# Chapter 6
# Multilevel mixture factor models for the evaluation of educational programs' effectiveness

Roberta Varriale and Caterina Giusti

## 6.1 Introduction

Factor models aim at explaining the associations among observed random variables in terms of fewer unobserved random variables, called common *factors*. When data have a hierarchical structure, multilevel mixture factor models are a powerful and flexible tool useful to correctly take into account the correlation between first-level units due to the data structure, and to evaluate the presence of latent sub-populations of units with some typical profile at different levels of the analysis.

In the Chapter, we describe the specification of a multilevel mixture factor model with continuous latent variables at the lower level of the analysis and a discrete latent variable at the higher level, focusing on some technical and applied features of the analysis. The theory will be illustrated by means of an application on the job satisfaction of the graduates of the University of Florence. The main aim of the analysis is to describe and summarize some aspects of job satisfaction measured at the individual level and, at the same time, to cluster higher level units (degree courses) in classes with some typical characteristics, in order to analyse their effectiveness.

The Chapter is organized as follows. In Sect. 6.2 we introduce the multilevel mixture factor model, and in Sect. 6.3 we collocate it in the Generalized Latent Variable framework. The details of estimation procedures and of model selection for multilevel mixture factor models are described in Sects. 6.4 and 6.5. Finally, in Sect. 6.6 we present and comment the main results of the case study on the evaluation of University effectiveness.

Roberta Varriale
Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands;
Department of Statistics "G. Parenti", University of Florence, Florence, Italy,
e-mail: roberta.varriale@ds.unifi.it

Caterina Giusti
Department of Statistics and Mathematics Applied to Economics, University of Pisa, Via C. Ridolfi 10, 56124, Pisa, Italy, e-mail: caterina.giusti@ec.unipi.it

M. Bini et al. (eds.), *Statistical Methods for the Evaluation of Educational Services and Quality of Products*, Contributions to Statistics,
DOI 10.1007/978-3-7908-2385-1_6, 

## 6.2 The multilevel mixture factor model

Factor models aim at finding a set of continuous latent variables, called *factors*, that contains the same information of a given set of observed variables (Bartholomew and Knott, 1999). One basic assumption of factor models states that the observed variables are measured on a set of independent units. This assumption is inadequate when units are nested in clusters having a hierarchical structure, sharing common environments, experiences and interactions: in these cases multilevel techniques are necessary in order to correctly take into account the correlation between first-level units due to the data structure. In this Chapter, attention is limited to datasets with two hierarchical levels, since the extension to more than two levels is conceptually straightforward.

The basic idea of a factor model adapted to deal with multilevel data is that some model parameters – indicator intercepts or thresholds and residual variances, factor loadings, factor means and variances – are allowed to differ across the *observed* groups (higher level units). These differences can be modeled including group dummies in the model, as in the multigroup (or fixed-effects) approach, or can be modeled with a multilevel factor model with continuous latent variables at all levels of the analysis by assuming that the group coefficients are random-effects coming from a particular distribution whose parameters should be estimated (Searle et al., 1992; Vermunt, 2003).

In a confirmatory perspective, the multilevel mixture factor model is a useful model to take into account the hierarchical structure of the data and to compare the observed groups of units, by evaluating the existence of unobserved subpopulations (classes) of groups with similar features with respect to the factor model parameters and overcoming the production of over-detailed information of the multigroup factor model, which estimates as many group coefficients as the groups (Vermunt, 2003).

In one-level context, the term finite mixture (McLachlan and Peel, 2000) or latent class model (Lazarsfield and Henry, 1968; Goodman, 1974) is typically used for models including only a categorical latent variable, whereas the term factor mixture model is used for models including both continuous latent variables and a categorical latent variable (Lubke and Muthén, 2005). Both models are usually applied to classify individual units into $K$ latent classes with similar model parameters; in standard finite mixture models the clustering is based on the similarity of the observed item parameters (intercept or thresholds), in factor mixture models the clustering is based on the similarity of both the item parameters and/or the factor loadings. A discrete latent variable can also be used as a non parametric specification of a distribution of continuous latent variables (Aitkin, 1999; Vermunt and Magidson, 2005). Indeed, a finite mixture distribution results from the discretization of a continuous latent variable distribution into $K$ probability masses $\pi_k$ at mass points $z_k$; the nonparametric specification is so represented by a finite mixture model with the maximum number of identifiable latent classes.

Formally, a factor mixed model includes a categorical latent variable in the model with a multinomial distribution; besides the parameters of the factor model, also the parameters of the multinomial distribution have to be estimated.

In two-level context, finite mixture components, formally "represented" by a categorical latent variable, may be present at the lower or/and higher level. When there are mixture components at both levels of the analysis, the multilevel latent class model is obtained (Vermunt, 2003), otherwise we obtain the multilevel mixture factor model. In the Chapter, we only discuss two-level models characterized by continuous latent variables at the lower level and a categorical latent variable at the higher level. The main aims of this model are to analyse the underlying structure of the phenomenon at the lower level and, at the same time, classify higher level units in some latent classes with similar profiles.

