Nested hidden Markov chains for modeling dynamic unobserved heterogeneity in multilevel longitudinal data

Francesco Bartolucci and Monia Lupparelli

August 10, 2012

Abstract

In the context of multilevel longitudinal data, where sample units are collected in clusters, an important aspect that should be accounted for is the unobserved heterogeneity between sample units and between clusters. For this aim we propose an approach based on nested hidden (latent) Markov chains, which are associated to every sample unit and to every cluster. The approach allows us to account for the mentioned forms of unobserved heterogeneity in a dynamic fashion; it also allows us to account for the correlation which may arise between the responses provided by the units belonging to the same cluster. Given the complexity in computing the manifest distribution of these response variables, we make inference on the proposed model through a composite likelihood function based on all the possible pairs of subjects within every cluster. The proposed approach is illustrated through an application to a dataset concerning a sample of Italian workers in which a binary response variable for the worker receiving an illness benefit was repeatedly observed.

KEYWORDS: composite likelihood, EM algorithm, latent Markov model, pairwise likelihood

1 Introduction

In modeling longitudinal data, it is common to account for the unobserved heterogeneity between sample units, that is, the heterogeneity that cannot be explained on the basis of the observable covariates (Diggle et al., 2002; Hsiao, 2003; Frees, 2004; Fitzmaurice et al., 2009). This is normally accomplished by the introduction of latent variables or random effects. For instance, a typical approach consists of associating a random intercept to every sample unit which affects the distribution of each occasion-specific response in the same fashion. This allows us to account for a form of *time-constant* unobserved heterogeneity which is due to unobservable covariates and related factors.

More recent approaches for longitudinal data are based on allowing for a form of *time-varying* unobserved heterogeneity, relaxing in this way the assumption that the effect of unobservable covariates on the response variables is constant in time. This is sensible in many applied contexts, especially in the presence of long panels and with a limited set of observable covariates. Among these time-varying approaches, it is worth mentioning the one described in Heiss (2008), which is based on random effects having an AR(1) structure, and that proposed by Bartolucci and Farcomeni (2009), which is based on a hidden (latent) Markov chains for capturing the unobserved heterogeneity in a dynamic fashion. For a comparison between the two approaches see Bartolucci et al. (2010a).

The above considerations are obviously pertinent when we deal with multilevel longitudinal data, where sample units are collected in clusters, with the addiction that it is also appropriate modeling the unobserved heterogeneity between clusters and the correlation between the responses provided by the units in the same cluster. Note that multilevel longitudinal data are more and more easily encountered in socio-economic contexts. In particular, the dataset motivating this paper, that will be described in detail in the following, concerns a sample of workers (sample units) in different firms (clusters), who are longitudinally observed. As response we have a binary variable equal to 1 if the employee receives illness benefits in a certain year and to 0 otherwise. Datasets having a similar structure are nowadays available, for instance, in educational contexts, where students are collected in classes and are followed for a certain number of years of schooling. In

these datasets we typically have a limited set of observable covariates and the need arises for an appropriate modeling of the unobserved heterogeneity between both sample units and clusters.

For the aim described above, we propose an approach based on nested hidden Markov chains which may be seen as an extension of the approach proposed by Bartolucci and Farcomeni (2009) for longitudinal data. In particular, we associate a first-order homogeneous hidden Markov chain to every sample unit and to every cluster. The time-specific realizations of these two chains go to affect the distribution of the response variables together with the covariates observed at unit and cluster levels. Coming back to the above example about the sample of employees, the different states of the unit-level Markov chain correspond to different levels of the residual (not explained by the unit-level observable covariates) tendency to require an illness benefit by an employee. A similar interpretation may be found for the different states of the cluster-level Markov chain, which affect the behavior of the employees in the same firm. Moreover, the possibility that the unit-level state changes may be due to events of the employee's life that are not recorded in the dataset, such as a sudden worsening of his/her health status. Similarly, a change in the cluster-level state may be due to events about the firm, such as the change of the management. In any case, we can test if the latent effects are indeed dynamic or not on the basis of the dataset at hand.

