}\phi^{(p)}(\mathbf{y}_{ij}; \boldsymbol{\Lambda}\boldsymbol{\mu}_k, \boldsymbol{\Lambda}\boldsymbol{\Sigma}_k\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi})} \end{aligned} \tag{16}$$

and

$$\begin{aligned} f(s_{jl}=1|\mathbf{y}_j) &= \frac{f(s_{jl}=1)f(\mathbf{y}_j|s_{jl}=1)}{\sum_{l=1}^{L} f(s_{jl}=1)f(\mathbf{y}_j|s_{jl}=1)} \\ &= \frac{w_l\prod_{i=1}^{n_j} f(\mathbf{y}_{ij}|s_{jl}=1)}{\sum_{l=1}^{L} w_l\prod_{i=1}^{n_j} f(\mathbf{y}_{ij}|s_{jl}=1)}, \end{aligned}$$

where $f(\mathbf{y}_{ij}|s_{jl}=1)$ is the denominator of the fraction in Eq. (16).

Eq. (15) is exploited also in the derivation of $f(r_{ijk}=1|\mathbf{y}_j)$ as follows:

$$\begin{aligned} f(r_{ijk}=1|\mathbf{y}_j) &= \sum_{l=1}^{L} f(r_{ijk}=1|s_{jl}=1, \mathbf{y}_j)f(s_{jl}=1|\mathbf{y}_j) \\ &= \sum_{l=1}^{L} f(r_{ijk}=1|s_{jl}=1, \mathbf{y}_{ij})f(s_{jl}=1|\mathbf{y}_j). \end{aligned}$$

In order to evaluate the first and second moments of the conditional density $f(\mathbf{z}_{ij}|r_{ijk}=1,\mathbf{y}_j)$, it is worth noting that $f(\mathbf{z}_{ij}|r_{ijk}=1,\mathbf{y}_j)=f(\mathbf{z}_{ij}|r_{ijk}=1,\mathbf{y}_{ij})$ because according to (5) the factor score of a subject ($\mathbf{z}_{ij}$) is independent of the observed data of other subjects. Moreover,

$$f(\mathbf{z}_{ij}|r_{ijk}=1,\mathbf{y}_{ij})=\frac{f(\mathbf{y}_{ij}|\mathbf{z}_{ij})f(\mathbf{z}_{ij}|r_{ijk}=1)}{f(\mathbf{y}_{ij}|r_{ijk}=1)}, \tag{17}$$

where

$$f(\mathbf{y}_{ij}|\mathbf{z}_{ij})=\phi^{(p)}(\mathbf{\Lambda}\mathbf{z}_{ij},\mathbf{\Psi})$$
$$f(\mathbf{z}_{ij}|r_{ijk}=1)=\phi^{(q)}(\boldsymbol{\mu}_k,\mathbf{\Sigma}_k)$$
$$f(\mathbf{y}_{ij}|r_{ijk}=1)=\phi^{(p)}(\mathbf{\Lambda}\boldsymbol{\mu}_k,\mathbf{\Lambda}\mathbf{\Sigma}_k\mathbf{\Lambda}^\top+\mathbf{\Psi}).$$

After some simple algebra, it is possible to show that

$$f(\mathbf{z}_{ij}|r_{ijk}=1,\mathbf{y}_{ij})=\phi^{(q)}(\boldsymbol{\rho}_k(\mathbf{y}_{ij}),\boldsymbol{\xi}_k)$$

with

$$\boldsymbol{\xi}_k=\left(\mathbf{\Sigma}_k^{-1}+\mathbf{\Lambda}^\top\mathbf{\Psi}^{-1}\mathbf{\Lambda}\right)^{-1},$$
$$\boldsymbol{\rho}_k(\mathbf{y}_{ij})=\boldsymbol{\xi}_k\left(\mathbf{\Sigma}_k^{-1}\boldsymbol{\mu}_k+\mathbf{\Lambda}^\top\mathbf{\Psi}^{-1}\mathbf{y}_{ij}\right).$$

Finally, first and second moments of $f(\mathbf{z}_{ij}|\mathbf{y}_j)$ can be computed as weighted average of the corresponding moments of (17) with weights given by $f(\mathbf{r}_{ij}|\mathbf{y}_{ij})$.

