Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates

Bartolucci, Francesco and Farcomeni, Alessio and Pennoni, Fulvia

13 April 2012

Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates

Francesco Bartolucci[*], Alessio Farcomeni[†], Fulvia Pennoni[‡]

April 13, 2012

Abstract

We provide a comprehensive overview of latent Markov (LM) models for the analysis of longitudinal data. The main assumption behind these models is that the response variables are conditionally independent given a latent process which follows a first-order Markov chain. We first illustrate the more general version of the LM model which includes individual covariates. We then illustrate several constrained versions of the general LM model, which make the model more parsimonious and allow us to consider and test hypotheses of interest. These constraints may be put on the conditional distribution of the response variables given the latent process (measurement model) or on the distribution of the latent process (latent model). For the general version of the model we also illustrate in detail maximum likelihood estimation through the Expectation-Maximization algorithm, which may be efficiently implemented by recursions known in the hidden Markov literature. We discuss about the model identifiability and we outline methods for obtaining standard errors for the parameter estimates. We also illustrate methods for selecting the number of states and for path prediction. Finally, we illustrate Bayesian estimation method. Models and related inference are illustrated by the description of relevant socio-economic applications available in the literature.

Keywords: EM algorithm, Bayesian framework, Forward-Backward recursions, Hidden Markov models, Measurement errors, Panel data, Unobserved heterogeneity

[*] Department of Economics, Finance and Statistics, University of Perugia, Via A. Pascoli, 20, 06123 Perugia, Italy, email: bart@stat.unipg.it

[†] Department of Public Health and Infectious Diseases, Sapienza - University of Rome, Piazzale Aldo Moro, 5, 00185 Roma, Italy, email: alessio.farcomeni@uniroma1.it

[‡] Department of Statistics, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy, email: fulvia.pennoni@unimib.it

1 Introduction

In many applications involving longitudinal data, the interest is often focused on the evolution of a latent characteristic of a group of individuals over time, which is measured by one or more occasion-specific response variables. This characteristic may correspond, for instance, to the quality-of-life of subjects suffering from a certain disease, which is indirectly assessed on the basis of responses to a set of suitably designed items that are repeatedly administered during a certain period of time.

Many models are proposed in the statistical literature for the analysis of longitudinal data; the choice mainly depends on the context of application. For a review see, among others, Fitzmaurice et al. (2009). In this literature, latent Markov (LM) models, on which the present paper is focused, have a special role. These models assume the existence of a latent process which affects the distribution of the response variable. The main assumption behind this approach is that the response variables are conditionally independent given this latent process, which follows a Markov chain with a finite number of states. The basic idea related to this assumption, which is referred to as *local independence*, is that the latent process fully explains the observable behavior of a subject together with possibly available covariates. Therefore, in studying LM models, it is important to distinguish between two components: the *measurement model*, i.e. the conditional distribution of the response variables given the latent process, and the *latent model*, i.e. the distribution of the latent process.

From many points of view, longitudinal data are similar to time-series data. The main difference is that time series are usually made of many repeated measures referred to a single unit, whereas only few repeated measures are typically available in a longitudinal dataset, but for many statistical units. However, inferential approaches arising within the time series literature cannot be directly extended to models for longitudinal data. In the context of time-series data analysis, for example, the asymptotic properties of an estimator are studied assuming that the number of repeated measures grows to infinity. On the contrary, in the context of longitudinal data analysis, asymptotic properties are studied assuming that the sample size tends to infinity, while the number of occasions of observation is held fixed. In the statistical and econometric literatures, one of the most interesting model for the analysis of time-series data is the Hidden Markov (HM) model (Baum and Petrie, 1966). The model is based on the same assumptions and estimation methods of the LM model, but the structure of the data they aim to analyze is different. However, the terminology is not univocal and the name HM model is sometimes adopted even for models applied to the analysis of longitudinal data.

Many relevant results in the HM literature have been developed in the 60's and 70's; one of the most relevant paper is due to Baum et al. (1970). Then, many advances have been developed in connection with engineering, informatics, and bioinformatics applications; consider, for instance, the papers by Levinson et al. (1983) and Ghahramani and Jordan (1997). For general reviews see the monographs of MacDonald and Zucchini (1997), Koski (2001), and Zucchini and MacDonald (2009), whereas for reviews and recent advances about the estimation methods see Bickel et al. (1998), Cappé et al. (2005), and Andersson and Rydén (2009).

In the following, we refer to the LM model based on a first-order Markov chain, non-homogeneous transition probabilities, and covariates as the *general* LM model. This model is presented in the case of multivariate data when we observe more response variables at each occasion. For this LM model we discuss in detail maximum likelihood estimation through the Expectation-Maximization (EM) algorithm (Baum et al., 1970; Dempster et al., 1977). We also briefly outline the Bayesian estimation method even though we acknowledge that other estimation methods are available (Archer and Titterington, 2009; Cappé et al., 1989; Künsch, 2005; Turner, 2008). For the implementation of the EM algorithm, we illustrate suitable recursions which allow us to strongly reduce the computational effort.

The paper also focuses on some constrained versions of the general LM model. Constraints have the aim of making the model more parsimonious, easier to interpret, and in correspondence with certain hypotheses that may be interesting to test. These constraints may be posed on the measurement model, that is on the conditional distribution of the response variables given the latent process, or on the latent model, that is on the distribution of the latent process. Regarding the former, we discuss in detail parametrizations which makes the latent states interpretable in terms of ability or propensity levels. Regarding the latter, we outline several simplifications of the transition matrix, mostly based on constraints of equality between certain elements of this matrix and/or on the constraint that some elements are equal to 0.

One of the main problems is how to test for the above restrictions. For this aim, we make use of the likelihood ratio (LR) statistic. It is important to note that, when constraints concern the transition matrix, the null asymptotic distribution has not necessarily an asymptotic chi-squared distribution, but a distribution of chi-bar-squared type (Bartolucci, 2006). We also illustrate an extension of the model to relax local independence. It is made in two ways: including among the covariates the lagged response variable, and allowing the response variables at the same time occasion to be dependent even conditionally on the corresponding latent variable. We discuss model identifiability and we review methods for obtaining standard errors for the model parameters. We illustrate the main approaches for selecting the number of latent states and we also discuss the problem of path prediction through the Viterbi algorithm (Viterbi, 1967; Juang and Rabiner, 1991).

The paper is organized as follows. In Section 2 we illustrate the main cases of application of the LM model and we briefly outline the model history. In Section 3 we outline the general LM model and constrained versions of the model based on parsimonious and interpretable parameterizations. In Section 4 we show how to obtain the manifest distribution of the response variables and the posterior distribution of the latent states by the Baum-Welch recursions (Baum et al., 1970; Welch, 2003). Then, we discuss maximum likelihood estimation for the general LM model based on the EM algorithm in Section 5, where we also deal with model identifiability and standard errors for the parameter estimates. Section 6 illustrates methods for the selection of the number of states and path prediction. Section 7 briefly outlines Bayesian estimation as an alternative to maximum likelihood estimation. Section 8 illustrates different types of LM model through various examples involving longitudinal categorical data, summarizing the results from other papers. The paper ends with a section where we draw main conclusions and discuss further developments of the present framework.

2 Model motivations and historical review

The use of the LM models finds justification in different types of analysis that one may be interested to perform. We illustrate in the following three main cases where it is sensible to apply LM models.

1. *Accounting for measurement errors.* In such cases the adopted LM Model is seen as an extension of a Markov chain model Anderson (1951, 1954), which represents a fundamental model for stochastic processes. As an example, consider the case in which the response variables correspond to different items and indicators reflecting the quality-of-life of an elderly subject. The items may concern the activity of daily living, whereas the indicators derive from certain clinical measures. In the present framework, the dependence structure between these variables is simplified by introducing one or few latent variables. The different configurations of latent variables are interpreted as different levels of the quality-of-life, an individual characteristic which is only indirectly observable through the response variables. This interpretation is enforced by the assumption of local independence.

2. *Accounting for unobserved heterogeneity between subjects.* The latent variable may have the role of accounting for the unobserved heterogeneity between subjects. This aspect is in

connection with the inclusion of the observed covariates in the measurement model which do not fully explain the heterogeneity between the responses provided by different individuals. The advantage with respect to a standard random effect or latent class model with covariates is that we admit that the effect of unobservable covariates has its own dynamics. As an example, consider the case in which we observe, at repeated occasions, a binary response variable indicating if an individual has a job position. Then, through an LM model we can take into account the effect of unobservable factors, such as motivations or intelligence, which affect this variable.

3. *Finding latent subpopulations of the population of interest.* The latent states are identified as different subpopulations, with units in the same subpopulation having a common distribution for the response variables. In this context, an LM model may be seen as an extension of the latent class (LC) model (Lazarsfeld, 1950; Lazarsfeld and Henry, 1968) in which subjects are allowed to move between the latent classes during the period of observation. If available, covariates are included in the latent model and then may affect the initial and transition probabilities of the Markov chain. As an example, consider the case of the analysis of criminal data in which for each age band (period of time) we know which types of crime are committed by subjects in a certain cohort. Then, we can first classify subjects in different clusters and then study the transition between different clusters. Each cluster corresponds to a latent state, with subjects in the same state having the same criminal behavior.

In the following we outline a brief historical review of the LM approach. This approach dates back to Wiggins' Ph.D. thesis, Wiggins (1955), who introduced a version of this model based on a homogeneous Markov chain, a single outcome at each occasion, and did not account for individual covariates. Wiggins (1955) formulated the model so that a manifest transition is a mixture of a true change and a spurious change due to measurement errors in the observed states. See also Wiggins (1973) for a deep illustration of this model and some simple generalizations.

The most natural extension of the basic LM model of Wiggins is for the inclusion of individual covariates. The formulation proposed by Vermunt et al. (1999) is based on the parameterization of initial and transition probabilities of the latent Markov process through a series of multinomial logit regression models depending on time constant and time-varying covariates. More recently, Bartolucci et al. (2007) extended this approach to the case of more than one response variable. Bartolucci and Pennoni (2007) also allowed transition probabilities to depend on lagged response variables. The formulation proposed by Bartolucci and Farcomeni (2009) includes the covariates in the measurement part of the model in the presence of multivariate responses. They propose to reparametrize the conditional distribution of the response variables given the latent process through a multivariate logistic transformation (McCullagh and Nelder, 1989; Glonek and McCullagh, 1995).