Assume that there are $J$ groups with a different number of individual units $n_j$, whose total number is equal to $N = \sum_{i=1}^{J} n_j$. For each individual, $H$ items are observed. Conditional on the latent variables, the response model for the observed variables is a generalized linear model specified via a linear predictor, a link, and a distribution from the exponential family. Let $y_{hij}$ denote the observed response on indicator $h$ $(h = 1,\ldots,H)$ of individual $i$ $(i = 1,\ldots,n_j)$ within group $j$ $(j = 1,\ldots,J)$ and let $v_{hij}$ be the linear predictor of the response model. The conditional expectation of the response $y_{hij}$ given the latent variables at different levels is "linked" to the linear predictor $v_{hij}$ via a link function:

$$g(E(y_{hij}|\boldsymbol{\eta}_j)) = v_{hij} \tag{6.1}$$

where $\boldsymbol{\eta}_j = \left(\boldsymbol{\eta}_j^{(2)'},\ldots,\boldsymbol{\eta}_j^{(L)'}\right)'$ represents all latent variables, $\boldsymbol{\eta}_j^{(l)} = \left(\eta_{1j}^{(l)},\ldots,\eta_{M_l j}^{(l)}\right)'$ indicates all the latent variables varying at level $l$ and $M_l$ denotes the number of these latent variables. In particular, the latent variables varying at the individual and cluster level are denoted, respectively, with $\boldsymbol{\eta}_j^{(2)}$ and $\boldsymbol{\eta}_j^{(3)}$; indeed, since we are analysing models for datasets with one level of hierarchy, $l = 2,3$. Following the conventions, these models are called two-level models: the individual units $i$ are the level-1 units, and the group level units $j$ are the level-2 units. If the items are treated as level-1 units, the models become three-level models with individual units at level 2 and groups at level 3.

Different distributional forms are allowed for each indicator and the choice among different link functions naturally follows from the scale types of the observed variables. In particular, while in the traditional literature different terms are used depending on the nature of both latent and observed variables (Bartholomew and Knott, 1999), in the following we will use only the general term factor models. Recent developments in computational statistics extended the use of estimation methods traditionally used for models with only continuous indicators to the analysis of models with any kind of response variables.

As an example, with continuous responses an identity link and a normal distribution are usually assumed, so (we do not use the subscript $j$, for simplicity):

$$y_{hi} = v_{hi} + e_{hi}$$

with $f(e) \sim N(0,\sigma^2)$; therefore, the conditional density of $y_{hi}$ given the latent variables becomes:

$$f(y_{hi}|\boldsymbol{\eta}_j) = \sigma^{-1}\phi(v\sigma^{-1})$$

where $\phi$ represents the standard normal density. As another example, with ordinal responses several model specifications are possible. Let $s$, $s = 1,\ldots,S$ be the category of the ordinal response $y_{hi}$, the model for the cumulative probabilities is expressed by:

$$g[P(y_{hi} \leq s|\boldsymbol{\eta}_j)] = \alpha_s - v_{hi} \quad s = 1,\ldots,S-1 \tag{6.2}$$

where $\alpha_s$ with $\alpha_1 < \ldots < \alpha_{S-1}$ are the thresholds to be estimated. Typical choices of link function include the probit, logit and complementary log-log.

The two-level mixture factor model for continuous indicators and with one categorical latent variable at the highest level of analysis is:

$$y_{hij} = \mu_{hj} + \sum_{m=1}^{M_2} \lambda_{mh}^{(2)} \eta_{mij}^{(2)} + e_{hij}^{(2)} \tag{6.3}$$

$$\mu_{hj} = \sum_{k=1}^{K} \lambda_{kh}^{(3)} \eta_{kj}^{(3)} + e_{hj}^{(3)} \tag{6.4}$$

$$\eta_{mij}^{(2)} = \sum_{k=1}^{K} \beta_{km}^{(3)} \eta_{kj}^{(3)} + e_{mij}^{(2)} \tag{6.5}$$

where $\eta_{mij}^{(2)}$ denotes the $m$th common factor at individual level, $\lambda_{mh}^{(2)}$ represents the factor loading for factor $m$ and item $h$ and $\mu_{hj}$ is the item $h$ intercept for each group $j$. The two terms $e_{hij}^{(2)}$ and $e_{hj}^{(3)}$ represent the item-specific errors at lower and higher level. The variable $\eta_{kj}^{(3)}$ in Eqs. (6.4) and (6.5) is an indicator variable taking value 1 if unit $i$ belongs to latent class $k$ of the categorical latent variable $\boldsymbol{\eta}_j^{(3)}$ and 0 otherwise, and $\lambda_{kh}^{(3)}$ and $\beta_{km}^{(3)}$ represent the coefficients for each class $k$. The classes are mutually exclusive and, for the identification of the model, $\sum_{k=1}^{K} \lambda_{kh}^{(3)} = 0$ or $\lambda_{1h}^{(3)} = 0$ and $\sum_{k=1}^{K} \beta_{km}^{(3)} = 0$ or $\beta_{1m}^{(3)} = 0$. The term $e_{mij}^{(2)}$ represents a residual component of the relationship between $\eta_{mij}^{(2)}$ and $\boldsymbol{\eta}_j^{(3)}$.

The variable $\boldsymbol{\eta}_j^{(3)} = \left(\eta_{1j}^{(3)},\ldots,\eta_{Kj}^{(3)}\right)$ has a multinomial distribution, with:

$$\pi_k = P\left(\eta_j^{(3)} = k\right) = P\left(\eta_{kj}^{(3)} = 1\right) = \frac{\exp(\gamma_k)}{\sum_{t=1}^{K} \exp(\gamma_t)} \tag{6.6}$$

with

$$\sum_{k=1}^{K} \pi_k = 1. \tag{6.7}$$

The term $\gamma_k$ in Eq. (6.6) represents the intercept term of the linear predictor of the logit model for the expectation of the latent distribution ($\pi_k$); models with covariate effects on class membership can be defined by including covariate effects in this linear term.

The basic assumptions of multilevel mixture factor models are that each group belongs to no more than one latent class $k$, the individuals are independent inside each group conditional on the latent class $k$ at the higher level and the $H$ responses of individual $i$ are independent of each other given the continuous latent variables at the individual level and the group latent class membership, which is often referred to as the local independence assumption (Bartholomew and Knott, 1999).