5.3. M-step

In the M-step, the first derivatives with respect to $\boldsymbol{\theta}$ of the four terms in (14) have to be evaluated.

The optimal values for $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are obtained by computing the score function of

$$E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}'}\left[\sum_{j=1}^J\sum_{i=1}^{n_j}\log f(\mathbf{y}_{ij}|\mathbf{z}_{ij};\boldsymbol{\theta})\right]=\sum_{j=1}^J\sum_{i=1}^{n_j}\int f(\mathbf{z}_{ij}|\mathbf{y}_j;\boldsymbol{\theta}')\log f(\mathbf{y}_{ij}|\mathbf{z}_{ij};\boldsymbol{\theta})\mathrm{d}\mathbf{z}_{ij}$$

where $f(\mathbf{y}_{ij}|\mathbf{z}_{ij};\boldsymbol{\theta})=\phi^{(p)}(\mathbf{\Lambda}\mathbf{z}_{ij},\mathbf{\Psi})$. After some passages, we derive

$$E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}'}\left[\sum_{j=1}^J\sum_{i=1}^{n_j}\log f(\mathbf{y}_{ij}|\mathbf{z}_{ij};\boldsymbol{\theta})\right]\propto\sum_{j=1}^J\sum_{i=1}^{n_j}-\frac{1}{2}\log(\det\mathbf{\Psi})-\frac{1}{2}\mathrm{tr}\,\mathbf{\Psi}^{-1}(\mathbf{y}_{ij}\mathbf{y}_{ij}^\top-2\mathbf{y}_{ij}E[\mathbf{z}_{ij}^\top|\mathbf{y}_j]\mathbf{\Lambda}^\top+\mathbf{\Lambda}E[\mathbf{z}_{ij}\mathbf{z}_{ij}^\top|\mathbf{y}_j]\mathbf{\Lambda}^\top)$$

from which the estimates for the new parameters given the provisional ones are

$$\hat{\mathbf{\Lambda}}=\sum_{j=1}^J\sum_{i=1}^{n_j}\mathbf{y}_{ij}E[\mathbf{z}_{ij}^\top|\mathbf{y}_j]E[\mathbf{z}_{ij}\mathbf{z}_{ij}^\top|\mathbf{y}_j]^{-1},\qquad\hat{\mathbf{\Psi}}=\sum_{j=1}^J\sum_{i=1}^{n_j}\mathbf{y}_{ij}\mathbf{y}_{ij}^\top-\mathbf{y}_{ij}E[\mathbf{z}_{ij}^\top|\mathbf{y}_j]\mathbf{\Lambda}^\top.$$

Estimates for $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ can be obtained by maximizing the second term of (14):

$$E_{\mathbf{z},\mathbf{r}|\mathbf{y},\boldsymbol{\theta}'}\left[\sum_{j=1}^J\sum_{i=1}^{n_j}\log f(\mathbf{z}_{ij}|r_{ijk}=1;\boldsymbol{\theta})\right]=\sum_{j=1}^J\sum_{i=1}^{n_j}\sum_{k=1}^K\int f(\mathbf{z}_{ij},r_{ijk}=1|\mathbf{y}_j;\boldsymbol{\theta}')\log f(\mathbf{z}_{ij}|r_{ijk}=1;\boldsymbol{\theta})\mathrm{d}\mathbf{z}_{ij}$$

where $f(\mathbf{z}_{ij}|\mathbf{r}_{ij};\boldsymbol{\theta})=\phi^{(q)}(\boldsymbol{\mu}_k,\mathbf{\Sigma}_k)$, as previously observed. The estimates of the new parameters in terms of provisional ones are