The proposed approach may be cast in the literature about latent Markov (LM) models for longitudinal data, as described by Bartolucci et al. (2010b). It is worth noting that other multilevel extensions of the latent (or hidden) Markov approach for longitudinal data are available in the literature. We mention, in particular, the extensions proposed by Bartolucci et al. (2009) and Bartolucci et al. (2011). About multilevel extensions see also Asparouhov and Muthén (2008) and about related models including random effects, but not in a context of analysis of multilevel data, see van de Pol and Langeheine (1990), Altman (2007), and Maruotti (2011). In these cases the effects (fixed or random) associated to every cluster are time-constant. However, an extension in which these effects are time-varying has not been proposed yet, at least to our knowledge.

Under the proposed model, the manifest distribution of the response variables is com-

putationally intractable 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 using the same recursion exploited by Baum et al. (1970) to deal with hidden Markov models and how to maximize this likelihood by an Expectation-Maximization (EM) algorithm similar to the one they suggest and implemented along the same lines as in Bartolucci and Farcomeni (2009). 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 (CLIC) developed by Varin and Vidoni (2005). An `R` implementation of the functions used for the estimation of the model in the presence of binary response variables is available to the reader upon request.

The paper is organized as follows. In the next section we briefly review the LM model with covariates (Bartolucci and Farcomeni, 2009) and its maximum likelihood estimation. Section 3 illustrates the proposed multilevel extension dealing with the case of continuous and binary response variables. Pairwise likelihood inference for this model is described in Section 4. In Section 5 we illustrate the model by an application based on the dataset concerning the sample of workers mentioned above. Finally, in Section 6 we draw the main conclusions.

2   Using hidden Markov chains for modeling unob-served heterogeneity

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, $i = 1, \ldots, n$, $t = 1, \ldots, T$, and let $\boldsymbol{Z}_i^{(t)}$ be the corresponding column vector of covariates, which may also include the lagged responses. In the context of our application, the response variables are binary, although the LM

model may be also applied to variables having a different nature.

In the following, we outline how to model these data accounting for unobserved heterogeneity in a dynamic fashion, by introducing a hidden Markov chain, as suggested by Bartolucci and Farcomeni (2009).

2.1 Model assumptions

We assume that, for $i = 1, \ldots, n$, the response variables $Y_i^{(1)}, \ldots, Y_i^{(T)}$ are conditionally independent given the covariate vectors $\boldsymbol{Z}_i^{(1)}, \ldots, \boldsymbol{Z}_i^{(T)}$ and a latent process $\boldsymbol{V}_i = (V_i^{(1)}, \ldots, V_i^{(T)})$, which follows a first-order homogeneous Markov chain and is independent of the covariates.

This chain has k states, labeled from 1 to k, with initial and transition probabilities

$$
\begin{aligned}
\pi_v &= p(V_i^{(1)} = v), \quad v = 1, \ldots, k, \\
\pi_{v|\bar{v}} &= p(V_i^{(t)} = v | V_i^{(t-1)} = \bar{v}), \quad t = 2, \ldots, T, \ \bar{v}, v = 1, \ldots, k.
\end{aligned}
$$

Note that, in the above definitions, v refers to the current state, whereas $\bar{v}$ refers to the previous one. This convention will be used throughout the paper. Moreover, the initial probabilities are collected in the k-dimensional column vectors $\boldsymbol{\pi}$, whereas the transition probabilities are collected in the $k \times k$ transition matrix $\boldsymbol{\Pi}$. Note that these probabilities are the same for all sample units and, in particular, the transition probabilities are time homogenous. Moreover, in order to make the model more parsimonious, different constraints may be imposed on the matrix $\boldsymbol{\Pi}$; see also Bartolucci (2006). For instance, we may assume that this matrix is tridiagonal, with constant off-diagonal elements, so that with $k = 3$ we have

$$
\boldsymbol{\Pi} = \begin{pmatrix} 1 - \rho & \rho & 0 \\ \rho & 1 - 2\rho & \rho \\ 0 & \rho & 1 - \rho \end{pmatrix}, \tag{1}
$$

where ρ is a parameter between 0 and 0.5 to be estimated.

