CEIS Tor Vergata
Centre for International Studies on Economic Growth

CEIS Tor Vergata
RESEARCH PAPER SERIES

Working Paper No. 97 March 2007

Maximum likelihood estimation of an extended latent Markov model for clustered binary panel data

Francesco Bartolucci Valentina Nigro

CEIS Tor Vergata - Research Paper Series, Vol. 33, No. 97 , March 2007

Maximum likelihood estimation of an extended latent Markov model for clustered binary panel data

Francesco Bartolucci* and Valentina Nigro†

November 2006

Abstract

Computational aspects concerning a model for clustered binary panel data are analysed. The model is based on the representation of the behavior of a subject (individual panel member) in a given cluster by means of a latent process that is decomposed into a cluster-specific component, which follows a first-order Markov chain, and an individual-specific component, which is time-invariant and is represented by a discrete random variable. In particular, an algorithm for computing the joint distribution of the response variables is introduced. The algorithm may be used even in the presence of a large number of subjects in the same cluster. Also an Expectation-Maximization (EM) scheme for the maximum likelihood estimation of the model is described showing how the Fisher information matrix can be estimated on the basis of the numerical derivative of the score vector. The estimate of this matrix is used to compute standard errors for the parameter estimates and to check the identifiability of the model and the convergence of the EM algorithm. The approach is illustrated by means of an application to a dataset concerning Italian employees' illness benefits.

Keywords: EM algorithm; Finite mixture models; Heterogeneity; Latent class model; State dependence.

JEL classification codes: C23, C25, C51, C63.

*Dipartimento di Economia, Finanza e Statistica, Università di Perugia, 06123 Perugia, Italy, *e-mail:* bart@stat.unipg.it

†Dipartimento di Studi Economico-Finanziari e Metodi Quantitativi, Università di Roma "Tor Vergata", Via Columbia 2, 00133 Roma, Italy, *e-mail:* Valentina.Nigro@uniroma2.it

1 Introduction

Many econometric and statistical models are now available for the analysis of binary panel data (for a review see Hsiao (1986); Arellano and Honoré (2001); Langeheine and van de Pol (2002)). Among these models, it is worthwile mentioning those based on the assumption that the response variables have a simplified dependence structure given a latent process which may be discrete or continuous.

The first model based on a discrete latent process was introduced by Wiggins (1973). It may be seen as an extension of the latent class model (Lazarsfeld and Henry (1968); Goodman (1974)) in which the response variables are conditionally independent given a latent Markov (LM) chain. This model has been applied in many fields, especially in psychological and educational measurement (Langeheine *et al.* (1994); Vermunt *et al.* (1999)) and sociology (Van de Pol and Langeheine (1990); Mannan and Koval (2003)). Likelihood inference for the LM model was studied by Bartolucci (2006) who considered, in particular, the problem of testing linear hypotheses on the transition probabilities of the latent process. The model has also been extended in several directions. In particular, Van de Pol and Langeheine (1990) proposed the latent mixed Markov model. It is in practice a finite mixture of LM models which allows the parameters of the latent process to be different between subjects. Moreover, Vermunt *et al.* (1999) introduced a version of the LM model in which the initial and transition probabilities of the latent process depend on individual covariates.

Among the first authors dealing with models based on a continuous latent process, it is worthwhile mentioning Heckman and Willis (1977), Heckman (1981a) and Butler and Moffitt (1982). These models are often used in economic contexts as, for instance, in the analysis of labor market data (Hyslop (1999)), consumer choices (Chintagunta *et al.* (2001)) and debt repayments (Hajivassiliou and McFadden (1998)). Most of them are based on the assumption that there exists an AR(1) latent process given which the response variables satisfy a first-order Markovian dependence structure, so that state dependence (Heckman (1981b)) is properly taken into account together with available covariates. In economic contexts, state dependence is normally due to habit, risk aversion or transition costs.

All the models mentioned above are based on the assumption that the subjects (individual panel members) are independent of each other. In several situations, however, this assumption may not be realistic. We are referring, in particular, to situations in which the panel members are clustered according to specific criteria, such as residence in a certain region or being employed

in a certain firm. This obviously gives rise to a within-cluster correlation that, if ignored, may lead to less efficient estimators of the parameters, the standard errors of which may be strongly underestimated. A simple way to take this correlation into account could be to include in one of the models mentioned above a dummy explanatory variable for being in a certain cluster. This measure, however, is not completely satisfactory because we would add as many parameters as the number of clusters, making the inference unreliable in most situations. The number of parameters to add would be even larger if we want to assume that each cluster-specific effect has its own dynamics.

In this paper, we deal with an extended version of the LM model of Wiggins (1973) for the analysis of clustered binary panel data in which the latent process describing the behavior of a subject in a cluster is decomposed into a dynamic component common to all the subjects in the cluster and a time-invariant component for the presence of unobserved individual heterogeneity. The dynamic component is represented by a first-order Markov process, whereas, following Heckman and Singer (1984), the time-invariant component is represented by a discrete random variable. In this way, the correlation between the responses provided by different subjects in the same cluster is properly taken into account. The model also allows us to take into account state dependence and available covariates by means of a logit parameterization of the distribution of every response variable given the latent variables. For this model, we introduce an efficient algorithm for computing the joint distribution of the response variables and particular moments of the conditional distribution of the latent variables given the response variables. Note that direct evaluation of the joint distribution of the response variables requires a number of operations which grows exponentially with the sample size of the largest cluster, whereas the proposed algorithm has a numerical complexity which grows linearly with the overall sample size. As we show, the algorithm also allows us to implement an Expectation-Maximization (EM) scheme (Dempster *et al.* (1977)) for the maximum likelihood (ML) estimation of the parameters of the proposed model. We also deal with the estimation of the Fisher information matrix on the basis of the numerical derivative of the score vector. The latter may be computed by means of an algorithm related to that used to compute the joint distribution of response variables. On the basis of the estimated information matrix we can compute standard errors for the parameter estimates and check the identifiability of the model and the convergence of the EM algorithm. The algorithms described in this paper have been implemented in a series of MATLAB functions which are available from the web page `www.stat.unipg.it/~bart`.

To our knowledge, models with a structure similar to that of the model here discussed have not been previously considered. The literature on clustered binary panel data is in fact rather

scarce; one of the few contributions is due to Goldstein *et al.* (2000). Nevertheless, the issues of unobserved individual heterogeneity and cluster-specific dynamics apply to many macroeconomic settings. For instance, a potentially fruitful application of the model and estimation approach here proposed is the analysis of sovereign debt crises on which there is renewed interest (see Fuertes and Kalotychou, 2006).

To illustrate our approach, we discuss the analysis of a dataset on individual work histories derived from the administrative archives of the Italian National Institute of Social Security (INPS). The response variable of interest is a binary variable for an employee receiving illness benefits in a certain year. The analysis shows positive state dependence, strong persistence in the latent process and significant covariates.

The paper is organized as follows. In the next Section we introduce the basic notation and we describe the proposed extension of the LM model for clustered binary panel data. In Section 3 we illustrate the algorithm for computing the joint distribution of the response variables and of certain moments of the conditional distribution of the latent variables given the response variables. The EM algorithm for the ML estimation of the parameters is described in Section 4 where we also deal with the estimation of the information matrix. The application to labor market data is described in Section 5, whereas in Section 6 we draw the main conclusions.