$$\hat{\boldsymbol{\mu}}_k=\frac{\sum_{j=1}^J\sum_{i=1}^{n_j}f\left(r_{ijk}=1|\mathbf{y}_{ij};\boldsymbol{\theta}'\right)E[\mathbf{z}_{ij}|r_{ijk}=1,\mathbf{y}_{ij};\boldsymbol{\theta}']}{\sum_{j=1}^J\sum_{i=1}^{n_j}f\left(r_{ijk}=1|\mathbf{y}_{ij};\boldsymbol{\theta}'\right)},$$

$$\hat{\mathbf{\Sigma}}_k=\frac{\sum_{j=1}^J\sum_{i=1}^{n_j}f\left(r_{ijk}=1|\mathbf{y}_{ij};\boldsymbol{\theta}'\right)\left(E[\mathbf{z}_{ij}\mathbf{z}_{ij}^\top|r_{ijk}=1,\mathbf{y}_{ij};\boldsymbol{\theta}']-\boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top\right)}{\sum_{j=1}^J\sum_{i=1}^{n_j}f\left(r_{ijk}=1|\mathbf{y}_{ij};\boldsymbol{\theta}'\right)}.$$

Finally, the estimates for the mixing proportions $\boldsymbol{\pi}$ and $\mathbf{w}$ of the mixture can be computed by evaluating the score functions of

Table 1
Chironomus data. Description of the $p = 17$ variables.

X_1	Width of labial tooth C_2
X_2	Width of central labial teeth C_1 and C_2
X_3	Height of labial tooth C_1 from base of C_2
X_4	Height of labial tooth C_2
X_5	Height of labial tooth L_1
X_6	Height of central labial tooth
X_7	Width between apices of C_2 teeth
X_8	Number of pecten epipharyngeal teeth
X_9	Ventral head length
X_{10}	Length antennal segment 1
X_{11}	Length antennal segment 1 to ring organ
X_{12}	Length antennal segment 2
X_{13}	Length antennal segment 3
X_{14}	Length antennal segment 4
X_{15}	Width of antenna at ring organ
X_{16}	Mandibible length
X_{17}	Frontoclypeus width

$$E_{\mathbf{r},\mathbf{s}|\mathbf{y},\boldsymbol{\theta}'}\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{k=1}^{K}\log f(r_{ijk}=1|s_{jl}=1;\boldsymbol{\theta})\right]=\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{l=1}^{L}\sum_{k=1}^{K}f(r_{ijk}=1,s_{jl}=1|\mathbf{y}_j;\boldsymbol{\theta}')\log f(r_{ijk}=1|s_{jl}=1;\boldsymbol{\theta})$$

and

$$E_{\mathbf{s}|\mathbf{y},\boldsymbol{\theta}'}\left[\sum_{j=1}^{J}\sum_{l=1}^{L}\log f(s_{jl}=1;\boldsymbol{\theta})\right]=\sum_{j=1}^{J}\sum_{l=1}^{L}f(s_{jl}=1|\mathbf{y}_j;\boldsymbol{\theta}')\log f(s_{jl}=1;\boldsymbol{\theta})$$

under the constraints that they are positive and sum to one. The obtained estimates are

$$\hat{\pi}_{k|l}=\sum_{j=1}^{J}\sum_{i=1}^{n_j}f\left(r_{ijk}=1|s_{jl}=1,\mathbf{y}_{ij};\boldsymbol{\theta}'\right)f\left(s_{jl}=1|\mathbf{y}_j;\boldsymbol{\theta}'\right)$$

and

$$\hat{w}_l=\sum_{j=1}^{J}f\left(s_{jl}=1|\mathbf{y}_j;\boldsymbol{\theta}'\right)$$

where all the posterior densities have been derived in the E-step.

5.4. Initialization of the algorithm

A crucial point associated with the application of the EM algorithm is the choice of the starting values for the parameters of the model, since the algorithm is particularly sensitive to its initialization. In this procedure an ordinary factor analysis with q factors was used in order to derive the starting values for the factor loading matrix. The diagonal covariance matrix $\boldsymbol{\Psi}$ has been initialized by taking a given proportion of the variances of the observed variables (for instance, 0.2). The starting values for the component means and covariances were set by the method of moments, on the basis of an MCLUST classification with K clusters (see [7,8]) performed on the whole data set. Finally, a uniform random initialization has been used for the mixture weights.