The basic LM approach has been generalized into the mixed LM approach by van de Pol and Langeheine (1990), who also transposed the general approach of Markov chains into the framework of mixture distribution models (see also Rost (2002)). This extended approach has also been considered by Altman (2007). In the latter, both fixed and continuous random effect are introduced in the measurement or in the latent part of the model. An interesting review concerning the mixed hidden Markov model in which random effects are assumed to have a discrete distribution may be found in Maruotti (2011). More recently, a method based on fixed effects to represent the factors common to all units in the same cluster was proposed by Bartolucci et al. (2009). A formulation based on random parameters having a discrete distribution was instead proposed by Bartolucci et al. (2011) in which the context of application is on the evaluation of the student math achievement. A related approach is the *latent transition analysis* (Bye and Schechter, 1986; Langeheine, 1988; Collins and Wugalter, 1992; Kaplan, 2008) in which the parameters of the LM models are allowed to vary in different latent subpopulations.

It is important to also mention alternative, but related, approaches such as the *latent growth approach* and the *mixture Markov approach*. The latent growth approach is based on a parameterization of the response variables on individual-specific parameters and covariates associated to the time occasions. The individual-specific parameters have a Gaussian distribution with a mean which may also depend on the observable covariates. Models formulated on the basis of this approach have different names, such as *latent growth models* or *latent curve models*. For a review see Muthén (2004), Nagin (1999), and Vermunt (2010); see also Bollen and Curran (2006). The mixture Markov approach is based on separate MC models for latent subpopulations. Each subpopulation has specific initial and transition probabilities of the Markov chain and the probability of belonging to a certain subpopulation typically depends on individual covariates. This approach was developed by Dias and Vermunt (2007) for an application in market segmentation. A particular case is the mover-stayer model of Goodman (1961).

3 THE General latent Markov model framework

In the following, we illustrate the general LM model framework and some constrained versions of the model.

3.1 Model formulation

Let T denote the number of occasions of observation and suppose that, at each time occasion, we observe r response variables denoted by $Y_j^{(t)}$, with $j = 1, \ldots, r$ and $t = 1, \ldots, T$. These response variables are collected, for each time occasion, in the random vector $\boldsymbol{Y}^{(t)}$, $t = 1, \ldots, T$; we also denote by $\tilde{\boldsymbol{Y}}$ the overall vector of the rT response variables. When available, we also denote by $\boldsymbol{X}^{(t)}$ a vector of covariates corresponding to $\boldsymbol{Y}^{(t)}$ and by $\tilde{\boldsymbol{X}}$ we denote the vector of all the individual covariates which is obtained by stacking the vectors $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)}$.

In order to model the distribution of $\tilde{\boldsymbol{Y}}$ given $\tilde{\boldsymbol{X}}$, the LM approach assumes the existence of a latent process $\boldsymbol{U} = (U^{(1)}, \ldots, U^{(T)})$, which is assumed to follow a first-order Markov chain with state space $\{1, \ldots, k\}$, where k is the number of latent states. Under the assumption local independence, the random vectors $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T)}$ are conditionally independent given the latent process. This assumption leads to a strong simplification of the model, but it can relaxed as we illustrate in the following. Moreover, it is assumed that the distribution of each response vector $\boldsymbol{Y}^{(t)}$ only depend on the corresponding latent variable $U^{(t)}$ and the possible available covariates in $\boldsymbol{X}^{(t)}$.

In order to clarify the above assumptions and provide further developments, we introduce a general notation according to which, for a random variable A, by $f_A(a)$ we denote the probability mass (or density) function for the distribution of A and, given two random variables A and B, by $f_{B|A}(b|a)$ we denote that for the conditional distribution of B given A. Then, with reference to LM framework introduced above, we let

$$\begin{aligned} f^{(t)}(\boldsymbol{y}|u,\boldsymbol{x}) &= f_{Y^{(t)}|U^{(t)},\boldsymbol{X}^{(t)}}(\boldsymbol{y}|u,\boldsymbol{x}), \quad t = 1, \ldots, T,\ u = 1, \ldots, k, \\ p(u|\boldsymbol{x}) &= f_{U^{(1)}|\boldsymbol{X}^{(1)}}(u|\boldsymbol{x}), \quad u = 1, \ldots, k, \\ p^{(t)}(u|\bar{u},\boldsymbol{x}) &= f_{U^{(t)}|U^{(t-1)},\boldsymbol{X}^{(t)}}(u|\bar{u},\boldsymbol{x}), \quad t = 2, \ldots, T,\ \bar{u}, u = 1, \ldots, k. \end{aligned}$$

In the above expressions, by u we indicate the current latent state, by $\bar{u}$ the previous latent state, by $\boldsymbol{x}$ a realization of $\boldsymbol{X}^{(t)}$, and by $\boldsymbol{y}$ a realization of $\boldsymbol{Y}^{(t)}$. It has to be clear since now that, when we use an LM model for a specific application, we formulate suitable assumptions on latent model and the measurement model. In practice, formulating assumptions on measurement model amounts to express $f^{(t)}(\boldsymbol{y}|u,\boldsymbol{x})$ as a function depending on suitable parameters to be estimated. Accordingly, formulating assumptions on the latent model amounts to suitably parametrize the initial probability function $p(u|\boldsymbol{x})$ and the transition probability function $p^{(t)}(u|\bar{u},\boldsymbol{x})$. How to express these parametrizations is clarified in the following.

3.1.1 Formulation of the measurement model.

In order to parametrize the probability (or density) function $f^{(t)}(\boldsymbol{y}|u,\boldsymbol{x})$, it is necessary to disentangle the univariate case, in which only one response variable is observed at each time occasion, and then $\boldsymbol{Y}^{(t)} = Y^{(t)}$ and we use the notation $f^{(t)}(y|u,\boldsymbol{x})$, from the multivariate case in which we observe more response variables at each occasion.

In the univariate case, we have first to formulate a specific assumption on the distribution of $Y^{(t)}$ given $U^{(t)}$ and $\boldsymbol{X}^{(t)}$, which may belong to an arbitrary family depending on a set of parameters $\boldsymbol{\eta}_{u\boldsymbol{x}}^{(t)}$. Then, we assume that

$$\boldsymbol{\eta}_{u\boldsymbol{x}}^{(t)} = \boldsymbol{W}_{u\boldsymbol{x}}^{(t)}\boldsymbol{\beta}, \quad u = 1,\dots,k, \tag{1}$$

where $\boldsymbol{W}_{u\boldsymbol{x}}^{(t)}$ is a design matrix depending on u and $\boldsymbol{x}$ and $\boldsymbol{\beta}$ is a reduced vector of parameters. In practice, $\boldsymbol{\eta}_{u\boldsymbol{x}}^{(t)}$ is obtained by a suitable link function of the mean of this distribution; then it reduces to a scalar which is denoted by $\eta_{u\boldsymbol{x}}^{(t)}$. We are referring to the link functions used within Generalized Linear Model (GLM) literature (McCullagh and Nelder, 1989). However, when $Y^{(t)}$ is a categorical variable with more than two categories, whose number is denoted by c, more complex link functions have to be used, which also take into account the ordinal nature of the response variables. Following Colombi and Forcina (2001), these link functions may be expressed as

$$\boldsymbol{\eta}_{u\boldsymbol{x}}^{(t)} = \boldsymbol{C}\log[\boldsymbol{M}\boldsymbol{f}^{(t)}(u,\boldsymbol{x})], \tag{2}$$

where $\boldsymbol{C}$ is a suitable matrix of constraints, $\boldsymbol{M}$ is a marginalization matrix with elements 0 and 1, and $\boldsymbol{f}^{(t)}(u,\boldsymbol{x})$ is a c-dimensional column vector with elements $f^{(t)}(y|u,\boldsymbol{x})$ for all possible values of y. In this case we denote by $\eta_{y|u\boldsymbol{x}}^{(t)}$, $y = 1,\dots,c-1$, each element of $\boldsymbol{\eta}_{u\boldsymbol{x}}^{(t)}$.

The two following examples clarify the possible formulations.

Example 1 *In the case of continuous response variables, we can simply assume that, given $U^{(t)}$ and $\boldsymbol{X}^{(t)}$, $Y^{(t)}$ has normal distribution with mean $\eta_{u\boldsymbol{x}}^{(t)}$. This amounts to use an identity link function in the GLM terminology. The resulting model then assumes that*

$$Y^{(t)} = \beta_{1u} + \boldsymbol{x}'\boldsymbol{\beta}_2 + \varepsilon_t, \quad t = 1,\dots,T,$$

where $\varepsilon_t \sim N(0,\sigma^2)$, where σ^2 is a variance parameter to be estimated. This model is then an extension of the linear regression model, with time-varying discrete random effects. □

Example 2 *In the case of binary response variables, it is natural to assume that, given $U^{(t)}$ and $\boldsymbol{X}^{(t)}$, $Y^{(t)}$ has a Bernoulli distribution with a certain "success" probability, the logit of which cor-*

responds to $\eta^{(t)}_{u\boldsymbol{x}}$*. The resulting model then assumes that*

$$\log \frac{f^{(t)}(1|u,\boldsymbol{x})}{f^{(t)}(0|u,\boldsymbol{x})} = \beta_{1u} + \boldsymbol{x}'\boldsymbol{\beta}_2, \quad t = 1,\dots,T.$$

In this case, we have an extension of the standard logit model with random effects. □

Example 3 *Suppose that* $Y^{(t)}$ *is an ordinal variable with* c *categories, which are labelled from 1 to* c*. Then, the conditional distribution of this variable, given* $U^{(t)}$ *and* $\boldsymbol{X}^{(t)}$*, is assumed to be of multivariate Bernoulli type. This distribution may be parametrized by the so-called global logits, which are computed as*

$$\eta_{y|u\boldsymbol{x}} = \log \frac{f(y+1|u,\boldsymbol{x}) + \cdots + f(c|u,\boldsymbol{x})}{f(1|u,\boldsymbol{x}) + \cdots + f(y|u,\boldsymbol{x})}, \quad y = 1,\dots,c-1. \tag{3}$$

Once these logits are collected in the $c-1$ *dimensional vector* $\boldsymbol{\eta}_{u\boldsymbol{x}}$*, we can formulate an assumption of type*

$$\eta_{y|u\boldsymbol{x}} = \beta_{1y} + \beta_{2u} + \boldsymbol{x}'\boldsymbol{\beta}_4, \quad y = 1,\dots,c-1,\ u = 1,\dots,k,$$

where β_{1y} *are cutpoints,* β_{2u} *are intercepts specific of the corresponding latent state,* $\boldsymbol{\beta}_3$ *is a vector of parameters for the covariates. In this way, we are generalizing the Proportional Odds Model of Glonek and McCullagh (1995), by including time-varying discrete random effects.*

It is interesting to note that the link function may be expressed as in (2). With $c = 4$*, for instance, we obtain global logits by letting*

$$\boldsymbol{C} = \begin{pmatrix} -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 \end{pmatrix}, \quad \boldsymbol{M} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

□

In the case of multivariate responses, it is common to formulate an extended version of the assumption of local independence, according to which the single response variables $Y^{(t)}_j$ in every vector $\boldsymbol{Y}^{(t)}$ are conditionally independent given $U^{(t)}$ and $\boldsymbol{X}^{(t)}$. Adopting an obvious notation, we have that

$$f^{(t)}(\boldsymbol{y}|u,\boldsymbol{x}) = \prod_j f^{(t)}_j(y_j|u,\boldsymbol{x}). \tag{4}$$

Each univariate distribution $f_j^{(t)}(y|u,\boldsymbol{x})$ may be formulated as above depending on a specific parameters $\boldsymbol{\eta}_{j|u\boldsymbol{x}}^{(t)}$. Then, we assume that

$$\boldsymbol{\eta}_{j|u\boldsymbol{x}}^{(t)} = \boldsymbol{W}_{j|u\boldsymbol{x}}^{(t)}\boldsymbol{\beta}, \quad j = 1,\dots,r,\ u = 1,\dots,k, \tag{5}$$

where $\boldsymbol{W}_{j|u\boldsymbol{x}}^{(t)}$ is a suitable design matrix.