2 The model

Let T denote the number of time periods, n denote the number of subjects in the panel and suppose that these subjects are clustered, according to some criteria, into H clusters of size $n_1, \ldots, n_H$ respectively. Also let y_{hit}, $h = 1, \ldots, H$, $i = 1, \ldots, n_h$, $t = 1, \ldots, T$, denote the binary response variable of interest for subject i in cluster h at time period t and $\boldsymbol{x}_{hit}$ denote the corresponding vector of fixed covariates, or equivalently the observed value of strictly exogenous covariates if these are random. The response variables referring to the same subject i in cluster h will be collected into the vector $\boldsymbol{y}_{hi} = \{y_{hit},\, t = 1, \ldots, T\}$, whereas the response variables referred to the same cluster h will be collected in the matrix $\boldsymbol{Y}_h = \{y_{hit},\, i = 1, \ldots, n_h,\, t = 1, \ldots, T\}$. Finally, the set of all the response variables will be denoted by $\mathcal{Y} = \{y_{hit},\, h = 1, \ldots, H,\, i = 1, \ldots, n_h,\, t = 1, \ldots, T\}$.

Following a standard econometric approach, we assume that every response variable y_{hit} is a discretized version of a latent continuous variable which is interpretable as a measure of the utility

or propensity to experience a certain situation. More precisely, we assume that

$$y_{hit} = \begin{cases} 1(\boldsymbol{x}'_{hit}\boldsymbol{\beta}_1 + \eta_{hi1} > 0), & t = 1, \\ 1(\boldsymbol{x}'_{hit}\boldsymbol{\beta}_2 + y_{hi,t-1}\gamma + \eta_{hit} > 0), & t = 2, \ldots, T, \end{cases}$$

where $1(\cdot)$ is the indicator function, which takes value 1 if its argument is true and 0 otherwise, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of regression coefficients for the observed covariates and γ is a regression coefficient for the lagged response variable, which therefore provides a measure of the true state dependence effect (Heckman (1981b)). Note that, in order to take into account the initial condition problem (Heckman, 1981c), we allow the regression coefficients for the first time period ($\boldsymbol{\beta}_1$) to be different from those for the following periods ($\boldsymbol{\beta}_2$). We also assume the following decomposition of the error terms:

$$\eta_{hit} = \alpha_{1ht} + \alpha_{2hi} + \varepsilon_{hit}, \qquad h = 1, \ldots, H, \quad i = 1, \ldots, n_h, \quad t = 1, \ldots, T,$$

where the random variables α_{1ht} capture the cluster-specific dynamic effects, the random variables α_{2hi} capture the unobserved heterogeneity between subjects and the random variables ε_{hit} represent residual error components. All these random variables are assumed to be mutually independent and independent of the covariates, with the exception of those in $\boldsymbol{\alpha}_{1h} = \{\alpha_{1ht},\ t = 1, \ldots, T\}$ which, for $h = 1, \ldots, H$, are assumed to follow a first-order Markov chain with states ϕ_c, $c = 1, \ldots, C$, initial probabilities π_c, $c = 1, \ldots, C$, and transition probabilities $\pi_{c|b}$, $b, c = 1, \ldots, C$. We also assume that each random variable α_{2hi}, $h = 1, \ldots, H$, $i = 1, \ldots, n_h$, has a discrete distribution with support points ψ_s, $s = 1, \ldots, S$, and probabilities ρ_s, $s = 1, \ldots, S$. Finally, we assume that the random variables ε_{hit}, $h = 1, \ldots, H$, $i = 1, \ldots, n_h$, $t = 1, \ldots, T$, have a standard logistic distribution.

Note that the error terms η_{hit} referred to the same subject i in cluster h are allowed to be dependent across time with a dynamics depending on the cluster to which the subject belongs. Also the error terms referred to different subjects but to the same time period t are allowed to be dependent when these subjects belong to the same cluster. In this way, our model takes into account the cluster effect beyond the effect of unobserved individual heterogeneity. On the other hand, the above assumptions imply that the error terms are independent of the covariates. In certain contexts, it may be worthwhile to relax this assumption. For instance, a correlated random effect specification in which the latent variables are specified as a function of the complete history of the covariates (Chamberlain (1984)) could be explored. However, this is beyond the scope of the present paper.

It is also interesting to note that the model based on the above assumptions can be seen as a hierarchical random effect model in which the intercept is a random parameter having a discrete

distribution. We in fact have three nested levels of data. At the lowest level we consider the response variable for every time period t and every subject hi. At the intermediate level we consider the complete response pattern of every subject hi; this pattern is affected by the time-invariant latent variable α_{2hi}. At the highest level we consider the response patterns of all the subjects in cluster h; these patterns are affected by the latent process $\boldsymbol{\alpha}_{1h}$, which allows us to take into account, for instance, unobserved economic shocks. It is worthwhile to emphasize the difference with respect to a conventional LM model based on an independent Markov chain for each subject. This model obviously ignores the within-cluster correlation, but allows the latent factors affecting the response variables to have their own dynamics at individual level. Nevertheless, the conventional LM model may be seen as a particular case of our model in which there is one subject in each cluster and the individual-specific random variables have only one support point. The choice between the two models may depend on the object of the study and on the data availability.

Now let $p(y_{hit}|c_{ht}, s_{hi}, y_{hi,t-1})$, with the last argument vanishing for $t = 1$, denote the conditional probability of y_{hit} given $\alpha_{1ht} = \phi_{c_{ht}}$, $\alpha_{2hi} = \psi_{s_{hi}}$ and the lagged response variable $y_{hi,t-1}$. Note that we indicate the specific support point of α_{1ht} on which we are conditioning by c_{ht}, with $c_{ht} = 1, \ldots, C$; similarly, we indicate a specific support point of α_{2hi} on which we are conditioning by s_{hi}, with $s_{hi} = 1, \ldots, S$. The assumptions made on the distribution of the error terms η_{hit} imply that

$$p(y_{hit}|c_{ht}, s_{hi}, y_{hi,t-1}) = \begin{cases} \dfrac{\exp[y_{hi1}(\phi_{c_{ht}} + \psi_{s_{hi}} + \boldsymbol{x}'_{hi1}\boldsymbol{\beta}_1)]}{1 + \exp(\phi_{c_{ht}} + \psi_{s_{hi}} + \boldsymbol{x}'_{hi1}\boldsymbol{\beta}_1)}, & t = 1, \\ \dfrac{\exp[y_{hit}(\phi_{c_{ht}} + \psi_{s_{hi}} + \boldsymbol{x}'_{hit}\boldsymbol{\beta}_2 + y_{hi,t-1}\gamma)]}{1 + \exp(\phi_{c_{ht}} + \psi_{s_{hi}} + \boldsymbol{x}'_{hit}\boldsymbol{\beta}_2 + y_{hi,t-1}\gamma)}, & t = 2, \ldots, T. \end{cases}$$

