Journal of Statistical Software

October 2017, Volume 81, Issue 4. doi: 10.18637/jss.v081.i04

LMest: An R Package for Latent Markov Models for Longitudinal Categorical Data

Francesco Bartolucci
University of Perugia

Silvia Pandolfi
University of Perugia

Fulvia Pennoni
University of Milano-Bicocca

Abstract

Latent Markov (LM) models represent an important class of models for the analysis of longitudinal data, especially when response variables are categorical. These models have a great potential of application in many fields, such as economics and medicine. We illustrate the R package **LMest** that is tailored to deal with the basic LM model and some extended formulations accounting for individual covariates and for the presence of unobserved clusters of units having the same initial and transition probabilities (mixed LM model). The main functions of the package are tailored to parameter estimation through the expectation-maximization algorithm, which is based on suitable forward-backward recursions. The package also permits local and global decoding and to obtain standard errors for the parameter estimates. We illustrate the use of the package and its main features through some empirical examples in the fields of labour market, health, and criminology.

Keywords: expectation-maximization algorithm, forward-backward recursions, hidden Markov model, missing data.

1. Introduction

In this paper we illustrate the R (R Core Team 2017) package **LMest** (Bartolucci and Pandolfi 2017), available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=LMest`, which provides a collection of functions that can be used to estimate latent Markov (LM) models for longitudinal categorical data. The package is related to the book by Bartolucci, Farcomeni, and Pennoni (2013), where these models are illustrated in detail from the methodological point of view. Additional insights are given in the discussion paper by Bartolucci, Farcomeni, and Pennoni (2014b).

LM models are designed for the analysis of univariate and multivariate longitudinal/panel

data based on the repeated observation of a sample of units across time. These models are specially tailored to study the evolution of an individual characteristic of interest, when this characteristic is not directly observable. For this aim, the models at issue rely on a latent process following a Markov chain. Another reason for using LM models is to account for time-varying unobserved heterogeneity in addition to the effect of observable covariates on the response variables.

The initial LM formulation introduced by Wiggins (1973), see also Wiggins (1955), has been developed in several directions and in connection with applications in many fields, such as economics, medicine, psychology, and sociology. In particular, the basic LM model, which relies on a homogeneous Markov chain of first order, has been extended in several ways on the basis of parameterizations that allow us to incorporate certain hypotheses in the model. The most relevant extended version includes individual covariates that may affect the distribution of the latent process (Vermunt, Langeheine, and Böckenholt 1999; Bartolucci, Pennoni, and Francis 2007) or the conditional distribution of the response variables given this process (Bartolucci and Farcomeni 2009). LM models may be also formulated to take into account certain types of unobserved heterogeneity. In particular, we are referring to the mixed LM model (Van de Pol and Langeheine 1990) and to the LM model with random effects (Altman 2007; Maruotti 2011; Bartolucci, Pennoni, and Vittadini 2011). Within the first formulation, the initial and transition probabilities of the latent process are allowed to vary across different latent subpopulations.

LM models are conceived quite similarly to hidden Markov (HM) models (Zucchini and MacDonald 2009) for time-series data, but they are tailored to longitudinal data where many individuals are observed at only a few occasions, typically no more than ten. Differently from LM models, HM models with covariates or for complex data structures are rarely applied because these structures are typical of longitudinal studies.

Some R packages already exist that can handle LM and related models. In particular, HM models can be estimated by using packages **HMM** (Himmelmann 2010), **HiddenMarkov** (Harte 2017), or **depmixS4** (Visser and Speekenbrink 2010). The last one is the most closely related to our package **LMest** as it is tailored to deal with HM models based on a generalized linear formulation that can include individual covariates. On the other hand, package **depmixS4** is designed to deal with repeated measurements on a single unit, as in a time-series context. Packages **mhsmm** (O'Connell and Højsgaard 2011) and **hsmm** (Bulla and Bulla 2013) may be used to estimate hidden semi-Markov models. Package **msm** (Jackson 2011) is tailored to deal with HM and related continuous time models. Finally, it is worth mentioning package **hmm.discnp** (Turner 2016), which can be used to fit multiple hidden Markov models, and package **seqHMM** (Helske and Helske 2017), which also includes graphical tools to visualize sequence data and categorical time series with multiple units and covariates. The latter one is related to package **LMest** as it can be used to estimate certain versions of mixed hidden Markov models. A commercial software to perform data analyses based on certain types of LM models is **Latent GOLD** (Vermunt and Magidson 2016). Additionally, some MATLAB (The MathWorks Inc. 2014) toolboxes are available for this aim; see for example the **HMM** toolbox implemented by Murphy (1998).

The distinguishing features of the **LMest** package with respect to the packages mentioned above are the following:

- **LMest** is designed to work with longitudinal data, that is, with (even many) i.i.d. replicates of (usually short) sequences of data.
- It can deal with univariate and multivariate categorical outcomes.
- It allows for missing responses, drop-out, and non-monotonic missingness, under the missing-at-random assumption (Little and Rubin 2002).
- Standard errors for the parameter estimates are obtained by exact computation or through reliable approximations of the observed information matrix.
- Individual covariates are included through suitable parameterizations.
- Additional discrete random effects can be used to formulate mixed LM models.
- Computationally efficient algorithms are implemented for estimation and prediction of the latent states, also by relying on certain Fortran routines.

The present article is organized as follows. Section 2 briefly outlines the general formulation of LM models and deals with their maximum likelihood estimation. Section 3 describes the use of the **LMest** package to estimate the basic LM model without covariates. Section 4 is focused on LM models with individual covariates included in the measurement model, while Section 5 is focused on the case of individual covariates affecting the distribution of the latent process. In Section 6 we introduce the mixed LM model and we describe the R function for its estimation. Finally, Section 7 summarizes the main conclusions.

2. Latent Markov models for longitudinal data

In the following we provide a brief review of the statistical methodology related to LM models. The illustration closely follows the recent paper by Bartolucci *et al.* (2014b). We also focus on maximum likelihood estimation of these models on the basis of the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). Moreover, we deal with more advanced topics which are important for applications, such as selection of the number of latent states and prediction of these states via local or global decoding (Viterbi 1967; Juang and Rabiner 1991).

2.1. The general latent Markov model formulation

Consider the multivariate case where for a generic unit we observe a vector $\boldsymbol{Y}^{(t)}$ of r categorical response variables at T occasions, so that $t = 1, \ldots, T$. Each response variable is denoted by $Y_j^{(t)}$ and has c_j categories, labeled from 0 to $c_j - 1$, with $j = 1, \ldots, r$. Also let $\tilde{\boldsymbol{Y}}$ be the vector obtained by stacking $\boldsymbol{Y}^{(t)}$ for $t = 1, \ldots, T$; this vector has then rT elements. Obviously, in the univariate case we have a single response variable $Y^{(t)}$ for each time occasion, and $\tilde{\boldsymbol{Y}}$ is composed of T elements. When available, we also denote by $\boldsymbol{X}^{(t)}$ the vector of individual covariates available at the t-th time occasion and by $\tilde{\boldsymbol{X}}$ the vector of all the individual covariates, which is obtained by stacking vectors $\boldsymbol{X}^{(t)}$ for $t = 1, \ldots, T$.

The general LM model formulation assumes the existence of a latent process, denoted by $\boldsymbol{U} = (U^{(1)}, \ldots, U^{(T)})$, which affects the distribution of the response variables. Such a process is

assumed to follow a first-order Markov chain with state space $\{1, \ldots, k\}$, where k is the number of latent states. Under the *local independence* assumption, the response vectors $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T)}$ are assumed to be conditionally independent given the latent process $\boldsymbol{U}$. Moreover, the elements $Y_j^{(t)}$ of $\boldsymbol{Y}^{(t)}$, with $t = 1, \ldots, T$, are conditionally independent given $U^{(t)}$. This assumption leads to a strong simplification of the model, but it can be relaxed by allowing serial dependence through the inclusion of the lagged response variable among covariates, as in Bartolucci and Farcomeni (2009).

The parameters of the *measurement model* are the conditional response probabilities

$$\phi_{jy|u\boldsymbol{x}}^{(t)} = \mathsf{P}(Y_j^{(t)} = y | U^{(t)} = u, \boldsymbol{X}^{(t)} = \boldsymbol{x}), \quad j = 1, \ldots, r, \;\; y = 0, \ldots, c_j - 1,$$

which reduce to

$$\phi_{y|u\boldsymbol{x}}^{(t)} = \mathsf{P}(Y^{(t)} = y | U^{(t)} = u, \boldsymbol{X}^{(t)} = \boldsymbol{x}), \quad y = 0, \ldots, c - 1,$$

in the univariate case, with $t = 1, \ldots, T$ and $u = 1, \ldots, k$.

The parameters of the latent process are the initial probabilities

$$\pi_{u|\boldsymbol{x}} = \mathsf{P}(U^{(1)} = u | \boldsymbol{X}^{(1)} = \boldsymbol{x}), \quad u = 1, \ldots, k,$$

and the transition probabilities

$$\pi_{u|\bar{u}\boldsymbol{x}}^{(t)} = \mathsf{P}(U^{(t)} = u | U^{(t-1)} = \bar{u}, \boldsymbol{X}^{(t)} = \boldsymbol{x}), \quad t = 2, \ldots, T, \;\; \bar{u}, u = 1, \ldots, k,$$

where $\boldsymbol{x}$ denotes a realization of $\boldsymbol{X}^{(t)}$, y a realization of $Y_j^{(t)}$, u a realization of $U^{(t)}$, and $\bar{u}$ a realization of $U^{(t-1)}$.

On the basis of the above parameters, the conditional distribution of $\boldsymbol{U}$ given $\tilde{\boldsymbol{X}}$ may be expressed as

$$\mathsf{P}(\boldsymbol{U} = \boldsymbol{u} | \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}) = \pi_{u^{(1)}|\boldsymbol{x}^{(1)}} \prod_{t=2}^{T} \pi_{u^{(t)}|u^{(t-1)}\boldsymbol{x}^{(t)}},$$

where $\boldsymbol{u} = (u^{(1)}, \ldots, u^{(T)})$ and $\tilde{\boldsymbol{x}}$ denotes a realization of the vector of all response variables $\tilde{\boldsymbol{X}}$. Moreover, the conditional distribution of $\tilde{\boldsymbol{Y}}$ given $\boldsymbol{U}$ and $\tilde{\boldsymbol{X}}$ may be expressed as

$$\mathsf{P}(\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}} | \boldsymbol{U} = \boldsymbol{u}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}) = \prod_{t=1}^{T} \phi_{\boldsymbol{y}^{(t)}|u^{(t)}\boldsymbol{x}^{(t)}}^{(t)},$$

where, in general, we define $\phi_{\boldsymbol{y}|u\boldsymbol{x}}^{(t)} = \mathsf{P}(\boldsymbol{Y}^{(t)} = \boldsymbol{y} | U^{(t)} = u, \boldsymbol{X}^{(t)} = \boldsymbol{x})$ and, due to the assumption of local independence, we have

$$\phi_{\boldsymbol{y}|u\boldsymbol{x}}^{(t)} = \prod_{j=1}^{r} \phi_{jy_j|u\boldsymbol{x}}^{(t)}.$$

In the above expressions, $\tilde{\boldsymbol{y}}$ is a realization of $\tilde{\boldsymbol{Y}}$ made by the subvectors $\boldsymbol{y}^{(t)} = (y_1^{(t)}, \ldots, y_r^{(t)})$ whereas $\boldsymbol{y}$ is a realization of $\boldsymbol{Y}^{(t)}$ with elements y_j, $j = 1, \ldots, r$.