The $\boldsymbol{\eta}_j^{(2)}$ are usually assumed to be normally independent and identically distributed with:

$$\boldsymbol{\eta}_j^{(2)} \sim MN(\mathbf{0}, \boldsymbol{\Psi}^{(2)})$$

where MN indicates the Multivariate Normal distribution and $\boldsymbol{\Psi}^{(2)}$ is the $M_2 \times M_2$ variance and covariance matrix with elements $\psi_{mm'}^{(2)}$.

It is also assumed that the item-specific error at both levels of the analysis, $e_{hij}^{(2)}$ and $e_{hj}^{(3)}$, are mutually independent and identically normally distributed.

In the most general case of multilevel mixture factor analysis, both $\lambda_{kh}^{(3)}$ and $\beta_{km}^{(3)}$ in Eqs. (6.4) and (6.5) may differ across higher-level mixture components in order to capture the differences between individuals due to the hierarchical data structure. Two special cases of the model are obtained by constraining these terms. In the first case, $\lambda_{kh}^{(3)} = 0$, therefore the outcome variables are not directly affected by the higher level latent class and the item intercepts do not vary across group-level classes; in the second case, $\beta_{km}^{(3)} = 0$, so the individual-level latent variable does not vary across group-level classes. The first case is typically used when the researchers' interest is in classifying the higher level units and comparing the obtained groups with a confirmative approach, "pushing" up the information collected at the individual-level to the group-level through the different "steps" of the model. The second case is typically used with an exploratory approach, aiming at analysing separately the lower and higher structure of the data.

The model is represented in Fig. 6.1. Following the conventions, circles represent latent variables and rectangles represent observed variables. The latent categorical variable are indicated with a filled circle. The arrows connecting latent and/or observed variables do not necessarily represent linear relations and possible correlations among latent variables or among items are represented with dotted lines. The nested frames represent the nested levels, for example, variables located within the outer frame labeled $j$ vary between clusters and have a $j$ subscript (Skrondal and Rabe-Hesketh, 2004).

## 6.3 The Generalized Latent Variable framework

The two-level mixture factor model described so far belongs to the Generalized Latent Variable framework introduced by Muthén (2008) and Vermunt (2007). This

general framework integrates specific methodologies for latent variable modelling, such as multilevel, longitudinal and structural equation models as well as item response models, factor models and so on, in a global theoretical context and allows to define models with any combination of categorical and continuous latent variables at each level of the hierarchy.

The generalized latent variable model is formally described by two elements: the response model for the observed variables conditional on the latent variables and the model for the latent variables. Using the index $j$ to denote an independent observation corresponding to the highest level of the hierarchy, the two-level mixture factor model is expressed by:

$$g[E(\boldsymbol{y}_j|\boldsymbol{\eta}_j)] = \mathbf{Z}_j\boldsymbol{\beta} + \boldsymbol{\Lambda}^{(1)}\boldsymbol{\eta}_j \tag{6.8}$$

$$h[E(\boldsymbol{\eta}_j^{(2)})] = \mathbf{X}_j\boldsymbol{\gamma} + \boldsymbol{\Lambda}^{(2)}\boldsymbol{\eta}_j^{(3)} \tag{6.9}$$

where $\mathbf{y}_j$ denotes the response vector with elements $y_{hij}$ representing the response to indicator $h$ of each individual $i$ belonging to group $j$.

In the two-level framework, the vector $\boldsymbol{\eta}_j = \left(\boldsymbol{\eta}_j^{(2)'}, \boldsymbol{\eta}_j^{(3)'}\right)'$ in Eq. (6.8) denotes the latent variables varying at the $i$-th and $j$-th level of the analysis affecting directly the observed responses. The vector $\boldsymbol{\eta}_j^{(3)}$ in Eq. (6.9) denotes the latent variables at the $j$-th level affecting the latent variables at the $i$-th level.

The two matrices $\mathbf{Z}_j$ and $\mathbf{X}_j$ with the corresponding coefficient vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ denote the fixed part of the model affecting, respectively, the observed items and the latent structure at level 2. Different links and distributions can be specified for different responses. The matrices $\boldsymbol{\Lambda}$, which elements do not vary depending on $j$, represent the factor loading matrix of the generalized latent variable model. In particular, $\boldsymbol{\Lambda}^{(1)}$ indicates the factor loading matrix relating the latent variables directly

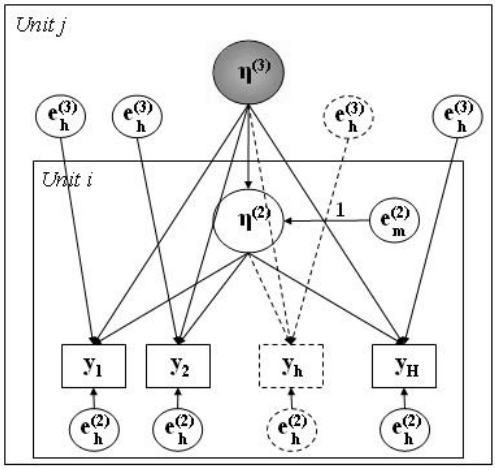


**Fig. 6.1** Two-level mixture factor model.

to the outcomes and $\boldsymbol{\Lambda}^{(2)}$ indicates the factor loading matrix relating level 3 to level 2 latent variables.

Table 6.1 schematically represents different specifications of the two-level mixture factor model. In particular, a model with continuous latent variables at both levels of the analysis is called two-level factor model, while models with both continuous and categorical latent variables are called two-level mixture factor models. Which model should be selected depends on the aims of the specific research and on the substantive reason to believe in the nature, continuous or categorical, of the latent variables.