In the case of multivariate data, it may be also reasonable to assume that the response variables in $\boldsymbol{Y}^{(t)}$ are not conditionally independent given $U^{(t)}$ and $\boldsymbol{X}^{(t)}$, allowing then for a form of *contemporary dependence*. In this case, we adopt a single link function for all the multivariate distribution $f^{(t)}(\boldsymbol{y}|u,\boldsymbol{x})$ depending on a single vector of parameters $\boldsymbol{\eta}_{u\boldsymbol{x}}$ which is parametrized as in (1). Such an approach, has been proposed for categorical response variable by Bartolucci and Farcomeni (2009). However, the above approach is difficult to adopt outside from the case of categorical responses and, even for this reason, it is in general assumed that assumption (4) holds.

It is worth noting that in certain applications we want to make the conditional distribution of $\boldsymbol{Y}^{(t)}$ given $U^{(t)}$ independent of the covariates $\boldsymbol{X}^{(t)}$ or these covariates are simply not available. In these cases, every function $f^{(t)}(\boldsymbol{y}|u,\boldsymbol{x})$ depends on a vector of parameters of type $\boldsymbol{\eta}_u$ or on vectors of parameters $\boldsymbol{\eta}_{j|u}$ in the multivariate case under assumption (4). These parameters may be considered as unconstrained or may be constrained by a formulation which recalls those in (1) and (5), in which the design matrix does not depend on the covariates.

This is clarified in the following example.

Example 4 *As in Example 3, suppose that $Y^{(t)}$ is an ordinal variable with c categories. Then, we may directly parametrize the conditional distribution of each variable $Y^{(t)}$, given $U^{(t)}$ and $\boldsymbol{X}^{(t)}$, by letting $\boldsymbol{\eta}_u$ equal to the vector of global logits defined in (3), which are denoted by $\eta_{y|u}$. Otherwise, we may assume that*

$$\eta_{y|u} = \beta_{1y} + \beta_{2u}, \quad y = 1,\dots,c-1,\ u = 1,\dots,k,$$

where the parameter β_{2u} is a measure of the effect of the latent state on the probability of responding by a higher category, so that the latent states are easily interpretable in terms of tendency or propensity towards a certain behavior. An example of parametrizations of this type is for modeling data about drug consumption measured by an ordinal variables, where the parameter β_{2u} directly measures the tendency towards this behavior for subject in latent state u; see Section 8.1 for an description of an application of this type. □

Finally, it is interesting to note that by including in each vector of covariates $\boldsymbol{X}^{(t)}$ the lagged response variables, that is the vector $\boldsymbol{Y}^{(t-1)}$, we can easily relax the hypothesis of local independence by allowing *serial dependence*; see Bartolucci and Farcomeni (2009) for details.

3.1.2 Formulation of the latent model.

Here we illustrate different parameterizations of the initial and transition probabilities of the latent Markov chain.

For the initial probabilities, in particular, we assume

$$\boldsymbol{\lambda}_{\boldsymbol{x}} = \boldsymbol{Z}_{\boldsymbol{x}}\boldsymbol{\gamma},$$

where is a vector obtained by a particular link function of the probabilities $p(1|\boldsymbol{x}),\dots,p(k|\boldsymbol{x})$, $\boldsymbol{Z}_{\boldsymbol{x}}$ is a design matrix depending on the covariates in $\boldsymbol{x}$, and $\boldsymbol{\gamma}$ is the corresponding parameter vector. Similarly, for the transition probabilities we have

$$\boldsymbol{\lambda}_{\bar{u}\boldsymbol{x}}^{(t)} = \boldsymbol{Z}_{\bar{u}\boldsymbol{x}}^{(t)}\boldsymbol{\delta},$$

where $\boldsymbol{\lambda}_{\bar{u}\boldsymbol{x}}^{(t)}$ is a link function of $p^{(t)}(1|\bar{u}),\dots,p^{(t)}(k|\bar{u})$, $\boldsymbol{Z}_{\bar{u}\boldsymbol{x}}^{(t)}$ is a suitable design matrix depending on the previous latent state $\bar{u}$, and $\boldsymbol{\delta}$ is the corresponding vector of parameters. The above link functions that may be formulated on the basis on different types of logit. Typically, we use multinomial logits or global logits. As reference category, the multinomial logits have the first category when modeling the initial probabilities and category u when modeling the transition probabilities. Global logits are used when the latent states are ordered on the basis of a suitable parametrization of the conditional distribution of the response variables given the latent process. In any case, these link functions may be expressed as in (2). Moreover, it is worth noting that, in certain cases, it is reasonable to parametrize only certain transition probabilities, while leaving the others constrained to 0, so that a more parsimonious model results.

The following example helps to clarify the above parametrizations described above.

Example 5 - Tridiagonal transition matrix with logit parametrization. *Assuming that the latent states are ordered, we may require that the initial probabilities depend on the covariates in* $\boldsymbol{X}^{(1)}$ *through a global logits link function as follows*

$$\lambda_{u|\boldsymbol{x}} = \log\frac{p(u|\boldsymbol{x})+\cdots+p(k|\boldsymbol{x})}{p(1|\boldsymbol{x})+\cdots+p(u-1|\boldsymbol{x})} = \gamma_{1u} + \boldsymbol{x}'\boldsymbol{\gamma}_2, \quad u = 2,\dots,k.$$

Moreover, it is reasonable to assume a particular structure for the transition matrices in order to achieve a reduction of the model parameters. An example is when they are tridiagonal, so that *transition from state $\bar{u}$ is only allowed to state $u = \bar{u} - 1, \bar{u} + 1$; with $k = 4$, for instance, we have the transition matrix*

$$\boldsymbol{P}_{\boldsymbol{x}}^{(t)} = \begin{pmatrix} p^{(t)}(1|1,\boldsymbol{x}) & p^{(t)}(2|1) & 0 & 0 \\ p^{(t)}(1|2,\boldsymbol{x}) & p^{(t)}(2|2,\boldsymbol{x}) & p^{(t)}(3|2,\boldsymbol{x}) & 0 \\ 0 & p^{(t)}(2|3,\boldsymbol{x}) & p^{(t)}(3|3,\boldsymbol{x}) & p^{(t)}(4|3,\boldsymbol{x}) \\ 0 & 0 & p^{(t)}(3|4,\boldsymbol{x}) & p^{(t)}(4|4,\boldsymbol{x}) \end{pmatrix}. \tag{6}$$

This constraint only makes sense if the latent states are suitably ordered. In this case, we may also assume a parameterization based on logits of the following type

$$\lambda_{u|\bar{u}\boldsymbol{x}}^{(t)} = \log \frac{p^{(t)}(u|\bar{u},\boldsymbol{x})}{p^{(t)}(\bar{u}|\bar{u},\boldsymbol{x})} = \delta_{1u} + \delta_{2,(3+u-\bar{u})/2} + \boldsymbol{x}'\boldsymbol{\delta}_3, \quad t = 2,\ldots,T,\ \bar{u} = 1,\ldots,k,$$

where $u = 2$ for $\bar{u} = 1$, $u = k-1$ for $\bar{u} = k$, and $u = \bar{u}-1, \bar{u}+1$ for $\bar{u} = 2,\ldots,k-1$. □

Finally, we have to clarify that, when covariates are assumed to enter in the latent model, they are typically excluded from the measurement model and vice-versa. This simplifies the interpretation of the model and reduce the possibilities that the result model is unidentifiable. See also Section 2 for further comments about where to include the individual covariates depending on the aim of the application of interest and the type of data.

4 Manifest and posterior distributions

Regardless of the specific model formulation, the assumption of local independence implies that for conditional distribution of $\tilde{\boldsymbol{Y}}$, given $\tilde{\boldsymbol{X}}$ and $\boldsymbol{U}$, we have

$$f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U},\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\boldsymbol{u},\tilde{\boldsymbol{x}}) = \prod_t f^{(t)}(\boldsymbol{y}^{(t)}|u^{(t)},\boldsymbol{x}^{(t)}),$$

for any realization $\tilde{\boldsymbol{y}}$ of $\tilde{\boldsymbol{Y}}$, $\tilde{\boldsymbol{x}}$ of $\tilde{\boldsymbol{X}}$, $\boldsymbol{u}$ of $\boldsymbol{U}$. Moreover, since we assume that the latent process follows a first-order Markov chain, we have that

$$f_{\boldsymbol{U}|\tilde{\boldsymbol{X}}}(\boldsymbol{u}|\tilde{\boldsymbol{x}}) = p(u^{(1)}|\boldsymbol{x}^{(1)}) \prod_{t=2}^{T} p^{(t)}(u^{(t)}|u^{(t-1)},\boldsymbol{x}^{(t)}).$$

Now let

$$f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) = f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}})$$

denote the probability mass (or density) function for the *manifest distribution* of $\tilde{\boldsymbol{Y}}$ given $\tilde{\boldsymbol{X}}$. We have that

$$f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) = \sum_{\boldsymbol{u}} f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U},\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\boldsymbol{u},\tilde{\boldsymbol{x}}) f_{\boldsymbol{U}|\tilde{\boldsymbol{X}}}(\boldsymbol{u}|\boldsymbol{x}), \tag{7}$$

where the sum $\sum_{\boldsymbol{u}}$ is extended to all possible configurations $\boldsymbol{u}$ of the latent process.