In the presence of individual covariates, the *manifest distribution* of the response variables corresponds to the conditional distribution of $\tilde{\boldsymbol{Y}}$ given $\tilde{\boldsymbol{X}}$, which is defined as

$$\begin{aligned}\mathsf{P}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) &= \mathsf{P}(\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}) \\ &= \sum_{\boldsymbol{u}} \pi_{u^{(1)}|\boldsymbol{x}^{(1)}} \pi^{(2)}_{u^{(2)}|u^{(1)}\boldsymbol{x}^{(2)}} \cdots \pi^{(T)}_{u^{(T)}|u^{(T-1)}\boldsymbol{x}^{(T)}} \times \phi^{(1)}_{\boldsymbol{y}^{(1)}|u^{(1)}\boldsymbol{x}^{(1)}} \cdots \phi^{(T)}_{\boldsymbol{y}^{(T)}|u^{(T)}\boldsymbol{x}^{(T)}}. \end{aligned} \tag{1}$$

In the basic version of the model, individual covariates are ruled out; therefore, we use symbol $\mathsf{P}(\tilde{\boldsymbol{y}})$ to refer to the manifest distribution. Moreover, when these covariates are available, we suggest to avoid that they simultaneously affect the distribution of the latent process and the conditional distribution of the response variables given this process. In fact, the two formulations have different interpretations, as explained in more detail in the following, and the resulting model would be difficult to interpret and estimate.

Finally, it is important to note that computing $\mathsf{P}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}})$, or $\mathsf{P}(\tilde{\boldsymbol{y}})$ for the basic LM model, involves a sum extended to all possible configurations of the vector $\boldsymbol{u}$, which are k^T; this typically requires a considerable computational effort. However, in order to efficiently compute such a probability we can use a forward recursion due to Baum, Petrie, Soules, and Weiss (1970), as illustrated in Bartolucci *et al.* (2013, Chapter 3).

2.2. Maximum likelihood estimation

We illustrate maximum likelihood estimation in the general case in which covariates are available. In this case, for a sample of n independent units that provide the response vectors $\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n$ and given the corresponding vectors of covariates $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n$, the model log-likelihood has the following expression:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \mathsf{P}(\tilde{\boldsymbol{y}}_i|\tilde{\boldsymbol{x}}_i).$$

Each vector $\tilde{\boldsymbol{y}}_i$ is a realization of $\tilde{\boldsymbol{Y}}$ that, in the multivariate case, is made up of the subvectors $\boldsymbol{y}_i^{(t)}$, $t = 1, \ldots, T$, having elements $y_{ij}^{(t)}$, $j = 1, \ldots, r$; similarly, $\tilde{\boldsymbol{x}}_i$ may be decomposed into the time-specific subvectors $\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(T)}$. Moreover, $\mathsf{P}(\tilde{\boldsymbol{y}}_i|\tilde{\boldsymbol{x}}_i)$ corresponds to the manifest probability of the responses provided by subject i, see Equation 1, and $\boldsymbol{\theta}$ is the vector of all free parameters affecting $\mathsf{P}(\tilde{\boldsymbol{y}}_i|\tilde{\boldsymbol{x}}_i)$.

The above log-likelihood function can be maximized by the EM algorithm (Baum *et al.* 1970; Dempster *et al.* 1977), as described in the following section.

Expectation-maximization algorithm

The EM algorithm is based on the complete data log-likelihood that, with multivariate categorical data, has the following expression:

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) = \sum_{j=1}^{r} \sum_{t=1}^{T} \sum_{u=1}^{k} \sum_{\boldsymbol{x}} \sum_{y=0}^{c_j-1} a_{ju\boldsymbol{x}y}^{(t)} \log \phi_{jy|u\boldsymbol{x}}^{(t)} + \sum_{u=1}^{k} \sum_{\boldsymbol{x}} b_{u\boldsymbol{x}}^{(1)} \log \pi_{u|\boldsymbol{x}} + \\ \sum_{t=2}^{T} \sum_{\bar{u}=1}^{k} \sum_{u=1}^{k} \sum_{\boldsymbol{x}} b_{\bar{u}u\boldsymbol{x}}^{(t)} \log \pi_{u|\bar{u}\boldsymbol{x}}^{(t)}, \end{aligned} \tag{2}$$

where, with reference to occasion t and covariate configuration $\boldsymbol{x}$, $a_{ju\boldsymbol{x}y}^{(t)}$ is the number of individuals that are in the latent state u and provide response y to variable $Y_j^{(t)}$, $b_{u\boldsymbol{x}}^{(t)}$ is the frequency of latent state u, and $b_{\bar{u}u\boldsymbol{x}}^{(t)}$ is the number of transitions from state $\bar{u}$ to state u.
The EM algorithm alternates the following two steps until convergence:

- **E-step**: This step consists in computing the posterior (given the observed data) expected value of each frequency involved in Equation 2 by suitable forward-backward recursions (Baum *et al.* 1970); these expected values are denoted by $\hat{a}_{ju\boldsymbol{x}y}^{(t)}$, $\hat{b}_{u\boldsymbol{x}}^{(t)}$, and $\hat{b}_{\bar{u}u\boldsymbol{x}}^{(t)}$.
- **M-step**: This step consists in maximizing the complete data log-likelihood expressed as in Equation 2, with each frequency substituted by the corresponding expected value. How to maximize this function depends on the specific formulation of the model and, in particular, on whether the covariates are included in the measurement or in the latent model.

The convergence of the EM algorithm is checked on the basis of the relative log-likelihood difference, that is,

$$\left[\ell(\boldsymbol{\theta}^{(s)}) - \ell(\boldsymbol{\theta}^{(s-1)})\right] / |\ell(\boldsymbol{\theta}^{(s)})| < \epsilon, \tag{3}$$

where $\boldsymbol{\theta}^{(s)}$ is the parameter estimate obtained at the end of the s-th M-step and ϵ is a suitable tolerance level (e.g., 10^{-8}).
The EM algorithm could converge to a mode of the log-likelihood that does not correspond to the global maximum, due to the multimodality of this function. In order to avoid this problem, we suggest to use different initializations of this algorithm, either deterministic or random, and to take as final estimate the one corresponding to the highest log-likelihood; this estimate is denoted by $\hat{\boldsymbol{\theta}}$. In particular, for LM models without covariates, the random initialization is based on suitably rescaled random numbers drawn from a uniform distribution from 0 to 1 for the initial and transition probabilities of the Markov chain and for the conditional response probabilities.
We refer the reader to Bartolucci *et al.* (2013) for a detailed description of the EM algorithm and its initialization.

Standard errors

After the model is estimated, standard errors for the parameter estimates may be obtained on the basis of the observed information matrix, denoted by $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$. In particular, each standard error is obtained as the square root of the corresponding diagonal element of the inverse of this matrix, $\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}$. The **LMest** package computes the observed information matrix, and then provides the standard errors, by using either the exact computation method proposed by Bartolucci and Farcomeni (2015) or the numerical method proposed by Bartolucci and Farcomeni (2009), depending on the complexity of the model of interest.
The exact computation of $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ is based on the Oakes' identity (Oakes 1999). This method uses the complete data information matrix, produced by the EM algorithm, and a correction matrix computed on the basis of the first derivative of the posterior probabilities obtained from the backward-forward recursions. On the other hand, with the approximate method,

$\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ is obtained as minus the numerical derivative of the score vector $\boldsymbol{s}(\hat{\boldsymbol{\theta}})$ at convergence. The score vector, in turn, is computed as the first derivative of the conditional expected value of the complete data log-likelihood, which is based on the expected frequencies $\hat{a}_{ju\boldsymbol{xy}}^{(t)}$, $\hat{b}_{u\boldsymbol{x}}^{(t)}$, and $\hat{b}_{\bar{u}u\boldsymbol{x}}^{(t)}$ corresponding to the final parameter estimates $\hat{\boldsymbol{\theta}}$, that is,

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

where $\mathcal{X}$ and $\mathcal{Y}$ stand for the observed data; see Bartolucci *et al.* (2013) and Pennoni (2014) for details.

For the basic LM model and for the model with individual covariates affecting the distribution of the latent process, the **LMest** package also provides functions to obtain standard errors by parametric bootstrap (Davison and Hinkley 1997).

2.3. Criteria for selecting the number of latent states

In certain applications, the number of latent states, k, can be *a priori* defined, as in the univariate case in which it is reasonable to fix k equal to the number of response categories. Otherwise, the following criteria are typically used to select the number of latent states: the Akaike information criterion (AIC) of Akaike (1973) and the Bayesian information criterion (BIC) of Schwarz (1978). They are based on the indices

$$\begin{aligned} \text{AIC} &= -2\hat{\ell} + 2\ \#\text{par}, \\ \text{BIC} &= -2\hat{\ell} + \log(n)\ \#\text{par}, \end{aligned}$$

where $\hat{\ell}$ denotes the maximum of the log-likelihood of the model of interest and #par denotes the number of free parameters.

According to each of the above criteria, the optimal number of latent states is the one corresponding to the minimum value of AIC or BIC; this model represents the best compromise between goodness-of-fit and complexity. If the two criteria lead to selecting a different number of states, the second one is usually preferred. However, other criteria may be used, such as those taking into account the quality of the classification; for a review see Bacci, Pandolfi, and Pennoni (2014) and Bartolucci, Bacci, and Pennoni (2014a).

2.4. Local and global decoding

The **LMest** package allows us to perform decoding, that is, prediction of the sequence of the latent states for a certain sample unit on the basis of the data observed for this unit.

In particular, the EM algorithm directly provides the estimated posterior probabilities of $U^{(t)}$, denoted by $\mathsf{P}(U^{(t)} = u|\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}})$, for $t = 1, \ldots, T$, $u = 1, \ldots, k$, and for every covariate and response configuration $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ observed at least once. These probabilities can be directly maximized to obtain a prediction of the latent state of each subject at each time occasion t; this is the so-called *local decoding*. Note that this type of decoding minimizes the classification error at each time occasion, but may yield sub-optimal predictions of $U^{(1)}, \ldots, U^{(T)}$.

In order to track the latent state of a subject across time, the most *a posteriori* likely sequence of states must be obtained through the so-called *global decoding*, which is based on

an adaptation of the Viterbi (1967) algorithm; see also Juang and Rabiner (1991). The algorithm proceeds through a forward-backward recursion of a complexity similar to the recursions adopted for maximum likelihood estimation within the EM algorithm, so that global decoding may be even performed for long sequences of data; see Bartolucci *et al.* (2013, Chapter 7).

3. Basic latent Markov model

Following Bartolucci *et al.* (2013), and as already mentioned above, the basic LM model rules out individual covariates and assumes that the conditional response probabilities are time homogeneous. In symbols, we have that $\phi^{(t)}_{y|u\boldsymbol{x}} = \phi_{y|u}$ in the univariate case and $\phi^{(t)}_{jy|u\boldsymbol{x}} = \phi_{jy|u}$ in the multivariate case; we also have $\pi_{u|\boldsymbol{x}} = \pi_u$ and $\pi^{(t)}_{u|\bar{u}\boldsymbol{x}} = \pi^{(t)}_{u|\bar{u}}$.

In order to fit this model, we use function `est_lm_basic` as illustrated, through a specific application, in the following.

