

Pairwise likelihood inference for multilevel latent Markov models

Francesco Bartolucci and Monia Lupparelli

Abstract We propose an extended latent Markov model for categorical longitudinal data with a multilevel structure. This extension allows us to take into account the correlation which may arise between the responses provided by individuals belonging to the same cluster and to model the cluster effect in a dynamic fashion. Given the complexity of computing the manifest distribution, we make inference on the model through a composite likelihood function based on all the possible pairs of subjects within every cluster. The resulting approach is illustrated through an application to a dataset concerning a sample of Italian workers.

Key words: composite likelihood information criterion, hidden Markov model, EM algorithm

1 Introduction

The Latent Markov (LM) model (see Wiggins, 1973; Langeheine and van de Pol, 1990) represents a standard tool for the analysis of categorical longitudinal data, especially when the main aim of the analysis is describing the individual change with respect to a certain latent status over the time; for a review see Bartolucci et al. (2010a). The basic assumption of this model is that the response variables corresponding to the different time occasions are conditionally independent given a discrete latent process which follows a first-order Markov chain.

The LM model has been extended in several ways. One of the most interesting extensions is for the inclusion of time-fixed and time-varying individual covariates.

Francesco Bartolucci

Department of Economics, Finance, and Statistics, University of Perugia,
e-mail: bart@stat.unipg.it

Monia Lupparelli

Department of Statistical Sciences, University of Bologna,
e-mail: monia.lupparelli@unibo.it

Another interesting extension is to multilevel data. This kind of extension has been developed by Bartolucci and Lupparelli (2007) and Bartolucci et al. (2010b) on the basis of a set of cluster-level discrete latent variables which affect the initial and transition probabilities of the individual-level latent processes. A similar extension based on random effects having a continuous distribution has been developed by Asparouhov and Muthén (2008); see also Altman (2007). All these extensions are based on the assumption that the clusters have a time-constant effect on the distribution of the response variables of the subjects they include.

In this paper, we propose a multilevel extension of the LM model in which the cluster effect is not restricted to be time-constant. In particular, we model the effect of each cluster by a sequence of latent variables which are time-specific. Each cluster-level latent process follows a first-order Markov chain with initial and transition probabilities that may possibly depend on available covariates. A model results which is based on a set of nested latent processes. The latent processes referred to the subjects in the same cluster are conditionally independent given the cluster-level latent process.

Under the proposed model, the manifest distribution of the response variables is computationally untractable in most applications. Therefore, to make inference on the model we exploit an approach based on a composite likelihood (Lindsay, 1988; Cox and Reid, 2004), which is computed on the basis of the joint distribution of the response variables for each pair of subjects in the same cluster. A similar approach was followed by Renard et al. (2004) to deal with a multilevel probit model; for applications of this inferential approach to similar contexts see Hjort and Varin (2008) and Varin and Czado (2010). In particular, we show how to compute the pairwise likelihood by the same recursion exploited by Baum et al. (1970) to deal with hidden Markov model and how to maximize this likelihood by an EM algorithm similar to the one they suggest. We also show how to obtain standard errors for the parameter estimates and how to make model selection on the basis of the composite likelihood information criterion developed by Varin and Vidoni (2005).

The paper is organized as follows. In the next section we briefly review the basic LM model and its maximum likelihood estimation. Section 3 illustrates the proposed multilevel extension, whereas pairwise likelihood inference for this model is described in Section 4. In Section 5 we illustrate the model by an application based on a dataset concerning a sample of individuals working in different Italian firms. Finally, in Section 6 we describe future extensions of the proposed approach.

2 The latent Markov model

Consider a panel of n subjects observed at T occasions and let $Y_i^{(t)}$ denote the response variable of interest for subject i at occasion t . In our context, the response variables are typically categorical, although the LM model may be also applied to the case of continuous response variables. Then, we denote by l the number of response categories. These categories are labeled from 0 to $l - 1$.

2.1 Model assumptions

We assume that, for each subject i , $i = 1, \dots, n$, the response variables $Y_i^{(1)}, \dots, Y_i^{(T)}$ are conditionally independent given a latent process $\mathbf{V}_i = (V_i^{(1)}, \dots, V_i^{(T)})'$, which follows a first-order Markov chain. This chain has k states, labeled from 1 to k , with initial probabilities $\pi_i(v) = p(V_i^{(1)} = v)$, $v = 1, \dots, k$, and transition probabilities $\pi_i^{(t)}(v^{(1)}|v^{(0)}) = p(V_i^{(t)} = v^{(1)} | V_i^{(t-1)} = v^{(0)})$, $t = 2, \dots, T$, $v^{(0)}, v^{(1)} = 1, \dots, k$. The initial probabilities are collected in the vectors $\boldsymbol{\pi}_i$, whereas the transition probabilities are collected in the matrices $\boldsymbol{\Pi}_i^{(t)}$.