6. Some applications

6.1. A toy example: Chironomus larvae data

The aim of this example is to illustrate the classification performance of the proposed model on two-level multivariate data with varying higher level clustering structures.

The data are taken from a study on some morphometric attributes of *Chironomus larvae* [4]. Three species of larvae named *nepeanensis*, *staegeri* and *tepperi* have been considered, with sample sizes equal to 53, 52 and 57, respectively. For each larva, $p = 17$ characteristics of the larval head capsula have been measured. A list of these morphometric attributes is given in Table 1.

Eight different artificial two-level multivariate data set were considered, assuming $L = 2$ higher level classes. These data sets were generated so that they share the same lower level structure, given by the multivariate data described above, and different structures in the higher level. The following factors were manipulated in order to generate the different higher

Table 2
Chironomus data. Average misclassification error rates of higher level units (with standard deviation in brackets).

	$\mathbf{w} = (0.50, 0.50)$		$\mathbf{w} = (0.33, 0.66)$	
	$n_j = 3$	$n_j = 9$	$n_j = 3$	$n_j = 9$
$\boldsymbol{\pi}_1 = (0.14, 0.44, 0.42)$ $\boldsymbol{\pi}_2 = (0.52, 0.20, 0.28)$	0.297 (0.059)	0.132 (0.067)	0.313 (0.039)	0.242 (0.075)
$\boldsymbol{\pi}_1 = (0.17, 0.59, 0.23)$ $\boldsymbol{\pi}_2 = (0.48, 0.05, 0.47)$	0.143 (0.053)	0.009 (0.040)	0.259 (0.036)	0.112 (0.019)

Table 3
Il sole 24 ore data. Description of the $p = 12$ indicators.

X_1	Per capita deposit in euros
X_2	Average monthly pension in euros
X_3	Consumer price index
X_4	Per capita gross domestic product index
X_5	House price in euros per sq.m in semi centre
X_6	Per capita expenditure on furniture and household appliances
X_7	Infrastructure equipment index (Tagliacarne Institute)
X_8	Road accidents per 10,000 inhabitants
X_9	Annual thermal excursion in Celsius degrees
X_{10}	Ecosystem indicator (Legambiente)
X_{11}	Ratio of closed legal actions to new and pending ones
X_{12}	High school drop-out rate

level structures: the number n_j of specimens per higher level unit; the sets $\boldsymbol{\pi}_1 = [\pi_{k|1}]_{k=1,\ldots,3}$ and $\boldsymbol{\pi}_2 = [\pi_{k|2}]_{k=1,\ldots,3}$ of the proportions of the three species in the two higher level classes; and finally, balanced and unbalanced higher level class proportions were considered.

More specifically, the following settings were considered for the three design factors listed above: $n_j = 3$ and $n_j = 9$; the situation $\{\boldsymbol{\pi}_1 = (0.14, 0.44, 0.42), \boldsymbol{\pi}_2 = (0.52, 0.20, 0.28)\}$ was compared to a situation where lower level class membership probabilities are more different across higher level classes, namely $\{\boldsymbol{\pi}_1 = (0.17, 0.59, 0.23), \boldsymbol{\pi}_2 = (0.48, 0.05, 0.47)\}$; finally, the setting $\mathbf{w} = \left(\frac{1}{2}, \frac{1}{2}\right)$ was compared with the unbalanced one $\mathbf{w} = \left(\frac{1}{3}, \frac{2}{3}\right)$.

A total of $8 = 2 \times 2 \times 2$ possible combinations of the three factors have been considered. For each of these settings, the proposed model has been estimated starting from 100 different initializations of the EM algorithm. The values L and K were set equal to 2 and 3, respectively, since the aim of the experiment was to investigate the classification performance of the model at the two levels of the hierarchy. The number of factors, q, was set equal to 3, as suggested by the Bayesian information criterion in a preliminary application of HFMM on the original data.