**Table 6.1** Matrix of potential two-level models with underlying latent variables

| | Higher level | latent variables | |
|---|---|---|---|
| Lower level latent variables | Continuous | Categorical | Combination |
| Continuous | A1 | A2 | A3 |
| Categorical | B1 | B2 | B3 |
| Combination | C1 | C2 | C3 |

Model A1, in which both the lower and higher level latent variables are continuous, is represented by the multilevel factor model, as described by Goldstein and McDonald (1988) and Longford and Muthén (1992); its extension to ordinal indicators is given by Grilli and Rampichini (2007a). Model A1 contains also three-level regression models with continuous random effects. Model B2, in which both the lower and higher level latent variables are categorical, is the multilevel latent class model. Vermunt (2003) proposes a model where lower level units are clustered based on their observed responses and higher level units are clustered based on the likelihood of their members to be in one of the unit level clusters. Vermunt (2003) also proposes a multilevel latent class model with continuous random effects at the group level (B1). Palardy and Vermunt (2009) used specification A3 to define a multilevel extension of the mixture growth model (Muthén, 2004), where two-level units are classified into homogeneous groups based on properties of their mean growth trajectories.

This brief and incomplete review of the literature shows how modelling using a combination of continuous and categorical latent variables provides an extremely general and flexible framework of analysis. Furthermore, different traditions such as growth modelling, multilevel modelling, latent class analysis are brought together using the unifying theme of latent variables.

## 6.4 Likelihood, estimation and posterior analysis

Recent developments in computational statistics have enhanced the feasibility of a maximum likelihood analysis in the context of multilevel mixture factor models.

In this section we briefly present the formulation of the likelihood that has to be maximized.

In two-level models, the total marginal likelihood is:

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{J} L_j(\boldsymbol{\theta}) = \prod_{j=1}^{J} f^{(j)}(\mathbf{y}_{(j)}|\boldsymbol{\theta}) \tag{6.10}$$

where $L_j$ indicates the likelihood of group $j$, the groups are assumed to be independent and $\boldsymbol{\theta}$ represents the complete set of unknown parameters to be estimated. The complete likelihood can be derived recursively. In a model with $\boldsymbol{\eta}^{(2)}$ and $\boldsymbol{\eta}^{(3)}$ being, respectively, continuous latent variables at the first and second level of the analysis (not using the subscript $j$ for the latent variables hereafter, for simplicity), the likelihood for each group $j$ is given by:

$$L_j(\boldsymbol{\theta}) = \int_{\boldsymbol{\eta}^{(3)}} \prod_{i=1}^{n_j} L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(3)}) f(\boldsymbol{\eta}^{(3)}) d\boldsymbol{\eta}^{(3)} \tag{6.11}$$

where the $n_j$ level-1 units within level-2 units are assumed to be independent given the random coefficients $\boldsymbol{\eta}^{(3)}$. For each first-level unit, controlling for the effect of the latent variables at the highest level, the likelihood is expressed by:

$$L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(3)}) = \int_{\boldsymbol{\eta}^{(2)}} L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)}) f(\boldsymbol{\eta}^{(2)}|\boldsymbol{\eta}^{(3)}) d\boldsymbol{\eta}^{(2)}. \tag{6.12}$$

Finally, considering the local independence assumption, the observed indicators are assumed to be independent given the latent variables, so:

$$L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)}) = \prod_{h=1}^{H} f(y_{hij}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)}) \tag{6.13}$$

where $f(y_{hij}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)})$ indicates the distribution of the response variables.

When the latent variables are categorical, the multiple integrals are replaced by multiple sums. In a model with $\boldsymbol{\eta}^{(3)}$ and $\boldsymbol{\eta}^{(2)}$ being, respectively, a categorical and continuous latent variables, the likelihood is expressed by:

$$\begin{aligned}
L_j(\boldsymbol{\theta}) &= \sum_{k=1}^{K} P(\boldsymbol{\eta}^{(3)} = k) \prod_{i=1}^{n_j} L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(3)} = k) \\
L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(3)} = k) &= \int_{\boldsymbol{\eta}^{(2)}} L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)} = k) f(\boldsymbol{\eta}^{(2)}|\boldsymbol{\eta}^{(3)} = k) d\boldsymbol{\eta}^{(2)} \\
L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)} = k) &= \prod_{h=1}^{H} f(y_{hij}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)} = k).
\end{aligned}$$

Maximum Likelihood estimation involves finding the estimates for $\boldsymbol{\theta}$ that maximize the marginal likelihood function (or the log-likelihood function).

In maximizing the likelihood, two separated problems must be considered: solving the integrals involved in the likelihood and maximizing the likelihood function. With respect to the first aspect, while a closed form expression for these integrals is available when all responses and latent variables are continuous and normally distributed, in the other cases there are several approaches to approximating the integrals, as Laplace approximation, numerical integration using quadrature or adaptive quadrature, Monte Carlo integration (Skrondal and Rabe-Hesketh, 2004). With respect to the second aspect, several methods were proposed for maximizing the likelihood, the most common being the Expectation-Maximization (EM) algorithm and Newton-Raphson or Fisher scoring algorithms. Of course, each integration method may be combined with some maximization methods.

The main aim of a researcher using factor models is in what can be known about the latent variables after the indicators have been observed (Bartholomew and Knott, 1999). At each level of the analysis, this information is represented by the conditional density:

$$f(\boldsymbol{\eta}|\mathbf{y}) = f(\boldsymbol{\eta})f(\mathbf{y}|\boldsymbol{\eta})/f(\mathbf{y}). \tag{6.14}$$

From the point of view of social behavioral scientists, this means locating units on the dimensions of the latent space (*factor scores*), or classifying units in different *classes* representing some typical profile. Obviously, units with the same response pattern will be assigned the same factor score or class.

Some scoring methods are the ones based on the empirical Bayesian posterior distribution and the maximum likelihood method (Skrondal and Rabe-Hesketh, 2004). Usually, the firsts are the most used; indeed, while the maximum likelihood approach produces scores that are conditionally unbiased, it is not consistent with the modelling assumptions since it requires that the latent variables are considered fixed parameters and does not yield predictions for clusters with insufficient information. For this reason, we only present the two Bayesian posterior distribution methods.