Another distribution of fundamental importance is the *posterior distribution* of the latent process, which corresponds to the conditional distribution of $\boldsymbol{U}$ given $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$. In this case, we adopt the notation

$$q(\boldsymbol{u}|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = f_{\boldsymbol{U}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(\boldsymbol{u}|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}),$$

which corresponds to the probability mass function

$$q(\boldsymbol{u}|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = \frac{f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U},\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\boldsymbol{u}, \tilde{\boldsymbol{x}}) f_{\boldsymbol{U}|\tilde{\boldsymbol{X}}}(\boldsymbol{u}|\boldsymbol{x})}{f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}})}. \tag{8}$$

This distribution will be used for parameter estimation and for predicting the latent state at each time occasion.

It has to be clear that the sum in (7) may be extended over a huge number of configurations $\boldsymbol{u}$. The number of these configurations is k^T. Therefore, computing the manifest distribution of $\tilde{\boldsymbol{Y}}$, as expressed in (7), may be infeasible in real applications. The same also happens for the posterior distribution in (8), since it involves the manifest distribution at the denominator. However, in order to deal with these distributions, we can exploit certain recursions which have been developed, in a series of fundamental papers in the HM literature, by Baum and Welch and their coauthors; see Baum et al. (1970) and Welch (2003).

The first recursion of Baum and Welch is a forward recursion which allows us to compute the manifest distribution of $\tilde{\boldsymbol{Y}}$. Let

$$l^{(t)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) = f_{U^{(t)},\boldsymbol{Y}^{(1)},\ldots,\boldsymbol{Y}^{(t)}|\boldsymbol{X}^{(1)},\ldots,\boldsymbol{X}^{(t)}}(u, \boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(t)}|\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(t)}),$$

for $t = 1, \ldots, T$ and $u = 1, \ldots, k$. Then, this recursion is initialized with

$$l^{(1)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) = p(u|\boldsymbol{x}^{(1)}) f^{(1)}(\boldsymbol{y}^{(1)}|u, \boldsymbol{x}^{(1)}), \quad u = 1, \ldots, k.$$

and then is based on the following step

$$l^{(t)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) = \sum_{\bar{u}} l^{(t-1)}(\bar{u}, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) p^{(t)}(u|\bar{u}, \boldsymbol{x}^{(t)}) f^{(t)}(\boldsymbol{y}^{(t)}|u, \boldsymbol{x}^{(t)}), \quad u = 1, \ldots, k,$$

to be performed for $t = 2, \ldots, T$. At the end of the forward recursion, the manifest distribution is simply obtained as

$$f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) = \sum_{u} l^{(T)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}).$$

The other recursion introduced by Baum and Welch is a backward recursion, which allows us to obtain the posterior distribution of every latent state and or every pair of consecutive latent states. Let

$$m^{(t)}(\tilde{\boldsymbol{y}}|\bar{u}, \tilde{\boldsymbol{x}}) = f_{\boldsymbol{Y}^{(t+1)},\ldots,\boldsymbol{Y}^{(T)}|U^{(t)},\boldsymbol{X}^{(t+1)},\ldots,\boldsymbol{X}^{(T)}}(y^{(t+1)}, \ldots, y^{(T)}|\bar{u}, \boldsymbol{x}^{(t+1)}, \ldots, \boldsymbol{x}^{(T)}),$$

for $t = 1, \dots, T-1$ and $\bar{u} = 1, \dots, k$. This recursion, it initialized with

$$m^{(T)}(\tilde{\boldsymbol{y}}|\bar{u}, \tilde{\boldsymbol{x}}) = 1, \quad \bar{u} = 1, \dots, k,$$

and it is then based on the following steps

$$m^{(t)}(\tilde{\boldsymbol{y}}|\bar{u}, \tilde{\boldsymbol{x}}) = \sum_u m^{(t+1)}(\tilde{\boldsymbol{y}}|u, \tilde{\boldsymbol{x}}) p^{(t+1)}(u|\bar{u}, \boldsymbol{x}^{(t+1)}) f^{(t+1)}(\boldsymbol{y}^{(t+1)}|u), \tag{9}$$

where $\bar{u} = 1, \dots, k$, to be performed in reverse order, that is from $t = T-1$ to $t = 1$.

From the results of above recursion, and also using the results of the forward recursion, we obtain the following posterior distributions

$$q^{(t)}(u|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = f_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}}) = \frac{l^{(t)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) m^{(t)}(\tilde{\boldsymbol{y}}|u, \tilde{\boldsymbol{x}})}{f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}})}, \quad u = 1, \dots, k, \tag{10}$$

for $t = 1, \dots, T$. We also obtain

$$\begin{aligned} & q^{(t)}(\bar{u}, u|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = f_{U^{(t-1)}, U^{(t)}|\tilde{\boldsymbol{Y}}}(\bar{u}, u|\tilde{\boldsymbol{y}}) = \\ = \quad & \frac{l^{(t-1)}(\bar{u}, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) p^{(t)}(u|\bar{u}, \boldsymbol{x}^{(t)}) f^{(t)}(\boldsymbol{y}^{(t)}|u, \boldsymbol{x}^{(t)}) m^{(t)}(\tilde{\boldsymbol{y}}|u, \tilde{\boldsymbol{x}})}{f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}})}, \end{aligned} \tag{11}$$

with $\bar{u}, u = 1, \dots, k$ and for $t = 2, \dots, T$.

In order to efficiently implement the above recursions, it is important to express them using the matrix notation. At this regard we refer to Bartolucci (2006), Bartolucci and Pennoni (2007), and Zucchini and MacDonald (2009).

5 Likelihood inference

In the presence of individual covariates, the observed data correspond to the vectors $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{y}}_i$, for $i = 1, \dots, n$. In particular, $\tilde{\boldsymbol{x}}_i$ may be decomposed into the time-specific subvectors of covariates $\boldsymbol{x}_i^{(1)}, \dots, \boldsymbol{x}_i^{(T)}$. Similarly, we recall that $\tilde{\boldsymbol{y}}_i$ is made of the subvectors $\boldsymbol{y}_i^{(1)}, \dots, \boldsymbol{y}_i^{(T)}$, in the multivariate case.

Using the above notation, we have the following expression for the model log-likelihood:

$$\ell(\boldsymbol{\theta}) = \sum_i \log f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}_i|\tilde{\boldsymbol{x}}_i),$$

where $\boldsymbol{\theta}$ is the vector of the parameters. An equivalent expression that is computationally more convenient is

$$\ell(\boldsymbol{\theta}) = \sum_{\tilde{\boldsymbol{x}}} \sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} \log f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}),$$

where $n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}}$ is the joint frequency of the covariate configuration $\tilde{\boldsymbol{x}}$ and the response configuration $\tilde{\boldsymbol{y}}$.

The likelihood function can be maximized by an EM algorithm.

5.1 Expectation-Maximization algorithm

The EM algorithm is based on the complete data log-likelihood that, in the univariate categorical case, has expression

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) &= \sum_t\sum_u\sum_y\sum_{\boldsymbol{x}} a_{u\boldsymbol{x}y}^{(t)}\log f^{(t)}(y|u,\boldsymbol{x}) + \\ &+ \sum_u\sum_{\boldsymbol{x}} b_{u\boldsymbol{x}}^{(1)}\log p(u|\boldsymbol{x}) + \sum_{t>2}\sum_{\bar{u}}\sum_u\sum_{\boldsymbol{x}} b_{\bar{u}u\boldsymbol{x}}^{(t)}\log p^{(t)}(u|\bar{u},\boldsymbol{x}), \qquad (12)\end{aligned}$$

where $b_{u\boldsymbol{x}}^{(t)}$ is a frequency of the latent state u and covariate configuration $\boldsymbol{x}$ at occasion t; with reference to the same occasion and the covariate configuration, $b_{\bar{u}u\boldsymbol{x}}^{(t)}$ is the number of transitions from state $\bar{u}$ to state u, whereas $a_{u\boldsymbol{x}y}^{(t)}$ is the number of subjects that are in latent state u and provide response y. Expressions similar to that in (12) results for the complete data log-likelihood in more general cases, such as that of multivariate responses.

Based on the complete data log-likelihood, the two steps of the EM algorithm are as follows:

- **E-step**: it computes the conditional expected value of each frequency involved in (12); these expected values are denoted by $\hat{a}_{u\boldsymbol{x}y}^{(t)}$, $\hat{b}_{u\boldsymbol{x}}^{(t)}$, and $\hat{b}_{\bar{u}u\boldsymbol{x}}^{(t)}$. In particular, we have

$$\begin{aligned}\hat{a}_{u\boldsymbol{x}y}^{(t)} &= \sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}}q^{(t)}(u|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})I(\boldsymbol{x}^{(t)}=\boldsymbol{x},y^{(t)}=y),\\ \hat{b}_{u\boldsymbol{x}}^{(t)} &= \sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}}q^{(t)}(u|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})I(\boldsymbol{x}^{(t)}=\boldsymbol{x}),\\ \hat{b}_{\bar{u}u\boldsymbol{x}}^{(t)} &= \sum_{\tilde{\boldsymbol{x}}}\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}}q^{(t)}(\bar{u},u|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})I(\boldsymbol{x}^{(t)}=\boldsymbol{x}).\end{aligned}$$

  These expressions are based on the posterior probabilities introduced in Section 4; see, in particular, expressions (10) and (11).

- **M-step**: it consists of maximizing the complete data log-likelihood expressed as in (12), with each frequency substituted by the corresponding expected value. How to maximize this function depends on the specific formulation of the model and, in particular, on whether the covariates are included in the measurement model or in the latent model. We refer to Bartolucci (2006) and Bartolucci and Farcomeni (2009) for details.

5.1.1 Initialization of the algorithm.

As typically happens for latent variable and mixture models, the likelihood function may be multimodal. In particular, the EM algorithm could converge to a mode of the likelihood which does not correspond to the global maximum. In order to increase the chance of reaching the global maximum, the EM shall be initialized properly. We give guidelines below on deterministic starting solutions and suggest to compare the value at convergence with values obtained starting from randomly chosen initial values. For a similar multi-start strategy for mixture models see Berchtold (2004).

We illustrate in detail, in the case of univariate categorical responses, the deterministic rule we suggest to use for initializing the EM algorithm. It consists of computing the global logits for the observed distribution of the response variables. In practice, this amounts to computing:

$$\eta_y = \log \frac{\sum_{i=1}^n \sum_{t=1}^T I(y_i^{(t)} \geqslant y)}{\sum_{i=1}^n \sum_{t=1}^T I(y_i^{(t)} < y)}, \quad y = 1, \ldots, l-1.$$

Moreover, once a grid of points $\nu_1, \ldots, \nu_k$ is chosen (we suggest a grid of equispaced points between $-k$ and k), the conditional response probabilities are initialized as follows:

$$f(y|u, \boldsymbol{x}) = \phi_{y|u} - \phi_{y+1|u}, \quad y = 0, \ldots, c-1,$$

where

$$\phi_{y|u} = \begin{cases} 1 & y = 0, \\ \frac{\exp(\eta_{y-1}+\nu_u)}{1+\exp(\eta_{y-1}+\nu_u)} & y = 1, \ldots, c-1, \\ 0 & y = c, \end{cases}$$

for $u = 1, \ldots, k$. Note that $\phi_{y|u}$ are the values of the corresponding survival function, that is $\phi_{y|u} = p(Y^{(t)} \geqslant y | U^{(t)} = u)$. This rule guaranties that the conditional probabilities $\phi_{y|u}$ sum up to 1 and the resulting distribution of the response variables is statistically increasing with u. In the multivariate case, we suggest to apply this rule separately for each response variable, so as to initialize the conditional response probabilities $\phi_{jy|u}$, $j = 1, \ldots, r$, $u = 1, \ldots, k$, $y = 0, \ldots, c_j - 1$.

