Post-print version of the paper published in the Journal of the American Statistical Association;
https://doi.org/10.1080/01621459.2014.998935

Pairwise likelihood inference for nested hidden Markov chain models for multilevel longitudinal data

Francesco Bartolucci
University of Perugia (IT)
bart@stat.unipg.it

Monia Lupparelli
University of Bologna (IT)
monia.lupparelli@unibo.it

November 29, 2014

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 previously 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. Under the assumed model, computing the manifest distribution of these response variables is infeasible even with few units per cluster. Therefore, we make inference on this model through a composite likelihood function based on all the possible pairs of subjects within each cluster. Properties of the composite likelihood estimator are assessed by simulation. 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: CL-BIC, composite likelihood, EM algorithm, latent Markov model, random effects, unobserved heterogeneity.

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 by means of 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.

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 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 chain for capturing the unobserved heterogeneity in a dynamic fashion. For a comparison between the two approaches see Bartolucci et al. (2014).

The above considerations are obviously pertinent when we deal with multilevel longitudinal data, where sample units are collected in clusters, with the addition 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 article concerns a sample of workers (sample units) in different firms (clusters), who are longitudinally observed for 10 years. Each response is binary and it is equal to 1 if the employee receives illness benefits in a certain year and to 0 otherwise. In many applications, as the present one, long panel datasets have a limited set of observable covariates and this also motivates the need of approaches for modeling unobserved heterogeneity between both sample units and clusters.

Standard multilevel techniques for clustered data have been developed, among others,

by Goldstein (2011). These techniques have been also extended for modeling longitudinal clustered data in several ways within the wide class of mixture models; see among others Muthén and Asparouhov (2009). Algorithms for the implementation and the maximum likelihood estimation of these models are developed in the Mplus software (Muthén and Asparouhov, 2011). Furthermore, multilevel data have been modeled using latent class analysis (Vermunt, 2003), which can be implemented and fitted using the Latent GOLD software (Vermunt and Magidson, 2005).

We propose an approach for multilevel longitudinal data 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 affect the distribution of the response variables together with the covariates observed at unit and cluster level. With reference to the mentioned application, 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 provided for the different states of the cluster-level Markov chain that affects the behavior of all employees in the same firm. Moreover, the possibility is taken into account that the unit-level state changes are 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 for 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. (2013). 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. Nevertheless, it is worth mentioning the recent paper by Lagona et al. (2015), which proposes an extension of a hidden Markov model for bivariate time series law data in which there is a nested structure in time (months nested in government periods nested in legislatures). However, the context of this work is rather different from the longitudinal context here considered.

Under the extended model we propose, the manifest distribution of the response variables is computationally intractable in most applications; see Section 4.2 and 5.1 for a discussion about the computational complexity. Therefore, to make inference on the model we employ a composite likelihood method (Lindsay, 1988; Cox and Reid, 2004) based on 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) for a multilevel probit model; see also Hjort and Varin (2008) and Varin and Czado (2010). In particular, we show how to compute the pairwise likelihood by using the same recursion employed 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 propose and implemented along the same lines as in Bartolucci and Farcomeni (2009).

We also outline a weighted composite likelihood approach which may be in general more suitable in the presence of data grouped in clusters of different size. Through a simulation study, we compare the performance of the unweighted and weighted approaches under different scenarios in terms of efficiency of the estimators for the class of models developed in this article. We also consider some scenarios in which the full likelihood estimation is viable, due to a reduced complexity. In this way we can also evaluate the loss of efficiency of the proposed estimators with respect to the (typically infeasible) full likelihood estimator. Furthermore, we show how to obtain standard errors for the parameter estimates and how to make model selection using the CL-BIC approach formulated by Gao and Song (2010) which is a composite likelihood based selection criterion for high dimensional data models.

The paper is organized as follows. In Section 2 we briefly review the LM model with covariates and maximum likelihood estimation of this model. Section 3 provides a summary

of existing approaches related with the approach proposed here to deal with multilevel longitudinal data. This section also contains a brief overview of the recent literature about composite likelihood inference. The multilevel extension for LM models is illustrated in Section 4 for the case of continuous and binary response variables. Pairwise likelihood inference for this model is described in Section 5 and studied by simulation in Section 6. In Section 7 we illustrate the approach by an application based on the dataset concerning the sample of workers mentioned above and in Section 8 we draw the main conclusions.

An R implementation of the functions used for the estimation of the proposed model in the presence of binary response variables is available to the reader upon request.

2 Using hidden Markov chains for modeling time-varying unobserved heterogeneity

Consider a panel of n subjects observed at T occasions, 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 with 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 in 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

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

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 the transition probabilities are time homogenous.

More parsimonious models can be defined by imposing different constraints on the initial and transition probabilities; see also Bartolucci (2006). In particular, we may assume a stationary hidden Markov model, in the sense that the initial distribution is equal to the stationary distribution. This assumption is rather common in the hidden Markov literature because it provides more interpretable results in the presence of covariates also affecting the response variables; see Section 4 and Zucchini and MacDonald (2009) for a discussion about this constraint. Moreover, we may assume constant off-diagonal elements for each row of the transition matrix. This means that the probability of moving to a different state is invariant with respect to the state of destination, depending only on the current state. This additional constraint makes the model more parsimonious without loss of interpretability.

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 $\boldsymbol{Z}_i^{(t)}$:

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

where β_v is an intercept related to the latent state and $\boldsymbol{\delta}$ is a vector of regression coefficients. These parameters, including the variance σ^2, can be estimated together with the initial and transition probabilities. For binary response variables it is instead natural assuming that

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

where

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

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

The above approach may be extended to response variables having a different nature through a Generalized Linear Model (GLM) parametrization (McCullagh and Nelder, 1989) or similar parametrizations for multivariate categorical data (Bartolucci et al., 2007). We refer the reader to Bartolucci and Farcomeni (2009) and Maruotti (2011) for further details.

2.2 Maximum likelihood estimation

In an application, we observe the response configuration $\boldsymbol{y}_i = (y_i^{(1)}, \ldots, y_i^{(T)})'$ and the sequence of covariate vectors $\boldsymbol{z}_1^{(1)}, \ldots, \boldsymbol{z}_i^{(T)}$ for $i = 1, \ldots, n$, where n is the sample size; we collect the covariates in the unique vector $\boldsymbol{z}_i$ (for all time occasions). 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,

$$p(\boldsymbol{y}_i|\boldsymbol{z}_i) = \sum_{\boldsymbol{v}} p(\boldsymbol{y}_i|\boldsymbol{V}_i = \boldsymbol{v}, \boldsymbol{z}_i) p(\boldsymbol{V}_i = \boldsymbol{v}), \tag{1}$$

where the sum $\sum_{\boldsymbol{v}}$ is over all the possible configurations $\boldsymbol{v} = (v^{(1)}, \ldots, v^{(T)})'$ of the latent process $\boldsymbol{V}_i$.