3.1. Application to job satisfaction data

The illustration is based on data coming from the Russia Longitudinal Monitoring Survey (RLMS)[1]. The data are obtained by the "adult questionnaire", which is also focused on aspects concerning primary and secondary employment. In particular, we consider the question concerning job satisfaction related to the primary work. The resulting response variable, named IKSJ, has five ordered categories: "absolutely satisfied", "mostly satisfied", "neutral", "not very satisfied", "absolutely unsatisfied", which are coded from 1 to 5. The data we use are referred to a sample of $n = 1{,}718$ individuals followed for $T = 7$ years from 2008 to 2014. According to the economic theory proposed by Stiglitz, Amartya, and Fitoussi (2010), job satisfaction connects two main latent factors concerning individual characteristics and working environment that may evolve in time. Therefore, the LM approach is particularly suitable for the analysis of the data at issue.

The data are already contained in the data frame `RLMSdat` included in the **LMest** package. As illustrated in the following, each line of this data frame refers to an individual and each column contains the observed responses from the first to the last year of interview.

```
R> library("LMest")
R> data("RLMSdat", package = "LMest")
R> head(RLMSdat)

  IKSJQ IKSJR IKSJS IKSJT IKSJU IKSJV IKSJW
1     2     2     2     2     1     1     2
2     2     2     3     2     2     2     2
3     2     4     4     2     3     4     2
4     2     3     2     2     2     2     2
5     2     2     3     2     2     2     2
6     3     4     3     2     2     4     3
```

Function `est_lm_basic` requires the following main input arguments:

[1] For more details on the study see http://www.cpc.unc.edu/projects/rlms-hse, http://www.hse.ru/org/hse/rlms.

- `S`: Design array for the response configurations (of dimension `n` × `TT` × `r`, where `TT` corresponds to the number of time occasions) with categories starting from `0`; missing responses are allowed, coded as `NA`.
- `yv`: Vector of frequencies of the available configurations.
- `k`: Number of latent states.
- `mod`: Model on the transition probabilities; `mod = 0` when these probabilities depend on time, `mod = 1` when they are independent of time (i.e., the latent Markov chain is time homogeneous), and `mod` from `2` to `TT` when the Markov chain is partially homogeneous of order equal to `mod`.
- `tol`: Tolerance level for checking convergence, which corresponds to ϵ in definition (3); the default value is `1e-8`.
- `maxit`: Maximum number of iterations of the algorithm; the default value is `1000`.
- `start`: Equal to `0` for deterministic starting values of the model parameters (default value), to `1` for random starting values, and to `2` for initial values provided as input arguments.
- `piv`, `Pi`, `Psi`: Initial values of the initial probability vector, of the transition probability matrix, and of the conditional response probabilities, respectively, when `start = 2`.
- `out_se`: Equal to `TRUE` to require the computation of the information matrix and the standard errors; `FALSE` is the default option.

In order to obtain the estimates for the data reported above, we use function `aggr_data` that, starting from a unit-by-unit data frame, returns the set of distinct observed response patterns and the corresponding frequencies:

```
R> out <- aggr_data(RLMSdat)
R> yv <- out$freq
R> S <- 5 - out$data_dis
```

Note that the response categories must start from `0` in order to be used in `est_lm_basic`; therefore, these categories are rescaled in a way that also accounts for their reverse order in the initial dataset. In this way, level `0` corresponds to category "absolutely unsatisfied" and level `4` corresponds to category "absolutely satisfied".

In this illustrative application, we estimate the basic LM model under the assumption of time homogeneous transition probabilities (`mod = 1`) with a fixed number of states, $k = 3$, so as to obtain three groups of individuals clustered on the basis of the level of job satisfaction. For this aim, we use the following command:

```
R> mod1 <- est_lm_basic(S, yv, k = 3, mod = 1, start = 0, out_se = TRUE)
```

In this application we use the deterministic initialization (`start = 0`). Moreover, option `out_se = TRUE` is used to obtain the standard errors for the parameter estimates on the basis of the observed information matrix, which is exactly computed as described in Section 2.2.

The running time of the above command is around 13 seconds when run, as the other codes illustrated in the paper, on an Intel Core i7 processor with 2.7 GHz.

In the following, we show the estimation results, by using the `print` method, which provides the maximum log-likelihood, the number of free parameters, and values of AIC and BIC indices:

```
R> mod1

Call:
est_lm_basic(S = S, yv = yv, k = 3, start = 0, mod = 1, out_se = TRUE)

Convergence info:
        LogLik np      AIC      BIC
[1,] -13557.21 20 27154.41 27263.39
```

The main outputs of function `est_lm_basic` may be displayed using the following command:

```
R> summary(mod1)

Call:
est_lm_basic(S = S, yv = yv, k = 3, start = 0, mod = 1, out_se = TRUE)

Coefficients:

Initial probabilities:
     est_piv
[1,]  0.3638
[2,]  0.4623
[3,]  0.1739

Standard errors for the initial probabilities:
     se_piv
[1,] 0.0185
[2,] 0.0244
[3,] 0.0189

Transition probabilities:
     state
state      1      2      3
    1 0.8863 0.0991 0.0145
    2 0.0472 0.9335 0.0193
    3 0.0469 0.0774 0.8757

Standard errors for the transition probabilities:
     state
state      1      2      3
    1 0.0105 0.0109 0.0051
```

```
    2 0.0069 0.0088 0.0059
    3 0.0105 0.0198 0.0196

Conditional response probabilities:
, , item = 1

        state
category      1      2      3
       0 0.0731 0.0000 0.0031
       1 0.2442 0.0224 0.0233
       2 0.4421 0.0965 0.0448
       3 0.2197 0.8166 0.3480
       4 0.0209 0.0645 0.5808


Standard errors for the conditional response probabilities:
, , item = 1

        state
category      1      2      3
       0 0.0048 0.0001 0.0019
       1 0.0092 0.0035 0.0056
       2 0.0103 0.0079 0.0087
       3 0.0137 0.0105 0.0245
       4 0.0038 0.0075 0.0256
```

According to the estimated conditional probabilities $\hat{\phi}_{y|u}$ (returned in `Psi`), we can interpret the first latent state as the one corresponding to a low level of satisfaction (high probability of responding "not very satisfied" and "neutral"), the second state to an intermediate level of this characteristic (probability of around 0.82 of responding "mostly satisfied"), whereas the last state corresponds to the highest level of satisfaction. The output displayed above also contains the estimated initial probability vector (`piv`), with elements $\hat{\pi}_u$, $u = 1, 2, 3$, that may be easily interpreted as quantities proportional to the size of each latent state at the beginning of the period of observation. Accordingly, we conclude that in 2008 most individuals belong to the second latent state, 36% of them belong to the first state, and only 17% to the last latent state. Moreover, according to the estimated transition probabilities $\hat{\pi}_{u|\bar{u}}$ (returned in `Pi`), the selected model leads to the conclusion that there is a quite high persistence in the same state during the years of the survey.

Selection of the number of states

It is important to recall that, when the value of k is not a priori known, it must be selected on the basis of the observed data. Moreover, different initializations of the EM algorithm must be attempted in order to prevent the problem of multimodality of the likelihood function. Both issues, that is, model selection and multimodality, can be addressed by using function `search.model.LM` that may be also used for the more complex models that will be illustrated later on. In the present application, we use this function to estimate the basic LM model for

increasing values of k from 1 to 5 so as to select the optimal number of latent states using BIC. Moreover, considering that the likelihood function may be multimodal, `search.model.LM` uses one deterministic initialization (`start = 0`) and a number of random initializations (`start = 1`) proportional to the number of latent states. In this preliminary exploration, the tolerance level is set equal to `1e-5` to reduce the computing time. Starting from the best solution obtained in this way, a final run is performed (`start = 2`), with a default tolerance level equal to `1e-10`.

Function `search.model.LM` requires the following main input arguments (for additional details we refer to the help page of the function):

- `version`: Model to be estimated (`"basic"` = basic LM model – parameters are estimated by function `est_lm_basic`; `"manifest"` = LM model with covariates in the measurement model – function `est_lm_cov_manifest`; `"latent"` = LM model with covariates in the distribution of the latent process – function `est_lm_cov_latent`).
- `kv`: Vector of possible number of latent states.
- `nrep`: To fix the number of random initializations for each element of `kv`; this number is equal to `nrep`$\times$`(k - 1)` and the default value is `nrep = 2`.
- `...`: Additional arguments for functions `est_lm_basic`, `est_lm_cov_manifest`, or `est_lm_cov_latent`.

Using the following commands, we obtain the results of the model selection strategy illustrated above:

```
R> set.seed(14326)
R> res1 <- search.model.LM(version = "basic", kv = 1:5, S, yv, mod = 1,
+    out_se = TRUE)
R> summary(res1)

Call:
search.model.LM(version = "basic", kv = 1:5, S, yv, mod = 1,
    out_se = TRUE)
     states        lk np      AIC      BIC
[1,]      1 -14943.73  4 29895.45 29917.25
[2,]      2 -13921.09 11 27864.18 27924.12
[3,]      3 -13557.20 20 27154.41 27263.39
[4,]      4 -13392.93 31 26847.86 27016.78
[5,]      5 -13369.45 44 26826.90 27066.65
```

The computing time required to run the above model selection strategy is around 159 seconds. Note that we fix the seed, by command `set.seed(14326)`, so that the reader can reproduce exactly the same results. On the basis of the above output, the model corresponding to the minimum BIC is that with $k = 4$ latent states. Function `search.model.LM` returns `out.single` as main output, which contains, in a list format, the output of each model for every k in `kv`. Therefore, `summary` method may be used to show the results for the model with the selected number of states:

```
R> summary(res1$out.single[[4]])

Call:
est_lm_basic(S = ..1, yv = ..2, k = k, start = 2, mod = 1, tol = tol2,
    out_se = out_se, piv = out[[k]]$piv, Pi = out[[k]]$Pi,
    Psi = out[[k]]$Psi)

Coefficients:

Initial probabilities:
     est_piv
[1,]  0.1863
[2,]  0.2021
[3,]  0.4400
[4,]  0.1717

Standard errors for the initial probabilities:
     se_piv
[1,] 0.0211
[2,] 0.0255
[3,] 0.0250
[4,] 0.0187

Transition probabilities:
     state
state      1      2      3      4
    1 0.7645 0.1568 0.0642 0.0145
    2 0.0183 0.8727 0.0925 0.0165
    3 0.0088 0.0520 0.9219 0.0173
    4 0.0266 0.0250 0.0675 0.8809

Standard errors for the transition probabilities:
     state
state      1      2      3      4
    1 0.0251 0.0299 0.0199 0.0103
    2 0.0119 0.0208 0.0171 0.0077
    3 0.0047 0.0093 0.0105 0.0064
    4 0.0085 0.0110 0.0198 0.0197

Conditional response probabilities:
, , item = 1

        state
category      1      2      3      4
       0 0.1748 0.0111 0.0009 0.0030
       1 0.4541 0.1035 0.0248 0.0222
       2 0.1851 0.5576 0.0635 0.0493
```

```
       3 0.1679 0.3040 0.8413 0.3463
       4 0.0181 0.0239 0.0694 0.5793


Standard errors for the conditional response probabilities:
, , item = 1

        state
category      1      2      3      4
       0 0.0175 0.0042 0.0010 0.0019
       1 0.0244 0.0137 0.0038 0.0055
       2 0.0277 0.0212 0.0082 0.0091
       3 0.0216 0.0218 0.0115 0.0243
       4 0.0072 0.0055 0.0080 0.0255
```

The specific output for the selected model with 4 classes may be interpreted starting from the estimated conditional response probabilities to identify the different latent states and then considering the initial and transition probabilities to have a picture of the distribution of these states across time.