With the empirical Bayesian approach, according to Bayes' theorem, the conditional posterior distribution of the latent variables given the observed variables is expressed by:

$$f(\boldsymbol{\eta}|\mathbf{y},\hat{\boldsymbol{\theta}}) = \frac{f(\mathbf{y},\boldsymbol{\eta}|\hat{\boldsymbol{\theta}})}{f(\mathbf{y}|\hat{\boldsymbol{\theta}})} = \frac{f(\mathbf{y}|\boldsymbol{\eta},\hat{\boldsymbol{\theta}})f(\boldsymbol{\eta}|\hat{\boldsymbol{\theta}})}{\int_{\boldsymbol{\eta}} f(y|\boldsymbol{\eta},\hat{\boldsymbol{\theta}})f(\boldsymbol{\eta}|\hat{\boldsymbol{\theta}})} \tag{6.15}$$

where $\hat{\boldsymbol{\theta}}$ represent the estimated parameters, $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ is the distribution of the observed variables and $f(\mathbf{y},\boldsymbol{\eta}|\hat{\boldsymbol{\theta}})$ is the joint distribution of the observed and latent variables. This approach uses the term "Bayesian" since both the latent and observed variables are treated as random variables. Actually, the full Bayesian approach would assume a prior distribution for $\boldsymbol{\theta}$ in addition to the distribution for $\boldsymbol{\eta}$ and the $\boldsymbol{\theta}$ in Eq. (6.15) would be treated as fixed constants.

The computation of the posterior distribution is strictly related to the specification of the prior distribution of the latent variables. Usually, the posterior distribution cannot be expressed in closed form and heavy numerical integration is required. In

factor models with continuous random variables, it follows from standard results on conditional multivariate normal densities that the posterior density is multivariate normal; for other response types, the posterior density tends to multinormality as the number of units in the clusters increases (Skrondal and Rabe-Hesketh, 2004).

After estimating the empirical Bayesian posterior distribution, two approaches can be used to estimate the factor scores (or latent class) associated to each unit: the prediction using empirical Bayes (also called a posteriori) and the prediction using empirical Bayes modal (also known as *modal* a posteriori).

The empirical Bayes prediction is the most widely used method for scoring. The predictors are represented by the mean of the posterior empirical Bayesian latent variables distribution in Eq. (6.15), so:

$$\boldsymbol{\eta}^{\mathrm{EB}} = E(\boldsymbol{\eta}|\mathbf{y}, \hat{\boldsymbol{\theta}}). \tag{6.16}$$

With continuous normal latent variables, the empirical Bayes predictor is the best linear unbiased predictor BLUP (Skrondal and Rabe-Hesketh, 2004).

The prediction using empirical Bayes modal uses the posterior mode instead of the posterior mean for the prediction of the factor scores:

$$\boldsymbol{\eta}^{\mathrm{EBM}} = \overset{\max \arg}{\boldsymbol{\eta}} (\boldsymbol{\eta}|\mathbf{y}, \hat{\boldsymbol{\theta}}). \tag{6.17}$$

This method does not require numerical integration, so when the posterior density is approximately multivariate normal it is often used as an approximation of the empirical Bayes solutions. In particular, this method represents the standard classification method in latent class modelling since it minimize the expected misclassification rate (Skrondal and Rabe-Hesketh, 2004). Obviously, in standard factor models the predictors obtained with the empirical Bayes and empirical Bayes modal coincide.

## 6.5 Model selection

A number of overall and individual statistical measures of fit has been proposed in order to evaluate a specified model on the basis of empirical data. In the following, some tests based on the likelihood theory and some information criteria useful to choose between different multilevel and multilevel mixture factor models are briefly introduced.

One method to compare nested models is based on the likelihood ratio test (Agresti, 2002). However, standard asymptotic results for the test do not hold if the null hypothesis is on the boundary of the parameter space since regularity conditions would be violated; well-known examples are testing the null hypothesis relating to random effects (Self and Liang, 1987) and testing the hypothesis on the variability of the latent factors. In these cases, a rule of thumb is to divide by two the asymptotic $p$-value of the Chi-squared likelihood ratio test statistic distribution (Skrondal and

Rabe-Hesketh, 2004). Also in the mixture models framework the likelihood ratio statistic cannot be used to compare two nested models, one with $k_0$ classes and one with $k_1$ classes ($k_0 < k_1$). Indeed, under the null hypothesis of $k_0$ groups, some of the parameters of the model with $k_1$ classes lie on the boundary of the parameter space so that regularity conditions for likelihood ratio statistic to be asymptotically Chi-squared are not fulfilled. In particular, the correct null distribution of the likelihood ratio statistic is unknown (Everitt, 1988) but a lot of conjectures and simulations have been published on this topic (McLachlan and Peel, 2000).

Another approach for comparing models is based on the computation of some indexes representing a penalized form of the likelihood: as the likelihood increases with the addiction of some parameters, it is penalized by the subtraction of a term related to the number of parameters. These information criteria are generally expressed in terms of:

$$-2\log L(\boldsymbol{\theta}) + C \tag{6.18}$$

where the first term measures the lack of fit of the model and $C$ is the penalty term that measures the complexity of the model. The intent is therefore to choose a model to minimize this criterion.

Relating to the problem of choosing between models with different number of latent classes, a variety of textbooks and articles suggest the use of the Bayesian Information Criterion (BIC) (Schwarz, 1978) as a good indicator (Nylund et al., 2007). The BIC is expressed by:

$$BIC = -2\log L + p \times \log(N) \tag{6.19}$$

where $\log L$ is the loglikelihood value, $p$ is the number of parameters and $N$ is the number of observations for the fitted model. In two-level models the number of observations can refer to both within and between level; this distinction can make a substantial difference when determining the number of classes of a multilevel mixture model. To our knowledge, while there is a wide variety of literature available on the performance of model selection statistics for determining the number of mixture components in one-level mixture models, there are no works in the two-level context, except that of Lukočiené and Vermunt (2004). In their paper, the authors show the results of a simulation study on multilevel latent class analysis with a fixed number of classes at the lower level, aiming at individuating the best index for determining the number of mixture components at the higher level.