Efficient computation of the probability in (1) may be performed by a forward recursion available in the hidden Markov literature (see Baum et al., 1970; Levinson et al., 1983; Zucchini and MacDonald, 2009). For every $i = 1, \ldots, n$, $t = 1, \ldots, T$, and $v = 1, \ldots, k$, the probability (or density) function

$$q_i^{(t)}(v) = p(y_i^{(1)}, \ldots, y_i^{(t)}, V_i^{(t)} = v|\boldsymbol{z}_i^{(1)}, \ldots, \boldsymbol{z}_i^{(t)})$$

is obtained by the recursion

$$q_i^{(t)}(v) = \sum_{\bar{v}=1}^{k} q_i^{(t-1)}(\bar{v}) \pi_{v|\bar{v}} p(y_i^{(t)}|V_i^{(t)} = v, \boldsymbol{z}_i^{(t)}), \quad v = 1, \ldots, k, \quad t = 1, \ldots, T, \tag{2}$$

with initialization $q_i^{(1)}(v) = \pi_v p(y_i^{(1)}|V_i^{(1)} = v, \boldsymbol{z}_i^{(1)})$. The manifest probability of $\boldsymbol{y}_i$ is finally obtained as

$$p(\boldsymbol{y}_i|\boldsymbol{z}_i) = \sum_{v=1}^{k} q_i^{(T)}(v).$$

This forward recursion requires a number of iterations of order $O(k^2 T)$ which linearly increases with the panel length. The memory requirement to store all values of $q_i^{(t)}(v)$,

$t = 1, \ldots, T$, $v = 1, \ldots, k$, is instead of order $O(kT)$. For further technical details, see Khreich et al. (2010).

Maximum likelihood estimation is performed on the basis of the log-likelihood $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{y}_i|\boldsymbol{z}_i)$, where $\boldsymbol{\theta}$ denotes the parameter vector specified according to the model of interest. We maximize this function by an EM algorithm (Baum et al., 1970; Dempster et al., 1977) 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 consists of computing 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 consists of maximizing this expected value with respect to $\boldsymbol{\theta}$, so that this parameter vector is 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). Nevertheless, the implementation of the EM algorithm must include further adjustments required to account for the stationarity assumption of the latent process; see Bulla and Berzel (2008).

3 Overview of the relevant literature

We provide a brief description of the approaches more closely related to the one proposed in this article, together with an overview about composite likelihood methods.

3.1 Models for multilevel longitudinal data

Latent variable models have been formulated in several ways to account for the additional unobserved correlation which may arise in the presence of clustered data. In the multilevel literature, clustering of longitudinal data is generally related to sets of individual observations grouped across the time. However, in this article we consider longitudinal data with a multilevel structure where clustering is related to sets of individuals grouped according to some criteria. Therefore, we are interested in suitable statistical tools for modeling two types of correlation that cannot be explained by means of the observable covariates: (*i*) correlation between responses of a single individual across time and (*ii*) correlation between

individuals belonging to the same cluster.

The standard LM approach, formulated as in Bartolucci and Farcomeni (2009) and in related papers, provides a well-established tool for modeling the first kind of correlation. We also have to mention that some attempts have been already accomplished to extend this approach to deal with the correlation at cluster level in the presence of multilevel longitudinal data. In particular, Bartolucci et al. (2009) considered an LM model with covariates for assessing the evolution of the health status of patients admitted in certain nursing homes. In this model, the correlation between individuals in the same facility is explained by including suitable fixed effects at cluster level. Later, Bartolucci et al. (2011) proposed an extension of LM model with covariates for the analysis of the ability in a sample of students belonging to different classes in different schools; the effect of each cluster is modeled by a discrete latent variable. Both approaches assume time-constant cluster effects. Moreover, the use of fixed effects at cluster level as in Bartolucci et al. (2009) is suitable only for analysis of samples with a limited number of large clusters.

We propose a multilevel LM extension based on nested Markov chains for modeling time-varying heterogeneity between individuals and between groups of individuals which allows for time-varying effects of the clusters and may also be used in the presence of many clusters of not large dimension. This is the main novelty of the proposed approach together with the use of composite likelihood methods required by the complexity of the resulting model.

It is also worth recalling the work Altman (2007) based on the mixed LM models where the cluster effect is modeled by including a continuous latent variables in an LM model. This approach permits modeling unobserved heterogeneity still in the presence of a large number of clusters, but it is not explicitly developed to deal with multilevel data. A related approach was developed by Maruotti (2011) on the basis of discrete latent variables.

3.2 Composite likelihood inference

Composite likelihood methods provide an alternative approach when full likelihood inference is not feasible. Problems in computing the full likelihood may arise for high dimensional data with a complex dependence structure. Nowadays composite likelihood techniques are widely employed in several contexts; for a rather recent review see Varin et al. (2011).

Following Lindsay (1988) and Cox and Reid (2004), the composite likelihood is a function obtained by means of the product of likelihood components referred to specific subsets of data. These components may correspond to marginal or conditional densities, so that the literature typically distinguishes between the conditional and marginal version.

Composite conditional likelihood, originally proposed by Besag (1974), is the product of the conditional density of each observation given its neighbors. Other versions have been later developed, for instance by Molenberghs and Verbeke (2004). Composite marginal likelihood is the product of the marginal density of subsets of observations. In particular, the pairwise likelihood, which considers densities of pairs of observations, represents a relevant case within the marginal approach (see Cox and Reid, 2004; Varin, 2008) and corresponds to the composite likelihood method we employ in this article. For related works on the pairwise likelihood method and the asymptotic properties of the estimators based on this approach, see also Hanfelt (2004), Varin and Vidoni (2005), and Gao and Song (2011).

Composite likelihood techniques inherit many of the interesting properties of the full likelihood inference which have been well established for several classes of models, mainly consistency and asymptotic normality of the resulting estimators. Nevertheless, there are still many practical and technical issues which represent open problems and need to be better explored. At this stage we recall three crucial aspects related to our application.

First, the sample estimate of the score variance is generally computed by the outer product of the composite scores. When this yields to a biased estimate of the Hessian matrix, the pairwise inference could not be asymptotically efficient; see Lindsay (1982). Nevertheless, the loss of efficiency might be reduced if the score variance is estimated by means of bootstrap methods or by Monte Carlo simulations; see Varin et al. (2011, Sec. 5.1) for an illustration of this method. The Monte Carlo solution is better illustrated in Section 5.2 for the class of models discussed in this paper.

Second, model selection also represents an important aspect because well-known criteria, such as AIC (Akaike, 1973) or BIC (Schwarz, 1978), cannot be directly used for composite likelihood inference. In this regard, Varin and Vidoni (2005) propose a method based on CLIC, a composite version of AIC, where the term of penalization depends on the estimation of the variance-covariance matrix of the specific model parameters discussed above.

Successively, Gao and Song (2010) propose a composite version of BIC, named CL-BIC, based on increasing the penalization term for the selection of high dimensional data models. The latter is the criterion adopted in this paper because, as shown by Gao and Song (2010) through a series of simulations, CL-BIC generally outperforms CLIC, in particular for complex models.

Finally, another important issue is whether a weighted approach provides useful insights for composite likelihood methods. This may be crucial in the presence of unequal cluster sizes as the composite approach we initially propose gives more weight to observations coming from larger clusters than to those coming from smaller clusters. See Renard et al. (2004) and Kuk and Nott (2000) for a discussion about this issue. Therefore, in Section 5.3 we also propose a weighted pairwise likelihood method for the class of models developed in this work and in Section 6 we compare the performance between the weighted and unweighted approach through a simulation study.

4 The proposed multilevel extension for LM models

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 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 LM approach outlined in Section 2.

4.1 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 unit 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 type of marginal independence only holds for subjects belonging to 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 $\boldsymbol{V}_{hi}$, we substantially adopt the same notation as in Section 2, and then we let

$$\begin{aligned} \pi_v &= p(V_{hi}^{(1)} = v), \quad v = 1, \ldots, k_2, \\ \pi_{v|\bar{v}} &= p(V_{hi}^{(t)} = v | V_{hi}^{(t-1)} = \bar{v}), \quad t = 2, \ldots, T, \ \bar{v}, v = 1, \ldots, k_2; \end{aligned}$$

these initial and transition probabilities are still collected in the vector $\boldsymbol{\pi}$ and in the matrix $\boldsymbol{\Pi}$, respectively.

