PACKAGE LMest FOR LATENT MARKOV ANALYSIS OF LONGITUDINAL CATEGORICAL DATA

Francesco Bartolucci[1], Silvia Pandofi[1], and Fulvia Pennoni[2]

[1] Department of Economics, University of Perugia (e-mail: francesco.bartolucci@unipg.it, silvia.pandolfi@unipg.it)

[2] Department of Statistics and Quantitative Methods, University of Milano-Bicocca (e-mail: fulvia.pennoni@unimib.it)

ABSTRACT: Latent Markov (LM) models represent an important tool of analysis of longitudinal data. We illustrate the main functions of the R package LMest that is tailored to fit the basic LM model, and some of its extended formulations, on longitudinal categorical data. The illustration is based on empirical analyses of datasets from a socio-economic perspective.

KEYWORDS: Expectation-Maximization algorithm, forward-backward recursions, mixed models, time-varying unobserved heterogeneity.

1 Introduction

We illustrate the R package LMest (V2.3, available from http://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 strongly related to the book of Bartolucci *et al.* , 2013, where these models are illustrated from the methodological point of view. The package is described in detail in the devoted paper of Bartolucci *et al.* , 2017, to which we refer the reader for a more comprehensive overview.

The LMest package has several distinguishing features over the existing R packages for similar models. In particular, it is designed to deal with longitudinal data, that is, with (even many) i.i.d. replicates of (usually short) sequences of data, and it can be used with univariate and multivariate categorical outcomes. The package also allows us to deal with missing responses, including drop-out and non-monotonic missingness, under the missing-at-random assumption. Moreover, standard errors for the parameter estimates are obtained by exact computation of the information matrix or through reliable numerical approximations of this matrix. Finally, computationally efficient algorithms are implemented for estimation and prediction of the latent states, by relying on suitable Fortran routines.

In the next sections we show how, through the main functions of the `LMest` package, we can estimate the basic LM model and LM models with individual covariates; these covariates are included in the model through suitable parameterizations. In addition, we briefly show how to perform model selection and local and global decoding. For reasons of space we just mention that additional discrete random effects can be used to formulate mixed LM models, which are estimable through the `R` function `est_lm_mixed`. In this case, the initial and transition probabilities of the latent process are allowed to vary across different latent subpopulations defined by an additional discrete latent variable.

2 The general latent Markov model formulation

In the following we provide a brief review of the main assumptions of LM models for categorical longitudinal data. For a generic sample unit we consider a vector $\mathbf{Y}^{(t)}$ 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 $\tilde{\mathbf{Y}}$ be the vector obtained by stacking $\mathbf{Y}^{(t)}$ for $t = 1,\ldots,T$. When available, we denote by $\mathbf{X}^{(t)}$ the vector of individual covariates available at the t-th time occasion and by $\tilde{\mathbf{X}}$ the vector of all the individual covariates. As usual, capital letters are used to denote random variables or vectors and small letters for their realizations.

The general LM model formulation assumes the existence of a latent process, denoted by $\mathbf{U} = (U^{(1)},\ldots,U^{(T)})$, 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 $\mathbf{Y}^{(1)},\ldots,\mathbf{Y}^{(T)}$ are assumed to be conditionally independent given the latent process $\mathbf{U}$. Moreover, the elements $Y_j^{(t)}$ of $\mathbf{Y}^{(t)}$ are conditionally independent given $U^{(t)}$. Parameters of the *measurement model* are the conditional response probabilities $\phi_{jy|u\mathbf{x}}^{(t)} = p(Y_j^{(t)} = y|U^{(t)} = u, \mathbf{X}^{(\mathbf{t})} = \mathbf{x})$, whereas parameters of the *structural model* are the initial and transition probabilities of the latent process: $\pi_{u|\mathbf{x}} = p(U^{(1)} = u|\mathbf{X}^{(1)} = \mathbf{x})$, $\pi_{u|\bar{u}\mathbf{x}}^{(t)} = p(U^{(t)} = u|U^{(t-1)} = \bar{u}, \mathbf{X}^{(t)} = \mathbf{x})$.

Maximum likelihood estimation of the LM models is performed through the Expectation-Maximization (EM) algorithm based on forward-backward recursions (Baum *et al.* , 1970; Dempster *et al.* , 1977). This algorithm is based on alternating two steps, consisting in obtaining the posterior distribution of the latent states given the observed data (E-step) and updating the parameters by maximizing the expected value of the *complete data log-likelihood* (M-step).

3 Basic latent Markov model

The basic LM model rules out individual covariates and assumes that the conditional response probabilities are time homogenous. In symbols, we have that $\phi^{(t)}_{jy|u\mathbf{x}} = \phi_{jy|u}$, $\pi_{u|\mathbf{x}} = \pi_u$, and $\pi^{(t)}_{u|\bar{u}\mathbf{x}} = \pi^{(t)}_{u|\bar{u}}$. The model is fitted by function `est_lm_basic`, which requires the following main input arguments:

- `S`: array of response configurations of dimension `n` $\times$ `TT` (number of time occasions) $\times$ `r`; missing responses are indicated with `NA`;
- `yv`: vector of frequencies of the response 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;
- `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.

The output may be shown through the usual `print` and `summary` commands, which display, among others, the maximum log-likelihood, the estimated conditional response probabilities (`Psi`) and the estimated initial (`piv`) and transition probabilities (`Pi`). The illustration of this function is based on the survey data provided by the Russia Longitudinal Monitoring Survey*, by considering an ordinal response variable related to job satisfaction measured on a scale ranging from 1 ("absolutely satisfied") to 5 ("absolutely not satisfied").