## 6.6 Case study

In this section a multilevel mixture factor model is used in order to evaluate the university external effectiveness of the degree courses of the University of Florence. As suggested by Chiandotto (2004), students' perception of the quality of the services provided by an institution can be evaluated both at the time of the degree (internal

effectiveness) and some date later (external effectiveness). In particular, we evaluate the University performance from the users' subjective point of view, as perceived three years after the degree.

Different proposals on the use of multilevel methodologies to analyse both the external and internal effectiveness of the university system can be found, as some examples, in Giusti and Varriale (2008); Chiandotto et al. (in press); Chiandotto and Varriale (2006); Chiandotto and Giusti (2006). In the present application the use of multilevel mixture factor models, with a combination of continuous and categorical latent variables at different levels of the analysis, allows to fulfill two objectives, corresponding to the levels of the analysis. The "first level objective" is to understand the latent constructs underlying the phenomenon of job satisfaction using the information available at the individual level, that is the satisfaction expressed by graduated students that are employed three years after the degree. At the same time this individual information can be used to fulfill a "second level objective", to classify the study programs attended by the graduates into a small number of classes representing some typical profiles, that is to identify those programs with similar characteristics with respect to job satisfaction.

The job satisfaction is a complex process naturally considered as a latent construct not directly observable but measured by some indicators. Data come from the AlmaLaurea survey "Employment opportunities, 2005" (Almalaurea, 2006) and they concern graduates of the University of Florence. Data have a hierarchical structure, with graduates nested in different degree courses; in particular, it is interesting to investigate the effect of this level of aggregation on job satisfaction.

We consider the graduates with the old Italian university system during the summer session of the solar year 2002 who are employed at the moment of the interview, 3 years after the degree. We focus on the analysis of job satisfaction three years after the degree since it is reasonable that after that time all graduates find the job they have studied for and they are usually no more involved in specialization and training courses, except for the graduates in medicine. Obviously, as a confirmation of the results obtained with the present work, it would be interesting to repeat the same analyses when data referring to the graduates' occupational status five years after their degree will be available. For reasons of representativeness, we only consider those degree courses with at least eight employed graduates. The 1,025 graduates we include in the analysis represent almost 60% of the graduates at the University of Florence in the summer session of 2002; the total and percentage numbers of graduates in each degree course are in Table 6.2.

The questionnaire used for the Almalaurea survey "Employment opportunities" is very comprehensive, since it deals with many aspects related to the current job or the search for a job. The questionnaire section on the satisfaction with the actual job consists in 14 items. Through a correlation analysis and other preliminary considerations, we selected five of these items, measuring the satisfaction with: earnings, career opportunities, coherence with the University studies, professionalism and cultural interests. All these items are expressed on an ordinal scale with 10 categories; the items are considered as continuous variables because of the number of the categories. The average evaluation for each of the 5 items is in Table 6.3.

**Table 6.2** Number of graduates employed three years after the degree, by degree course. Students graduated (old system degree) at the University of Florence, summer session, year 2002

| Degree course | Number of employed graduates | Percentage |
|---|---|---|
| Architecture | 216 | 21.07 |
| Chemistry | 9 | 0.88 |
| Business economics | 26 | 2.54 |
| Economics | 67 | 6.54 |
| Philosophy | 16 | 1.56 |
| Law | 106 | 10.34 |
| Civil engineering | 31 | 3.02 |
| Electronic engineering | 29 | 2.83 |
| Mechanical engineering | 23 | 2.24 |
| Literature | 78 | 7.61 |
| Foreign lang. and literature | 48 | 4.68 |
| Mathematics | 11 | 1.07 |
| Medicine | 17 | 1.66 |
| Psychology | 51 | 4.98 |
| Biology | 11 | 1.07 |
| Political sciences | 131 | 12.78 |
| History | 8 | 0.78 |
| Informatics engineering | 10 | 0.98 |
| Environmental engineering | 21 | 2.05 |
| Educational sciences | 102 | 9.95 |
| Forest and environ. sciences | 14 | 1.37 |
| | 1,025 | 100 |

As we can see, there are some differences between the degree courses in the mean evaluations expressed by the graduates. For example, the graduates in philosophy and history express the lowest mean evaluations for the aspects coherence and cultural interests; moreover, they give low scores to the other three aspects. At the opposite, the graduates in architecture and law are the most overall satisfied. For the graduates in medicine we observe a really high evaluation for coherence, professionalism and cultural interests, as expected, but lower mean values for career and earnings, probably because these graduates are still involved in some specialization courses. There are also some differences between similar degree courses, like the ones in engineering; for example, the interviewed who graduated in electronic engineering seem to be less satisfied with their careers with respect to their colleagues. The differences in graduates' satisfaction between degree courses show an important influence of the hierarchical data structure on job satisfaction.

Due to the results of the preliminary analyses on the correlation structure between the items and to the latent (non observable) nature of the job satisfaction, we proceeded with an exploratory (EFA) and a confirmatory one-level factor analysis (CFA). As illustrated in Sect. 6.5, likelihood ratio tests have been used to compare models with different factor loadings, while BICs have been used to compare models with different number of latent factors. In particular, with EFA we compared models with 2 and 3 latent factors measured, at the same time, by all the indicators.