To maintain a parsimonious parametrization, with respect to more standard versions of the LM model, we adopt two constraints on each hidden Markov chain at cluster and unit level: (*i*) the initial distribution coincides with the stationary distribution (stationarity); (*ii*) the transition probabilities from one latent state to another only depend on the previous state. With reference to the cluster-level processes, the first constraint amounts to assuming that $\boldsymbol{\Lambda}'\boldsymbol{\lambda} = \boldsymbol{\lambda}$, so that the marginal distribution is the same for any time occasion, that is, $p(U_h^{(t)} = u) = \lambda_u$, $u = 1, \ldots, k_1$. This constraint makes particularly sense if, as in the application illustrated in Section 7, the common time trend is modeled by including suitable explanatory variables among the covariates. Therefore, the latent states may be interpreted in terms of residuals with respect to this common trend that has always the same

distribution, although they are allowed to be serially dependent. The second constraint may be formally expressed as

$$\begin{aligned} \lambda_{u|\bar{u}} &= \lambda^*_{\bar{u}}, \quad u = 1 \ldots, k_1,\ u \neq \bar{u}, \\ \lambda_{\bar{u}|\bar{u}} &= 1 - (k_1 - 1)\lambda^*_{\bar{u}}, \end{aligned}$$

for $\bar{u} = 1, \ldots, k_1$, where $\lambda^*_{\bar{u}}$ are common transition probabilities between 0 and $1/(k_1 - 1)$. For instance, with $k_1 = 3$ latent states we have

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 - 2\lambda^*_1 & \lambda^*_1 & \lambda^*_1 \\ \lambda^*_2 & 1 - 2\lambda^*_2 & \lambda^*_2 \\ \lambda^*_3 & \lambda^*_3 & 1 - 2\lambda^*_3 \end{pmatrix}. \tag{3}$$

In this way, the Markov chain is parametrized on the basis of k_1 free parameters. Note that for the case of two latent states, $k_1 = 2$, there is no restriction on the transition matrix. Concerning the unit-level Markov chains, the same constraints (i) and (ii) as above are expressed on the basis of the common transition probabilities $\pi^*_{\bar{v}}$, $\bar{v} = 1, \ldots, k_2$, which are between 0 and $1/(k_2 - 1)$.

For the conditional response probabilities, the same considerations expressed in Section 2 still hold. Then, in the case of continuous response variables we assume that

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

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

$$Y^{(t)}_{hi} | U^{(t)}_h = u, V^{(t)}_{hi} = v, \boldsymbol{X}^{(t)}_h = \boldsymbol{x}, \boldsymbol{Z}^{(t)}_{hi} = \boldsymbol{z} \ \sim \ Bern(\psi^{(t)}_{hi}(u, v, \boldsymbol{x}, \boldsymbol{z})), \tag{4}$$

where

$$\log \frac{\psi^{(t)}_{hi}(u, v, \boldsymbol{x}, \boldsymbol{z})}{1 - \psi^{(t)}_{hi}(u, v, \boldsymbol{x}, \boldsymbol{z})} = \alpha_u + \beta_v + \boldsymbol{x}'\boldsymbol{\gamma} + \boldsymbol{z}'\boldsymbol{\delta}, \tag{5}$$

with parameters having the same interpretation as above. Obviously, this formulation may be extended also to other cases by a GLM parametrization or similar parametrizations, as illustrated at the end of Section 2.1.

4.2 Manifest distribution

When we observe a set of multilevel longitudinal data, we have a sequence of responses $\boldsymbol{y}_{hi} = (y_{hi}^{(1)}, \ldots, y_{hi}^{(T)})'$ 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 $\boldsymbol{z}_{hi}^{(1)}, \ldots, \boldsymbol{z}_{hi}^{(T)}$; these covariates are collected in the unique vector $\boldsymbol{z}_{hi}$ when referred to 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:

$$p(\boldsymbol{y}_h|\boldsymbol{x}_h, \boldsymbol{z}_h) = \sum_{\boldsymbol{u}} p(\boldsymbol{U}_h = \boldsymbol{u}) p(\boldsymbol{y}_h|\boldsymbol{U}_h = \boldsymbol{u}, \boldsymbol{x}_h, \boldsymbol{z}_h), \tag{6}$$

where

$$p(\boldsymbol{y}_h|\boldsymbol{U}_h = \boldsymbol{u}, \boldsymbol{x}_h, \boldsymbol{z}_h) = \prod_{i=1}^{n_h} \Big[\sum_{\boldsymbol{v}} p(\boldsymbol{y}_{hi}|\boldsymbol{U}_h = \boldsymbol{u}, \boldsymbol{V}_{hi} = \boldsymbol{v}, \boldsymbol{x}_h, \boldsymbol{z}_h) p(\boldsymbol{V}_{hi} = \boldsymbol{v}) \Big], \tag{7}$$

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

According to the above expression, the manifest distribution $p(\boldsymbol{y}_h|\boldsymbol{x}_h, \boldsymbol{z}_h)$ is obtained by computing first the conditional distribution $p(\boldsymbol{y}_h|\boldsymbol{U}_h = \boldsymbol{u}, \boldsymbol{x}_h, \boldsymbol{z}_h)$, and then marginalizing the latter with respect to $\boldsymbol{U}_h$. However, even if we employ recursion (2) instead of the expression in (7) for computing $p(\boldsymbol{y}_h|\boldsymbol{U}_h = \boldsymbol{u}, \boldsymbol{x}_h, \boldsymbol{z}_h)$, the number of operations required for computing $p(\boldsymbol{y}_h|\boldsymbol{x}_h, \boldsymbol{z}_h)$ my be huge. In fact, this number is of order $O(k_1^T k_2^2 n_h T)$ which linearly increases with the cluster size n_h and exponentially with the panel length T. The memory requirement to store all the forward probabilities is instead of order $O(k_1^T k_2 n_h T)$. To clarify this point, Table S-1 in the Supplementary Material shows the order of the number of operations for different combinations of k_1, k_2 $(1, \ldots, 5)$, T $(5, 10)$, n_h $(5, 10, 20)$.

The results in Table S-1 confirm that the numerical complexity rapidly increases with T; in particular, with $T = 10$ this complexity makes prohibitive to compute $p(\boldsymbol{y}_h|\boldsymbol{x}_h, \boldsymbol{z}_h)$ even for a small cluster size and a limited number of latent states. Indeed, we experienced that

computing this distribution is infeasible even with values of T smaller than 10, which may be easily encountered in social and economic applications, and in certain medical studies. Therefore, maximum likelihood estimation for the multilevel LM model parameters cannot be generally performed. This surely happens for the application motivating this article, which is described in Section 7. For this reason, we suggest below a composite likelihood approach based on a likelihood function with components corresponding to all possible pairs of units in each cluster. This is then a pairwise likelihood function across sample units.

Nevertheless, when computing the manifest distribution $p(\boldsymbol{y}_h|\boldsymbol{x}_h, \boldsymbol{z}_h)$ is feasible, because the number of time occasions is limited, on the basis of this distribution we may compute a full likelihood function. This function may be maximized by an extension of the algorithm described in Bartolucci and Farcomeni (2009) and related to the extension used by Bartolucci et al. (2011) for estimating a simpler multilevel version of the LM model. We skip details about this algorithm here because the resulting full maximum likelihood estimator is only used as a comparison for the composite likelihood estimator in the simulation study dealt with in Section 6, when this estimator may be used, and it is not suggested in general.