For subject i at occasion t , the latent variable $V_i^{(t)}$ corresponds to the level of the unobservable characteristic of interest. The way in which this characteristic affects the corresponding response variable $Y_i^{(t)}$ depends on the conditional probabilities $\psi_i^{(t)}(y|v) = p(Y_i^{(t)} = y | V_i^{(t)} = v)$, $v = 1, \dots, k$, $y = 0, \dots, l-1$. These conditional response probabilities are collected in the matrices $\boldsymbol{\Psi}_i^{(t)}$.

In order to achieve a more parsimonious model, some constraints must be adopted on the parameters of the distribution of the latent process or on those involved in the distribution of the response variables given this process.

A natural assumption concerning the latent process is that the initial and transition probabilities are common to all subjects, i.e. $\boldsymbol{\pi}_i = \boldsymbol{\pi}$, $\boldsymbol{\Pi}_i^{(t)} = \boldsymbol{\Pi}^{(t)}$, $i = 1, \dots, n$, $t = 2, \dots, T$. These constraints may be assumed together with the constraint that the transition probabilities are time-homogeneous, and then we have one transition matrix $\boldsymbol{\Pi}$ common to all subjects and occasions. Alternatively, the initial and the transition probabilities of the latent process can be modeled through a linear or a generalized linear parametrization which may involve time-fixed and time-varying individual covariates. For instance, Bartolucci et al. (2009) formulated an LM model to study the evolution of the health status of elderly people, in which initial and transition probabilities are parametrized by cumulative-logits. How to formulate a linear parametrization on the transition probabilities has been explored by Bartolucci (2006). This type of parametrization allows us to impose the constraint that certain elements of the transition matrix are equal to 0, so that transition is not allowed between certain pairs of states.

Constraints can also be assumed on the conditional response probabilities. For instance, these probabilities are generally assumed to be equal for each individual and to be time-homogeneous, i.e. $\boldsymbol{\Psi}_i^{(t)} = \boldsymbol{\Psi}$, $i = 1, \dots, n$, $t = 1, \dots, T$. Therefore, we have a single probability matrix $\boldsymbol{\Psi}$, the elements of which allow us to easily interpret the latent states. An ordinal constraint for these probabilities is assumed by Bartolucci et al. (2009), whereas Bartolucci et al. (2008) proposed a Rasch model for the conditional probabilities. Whenever the available set of individual covariates does not directly affect the latent states, it can be used for modeling the conditional response probabilities in order to explain the unobserved individual heterogeneity over the time.

2.2 Maximum likelihood estimation

Obviously, the maximum likelihood method requires to compute the manifest distribution of each observed response configuration $\mathbf{y}_i = (y_i^{(1)}, \dots, y_i^{(T)})'$, $i = 1, \dots, n$, that is

$$p(\mathbf{y}_i) = \sum_{\mathbf{v}} p(\mathbf{y}_i | \mathbf{V}_i = \mathbf{v}) p(\mathbf{V}_i = \mathbf{v}), \quad (1)$$

where the sum $\sum_{\mathbf{v}}$ is extended to all the possible configurations $\mathbf{v} = (v_i^{(1)}, \dots, v_i^{(T)})'$ of the latent process $\mathbf{V}_i$.

Efficient computation of the probability in (1) may be performed by exploiting a forward recursion available in the hidden Markov literature (see Baum et al., 1970; Levinson et al., 1983; MacDonald and Zucchini, 1997). As in Bartolucci (2006), it is convenient to express this recursion by using the matrix notation on the basis of the initial probability vectors $\boldsymbol{\pi}_i$ and transition probability matrices $\boldsymbol{\Pi}_i^{(t)}$. For this aim, consider the column vector $\mathbf{q}_i^{(t)}(\mathbf{y}_i)$ with elements $p(V_i^{(t)} = v, \mathbf{y}_i)$ for $v = 1, \dots, k$. This vector may be computed as follows

$$\mathbf{q}_i^{(t)}(\mathbf{y}_i) = \begin{cases} \text{diag}[\mathbf{m}_i^{(1)}(\mathbf{y}_i)] \boldsymbol{\pi}_i, & \text{if } t = 1, \\ \text{diag}[\mathbf{m}_i^{(t)}(\mathbf{y}_i)] [\boldsymbol{\Pi}_i^{(t)}]' \mathbf{q}_i^{(t-1)}(\mathbf{y}_i), & \text{otherwise,} \end{cases} \quad (2)$$

where $\mathbf{m}_i^{(t)}(\mathbf{y}_i)$ is the column vector with elements $\psi_i^{(t)}(y_i^{(t)} | v)$ for $v = 1, \dots, k$. Once this recursion has been performed for $t = 1, \dots, T$, we may obtain $p(\mathbf{y}_i)$ as the sum of the elements of the vector $\mathbf{q}_i^{(T)}(\mathbf{y}_i)$.