Finally, regardless of the nature of the response variables, we suggest to use the following starting values for the initial probabilities

$$p(u|\boldsymbol{x}) = \frac{1}{k}, \quad u = 1, \ldots, k,$$

whereas, for the transition probabilities, we suggest to use

$$p^{(t)}(u|\bar{u}, \boldsymbol{x}) = \frac{1}{h+k} \begin{cases} h+1, & u = \bar{u}, \\ 1, & u \neq \bar{u}, \end{cases}$$

for $t = 2, \ldots, T$, where h is a suitable constant (we generally use $h = 9$ in our applications).

The random starting rule that we propose is based on suitably normalized random numbers drawn from a uniform distribution from 0 and 1. In particular, in the univariate case we first draw each $\phi_{y|u}$, $u = 1, \ldots, k$, $y = 0, \ldots, c-1$, from this distribution and then we normalize these probabilities so that the constraint $\sum_{y=0}^{c-1} \phi_{y|u} = 1$ is satisfied for $u = 1, \ldots, k$. In a similar way, we suggest to choose each initial probability $p(u|\boldsymbol{x})$, $u = 1, \dots, k$, as a random number drawn from uniform distribution between 0 and 1 which is suitably normalized. The same may applied, for $t = 2, \dots, T$ and $=1, \dots, k$, to draw the transition probabilities $p^{(t)}(u|\bar{u}, \boldsymbol{x})$, $u = 1, \dots, k$, which must sum up to 1.

Inference is finally based on the solution corresponding to the largest value of the likelihood at convergence, which hopefully corresponds to the global optimum. The other local modes of the likelihood are not of interest for inference. In fact, the random rule allows us to adequately explore the parameter space, when its application is repeated a suitable number of times. Therefore, once an estimate is obtained starting with the deterministic rule, say $\hat{\boldsymbol{\theta}}_0$, we suggest to perform again the algorithm starting from a suitable number R of randomly chosen points of the parameters space, obtaining the estimates $\hat{\boldsymbol{\theta}}_1, \dots, \hat{\boldsymbol{\theta}}_R$. Then, we compare $\ell(\hat{\boldsymbol{\theta}}_0)$ with the maximum of $\ell(\hat{\boldsymbol{\theta}}_1), \dots, \ell(\hat{\boldsymbol{\theta}}_R)$. Provided that R is large enough, if $\ell(\hat{\boldsymbol{\theta}}_0)$ is not smaller than this maximum (up to a negligible tolerance level), then we can be confident that the solution based on the deterministic starting rule corresponds to the global maximum of $\ell(\boldsymbol{\theta})$. Otherwise, this rule needs to be somehow improved. In any case, we can take, as final estimate of the parameters, denoted by $\hat{\boldsymbol{\theta}}$, the one corresponding to the highest log-likelihood among $\hat{\boldsymbol{\theta}}_0, \dots, \hat{\boldsymbol{\theta}}_R$.

5.2 Information matrix, local identifiability, and standard errors

The EM algorithm uses neither the observed nor the expected information matrix which are suitable transformations of the second derivative matrix of the (incomplete) log-likelihood. From the inverse of one of these matrices, we obtain standard errors for the parameter estimates. Then, from the output of the EM algorithm, we have not an obvious mean of assessing the precision of these maximum likelihood estimates. This motivated Louis (1982) to develop a procedure for calculating the observed information matrix from the EM algorithm output. The proposed procedure involves expressing this matrix as the difference between two matrices corresponding to the total information, which we would have if we knew the latent states, and the missing information. However, the expression of the latter may be cumbersome to compute; see also Oakes (1999).

Several methods have been proposed to overcome this difficulty. Most of these methods have been developed within the literature on hidden Markov models; for a concise review see Lystig and Hughes (2002). The more interesting methods are based on the information matrix obtained from the EM algorithm by the technique of Louis (1982) or related techniques; see for instance Turner et al. (1998) and Bartolucci and Farcomeni (2009). Other interesting methods obtain the information matrix on the basis of the second derivative of the manifest probability of the response variables by a recursion similar to (9); see Lystig and Hughes (2002) and Bartolucci (2006). Among the methods related to the EM algorithm, that proposed by Bartolucci and Farcomeni (2009) is very simple to implement and requires a small extra code over that required for the maximum likelihood estimation. The method exploits a well-known result according to which the score of the complete data log-likelihood computed at the E-step of this algorithm corresponds to the score of the incomplete data log-likelihood. More precisely, we have

$$\boldsymbol{s}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left.\frac{\partial \mathrm{E}_{\bar{\boldsymbol{\theta}}}[\ell^*(\boldsymbol{\theta})|\boldsymbol{Y}]}{\partial \boldsymbol{\theta}}\right|_{\bar{\boldsymbol{\theta}}=\boldsymbol{\theta}},$$

where $\mathrm{E}_{\bar{\boldsymbol{\theta}}}[\ell^*(\boldsymbol{\theta})|\boldsymbol{Y}]$ denotes the conditional expected value of the complete-data log-likelihood computed at the parameter value $\bar{\boldsymbol{\theta}}$. Then, the observed information matrix, denoted by $\boldsymbol{J}(\boldsymbol{\theta})$, is obtained as minus the numerical derivative of $\boldsymbol{s}(\boldsymbol{\theta})$. The standard error of each parameter estimate is then obtained as the square root of the corresponding diagonal element of $\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}$. These standard errors may be used for hypothesis testing and for obtaining confidence intervals in the usual way.

The information matrix can also be used for checking local identifiability at $\hat{\boldsymbol{\theta}}$. We consider the model to be local identifiable if the matrix $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ is of full rank. See also McHugh (1956) and Goodman (1974).

6 Model choice and path prediction

A fundamental point in the LM model is the choice of the number of latent states, denoted by k. In certain applications, this number is a priori defined by the nature of the problem or the particular interest of the user. In most cases, however, it is selected on the basis of the observed data. At this aim, two main approaches are commonly used, which are based on likelihood ratio testing and on information criteria. The other aspect of interest is the prediction of the sequence of latent states for a given subject on the basis of his/her observable covariates and response variables. We discuss how to find the maximum a posteriori sequence of latent states for a given subject.

6.1 Model selection

A first approach is based on performing a likelihood ratio test between the model with k classes and that with $k+1$ classes for increasing values of k, until the test is not rejected. The adopted test statistic may be expressed as

$$LR = -2(\hat{\ell}_0 - \hat{\ell}_1),$$

where $\hat{\ell}_0$ is the maximum log-likelihood of the smaller model and $\hat{\ell}_1$ is that of the larger model. The problem of this approach is that, in order to obtain a p-value for LR, we need to use a parametric bootstrap procedure (Feng and McCulloch, 1996) based on a suitable number of samples simulated from the estimated model with k classes. This is because the standard regularity conditions, required to validly use the chi-squared distribution to compute these p-values, are not met in this case. It is important to note that Cheng and Liu (2001) propose a new consistency statistics to be evaluated with the bootstrap procedure to test hypotheses and find the best parsimonious model.

Another approach in order to choose the number of latent states is based on information criteria. The two most common criteria are the Akaike Information Criterion (AIC), see Akaike (1973), and the Bayesian Information Criterion (BIC), see Schwarz (1978). According to first criterion, we choose the number of states corresponding to the minimum of $AIC = -2\ell(\hat{\boldsymbol{\theta}}) + 2g$, where g is the number of free parameters; according to the second, we choose the model with the smallest value of $BIC = -2\ell(\hat{\boldsymbol{\theta}}) + g\log(n)$.

The performance of the two approaches above have been deeply studied in the literature on mixture models; see McLachlan and Peel (2000), Chapter 6. These criteria have also been studied in the hidden Markov literature for time series, where the two indices above are penalized with a term depending on the number of time occasions; see Boucheron and Gassiat (2007). From these studies, it emerges that BIC is usually preferable to AIC, as the latter tends to overestimate the number of latent states. The use of BIC is also criticized see among others Rusakov and Geiger (2005). The theoretical properties of AIC and BIC applied to the LM models are less studied. However, BIC is a commonly accepted model choice criterion even for these models. It has been applied by many authors, such as Langeheine (1994), Langeheine and Van de Pol (1994), and Magidson and Vermunt (2001). In particular, Bartolucci et al. (2009) suggested the use of this criterion together with that of diagnostic statistics measuring the goodness-of-fit and goodness-of-classification, whereas a simulation study may be found in Bartolucci and Farcomeni (2009).

Other indexes have been proposed in the literature to assess the quality of the classification. Bartolucci et al. (2009) propose to use

$$S = \frac{\sum_{\tilde{\boldsymbol{y}}} n_{\tilde{\boldsymbol{y}}} \sum_{t=1}^{T} [\hat{f}^{*(t)}(\tilde{\boldsymbol{y}}|u) - 1/k]}{(1-1/k)nT},$$

where $f^{*(t)}(\tilde{\boldsymbol{y}}|u)$ is the maximum, with respect to u, of the posterior probabilities $f^{(t)}_{U^{(t)}|\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{y}})$. Index S is always between 0 and 1, with 1 corresponding to the situation of absence of uncertainty in the classification, since one of such posterior probabilities is equal to 1 for every $\tilde{\boldsymbol{y}}$ and t, with all the other probabilities equal to 0.

6.2 Path prediction

Once the model has been estimated, a relevant issue is that of path prediction, i.e. finding the most likely sequence of latent states for a given subject on the basis of the responses he/she provided. For each subject in the sample, this is the sequence $\hat{\boldsymbol{u}} = (\hat{u}^{(1)}, \ldots, \hat{u}^{(T)})$ maximizing

$$f_{\boldsymbol{U}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}). \tag{13}$$

The problem above is different from that of finding the most likely state occupied by a subject at certain occasion, which is found maximizing $f_{U^{(t)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}}(u|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \propto q^{(t)}(u, \tilde{\boldsymbol{y}})$. This problem may be simply solved on the basis of the posterior probabilities computed at the last iteration of the EM algorithm (see Section 5) but does not correspond to maximize the joint probability in (13). In order to do so, we employ an efficient algorithm which avoids the evaluation of the posterior probability (13) at every configuration $\boldsymbol{u}$ of the latent process. This algorithm is known as the Viterbi algorithm (Viterbi, 1967; Juang and Rabiner, 1991), and is illustrated in the following.