4. Covariates in the measurement model

When the individual covariates are included in the measurement model, the conditional distribution of the response variables given the latent states may be parameterized by generalized logits. In such a situation, the latent variables account for the unobserved heterogeneity, that is, the heterogeneity between individuals that we cannot explain on the basis of the observable covariates. The advantage with respect to a standard random-effects or latent class model with covariates is that the unobservable heterogeneity is allowed to be time-varying; for a deeper discussion see Bartolucci and Farcomeni (2009) and Pennoni and Vittadini (2013).

4.1. Assumptions

In dealing with univariate data in which each response variable has an ordinal nature, as in the next illustrative example, we denote the number of its response categories by c. In formulating the model we rely on a parameterization based on global logits of the following type:

$$\log \frac{\mathsf{P}(Y^{(t)} \geq y | U^{(t)} = u, \boldsymbol{X}^{(t)} = \boldsymbol{x})}{\mathsf{P}(Y^{(t)} < y | U^{(t)} = u, \boldsymbol{X}^{(t)} = \boldsymbol{x})} = \log \frac{\phi^{(t)}_{y|u\boldsymbol{x}} + \ldots + \phi^{(t)}_{c-1|u\boldsymbol{x}}}{\phi^{(t)}_{0|u\boldsymbol{x}} + \ldots + \phi^{(t)}_{y-1|u\boldsymbol{x}}} = \mu_y + \alpha_u + \boldsymbol{x}^\top \boldsymbol{\beta}, \quad (4)$$

with $t = 1, \ldots, T$, $u = 1, \ldots, k$, and $y = 1, \ldots, c-1$. Note that these logits reduce to standard logits in the case of binary variables, that is, when $c = 2$. In the above expression, μ_y are the cut-points, α_u are the support points corresponding to each latent state, and $\boldsymbol{\beta}$ is the vector of regression parameters for the covariates.

As mentioned in Section 2.1, the inclusion of individual covariates in the measurement model is typically combined with the constraints $\pi_{u|\boldsymbol{x}} = \pi_u$ and $\pi^{(t)}_{u|\bar{u},\boldsymbol{x}} = \pi_{u|\bar{u}}$, $t = 1, \ldots, T$, $\bar{u}, u =$

$1, \ldots, k$, in order to avoid interpretability problems of the resulting model. Also note that, under these constraints, the transition probabilities are assumed to be time homogeneous so as to reduce the number of free parameters.

The LM model with individual covariates in the measurement model may be estimated by function `est_lm_cov_manifest` that is illustrated in the following.

4.2. Application to health related data

We provide an illustration based on data collected within the Health and Retirement Study (HRS) conducted by the University of Michigan[2]. The data concern aspects related to retirement and health among elderly individuals in the USA. The sample is nationally representative of the population aged over 50 years, whereas the response variable is the Self-Reported Health Status (named SRHS) and it is measured on a scale based on five ordered categories: "excellent", "very good", "good", "fair", and "poor". The sample we use includes $n = 7,074$ individuals interviewed at $T = 8$ occasions every two years. Therefore, it is reasonable to expect that unobserved factors affecting the health status may change during a so long period, and then time-invariant latent variables are not suitable to represent these factors. The LM model with covariates directly takes this issue into account.

The data are reported in long format and, therefore, for each subject the number of records is equal to the number of occasions. There are no missing responses or dropout in the dataset.

```
R> data("data_SRHS_long", package = "LMest")
R> data_SRHS <- data_SRHS_long
R> data_SRHS[1:10, ]
```

```
    id gender race education age srhs
1   1      1    1         3  56    4
2   1      1    1         3  58    4
3   1      1    1         3  60    3
4   1      1    1         3  62    3
5   1      1    1         3  64    4
6   1      1    1         3  66    3
7   1      1    1         3  68    3
8   1      1    1         3  70    3
9   2      2    1         5  54    3
10  2      2    1         5  55    3
```

The first column contains the `id` code of each subject, then there are columns for the available covariates. Gender is coded as 1 for male and 2 for female, whereas race has three categories coded as 1 for white, 2 for black, and 3 for others. Educational level is represented by five ordered categories coded as 1 for high school, 2 for general educational diploma, 3 for high school graduate, 4 for some college, and 5 for college and above. Finally, age is measured in years for each time occasion. The last column is related to the categorical response variable with $c = 5$ categories that are originally coded from 1 to 5. For instance, the individual with `id` equal to `1` provides responses "good" or "fair" at each time occasion.

[2]For more details on the study see http://hrsonline.isr.umich.edu/.

Function `est_lm_cov_manifest` requires the following main input arguments (see the help page of the function for additional arguments):

- `S`: Design array for the response configurations (of dimension `n` $\times$ `TT`) with categories starting from `0`.
- `X`: Array of covariates (of dimension `n` $\times$ `TT` $\times$ `nc`, where `nc` corresponds the number of covariates).
- `k`: Number of latent states.
- `mod`: Type of model to be estimated, coded as `mod = "LM"` for the model illustrated in Section 4.1 based on parameterization (4). In such a context, the latent process is of first order with initial probabilities equal to those of the stationary distribution of the chain. When `mod = "FM"`, the function estimates a model relying on the assumption that the distribution of the latent process is a mixture of AR(1) processes with common variance σ^2 and specific correlation coefficients ρ_u. This model strictly follows the one proposed by Bartolucci *et al.* (2014a); see also Pennoni and Vittadini (2013) for a comparison between the two types of model and the help page of the function for further details.
- `q`: Number of support points of the AR(1) structure described above.
- `tol`: Tolerance level for checking convergence; the default value is `1e-8`.
- `maxit`: Maximum number of iterations of the algorithm; the default value is `1000`.
- `start`: Equal to `0` for deterministic starting values of the model parameters (default value), to `1` for random starting values, and to `2` for initial values in input.
- `mu`, `al`, `be`, `la`, `PI`: Initial values of the model parameters when `start = 2` (vector of cut-points, vector of support points for the latent states, vector of regression parameters, vector of initial probabilities, and transition probability matrix, respectively).
- `output`: Equal to `TRUE` to print additional output; `FALSE` is the default option.
- `out_se`: Equal to `TRUE` to calculate the information matrix and the standard errors; `FALSE` is the default option.

Function `est_lm_cov_manifest`, as well as other estimation functions for LM models, require the data to be in array format. To give this structure to data that are originally in long format, we can use function `long2matrices` that is included in the package.

```
R> out <- with(data_SRHS, long2matrices(id = id, X = cbind(gender - 1,
+    race == 2 | race == 3, education == 4, education == 5, age - 50,
+    (age - 50)^2/100), Y = srhs))
```

Function `long2matrices` mainly requires, as input arguments, the vector of the individual labels, `id`, the matrix of the covariates, `X`, and the vector of the responses, `Y`; see the help page of the function for details. Note that, in using the previous command, gender is included as a dummy variable equal to 1 for female, race is included as a dummy variable equal to 1 for non-white, and two dummy variables are used for the educational level: the first is equal to 1 for

some college education and the second is equal to 1 for college education and above. Moreover, age is scaled by 50, and age squared is also included (suitably rescaled). An alternative formulation, which may result in a simpler interpretation of the parameter estimates, is based on considering as covariates the baseline age and the time since the beginning of the longitudinal study.

Function `long2matrices` generates the array of the responses `YY`, which are rescaled to vary from `0` ("poor") to `4` ("excellent"), as required by `est_lm_cov_manifest`:

```
R> S <- 5 - out$YY
```

It also returns the array `XX` containing the covariate configurations of each individual for every time occasion. For example, individual with `id == 3994` has the following covariate configuration:

```
R> X <- out$XX
R> X[3994, , ]

     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    0    0    0    1 0.01
[2,]    1    0    0    0    3 0.09
[3,]    1    0    0    0    5 0.25
[4,]    1    0    0    0    7 0.49
[5,]    1    0    0    0    9 0.81
[6,]    1    0    0    0   11 1.21
[7,]    1    0    0    0   13 1.69
[8,]    1    0    0    0   15 2.25
```

The individual considered above is a female (see first column), white (second column), with high school diploma (third and fourth columns); she is 51 years old at the first interview, 53 years old at the second interview, and so on (fifth column).

As a preliminary example, to estimate the model at issue with a given number of states, say $k = 3$, the command is

```
R> mod2 <- est_lm_cov_manifest(S, X, k = 3, mod = "LM", tol = 1e-8,
+    start = 1, output = TRUE, out_se = TRUE)
```

However, instead of showing and commenting the output produced in this way, we prefer to directly illustrate how to deal with this model by function `search.model.LM`, which addresses the problem of model selection, in terms of k, and that of multimodality of the likelihood function. Moreover, as the sample size is particularly large and the model selection strategy may require a considerable amount of computing time, we extract a subset of observations to make the results easily reproducible. Accordingly, we consider only those individuals who are 70 years old at the third interview so as to account for the influence of the covariates on the oldest individuals. Then, we run the `search.model.LM` function for a number of states k from `1` to `5`, as follows:

```
R> ind <- (X[, 3, 5] >= 20)
R> Sn <- S[ind, , ]
```

```
R> Xn <- X[ind, , ]
R> set.seed(71432)
R> res2 <- search.model.LM(version = "manifest", kv = 1:5, Sn, Xn,
+    mod = "LM", out_se = TRUE, tol2 = 1e-8)
R> summary(res2)

Call:
search.model.LM(version = "manifest", kv = 1:5, Sn, Xn, mod = "LM",
    tol2 = 1e-08, out_se = TRUE)
     states        lk np      AIC      BIC
[1,]      1 -1557.692 10 3135.384 3164.363
[2,]      2 -1419.621 13 2865.242 2902.914
[3,]      3 -1354.151 18 2744.302 2796.463
[4,]      4 -1332.653 25 2715.306 2787.752
[5,]      5 -1323.122 34 2714.244 2812.770
```

Under this setting, the function requires about 5,041 seconds to run. Note that option `out_se = TRUE` allows us to obtain standard errors for the parameter estimates, by means of the numerical method described in Section 2.2, so as to evaluate the effect of the covariates on the probability of responding in a certain way. According to these results, the minimum BIC index corresponds to the model with $k = 4$ latent states; the estimation results are illustrated in the following:

```
R> summary(res2$out.single[[4]])

Call:
est_lm_cov_manifest(S = ..1, X = ..2, k = k, mod = "LM", tol = tol2,
    start = 2, mu = out[[k]]$mu, al = out[[k]]$al, be = out[[k]]$be,
    si = out[[k]]$si, rho = out[[k]]$rho, la = out[[k]]$la, PI = out[[k]]$PI,
    out_se = out_se)

Coefficients:

  Vector of cut-points:
[1]  5.4921  2.4152 -0.1917 -2.8573

 Support points for the latent states:
[1] -1.6147 -5.0482  0.7809  4.0389

 Estimate of the vector of regression parameters:
[1] -0.9669  0.7219  2.1244  2.4585  0.0312 -0.3574

 Vector of initial probabilities:
[1] 0.2810 0.1130 0.4368 0.1692

 Transition matrix:
       [,1]   [,2]   [,3]   [,4]
```

```
[1,] 0.9032 0.0334 0.0245 0.0388
[2,] 0.1788 0.7962 0.0000 0.0250
[3,] 0.0092 0.0301 0.9606 0.0002
[4,] 0.0177 0.0029 0.0611 0.9184

 Standard errors for the regression parameters:
[1] 0.3257 0.3945 0.3784 0.3489 0.1424 0.2730
```

The above output contains the estimated cut-points (`mu`), corresponding to $\hat{\mu}_y$, the estimated support points for the latent states (`al`), corresponding to $\hat{\alpha}_u$, and the estimated vector of regression parameters (`be`), $\hat{\boldsymbol{\beta}}$, as in expression (4). Note that the support points could be sorted so that the latent states result ordered from the worst to the best perceived health conditions. The estimated coefficients in $\hat{\boldsymbol{\beta}}$ are reported in the same order adopted to define the array `X` of covariates. The list of objects returned by the function, contained in `res2$out.single[[4]]`, may also be displayed in the usual way; for a complete list of the returned arguments, we refer to the help pages of the package. As an example, the standard errors for the estimated regression coefficients (`sebe`) may be obtained as

```
R> round(res2$out.single[[4]]$sebe, 3)
```

```
[1] 0.326 0.394 0.379 0.349 0.142 0.273
```

On the basis of the estimated regression parameters, we can evaluate the effect of the covariates on the probability of reporting a certain level of the health status. In particular, women tend to report worse health status than men (the odds ratio for females versus males is equal to $\exp(-0.967) = 0.380$), whereas individuals having a higher number of years of schooling tend to have a better opinion about their health status than subjects having lower education (the odds ratio for college education and above is equal to $\exp(2.458) = 11.682$). We also observe that white individuals have a lower probability of reporting a good health status with respect to non-white individuals, but the coefficient is not significant. Among the selected individuals aged 70 years and over, the effect of age is positive but it is not significant. Also the quadratic term of age is not significant.