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 assumed *measurement model*. For instance, in the case of continuous response variables, it is natural to formulate the following

assumption on the conditional distribution of $Y_i^{(t)}$ given $V_i^{(t)}$ and $\boldsymbol{Z}_i^{(t)}$:

$$Y_i^{(t)}|V_i^{(t)} = v, \boldsymbol{Z}_i^{(t)} = \boldsymbol{z} \sim N(\beta_v + \boldsymbol{z}'\boldsymbol{\delta}, \sigma^2),$$

where β_v is an intercept related to the latent state and $\boldsymbol{\delta}$ is a vector of regression coefficients. Obviously, these parameters, including the variance σ^2, can be estimated together with the above initial and transition probabilities.

With binary response variables, instead, it is natural to assume that

$$Y_i^{(t)}|V_i^{(t)} = v, \boldsymbol{Z}_i^{(t)} = \boldsymbol{z} \sim Bern(\psi_i^{(t)}(v, \boldsymbol{z})),$$

where

$$\log \frac{\psi_i^{(t)}(v, \boldsymbol{z})}{1 - \psi_i^{(t)}(v, \boldsymbol{z})} = \beta_v + \boldsymbol{z}'\boldsymbol{\delta},$$

with $\psi_{hi}^{(t)}(v, \boldsymbol{z})$ corresponding to the conditional "probability of success", that is $\psi_{hi}^{(t)}(v, \boldsymbol{z}) = p(Y_i^{(t)} = 1|V_i^{(t)} = v, \boldsymbol{Z}_i^{(t)} = \boldsymbol{z})$.

The above approach may be extended to response variables having a different nature, even ordinal variables, and also to multivariate contexts, in which we observe more response variables at each time occasions. We refer the reader to Bartolucci and Farcomeni (2009) for details on the resulting LM model.

2.2 Maximum likelihood estimation

When we deal with an observed sample, for $i = 1, \ldots, n$ we have an observed response configuration $\boldsymbol{y}_i = (y_i^{(1)}, \ldots, y_i^{(T)})$ and an observed sequence of covariates vectors $\boldsymbol{z}_1^{(1)}, \ldots, \boldsymbol{z}_i^{(T)}$; we collect these covariates in the unique vector $\boldsymbol{z}_i$ (for all time occasion). In order to perform maximum likelihood estimation of the above model on the basis of these data, the need arises of computing the *manifest distribution* of $\boldsymbol{y}_i$ given $\boldsymbol{z}_i$, that is,

$$p(\boldsymbol{y}_i|\boldsymbol{z}_i) = \sum_{\boldsymbol{v}} p(\boldsymbol{y}_i|\boldsymbol{V}_i = \boldsymbol{v}, \boldsymbol{z}_i)p(\boldsymbol{V}_i = \boldsymbol{v}), \qquad (2)$$

where the sum $\sum_{\boldsymbol{v}}$ is over all the possible configurations $\boldsymbol{v} = (v_i^{(1)}, \ldots, v_i^{(T)})$ of the latent process $\boldsymbol{V}_i$.

Efficient computation of the probability in (2) 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) and Bartolucci and Farcomeni (2009), it is convenient to express this recursion by using the matrix notation on the basis of the initial probability vectors $\boldsymbol{\pi}$ and transition matrix $\boldsymbol{\Pi}$. For this aim, consider the column vector $\boldsymbol{q}_i^{(t)}$ with elements

$$p(y_i^{(1)}, \ldots, y_i^{(t)}, V_i^{(t)} = v, \boldsymbol{z}_i^{(1)}, \ldots, \boldsymbol{z}_i^{(t)}), \quad v = 1, \ldots, k.$$

This vector may be recursively computed as follows:

$$\boldsymbol{q}_i^{(t)} = \begin{cases} \operatorname{diag}(\boldsymbol{m}_i^{(1)})\boldsymbol{\pi}, & \text{if } t = 1, \\ \operatorname{diag}(\boldsymbol{m}_i^{(t)})\boldsymbol{\Pi}'\boldsymbol{q}_i^{(t-1)}, & \text{otherwise,} \end{cases} \tag{3}$$

where $\boldsymbol{m}_i^{(t)}$ is the column vector with elements $p(y_i^{(t)}|V_i^{(t)} = v, \boldsymbol{z}_i^{(t)})$, for $v = 1, \ldots, k$, which is defined on the basis of the assumed measurement model. Once this recursion has been performed for $t = 1, \ldots, T$, we may obtain $p(\boldsymbol{y}_i)$ as the sum of the elements of the vector $\boldsymbol{q}_i^{(T)}$.