Maximum likelihood estimation is performed by maximizing the log-likelihood $\ell(\boldsymbol{\theta}) = \sum_i \log[p(\mathbf{y}_i)]$, where $\boldsymbol{\theta}$ denotes the vector of the model parameters. We maximize this function by an EM algorithm (Baum et al., 1970; Dempster et al., 1977), which is based on the *complete data log-likelihood* denoted by $\ell^*(\boldsymbol{\theta})$, i.e. the log-likelihood that we could compute if we knew the latent state of each subject at every occasion.

The EM algorithm alternates two steps (E and M) until convergence: the *E*-step computes the conditional expectation of $\ell^*(\boldsymbol{\theta})$ given the observed data and the current value of $\boldsymbol{\theta}$ using recursions similar to the one illustrated above; the *M*-step maximizes this expected value with respect to $\boldsymbol{\theta}$, so that this parameter vector results updated. The latter may require simple iterative algorithms of Newton-Raphson type.

3 The proposed multilevel extension

Let n denote the number of subjects which, according to some criteria, are grouped in H clusters of size $n_1, \dots, n_H$. For each subject i in cluster h , data are available at T occasions and, for each occasion t , we denote by $Y_{hi}^{(t)}$ the corresponding categorical response variable.

3.1 Model assumptions

Our extension assumes the existence of a latent process $\mathbf{U}_h = (U_h^{(1)}, \dots, U_h^{(T)})'$ for each cluster $h, h = 1, \dots, H$, and a latent process $\mathbf{V}_{hi} = (V_{hi}^{(1)}, \dots, V_{hi}^{(T)})'$ for each subject $i, i = 1, \dots, n_h$, in the cluster. Both processes follow a first-order Markov chain with k_1 states at cluster-level and k_2 at individual level. The basic assumption is that, for any h , $\mathbf{V}_{h1}, \dots, \mathbf{V}_{hn_h}$ are conditionally independent given $\mathbf{U}_h$. This implies that the response vectors for two subjects in the same cluster are conditionally independent given $\mathbf{U}_h$, but, in general, they are not marginally independent. Independence holds when these two subjects belong to two different clusters.

The initial probabilities of each cluster-level latent process are denoted by $\lambda_h(u) = p(U_h^{(1)} = u), u = 1, \dots, k_1$, and are collected in the vectors $\boldsymbol{\lambda}_h$, whereas the transition probabilities are denoted by $\lambda_h^{(t)}(u^{(1)}|u^{(0)}) = p(U_h^{(t)} = u^{(1)}|U_h^{(t-1)} = u^{(0)}), t = 2, \dots, T, u^{(0)}, u^{(1)} = 1, \dots, k_1$, and are collected in the matrices $\boldsymbol{\Lambda}_h^{(t)}$. About the individual-level processes, we have to consider the dependence on the cluster-level latent status and then we have the initial probabilities $\pi_{hi}(v|u) = p(V_{hi}^{(1)} = v|U_h^{(1)} = u), u = 1, \dots, k_1, v = 1, \dots, k_2$, collected in the vectors $\boldsymbol{\pi}_{hi}(u)$ and the transition probabilities $\pi_{hi}^{(t)}(v^{(1)}|u, v^{(0)}) = p(V_{hi}^{(t)} = v^{(1)}|U_h^{(t)} = u, V_{hi}^{(t-1)} = v^{(0)}), t = 2, \dots, T, u = 1, \dots, k_1, v^{(0)}, v^{(1)} = 1, \dots, k_2$, collected in the transition matrices $\boldsymbol{\Pi}_{hi}^{(t)}(u)$. Conditioning on $\mathbf{U}_h$ is also necessary when we express the conditional response probabilities given the latent state. Then, we have $\psi_{hi}^{(t)}(y|u, v) = p(Y_{hi}^{(t)} = y|U_h^{(t)} = u, V_{hi}^{(t)} = v), t = 1, \dots, T, u = 1, \dots, k_1, v = 1, \dots, k_2, y = 0, \dots, l - 1$. These probabilities are collected in the matrices $\boldsymbol{\Psi}_{hi}^{(t)}(u)$.

It is important to formulate constraints on the distribution of the latent processes or on the conditional distribution of the response variables given these processes, so that a parsimonious model results.