Our assumptions also imply that for every h and i, the conditional probability of $\boldsymbol{y}_{hi}$ given $\alpha_{1ht} = \phi_{c_{ht}}$, $t = 1, \ldots, T$, and $\alpha_{2hi} = \psi_{s_{hi}}$ may be expressed as

$$p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi}) = \prod_t p(y_{hit}|c_{ht}, s_{hi}, y_{i,t-1}),$$

with $\boldsymbol{c}_h = \{c_{ht},\ t = 1, \ldots, T\}$. Moreover, for every h, the distribution of $\boldsymbol{Y}_h$ may be expressed as

$$p(\boldsymbol{Y}_h) = \sum_{\boldsymbol{c}_h} \sum_{\boldsymbol{s}_h} p(\boldsymbol{c}_h) p(\boldsymbol{s}_h) \prod_i p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi}), \tag{1}$$

where the first sum is extended to all the possible configurations of $\boldsymbol{c}_h$ and the second to all the possible configurations of $\boldsymbol{s}_h = \{s_{hi},\ i = 1, \ldots, n_h\}$, whereas

$$p(\boldsymbol{c}_h) \quad = \quad \pi_{c_{h1}} \prod_{t>1} \pi_{c_{ht}|c_{h,t-1}}, \tag{2}$$

and

$$p(\boldsymbol{s}_h) \quad = \quad \prod_i \rho_{s_{hi}}. \tag{3}$$

Finally, since the random matrices $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_H$ are mutually independent, we can write

$$p(\mathcal{Y}) = \prod_h p(\boldsymbol{Y}_h). \tag{4}$$

Direct evaluation of (1) requires a number of operations which increases exponentially with the number of time periods (T) and the size of the largest cluster (n_h). Although in a typical panel study T is small, n_h may be large and hence evaluation of (1), and then that of (4), may be infeasible.

3 Efficient computation of joint and conditional distributions

We now introduce an algorithm which allows us to compute efficiently $p(\boldsymbol{Y}_h)$ for each cluster h, and then $p(\mathcal{Y})$, along with certain conditional probabilities which are required for the EM algorithm that will be outlined in the next Section. This algorithm exploits the fact that for every h, the random vectors $\boldsymbol{y}_{h1}, \ldots, \boldsymbol{y}_{hn_h}$ are conditionally independent given $\boldsymbol{\alpha}_{1h}$. Therefore, we can write

$$p(\boldsymbol{c}_h, \boldsymbol{Y}_h) = p(\boldsymbol{c}_h) \prod_i p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h), \tag{5}$$

where

$$p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h) = \sum_{s_{hi}} p(s_{hi}) p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi}), \tag{6}$$

and then

$$p(\boldsymbol{Y}_h) = \sum_{\boldsymbol{c}_h} p(\boldsymbol{c}_h, \boldsymbol{Y}_h).$$

Once $p(\boldsymbol{Y}_h)$ is computed for every cluster h, $p(\mathcal{Y})$ can be obtained by using (4). Note, in this case, that the number of operations grows linearly with the sample size.

The above rules may be simply represented by using matrix notation. So let $\boldsymbol{M}_{hi}$ be a $C^T \times S$ matrix with elements $p(s_{hi})p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi})$ arranged by letting $\boldsymbol{c}_h$ run by row in lexicographical order and s_{hi} by column. Then the column vector $\boldsymbol{p}_h$ with elements $p(\boldsymbol{c}_h, \boldsymbol{Y}_h)$ for every $\boldsymbol{c}_h$ may be obtained as

$$\boldsymbol{p}_h = \mathrm{diag}(\boldsymbol{q}) \prod_i \boldsymbol{m}_{hi}, \qquad \boldsymbol{m}_{hi} = \boldsymbol{M}_{hi} \mathbf{1}_S, \tag{7}$$

where $\boldsymbol{q}$ is a column vector with elements $p(\boldsymbol{c}_h)$, the product $\prod_i$ is elementwise and $\mathbf{1}_S$ denotes a column vector of S ones. Finally, we have that

$$p(\boldsymbol{Y}_h) = \boldsymbol{p}_h' \mathbf{1}_{C^T}.$$

For the EM algorithm, we need to compute the conditional probability, given $\boldsymbol{Y}_h$, of $(\alpha_{1h,t-1} = \phi_{c_{h,t-1}}, \alpha_{1ht} = \phi_{c_{ht}})$, for $h = 1, \ldots, H$ and $t = 2, \ldots, T$, and of $(\alpha_{1ht} = \phi_{c_{ht}}, \alpha_{2hi} = \psi_{s_{hi}})$, for $h = 1, \ldots, H$, $i = 1, \ldots, n_h$ and $t = 1, \ldots, T$. The probabilities of the first type may be expressed as

$$p(c_{h,t-1}, c_{ht}|\boldsymbol{Y}_h) = \sum_{\boldsymbol{c}^-_{h,t-1,t}} \frac{p(\boldsymbol{c}_h, \boldsymbol{Y}_h)}{p(\boldsymbol{Y}_h)}, \tag{8}$$

where $\boldsymbol{c}^-_{h,t-1,t}$ denotes the subvector of $\boldsymbol{c}_h$ without the elements $c_{h,t-1}$ and c_{ht}. Those of second type, instead, may be expressed as

$$p(c_{ht}, s_{hi}|\boldsymbol{Y}_h) = \sum_{\boldsymbol{c}^-_{ht}} \frac{p(\boldsymbol{c}_h, s_{hi}, \boldsymbol{Y}_h)}{p(\boldsymbol{Y}_h)}, \tag{9}$$

where $\boldsymbol{c}^-_{ht}$ denotes the subvector of $\boldsymbol{c}_h$ without the element c_{ht} and, because of (5) and (6),

$$p(\boldsymbol{c}_h, s_{hi}, \boldsymbol{Y}_h) = p(\boldsymbol{c}_h, \boldsymbol{Y}_h) \frac{p(s_{hi}) p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi})}{p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h)}.$$

Also (8) and (9) may be implemented by using matrix notation. In particular, the C^2-dimensional vector $\boldsymbol{r}_{ht}$ with elements $p(c_{h,t-1}, c_{ht}|\boldsymbol{Y}_h)$, arranged by letting $(c_{h,t-1}, c_{ht})$ run in lexicographical order, may be obtained as

$$\boldsymbol{r}_{ht} = \boldsymbol{G}_t \boldsymbol{p}_h / p(\boldsymbol{Y}_h), \tag{10}$$

where $\boldsymbol{G}_t$ is an aggregation matrix of dimension $C^2 \times C^T$ defined as

$$\boldsymbol{G}_t = \bigotimes_{j=1}^{T} \boldsymbol{G}_{jt}, \qquad \text{with} \qquad \boldsymbol{G}_{jt} = \begin{cases} \boldsymbol{I}_C & \text{if } j = t-1, t \\ \boldsymbol{1}'_C & \text{otherwise} \end{cases},$$

and $\boldsymbol{I}_C$ denoting an identity matrix of dimension C. The $C \times S$ matrix $\boldsymbol{E}_{hit}$ with elements $p(c_{ht}, s_{hi}|\boldsymbol{Y}_h)$, arranged by letting c_{ht} run by row and s_{hi} by column, may be computed as

$$\boldsymbol{E}_{hit} = \boldsymbol{R}_t \boldsymbol{D}_{hi} / p(\boldsymbol{Y}_h), \tag{11}$$

where

$$\boldsymbol{D}_{hi} = \text{diag}(\boldsymbol{p}_h) \text{diag}(\boldsymbol{M}_{hi} \boldsymbol{1}_S)^{-1} \boldsymbol{M}_{hi},$$

is a matrix with elements $p(\boldsymbol{c}_h, s_{hi}, \boldsymbol{Y}_h)$ arranged as in $\boldsymbol{M}_{hi}$ and $\boldsymbol{R}_t$ is an aggregation matrix of dimension $C \times C^T$ defined as

$$\boldsymbol{R}_t = \bigotimes_{j=1}^{T} \boldsymbol{R}_{jt}, \qquad \text{with} \qquad \boldsymbol{R}_{jt} = \begin{cases} \boldsymbol{I}_C & \text{if } j = t \\ \boldsymbol{1}'_C & \text{otherwise} \end{cases}.$$

From the matrix $\boldsymbol{E}_{hit}$ we can obtain the C-dimensional vector containing the probabilities $p(c_{ht}|\boldsymbol{Y}_h)$ simply as $\boldsymbol{E}_{hit} \boldsymbol{1}_S$. These probabilities will also be used in the EM algorithm illustrated in the following section.