Maximum likelihood estimation is performed by maximizing the log-likelihood $\ell(\boldsymbol{\theta}) = \sum_i \log[p(\boldsymbol{y}_i|\boldsymbol{z}_i)]$, where $\boldsymbol{\theta}$ denotes the vector of all 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 $\ell^*(\boldsymbol{\theta})$, that is, 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. A detailed description of this EM algorithm is available in Bartolucci and Farcomeni (2009).

3   Proposed multilevel extension

In the context of multilevel longitudinal data, the n sample units are grouped, according to some criteria, in H clusters of size $n_1, \ldots, n_H$. Then, for each subject i in cluster h, data are available at T consecutive occasions. In particular, we denote by $Y_{hi}^{(t)}$ the

corresponding response variable and by $\boldsymbol{Z}_{hi}^{(t)}$ the corresponding column vector of covariates, where $h = 1, \ldots, H$, $i = 1, \ldots, n_h$, and $t = 1, \ldots, T$. Moreover, by $\boldsymbol{X}_h^{(t)}$, with $h = 1, \ldots, H$ and $t = 1, \ldots, T$, we denote column vectors of cluster-level covariates, which may be time-varying.

In the following we show how multilevel longitudinal data, having the structure described above, may be analyzed by an extension of the approach outlined in Section 2.

3.1 Model assumptions

Our extension assumes the existence of a latent process $\boldsymbol{U}_h = (U_h^{(1)}, \ldots, U_h^{(T)})$ for each cluster h, $h = 1, \ldots, H$, and a latent process $\boldsymbol{V}_{hi} = (V_{hi}^{(1)}, \ldots, V_{hi}^{(T)})$ for each subject i, $i = 1, \ldots, n_h$, in the cluster. Both processes follow a first-order homogeneous Markov chain with k_1 states at cluster level and k_2 at individual level. These processes are assumed to be independent each other and also independent of the unit- and cluster-level covariates. Moreover, extending the assumptions formulated in Section 2, we impose that, for every sample unit hi (unit i in cluster h), the response variables $Y_{hi}^{(t)}$ are conditionally independent given $\boldsymbol{U}_h$, $\boldsymbol{V}_{hi}$ and the corresponding covariates. This implies that the response vectors for two subjects in the same cluster are conditionally independent given $\boldsymbol{U}_h$, but they are not marginally independent. This marginal independence holds for subjects belonging to two different clusters.

The initial and the transition probabilities of each cluster-level latent process are denoted by

$$\begin{aligned} \lambda_u &= p(U_h^{(1)} = u), \quad u = 1, \ldots, k_1 \\ \lambda_{u|\bar{u}} &= p(U_h^{(t)} = u | U_h^{(t-1)} = \bar{u}), \quad t = 2, \ldots, T, \; \bar{u}, u = 1, \ldots, k_1, \end{aligned}$$

and are collected in the vector $\boldsymbol{\lambda}$ and in the transition matrix $\boldsymbol{\Lambda}$. Moreover, for the unit-level latent processes $\boldsymbol{U}_{hi}$, we substantially adopt the same notation as in Section 2, and then we let $\pi_v = p(V_{hi}^{(1)} = v)$ and $\pi_{v|\bar{v}} = p(V_{hi}^{(t)} = v | V_{hi}^{(t)} = \bar{v})$; these initial and transition probabilities are still collected in the vector $\boldsymbol{\pi}$ and in the matrix $\boldsymbol{\Pi}$, respectively.

Finally, about the conditional response probabilities, the same considerations expressed in Section 2 still holds. Then, in the case of continuous response variables we

may assume that:

$$Y_{hi}^{(t)}|U_h^{(t)} = u, V_{hi}^{(t)} = v, \boldsymbol{X}_h^{(t)} = \boldsymbol{x}, \boldsymbol{Z}_{hi}^{(t)} = \boldsymbol{z} \sim N(\alpha_u + \beta_v + \boldsymbol{x}'\boldsymbol{\gamma} + \boldsymbol{z}'\boldsymbol{\delta}, \sigma^2),$$

where α_u is an intercept related to the cluster-level latent state, β_v is an intercepts related to the unit-level latent state, and $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are corresponding vectors of regression coefficients.