About the distribution of each cluster-level latent process, the simplest constraint is that $\boldsymbol{\lambda}_h = \boldsymbol{\lambda}$ and $\boldsymbol{\Lambda}_h^{(t)} = \boldsymbol{\Lambda}, h = 1, \dots, H, t = 2, \dots, T$, implying a time-homogenous Markov chain with the same parameters for all clusters. Other constraints may be formulated on the basis of linear or generalized linear parameterizations. We recall that through a linear parametrization we may formulate the constraint that certain transition probabilities are equal to 0. In particular, the constraint $\boldsymbol{\Lambda}_h^{(t)} = \mathbf{I}$, where $\mathbf{I}$ denotes an identity matrix of suitable dimension, implies that the transition between latent states is not allowed and then the cluster effect is time-constant. In a similar way we may constrain the initial and transition probabilities of each individual-level latent process. At this regard, an important choice is to assume that these probabilities, or alternatively the conditional response probabilities, are independent of the corresponding cluster-level latent process. The first constraint, together with the constraint of time-homogeneity, is formulated as $\boldsymbol{\pi}_{hi}(u) = \boldsymbol{\pi}$ and $\boldsymbol{\Pi}_{hi}^{(t)}(u) = \boldsymbol{\Pi}, h = 1, \dots, H, i = 1, \dots, n_h, u = 1, \dots, k_1, t = 2, \dots, T$. On the other hand, when the individual-level initial or transition probabilities are allowed to depend on the corresponding cluster-level latent process, we may use specific vectors

$\boldsymbol{\pi}(u)$ and matrices $\boldsymbol{\Pi}(u)$ for $u = 1, \dots, k_1$. Alternatively, we may adopt the following generalized linear parametrization

$$\begin{aligned} \log \frac{\pi_{hi}(v|u)}{\pi_{hi}(1|u)} &= \alpha_{uv} + [\mathbf{z}_{hi}^{(1)}]' \boldsymbol{\beta}_v, \quad v = 2, \dots, k_2, \\ \log \frac{\pi_{hi}^{(t)}(v^{(1)}|u, v^{(0)})}{\pi_{hi}^{(t)}(v^{(0)}|u, v^{(0)})} &= \alpha_{uv^{(0)}v^{(1)}} + [\mathbf{z}_{hi}^{(t)}]' \boldsymbol{\beta}_{v^{(0)}v^{(1)}}, \quad t = 2, \dots, T, \quad v^{(0)}, v^{(1)} = 1, \dots, k_2, \end{aligned}$$

for $u = 1, \dots, k_1$ and all h and i , where α_{uv} and $\alpha_{uv^{(0)}v^{(1)}}$ are support points corresponding to the different latent states at cluster and individual level and $\mathbf{z}_{hi}^{(t)}$ is a design vector which may include individual covariates.

Finally, about the conditional response probabilities, the constraint $\boldsymbol{\Psi}_{hi}^{(t)}(u) = \boldsymbol{\Psi}$ implies that there is no cluster effect on these probabilities and they do not depend neither on the subject nor on the time occasion. However, there may be an indirect effect through the distribution of the individual-level latent process. Conversely, if belonging to a certain cluster directly affects the response probabilities of the individuals, a generalized linear parametrization may be convenient. In the case of binary responses, for instance, we may assume the parametrization

$$\log \frac{\psi_{hi}^{(t)}(1|u, v)}{\psi_{hi}^{(t)}(0|u, v)} = \gamma_{1u} + \gamma_{2v} + [\mathbf{z}_{hi}^{(t)}]' \boldsymbol{\delta}, \quad t = 1, \dots, T, \quad v = 1, \dots, k_2, \quad (3)$$

for $u = 1, \dots, k_1$ and all h and i , where the support points γ_{1u} and γ_{2v} correspond, respectively, to the cluster-level and to the individual-level latent states.

3.2 Manifest distribution

Let $\mathbf{y}_{hi} = (y_{hi}^{(1)}, \dots, y_{hi}^{(T)})'$ denote the vector of all the responses provided by subject i in cluster h and $\mathbf{y}_h$ that obtained by collecting the responses of all subjects in cluster h , that is $y_{hi}^{(t)}$ for $i = 1, \dots, n_h$ and $t = 1, \dots, T$.