In this example, the time-varying random effects are used to account for the unobserved heterogeneity and the interpretation of the latent distribution is not of primary interest. The fact that the optimal number of states is $k = 4$ provides evidence for the presence of this type of heterogeneity, that is, that SRHS cannot be only explained on the basis of the few covariates we have used.

From the estimated initial probabilities π_u, returned in the vector `la`, we observe that many individuals start in the third latent class (44%), which corresponds to subjects with a quite good perceived health status. The estimated transition matrix (`PI`), with elements corresponding to $\pi_{u|\bar{u}}$, $\bar{u}, u = 1, \ldots, 4$, shows a quite high persistence in the same state. The highest transition probability is 0.18 and is observed from the second state, corresponding to the worst health condition, to the first state. The remaining transition probabilities are always lower than 0.07. For another application of the LM model to ordinal longitudinal data see Pennoni and Vittadini (2013).

5. Covariates in the latent model

When the covariates are included in the latent model, we suppose that the response variables measure the individual characteristic of interest (e.g., the quality of life) that is represented by the latent variables. This characteristic is not directly observable and may evolve over time. In such a case, the main research interest is in modeling the effect of covariates on the latent distribution, as is illustrated in the following; see also Bartolucci, Lupparelli, and Montanari (2009).

5.1. Assumptions

A natural way to allow the initial and transition probabilities of the LM chain to depend on individual covariates is by adopting a multinomial logit parameterization as follows:

$$\log \frac{\mathsf{P}(U^{(1)} = u|\boldsymbol{X}^{(1)} = \boldsymbol{x})}{\mathsf{P}(U^{(1)} = 1|\boldsymbol{X}^{(1)} = \boldsymbol{x})} = \log \frac{\pi_{u|\boldsymbol{x}}}{\pi_{1|\boldsymbol{x}}} = \beta_{0u} + \boldsymbol{x}^\top \boldsymbol{\beta}_{1u}, \quad u = 2, \ldots, k, \tag{5}$$

$$\log \frac{\mathsf{P}(U^{(t)} = u|U^{(t-1)} = \bar{u}, \boldsymbol{X}^{(t)} = \boldsymbol{x})}{\mathsf{P}(U^{(t)} = \bar{u}|U^{(t-1)} = \bar{u}, \boldsymbol{X}^{(t)} = \boldsymbol{x})} = \log \frac{\pi^{(t)}_{u|\bar{u}\boldsymbol{x}}}{\pi^{(t)}_{\bar{u}|\bar{u}\boldsymbol{x}}} = \gamma_{0\bar{u}u} + \boldsymbol{x}^\top \boldsymbol{\gamma}_{1\bar{u}u}, \tag{6}$$

for $t = 2, \ldots, T$ and $\bar{u}, u = 1, \ldots, k$, with $\bar{u} \neq u$. In the above expressions, $\boldsymbol{\beta}_u = (\beta_{0u}, \boldsymbol{\beta}_{1u}^\top)^\top$ and $\boldsymbol{\gamma}_{\bar{u}u} = (\gamma_{0\bar{u}u}, \boldsymbol{\gamma}_{1\bar{u}u}^\top)^\top$ are parameter vectors to be estimated which are collected in the matrices $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$.

For a more parsimonious model, instead of using (6) we can rely on the following parameterization for the transition probabilities, that is, a multinomial logit parameterization based on the difference between two sets of parameters:

$$\log \frac{\mathsf{P}(U^{(t)} = u|U^{(t-1)} = \bar{u}, \boldsymbol{X}^{(t)} = \boldsymbol{x})}{\mathsf{P}(U^{(t)} = \bar{u}|U^{(t-1)} = \bar{u}, \boldsymbol{X}^{(t)} = \boldsymbol{x})} = \gamma_{0\bar{u}u} + \boldsymbol{x}^\top (\boldsymbol{\gamma}_{1u} - \boldsymbol{\gamma}_{1\bar{u}}), \tag{7}$$

where $\boldsymbol{\gamma}_{11} = \mathbf{0}$ to ensure model identifiability. The parameterization used for modeling the initial probabilities is again based on standard multinomial logits, as defined in (5).

As already mentioned, when the covariates affect the distribution of the latent process, these covariates are typically excluded from the measurement model and we adopt the constraint $\phi^{(t)}_{y|u\boldsymbol{x}} = \phi_{y|u}$ in the univariate case or $\phi^{(t)}_{jy|u\boldsymbol{x}} = \phi_{jy|u}$ in the multivariate case. We also rely on the assumption of time homogeneity for the conditional response probabilities. Parameterizations based on (5) and (6) or (7) are implemented in the R function `est_lm_cov_latent`, which allows us to estimate the resulting LM models.

5.2. Application to health related data

To illustrate function `est_lm_cov_latent`, we consider the HRS data introduced in Section 4.2. In such a context, an interesting research question concerns the relationship between SRHS and the covariates. When the latter ones are included in the latent model, the initial and transition probabilities are estimated accounting for the covariate configurations and this may be useful to identify clusters of individuals related to specific needs.

The R function is based on the following main input arguments:

- `S`: Array of observed response configurations (of dimension `n` × `TT` × `r`) with categories starting from `0`; missing responses are allowed, coded as `NA`.
- `X1`: Matrix of covariates affecting the initial probabilities (of dimension `n` × `nc1`, where `nc1` is the number of corresponding covariates).
- `X2`: Array of covariates affecting the transition probabilities (of dimension `n` × `(TT - 1)` × `nc2`, where `nc2` is the number of corresponding covariates).
- `k`: Number of latent states.
- `start`: Equal to `0` for deterministic starting values of the model parameters (default), to `1` for random starting values, and to `2` to define initial values as input arguments.
- `tol`: Tolerance level for checking convergence; the default value is `1e-8`.
- `maxit`: Maximum number of iterations of the algorithm; the default value is `1000`.
- `param`: Type of parameterization for the transition probabilities, coded as `param = "multilogit"` (default) for the model parameterization defined in (6) and as `param = "difflogit"` for the parameterization defined in (7).
- `Psi, Be, Ga`: Initial values of the matrix of the conditional response probabilities and of the parameters affecting the logits for the initial and transition probabilities, respectively, when `start = 2`.
- `output`: Equal to `TRUE` to obtain additional output arguments; `FALSE` is the default option.
- `out_se`: Equal to `TRUE` to compute the information matrix and standard errors; `FALSE` is the default option.
- `fixPsi`: Equal to `TRUE` if the matrix of conditional response probabilities is given in the input and is kept fixed during the estimation process; `FALSE` is the default option.

The model to be fitted is specified by means of two design matrices for the initial and transition probabilities as defined in the following:

```
R> data("data_SRHS_long", package = "LMest")
R> data_SRHS <- data_SRHS_long
R> out <- with(data_SRHS, long2matrices(id = id, X = cbind(gender - 1,
+    race == 2 | race == 3, education == 4, education == 5,
+    age - 50, (age - 50)^2/100), Y = srhs))
R> S <- 5 - out$YY
R> X <- out$XX
R> X1 <- X[, 1, ]
R> TT <- 8
R> X2 <- X[, 2:TT, ]
R> colnames(X1) <- c("gender", "race", "some college",
+    "college and above",  "age", "age^2")
R> dimnames(X2)[[3]] <- c("gender", "race", "some college",
+    "college and above", "age", "age^2")
```

Note that the array of the response configurations, `S`, is rescaled; this is because the first response category must be coded as `0`, corresponding to the worst self-reported health status. Moreover, matrix `X1` is referred to the first time occasion and it includes the covariates affecting the initial probabilities of the latent process, as in (5). Accordingly, for $t = 2, \ldots, T$, matrix `X2` includes the covariates affecting the transition probabilities of the latent process, as in (6).

We illustrate the function considering the full sample of $n = 7,074$ individuals, with $T = 8$ time occasions. In particular, we fit the model defined in Section 5.1, with a number of latent states equal to the number of response categories (i.e., $k = 5$), by using the following command:

```
R> mod3 <- est_lm_cov_latent(S = S, X1 = X1, X2 = X2, k = 5, start = 0,
+    param = "multilogit", fort = TRUE, output = TRUE)
```

Here, we rely on a deterministic initialization of the estimation algorithm. The computing time required to run the above function, again on an Intel Core i7, is around 30 seconds. The results can be displayed by using the `print` method, which returns the main output arguments:

```
R> mod3

Call:
est_lm_cov_latent(S = S, X1 = X1, X2 = X2, k = 5, start = 0,
    param = "multilogit", fort = TRUE, output = TRUE)

Convergence info:
        LogLik  np      AIC      BIC
[1,] -62426.58 188 125229.2 126519.6
```