5 Pairwise likelihood inference

In order to make inference on the model parameters, we propose the use of the following pairwise log-likelihood:

$$\begin{aligned} p\ell(\boldsymbol{\theta}) &= \sum_{h=1}^{H} \sum_{i=1}^{n_h-1} \sum_{j=i+1}^{n_h} p\ell_{hij}(\boldsymbol{\theta}), \\ p\ell_{hij}(\boldsymbol{\theta}) &= \log p(\boldsymbol{y}_{hi}, \boldsymbol{y}_{hj}|\boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj}), \end{aligned} \tag{8}$$

which recalls the pairwise log-likelihood used by Renard et al. (2004) in a simpler context. With binary response variables, the parameter vector $\boldsymbol{\theta}$ includes the support points α_u, $u = 1, \ldots, k_1$, and β_v, $v = 1, \ldots, k_2$, the regression coefficients for the covariates $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ (see expression (5)), and the parameters $\rho_{\bar{u}}$, $\bar{u} = 1, \ldots, k_1$, and $\tau_{\bar{v}}$, $\bar{v} = 1, \ldots, k_2$, defined on the basis of a logit-type transformation of the transition probabilities $\lambda^*_{\bar{u}}$ and $\pi^*_{\bar{v}}$. More

precisely, the transformations used for these probabilities are:

$$\rho_{\bar{u}} = \log \frac{\lambda^*_{\bar{u}}}{1-(k_1-1)\lambda_{\bar{u}^*}}, \quad \bar{u} = 1, \ldots, k_1, \tag{9}$$

$$\tau_{\bar{v}} = \log \frac{\pi^*_{\bar{v}}}{1-(k_2-1)\pi_{\bar{v}^*}}, \quad \bar{v} = 1, \ldots, k_2; \tag{10}$$

see Appendix for technical details. These transformations are useful because the resulting parameter space is $\Re^{|\boldsymbol{\theta}|}$, where $|\boldsymbol{\theta}|$ stands for the dimension of $\boldsymbol{\theta}$.

Note that, when the dimension of each cluster is two ($n_h = 2$, $h = 1, \ldots, H$), $p\ell(\boldsymbol{\theta})$ is the full log-likelihood of the model, since it is based on the manifest probability of the responses provided by all possible pairs of subjects in the same cluster. Moreover, this function must be suitably modified by adding a component of type

$$\sum_{\substack{h=1 \\ n_h=1}}^{H} \log p(\boldsymbol{y}_{h1}|\boldsymbol{x}_h, \boldsymbol{z}_{h1}), \tag{11}$$

whereas the sum in (8) is referred to all clusters including at least two units. However, in order to simplify the description of the algorithm to maximize $p\ell(\boldsymbol{\theta})$, in the following we consider component (11) as omitted. Nevertheless, the adjustments required in the presence of some clusters with only one unit are minimal.

5.1 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 employ recursion (2) already used for the model illustrated in Section 2. In fact, we have that

$$p(\boldsymbol{y}_{hi}, \boldsymbol{y}_{hj}|\boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj}) = p(\tilde{\boldsymbol{y}}^{(1)}_{hij}, \ldots, \tilde{\boldsymbol{y}}^{(T)}_{hij}|\boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj}), \tag{12}$$

where $\tilde{\boldsymbol{y}}^{(t)}_{hij}$ is a realization of the vector $\tilde{\boldsymbol{Y}}^{(t)}_{hij} = (Y^{(t)}_{hi}, Y^{(t)}_{hj})'$. 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 $\boldsymbol{W}^{(1)}_{hij}, \ldots, \boldsymbol{W}^{(T)}_{hij}$, where $\boldsymbol{W}^{(t)}_{hij} = (U^{(t)}_h, V^{(t)}_{hi}, V^{(t)}_{hj})'$, 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 $\boldsymbol{w} = (u, v_1, v_2)'$. It is simple to see that the initial probability of state $\boldsymbol{w}$ is

$$\phi_{\boldsymbol{w}} = p(\boldsymbol{W}^{(1)}_{hij} = \boldsymbol{w}) = \lambda_u \pi_{v_1} \pi_{v_2}, \tag{13}$$

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

$$\phi_{\boldsymbol{w}|\bar{\boldsymbol{w}}} = p(\boldsymbol{W}_{hij}^{(t)} = \boldsymbol{w} | \boldsymbol{W}_{hij}^{(t-1)} = \bar{\boldsymbol{w}}) = \lambda_{u|\bar{u}} \pi_{v_1|\bar{v}_1} \pi_{v_2|\bar{v}_2}. \tag{14}$$

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

$$p(\tilde{\boldsymbol{y}}_{hij}^{(t)} | \boldsymbol{W}_{hij}^{(t)} = \boldsymbol{w}, \boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj}) = p(y_{hi}^{(t)} | u, v_1, \boldsymbol{x}_h^{(t)}, \boldsymbol{z}_{hi}^{(t)}) p(y_{hj}^{(t)} | u, v_2, \boldsymbol{x}_h^{(t)}, \boldsymbol{z}_{hj}^{(t)}). \tag{15}$$

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 (2) is applied with the probability or density $p(\boldsymbol{y}_i^{(t)} | V_i^{(t)} = v, \boldsymbol{z}_i^{(t)})$ substituted by the probability or density $p(\tilde{\boldsymbol{y}}_{hij}^{(t)} | \boldsymbol{W}_{hij}^{(t)} = \boldsymbol{w}, \boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj})$ 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 $\phi_{\boldsymbol{w}|\bar{\boldsymbol{w}}}$. In this way, the order of the number of iterations representing the computational complexity is $O(k_1^2 k_2^4 n_h (n_h - 1) T/2)$ which linearly, rather than exponentially, increases with the panel length; see Table S-2 for an illustration.

From the comparison between Tables S-1 and S-2 in the Supplementary Material, we notice that the pairwise based approach generally leads to a complexity reduction as $k_1 > 2$. In particular, the ratio between the order of the complexity under the pairwise and the full likelihood approach when $T = 5$ is 0.30 for $k_1 = 3$, $k_2 = 2$ and $n_h = 5$, it is 0.63 for $k_1 = 4$, $k_2 = 3$, $n_h = 10$ and it is 0.59 for $k_1 = 4$, $k_2 = 2$, $n_h = 20$. This reduction is more evident in long panels with smaller cluster size; when $T = 10$, this ratio is always lower than 1 when $k_1 > 1$; moreover with $k_1 = k_2 = 2$ the ratio is 0.03 for $n_h = 5$, 0.07 for $n_h = 10$ and 0.15 for $n_h = 20$; with $k_1 = 2$ and $k_2 = 3$ the ratio is 0.07 for $n_h = 5$, 0.16 for $n_h = 10$ and 0.33 for $n_h = 20$; with $k_1 = 3$ and $k_2 = 2$ it becomes closer to zero even for larger cluster size. An emblematic case is for $k_1 = k_2 = 5$, $n_h = 5$, and $T = 10$, when the order of the number of operations required by the full likelihood approach is around 7,800 times that required by the pairwise likelihood method.