Under the above assumptions, the manifest probability of $\mathbf{y}_h$ is given by

$$p(\mathbf{y}_h) = \sum_{\mathbf{u}} p(\mathbf{U}_h = \mathbf{u}) \prod_{i=1}^{n_h} \sum_{\mathbf{v}} p(\mathbf{y}_{hi} | \mathbf{U}_h = \mathbf{u}, \mathbf{V}_{hi} = \mathbf{v}) p(\mathbf{V}_{hi} = \mathbf{v} | \mathbf{U}_h = \mathbf{u}),$$

where the sum $\sum_{\mathbf{u}}$ is extended to all the possible configurations of the latent process $\mathbf{U}_h$ and $\sum_{\mathbf{v}}$ is extended to all the possible configurations of $\mathbf{V}_{hi}$. For the cases in which computing $p(\mathbf{y}_h)$ is feasible, estimation of the model parameters can be performed by maximizing the log-likelihood $\ell(\boldsymbol{\theta}) = \sum_h \log[p(\mathbf{y}_h)]$. However, computation of $p(\mathbf{y}_h)$ is usually infeasible even if the probability $p(\mathbf{y}_{hi} | \mathbf{U}_h = \mathbf{u})$ is obtained by recursion (2). For this reason, we suggest below a pairwise likelihood based approach.

4 Pairwise likelihood inference

In order to make inference on the model parameters, we exploit the following pairwise log-likelihood

$$p\ell(\boldsymbol{\theta}) = \sum_{h=1}^H \sum_{i=1}^{n_h-1} \sum_{j=i+1}^{n_h} p\ell_{hij}(\boldsymbol{\theta}), \quad p\ell_{hij}(\boldsymbol{\theta}) = \log[p(\mathbf{y}_{hi}, \mathbf{y}_{hj})],$$

which closely recalls the pairwise log-likelihood used by Renard et al. (2004).

Note that, when the dimension of each cluster is two, this function is the exact log-likelihood of the model, since it is based on the manifest probability of the responses provided by all the possible pairs of subjects in the same cluster.

4.1 Computation and maximization of the pairwise likelihood

In order to efficiently compute the probability $p(\mathbf{y}_{hi}, \mathbf{y}_{hj})$ as a function of the parameters in $\boldsymbol{\theta}$, we exploit recursion (2) already used for the LM model illustrated in Section 2. In fact, $p(\mathbf{y}_{hi}, \mathbf{y}_{hj}) = p(\tilde{\mathbf{y}}_{hij}^{(1)}, \dots, \tilde{\mathbf{y}}_{hij}^{(T)})$, where $\tilde{\mathbf{y}}_{hij}^{(t)}$ is a realization of the vector $\tilde{\mathbf{Y}}_{hij}^{(t)} = (Y_{hi}^{(t)}, Y_{hj}^{(t)})'$. It may be simply proved that, for $t = 1, \dots, T$, these vectors follow a bivariate LM model since they are conditionally independent given the latent process $\mathbf{W}_{hij}^{(1)}, \dots, \mathbf{W}_{hij}^{(T)}$, where $\mathbf{W}_{hij}^{(t)} = (U_h^{(t)}, V_{hi}^{(t)}, V_{hj}^{(t)})'$. In particular, this latent process follows a Markov chain with $k = k_1 k_2^2$ states indexed by $\mathbf{w} = (u, v_1, v_2)$, initial probabilities

$$\phi_{hij}(\mathbf{w}) = \lambda_h(u) \pi_{hi}(v_1|u) \pi_{hj}(v_2|u), \quad (4)$$

and, for $t = 2, \dots, T$, transition probabilities

$$\phi_{hij}^{(t)}(\mathbf{w}^{(1)}|\mathbf{w}^{(0)}) = \lambda_h^{(t)}(u^{(1)}|u^{(0)}) \pi_{hi}^{(t)}(v_1^{(1)}|u^{(1)}, v_1^{(0)}) \pi_{hj}^{(t)}(v_2^{(1)}|u^{(1)}, v_2^{(0)}). \quad (5)$$

Moreover, the model assumptions imply that, given $\mathbf{W}_{hi}^{(t)} = \mathbf{w}$, the conditional probability of $\tilde{\mathbf{y}}_{hij}^{(t)}$ is equal to

$$\tau_{hij}^{(t)}(\tilde{\mathbf{y}}_{hij}^{(t)}|\mathbf{w}) = \psi_{hi}^{(t)}(y_{hi}^{(t)}|v_1) \psi_{hj}^{(t)}(y_{hj}^{(t)}|v_2). \quad (6)$$

In order to compute $p(\mathbf{y}_{hi}, \mathbf{y}_{hj})$, recursion (2) is applied with $\mathbf{m}_i(\mathbf{y}_i)$ substituted by the vector having elements $\tau_{hij}^{(t)}(\tilde{\mathbf{y}}_{hij}^{(t)}|\mathbf{w})$ for all $\mathbf{w}$. Similarly, $\boldsymbol{\pi}_i$ must be substituted by the initial probability vector with elements $\phi_{hij}(\mathbf{w})$ and $\boldsymbol{\Pi}_i^{(t)}$ by the transition matrix with elements $\phi_{hij}^{(t)}(\mathbf{w}^{(1)}|\mathbf{w}^{(0)})$.