With binary response variables, instead, it is natural to assume that

$$Y_{hi}^{(t)}|U_h^{(t)} = u, V_{hi}^{(t)} = v, \boldsymbol{X}_h^{(t)} = \boldsymbol{x}, \boldsymbol{Z}_{hi}^{(t)} = \boldsymbol{z} \sim Bern(\psi_{hi}^{(t)}(u, v, \boldsymbol{x}, \boldsymbol{z})), \tag{4}$$

where

$$\log \frac{\psi_{hi}^{(t)}(u, v, \boldsymbol{x}, \boldsymbol{z})}{1 - \psi_{hi}^{(t)}(u, v, \boldsymbol{x}, \boldsymbol{z})} = \alpha_u + \beta_v + \boldsymbol{x}'\boldsymbol{\gamma} + \boldsymbol{z}'\boldsymbol{\delta},$$

with parameters having the same interpretation as above.

3.2 Manifest distribution

When we observe a set of multilevel longitudinal data, we have a sequence of response $\boldsymbol{y}_{hi} = (y_{hi}^{(1)}, \ldots, y_{hi}^{(T)})$ for every sample unit hi, with $h = 1, \ldots, H$, $i = 1, \ldots, n_h$. We denote by $\boldsymbol{y}_h$ the vector obtained by collecting the responses of all subjects in cluster h, that is $y_{hi}^{(t)}$ for $i = 1, \ldots, n_h$ and $t = 1, \ldots, T$. Similarly, we observe the vectors of unit-level covariates $\boldsymbol{z}_{hi}^{(t)}, \ldots, \boldsymbol{z}_{hi}^{(T)}$; these covariates are collected in the unique vector $\boldsymbol{z}_{hi}$ when referred to the unit hi (for all time occasions) and in the vector $\boldsymbol{z}_h$ when referred to all units in the same cluster h. Finally, for every cluster h, we observe the vectors of cluster-level covariates $\boldsymbol{x}_h^{(t)}$, which are collected in the unique vector $\boldsymbol{x}_h$ (for all time occasions).

Under the above assumptions, the manifest probability of $\boldsymbol{y}_h$ given $\boldsymbol{x}_h$ and $\boldsymbol{z}_h$ has the following expression:

$$\begin{aligned} p(\boldsymbol{y}_h|\boldsymbol{x}_h, \boldsymbol{z}_h) &= \sum_{\boldsymbol{u}} p(\boldsymbol{U}_h = \boldsymbol{u}) \\ &\times \prod_{i=1}^{n_h} \left[\sum_{\boldsymbol{v}} p(\boldsymbol{y}_{hi}|\boldsymbol{U}_h = \boldsymbol{u}, \boldsymbol{V}_{hi} = \boldsymbol{v}, \boldsymbol{x}_h, \boldsymbol{z}_h) p(\boldsymbol{V}_{hi} = \boldsymbol{v}) \right], \end{aligned}$$

where the sum $\sum_{\boldsymbol{u}}$ is over all the possible configurations of the latent process $\boldsymbol{U}_h$ and $\sum_{\boldsymbol{v}}$ is over all the possible configurations of $\boldsymbol{V}_{hi}$.

9

For the cases in which computing $p(\boldsymbol{y}_h|\boldsymbol{x}_h, \boldsymbol{z}_h)$ is feasible, estimation of the model parameters can be performed by maximizing the log-likelihood $\ell(\boldsymbol{\theta}) = \sum_h \log[p(\boldsymbol{y}_h|\boldsymbol{x}_h, \boldsymbol{z}_h)]$. However, computation of $p(\boldsymbol{y}_h|\boldsymbol{x}_h, \boldsymbol{z}_{hi})$ is usually infeasible even if the conditional probability $p(\boldsymbol{y}_{hi}|\boldsymbol{U}_h = \boldsymbol{u}, \boldsymbol{V}_{hi} = \boldsymbol{v}, \boldsymbol{x}_h, \boldsymbol{z}_h)$ is obtained by recursion (3). 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:

$$
\begin{aligned}
p\ell(\boldsymbol{\theta}) &= \sum_{h=1}^{H} \sum_{i=1}^{n_h-1} \sum_{j=i+1}^{n_h} p\ell_{hij}(\boldsymbol{\theta}), \\
p\ell_{hij}(\boldsymbol{\theta}) &= \log[p(\boldsymbol{y}_{hi}, \boldsymbol{y}_{hj}|\boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj})],
\end{aligned}
$$