The pairwise log-likelihood $p\ell(\boldsymbol{\theta})$ defined in (8) 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

$$\begin{aligned} p\ell^*_{hij}(\boldsymbol{\theta}) &= \sum_{\boldsymbol{w}} d^{(1)}_{hij}(\boldsymbol{w}) \log \phi_{\boldsymbol{w}} + \sum_{t=2}^{T} \sum_{\bar{\boldsymbol{w}}} \sum_{\boldsymbol{w}} d^{(t)}_{hij}(\bar{\boldsymbol{w}}, \boldsymbol{w}) \log \phi_{\boldsymbol{w}|\bar{\boldsymbol{w}}} \\ &\quad + \sum_{t=1}^{T} \sum_{\boldsymbol{w}} d^{(t)}_{hij}(\boldsymbol{w}) \log p(\tilde{\boldsymbol{y}}^{(t)}_{hij} | \boldsymbol{W}^{(t)}_{hij} = \boldsymbol{w}, \boldsymbol{x}_h, \boldsymbol{z}_{hi}, \boldsymbol{z}_{hj}). \end{aligned} \tag{16}$$

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

The complete data pairwise log-likelihood may be simply expressed in terms of the parameters of the proposed multilevel model by substituting (13), (14), and (15) in the above expression. For instance, the first component becomes the sum over u of

$$\tilde{d}^{(1)}_{hij}(u) \log \lambda_h(u) + \sum_{v_1=1}^{k} \tilde{d}^{(1,1)}_{hij}(u, v_1) \log \pi_{hi}(v_1|u) + \sum_{v_2=1}^{k} \tilde{d}^{(1,2)}_{hij}(u, v_2) \log \pi_{hj}(v_2|u), \tag{17}$$

where the variables $\tilde{d}^{(1)}_{hij}(u)$, $\tilde{d}^{(1,1)}_{hij}(u, v_1)$, and $\tilde{d}^{(1,2)}_{hij}(u, v_2)$ are obtained by summing $d^{(1)}_{hij}(\boldsymbol{w})$ 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 $d^{(t)}_{hij}(\boldsymbol{w})$ and $d^{(t)}_{hij}(\bar{\boldsymbol{w}}, \boldsymbol{w})$ is computed by using the same recursions used in the algorithm of Baum et al. (1970). At the M-step, the model parameters are updated by maximizing the function resulting by substituting these expected values in (16) and using simplification (17) and similar simplifications. In any case, the EM algorithm is implemented along the same lines as the algorithm used for computing the full likelihood, taking into account that the parameters $\rho_{\bar{u}}$ and $\tau_{\bar{v}}$ for the transition probabilities, defined in (9) and (10), are updated by a numerical algorithm which also accounts for the constraint that the initial distribution of each latent process coincides with its stationary distribution (Bulla and Berzel, 2008).

In order to make the pairwise likelihood estimation faster, we combine the EM algorithm previously illustrated with the BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm which

is a well-known iterative method for solving unconstrained nonlinear optimization problems (Avriel, 2003). In practice, the first algorithm is run for a limited number of steps, and then the second algorithm is used to directly reach the maximum of $p\ell(\boldsymbol{\theta})$ starting from the EM solution. For the implementation of the BFGS algorithm, the score of the objective function is analytically computed and only the Hessian matrix is numerically derived. This procedure has proved to have very good performance because the second-order derivative is not re-computed at each step, but it is updated in a suitable way. This leads to a sharp estimate of the observed information matrix and to a faster convergence with respect to using only the EM algorithm until final convergence.

From the results of the EM algorithm and, in particular, on the basis of the posterior expected values computed at the E-step, a local decoding for each cluster and unit can be performed. This amounts to single out the state that, for each time occasion, has the greatest *a posteriori* probability given the observed sequences of responses and covariates. In particular, the quantities to be used are the posterior expected values of suitable sums of the indicator variables $d_{hij}^{(t)}(\boldsymbol{w})$ of same type used in equation (17). An alternative procedure, named global decoding, is based on the Viterbi algorithm (see Viterbi, 1967; Juang and Rabiner, 1991) that, in the present case, allows us to find in an efficient way the path of states with the highest *a posteriori* probability for each cluster or unit. The predicted paths, either found by the local or global decoding, may be represented in plots similar to those in Figures S-1 and S-2 in which, however, such latent trajectories are conditional on the observed responses.

5.2 Model selection and hypothesis testing

For model selection we adopt the criterion suggested by Gao and Song (2010) which may be seen as the counterpart of BIC for composite likelihood inference. Then, the model to be selected is the one which maximizes the following index:

$$\text{CL-BIC} = p\ell(\hat{\boldsymbol{\theta}}) - \text{tr}(\hat{\boldsymbol{J}}^{-1}\hat{\boldsymbol{K}})\frac{\log(H)}{2}, \tag{18}$$

where $\text{tr}(\hat{\boldsymbol{J}}^{-1}\hat{\boldsymbol{K}})$ is a penalty for the model complexity. The general CL-BIC proposed by Gao and Song (2010) also includes a further penalty term related to the model space

complexity which makes the criterion applicable to much broad ranges of likelihood or quasi-likelihood approaches. However we deem this further penalty term does not provide useful insights in the present context and it is here omitted. Matrices $\hat{\boldsymbol{J}}$ and $\hat{\boldsymbol{K}}$ are the estimates of

$$\boldsymbol{J} = \mathrm{E}\left[-\frac{\partial^2 p\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right], \qquad \boldsymbol{K} = \mathrm{V}\left[\frac{\partial p\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right].$$

In Renard et al. (2004) these matrices are estimated as

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

and, consequently, the variance-covariance matrix of the pairwise likelihood estimator $\hat{\boldsymbol{\theta}}$, and its standard errors, are obtained by the sandwich formula

$$\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{J}}^{-1}\hat{\boldsymbol{K}}\hat{\boldsymbol{J}}^{-1}. \tag{19}$$

In particular, $\hat{\boldsymbol{J}}$ is obtained from the Hessian provided by the numerical algorithm described at the end of Section 5.1.

On the basis of a series of simulated samples, we verified that the estimator $\hat{\boldsymbol{K}}$ of the variance of the score previously illustrated is often not stable and works out not coherent values of the penalty term for CL-BIC, because less penalty values may result for more complex models. More stable estimates are instead achieved for the empirical score expectation. Therefore we estimate $\boldsymbol{K}$ through a Monte Carlo method, as suggested among others by Varin et al. (2011). This method is based on: (*i*) drawing a suitable number of samples from the estimated model; (*ii*) computing for each simulated sample the pseudo-score $\partial p\ell(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$; and (*iii*) estimating $\boldsymbol{K}$ as the variance-covariance matrix of these score vectors. We experienced that in this way more stable estimates of the variance of the score are obtained and this, in turn, implies penalty terms more coherent with the model complexity computed according to expression (18).

We use the CL-BIC 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, this criterion can be also used for selecting one of the possible parametrizations illustrated in Section 4.

An important issue that must be considered when we deal with latent variable models is the parameter identifiability, as lack of identifiability may lead to complications for es-

timation and problems regarding the inferential properties of the estimators. As typically happens with reference to these models, we focus on local identifiability, that is, identifiability in a neighborhood of a given parameter value, rather than global identifiability that refers to the whole parameter space (e.g., McHugh, 1956; Rothenberg, 1971); see also Bartolucci (2006) with specific reference to a basic LM model. In order to establish that the model is locally identifiable, we check the rank of the observed information matrix. The condition that this matrix has full rank must be always verified in the inferential procedure described in this section (in particular for model selection and to derive standard errors for the parameter estimates), as these procedures are based on the inverse of such a matrix.

5.3 Weighted pairwise likelihood

In the current form, the pairwise likelihood defined in (8) gives more weight to the units belonging to larger clusters. This is because the number of components involving the data of a specific unit is equal to the number of other units in the same cluster. 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. Then, we propose the following weighted pairwise log-likelihood:

$$p\ell_\xi(\boldsymbol{\theta}) = \sum_{h=1}^{H} \xi_h p\ell_h(\boldsymbol{\theta}), \tag{20}$$

where $\xi_h = 1$ if the size of cluster h is $n_h = 1$, and $\xi_h = (n_h - 1)^{-1}$ otherwise, for $h = 1, \ldots, H$. Then, the weight given to each sample unit proportionally decreases with the cluster size, so that in models with $k_1 = k_2 = 1$ (no latent structure) we obtain from the maximization of (20) the same estimates obtained from the maximization of the full likelihood; the same does not happen with the unweighted pairwise likelihood initially defined. Pairwise likelihood estimations and model selection for the weighted approach can be carried out using the maximization procedure and CL-BIC discussed in Sections 5.1 and 5.2, respectively.