It is clear from the above description that the proposed algorithm has a numerical complexity that increases linearly with n, but exponentially with T. This is because it considers all the C^T realizations of the cluster-specific latent Markov chain. In certain situations, it may then be useful to apply a stochastic version of the algorithm which may be described as follows. Let $\boldsymbol{C}$ denote an $R \times T$ matrix, any row of which corresponds to a randomly drawn realization $\boldsymbol{c}_h$ of the latent Markov chain, and $\boldsymbol{M}_{hi}$ be the corresponding $R \times S$ matrix with elements $p(s_{hi})p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi})$. Then $p(\boldsymbol{Y}_h)$ may be computed as $\boldsymbol{p}_h'\boldsymbol{1}_R$ with $\boldsymbol{p}_h$ defined as in (7) with $\boldsymbol{q} = \boldsymbol{1}_R/R$. Moreover, the conditional probabilities $p(c_{h,t-1}, c_{ht}|\boldsymbol{Y}_h)$ may still be computed on the basis of (10), whereas $p(c_{ht}, s_{hi}|\boldsymbol{Y}_h)$ and $p(c_{ht}|\boldsymbol{Y}_h)$ may be computed on the basis of (11), with the aggregation matrices $\boldsymbol{G}_t$ and $\boldsymbol{R}_t$ suitably defined on the basis of $\boldsymbol{C}$.

As a final comment, note that when the size of a cluster, say the h-th, is large, the corresponding probability $p(\boldsymbol{Y}_h)$ could take extreme values. To avoid this problem, we can multiply each probability $p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi})$ by a suitable constant a. In this way, we obtain the probabilities $p^*(\boldsymbol{c}_h, s_{hi}, \boldsymbol{Y}_h)$, $p^*(\boldsymbol{c}_h, \boldsymbol{Y}_h)$ and $p^*(\mathcal{Y})$ which are equal, respectively, to $a^{n_h}p(\boldsymbol{c}_h, s_{hi}, \boldsymbol{Y}_h)$, $a^{n_h}p(\boldsymbol{c}_h, \boldsymbol{Y}_h)$ and $a^{n_h}p(\mathcal{Y})$. These, however, may be still used in (8) and (9) to obtain $p(c_{h,t-1}, c_{ht}|\boldsymbol{Y}_h)$ and $p(c_{h,t-1}, s_{hi}|\boldsymbol{Y}_h)$.

4 Maximum likelihood estimation

The log-likelihood of the model defined in Section 2 may be simply expressed as

$$\ell(\boldsymbol{\theta}) = \log[p(\mathcal{Y})],$$

where $p(\mathcal{Y})$ is the joint probability of the observed response variables computed as a function of the vector of all the identifiable parameters. This vector may be expressed as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \boldsymbol{\theta}_3')'$, where $\boldsymbol{\theta}_1$ contains the identifiability parameters of the conditional distribution of the response variables given the latent variables, $\boldsymbol{\theta}_2$ contains those of the distribution of the cluster-specific latent processes and $\boldsymbol{\theta}_3$ those of the distribution of the individual-specific latent variables. In particular:

- in order to take the intercept into account, the first element of each vector of covariates $\boldsymbol{x}_{hit}$ is equal to 1. To make the model identifiable, we then constrain the first support point of each random variable α_{1ht} and α_{2hi} to be equal to 0, i.e. $\phi_1 = 0$ and $\psi_1 = 0$, and only the remaining support points enter the vector $\boldsymbol{\theta}_1$ which, consequently, is defined as $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \gamma, \boldsymbol{\phi}', \boldsymbol{\psi}')'$, with $\boldsymbol{\phi} = \{\phi_c,\ c = 2, \ldots, C\}$ and $\boldsymbol{\psi} = \{\psi_s,\ s = 2, \ldots, S\}$;
- for the parameters of the distribution of the cluster-specific latent processes we have $\sum_c \pi_c = 1$ and $\sum_c \pi_{c|b} = 1$ for every b. We then exclude from the vector $\boldsymbol{\theta}_2$ the first initial probability,

π_1, and the first transition probability $\pi_{1|b}$ for $b = 1, \ldots, C$. This vector is therefore given by $\boldsymbol{\theta}_2 = (\boldsymbol{\pi}', \boldsymbol{\mu}')'$, with $\boldsymbol{\pi} = \{\pi_c,\ c = 2, \ldots, C\}$ and $\boldsymbol{\mu} = \{\pi_{c|b},\ b = 1, \ldots, C,\ c = 2, \ldots, C\}$;

- taking into account the constraint $\sum_s \rho_s = 1$, we have that $\boldsymbol{\theta}_3 = \boldsymbol{\rho}$, with $\boldsymbol{\rho} = \{\rho_s,\ s = 2, \ldots, S\}$.

In what follows, we show how, in order to estimate $\boldsymbol{\theta}$, we can maximize $\ell(\boldsymbol{\theta})$ by means of an EM algorithm. When dealing with latent variable models, this algorithm has the advantage of being simpler to implement and more stable with respect to more direct maximization algorithms, such as the Newton-Raphson.

4.1 The EM algorithm

The EM algorithm is based on the concept of complete data which, in the present context, may be represented by $(\boldsymbol{c}_h, \boldsymbol{s}_h, \boldsymbol{Y}_h)$, $h = 1, \ldots, H$, where $\boldsymbol{c}_h$ stands for a specific realization of the latent process $\boldsymbol{\alpha}_{1h}$ and $\boldsymbol{s}_h$ for a specific realization of the vector of latent variables $\boldsymbol{\alpha}_{2h} = \{\alpha_{2hi},\ i = 1, \ldots, n_h\}$ and $\boldsymbol{Y}_h$ represents the observed value of all the response variables referred to the same cluster h. Note that $(\boldsymbol{c}_h, \boldsymbol{s}_h)$, $h = 1, \ldots, H$, can not be observed and so they are considered as missing data. If these data were known, we could compute the *complete data log-likelihood*

$$\ell^*(\boldsymbol{\theta}) = \sum_h \log[p(\boldsymbol{Y}_h|\boldsymbol{c}_h, \boldsymbol{s}_h)p(\boldsymbol{c}_h)p(\boldsymbol{s}_h)] = \sum_h \sum_{\boldsymbol{c}} \sum_{\boldsymbol{s}} g_h(\boldsymbol{c}, \boldsymbol{s}) \log[p(\boldsymbol{Y}_h|\boldsymbol{c}, \boldsymbol{s})p(\boldsymbol{c})p(\boldsymbol{s})],$$

where $g_h(\boldsymbol{c}, \boldsymbol{s}) = 1(\boldsymbol{c}_h = \boldsymbol{c}, \boldsymbol{s}_h = \boldsymbol{s})$. Considering (1), (2) and (3), and after some algebra, the above log-likelihood may be expressed as