The pairwise log-likelihood $p\ell(\boldsymbol{\theta})$ can be maximized by an EM algorithm having a structure that closely recalls that outlined in Section 2.2. In this case, in particular, the *complete data pairwise log-likelihood* is

$$p\ell^*(\boldsymbol{\theta}) = \sum_{h=1}^H \sum_{i=1}^{n_h-1} \sum_{j=i+1}^{n_h} p\ell_{hij}^*(\boldsymbol{\theta}),$$

where

$$\begin{aligned} p\ell_{hij}^*(\boldsymbol{\theta}) = & \sum_{\mathbf{w}} d_{hij}^{(1)}(\mathbf{w}) \log[\phi_{hij}(\mathbf{w})] + \\ & + \sum_{t>1} \sum_{\mathbf{w}^{(0)}} \sum_{\mathbf{w}^{(1)}} d_{hij}^{(t)}(\mathbf{w}^{(0)}, \mathbf{w}^{(1)}) \log[\phi_{hij}^{(t)}(\mathbf{w}^{(1)} | \mathbf{w}^{(0)})] + \sum_t \sum_{\mathbf{w}} d_{hij}^{(t)}(\mathbf{w}) \log[\tau_{hij}^{(t)}(\tilde{\mathbf{y}}_{hij}^{(t)} | \mathbf{w})]. \end{aligned} \quad (7)$$

In the above expression, $d_{hij}^{(t)}(\mathbf{w})$ is a dummy variable equal to 1 if, at occasion t , cluster h is in latent state u , subject hi is in latent state v_1 , and subject hj is in latent state v_2 ; moreover, we have $d_{hij}^{(t)}(\mathbf{w}^{(0)}, \mathbf{w}^{(1)}) = d_{hij}^{(t-1)}(\mathbf{w}^{(0)}) d_{hij}^{(t)}(\mathbf{w}^{(1)})$.

The complete data pairwise log-likelihood may be simply expressed in terms of the parameters of the proposed multilevel model by substituting (4), (5), and (6) in the above expression. For instance, the first component becomes the sum over u of

$$\tilde{d}_{hij}^{(1)}(u) \log[\lambda_h(u)] + \sum_{v_1} \tilde{d}_{hij}^{(11)}(u, v_1) \log[\pi_{hi}(v_1 | u)] + \sum_{v_2} \tilde{d}_{hij}^{(12)}(u, v_2) \log[\pi_{hj}(v_2 | u)], \quad (8)$$

where the variables $\tilde{d}_{hij}^{(1)}(u)$, $\tilde{d}_{hij}^{(11)}(u, v_1)$, and $\tilde{d}_{hij}^{(12)}(u, v_2)$ are obtained by summing $d_{hij}^{(1)}(\mathbf{w})$ over suitable configurations of $\mathbf{w}$. In a similar way, we can express the other two components involving the transition and the conditional response probabilities.

At the E-step of the EM algorithm, the conditional expected value of each dummy variable $d_{hij}^{(t)}(\mathbf{w})$ and $d_{hij}^{(t)}(\mathbf{w}^{(0)}, \mathbf{w}^{(1)})$ is computed by using the same recursions exploited in the algorithm of Baum et al. (1970). At the M-step, the model parameters are updated by maximizing the function resulting by substituting the expected values in (7) and exploiting the simplification (8) and similar simplifications.

4.2 Model selection and hypothesis testing

As in Renard et al. (2004), we estimate the variance-covariance matrix of the pairwise likelihood estimator $\hat{\boldsymbol{\theta}}$ by the following sandwich formula

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{J}}^{-1} \hat{\mathbf{K}} \hat{\mathbf{J}}^{-1},$$

where

$$\hat{\mathbf{J}} = - \sum_h \frac{\partial^2 p\ell_h(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}, \quad \hat{\mathbf{K}} = \sum_h \frac{\partial p\ell_h(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \frac{\partial p\ell_h(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'}, \quad p\ell_h(\boldsymbol{\theta}) = \sum_{i=1}^{n_h-1} \sum_{j=i+1}^{n_h} p\ell_{hij}(\boldsymbol{\theta}).$$

We obtain the first derivative of $p\ell_h(\boldsymbol{\theta})$ as a by-product of the EM algorithm. The second derivative, instead, is obtained by a numerical method.