The `summary` method displays the estimation results:

```
R> summary(mod3)

Call:
est_lm_cov_latent(S = S, X1 = X1, X2 = X2, k = 5, start = 0,
    param = "multilogit", fort = TRUE, output = TRUE)

Coefficients:

 Be - Parameters affecting the logit for the initial probabilities:
                    logit
                            2       3       4       5
  intercept           0.7363  1.7189  1.6036  1.6193
  gender             -0.0047 -0.2998 -0.1043 -0.2342
  race                0.0203 -0.3720 -1.1436 -1.4381
  some college        0.4496  1.1250  1.4969  1.7694
  college and above   0.2917  1.7028  2.4876  3.0017
```

```
  age                -0.0357 -0.0338 -0.0426 -0.0727
  age^2               0.3170  0.2118  0.2071  0.2415

 Ga - Parameters affecting the logit for the transition probabilities:
, , logit = 1

                   logit
                         2        3       4        5
  intercept         -3.4506 -30.3104 -7.3260  -0.7017
  gender            -0.1625 -10.1897 -1.0196  -6.8384
  race              -0.2798  -8.4566  1.1111 -10.7134
  some college      -0.1829   0.3931 -8.2008  -1.3429
  college and above  0.7692   2.3875 -7.4341  -9.0080
  age                0.3068   1.8111  0.1387  -0.4990
  age^2             -1.4401  -2.9503 -0.0329   0.9238

, , logit = 2

                   logit
                         2       3        4        5
  intercept         -3.0686 -2.1749 -16.3128 -14.5490
  gender             0.2511 -0.1756  -1.4647  -0.4473
  race              -0.3063 -0.6950   0.5331   9.3810
  some college       0.0422  0.5414  -8.4763  -6.8829
  college and above  0.2865 -0.0138  -2.5061  -6.9650
  age                0.0230 -0.0457   1.7420   0.0202
  age^2             -0.0697  0.0633  -5.9457   0.0991

, , logit = 3

                   logit
                         2       3       4       5
  intercept         -4.6189 -2.0081 -3.4384 -4.0396
  gender            -0.4507 -0.2624 -0.1901 -1.8983
  race               0.2696 -0.0602 -0.0040  2.7118
  some college      -1.7920 -0.2221 -0.6352 -0.5840
  college and above -0.5082 -0.6282 -1.0302 -1.7189
  age               -0.0398 -0.0387  0.0223 -0.0833
  age^2              0.3219  0.1620  0.0080 -0.1805

, , logit = 4

                   logit
                         2       3       4       5
  intercept         -6.9963 -6.0078 -2.1665 -2.9657
  gender            -0.5585  0.2142 -0.1654 -0.7564
  race               0.8580  0.9215  0.3901  0.0079
```

```
  some college       0.8565 -0.4858 -0.2066 -0.3965
  college and above -1.0905  0.0214 -0.4433 -1.5825
  age                0.0658 -0.0476  0.0075 -0.0887
  age^2              0.1313  0.6135  0.0284  0.3115

, , logit = 5

                   logit
                         2       3       4       5
  intercept        -16.2170 -2.2457 -2.2154 -1.3555
  gender             1.4289 -2.3098 -0.7527 -0.2012
  race               3.3827  0.7291  2.0447  0.1940
  some college      -6.8843 -1.9841 -1.5632 -0.0773
  college and above -6.6738 -4.1917 -1.7000 -0.3858
  age                1.6782  0.0318 -0.0374 -0.0139
  age^2             -8.2599 -1.1975  0.0345  0.0545


 Psi - Conditional response probabilities:
, , item = 1

        state
category      1      2      3      4      5
       0 0.6981 0.0597 0.0043 0.0017 0.0004
       1 0.2670 0.6828 0.0844 0.0083 0.0014
       2 0.0255 0.2203 0.7131 0.1454 0.0314
       3 0.0083 0.0288 0.1813 0.7567 0.1871
       4 0.0012 0.0084 0.0169 0.0879 0.7797
```

The estimated conditional response probabilities (`Psi`), corresponding to $\hat{\phi}_{y|u}$, allow us to characterize the latent states: individuals in the first latent state have the highest probability of reporting the worst health status, whereas the last state corresponds to subjects having the highest probability of reporting the best health status. Individuals in the remaining states show intermediate levels of the perceived health status.

The argument `Be` returned by the function contains the estimated regression parameters affecting the distribution of the initial probabilities, and corresponds to $\hat{\boldsymbol{\beta}}$; see definition (5). The estimated positive intercepts indicate a general tendency to report a good health status at the beginning of the survey. Gender log-odds (second row of `Be`) are all negative, indicating that females report a worse health status than males at the first time occasion. The two log-odds corresponding to the educational level are positive, indicating that a higher educational level leads to better health. Finally, the negative estimates for age indicate that, at the beginning of the study, older individuals generally tend to report a poor health status.

The output `Ga` contains the estimated parameters affecting the distribution of the transition probabilities, and corresponds to matrix $\hat{\boldsymbol{\Gamma}}$; see definition (6). To offer some insights for the interpretation of this output, note that parameters in $\boldsymbol{\gamma}_{1\bar{u}u}$ refer to the transition from level $\bar{u}$ to level u of the latent process. Therefore, as an example, the first column of

`Ga[, , , 5]` contains the parameter estimates measuring the influence of each covariate on the transition from the fifth state, corresponding to the best health conditions, to the first state, corresponding to the worst conditions. On the basis of these results, we notice that the influence of gender is positive, meaning that for females the probability of this transition is higher with respect to males. The influence of education, measured on the logit scale, is negative meaning that individuals with a higher level of education tend to move from the fifth to the first state less frequently than those with a lower education. At the same time, age has a positive effect on the chance of transition from the highest to the lowest state.

Using option `output = TRUE`, the function also returns some additional outputs. In particular, it is possible to obtain the estimated initial probability matrix, `Piv`, and the estimated transition probabilities matrices, `PI`. On the basis of these estimates, it is possible to compute the average initial and transition probabilities for a group of individuals of interest. For example, if we select white females with high level of education (college education and above), we obtain the corresponding average estimated initial and transition probabilities by means of the following commands:

```
R> ind1 <- (X1[, 1] == 1 & X1[, 2] == 0 & X1[, 4] == 1)
R> piv1 <- round(colMeans(mod3$Piv[ind1, ]), 4)
R> piv1

       1      2      3      4      5
  0.0069 0.0192 0.1501 0.3449 0.4789

R> PI1 <- round(apply(mod3$PI[, , ind1, 2:TT], c(1, 2), mean), 4)
R> PI1

      state
state      1      2      3      4      5
    1 0.8429 0.1570 0.0000 0.0000 0.0000
    2 0.0788 0.8640 0.0570 0.0002 0.0000
    3 0.0041 0.0452 0.9389 0.0116 0.0002
    4 0.0005 0.0067 0.0656 0.9239 0.0033
    5 0.0000 0.0001 0.0060 0.1184 0.8755
```

In a similar way, it is possible to compute the average initial and transition probabilities for non-white females with the same educational level:

```
R> ind2 <- (X1[, 1] == 1 & X1[, 2] == 1 & X1[, 4] == 1)
R> piv2 <- round(colMeans(mod3$Piv[ind2, ]), 4)
R> piv2

     1      2      3      4      5
0.0191 0.0547 0.2898 0.3096 0.3268

R> PI2 <- round(apply(mod3$PI[, , ind2, 2:TT], c(1, 2), mean), 4)
R> PI2
```

```
     state
state      1      2      3      4      5
    1 0.8777 0.1223 0.0000 0.0000 0.0000
    2 0.0607 0.9086 0.0305 0.0003 0.0000
    3 0.0053 0.0426 0.9377 0.0114 0.0029
    4 0.0011 0.0145 0.0926 0.8886 0.0032
    5 0.0000 0.0001 0.0438 0.1350 0.8210
```

On the basis of the results above we conclude that, at the beginning of the period of observation, white females have a probability approximately equal to 0.82 of being in the last two states corresponding to very good or excellent health conditions. On the other hand, non-white females have a higher probability of being in the first three states, corresponding to worse health conditions, than white females. Moreover, the estimated transition probability matrices show a quite high persistence in the same latent state for both groups of individuals. However, the lower diagonal elements of the transition matrix for non-white females are higher than those referred to white females, except for some transitions, showing a worse perception of their health status.

Decoding

Local and global decoding are implemented in the R function `decoding`, which allows us to predict the sequence of latent states for a certain sample unit on the basis of the output of the main estimation functions in the package. Function `decoding` requires the following input arguments:

- `est`: Object containing the output of one of the following functions: `est_lm_basic`, `est_lm_cov_latent`, `est_lm_cov_manifest`, or `est_lm_mixed`.
- `Y`: Vector or matrix of responses.
- `X1`: Matrix of covariates affecting the initial probabilities (for `est_lm_cov_latent`) or affecting the distribution of the responses (for `est_lm_cov_manifest`).
- `X2`: Array of covariates affecting the transition probabilities (for `est_lm_cov_latent`).

For the application above, the most likely sequence of latent states for all sample units may be obtained by the following command:

```
R> out <- decoding(mod3, S, X1, X2)
R> head(out$Ug)
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    2    2    3    3    3    3    3    3
[2,]    3    3    3    3    3    3    3    3
[3,]    5    5    4    3    3    3    3    3
[4,]    3    3    2    2    2    2    2    2
[5,]    5    5    5    5    5    5    5    5
[6,]    4    4    4    4    4    4    4    4
```

For instance, we conclude that for the first subject there is only one transition, at the third time occasion, from the second to the third latent state.

6. Mixed latent Markov model

Another relevant extension of LM models may be formulated to take into account additional sources of (time-fixed) dependence in the data. In this paper, we provide an illustration of the mixed LM model (Van de Pol and Langeheine 1990) in which the parameters of the latent process are allowed to vary in different latent subpopulations defined by an additional discrete latent variable.

6.1. Assumptions

Let U be a (time-invariant) discrete latent variable that defines unobserved clusters (or latent classes) of units having the same initial and transition probabilities. The latent process is here denoted by $\boldsymbol{V} = (V^{(1)}, \ldots, V^{(T)})$, which substitutes the symbol $\boldsymbol{U}$ used in the previous sections. In such a context, the variables in $\boldsymbol{V}$ follow a first-order Markov chain only conditionally on U. This additional latent variable has k_1 support points (corresponding to the latent classes) and mass probabilities denoted by λ_u, $u = 1, \ldots, k_1$. Accordingly, we denote by k_2 the number of latent states, corresponding to the number of support points of every latent variable $V^{(t)}$, $t = 1, \ldots, T$.

Note that, under this approach, which may be useful from a perspective of clustering, the initial and transition probabilities of the latent Markov chain differ between sample units in a way that does not depend on the observable covariates.

The parameters to be estimated are the conditional response probabilities, denoted by

$$\phi_{jy|v} = \mathsf{P}(Y_j^{(t)} = y | V^{(t)} = v), \quad j = 1, \ldots, r,\ t = 1, \ldots, T,\ v = 1, \ldots, k_2,\ y = 0, \ldots, c_j - 1,$$

the initial probabilities

$$\pi_{v|u} = \mathsf{P}(V^{(1)} = v | U = u), \quad u = 1, \ldots, k_1,\ v = 1, \ldots, k_2,$$

and the transition probabilities

$$\pi_{v|u\bar{v}} = \mathsf{P}(V^{(t)} = v | U = u, V^{(t-1)} = \bar{v}), \quad t = 2, \ldots, T,\ u = 1, \ldots, k_1,\ \bar{v}, v = 1, \ldots, k_2.$$

This model relies on the assumption that the conditional response probabilities and the transition probabilities are time-homogeneous. Obviously, this formulation may be extended by also including observable covariates as illustrated in the previous sections; see Bartolucci *et al.* (2013, Chapter 6), for a detailed description.

We derive the manifest distribution of $\tilde{\boldsymbol{Y}}$ by extending the rules given in Section 2. In particular, the conditional distribution of $\boldsymbol{V}$ given U is equal to

$$\mathsf{P}(\boldsymbol{V} = \boldsymbol{v} | U = u) = \pi_{v^{(1)}|u} \prod_{t=2}^{T} \pi_{v^{(t)}|u\bar{v}^{(t-1)}},$$

where $\boldsymbol{v} = (v^{(1)}, \ldots, v^{(T)})$ denotes a realization of $\boldsymbol{V}$. Given the assumption of local independence that is maintained under this model, the conditional distribution of $\tilde{\boldsymbol{Y}}$ given U and $\boldsymbol{V}$

reduces to

$$\mathsf{P}(\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}}|U = u, \boldsymbol{V} = \boldsymbol{v}) = \mathsf{P}(\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}}|\boldsymbol{V} = \boldsymbol{v}) = \prod_{t=1}^{T} \phi_{\boldsymbol{y}^{(t)}|v^{(t)}} = \prod_{j=1}^{r}\prod_{t=1}^{T} \phi_{jy_j^{(t)}|v^{(t)}},$$

whereas the conditional distribution of $\tilde{\boldsymbol{Y}}$ given U is expressed as

$$\mathsf{P}(\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}}|U = u) = \sum_{\boldsymbol{v}} \pi_{v^{(1)}|u}\pi_{v^{(2)}|uv^{(1)}} \cdots \pi_{v^{(T)}|uv^{(T-1)}} \phi_{\boldsymbol{y}^{(1)}|v^{(1)}} \cdots \phi_{\boldsymbol{y}^{(T)}|v^{(T)}}.$$

Finally, the manifest distribution of $\tilde{\boldsymbol{Y}}$ is now obtained by the following sum

$$\mathsf{P}(\tilde{\boldsymbol{y}}) = \mathsf{P}(\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}}) = \sum_{u=1}^{k_1} \mathsf{P}(\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}}|U = u)\lambda_u,$$

which depends on the mass probabilities for the distribution of the latent variable U. Even in this case $\mathsf{P}(\tilde{\boldsymbol{y}})$ may be computed through a forward recursion (Baum *et al.* 1970).