Table 2 reports the misclassification error rates of higher level units, averaged across the 100 different EM initializations, in each of the 8 considered settings. As it could be expected, classification performance improves when the number n_j of lower level units per higher level unit is increased. The same thing occurs when lower level class proportions are made more different across higher level classes: this is due to the fact that under these latter circumstances higher level classes are made more separable. In addition, classification tasks seem to be more challenging in the case of unbalanced classes.

The corresponding results about lower level unit classification have not been reported since they do not reveal any remarkable difference across the explored settings: the model steadily performed very well in classifying larvae, with average error rates ranging from 0.018 to 0.05.

6.2. A real data example

Every year, the Italian financial newspaper *Il Sole 24 Ore* analyzes the quality of life in 103 provinces of Italy through several indicators collected in different thematic areas (www.ilsole24ore.com). This data set consists of $p = 12$ measurements for the Italian provinces collected and published in 2008. The first six indicators are related to wealth and standard of living; the remaining ones deal with environment and services. A description of these indicators is provided in Table 3.

Since the $n = 103$ provinces are nested within the $J = 20$ Italian regions, this study represents a typical example of three-level data set, where responses are the first-level units, provinces are the second-level units and regions are the third-level units. Our aim is to classify both regions and provinces into some clusters characterizing different levels of quality of life.

Italy is an economic reality characterized by a deep income inequality between the dynamic, industrialized Centre-North and the less developed, agricultural-based South. Almost all less-developed regions tend to be located in the subtropical Mediterranean climates of the country whereas the high-industrialized regions typically stand in the cooler temperate ones. For these reasons, we considered $L = 2$ clusters of regions. The proposed model has been estimated on these data with K ranging between 1 and 6 and the number of factors, q varying between 1 and 5. Table 4 reports the main characteristics of the optimal model chosen according to some information criteria: the Bayesian Information Criterion (BIC, [16]), the Akaike Information Criterion (AIC, [1]) and the variant that uses the penalizing constant 3 instead of 2 (AIC3, [5]). Besides

Table 4
Il sole 24 ore data. The estimated model results.

Model	log L	BIC	AIC	AIC3	h
$L = 2, K = 5, q = 3$	−1141.85	2700.82	2463.69	2553.69	90

Table 5
Il sole 24 ore data. Province cluster means and estimated weights.

	w_l		$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
$l = 1$	0.70	$\pi_{k\|1}$	0.00	0.34	0.54	0.13	0.00
$l = 2$	0.30	$\pi_{k\|2}$	0.81	0.00	0.00	0.00	0.19
	X_1		5 983.39	9 358.86	12 099.46	22 088.98	7 867.13
	X_2		556.08	648.15	721.87	837.54	608.93
	X_3		2.21	1.91	1.68	1.69	2.05
	X_4		16 909.08	23 084.06	28 203.16	33 379.21	19 772.76
	X_5		1 832.48	2 244.78	2 629.28	3 493.32	1 875.07
	X_6		764.81	1 069.01	1 307.90	1 379.77	951.86
	X_7		77.95	85.99	97.96	165.37	63.19
	X_8		227.43	333.40	418.39	557.05	287.83
	X_9		17.34	18.62	20.06	19.49	16.49
	X_{10}		43.09	50.31	56.40	57.09	45.96
	X_{11}		39.18	47.48	55.20	58.06	39.92
	X_{12}		1.81	1.71	1.08	1.74	3.76

Table 6
Il Sole 24 ore data. Estimated weights in the mixture model for multilevel data obtained by Latent GOLD 4.0.

	w_l		$k = 1$	$k = 2$	$k = 3$
$l = 1$	0.55	$\pi_{k\|1}$	0.88	0.00	0.12
$l = 2$	0.45	$\pi_{k\|2}$	0.00	0.97	0.03

these model selection indices, Lukočienė and Vermunt [11] have suggested to use the number of third-level units, J, as penalization term in the BIC formula instead of the number of second-level units, n, when the aim is to decide the number of components at the third level. However in the present application, we limited the attention to the classical BIC criterion because we focused on to $L = 2$ clusters of regions, for the previously explained reasons.