$$\ell^*(\boldsymbol{\theta}) = \ell_1^*(\boldsymbol{\theta}_1) + \ell_2^*(\boldsymbol{\theta}_2) + \ell_3^*(\boldsymbol{\theta}_3),$$

with

$$\begin{aligned}
\ell^*(\boldsymbol{\theta}_1) &= \sum_h \sum_i \sum_t \sum_c \sum_s w_{hit}(c, s) \log[p(y_{hit}|c, s, y_{i,t-1})], \\
\ell^*(\boldsymbol{\theta}_2) &= \sum_h \left[\sum_c w_h(c) \log(\pi_c) + \sum_b \sum_c \sum_{t>1} w_{ht}(b, c) \log(\pi_{c|b}) \right],
\end{aligned}$$

and

$$\ell^*(\boldsymbol{\theta}_3) = \sum_h \sum_i \sum_s z_{hi}(s) \log(\rho_s),$$

where $w_{hit}(c, s) = 1(c_{ht} = c, s_{hi} = s)$, $w_h(c) = 1(c_{h1} = c)$, $w_{ht}(b, c) = 1(c_{h,t-1} = b, c_{ht} = c)$ and $z_{hi}(s) = 1(s_{hi} = s)$. Since we do not know the true value of these dummy variables, at any iteration

of the EM algorithm they are replaced by suitable expected values; the resulting function is then maximized to update the parameter vector $\boldsymbol{\theta}$. More precisely, the EM algorithm alternates the following steps until convergence:

- **E-step**: compute the conditional expected value of the dummy variables listed above given the observed data $\mathcal{Y}$ and the current value of $\boldsymbol{\theta}$. Note that these expected values are equal to

$$\begin{aligned}
\tilde{w}_{hit}(c,s) &= E[w_{hit}(c,s)|\mathcal{Y}] = p(\alpha_{1ht}=\phi_c, \alpha_{2hi}=\psi_s|\boldsymbol{Y}_h), \\
\tilde{w}_{h}(c) &= E[w_{h}(c)|\mathcal{Y}] = p(\alpha_{1h1}=\phi_c|\boldsymbol{Y}_h), \\
\tilde{w}_{ht}(b,c) &= E[w_{ht}(b,c)|\mathcal{Y}] = p(\alpha_{1h,t-1}=\phi_b, \alpha_{1ht}=\phi_c|\boldsymbol{Y}_h),
\end{aligned}$$

  and

$$\tilde{z}_{hi}(s) = E[z_{hi}(s)|\mathcal{Y}] = p(\alpha_{2hi}=\psi_s|\boldsymbol{Y}_h),$$

  and therefore may be computed by using the algorithm described in Section 3. The complete data log-likelihood, with the dummy variables substituted by the above expected values, corresponds to the conditional expected value of $\ell^*(\boldsymbol{\theta})$ given the observed data, which will be denoted by $\tilde{\ell}^*(\boldsymbol{\theta}) = \tilde{\ell}_1^*(\boldsymbol{\theta}_1) + \tilde{\ell}_2^*(\boldsymbol{\theta}_2) + \tilde{\ell}_3^*(\boldsymbol{\theta}_3)$.

- **M-step**: maximize $\tilde{\ell}^*(\boldsymbol{\theta})$ by maximizing separately its components as follows:

  - **Maximization of** $\tilde{\ell}_1^*(\boldsymbol{\theta}_1)$: this may be performed by means of a Newton-Raphson algorithm. To implement this algorithm, we need the first derivative vector and the second derivative matrix of $\tilde{\ell}_1^*(\boldsymbol{\theta}_1)$. Using matrix notation, these derivatives may be expressed, respectively, as follows

$$\boldsymbol{s}_1^*(\boldsymbol{\theta}_1) = \sum_h \sum_i \sum_t \sum_c \sum_s \tilde{w}_{hit}(c,s)[y_{hit} - q_{hit}(c,s)]\boldsymbol{u}_{hit}(s,c),$$

    and

$$\boldsymbol{H}_1^*(\boldsymbol{\theta}_1) = -\sum_h \sum_i \sum_t \sum_c \sum_s \tilde{w}_{hit}(c,s) q_{hit}(c,s)[1-q_{hit}(c,s)]\boldsymbol{u}_{hit}(c,s)\boldsymbol{u}_{hit}(c,s)',$$

    where $q_{hit}(c,s,y_{hi,t-1}) = p(y_{hit}=1|\alpha_{1ht}=\phi_c, \alpha_{2hi}=\psi_s, y_{hi,t-1})$, with the last argument vanishing for $t=1$, and $\boldsymbol{u}_{hit}(s,c)$ is a column vector of the same dimension as $\boldsymbol{\theta}_1$ such that

$$\log \frac{q_{hit}(c,s,y_{hi,t-1})}{1-q_{hit}(c,s,y_{hi,t-1})} = \boldsymbol{u}_{hit}(c,s)'\boldsymbol{\theta}_1,$$

    for $h=1,\ldots,H$, $i=1,\ldots,n_h$ and $t=1,\ldots,T$. Obviously, $\boldsymbol{u}_{hit}(c,s)$ contains the vector of covariates $\boldsymbol{x}_{hit}$ together with suitable dummies for the latent variables.

– **Maximization of** $\tilde{\ell}_2^*(\boldsymbol{\theta}_2)$: in this case we have an explicit solution given by

$$\pi_c = \frac{1}{H}\sum_h \tilde{w}_h(c), \qquad c = 2, \ldots, C,$$

and

$$\pi_{c|b} = \sum_h \sum_{t>1} \tilde{w}_{ht}(b,c) \Big/ \sum_h \sum_{t>1} \sum_j \tilde{w}_{ht}(b,j), \qquad b = 1, \ldots, C, \quad c = 2, \ldots, C.$$

– **Maximization of** $\tilde{\ell}_3^*(\boldsymbol{\theta}_3)$: also in this case we have an explicit solution given by

$$\rho_s = \frac{1}{n}\sum_h \sum_i \tilde{z}_{hi}(s), \qquad s = 2, \ldots, S.$$

A crucial point concerns the initialization of the algorithm. Let $\tilde{\boldsymbol{\beta}}_1$, $\tilde{\boldsymbol{\beta}}_2$ and $\tilde{\gamma}$ denote the regression parameters obtained from a standard logistic model applied to the observed data. Once a suitable constant τ (e.g. 1) has been chosen, we suggest to use $\tilde{\boldsymbol{\beta}}_1 - (\tau[(C+S)/2-1], 0, \cdots, 0)'$ and $\tilde{\boldsymbol{\beta}}_2 - (\tau[(C+S)/2-1], 0, \cdots, 0)'$ as starting values for, respectively, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and $\tilde{\gamma}$ as starting value for γ. We then suggest to take $\{\tau(c-1),\ c = 2, \ldots, C\}$ as starting value for the vector $\boldsymbol{\phi}$ and $\{\tau(s-1),\ s = 2, \ldots, S\}$ for the vector $\boldsymbol{\psi}$. Finally, the initial values of the probability vectors $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ may be chosen as $\mathbf{1}_{C-1}/C$ and $\mathbf{1}_{S-1}/S$, respectively, whereas the transition probabilities $\pi_{c|b}$, collected in the vector $\boldsymbol{\mu}$, may be set equal to $1-\lambda$ if $b = c$ and to $\lambda/(C-1)$ otherwise, where λ is a suitable constant between 0 and 1 (e.g. 0.25).