Referred to the maximum likelihood estimation of the mixed LM model formulated above, we can extend the procedure illustrated in Section 2.2, where the complete data log-likelihood has now the following expression:

$$\ell^*(\boldsymbol{\theta}) = \sum_{j=1}^{r}\sum_{t=1}^{T}\sum_{v=1}^{k_2}\sum_{y=0}^{c_j-1} a_{jvy}^{(t)} \log \phi_{jy|v} + \sum_{u=1}^{k_1}\left(\sum_{v=1}^{k_2} b_{uv}^{(1)} \log \pi_{v|u} + \sum_{t=2}^{T}\sum_{\bar{v}=1}^{k_2}\sum_{v=1}^{k_2} b_{u\bar{v}v}^{(t)} \log \pi_{v|u\bar{v}}\right) + \sum_{u=1}^{k_1} c_u \log \lambda_u.$$

In the previous expression, $a_{jvy}^{(t)}$ is the number of sample units that are in latent state v at occasion t and provide response y to variable j. Moreover, with reference to latent class u and occasion t, $b_{uv}^{(t)}$ is the number of sample units in latent state v, and $b_{u\bar{v}v}^{(t)}$ is the number of transitions from state $\bar{v}$ to state v. Finally, c_u is the overall number of sample units that are in latent class u.

6.2. Application to data from criminology

The mixed LM model is illustrated by using a simulated dataset similar to the one analyzed in Bartolucci *et al.* (2007); see also Francis, Liu, and Soothill (2010) and Pennoni (2014). The data are related to the complete conviction histories of a cohort of offenders followed from the age of criminal responsibility, 10 years. The offense code has been reduced to 73 major offenses and they have been grouped according to the Research Development and Statistics Directorate (1998)[3] on the basis of the following ten typologies: "violence against the person", "sexual offenses", "burglary", "robbery", "theft and handling stolen goods", "fraud and forgery", "criminal damage", "drug offenses", "motoring offenses", and "other offenses". The main interest is in evaluating the patterns of criminal behavior among individuals.

For the simulated data, we consider $n = 10{,}000$ individuals (including the proportion of non-offenders): 4,800 females and 5,200 males. We also consider $T = 6$ age bands of length

[3]See http://discover.ukdataservice.ac.uk/catalogue/?sn=3935.

equal to five years (10–15, 16–20, 21–25, 26–30, 31–35, and 36–40 years) and $r = 10$ binary response variables, corresponding to the typologies of offenses defined above. For every age band, each response variable is equal to 1 if the subject has been convicted for a crime of the corresponding offense group and to 0 otherwise.

Then, the data matrix, reported below in long format, has been simulated on the basis of the same parameter estimates reported in Bartolucci *et al.* (2007):

```
R> data("data_criminal_sim", package = "LMest")
R> head(data_criminal_sim)
```

```
      id sex time y1 y2 y3 y4 y5 y6 y7 y8 y9 y10
 [1,]  1   1    1  0  0  0  0  0  0  0  0  0   0
 [2,]  1   1    2  0  0  0  0  0  0  0  0  0   0
 [3,]  1   1    3  0  0  0  0  0  0  0  0  0   0
 [4,]  1   1    4  0  0  0  0  0  0  0  0  0   0
 [5,]  1   1    5  0  0  0  0  0  0  0  0  0   0
 [6,]  1   1    6  0  0  0  0  0  0  0  0  0   0
```

The first column of the data matrix contains the `id` code of each subject, whereas the covariate gender (second column named `sex`) is coded as 1 for male and 2 for female, the column named `time` is referred to the age band, and the last ten columns are related to the binary response variables.

The R function aimed at estimating the mixed LM models is `est_lm_mixed`, which requires the following input arguments:

- `S`: Array of response configurations (`n` $\times$ `TT` $\times$ `r`) with categories starting from `0`.
- `yv`: Vector of frequencies of the configurations.
- `k1`: Number of support points, corresponding to latent classes, of the distribution of the latent variable U.
- `k2`: Number of support points, corresponding to latent states, of the distribution of the latent process $\boldsymbol{V}$.
- `start`: Equal to `0` for deterministic starting values of the model parameters (default value) and to `1` for random starting values.
- `tol`: Tolerance level for checking convergence; the default value is `1e-8`.
- `maxit`: Maximum number of iterations of the algorithm; the default value is `1000`.
- `out_se`: Equal to `TRUE` to calculate the information matrix and the standard errors; `FALSE` is the default option.

For this example we also use function `long2wide` that allows us to convert the data from the long to the wide format. It requires to specify the name of the data matrix, the column referred to the identification number of the individuals, the column of the age band (time occasions), and the names of the columns of the covariates and of the responses, as follows:

```
R> out <- long2wide(data = data_criminal_sim, nameid = "id", namet = "time",
+    colx = "sex", coly = paste0("y", 1:10))
R> YY <- out$YY
R> XX <- out$XX
R> freq <- out$freq
```

Other options can be found in the help page of the function. The main objects in output are the array `YY` of response configurations, the array `XX` of covariate configurations, and the vector `freq` of the corresponding frequencies. For the data at hand, the design matrix for the responses, `YY`, contains 915 different response configurations, for $T = 6$ age bands, and $r = 10$ response variables. Similarly, for the covariate gender, matrix `XX` contains 915 configurations for $T = 6$ age bands. In the following, we show two response patterns with the associated covariate configurations and the corresponding frequency in the sample:

```
R> YY[148, , ]

     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    0    0    0    0    0    0    0    0    0     0
[2,]    0    0    0    0    0    0    0    0    0     0
[3,]    0    0    0    0    0    0    0    0    0     0
[4,]    0    0    0    0    0    0    0    0    0     0
[5,]    0    0    0    0    0    0    0    0    0     0
[6,]    1    0    0    0    0    0    0    0    0     0

R> XX[148, ]

[1] 2 2 2 2 2 2

R> freq[148]

[1] 3

R> YY[149, , ]

     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    0    0    0    0    1    0    0    0    0     0
[2,]    0    0    0    0    0    0    0    0    0     0
[3,]    0    0    0    0    0    0    0    0    0     0
[4,]    0    0    0    0    0    0    0    0    0     0
[5,]    0    0    0    0    0    0    0    0    0     0
[6,]    0    0    0    0    0    0    0    0    0     0

R> XX[149, ]

[1] 2 2 2 2 2 2

R> freq[149]
```

```
 [1] 113
```

From the above configurations, we observe that only 3 females have been convicted for violence against the person (first response variable) in the last age band, which is from 36 to 40 years old, whereas 113 females committed a theft (fifth variable) during the first time window, related to age 10–15.

To illustrate the use of function `est_lm_mixed`, we fit the model in Section 6 on such data with $k_1 = 2$ latent classes, $k_2 = 2$ latent states, restricting the analysis to females. We use the following commands in R:

```
R> YY <- YY[XX[, 1] == 2, , ]
R> freq <- freq[XX[, 1] == 2]
R> mod4 <-  est_lm_mixed(S = YY, yv = freq, k1 = 2, k2 = 2)
```

Using an Intel Core i7 processor, the above function takes around 46 seconds to converge. Then, we obtain the value of the log-likelihood at convergence by the `print` method:

```
R> mod4

Call:
est_lm_mixed(S = YY, yv = freq, k1 = 2, k2 = 2)

Convergence info:
        LogLik np      BIC
[1,] -18347.08 27 36925.18
```

Moreover, the `summary` command provides the estimated mass probability vector (`la`), with elements corresponding to $\hat{\lambda}_u$, the estimated initial (`Piv`) and transition probability matrices (`Pi`), with elements $\hat{\pi}_{v|u}$ and $\hat{\pi}_{v|u\bar{v}}$, respectively, and the array of the estimated conditional response probabilities (`Psi`), containing $\hat{\phi}_{jy|v}$, for $j = 1, \ldots, 10$, $y = 0, 1$, and $v = 1, 2$:

```
R> summary(mod4)

Call:
est_lm_mixed(S = YY, yv = freq, k1 = 2, k2 = 2)

Coefficients:

Mass probabilities:
[1] 0.2175 0.7825

Initial probabilities:
   u
v   1      2
  1 1 0.9087
  2 0 0.0913
```

```
Transition probabilities:
, , u = 1

   v1
v0       1      2
  1 0.8525 0.1475
  2 0.6414 0.3586

, , u = 2

   v1
v0       1      2
  1 1.0000 0.0000
  2 0.3382 0.6618


Conditional response probabilities:
, , j = 1

   v
y        1      2
  0 0.9952 0.8242
  1 0.0048 0.1758

, , j = 2

   v
y        1      2
  0 0.9983 0.9809
  1 0.0017 0.0191

, , j = 3

   v
y        1      2
  0 0.9963 0.7436
  1 0.0037 0.2564

, , j = 4

   v
y        1      2
  0 0.9999 0.9737
  1 0.0001 0.0263
```

```
, , j = 5

   v
y        1      2
  0 0.9773 0.4546
  1 0.0227 0.5454

, , j = 6

   v
y        1      2
  0 0.9982 0.8892
  1 0.0018 0.1108

, , j = 7

   v
y        1      2
  0 0.9957 0.8177
  1 0.0043 0.1823

, , j = 8

   v
y        1      2
  0 0.9976 0.9105
  1 0.0024 0.0895

, , j = 9

   v
y        1      2
  0 0.9999 0.9815
  1 0.0001 0.0185

, , j = 10

   v
y        1      2
  0 0.9987 0.7912
  1 0.0013 0.2088
```

The estimated conditional probability of committing each type of crime, $\hat{\phi}_{j1|v}$, may be also displayed as follows:

```
R> round(mod4$Psi[2, , ], 3)
```

```
    j
v       1     2     3     4     5     6     7     8     9    10
  1 0.005 0.002 0.004 0.000 0.023 0.002 0.004 0.002 0.000 0.001
  2 0.176 0.019 0.256 0.026 0.545 0.111 0.182 0.090 0.019 0.209
```

According to the above probabilities, we can identify the first latent state as that of those females with null or very low tendency to commit crimes, whereas the second latent state corresponds to criminals having mainly as type of activity: "theft", "burglary", and "other offenses".

The model formulation allows us to characterize the two clusters of individuals at the beginning of the period of observation and to follow their evolution over time. According to the estimated mass probabilities, the first cluster, which includes around 22% of females, is characterized by individuals having, at the beginning of the period of observation, a probability equal to 1 to be in the first latent state (corresponding to null tendency to commit a crime). On the other hand, females included in the second cluster (78%) are characterized by an initial probability of being in the second latent state of around 0.09. Comparing the estimated transition probability matrices we observe, within each cluster, a very high level of persistence in the first latent state. Moreover, females classified in the first cluster present a higher probability (of around 0.64) to move from the second to the first state than those assigned to the second cluster (0.34), revealing a more pronounced tendency to improve in their behavior.