The estimated model is characterized by $K = 5$ clusters of provinces and $q = 3$ factors synthesizing the $p = 12$ indicators, thus leading to $h = 90$ estimated parameters. The optimal log-likelihood is -1141.85 with a value for the BIC of 2700.82, a value for the AIC and AIC3 of 2463.69 and 2553.69, respectively.

In line with the economic and territorial differences mentioned above, the $L = 2$ clusters of the third-level units correspond to the regions of north and center of Italy ($l = 1$) and to those of the South ($l = 2$) including the two islands of Sicily and Sardinia. The obtained classification of regions is shown in the upper-left panel of Fig. 2.

In order to characterize the $K = 5$ clusters of provinces, the conditional means $\boldsymbol{\Lambda}\boldsymbol{\mu}_k$ (for $k = 1, \ldots, 5$) have been computed and reported in Table 5, together with the estimated weights of the mixture model.

Results in table show that provinces in the first group are characterized by the lowest average level of per-capita deposit, pension and gross domestic product index. Moreover, they show the lowest level of infrastructure equipment and services. These provinces are mainly collocated in the southern cluster of regions (see the bottom-left panel of Fig. 2). In other terms these are the poorest Italian provinces. On the contrary, the second group of provinces shows slightly higher values than the previous one for most indicators. These are provinces mainly located in the center of Italy, characterized by a medium standard of living and a reasonably good equipment of services and infrastructures. Cluster 3 consists of industrialized and rich provinces in the center and north of Italy. Cluster 4 includes only seven big and wealthy provinces, like Milan and Rome. The last group of provinces seems to be quite similar to the first one but exhibits a high value for the high-school drop-out average rate. It consists of some provinces, mostly located in Sardinia, which are particularly characterized by this scholar problem.

Fig. 1 shows the three-dimensional scatterplot of the factor scores of the 103 provinces, distinguished by cluster. As shown by the graph, the five clusters are quite separated in the estimated latent space.

For comparative purposes, we have estimated the multilevel finite mixture model proposed in Vermunt [22] under the local independence assumption. Different models with $K = 1, \ldots, 6$ have been estimated using `Latent GOLD 4.0` and the best model according to the Bayesian information criterion consisted of $K = 3$ clusters. Since the estimated model showed some high residuals, the local independence assumption has been partially relaxed by introducing eight correlation parameters between the most relevant paired residuals ($X_4 - X_6, X_1 - X_5, X_2 - X_8, X_9 - X_{12}, X_7 - X_8, X_5 - X_7, X_7 - X_{10}, X_1 - X_4$), for a total of $h = 101$ estimated parameters. The log-likelihood of the resulting model is equal to −1141.43 with a value of BIC equal to 2750.98, a value of AIC of 2484.87 and AIC3 of 2585.87.

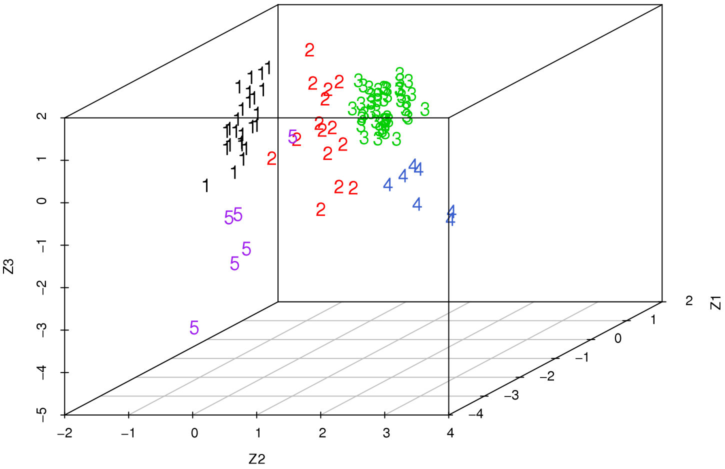

Fig. 1. *Il Sole 24 ore* data. The three-dimensional scatterplot of the factor scores of the 103 Italian provinces.