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

Note that, when the dimension of each cluster is two ($n_h = 2$, $h = 1, \ldots, H$), 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 $p(\boldsymbol{y}_{hi}, \boldsymbol{y}_{hj}|\boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj})$ as a function of the parameters in $\boldsymbol{\theta}$, we exploit recursion (3) already used for the model illustrated in Section 2. In fact, we have that

$$
p(\boldsymbol{y}_{hi}, \boldsymbol{y}_{hj}) = p(\tilde{\boldsymbol{y}}_{hij}^{(1)}, \ldots, \tilde{\boldsymbol{y}}_{hij}^{(T)}|\boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj}),
$$

where $\tilde{\boldsymbol{y}}_{hij}^{(t)}$ is a realization of the vector $\tilde{\boldsymbol{Y}}_{hij}^{(t)} = (Y_{hi}^{(t)}, Y_{hj}^{(t)})'$. It may be simply proved that, for $t = 1, \ldots, T$, these vectors follow a bivariate LM model with covariates since they are conditionally independent given the latent process $\boldsymbol{W}_{hij}^{(1)}, \ldots, \boldsymbol{W}_{hij}^{(T)}$, where $\boldsymbol{W}_{hij}^{(t)} = (U_h^{(t)}, V_{hi}^{(t)}, V_{hj}^{(t)})$, and the corresponding covariates. In particular, this latent process follows a Markov chain with an augmented space of $k = k_1 k_2^2$ states indexed by $\boldsymbol{w} = $

10

(u, v_1, v_2). It is simple to see that the initial probability of state $\boldsymbol{w}$ is

$$\phi_{\boldsymbol{w}} = p(\boldsymbol{W}_{hij}^{(1)} = \boldsymbol{w}) = \lambda_u \pi_{v_1} \pi_{v_2}, \tag{5}$$

whereas, for $t = 2, \ldots, T$, transition probability from state $\bar{\boldsymbol{w}} = (\bar{u}, \bar{v}_1, \bar{v}_2)$ to $\boldsymbol{w}$ is

$$\phi_{\boldsymbol{w}|\bar{\boldsymbol{w}}} = p(\boldsymbol{W}_{hij}^{(t)} = \boldsymbol{w} | \boldsymbol{W}_{hij}^{(t-1)} = \bar{\boldsymbol{w}}) = \lambda_{u|\bar{u}} \pi_{v_1|\bar{v}_1} \pi_{v_2|\bar{v}_2}. \tag{6}$$

Moreover, in the case of discrete or categorical response variables, the model assumptions imply that, given $\boldsymbol{W}_{hi}^{(t)} = \boldsymbol{w}$, the conditional probability of $\tilde{\boldsymbol{y}}_{hij}^{(t)}$ is equal to

$$p(\tilde{\boldsymbol{y}}_{hij}^{(t)} | \boldsymbol{W}_{hi}^{(t)} = \boldsymbol{w}, \boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj}) = p(y_{hi}^{(t)} | u, v_1, \boldsymbol{x}_h^{(t)}, \boldsymbol{z}_{hi}^{(t)}) p(y_{hj}^{(t)} | u, v_2, \boldsymbol{x}_h^{(t)}, \boldsymbol{z}_{hj}^{(t)}). \tag{7}$$

A similar expression holds for continuous response variables, based on the corresponding density functions.

In order to compute $p(\boldsymbol{y}_{hi}, \boldsymbol{y}_{hj} | \boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj})$, recursion (3) is applied with $\boldsymbol{m}_i$ substituted by the vector $\tilde{\boldsymbol{m}}_{hij}$ having elements $p(\tilde{\boldsymbol{y}}_{hij}^{(t)} | \boldsymbol{W}_{hi}^{(t)} = \boldsymbol{w}, \boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj})$ for all $\boldsymbol{w}$. Similarly, $\boldsymbol{\pi}$ must be substituted by the initial probability vector $\boldsymbol{\phi}$ with elements $\phi_{\boldsymbol{w}}$ and $\boldsymbol{\Pi}$ by the transition matrix $\boldsymbol{\Phi}$ with elements $\phi_{\boldsymbol{w}|\bar{\boldsymbol{w}}}$.