Notice that Renard et al. (2004) discuss how the weighted approach should be in general desirable, but in some occasions the unweighted pairwise may perform better. Then, the choice strictly depends on the problem at hand. The simulation study presented in the following section compares the performance of the unweighted and weighted pairwise

likelihood methods with the performance of the full likelihood method when the latter may be used.

6 Simulation study

In this section, we illustrate and discuss the results of a simulation study aimed at assessing the properties of the pairwise maximum likelihood estimator outlined in Section 5 for the proposed multilevel LM model in its version for binary variables. This study is based on scenarios that recall the application described in Section 7.

Given a model with k_1 and k_2 latent states, 100 datasets are simulated for each scenario obtained combining different values of number of occasions T, number of clusters H, and sample size. In detail, we consider $T = 5, 10$, $H = 200, 400$, whereas the number of units per cluster is generated either from the uniform distribution $U(1, 10)$ or from $U(1, 20)$. In this way, the average sample size is 1100 for $H = 200$ with $U(1, 10)$, 2200 for $H = 400$ with $U(1, 10)$, 2100 for $H = 200$ with $U(1, 20)$, and 4200 for $H = 400$ with $U(1, 20)$. Moreover two covariates at unit level are generated from an AR(1) process with autocorrelation equal to 0.5.

The simulation design contemplates eight different scenarios which are replicated under three multilevel LM models characterized by three different numbers of latent states; in particular, data are generated for the model with $k_1 = k_2 = 2$, with $k_1 = 2$ and $k_2 = 3$, and with $k_1 = 3$ and $k_3 = 2$. Data simulated under the three different models are fitted using the pairwise likelihood inference illustrated in Section 5 based both on the unweighted and on the weighted approach described Section 5.3. The simulated data are also fitted using the full likelihood inference illustrated in Section 2, only for the scenarios in which computing the full likelihood is feasible, that is, for $T = 5$ with $k_1 = k_2 = 2$ or $k_1 = 2$ and $k_2 = 3$. The results displayed in Tables S-3 to S-8 of the Supplementary Material allow us to compare the performance of the unweighted and weighted pairwise inference also with respect to the full likelihood approach.

Table S-3 contains the results related to the four different scenarios for the model with $k_1 = k_2 = 2$ latent states and $T = 5$. The first row of the table shows the true parameters of

the model which are, for the cluster-level latent process $\boldsymbol{U}_h$, the logit-type transformation ρ_1 and ρ_2 of the transition probabilities. For the unit level latent process $\boldsymbol{V}_{hi}$, we use the logit transformations τ_1 and τ_2 of the transition probabilities. The intercept and the regression coefficients for the two covariates generated at unit level are denoted by δ_0, δ_1, δ_2 and the support points are denoted by α_2 and β_2 (given that $\alpha_1 = \beta_1 = 0$). For each scenario we show the bias, the standard deviation (sd), and the root mean square error (rmse) of the estimates computed on the basis of the 100 replications. Furthermore, for the pairwise likelihood inference we compare the standard deviation of the estimates with the mean of the estimated standard errors (mean-se) of the estimates computed on the basis of sandwich formula in (19).

The results in Table S-3 show that, for both pairwise and full likelihood approaches, the highest bias values are for the parameters involved in the individual latent process (τ_1, τ_2) and the support point β_2. The bias decreases as the sample size increases in the second and, in particular, in the fourth scenario. In terms of bias the weighted pairwise approach performs better than the unweighted one; for instance, for τ_1 in the first scenario the bias under the former case is nearly the 50% of the bias under the latter. More generally, the table shows that the weighted pairwise likelihood inference provides a better approximation of the full likelihood inference with respect to the unweighted approach. For instance, for τ_1 in the first scenario, rmse with the unweighted approach is 2.6 times that with the full likelihood approach; this ratio decreases to 1.6 using the weighted approach. In the pairwise likelihood approaches, the difference between mean-se and sd mainly increases for the parameters involved in the individual latent process, even if these differences are more evident for the unweighted rather than for the weighted approach. For instance, for τ_2 in the first scenario, the ratio between mean-se and sd is nearly 1 under the weighted approach, that is 0.926, whereas it is 0.575 under the unweighted one. Nevertheless, the sandwich formula (19) provides good estimates of the true standard error as the sample size increases (in particular for the unweighted approach); for the same parameter τ_2 in the fourth scenario, the same ratio becomes 1.109 and 1.143 for the weighted and unweighted approach, respectively.

Similar considerations also hold for the scenarios considered in Table S-4 and referred

to models with latent states $k_1 = 2$, $k_2 = 3$, and $T = 5$. The first row of this table includes one more parameter τ_3 for the latent chain $\boldsymbol{V}_{hi}$ and one more support point β_3. In models with higher values of k_2, estimation of the parameters for $\boldsymbol{V}_{hi}$ seems to be more critical; however, the weighted pairwise approach is closer to the full likelihood inference, compared to the unweighted one, both in terms of bias and efficiency. For τ_3 in the second scenario, rmse is 0.833, 0.548 and 0.370 under unweighted pairwise, weighted pairwise, and full likelihood inference, respectively. In these scenarios, the standard error estimates of the parameters for $\boldsymbol{V}_{hi}$ under the pairwise approach are especially critical when the sample size is small, whereas for the other parameter estimates good results are obtained. For instance, consider τ_2 in the first scenario; the ratio between mean-se and sd is 0.392 for the unweighted approach, however the same ratio improves and becomes 0.598 under the weighted approach.

Table S-5 contains the results referred to the four different scenarios for the model with $k_1 = 3$ and $k_2 = 2$ latent states and $T = 5$. The results are only referred to the pairwise approaches because we noticed that, in this case, computing the full likelihood estimator is very slow even with 5 time occasions. The first row includes one more parameter ρ_3 for the latent chain $\boldsymbol{U}_h$ and one more support point α_3. As in the previous tables, the highest bias and rmse are generally for the estimates of the individual latent state model. The weighted approach always outperforms the unweighted one, especially for $H = 200$ with a smaller sample size, even for the standard error estimates; in the first scenario, the ratio between mean-se and sd for τ_2 is 0.622 and 0.994 under the unweighted and weighted pairwise, respectively.

The simulation study shows very good results for the pairwise inference both in terms of bias and standard deviation when the number of occasion increases; see the results for $T = 10$ in Table S-6 for models with latent states $k_1 = k_2 = 2$, in Table S-7 for $k_1 = 2$ and $k_2 = 3$, and in Table S-8 for $k_1 = 3$ and $k_2 = 2$. These tables show the performance only for the pairwise inference because computing the full likelihood is not feasible in these cases. We observe that the results are more positively affected by an increase of the number of time occasions than from an increase of the sample size. In Tables S-7 and S-8, in particular in scenarios with $H = 200$, we notice that the estimates under the weighted approach are more

stable (lower standard error) whereas the estimates under the unweighted approach are more accurate (lower bias). However in all of these three tables, the weighted pairwise inference generally performs better in terms of rmse than the unweighted one. For instance, in the fourth scenario considered in Table S-6, rmse for τ_1 is 0.136 and 0.142 under the weighted and unweighted approach, respectively. In the fourth scenario (Table S-7), the rmse for τ_3 is 0.158 and 0.181 under the weighted and unweighted approach, respectively. In the fourth scenario (Table S-8), the rmse for τ_2 is 0.166 and 0.178 under the weighted and unweighted approach, respectively. In particular, we notice that in models with $T = 10$, the estimates of the standard errors and the standard deviations are very close. For instance, for τ_1 in the first scenario, the ratio between mean-se and sd is 0.850 and 0.842 in Table S-6, 0.874 and 0.861 in Table S-7 for the weighted and unweighted pairwise inference, respectively. For τ_2 in the first scenario of Table S-8, the same ratio is 1 and 0.959 for the weighted and unweighted pairwise inference, respectively.