**Table 6.3** Mean evaluations with the selected items, by degree course. Students graduated (old system degree) at the University of Florence, summer session, year 2002

| Degree course | Coherence | Professionalism | Cultural interests | Earnings | Career |
|---|---|---|---|---|---|
| Architecture | 7.39 | 7.65 | 7.38 | 6.81 | 7 |
| Chemistry | 6.56 | 7.44 | 6.78 | 6.56 | 6.56 |
| Business economics | 7.85 | 7.85 | 6.85 | 7.19 | 6.92 |
| Economics | 6.91 | 7.34 | 6.46 | 6.85 | 6.78 |
| Philosophy | 4.81 | 6.63 | 5.06 | 5.63 | 5.73 |
| Law | 7.21 | 7.73 | 7.45 | 7.03 | 7.24 |
| Civil engineering | 7.74 | 7.68 | 7.29 | 6.55 | 6.41 |
| Electronic engineering | 6.55 | 7.14 | 6.83 | 6.28 | 5.97 |
| Mechanical engineering | 7.35 | 7.61 | 7.65 | 6.96 | 6.65 |
| Literature | 5.73 | 7.33 | 6.67 | 5.68 | 5.51 |
| Foreign lang. and literature | 5.71 | 7.15 | 6.4 | 6.08 | 6 |
| Mathematics | 5.36 | 7 | 6.82 | 6.64 | 5.64 |
| Medicine | 9.41 | 8.06 | 8.71 | 6.88 | 6.12 |
| Psychology | 6.2 | 7.12 | 6.75 | 5.59 | 5.82 |
| Biology | 7.82 | 8.55 | 7.18 | 5 | 5.18 |
| Political sciences | 5.53 | 7.16 | 6.57 | 6.29 | 6.43 |
| History | 3.25 | 7.13 | 5.75 | 6 | 5.75 |
| Informatics engineering | 7.4 | 7.2 | 7.1 | 6.3 | 6 |
| Environmental engineering | 7.76 | 8 | 7.38 | 6.9 | 6.45 |
| Educational sciences | 7.26 | 7.56 | 7.49 | 5.84 | 5.99 |
| Forest and environ. sciences | 6.43 | 7.21 | 7.21 | 5.79 | 5.93 |
| | 6.76 | 7.47 | 7.03 | 6.42 | 6.44 |

Subsequently, we run a CFA following what suggested by the correlation structure of the items and constraining to zero the loadings that resulted to be close to zero with EFA. The results of these analyses suggest the presence of two factors: one factor related to the *Cultural* features of the job, measured by career, professionalism, coherence and cultural interest, and one factor related to the *Status* of the job, measured by earnings, career and professionalism.

In order to take into account the two-level data hierarchy and to classify the degree courses in some latent classes with different profiles, we applied a two-level mixture factor model. The final model is:

$$y_{hi} = \mu_h + \sum_{m=1}^{M_2} \lambda_{mh}^{(2)} \eta_{mij}^{(2)} + e_{hij}^{(2)} \tag{6.20}$$

$$\eta_{mij}^{(2)} = \sum_{k=1}^{K} \beta_{km}^{(3)} \eta_{kj}^{(3)} + e_{mij}^{(2)} \tag{6.21}$$

At the program level, $\lambda_{kh}^{(3)} = 0$, therefore it is assumed that the degree courses differ only in the mean level of latent factors at the individual level $\left(\eta_{mij}^{(2)}\right)$ and the outcome variables are not directly affected by the higher level latent variable. In

other words, $\beta_{km}^{(3)}$ represents the mean of the $m$-th factor at individual level for the degree courses belonging to the $k$-th latent class.

In the model, the items coherence and earnings are the reference items (factor loading equal to 1), respectively, for the factors *Cultural* and *Status*. At the second level of the analysis, in Eq. (6.21) $\beta_{1m}^{(3)}$ are constrained to 0 for each $m$, $m = 1, 2$, in order to ensure the identification.

The Bayesian Information Criterion index calculated with $N$ equal to the number of groups is used to choose between models with different number of classes at group level. Table 6.4 shows BIC values for models composed of 1–4 classes.

**Table 6.4** Two-level mixture factor model: loglikelihood and fit indexes. Students graduated (old system degree) at the University of Florence, summer session, year 2002

| N classes | N param. | Log-likelihood | BIC ($N$ obs.) | BIC ($N$ groups) |
|---|---|---|---|---|
| 1 | 18 | −9775.29 | 19675.37 | 19605.38 |
| 2 | 21 | −9750.12 | 19645.82 | 19564.17 |
| 3 | 24 | −9737.45 | 19641.27 | 19547.97 |
| 4 | 27 | −9733.09 | 19653.36 | 19548.38 |

The final two-level mixture factor model is represented in Fig. 6.2.

At the individual level, the factor structure is very similar to that found with the one-level factor analysis. Again, we acknowledge the presence of two highly correlated latent factors (Table 6.6). The first factor (*Status*) is related to the satisfaction with earnings, career and professionalism; the second factor (*Cultural*) is related to career, professionalism and to the satisfaction with cultural interests and coherence of the job with the previous studies.

Factor loadings are shown in Table 6.5. All the loadings have the same sign. As is always the case, the latent dimension underlying the global satisfaction at the program level has an arbitrary scale, which means that factor scores must be interpreted relatively to each other. The most important aspects relating to factor *Status*

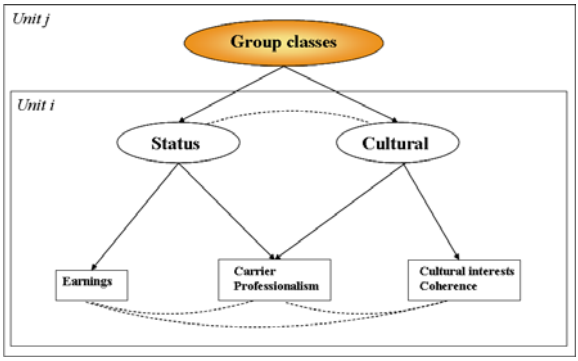


**Fig. 6.2** Two-level mixture factor model. Students graduated (old system degree) at the University of Florence, summer session, year 2002.

are earnings and career, while this is the case for coherence and cultural interests with the factor *Cultural*. In other words, for each degree course, the graduates' satisfaction with the job *Status* is measured mostly by their opinion on earnings and career and the graduates' satisfaction with the job *Cultural* is measured mostly by their opinion on coherence and cultural interests. Thus, the multilevel mixture factor model gives some insides on the dimensions influencing graduates' job satisfaction at the individual level.