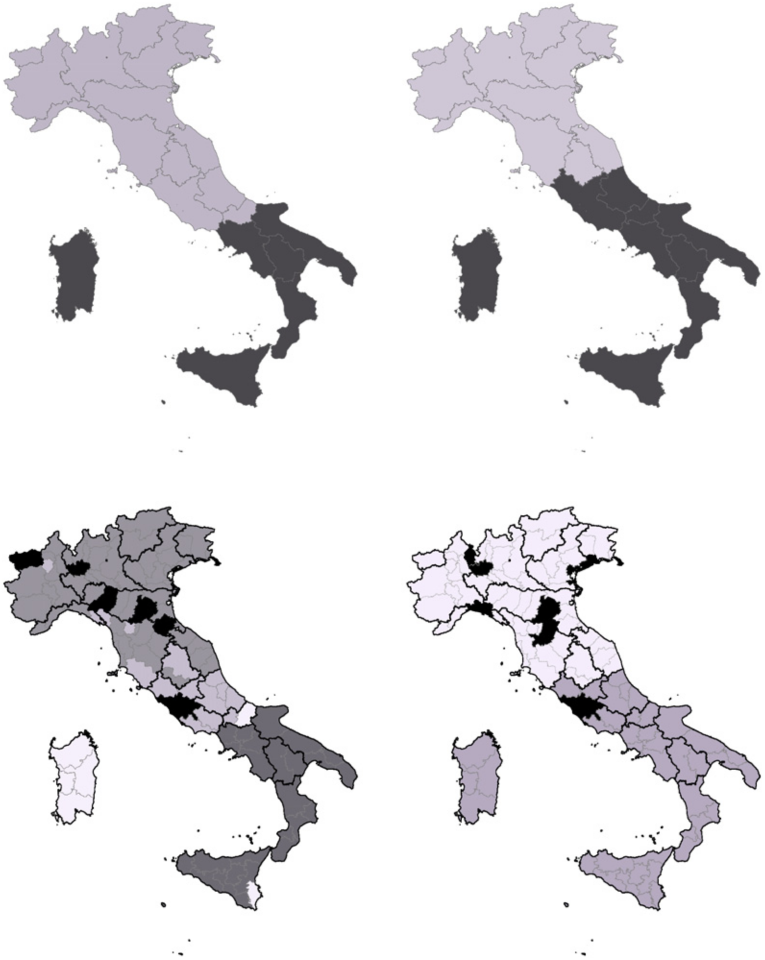

Fig. 2. *Il sole 24 ore* data. Region and province classification. Graph 1 (upper left) shows the region classification obtained with the proposed model. Graph 2 (upper right) contains the region classification according to the proposal by Vermunt. Graph 3 (bottom left) show the province classification into $K = 5$ clusters estimated with the proposed model and Graph 4 shows the province classification into $K = 3$ clusters obtained through Vermunt's approach.

Table 6 reports the estimated weights of this model. Regions are partitioned into $L = 2$ balanced clusters corresponding to the north ($l = 1$) and center-south ($l = 2$) of Italy, as shown in the upper-right panel of Fig. 2. Provinces in regions of the Center-South cluster into a whole group which corresponds to the poorest provinces ($\pi_{2|2} = 0.97$). Provinces in the north of Italy cluster into two groups, a numerous group of "middle" provinces ($\pi_{1|1} = 0.88$) and a small group of eight big and important provinces ($\pi_{1|3} = 0.12$) including Rome and Milan. This classification is shown in the bottom-right panel of Fig. 2.

7. Conclusions

In this paper, a model for clustering multilevel data has been introduced by combining tools from different backgrounds. Starting from the recent proposals on latent class analysis for multilevel data, we have defined and explored a dimensionally reduced solution. It is achieved by introducing a latent factor model in order to reduce the number of free parameters of the mixture model, without requiring the local independence assumption. In so doing the trade-off between model parsimony and flexibility is governed. In the illustrated real example, a meaningful clustering of both second- and third-level units has been obtained by estimating a model involving less parameters than those required by traditional multilevel mixture modeling.