7. Conclusions

We illustrate the R package **LMest** that allows us to efficiently fit latent Markov (LM) models for categorical longitudinal data. For a comprehensive overview about these models we refer the reader to Bartolucci *et al.* (2013) and Bartolucci *et al.* (2014b). Both manifest and latent distributions of the model can be parameterized so as to include the effect of individual covariates. The mixed formulation includes additional latent variables in these parameterizations. It shall be noted that all functions above can be used with multivariate categorical outcomes, with the only exception of `est_lm_cov_manifest`, which is restricted to univariate categorical outcomes. Functions `est_lm_basic` and `est_lm_cov_latent` also allow us to deal with missing data, non-monotone missingness, and dropout, under the missing-at-random assumption.

Overall, we consider this package as a relevant advance for applied researchers interested in longitudinal data analyses in the presence of categorical response variables. In particular, we recall that in this context LM models are particularly useful at least from three different perspectives: (*i*) to represent and study the evolution of an individual characteristic (e.g., quality of life) that is not directly observable; (*ii*) to account for unobserved heterogeneity due to omitted covariates in a time-varying fashion; and (*iii*) to account for measurement errors in observing a sequence of categorical response variables. We recall that, when covariates are available, they are typically included in the measurement model for applications of type (*ii*), so that the response variables are affected by observed covariates and latent variables that are considered on the same footing, whereas the covariates are included in the latent models for applications of type (*i*) and (*iii*), so that they affect the distribution of the latent process.

Further updates of the package will include the possibility to use multivariate outcomes in

function `est_lm_cov_manifest` and new functions with different formulations of mixed LM models, also for sample units collected in clusters (Bartolucci *et al.* 2011). We also plan to include estimation methods which are alternative to pure maximum likelihood estimation, as the three-step method proposed by Bartolucci, Montanari, and Pandolfi (2015) as well as the estimation procedure proposed by Bartolucci, Pennoni, and Vittadini (2016) to allow the model to be suitable in a potential outcome research framework (Rubin 1974).

Acknowledgments

F. Bartolucci and F. Pennoni acknowledge the financial support from the grant "Finite mixture and latent variable models for causal inference and analysis of socio-economic data" (FIRB – Futuro in ricerca) funded by the Italian Government (RBFR12SHVV). All authors thank, for providing the data, the RAND Center for the Study of Aging (`www.rand.org/labor/aging`) and the National Research University Higher School of Economics and ZAO Demoscope, together with Carolina Population Center, University of North Carolina at Chapel Hill and the Institute of Sociology RAS.

References

Akaike H (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In BN Petrov, F Csaki (eds.), *Second International Symposium on Information Theory*, pp. 267–281. Akademiai Kiado, Budapest.

Altman RM (2007). "Mixed Hidden Markov Models: An Extension of the Hidden Markov Model to the Longitudinal Data Setting." *Journal of the American Statistical Association*, **102**(477), 201–210. doi:10.1198/016214506000001086.

Bacci S, Pandolfi S, Pennoni F (2014). "A Comparison of Some Criteria for States Selection in the Latent Markov Model for Longitudinal Data." *Advances in Data Analysis and Classification*, **8**(2), 125–145. doi:10.1007/s11634-013-0154-2.

Bartolucci F, Bacci S, Pennoni F (2014a). "Longitudinal Analysis of Self-Reported Health Status by Mixture Latent Auto-Regressive Models." *Journal of the Royal Statistical Society C*, **63**(2), 267–288. doi:10.1111/rssc.12030.

Bartolucci F, 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**(486), 816–831. doi:10.1198/jasa.2009.0107.

Bartolucci F, Farcomeni A (2015). "Information Matrix for Hidden Markov Models with Covariates." *Statistics and Computing*, **25**(3), 515–526. doi:10.1007/s11222-014-9450-8.

Bartolucci F, Farcomeni A, Pennoni F (2013). *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC, Boca Raton.

Bartolucci F, Farcomeni A, Pennoni F (2014b). "Latent Markov Models: A Review of a General Framework for the Analysis of Longitudinal Data with Covariates." *TEST*, **23**(3), 433–465. doi:10.1007/s11749-014-0381-7.

Bartolucci F, Lupparelli M, Montanari GE (2009). "Latent Markov Model for Binary Longitudinal Data: An Application to the Performance Evaluation of Nursing Homes." *The Annals of Applied Statistics*, **3**(2), 611–636. doi:10.1214/08-aoas230.

Bartolucci F, Montanari GE, Pandolfi S (2015). "Three-Step Estimation of Latent Markov Models with Covariates." *Computational Statistics & Data Analysis*, **83**, 287–301. doi:10.1016/j.csda.2014.10.017.

Bartolucci F, Pandolfi S (2017). ***LMest**: Latent Markov Models with and without Covariates*. R package version 2.4.1, URL https://CRAN.R-project.org/package=LMest.

Bartolucci F, Pennoni F, Francis B (2007). "A Latent Markov Model for Detecting Patterns of Criminal Activity." *Journal of the Royal Statistical Society A*, **170**(1), 151–132. doi:10.1111/j.1467-985x.2006.00440.x.

Bartolucci F, Pennoni F, Vittadini G (2011). "Assessment of School Performance through a Multilevel Latent Markov Rasch Model." *Journal of Educational and Behavioural Statistics*, **36**(4), 491–522. doi:10.3102/1076998610381396.

Bartolucci F, Pennoni F, Vittadini G (2016). "Causal Latent Markov Model for the Comparison of Multiple Treatments in Observational Longitudinal Studies." *Journal of Educational and Behavioral Statistics*, **41**(2), 146–179. doi:10.3102/1076998615622234.

Baum LE, Petrie T, Soules G, Weiss N (1970). "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." *The Annals of Mathematical Statistics*, **41**(1), 164–171. doi:10.1214/aoms/1177697196.

Bulla J, Bulla I (2013). ***hsmm**: Hidden Semi Markov Models*. R package version 0.4, URL https://CRAN.R-project.org/package=hsmm.

Davison AC, Hinkley DV (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

Dempster AP, Laird NM, Rubin DB (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society B*, **39**(1), 1–38.

Francis B, Liu J, Soothill K (2010). "Criminal Lifestyle Specialization: Female Offending in England and Wales." *International Criminal Justice Review*, **20**(2), 188–204. doi:10.1177/1057567710368942.

Harte D (2017). ***HiddenMarkov**: Hidden Markov Models*. R package version 1.8-8, URL https://CRAN.R-project.org/package=HiddenMarkov.

Helske J, Helske S (2017). ***seqHMM**: Hidden Markov Models for Life Sequences and Other Multivariate, Multichannel Categorical Time Series*. R Package Version 1.0.7, URL https://CRAN.R-project.org/package=seqHMM.

Himmelmann L (2010). ***HMM**: Hidden Markov Models*. R package version 1.0, URL https://CRAN.R-project.org/package=HMM.

Jackson CH (2011). "Multi-State Models for Panel Data: The **msm** Package for R." *Journal of Statistical Software*, **38**(8), 1–28. doi:10.18637/jss.v038.i08.

Juang BH, Rabiner LR (1991). "Hidden Markov Models for Speech Recognition." *Technometrics*, **33**(3), 251–272. doi:10.2307/1268779.

Little RJA, Rubin DB (2002). *Statistical Analysis with Missing Data*. 2nd edition. John Wiley & Sons, New York. doi:10.1002/9781119013563.

Maruotti A (2011). "Mixed Hidden Markov Models for Longitudinal Data: An Overview." *International Statistical Review*, **79**(3), 427–454. doi:10.1111/j.1751-5823.2011.00160.x.

Murphy K (1998). "Hidden Markov Model (**HMM**) Toolbox for MATLAB." URL http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html.

Oakes D (1999). "Direct Calculation of the Information Matrix via the EM Algorithm." *Journal of the Royal Statistical Society B*, **61**(2), 479–482. doi:10.1111/1467-9868.00188.

O'Connell J, Højsgaard S (2011). "Hidden Semi Markov Models for Multiple Observation Sequences: The **mhsmm** Package for R." *Journal of Statistical Software*, **39**(4), 1–22. doi:10.18637/jss.v039.i04.

Pennoni F (2014). *Issues on the Estimation of Latent Variable and Latent Class Models: With Applications in the Social Sciences*. Scholars' Press, Saarbücken.

Pennoni F, Vittadini G (2013). "Two Competing Models for Ordinal Longitudinal Data with Time-Varying Latent Effects: An Application to Evaluate Hospital Efficiency." *Quaderni di Statistica*, **15**, 53–68.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rubin DB (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, **66**(5), 688–701. doi:10.1037/h0037350.

Schwarz G (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, **6**(2), 461–464. doi:10.1214/aos/1176344136.

Stiglitz JE, Amartya S, Fitoussi JP (2010). "Report by the Commission on the Measurement of Economic Performance and Social Progress." *Technical Report 1*, Paris: Commission on the Measurement of Economic Performance and Social Progress.

The MathWorks Inc (2014). *MATLAB – The Language of Technical Computing, Version R2014b*. Natick. URL http://www.mathworks.com/products/matlab/.

Turner R (2016). ***hmm.discnp**: Hidden Markov Models with Discrete Non-Parametric Observation Distributions*. R Package Version 0.2-4, URL https://CRAN.R-project.org/package=hmm.discnp.

Van de Pol F, Langeheine R (1990). "Mixed Markov Latent Class Models." *Sociological Methodology*, **20**, 213–247. doi:10.2307/271087.

Vermunt JK, Langeheine R, Böckenholt U (1999). "Discrete-Time Discrete-State Latent Markov Models with Time-Constant and Time-Varying Covariates." *Journal of Educational and Behavioral Statistics*, **24**(2), 179–207. doi:10.2307/1165200.

Vermunt JK, Magidson J (2016). "Technical Guide for **Latent GOLD** 5.1: Basic, Advanced, and Syntax." URL http://www.statisticalinnovations.com/.

Visser I, Speekenbrink M (2010). "**depmixS4**: An R Package for Hidden Markov Models." *Journal of Statistical Software*, **36**(7), 1–21. doi:10.18637/jss.v036.i07.

Viterbi AJ (1967). "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm." *IEEE Transactions on Information Theory*, **13**(2), 260–269. doi:10.1109/tit.1967.1054010.

Wiggins LM (1955). "Mathematical Models for the Analysis of Multi-Wave Panels." In *Ph.D. Dissertation*. Columbia University, Ann Arbor.

Wiggins LM (1973). *Panel Analysis: Latent Probability Models for Attitude and Behaviour Processes*. Elsevier, Amsterdam.

Zucchini W, MacDonald IL (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC, Boca Raton. doi:10.1201/9781420010893.

Affiliation:

Francesco Bartolucci, Silvia Pandolfi
Department of Economics
University of Perugia
Via A. Pascoli, 06123, Perugia, Italy
E-mail: francesco.bartolucci@unipg.it, silvia.pandolfi@unipg.it
URL: https://sites.google.com/site/bartstatistics/
https://sites.google.com/site/spandolfihome/

Fulvia Pennoni
Department of Statistics and Quantitative Methods
Via Bicocca degli Arcimboldi 8, 20126, Milano, Italy
E-mail: fulvia.pennoni@unimib.it
URL: https://sites.google.com/view/fulviapennoni

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
October 2017, Volume 81, Issue 4 *Submitted:* 2015-01-26
doi:10.18637/jss.v081.i04 *Accepted:* 2016-08-26