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

$$
\begin{aligned}
p\ell_{hij}^*(\boldsymbol{\theta}) &= \sum_{\boldsymbol{w}} d_{hij}^{(1)}(\boldsymbol{w}) \log(\phi_{\boldsymbol{w}}) \\
&+ \sum_{t>1} \sum_{\bar{\boldsymbol{w}}} \sum_{\boldsymbol{w}} d_{hij}^{(t)}(\bar{\boldsymbol{w}}, \boldsymbol{w}) \log(\phi_{\boldsymbol{w}|\bar{\boldsymbol{w}}}) \\
&+ \sum_{t} \sum_{\boldsymbol{w}} d_{hij}^{(t)}(\boldsymbol{w}) \log[p(\tilde{\boldsymbol{y}}_{hij}^{(t)} | \boldsymbol{W}_{hi}^{(t)} = \boldsymbol{w}, \boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj})].
\end{aligned} \tag{8}
$$

In the above expression, $d_{hij}^{(t)}(\boldsymbol{w})$ 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 $d_{hij}^{(t)}(\bar{\boldsymbol{w}}, \boldsymbol{w}) = d_{hij}^{(t-1)}(\bar{\boldsymbol{w}}) d_{hij}^{(t)}(\boldsymbol{w})$.

11

The complete data pairwise log-likelihood may be simply expressed in terms of the parameters of the proposed multilevel model by substituting (5), (6), and (7) in the above expression. For instance, the first component becomes the sum over u of

$$\tilde{d}_{hij}^{(1)}(u)\log[\lambda_h(u)] + \sum_{v_1} \tilde{d}_{hij}^{(11)}(u,v_1)\log[\pi_{hi}(v_1|u)] + \sum_{v_2} \tilde{d}_{hij}^{(12)}(u,v_2)\log[\pi_{hj}(v_2|u)], \quad (9)$$

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 $d_{hij}^{(1)}(\boldsymbol{w})$ over suitable configurations of $\boldsymbol{w}$. In a similar way we can express the other two components involving the transition and the conditional response probabilities (or densities).

At the E-step of the EM algorithm, the conditional expected value of each dummy variable $d_{hij}^{(t)}(\boldsymbol{w})$ and $d_{hij}^{(t)}(\bar{\boldsymbol{w}},\boldsymbol{w})$ 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 (8) and exploiting the simplification (9) and similar simplifications. In any case, the final algorithm is implemented along the same lines as the algorithm implemented by Bartolucci and Farcomeni (2009). We make our R implementation available to the reader upon request.

4.2 Model selection and hypothesis testing

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

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

where

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

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).

12

In order to deal with model selection, Varin and Vidoni (2005) suggested CLIC. According to this criterion, the model to be selected is the one which maximizes the following index

$$CLIC = p\ell(\hat{\boldsymbol{\theta}}) - \mathrm{tr}(\hat{\boldsymbol{K}}\hat{\boldsymbol{J}}^{-1}). \tag{10}$$

We use this criterion to select the number of states k_1 and k_2 of each latent process $\boldsymbol{U}_h$ at cluster level and $\boldsymbol{V}_{hi}$ at unit 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-collars and white-collars) 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.

As already mentioned in Section 1, 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 unit- and cluster-level covariates: *gender* (dummy equal to 1 for woman), *age* in 1994, *area* (Noth-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; see equation (1). We also assume a logistic regression model as in (4) for the conditional probabilities. Then, the unit-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 require illness benefits.

The first step of the analysis is the choice of the number of states for the cluster- and unit-level latent processes, denoted by k_1 and k_2 respectively. This choice is based on CLIC, which is based on the index defined in (10). 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 unit level.

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

	k_2		
k_1	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.