For a given subject with response configuration $\boldsymbol{y}$, let $\tilde{r}^{(1)}(u) = f_{U^{(1)}|Y^{(1)}}(u^{(1)}, y^{(1)})$ and, for $t = 2, \ldots, T$, let

$$\tilde{r}^{(t)}(u) = \max_{u^{(1)},\ldots,u^{(t-1)}} f_{U^{(1)},\ldots,U^{(t)},Y^{(1)},\ldots,Y^{(t)}}(u^{(1)}, \ldots, u^{(t-1)}, u, y^{(1)}, \ldots, y^{(t)}).$$

A forward recursion can be used to compute the above quantities, and then a backward recursion based on these quantities can then be used for path prediction:

1. for $u = 1, \ldots, k$ let $\tilde{r}^{(1)}(u) = p(u)\phi_{y^{(1)}|u}$;
2. for $t = 2, \ldots, T$ and $v = 1, \ldots, k$ compute $\tilde{r}^{(t)}(v)$ as

$$\phi_{y^{(t)}|v} \max_u [\tilde{r}^{(t-1)}(u)p^{(t)}(u|\bar{u})];$$

3. find the optimal state $\tilde{u}^{(T)}$ as $\hat{u}^{(T)} = \arg\max_u \tilde{r}^{(T)}(u)$;
4. for $t = T-1, \ldots, 1$, find the optimal state $\tilde{u}^{(t)}$ as $\tilde{u}^{(t)} = \arg\max_u \tilde{r}^{(t)}(u)p^{(t+1)}(\hat{u}|\bar{u})$.

All the above quantities are computed on the basis of the ML estimate of the parameter $\boldsymbol{\theta}$ of the model of interest.

7 Bayesian estimation

In the Bayesian framework parameters are random variables, and information is summarized by approximating the posterior distribution which is proportional to the prior distribution and the likelihood.

An advantage of Bayesian inference is that one can easily include prior information, if available. Furthermore, there are computational advantages since there usually are much less problems from multimodality of the likelihood in approximating the posterior with respect to numerical maximization. For a general introduction on Bayesian inference see for instance Bernardo and Smith (1994). There are different approaches to Bayesian estimation for hidden Markov models (e.g., Robert et al. (2000), Cappé et al. (2005), Zucchini and MacDonald (2009), Spezia (2010)). Many of these approaches treat also k as an unknown, and efficiently derive a posterior on this quantity through transdimensional algorithms, e.g., reversible jump (Green, 1995). Simply removing the transimensional steps yields algorithms for fixed k.

In the Bayesian framework a first issue regards choice of prior inputs. When prior information is available, it should be summarized in the prior distributions. When prior information is not available, one can use a Dirichlet with parameter $\boldsymbol{e}$, where $\boldsymbol{e}$ denotes a vector of ones, for the initial probabilities and independently for each row of the transition matrix. This prior has been suggested as a universal default prior for multinomial parameters by Tuyl et al. (2009). When covariates are included in the model, a naturally arising prior distribution for the regression coefficients (and latent intercepts) is a zero-centered multivariate Gaussian, with covariance matrix either diagonal or proportional to the hat matrix. Zero-centered multivariate Gaussians can be also used for the log-odds parameters (Leonard, 1975; Nazaret, 1987) in the multivariate case. The posterior distribution is seldom available in closed form, and shall be approximated through a MCMC sampling scheme (Robert and Casella, 2010). We outline here an efficient Gibbs sampler for the basic LM model, in which there are not covariate and the conditional response probabilities are time-homogeneous. Our approach is based on the augmented likelihood (that is, the complete likelihood), is adapted from Chib (1996), and to the best of our knowledge it has not been considered previously in the LM literature. While simpler approaches work with time series in the HMM context, in the LM context where the number of occasions is much shorter efficient algorithms are necessary. For reasons of space we restrict to the basic LM model for fixed k, but the approach can be extended to more general cases.

Our Gibbs sampling scheme proceeds by iterating the following steps. First of all, the latent indicators $\boldsymbol{u}$ are updated through an *ff-bs* algorithm (Chib, 1996). In practice the usual forward-backward recursions are used to compute $q^{(t)}(u, \tilde{\boldsymbol{y}})$. We then compute $f_{U^{(t)}|Y}(u|\boldsymbol{Y}) \propto q^{(t)}(u, \tilde{\boldsymbol{y}})$, and sample latent indicators from this distribution. The sampled latent indicators are used to compute $b_u^{(t)}$ and $b_{\bar{u}u}^{(t)}$ for the current iteration, that is, the frequency of each latent state at each time and the frequency of each of the possible k^2 transitions at each time, for $t = 2, \ldots, T$. The initial probabilities shall then be sampled from a Dirichlet with parameter $\boldsymbol{e} + \boldsymbol{b}^{(1)}$. Assuming a time constant transition matrix, we then have that the $\overline{u}$-th row shall be sampled from a Dirichlet with parameter $\boldsymbol{e} + \sum_{t=2}^{T} \boldsymbol{b}_{\overline{u}}^{(t)}$, with $\boldsymbol{b}_{\overline{u}}^{(t)} = (b_{\overline{u}1}^{(t)}, \dots, b_{\overline{u}k}^{(t)})$.

At the end of each iteration, label switching is tackled by sorting conditional probabilities of a baseline category in increasing order. By fixing an order *a priori*, we are implicitly choosing one of the $k!$ modes of the likelihood that could be visited by the algorithm, and excluding all others. Frühwirth-Schnatter (2001) discussed in detail the advantages and disadvantages of this strategy.

A final issue regards choice of the number of latent states. This can be done with the parallel sampling approach of Congdon (2006), which in practice simply involves fitting the model repeatedly for all values of k in a range of plausible values. A posterior distribution on k is then approximated by post-processing the MCMC output, given certain assumptions of prior independence.

Possibilities for further work include development of a transdimensional strategy in the LM context, thereby treating k also as a random variable.

8 Empirical illustrations

In order to illustrate the approaches reviewed in this paper, we provide a synthetic overview of some interesting applications appeared in the literature. We describe different datasets, the aspects involved in practically fitting and using the LM models, and the interpretation of the results.

8.1 Marijuana Consumption dataset

The univariate version of the LM model without covariates was applied by Bartolucci (2006) to analyze a marijuana consumption dataset based on five annual waves of the "National Youth Survey" (Elliot et al., 1989). The dataset concerns $n = 237$ respondents who were aged 13 years in 1976. The use of marijuana was measured through five ordinal variables, one for each annual wave, with three categories corresponding to: "never in the past", "no more than once in a month in the past year", and "once a month in the past year". The substantive research question is whether there is an increase of marijuana use with age.

Using BIC, the selected LM model is with three latent states, homogeneous transition probabilities, and a parsimonious parameterization for the measurement model based on global logits as the one illustrated in Example 3. This parametrization is based on one parameter for each latent state, which may be interpreted as the tendency to use marijuana for a subject in this state, and one cutpoint for each response category. Then, the latent states may be ordered representing subjects with "no tendency to use marijuana", "incidental users of marijuana", and "subjects with high tendency to use marijuana". Also note that the cutpoints are common to all the response variables, since these variables correspond to repeated measurements of the same phenomenon under the same circumstances. This is because the dynamics of the marijuana consumption are only ascribed to the evolution of the underlying tendency of this consumption.

Bartolucci (2006) also tested different hypotheses on the transition matrix of the latent process. In particular, he found that the hypothesis that the transition matrix has a tridiagonal structure, like that in (6), cannot be rejected. The results under the selected LM model say that, at the beginning of the period of observation, 89.6% of the sample is in the first class (lowest tendency to marijuana consumption) and 1.5% is in the third class (highest tendency to marijuana consumption). An interesting interpretation of the pattern of consumption emerges from the estimated transition matrix. A large percentage of subjects remain in the same latent class, but almost the 25% of accidental users switch to the class of high frequency users. From the estimated marginal probabilities of the latent classes results that the tendency to use marijuana increases with age, since the probability of the third class increases across time.

8.2 Educational dataset

An interesting illustration of an LM model with individual covariates was given by Vermunt et al. (1999). They used data from an educational panel study conducted by the "Institute for Science Education in Kiel (Germany)" (Hoffmann et al., 1985). A cohort of secondary school pupils was interviewed once a year from grade 7 to grade 9 with respect to their interests in physics as a school subject. The response variables have been dichotomized with categories "low" and "high" to avoid sparseness of the observed frequency table. Based on these data, the LM model is used to draw conclusions on whether interest in physics depends on the interest in the previous period of observation and on two available covariates: sex and grade in physics at the present time.

Vermunt et al. (1999) estimated a univariate LM model with both initial and transition probabilities of the latent process depending on the available covariates according to a multinomial logit parametrization. In this model, the measurement error was constrained to be the same for all time points, meaning that the conditional distribution of the response variables given the latent state is the same for every occasion. Then, they relaxed some of the basic assumptions of the LM model, such as the assumption that the Markov chain is of first-order.

According to the parameter estimates of the selected model, there is a significant effect of sex and grade on the interest in physics. Pupils with higher grades are more interested in physics than pupils with lower grades, girls are less interested in physics than boys. Moreover, the interest has a positive effect on the grade at the next time occasion. For the boys, the probability of switching from “low” to “high interest” is larger than that for girls, as well as to keep their interest high.

8.3 Criminal dataset

The multivariate version of the LM model where both the initial and the transition probabilities of the latent process depend on time-constant covariates was illustrated by Bartolucci et al. (2007); for a related study see Roeder et al. (1999). They analyzed the conviction histories of a cohort of offenders who were born in England and Wales in 1953. The offenders were followed from the age of criminal responsibility, 10 years, until the end of 1993. They were grouped in 10 major categories and gender was included in the model as explanatory variable. The analysis was based on $T = 6$ age bands: 10-15, 16-20, 21-25, 26-30, 31-35 and 36-40 years. The adopted LM model allows to estimate trajectories for behavioral types which are determined by the criminal conviction grouping. It also allows to give rise to a general population sample by augmenting the observed sample with not-convicted subjects.

According to Bartolucci et al. (2007), the fit of the model is considerably improved by relaxing the assumption of the homogeneity of the latent Markov chain, but retaining the constraint that males and females have the same transition probabilities. In particular, they selected a model based on partial homogeneity, in which there are two transition probability matrices: the first for transitions up to time $\bar{t}$ and the second from time $\bar{t}$ and beyond. The choice of $\bar{t} = 2$ has been made on the basis of the BIC of the estimated models with different values for $\bar{t}$.