Since the log-likelihood $\ell(\boldsymbol{\theta})$ may have more local maxima, it is convenient to try different starting values for the EM algorithm. These may be chosen by randomly perturbating the starting values defined above. In this case, we take the value of $\boldsymbol{\theta}$ at which we have the highest level of $\ell(\boldsymbol{\theta})$ as the ML estimate of this parameter vector. It will be denoted by $\hat{\boldsymbol{\theta}}$.

Finally consider that when C^T is too large, it may be convenient to perform the E-step on the basis of the stochastic version of the algorithm described at the end of Section 3. We implemented the resulting version of the EM algorithm in a way which makes use of only one set of realizations of the latent Markov chain; these realizations are drawn just at the beginning of the algorithm. On the basis of a small simulation study, the results of which are not reported here, we verified that the parameter estimates produced by this stochastic EM algorithm are normally very close to those produced by the non stochastic algorithm. As we may expect, however, the mean square error of the ML estimator is larger when it is based on the stochastic EM algorithm. By simulation we also studied how the efficiency of the estimator varies with the ratio between the average number of subjects per cluster and the overall sample size. We noticed that, provided that the number of clusters does not become too small, the efficiency tends to increase with this ratio. In fact, when

the number of clusters is too small, it happens that there are just a few realizations of the latent Markov chain and hence there is little information on the corresponding parameters.

4.2 Information matrix estimation

Using the notation defined in Section 3, we can compute the first derivative vector of $p(\boldsymbol{Y}_h)$ as

$$\frac{\partial p(\boldsymbol{Y}_h)}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{p}_h'}{\partial \boldsymbol{\theta}} \mathbf{1}_{C^T},$$

with

$$\frac{\partial \boldsymbol{p}_h'}{\partial \boldsymbol{\theta}} = \left\{ \frac{\partial \boldsymbol{q}'}{\partial \boldsymbol{\theta}} + \left[\sum_i \frac{\partial \boldsymbol{m}_{hi}'}{\partial \boldsymbol{\theta}} \text{diag}(\boldsymbol{m}_{hi})^{-1} \right] \text{diag}(\boldsymbol{q}) \right\} \text{diag}\left(\prod_i \boldsymbol{m}_{hi} \right). \tag{12}$$

From this derivative, we can compute the score vector as

$$\boldsymbol{s}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_h \frac{1}{p(\boldsymbol{Y}_h)} \frac{\partial p(\boldsymbol{Y}_h)}{\partial \boldsymbol{\theta}}.$$

How to compute the derivatives of $\boldsymbol{q}$ and $\boldsymbol{m}_{hi}$ used in (12), is shown in an Appendix.

We estimate the Fisher information matrix of the model as minus the numerical derivative of the score vector. This matrix is denoted by $\hat{\boldsymbol{F}}(\boldsymbol{\theta})$. In particular, on the basis of $\hat{\boldsymbol{F}}(\hat{\boldsymbol{\theta}})$, i.e. the estimate of the information matrix at $\hat{\boldsymbol{\theta}}$, we can check if the model is locally identifiable and compute standard errors for the parameter estimates. More precisely, the model is considered locally identifiable if $\hat{\boldsymbol{F}}(\hat{\boldsymbol{\theta}})$ is of full rank (see also Rothenberg (1971)), whereas the standard error for an element of $\hat{\boldsymbol{\theta}}$ is computed as the square root of the corresponding diagonal element of $\hat{\boldsymbol{F}}(\hat{\boldsymbol{\theta}})^{-1}$.

Finally, the matrix $\hat{\boldsymbol{F}}(\boldsymbol{\theta})$ may also be used to check the convergence of the EM algorithm illustrated above. In particular, if the largest absolute value of $\hat{\boldsymbol{F}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{s}(\hat{\boldsymbol{\theta}})$ is smaller than a suitable tolerance level (e.g. 10^{-6}), we conclude that the algorithm has stopped close enough to the true maximum of $\ell(\boldsymbol{\theta})$. Note that this criterion is more reliable than that based on the difference between the values of $\ell(\boldsymbol{\theta})$ corresponding to two consecutive EM iterations, since the latter does not take the curvature of the log-likelihood into proper account. For a discussion on this point see, among others, McLachlan and Peel (2000, Sec. 2.14).

5 An application

We applied our model to a dataset extracted from the administrative INPS archives, downloadable from `www.laboratoriorevelli.it/whip/`. The dataset concerns 3341 employees (both blue-collar and white-collar) from private Italian firms with at least 1000 workers, aged between 18 and 60 in

1994. These employees were followed for 6 years from 1994 to 1999. Only those who continuously worked in the same firm have been considered in our analysis.

The response variable of interest is *illness*. This is a binary variable equal to 1 if the employee to which it is referred has received illness benefits in a certain year and to 0 otherwise. In the Italian system, illness benefits are given for at most 180 days a year and correspond to the 50% of the daily mean earnings for the first 20 days and to the 66.66% for the following days. As covariates we used *age* in 1994, *age squared*, *income* (total annual compensation in thousands of Euros), the dummies *area* (indicating one of the 5 Italian zones where the employee works: North-West, North-East, Center, South or Islands), *sex* (equal to 1 for a woman), *skill* (equal to 1 for a blue-collar), *ptime* (equal to 1 for an employee with a part-time job) and 4 temporal dummies for the years 1996 to 1999.

In Tables 1 and 2 we present some summary statistics of the dataset considered in our analysis. Table 1, in particular, reports the observed average of the response variable for each year between 1994 and 1999, together with its standard deviation and the standard deviation between firms. Note that the average corresponds to the frequency of subjects that in a certain year received illness benefits. It may be observed that this frequency has a positive trend and that the between firms variability represents a consistent part of the total variability. This suggests that the information about the firm to which an employee belongs is relevant for our analysis.

	Year					
	1994	1995	1996	1997	1998	1999
average	0.128	0.143	0.154	0.162	0.167	0.171
standard deviation (total)	0.334	0.350	0.361	0.369	0.373	0.377
standard deviation (between firms)	0.164	0.174	0.179	0.184	0.194	0.197

Table 1: *Summary statistics for the response variable referred to each year from 1994 to 1999.*

In Table 2 we report the average and the standard deviation of the covariates for the overall sample and for specific subsamples. The third column, in particular, concerns the subsample of employees who never received illness benefits, the forth column concerns employees who received illness benefits in each year and the remaining columns concern employees with one or more transitions from not receiving illness benefits to receiving illness benefits and vice versa. These results suggest that observable characteristics represent an important source of variability of the response variable. In particular, the frequency of white-collars among the employees who never received illness benefits is higher than in the overall sample and these employees have a higher income. Employees who always received illness benefits are older with respect to the overall sample, are more

likely to be women and blue-collars, work mostly in the North-West and in the South, have lower income and have less frequently a part-time position. Employees who experienced one transition to illness are older, have lower income, are more likely to be blue-collars and to work in the South. A similar profile may be observed for those who experienced a single transition from illness, but these are younger. Finally, employees with multiple transitions are younger than in the overall sample, more frequently are blue-collars, have lower income and more frequently have a part-time position.