General results on the asymptotic properties of the pairwise likelihood estimator $\hat{\boldsymbol{\theta}}$ can be derived along the lines of classical maximum likelihood estimators. However, the former is expected to be less efficient since it relies on a restricted amount of information (Renard et al., 2004).

In order to deal with model selection, Varin and Vidoni (2005) suggest a criterion named the *composite likelihood information* criterion. According to this criterion, the model to be selected is the one which maximizes the following index

$$\text{CLIC} = p\ell(\hat{\boldsymbol{\theta}}) - \text{tr}(\hat{\mathbf{K}}\hat{\mathbf{J}}^{-1}). \quad (9)$$

We use this criterion to select the number of states k_1 and k_2 of each latent process $\mathbf{U}_h$ at cluster-level and $\mathbf{V}_{hi}$ at individual-level. Moreover, it can be also used for selecting one of the possible parametrizations illustrated in Section 3.

5 Application

We illustrate the proposed approach by an application based on a dataset on individual work histories derived from the administrative archives of the Italian National Institute of Social Security (INPS). We consider a sample of 1,876 employees (both blue-collar and white-collar) from 249 private Italian firms with 1,000 to 10,000 workers. The subjects, continuously working in the same firm and aged between 18 and 60 in 1994, were followed for 6 years, from 1994 to 1999. See Bartolucci and Nigro (2007) for further details.

The binary response variable of interest is *illness* (equal to 1 if the employee received illness benefits in a certain year and to 0 otherwise). We also consider a set of individual covariates: *gender* (dummy equal to 1 for woman), *age* in 1994, *area* (North-West, North-East, Center, South, or Islands), *skill* (dummy equal to 1 for a blue-collar), *income* (total annual compensation in thousands of Euros), and *part-time* (dummy equal to 1 for a part-time employee). Among the covariates we also include the lagged response.

To this dataset, we fitted the model described in Section 3 under the constraint that the transition matrices for both processes are tridiagonal with constant off-diagonal elements. We also assume a logistic regression model as in (3) for the conditional probabilities. Then, the individual-level latent process is expected to capture the propensity (which is not explained by the observed covariates) to get ill of every subject, whereas the cluster-level latent process explains the effect of different firms on the propensity to request the benefits.

The first step of the analysis is the choice of the number of states for the cluster and individual-level latent processes, denoted by k_1 and k_2 respectively. This choice is based on CLIC, which is defined in (9). The value of this index is reported in Table 1 for different values of k_1 and k_2 . According to these results we select the model with $k_1 = 3$ states at cluster-level and $k_2 = 2$ at individual-level.

Table 1 Values of CLIC for different values of k_1 and k_2 (in boldface the largest CLIC value).

k_1	k_2		
	1	2	3
1	-30724	-30300	-29972
2	-30144	-29773	-29779
3	-30018	-29705	-29756
4	-30001	-29727	-29747

Table 2 collects the estimates of the regression parameters obtained with the selected number of states. We note that the probability of receiving illness benefits is positively related to being a blue-collar and to the lagged response, whereas it is negatively related to income and to having a part-time job. The effects of gender, age and age squared are not significant.

Table 2 Estimates of the logistic regression parameters affecting the conditional probabilities.

parameter	estimate	s.e.	t -stat	p -value
intercept	-3.474	1.364	-2.547	0.011
gender	0.161	0.184	0.876	0.382
age	-0.003	0.045	-0.067	0.947
age ² /100	0.038	0.060	0.633	0.527
area: North-East	0.145	0.257	0.564	0.573
area: Center	-0.096	0.284	-0.338	0.735
area: South	-0.427	0.355	-1.203	0.229
area: Islands	-1.046	0.485	-2.157	0.031
skill	2.037	0.423	4.816	0.000
income	-0.200	0.035	-5.714	0.000
part-time	-0.795	0.338	-2.352	0.019
lagged-response	0.600	0.172	3.480	0.000

About the distribution of each cluster and individual-level latent process, the estimates of the initial and transition probabilities are reported in Table 3 and 4. For both processes, we observe that the states are well separated and the second state is the one with the highest initial probability. Moreover, the estimates of the transition matrices show that the cluster-level latent process has a lower persistence than the individual-level latent process.

Finally, we tried to simplify the model selected above by restricting the transition matrix of each latent process to be diagonal, so that transition between latent states is not allowed. In particular, the model in which the transition matrix at cluster-level is diagonal has a slightly larger value of CLIC equal to -29,706. On the other hand, the restriction that the transition matrix at individual-level is diagonal leads to a strong decrease of CLIC, which is equal to -29,757. We then retain the model in which latent transition is allowed both at cluster and at individual level.