7 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 3,850 employees (both blue-collars and white-collars, 24% of them are women) grouped in 402 private Italian firms with more than 1,000 workers. The subjects, continuously working in the same firm and aged between 18 and 60 in 1994, were followed for 10 years, from 1994 to 2004.

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-level covariates: *gender* (dummy equal to 1 for woman), *age* in 1994, *time* (time occasion from 1 and 10), *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) and we also include the lagged response. The unit-level covariates are time dependent except *gender* and *age* in 1994. Then we consider a set of cluster-level covariates: *area* (North-West, North-East, Center, South, or Islands) and *size* (number of

observed workers in each firm assuming value between 1 and 252). In the application, the non-linear effect of age and time is assumed including age^2 and $time^2$ among the covariates.

Given the reduced number of covariates, we deem the response variable of interest may depend on unobservable latent variables related to the propensity of employees in asking for illness benefits and to the propensity of firms in allowing their employees to receive these benefits. Furthermore, we are interested in verifying the hypothesis that, across time, these latent traits could be influenced by dynamic (unobserved) aspects related to the worker's life and to the management.

The model described in Section 4 is fitted on this dataset under assumptions (4) and (5) for the conditional distribution of the binary response *illness*. Moreover, we assume that the two latent Markov chains at cluster and unit level are stationary with constant off-diagonal elements for each row of the transition matrices, as described in Section 4. Then, in this application, we estimate the effect of the covariates on the probability to get ill in addition to the unobservable heterogeneity modeled over time through two stationary Markov processes. The unit-level process is expected to capture the worker propensity (not explained by the observed covariates) to get ill, whereas the cluster-level latent process explains the effect of different firms on the propensity to allow for illness benefits. Model estimation is carried out using the weighted pairwise inference illustrated in Section 5.

The first step of the analysis is the choice of k_1 and k_2 for both processes. This choice is based on CL-BIC defined in (18). The value of this index is reported in Table 1 for different values of k_1 and k_2. According to these results, the model with $k_1 = 3$ latent states at cluster level and $k_2 = 3$ latent states at unit-level is selected.

[Table 1 about here.]

Table 2 collects the estimates of the regression parameters for the selected model with $k_1 = k_2 = 3$ states (displayed on the right side), and also for other two models of interest with $k_1 = k_2 = 1$ (model with fixed effects displayed on the left side) and with $k_1 = k_2 = 2$.

[Table 2 about here.]

From the comparison of the estimates derived under the three fitted models, we notice that gender and all cluster-specific covariates (size and area) are not significant; also the

quadratic effects of age and time are not significant. On the other hand, the probability of receiving illness benefits is positively related to age, time, to being a blue-collar, whereas it is negatively related to income and to having a part-time job. The positive effect of the lagged response requires a more detailed discussion. This effect is shown to be significant under the models with a lower number of latent states, whereas it becomes not significant under the selected model. We consider the significant effects found in the models with $k_1 = k_2 = 1$ and $k_1 = k_2 = 2$ to be a spurious association arising because unobserved heterogeneity is not properly accounted for. In fact, this association vanishes after conditioning on the latent processes with $k_1 = k_2 = 3$. Then, the probability of receiving illness benefits is explained by a set of covariates and also by unobservable aspects which cannot be ignored. This is confirmed by the fact that the model with no random effects, that is, with $k_1 = k_2 = 1$, provides a lower value of CL-BIC. In particular, the correlation between workers in the same firm seems to be completely explained by the cluster-level latent process given that there are no significant cluster-specific covariates.

The dependence on time of the latent variables at cluster and unit level is also considered. Then, the data are fitted assuming three additional constraints on the selected models: (*i*) identity transition matrix for the cluster-level process (time-constant cluster random effects), (*ii*) identity transition matrix for the unit-level process (time-constant unit random effects) and (*iii*) identity transition matrices for both the cluster and unit-level processes (time-constant cluster and unit random effects). Note that, when we constrain the transition matrix to be diagonal, either at cluster or unit level, we adopt free initial probabilities. The CL-BIC values obtained under these three models are (*i*) -6076.29, (*ii*) -6069.15 and (*iii*) -6089.32, respectively; all of them are lower than -6057.69 obtained for the selected model with time-varying latent variables at both levels.

About the distribution of the cluster- and unit-level latent processes, the estimates of the initial and transition probabilities are reported in Tables 3 and 4 which also include the support point estimates.

[Table 3 about here.]

[Table 4 about here.]

For both processes, we observe that the states are well separated, the first and second states have the highest probability for the cluster-level process, whereas the second state has the highest probability for the individual-level process. The estimates of the transition matrices show a high persistence, in particular for the cluster-level latent process. In order to illustrate the behavior of the latent processes at individual and cluster-level based on these parameters, in Figures S-1 and S-2 we represent 1000 simulated trajectories from the corresponding distributions.

Nevertheless, simpler models implying time-constant random effects provide a lower value of CL-BIC compared to the selected model. This may depend on the fact that we are dealing with a long panel with a large number of clusters of small size where the transitions between latent states, even if moderate, are not negligible both for employees and firms.

8 Conclusions

With reference to multilevel longitudinal data, where sample units are collected in clusters, we propose an approach to account for the unobserved heterogeneity between sample units and between clusters in a dynamic fashion. For this class of models, computing the full likelihood in often infeasible, in particular with long panel data. Therefore, we propose a pairwise likelihood approach for model estimation. We remark that the approach we propose is rather new in the literature for multilevel longitudinal data, because this type of composite likelihood is defined for subsets of individuals in each cluster and not for subsets of time occasions for each individual. Hence we assess the performance of the proposed pairwise likelihood inference by means of a simulation study. From this study we draw the following conclusions.

First, we notice that the estimates of the latent parameters are in general more crucial than the estimates of the regression parameters for the observed covariates which show a very low bias and mean square error. Second, pairwise likelihood inference becomes more efficient as the sample size increases, but in particular as the panel length increases. Third, the weighted pairwise approach generally performs better than the unweighted one because the former better approximates the maximum likelihood estimates (when computing the

full likelihood is feasible) and in general provides lower mean square errors, then reducing the loss of information and increasing the efficiency of the parameter estimates.

Pairwise likelihood inference for multilevel LM models is used for the analysis of a longitudinal dataset based on a sample of Italian workers employed in different firms. The results of this application show how the proposed model is suitable for long panels with a large number of clusters of small size, that is, in a context where computing maximum likelihood estimates is demanding (even not feasible) and the pairwise likelihood inference has shown to provide good asymptotic properties. It is worth noting that the analyzed dataset supports the hypothesis of time-dependent latent variables both at cluster and unit level. This is the main novelty of the proposed extension for LM model based on nested hidden Markov chains. This extension provides useful insights for panel data with a limited number of covariates, as the one considered in this application in which the unobservable attitude of employees and firms in asking and allowing for illness benefits are influenced by several aspects which may deeply vary over the considered period of 10 years.

Finally, there are some aspects that require special attention, among others the so-called initial condition problem (Skrondal and Rabe-Hesketh, 2013) that may lead to biased estimates. We deem that this problem is mitigated for the application discussed in this paper because most of the covariates (also the lagged response) are not significant and the simulation results show the bias strongly decreases in similar contexts. Nevertheless it is an aspect to consider in particular for the analysis of short panels. Finally, although we consider an application with binary response variables, the methodological approach illustrated in this paper is completely general in terms of type of response variables, which may be also categorical, ordinal or not ordinal.