**Table 6.5** Factor loadings. Students graduated (old system degree) at the University of Florence, summer session, year 2002

| | Status | Cultural |
|---|---|---|
| Earnings | 1 | |
| Career | 0.98 | 0.13 |
| Professionalism | 0.16 | 0.57 |
| Coherence | | 1 |
| Cultural interests | | 0.82 |

**Table 6.6** Variances, covariance and *correlation* of the factors. Students graduated (old system degree) at the University of Florence, summer session, year 2002

| | Status | Cultural |
|---|---|---|
| Status | 3.23 | *0.39* |
| Cultural | 1.32 | 3.63 |

As already underlined, besides these results referring to the first level of analysis, the model expressed by (6.20) and (6.21) allows also to interpret the effect of the degree courses on graduates' job satisfaction.

At the second level of analysis, the model classifies the courses in three classes. The sizes of the three classes are different: a degree course has a probability equal to 0.45 to be in the first class, of 0.36 to be in the second one and of 0.19 in the third one (Table 6.7, last row). Due to the constraints, the class-specific effects must be interpreted in terms of deviations from the "reference class" where the effects are equal to 0; in this analysis, the reference class is the first (Fig. 6.3). The three classes differ in the mean value of the two latent factors: the second class has a higher mean level of satisfaction both for *Status* and *Cultural* and the third class has a slightly lower mean value for the factor Status, while the satisfaction with *Cultural* is the highest between the three.

Using the empirical Bayes modal prediction, the degree courses can be assigned to the three classes (Table 6.7, column 2), so that we can better interpret the previous results. The main part of the courses, 11 out of 21, are attributed to the reference class. For some of these courses, in particular for chemistry, informatics

**Table 6.7** Two-level mixture factor model: study programs classification based on the empirical Bayesian posterior distribution. Students graduated (old system degree) at the University of Florence, summer session, year 2002

| Degree course | Class (modal) | Prob. Class 1 | Prob. Class 2 | Prob. Class 3 |
|---|---|---|---|---|
| Architecture | 2 | 0 | 1 | 0 |
| Chemistry | 1 | 0.52 | 0.43 | 0.05 |
| Business economics | 2 | 0.01 | 0.99 | 0 |
| Economics | 2 | 0.14 | 0.86 | 0 |
| Philosophy | 1 | 1 | 0 | 0 |
| Law | 2 | 0 | 1 | 0 |
| Civil engineering | 2 | 0.01 | 0.68 | 0.3 |
| Electronic engineering | 1 | 0.92 | 0.07 | 0.01 |
| Mechanical engineering | 2 | 0.02 | 0.95 | 0.03 |
| Literature | 1 | 1 | 0 | 0 |
| Foreign lang. and literature | 1 | 1 | 0 | 0 |
| Mathematics | 1 | 0.89 | 0.1 | 0.01 |
| Medicine | 3 | 0 | 0.04 | 0.96 |
| Psychology | 1 | 1 | 0 | 0 |
| Biology | 3 | 0.01 | 0.01 | 0.98 |
| Political sciences | 1 | 1 | 0 | 0 |
| History | 1 | 0.97 | 0.02 | 0 |
| Informatics engineering | 1 | 0.4 | 0.36 | 0.25 |
| Environmental engineering | 2 | 0.01 | 0.87 | 0.11 |
| Educational sciences | 3 | 0 | 0 | 1 |
| Forest and environ. sciences | 1 | 0.61 | 0.15 | 0.24 |
| Mean values | | 0.45 | 0.36 | 0.19 |

engineering and forest and environmental sciences, the posterior probabilities of belonging to a specific latent class at the group level are spread in the three classes (Table 6.7, columns 3 to 5). A more in-depth analysis could be useful in order to analyse the peculiarities of these courses. The degree courses belonging to the second class, the "best" for the satisfaction with both latent factors, are architecture, business economics, economics, law, civil engineering, mechanical engineering and

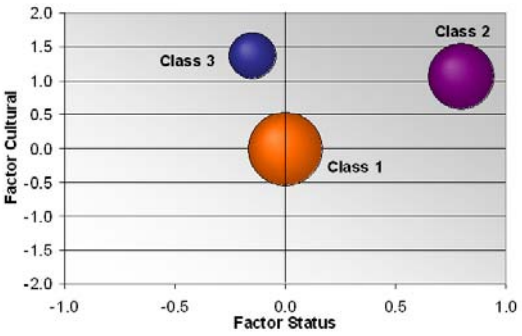


**Fig. 6.3** Latent classes features (latent factors *Cultural* and *Status*. Students graduated (old system degree) at the University of Florence, summer session year 2002.

environmental engineering. Graduates in these courses developed the skills and the possibility to choose a job which guarantees a high level of satisfaction with the different aspects we considered. Graduates in the courses belonging to the third class, namely medicine, biology and educational sciences, are instead more likely to have a job with a high correspondence to their cultural interests and previous studies, while the position or status of their jobs is maybe expected to increase in the future. In particular, graduates in medicine are probably still involved in specialization and training courses, while the other graduates can also be occupied in some occasional and temporary positions because they are encountering some difficulties to find the job they studied for.