Table 3 Support points and initial and transition probabilities of each cluster-level latent process.

latent state	support point	initial probability	transition probabilities		
1	0.000	0.2221	0.9130	0.0870	0.0000
2	0.444	0.7181	0.0870	0.8260	0.0870
3	2.931	0.0598	0.0000	0.0870	0.9130

Table 4 Support points and initial and transition probabilities of each individual-level latent process.

latent state	support point	initial probability	transition probabilities	
1	0.000	0.4122	0.9729	0.0271
2	2.718	0.5878	0.0271	0.9729

6 Future developments

The approach illustrated in this paper represents an initial attempt to apply the pairwise likelihood approach for making inference on a model based on nested latent Markov chains. This model has a complex structure and may be used for an effective analysis of multilevel longitudinal datasets.

The complete development of the proposed approach needs further investigation on the following aspects:

- *Statistical properties of the proposed estimator.* In particular, we plan to implement a deep simulation study of the finite sample properties.
- *Computational developments.* In order to achieve a faster convergence of the EM algorithm, we plan to combine it with a Newton-Raphson algorithm and to find a reasonable method of initialization in order to increase the chance to reach the global maximum.
- *Weighted pairwise likelihood.* The pairwise likelihood we use gives more weight to the clusters having a higher dimension. Then, as suggested by Renard et al. (2004), a weighted version of the pairwise log-likelihood may be more suitable when the clusters are strongly different in terms of dimension.
- *Model selection and hypothesis testing.* We plan to implement a simulation study of the reliability of the procedure to compute standard errors and of the power of the tests based on these standard errors. Moreover, an important aspect is to assess the performance of the composite likelihood information criterion for selecting, in particular, the number of states of the cluster and individual-level latent processes.

Acknowledgements We acknowledge the financial support from MIUR (PRIN 2007). F.Bartolucci acknowledges the financial support from the “Einaudi Institute for Economics and Finance”.

References

- 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.
- Asparouhov, T., Muthén, B. (2008). Multilevel mixture models. In: G.R. Hancock, K.M. Samuelson (Eds.), *Advances in latent variable mixture model*, Information Age Publishing, Charlotte, NC.
- 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., Lupparelli, M. (2007). The multilevel latent Markov model. In: J. del Castillo, A. Espinal, P. Puig (Eds.), *Proceedings of the 22nd International Workshop on Statistical Modelling*.
- Bartolucci, F., Nigro, V. (2007). Maximum likelihood estimation of an extended latent Markov model for clustered binary panel data. *Computational Statistics and Data Analysis* **51**, 3470–3483.
- Bartolucci, F., Pennoni, F., Lupparelli, M. (2008). Likelihood inference for the latent Markov Rasch model. In: C. Huber, N. Limnios, M. Mesbah, N. Nikulin (Eds.), *Mathematical Methods for Survival Analysis, Reliability and Quality of Life*, Wiley.
- Bartolucci, F., Lupparelli, M., 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., Farcomeni, A., Pennoni, F. (2010a). An overview of latent Markov models for longitudinal categorical data. *Technical report*, <http://arxiv.org/abs/10032804>.
- Bartolucci, F., Pennoni, F., Vittadini, G. (2010b). Assessment of school performance through a multilevel latent Markov Rasch model. *Technical report*, [arXiv:0909.4961v1](https://arxiv.org/abs/0909.4961v1).
- Baum, L., Petrie, T., Soules, G., Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41**, 164–171.
- Cox, D.R., Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729–737.
- Dempster, A.P., Laird, N.M., 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.
- Hjort, N.L., Varin, C. (2008). ML, PL, QL in Markov chain models. *Scandinavian Journal of Statistics* **35**, 64–82.
- Langeheine, R., van de Pol, F. (1990). A unifying framework for markov modeling in discrete space and discrete time. *Sociological Methods and Research* **18**, 416–441.
- Levinson, S.E., Rabiner, L.R., 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.
- Lindsay, B. (1988). Composite likelihood methods. In: N. Prabhu (Ed.), *Statistical Inference from Stochastic Process*, American Mathematical Society, Providence.
- MacDonald, I.L., Zucchini, W. (1997). *Hidden Markov and other Models for Discrete-Valued Time Series*. Chapman and Hall, London.
- Renard, D., Molenberghs, G., Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis* **44**, 649–667.
- Varin, C., Czado, C. (2010). A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics* **11**, 127–138.
- Varin, C., Vidoni, P. (2005). A note on the composite likelihood inference and model selection. *Biometrika* **92**, 519–528.
- Wiggins, L. (1973). *Panel Analysis: Latent probability models for attitude and behaviour processes*. Elsevier, Amsterdam.