About the distribution of each cluster and unit-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 unit-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 lower value of CLIC equal to -29,706. On the other hand, the

Table 2: *Estimates of the logistic regression parameters (collected in the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$) 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
$\text{age}^2/100$	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

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

latent state (u)	support point (α_u)	initial probability (λ_u)	transition probabilities ($\lambda_{u|\bar{u}}$)		
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 unit-level latent process.*

latent state (v)	support point (β_v)	initial probability (π_v)	transition probabilities ($\pi_{v|\bar{v}}$)	
1	0.000	0.4122	0.9729	0.0271
2	2.718	0.5878	0.0271	0.9729

restriction that the transition matrix at unit-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 unit levels.

6 Conclusions

With reference to multilevel longitudinal data, where sample units are collected in clusters, in this paper we propose an approach to account for the unobserved heterogeneity between sample units and between clusters in a dynamic fashion. The approach is based on associating a hidden (or latent) Markov chain to every sample unit and to every cluster. These Markov chains are assumed to be homogeneous and of the first-order, with transition probabilities that may be subjected to suitable constraints. The approach then extends the one proposed by Bartolucci and Farcomeni (2009), who proposed a latent Markov model with covariates for longitudinal data (not having a multilevel structure).

The complexity of the model formulated on the basis of the proposed approach does not allow us to make exact likelihood inference on its parameters. Therefore, we adopt a composite likelihood framework for making inference, which is based on considering all the possible pairs of units in every cluster, as suggested by Renard et al. (2004) in a simpler context. Within this framework, we also deal with model selection, based on the composite likelihood information criterion (Varin and Vidoni, 2005), and hypothesis testing. In an application based on data about a sample of Italian workers who are employed in different firms, we observed that this composite likelihood approach gives sensible estimates. In this application the response variable is binary, but the approach is

completely general in terms of type of response variable, which may be also continuous, discrete, or ordinal.

Possible further developments of the proposed approach may concern the implementation of faster algorithms for the maximization of the pairwise likelihood that we use. In fact, we maximize this function by an Expectation-Maximization (EM) algorithm which is implemented along the same lines as in Bartolucci and Farcomeni (2009). However, we think that this maximization may be made much faster by using, after a certain number of EM iterations, a Newton-Raphson algorithm. The implementation of this algorithm is made possible by the availability of the score and the observed information matrix (for the pairwise likelihood function), that we already are able to compute within the present approach.

Finally, another point that deserves attention is the use of alternative forms of composite likelihood for parameter estimation. In particular, in the current form, the adopted pairwise likelihood gives more weight to the data referred to the units belonging 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. Note, however, that in our application the clusters are not very different in terms of dimension and so, at least in the present case, we do not expect to obtain very different results on the basis of a weighted composite likelihood function.

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. and Muthén, B. (2008). Multilevel mixture models. In Hancock, G. R. and Samuelson, K. M., editors, *Advances in latent variable mixture models*. Charlotte, NC: Information Age Publishing.

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., Bacci, S., and Pennoni, F. (2010a). Mixture latent autoregressive models for longitudinal data. Technical Report 1108.1498v1, arXiv.

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., Farcomeni, A., and Pennoni, F. (2010b). An overview of latent markov models for longitudinal categorical data. *Statistical Science*, submitted.

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

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.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91:729–737.

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.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data.* Oxford University Press, New York.

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

Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences.* Cambridge University Press, Cambridge.

Heiss, F. (2008). Sequential numerical integration in nonlinear state space models for microeconometric panel data. *Journal of Applied Econometrics*, 23:373–389.

Hjort, N. L. and Varin, C. (2008). Ml, pl, ql in markov chain models. *Scandinavian Journal of Statistics*, 35:64–82.

Hsiao, C. (2003). *Analysis of Panel Data.* Cambridge University Press, New York.

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.

Lindsay, B. (1988). Composite likelihood methods. In Prabhu, N., editor, *Statistical Inference from Stochastic Process*, pages 221–239, Providence. American Mathematical Society.

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

Maruotti, A. (2011). Mixed hidden markov models for longitudinal data: An overview. *International Statistical Review*, 79:427–454.

Renard, D., Molenberghs, G., and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, 44:649–667.

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

Varin, C. and Czado, C. (2010). A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics*, 11:127–138.

Varin, C. and Vidoni, P. (2005). A note on the composite likelihood inference and model selection. *Biometrika*, 92:519–528.