	full sample	never illness	always illness	single transition to illness	single transition from illness	multiple transitions
sex	0.26	0.26	0.40	0.22	0.26	0.26
	(0.44)	(0.44)	(0.49)	(0.42)	(0.44)	(0.44)
age	38.24	38.24	39.81	40.03	37.92	37.72
	(8.67)	(8.35)	(8.36)	(9.23)	(9.71)	(9.16)
area: North-West	0.39	0.38	0.51	0.34	0.30	0.43
	(0.49)	(0.48)	(0.50)	(0.47)	(0.46)	(0.50)
area: North-East	0.15	0.14	0.17	0.17	0.16	0.19
	(0.36)	(0.35)	(0.38)	(0.38)	(0.37)	(0.39)
area: Center	0.24	0.29	0.10	0.13	0.26	0.17
	(0.43)	(0.45)	(0.30)	(0.34)	(0.44)	(0.37)
area: South	0.16	0.14	0.22	0.31	0.22	0.16
	(0.37)	(0.35)	(0.42)	(0.46)	(0.42)	(0.37)
area: Islands	0.06	0.06	0.00	0.05	0.06	0.06
	(0.23)	(0.23)	(0.00)	(0.23)	(0.23)	(0.23)
skill	0.45	0.24	0.97	0.85	0.77	0.81
	(0.49)	(0.42)	(0.17)	(0.35)	(0.41)	(0.38)
income	17.07	19.47	11.20	13.18	12.61	13.00
	(6.68)	(6.82)	(2.40)	(3.41)	(3.65)	(3.88)
ptime (at least one year)	0.08	0.07	0.04	0.08	0.14	0.12
	(0.28)	(0.25)	(0.20)	(0.28)	(0.35)	(0.32)
illness (number of years)	0.93	0.00	6.00	2.45	2.16	2.32
	(1.53)	(0.00)	(0.00)	(1.47)	(1.44)	(1.24)
number of observations	3341	2123	72	202	105	839

Table 2: *Average and standard deviation (in parentheses) of the observed covariates for certain subsamples of employees.*

To analyse the dataset described above, we used the extended LM model illustrated in Section 2. In applying this model, we considered the employees clustered according to the firm where they work. Note that the dataset do not include covariates for firms characteristics, such as quality environment or safety measures, which strongly affect the employees' health status. For this reason,

we relied on a latent variable process to take into account the variability of the response variable induced by being employed in different firms. As we already observed on the basis of results displayed in Table 1, this source of variability is expected to be relevant. On the other hand, the individual-specific latent variables may capture the propensity to get ill of every subject which is not explained by the observed covariates. The latter is assumed to be time-invariant. We recall that, in our approach, these latent variables are assumed to be independent of the individual covariates. This is perhaps a strong assumption, but this is a common drawback of most random effect models.

In the first step of our analysis we fitted the model with different levels of C and S. We recall that C indicates the number of states of the cluster-specific latent processes and S indicates the number of support points of the individual-specific latent variables. Table 3 shows the results obtained from this preliminary analysis in terms of the Bayesian Information Criterion (BIC; Schwarz, 1978). Note that, since the maximum number of subjects in the same cluster is 216, direct evaluation of $p(\mathcal{Y})$ would not be feasible. By using the algorithm illustrated in this paper, instead, we can efficiently compute this probability until $C = S = 4$ and obtain the corresponding parameter estimates as described in Section 4.1.

	S			
C	1	2	3	4
1	12,699	12,330	12,237	12,245
2	12,217	11,911	11,909	11,802
3	11,830	11,628	**11,604**	11,611
4	11,855	11,670	11,643	11,652

Table 3: *BIC index for different levels of C and S. Figures in bold are referred to the model with the smallest BIC index.*

On the basis of the results reported in Table 3, we chose $C = 3$ and $S = 3$; this choice, in fact, corresponds to the smallest BIC index. The estimates of the most relevant regression parameters, contained in $\boldsymbol{\beta}_2$, are displayed in Table 4, together with the corresponding standard errors, t-statistics and p-values. The signs of the estimates are consistent with the summary statistics displayed in Table 2. In particular, the probability of receiving illness benefits is positively related to being a *blue-collar*, whereas it is negatively related to the level of *income* and to having a *part-time* job. The effect of *sex*, *age* and *age squared* are not significant because the subsample of subjects who are always ill, though having a higher proportion of women than the overall sample, has a small size; moreover, age does not present a big variation between the subsamples. More surprisingly, *area* does not seem to be very significant. Finally, the coefficient measuring the *state*

dependence effect is significantly different from zero and presents a positive sign.

Covariate	estimate	standard error	t-statistics	p-value
constant	-12.509	0.861	-14.526	0.000
sex	0.171	0.112	1.527	0.127
age	-0.010	0.036	-0.278	0.781
age squared/100	0.034	0.048	0.719	0.472
area: North-East	0.169	0.112	1.507	0.132
area: Center	-0.103	0.118	-0.872	0.383
area: South	-0.131	0.115	-1.137	0.256
area: Islands	-0.197	0.194	-1.011	0.312
skill	3.034	0.160	19.012	0.000
income	-0.144	0.012	-12.507	0.000
ptime	-1.116	0.164	-6.823	0.000
lagged response	0.441	0.076	5.800	0.000

Table 4: *Estimates of the regression coefficients in $\boldsymbol{\beta}_2$ for the model with $C = S = 3$.*

Again for the $C = S = 3$ case, we report in Table 5 the estimates of the parameters characterizing the latent structure of the model. We can see that the states of the cluster-specific latent processes are well separated, with the second state having the highest initial probability; these processes are also highly persistent. Similarly, the support points of the individual-specific latent variables are well separated, with the second having the highest probability.

A final comment concerns the estimation of the state dependence effect. It is usually expected that this effect is overestimated when the heterogeneity between subjects is not adequately represented in the model. In these situations, in fact, the true state dependence is confounded with the spurious state dependence (Heckman (1981b)). To verify whether the same happens in our analysis, we show in Table 6 how the estimate of the parameter γ varies with C and S. These results are in accordance with the standard theory mentioned above. In particular, in order to

Latent process (cluster level)					Latent variable (individual level)	
	initial	transition			support	
states	probabilities	probabilities			points	probabilities
0.000	0.275	0.862	0.124	0.014	0.000	0.119
4.684	0.666	0.011	0.987	0.001	3.117	0.661
7.692	0.059	0.000	0.027	0.973	4.866	0.220

Table 5: *Estimates of the parameters characterizing the latent structure of the model with $C = S = 3$.*

avoid an overestimate of the state dependence effect, increasing the number of support points of the individual-specific latent variables is more effective than increasing the number of states of the cluster-specific latent processes.

	S		
C	1	2	3
1	1.676	0.806	0.538
2	1.415	0.567	0.445
3	1.228	0.557	0.441

Table 6: *Estimates of the parameter for the state dependence effect obtained under different levels of C and S.*

6 Conclusions

We presented a model for clustered binary panel data which allows us to take into account unobserved individual heterogeneity together with cluster-specific dynamics. The model is based on latent variables which have a discrete distribution. The use of discrete distributions has the advantage, in comparison to the use of continuous distributions, of allowing a more flexible formulation of the latent structure of the model (Heckman and Singer (1984)). In our approach, we also control for state dependence in a suitable way. For this model, we dealt with the computational aspects involved in the ML estimation. In particular, we introduced an algorithm which allows us to compute exactly the joint probability of the response variables, and then the log-likelihood of the model. The numerical complexity of the algorithm grows linearly with the sample size. Direct evaluation of this probability, instead, requires a number of operations which increases exponentially with the size of the largest cluster. We also described a stochastic version of the algorithm which may be used even when the number of latent states and/or the number of time periods is very large.