Appendix: logit-type transformation of the transition probabilities

As shown in Section 5, the model parameters includes a transformation of the transition probabilities based on (9) and (10). First of all, it is important to note that the inverse of

the first transformations is

$$\lambda^*_{\bar{u}} = \frac{\exp(\rho_{\bar{u}})}{1 + (k_1 - 1)\exp(\rho_{\bar{u}})}, \quad \bar{u} = 1, \ldots, k_1,$$

which ensures that $0 < \lambda^*_{\bar{u}} < 1/(k_1 - 1)$ for $\rho_{\bar{u}} \in \Re$. The inverse of the transformation for $\pi^*_{\bar{v}}$ is defined in a similar way on the basis of $\tau_{\bar{v}}$.

Obviously, on the basis of the transition probabilities $\lambda^*_{\bar{u}}$ and $\pi^*_{\bar{v}}$, we obtain in a simple way the corresponding transition matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Pi}$, see expression (3), and then the corresponding vectors of initial probabilities $\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$ are obtained by a simple rule given in Zucchini and MacDonald (2009), Sec. 4.2. In particular, we have that

$$\boldsymbol{\lambda} = (\boldsymbol{I}_{k_1} - \boldsymbol{\Lambda}' + \mathbf{1}_{k_1}\mathbf{1}'_{k_1})^{-1}\mathbf{1}_{k_1},$$

where $\mathbf{1}_{k_1}$ is a column vector k_1 ones and $\boldsymbol{I}_{k_1}$ is an identity matrix of the same dimension; in similar way we obtain $\boldsymbol{\pi}$ from $\boldsymbol{\Pi}$.

Acknowledgments

We acknowledge Laboratorio Riccardo Revelli, the Center for Employment Studies of Collegio Carlo Alberto in Torino (IT), who provided us the dataset analyzed in this paper; in particular, we thank Dr. Claudia Villosio for her helpful support. Francesco Bartolucci acknowledges the financial support from the grant RBFR12SHVV of the Italian Government (FIRB - Futuro in Ricerca - project "Mixture and latent variable models for causal inference and analysis of socio-economic data").

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.

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.

Avriel, M. (2003). *Nonlinear programming: analysis and methods*. Dover Publishing, New York.

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. (2014). Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. *Journal of the Royal Statistical Society, Series C*, 63:267–288.

Bartolucci, F., Colombi, R., and Forcina, A. (2007). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica*, 17:691–711.

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. (2013). *Latent Markov Models for Longitudinal Data*. Chapman and Hall/CRC Press, Boca Raton.

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., 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. E., 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.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236.

Bulla, J. and Berzel, A. (2008). Computational issues in parameter estimation for stationary hidden Markov models. *Computational Statistics*, 23:1–18.

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 Press, London.

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

Gao, X. and Song, P. X.-K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105:1531–1540.

Gao, X. and Song, P. X.-K. (2011). Composite likelihood EM algorithm with applications to multivariate hidden Markov model. *Statistica Sinica*, 21:165–185.

Goldstein, H. (2011). *Multilevel Statistical Models, 4th Edition*. John Wiley & Sons, Chichester, UK.

Hanfelt, J. J. (2004). Composite conditional likelihood for sparse clustered data. *Journal of the Royal Statistical Society, Series B*, 66:259–273.

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, Cambridge.

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

Khreich, W., Granger, E., Miri, A., and Sabourin, R. (2010). On the memory complexity of the forward-backward algorithm. *Pattern Recognition Letters*, 31:91–99.

Kuk, A. Y. C. and Nott, D. J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistics and Probability Letters*, 47:329–335.

Lagona, F., Maruotti, A., and Padovano, F. (2015). Multilevel multivariate modelling of legislative count data, with a hidden Markov chain. *Journal of the Royal Statistical Society, Series A*, (to appear).

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. G. (1982). Conditional score functions: Some optimality results. *Biometrika*, 69:503–512.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80:221–239.

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

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

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

Molenberghs, G. and Verbeke, G. (2004). Meaningful statistical model formulations for repeated measures. *Statistica Sinica*, 14:989–1020.

Muthén, B. and Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society, Series A*, 172:639–657.

Muthén, B. and Asparouhov, T. (2011). *LTA in Mplus: transition probabilities influenced by covariates*. Mplus Web Notes, No. 13.

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.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 39:577–591.

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

Skrondal, A. and Rabe-Hesketh, S. (2013). Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity. *Journal of the Royal Statistical Society, Series C*, 63:211–237.

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

Varin, C. (2008). On composite marginal likelihoods. *AStA*, 92:1–28.

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

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42.

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

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, 33:213–239.

Vermunt, J. K. and Magidson, J. (2005). *Latent GOLD 4.0 User's Guide*. Belmont, Massachusetts, Statistical Innovations Inc.

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

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC Press, Boca Raton, FL.

Table 1: *Values of CL-BIC computed using the Monte Carlo estimate of the score variance for different values of k_1 and k_2 (in boldface the largest CL-BIC value). For each model the penalization term is included (in brackets).*

	k_2			
k_1	1	2	3	4
1	-6398.94 (43.32)	-6158.62 (68.16)	-6100.29 (79.01)	-6085.09 (82.30)
2	-6211.61 (85.79)	-6076.73 (97.09)	-6069.87 (102.32)	-6078.90 (119.64)
3	-6179.11 (109.29)	-6072.93 (103.56)	**-6057.69** (107.21)	-6061.13 (119.62)
4	-6182.28 (120.29)	-6088.49 (123.32)	-6077.75 (139.41)	-6073.66 (142.19)

Table 2: *Estimates of the logistic regression parameters (collected in the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$) affecting the conditional probabilities under the model with $k_1 = k_2 = 1$, $k_1 = k_2 = 2$ and $k_1 = k_2 = 3$ latent states.*

	$k_1 = k_2 = 1$			$k_1 = k_2 = 2$			$k_1 = k_2 = 3$		
parameter	est.	se	p-value	est.	se	p-value	est.	se	p-value
intercept	-3.878	0.173	0.000	-7.169	0.461	0.000	-12.075	1.424	0.000
gender	0.096	0.094	0.245	0.155	0.152	0.246	0.269	0.191	0.170
age	0.019	0.005	0.003	0.039	0.011	0.003	0.038	0.009	0.001
age^2/100	-0.068	0.062	0.232	-0.050	0.141	0.371	-0.018	0.102	0.397
t	0.109	0.034	0.007	0.154	0.053	0.015	0.157	0.054	0.014
t^2	-0.002	0.003	0.352	-0.001	0.005	0.380	-0.002	0.005	0.376
size	0.006	0.011	0.348	0.006	0.017	0.369	0.012	0.017	0.311
area: North-East	0.205	0.117	0.113	0.555	0.215	0.028	0.508	0.221	0.048
area: Center	-0.077	0.205	0.367	-0.387	0.253	0.150	-0.523	0.301	0.115
area: South	0.020	0.208	0.408	-0.001	0.239	0.420	-0.083	0.258	0.375
area: Islands	-0.386	0.231	0.125	-0.386	0.268	0.165	-0.457	0.360	0.197
skill	0.947	0.151	0.000	1.843	0.201	0.000	2.059	0.246	0.000
income	-0.108	0.009	0.000	-0.150	0.013	0.000	-0.159	0.015	0.000
part-time	-0.788	0.146	0.000	-0.995	0.231	0.000	-1.057	0.207	0.000
lagged-response	1.542	0.094	0.000	0.291	0.096	0.011	0.168	0.104	0.135

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	-0.985	0.496	0.993	0.004	0.004
2	0.720	0.461	0.004	0.993	0.004
3	3.683	0.043	0.041	0.041	0.918

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	-5.067	0.135	0.928	0.036	0.036
2	0.369	0.698	0.007	0.986	0.007
3	2.569	0.166	0.029	0.029	0.941