From a computational point of view, the proposed model has been estimated using an EM-algorithm. The algorithm did not show any convergence problem when properly initialized. We have implemented it in R code and it is available from the authors upon request. In addition, we tried fitting the proposed model with the standard software for mixture models. As far as Mplus 5 [15] is concerned, the model can be framed in the context of two-level confirmatory factor analysis mixture modeling with continuous factor indicators; in our experiments some convergence problems have occurred when factors were allowed to be heteroscedastic. The GUI version of Latent Gold does not allow us to directly specify the model; however, the model could be specified and estimated using the syntax version of the software.

References

[1] H. Akaike, A new look at statistical model identification, IEEE Trans. Automat. Control 19 (1974) 716–723.

[2] S.S. Allua, Evaluation of single-and multilevel factor mixture model estimation, PhD thesis, University of Texas at Austin, 2007.

[3] T. Asparouhov, B. Muthen, Multilevel mixture models, in: G.R. Hancock, K.M. Samuelsen (Eds.), Advances in Latent Variable Mixture Models, Information Age Publishing Inc., Charlotte, NC, 2008, pp. 27–51.

[4] W.R. Atchley, J. Martin, A Morphometric analyisis of differential sexual dimorphism in larvae of Chironomus, Can. Entomol. 108 (1971) 819–827.

[5] H. Bozdogan, Determining the number of component clusters in the standard multivariate normal mixture model using model-selecton criteria, Technical Report No. UIC/DQM/A83-1, Quantitative Methods Department, University of Illinois at Chicago, 1983.

[6] N.M. Dempster, A.P. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1977) 1–38.

[7] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, J. Amer. Statist. Assoc. 97 (2002) 611–631.

[8] C. Fraley, A.E. Raftery, MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering, Technical Report No. 504, Department of Statistics, University of Washington, 2006.

[9] L.A. Goodman, Exploratory latent structure analysis using both identifiable and unidentifiable models, Biometrika 61 (1974) 215–231.

[10] G.H. Lubke, B. Muthén, Investigating population heterogeneity with factor mixture models, Psychol. Methods 10 (2005) 21–39.

[11] O. Lukočienė, J.K. Vermunt, Determining the number of components in mixture models for hierarchical data, in: A. Fink, et al. (Eds.), Advances in Data Analysis, Data Handling and Business Intelligence, Springer, Heidelberg, 2010, pp. 241–249.

[12] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, Oxford, 1976.

[13] G.J. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.

[14] A. Montanari, C. Viroli, Heteroscedastic factor mixture analysis, Stat. Model., 2010 (in press).

[15] L.K. Muthén, B.O. Muthén, Mplus Users Guide, fifth ed., Los Angeles, 2007.

[16] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.

[17] A. Skrondal, S. Rabe-Hesketh, Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models, Chapmal and Hall, New York, 2004.

[18] A.B. Snijders, R.J. Bosker, Multilevel Analysis, SAGE Publications, London, 1999.

[19] J.K. Vermunt, Multilevel latent class models, Sociol. Methodol. 33 (2003) 213–239.

[20] J.K. Vermunt, Mixed-effect logistic regression models for indirectly observed outcome variables, Multivariate Behav. Res. 40 (2005) 281–301.

[21] J.K. Vermunt, A hierarchical mixture model for clustering three-way data sets, Comput. Statist. Data Anal. 51 (2007) 5368–5376.

[22] J.K. Vermunt, Latent class and finite mixture models for multilevel data sets, Stat. Methods Med. Res. 17 (2008) 33–51.

[23] J.K. Vermunt, Mixture models for multilevel data sets, in: J. Hox, J.K. Roberts (Eds.), Handbook of Advanced Multilevel Analysis, Lawrence Erlbaum, 2010.

[24] J.K. Vermunt, J. Magidson, Hierarchical mixture models for nested data structures, in: C. Weihs, W. Gaul (Eds.), Classification: The Ubiquitous Challenge, Springer, Heidelberg, 2005, pp. 176–183.

[25] J.K. Vermunt, J. Magidson, LG-Syntax User's Guide: Manual for Latent GOLD 4.5, Syntax Module, Belmont, MA, Statistical Innovations Inc.