Future developments of the approach concern its extension to the case of ordinal response variables and to the multivariate case in which we have a vector of response variables for every time period. Moreover, in our approach we considered all the covariates as strictly exogenous. In order to take into account that some of them may not be strictly exogenous, the approach has to be appropriately extended. This may be done by allowing the distribution of the latent variables to depend on these explanatory variables in a suitable way.

Appendix: derivatives of $\boldsymbol{q}$ and $\boldsymbol{m}_{hi}$

We now show how to compute the derivatives of $\boldsymbol{q}$ and $\boldsymbol{m}_{hi}$ with respect to the parameters in $\boldsymbol{\theta}$; these derivatives are used in Section 4.2 to compute the score vector and to estimate the information matrix. We recall that $\boldsymbol{q}$ is a C^T-dimensional vector with elements $p(\boldsymbol{c}_h)$, which depend only on the parameters π_c and $\pi_{c|b}$ contained in $\boldsymbol{\theta}_2$. Taking into account the constraint $\sum_c \pi_c = 1$, we have that

$$\frac{\partial p(\boldsymbol{c}_h)}{\partial \pi_c} = p(\boldsymbol{c}_h)\frac{1(c_{h1}=c) - 1(c_{h1}=1)}{\pi_c}, \qquad c = 2, \ldots, C.$$

Moreover, taking into account that $\sum_c \pi_{c|b} = 1$ for every b, we have that

$$\frac{\partial p(\boldsymbol{c}_h)}{\partial \pi_{c|b}} = p(\boldsymbol{c}_h)\sum_{t>1}\frac{1(c_{h,t-1}=b)[1(c_{ht}=c) - 1(c_{ht}=1)]}{\pi_{c_{ht}|c_{h,t-1}}}, \qquad b = 1, \ldots, C, \quad c = 2, \ldots, C.$$

The derivative of $p(\boldsymbol{c}_h)$ with respect to each parameter not in $\boldsymbol{\theta}_1$ is obviously equal to 0.

We also recall that $\boldsymbol{m}_{hi} = \boldsymbol{M}_{hi}\boldsymbol{1}_{C^T}$, where $\boldsymbol{M}_{hi}$ is a $C^T \times S$ matrix with elements $p(s_{hi})p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi})$, which depend only on the parameters in $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_3$. The derivative of these elements with respect to $\boldsymbol{\theta}_1$ may be computed as

$$\begin{aligned}\frac{\partial[p(s_{hi})p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi})]}{\partial \boldsymbol{\theta}_1} &= p(s_{hi})\frac{\partial p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi})}{\partial \boldsymbol{\theta}_1} = \\ &= p(s_{hi})p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi})\sum_t (2y_{hit} - 1)[1 - p(y_{hit}|c_{ht}, s_{hi}, y_{ih,t-1})]\boldsymbol{u}_{hit}(c_{ht}, s_{hi}),\end{aligned}$$

where the vector $\boldsymbol{u}_{hit}(c, s)$ has been defined in Section 4.1. Finally, taking into account that $\sum_s \rho_s = 1$, we have that

$$\begin{aligned}\frac{\partial[p(s_{hi})p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi})]}{\partial \rho_s} &= \frac{\partial p(s_{hi})}{\partial \rho_s}p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi}) \\ &= [1(s_{hi}=s) - 1(s_{hi}=1)]p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi}), \qquad s = 2, \ldots, S.\end{aligned}$$

The derivative of $p(s_{hi})p(\boldsymbol{y}_{hi}|\boldsymbol{c}_h, s_{hi})$ with respect to each parameter in $\boldsymbol{\theta}_2$ is obviously equal to 0. By summing appropriately the derivatives above, we can obtain the derivative of $\boldsymbol{m}'_{hi}$ with respect to $\boldsymbol{\theta}$.

References

Arellano, M. and Honoré, B.E., 2001. Panel data models: some recent developments. In: J.J. Heckman and E. Leamer (Eds), Handbook of Econometrics, vol. 5, Elsevier, Amsterdam.

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.

Butler, J. S. and Moffitt, R., 1982. A computationally efficient quadrature procedure for one-factor multinomial probit models. Econometrica, 50, 761-768.

Chamberlain, G., 1984. Panel data. In: Z. Griliches and M.D. Intrilligator (Eds.), Handbook of Econometrics, Elsevier Science Publisher, Amsterdam.

Chintagunta, P., Kyriazidou, E. and Perktold, J., 2001. Panel data analysis of household brand choices. Journal of Econometrics, 103, 111-153.

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.

Fuertes, A. M. and Kalotychou, E., 2006. Early warning systems for sovereign debt crises: the role of heterogeneity. Computational Statistics and Data Analysis, forthcoming.

Goldstein, M.Y., Goldstein, H. and Heath, A., 2000. Multilevel models for repeated outcomes: attitudes and voting over the electoral cycle. Journal of Royal Statistical Society A, 163, 49-62.

Goodman, L. A., 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 61, 215-231.

Gourieroux, C. and Monfort, A., 1996. Simulation-Based Econometric Methods. Oxford University Press, Oxford.

Hajivassiliou, V. A. and McFadden, D.L., 1998. The method of simulated scores for the estimation of LDV models. Econometrica, 66, 863-896.

Heckman, J. J., 1981a. Statistical models for discrete panel data. In: D.L. McFadden and C.F. Manski (Eds.), Structural Analysis of Discrete Data, MIT Press, Cambridge, MA.

Heckman, J. J., 1981b. Heterogeneity and state dependence. In: Rosen, S. (Ed.), Studies in Labor Markets. University of Chicago Press, Chicago, pp. 91139.

Heckman, J. J., 1981c. The incidental parameter problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In: D.L. McFadden and C.F. Manski (Eds.), Structural Analysis of Discrete Data, MIT Press, Cambridge, MA.

Heckman, J. J. and Willis, R. J., 1977. A beta-logistic model for the analysis of sequential labor force participation by married women. Journal of Political Economy, 85, 2758.

Heckman, J. J. and Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica, 52, 271320.

Hsiao, C., 1986. Analysis of Panel Data. Cambridge University Press, Cambridge, UK.

Hyslop, D. R., 1999. State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. Econometrica, 67, 1255-1294.

Langeheine R. and van de Pol F., 2002. Latent Markov chains. In: J.A. Hagenaars and A.L. McCutcheon (Eds.), Applied Latent Class Analysis, Cambridge University Press, Cambridge, UK, 304-341.

Langeheine, R., Stern, E. and van de Pol, F., 1994. State mastery learning: dynamic models for longitudinal data. Applied Psychological Measurement, 18, 277-291.

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

Mannan, H. R. and Koval, J. J., 2003. Latent mixed Markov modelling of smoking transitions using Monte Carlo bootstrapping. Statistical Methods in Medical Research, 12 125-146.

McLachlan, G. and Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, New York.

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.

Van de Pol, F. and Langeheine, R., 1990. Mixed Markov latent class models. In: C.C. Clogg (Eds.), Sociological Methodology, Blackwell, Oxford, 213-247.

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

Wiggins, L. M., 1973. Panel Analysis: Latent Probability Models for Attitudes and Behavior Processes. Elsevier, Amsterdam.