In summary, the selected model is based on a partially homogeneous Markov chain with five latent states, different initial and equal transition probabilities for males and females. From the estimated conditional probabilities of conviction for any offense group and any latent state, it was possible to determine classes of criminal activity. In accordance to the typologies found in the criminological latent class literature, these classes are interpreted as: “non-offenders”, “incidental offenders”, “violent offenders”, “theft and fraud offenders” and “high frequency and varied offenders”.

According to the estimated initial probabilities in the first age band the percentage of males who are incidental offenders is higher than that of females. The common estimated transition probabilities for males and females from age band 10-15 to age band 16-20, and from one age band to the others for offenders over 16, show that the first transition occurs at an early age, 16 years, which in western society represents the peak of the age-crime curve.

At the first time occasions, “incidental offenders” have a quite high probability of persistence when moving from the age band 10-15 to age band 16-20. Moreover, “theft and fraud offenders” are mainly females and they have a high chance of moving to the class of non-offenders. The “high frequency and varied offenders” are mainly males and they have a high persistence.

According to the estimated transition probabilities from age band 16-20 to the subsequent age bands the subjects belonging to the latent state of “non-offenders” have a very low chance of becoming offenders; “theft and fraud offenders” and “violent offenders” have a high probability of dropping out of crime. From the estimated proportion of males and females in each latent state at every time occasion resulted that 7% of males are “violent offenders” at age 16-20 years and 32% are “incidental offenders” at the same age. Only 3% of females are “theft and fraud offenders” at age 16-20 years.

8.4 Dataset on financial products preferences

An interesting analysis of data obtained from face to face interviews of the household ownership of 12 financial products was offered by Paas et al. (2009). The panel was conducted by a market research company among 7676 Dutch households in 1996, 1998, 2000 and 2002. To have an accurate representation of the products portfolio, the households were asked to retrieve their bank and insurance records in order to check which product they own. Households that drop out were replaced to ensure the representativeness of the sample for the population with respect to demographic variables, such as age, income and marital status. The aim of the study was to get insights on the developments of the individual household product portfolio and on the effect of demographic covariates on such development. It also concentrates on predicting future behaviors of acquisition.

The authors proposed to use a time homogeneous multivariate LM model, with time-varying covariates affecting the latent process. They added additional assumptions to the model, such as constant conditional probabilities of the response variables given the latent process. This is done to avoid manifest changes, so that the product penetration levels are consistent in latent states over measurement occasions. Moreover, they formulated the model with a time-constant effect of the covariates on the transition probabilities.

The model selected on the basis of BIC is based on nine latent states. These states can be ordered according to increasing penetration levels across the analyzed products, which range from bonds, the most common owned product, to saving accounts. The results highlight some divergences from common order of acquisitions, such as the acquisition of a mortgage before owing a credit card or vice-versa. Loans and unemployment insurance are the most often acquired products. According to the estimated transition matrix there is a high persistence in the same latent state: only 14% of the households changed latent state in the period of the study. The most common switch is from latent class 7 to 8, where latent state 7 is characterized by the acquisition of mortgage, life insurance, pension found, car insurance, and saving accounts. Whereas latent state 8 for all the previous products plus the credit card. Another common switch is from latent state 4 to 7, where the first is characterized by the acquisition of life insurance, pension found, car insurance, and savings accounts. This means that multiple products were acquired between consecutive measurement occasions.

Income, age of the head of the household, and household size have a significant effect on the initial and transition probabilities according to the Wald test. The covariate values implying a larger probability of belonging to an initial latent class also imply a greater probability of switching into the same latent class. For example larger households are relatively often found in latent states where overall product ownership probabilities are relatively low.

The prediction of future purchase of a financial product was performed on the basis of the posterior latent state membership probabilities for each household at the last occasion, given all other observed information. To assess the accuracy of the forecasting, the authors used the Gini coefficient as a measure of concentration. Considering the empirical results in the last wave refereed to year 2002, which was not considered when estimating the model, the authors showed that, for most products, the prediction equations are effective for forecasting household acquisition.

8.5 Job position dataset

The multivariate LM model with covariates affecting the manifest probabilities proposed by Bartolucci and Farcomeni (2009) was applied by these authors to data extracted from the “Panel Study of Income Dynamics” database (University of Michigan). These data concern $n = 1446$ women who were followed from 1987 to 1993. The binary response variables are fertility, indicating whether a woman had given birth to a child in a certain year and employment, indicating whether she was employed. The covariates are: race (dummy variable equal to 1 for a black woman), age (in 1986), education (year of schooling), child 1-2 (number of children in the family aged between 1 and 2 years, referred to the previous year), child 3-5, child 6-13, child 14-, income of the husband (in dollars, referred to the previous year).

The main issue concerns the direct effect of fertility on employment. Also of interest are the strength of the state dependence effect for both response variables and how these variables depend on the covariates.

Bartolucci and Farcomeni (2009) used an LM model with covariates affecting the manifest probabilities since they were interested in separately estimating the effect of each covariate on each outcome. The proposed LM model allows to separate these effects from the unobserved heterogeneity, by modeling the latter with a latent Markov process. In this way, unobserved heterogeneity effects on the response variables are allowed to be time-varying; this is not allowed neither within a latent class model with covariates nor in the most common random effect models.

The model selected using AIC and BIC is with three latent states. Under this model, race has a significant effect on fertility, but not on employment according to the estimates of the parameters affecting the marginal logits of fertility and employment and the log-odds ratio between these variables. Age has a stronger effect on fertility than on employment. Education has a significant effect on both fertility and employment, whereas the number of children in the family strongly affects only the first response variable and income of the husband strongly affects only the second one.

The log-odds ratio between the two response variables, given the latent state, is negative and highly significant, meaning that the response variables are negatively associated when referred to the same year. On the other hand, lagged fertility has a significant negative effect on both response variables and lagged employment has a significant effect, which is positive, on both response variables. Therefore, fertility has a negative effect on the probability of having a job position in the same year of the birth, and the following one. Employment is serially positively correlated (as consequence of the state dependence effect) and fertility is negatively serially correlated.

From the estimates of the support points for each latent state it may be deduced that the latent states correspond to different levels of propensity to give birth to a child and to have a job position. The first latent state corresponds to subjects with the highest propensity to fertility and the lowest propensity to have a job position. On the contrary, the third latent state corresponds to subjects with the lowest propensity to fertility and the highest propensity to have a job position. Finally, the second state is associated to intermediate levels of both propensities. The two propensities are negatively correlated.

Overall, it results that the 78.5% of women started and persisted in the same latent state for the entire period, whereas for the 21.5% of women had one or more transitions between states. The presence of these transitions is in accordance to the rejection of the hypothesis that a latent class model is suitable for these data.

8.6 Dataset on anorectic patients

An interesting extension of the model to account for a hierarchical structure of the data has been recently proposed by Rijmen et al. (2007). They illustrated the model by a novel application using a data set form an ecological momentary assessment study (Vansteelandt et al., 2007) on the course of emotions among anorectic patients. At nine occasions for each of the seven days of observations, 32 females with eating disorders received a signal and were asked to rate themselves on a 7-point scale with respect to the intensity with which they experienced 12 emotional states. These were taken from the following emotional categories: “anger and irritation”, “shame and guilt”, “anxiety and tension”, “sadness and loneliness”, “happiness and joy”, “love and appreciation”. The response has been dichotomized (0-2 vs. 3-6) and the signal has been considered equally spaced. The aim of the study was to detect the course of emotion among the patients.

As a preliminary analysis, Rijmen et al. (2007) used the univariate version of the LM model without covariates. They treated each person by day combination as a separate case assuming that the data stemming from different days were independent and that the parameters were constant over days. On the basis of such model the authors selected four latent states. The first state is interpreted as positive mood, the third as negative mood, the second as low intensity for all emotions except tension, and the fourth as neutral to moderately positive mood. According to the estimated transition matrix there is high persistence in the same state. The probability of moving from state 1 to state 2 is 0.14, from state 3 to 4 is 0.18 and from state 4 to 3 is 0.14 . They noted that there is an indirect transition from state 3 to state 1 via the emotionally more neutral state 4. Over the days there is an increase of the marginal probabilities of states 1 and 4 indicating that the mood of patients tends to become better later on in the day.

Then the authors also used a hierarchical LM model by introducing a latent variable at day level to account for the fact that data stemming from different days are not independent. They modeled the transition between latent states at day and signal levels by a first-order time homogeneous Markov chain. They estimated a model with two states at the day level and two signal states within each day-state. For the first day state, the signal state 1 is characterized by high probabilities of experiencing positive emotions and low probabilities of negative emotions. Therefore this state is interpreted as positive mood. Instead the second day state is interpreted as negative mood. In the signal state 2 positive emotions are not well separated from negative emotions and the state is considered as an emotionally neutral to moderately positive state. A tendency to experience more positive emotions emerges from the estimated initial and conditional probabilities of the chain over day and signal.

9 Conclusions and further developments

In this paper, we presented a review of the latent Markov model which starts with the more general model for multivariate data with the inclusion of covariates. Then, we illustrated several constrained versions of the general LM model, and outlined maximum likelihood estimation and related inferential methods. We tried to keep the presentation simple and provide references to methods which are of easy implementation, taking ideas from methods widely applied in the literature on hidden Markov models for time-series data. We pointed out that longitudinal data are similar to time-series data but they are referred to several units observed at few time occasions. The use of longitudinal data is witnessed by the widespread use of these data in many research fields. Moreover, in order to retain a simple presentation we dealt with the case of balanced data in which all subjects are observed at the same number of occasions, the same response variables are obtained at every occasion, and there are no missing responses.

The potential of the LM approach is illustrated by summarizing the results of several applications available in the literature. It is worth noting that the number of applications available in the literature is growing; recent application also involve situations which are not typically those of longitudinal data. Examples concern the estimation of closed-populations on the basis of capture-recapture data (Bartolucci and Pennoni, 2007) and the assessment of the peer review process of grant applications (Bornmann et al., 2008).

In this section we would like to summarize further developments of the LM models that, in our opinion, deserve particular attention:

- *More flexible parametrizations for the conditional distribution of the response variables given the latent process*: We refer in particular to parametrizations which address the problem of multidimensionality when the model is applied, for instance, to data coming from the administration of test items assessing different types of ability. This is a problem well known in the Item Response Theory, which may concern different fields of applications such as psychology and assessment of the quality of life. Flexible parametrizations are also of interest when the response variables are of mixed nature; for instance, we may have a mixture of binary, ordinal, and continuous variables.

- *Higher-order Markov chains*: the assumptions that the latent Markov chain is of first-order may be restrictive when there is "memory effect" such that the latent state of a subject at a given occasion depends on two or more previous occasions. The extension to second or higher-order Markov chains is of interest. This extension is rather simple to implement and has been already applied in certain contexts. However, to our knowledge, a systematic illustration of this higher-order approach does not seem to be available in the literature.

- *Missing responses*: It is well known that longitudinal datasets typically suffer from the problem of missing responses due to several reasons. In applying an LM model, a trivial solution to this problem consists in eliminating from the sample subjects with at least one missing response. Obviously, this solution may lead to a strong bias of the parameter estimates when the assumption that these responses are missing at random is not plausible. The interest here is in formulating an LM approach in which the event of missing response is explicitly modeled given the latent state and in which the evolution of the latent process takes into account this possibility by adopting a suitable parametrization of the transition matrices. Another possible solution can in principle be given by the EM algorithm, which can be used to impute missing data.

- *Time-dependent covariates and causality*: We have made the assumption in this paper that covariates are strictly exogenous. There are cases in which time-dependent covariates may be influenced by other covariates not included and/or from the outcomes at previous occasions. This case of *time-dependent confounding* can possibly lead to biased estimates. A modification of the EM algorithm to account for the true underlying causal structure would lead to unbiased estimates and is ground for further work.

Acknowledgements

The authors are supported by the EIEF research grant ”Advances in non-linear panel models with socio-economic applications”. F. Bartolucci and F. Pennoni also acknowledge the financial support from PRIN 2007.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In Petrov, B. N. and F., C., editors, *Second International symposium on information theory*, pages 267–281, Budapest. Akademiai Kiado.

Altman, R. M. (2007). Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102:201–210.

Anderson, T. W. (1951). Probability models for analysing time changes in attitudes. In Corporation, T. R., editor, *The use of mathematical models in the measurement of the attitudes*. Lazarsfelsd P. F.

Anderson, T. W. (1954). Probability models for analysing time changes in attitudes. In F., L. P., editor, *Mathematical Thinking in the Social Science*. The Free press.

Andersson, S. and Rydén, T. (2009). Subspace estimation and prediction methods for hidden markov models. *The Annals of Statistics*, 37:4131–4152.

Archer, G. E. B. and Titterington, D. M. (2009). Parameter estimation for hidden markov chains. *J. Statist. Plann. Inference*, 108:365–390.

Bartolucci, F. (2006). Likelihood inference for a class of latent markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society, series B*, 68:155–178.

Bartolucci, F. and Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent markov heterogeneity structure. *Journal of the American Statistical Association*, 104:816–831.

Bartolucci, F., Lupparelli, M., and Montanari, G. E. (2009). Latent Markov model for binary longitudinal data: an application to the performance evaluation of nursing homes. *Annals of Applied Statistics*, 3:611–636.

Bartolucci, F. and Pennoni, F. (2007). A class of latent markov models for capture-recapture data allowing for time, heterogeneity and behavior effects. *Biometrics*, 63:568–578.

Bartolucci, F., Pennoni, F., and Francis, B. (2007). A latent markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society, Series A*, 170:151–132.

Bartolucci, F., Pennoni, F., and Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioural Statistics, in press*, 36:491–522.

Baum, L. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37:1554–1563.

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171.

Berchtold, A. (2004). Optimization of mixture models: Comparison of different strategies. *Computational statistics*, 19:385–406.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* Wiley, Chichester.

Bickel, P. J., Ritov, Y., and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals of Statistics*, 26:1614–1635.

Bollen, K. A. and Curran, P. J. (2006). *Latent curve models: A structural equation perspective.* Wiley, Hoboken, NJ.

Bornmann, L., Mutz, R., and Daniel, H.-D. (2008). Latent markov modeling applied to grand peer review. *Journal of Informetrics*, 2:217–228.

Boucheron, S. and Gassiat, E. (2007). An information-theoretic perspective on order estimation. In O. Cappé, E. Moulines, T. R., editor, *Inference in Hidden Markov Models*, pages 565–602. Springer.

Bye, B. V. and Schechter, E. S. (1986). A latent Markov model approach to the estimation of response error in multiwave panel data. *Journal of the American Statistical Association*, 81:375–380.

Cappé, O., Moulines, E., and Rydén, T. (1989). *Inference in Hidden Markov Models.* Springer-Verlag, New York.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models.* Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Cheng, R. C. H. and Liu, W. B. (2001). The consistency of estimators in finite mixture models. *Scandinavian Journal of Statistics*, 28:603–616.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, 75:79–97.

Collins, L. M. and Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, 27:131–157.

Colombi, R. and Forcina, A. (2001). Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, 88:1007–1019.

Congdon, P. (2006). Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Computational Statistics & Data Analysis*, 50:346–357.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Dias, J. G. and Vermunt, J. K. (2007). Latent class modeling of website users' search patterns: Implications for online market segmentation. *Journal of Retailing and Consumer Services*, 14:359–368.

Elliot, D. S., Huizinga, D., and Menard, S. (1989). *Multiple Problem Youth: Delinquency, Substance Use, and Mental Health Problems*. Springer-Verlag, New York.

Feng, Z. and McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *J. R. Statist. Soc.*, 58:609–617.

Fitzmaurice, G., Davidian, M., Verbeke, G., and G., M., editors (2009). *Longitudinal data analysis*. Chapman and Hall, CRC, London.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden markov models. *Machine Learning*, 29:245–273.

Glonek, G. F. V. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society B*, 57:533–546.

Goodman, L. A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56:841–868.

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

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and bayesian model determination. *Biometrika*, 82:711–732.

Hoffmann, L., Lehrke, M., and Todt, E. (1985). Development and changes in pupils' interest in physics (grade 5 to 10): Design of a longitudinal study. In Lehrke, M., Hoffmann, L., and Gardner, P. L., editors, *Interest in Science and Technology Education*, pages 71–80, Kiel. IPN.

Juang, B. and Rabiner, L. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33:251–272.

Kaplan, D. (2008). An overview of markov chain methods for the study of stage-sequential developmental processes. *Developmental Psychology*, 44:457–467.

Koski, T. (2001). *Hidden Markov Models for Bioinformatics.* Kluwer, Dordrecht.

Künsch, H. R. (2005). State space and hidden markov models. In O. E. Barndorff-Nielsen, D. R. C. and C. Klüppelberg, e., editors, *Complex Stochastic Systems*, pages 109–173, Boca Raton, FL. Chapman and Hall/CRC.

Langeheine, R. (1988). New development in latent class theory. In Langeheine, R. and J., R., editors, *Latent trait and latent class models*, pages 77–108. New York: Plenum Press.

Langeheine, R. (1994). Latent variables markov models. In von Eye, A. and Clogg, C., editors, *Latent variables analysis: Applications for developmental research*, pages 373–395, Thousand Oaks, CA. Sage.

Langeheine, R. and Van de Pol, F. (1994). Discrete-time mixed markov latent class models. In Dale, A. and Davies, R., editors, *Analyzing Social and Political Change: a Casebook of Methods*, pages 171–197, London. Sage Publications.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. S., editor, *Measurement and Prediction*, New York. Princeton University Press.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis.* Houghton Mifflin, Boston.

Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society (Ser. B)*, 37:23–37.

Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell System Technical Journal*, 62:1035–1074.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B: Methodological*, 44:226–233.

Lystig, T. C. and Hughes, J. (2002). Exact computation of the observed information matrix for hidden markov models. *Journal of Computational and Graphical Statistics*, 11:678–689.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and other Models for Discrete-Valued Time Series*. Chapman and Hall, London.

Magidson, J. and Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology*, 31:223–264.

Maruotti, A. (2011). Mixed hidden markov models for longitudinal data: An overview. *International Statistical Review*, 79.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd Edition*. Chapman and Hall, CRC, London.

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika*, 21:331–347.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data growth mixture modeling and related techniques for longitudinal data. In Kaplan, D., editor, *Handbook of quantitative methodology for the social sciences*, pages 345–368. Sage Publications, Newbury Park, CA.

Nagin, D. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods*, 4:139–157.

Nazaret, W. (1987). Bayesian log-linear estimates for three-way contingency tables. *Biometrika*, 74:401–410.

Paas, L. J., Vermunt, J. K., and Bijlmolt, T. H. A. (2009). Discrete time, discrete state latent markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society, Series A*, 170:955–974.

Rijmen, F., Vansteelandt, K., and De Boeck, P. (2007). Latent class models for diary methods data: parameter estimation by local computations. *Psychometrika*, 73:167–182.

Robert, C., Ryden, T., and Titterington, D. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B*, 62:57–75.

Robert, C. P. and Casella, G. (2010). *Monte Carlo Statistical Methods, 2nd Edition*. Springer-Verlag, New York.

Roeder, K., Lynch, K. G., and Nagin, D. S. (1999). Modeling uncertainty in latent class membership: a case study in criminology. *Journal of the American Statistical Association*, 94:766–776.

Rost, J. (2002). Mixed and latent markov models as item response models. *Methods of Psychological Research Online, Special Issue*, pages 53–70.

Rusakov, D. and Geiger, D. (2005). Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research*, 6.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Spezia, L. (2010). Bayesian analysis of multivariate gaussian hidden Markov models with an unknown number of regimes. *Journal of Time Series Analysis*, 31:1–11.

Turner, R. (2008). Direct maximization of the likelihood of a hidden Markov model. *Computational Statistics and Data Analysis*, 52:4147–4160.

Turner, T. R., Cameron, M. A., and Thomson, P. J. (1998). Hidden Markov chains in generalized linear models. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 26:107–125.

Tuyl, F., Gerlach, R., and Mengersen, K. (2009). Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. *Bayesian Analysis*, 4:151–158.

van de Pol, F. and Langeheine, R. (1990). Mixed markov latent class models. *Sociological Methodology*, 20:213–247.

Vansteelandt, K., Rijmen, F., Pieters, G., and Vanderlinden, J. (2007). Drive for thinness, affect regulation and physical activity in eating disorders: a daily life study. *Behaviour Research and Therapy*, 45:1717–1734.

Vermunt, J. (2010). Longitudinal research with latent variables. In van Montfort, K., Oud, J., and Satorra, A., editors, *Handbook of Advanced Multilevel Analysis*, pages 119–152. Springer, Heidelberg, Germany.

Vermunt, J. K., Langeheine, R., and Böckenholt, U. (1999). Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24:179–207.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269.

Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, 53:1–13.

Wiggins, L. (1973). *Panel Analysis: Latent probability models for attitude and behaviours processes.* Elsevier, Amsterdam.

Wiggins, L. M. (1955). *Mathematical models for the Analysis of Multi-wave Panels.* PhD thesis, Columbia University, Ann Arbor.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for time series: an introduction using R.* Springer-Verlag, New York.