A suitable function `search.model.LM` allows us to select the value of k on the basis of the observed data, by considering different initializations of the EM algorithm which is used to maximize the log-likelihood function. In this way, we can address jointly the problems of model selection and the multimodality of the likelihood function. This function can also be applied to the models illustrated in the following two sections.

4 Covariates in the measurement model

When the individual covariates are included in the measurement model, the latent variables account for the unobserved heterogeneity, that is, the heterogene-

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

ity between individuals that we cannot explain on the basis of the observable covariates. In this case, the conditional distribution of the response variables given the latent states may be parameterized by generalized logits.

In formulating the model we can rely on the following parameterization based on global logits for a single ordinal response variable with c categories:

$$\log \frac{\phi^{(t)}_{y|u\mathbf{x}} + \ldots + \phi^{(t)}_{c-1|u\mathbf{x}}}{\phi^{(t)}_{0|u\mathbf{x}} + \ldots + \phi^{(t)}_{y-1|u\mathbf{x}}} = \mu_y + \alpha_u + \mathbf{x}'\boldsymbol{\beta}, \quad u = 1,\ldots,k,\ y = 1,\ldots,c-1. \quad (1)$$

In the above expression, the μ_y are cut-points, the α_u are the support points corresponding to each latent state, and $\boldsymbol{\beta}$ is the vector of regression parameters for the covariates. The model is estimated by function `est_lm_cov_manifest`, which requires the following main input arguments:

- `S`: matrix of the observed 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 based on parameterization (1). 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 (Bartolucci *et al.* , 2014);
- `q`: number of support points of the AR(1) structure mentioned above.

We illustrate the above functions through the analysis of the survey data provided by Health and Retirement Study conducted by the University of Michigan[†]. The main response of interest is an ordinal variable related to self evaluation of the health status measured on a scale ranging from 1 ("excellent") to 5 ("poor").

5 Covariates in the latent model

When the covariates are included in the latent model, we suppose that the response variables measure a certain individual characteristic of interest (e.g.,

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

well-being), the evolution of which is represented by the latent Markov process. In fact, this characteristic is not directly observable and we assume it may evolve over time. In such a case, the main research interest is in measuring the effect of covariates on the latent distribution. In particular, the individual covariates are assumed to affect the initial and transition probabilities of the LM chain through the following multinomial logit parameterization:

$$\begin{aligned} \log(\pi_{u|\mathbf{x}}/\pi_{1|\mathbf{x}}) &= \beta_{0u} + \mathbf{x}'\boldsymbol{\beta}_{1u}, \quad u = 2,\ldots,k, \\ \log(\pi^{(t)}_{u|\bar{u}\mathbf{x}}/\pi^{(t)}_{\bar{u}|\bar{u}\mathbf{x}}) &= \gamma_{0\bar{u}u} + \mathbf{x}'\boldsymbol{\gamma}_{1\bar{u}u}, \quad t = 2,\ldots,T,\ \bar{u},u = 1,\ldots,k,\ \bar{u} \neq u, \end{aligned}$$

where $\boldsymbol{\beta}_u = (\beta_{0u}, \boldsymbol{\beta}'_{1u})'$ and $\boldsymbol{\gamma}_{\bar{u}u} = (\gamma_{0\bar{u}u}, \boldsymbol{\gamma}'_{1\bar{u}u})'$ are parameter vectors to be estimated, which are collected in the matrices $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$. A more parsimonious model for the transition probabilities is also allowed which is based on the difference between two sets of parameters of the type

$$\log(\pi^{(t)}_{u|\bar{u}\mathbf{x}}/\pi^{(t)}_{\bar{u}|\bar{u}\mathbf{x}}) = \gamma_{0\bar{u}u} + \mathbf{x}'(\boldsymbol{\gamma}_{1u} - \boldsymbol{\gamma}_{1\bar{u}}). \tag{2}$$

In the present case, the covariates are excluded from the measurement model and we adopt the constraint $\phi^{(t)}_{jy|u\mathbf{x}} = \phi_{jy|u}$. The above parameterizations are implemented in the `R` function `est_lm_cov_latent`, which 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 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;
- `param`: type of parameterization for the transition probabilities, coded as `param = "multilogit"` (default) for the multinomial logit parameterization and as `param = "difflogit"` for the parameterization based on the difference between two sets of parameters as in (2).

6 Local and global decoding

The prediction of the sequence of the latent states for a certain sample unit on the basis of the data observed for this unit can be performed by using function

`decoding`. In particular, the EM algorithm directly provides the estimated posterior probabilities of $U^{(t)}$, namely $p(U^{(t)} = u|\tilde{\mathbf{X}} = \tilde{\mathbf{x}}, \tilde{\mathbf{Y}} = \tilde{\mathbf{y}})$, for every covariate and response configuration $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ observed at least once. These probabilities can be directly maximized to obtain a prediction of the latent state of every subject at each time occasion t; this is the so-called *local decoding*. In order to track the latent state of a subject across time, the *a posteriori* most likely sequence of states must be obtained, through the so-called *global decoding*, which is based on an adaptation of the Viterbi algorithm (Viterbi, 1967).

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 function `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 function `est_lm_cov_latent`).

References

BARTOLUCCI, F., FARCOMENI, A., & PENNONI, F. 2013. *Latent Markov Models for Longitudinal Data*. Boca Raton: Chapman and Hall/CRC press.

BARTOLUCCI, F., BACCI, S., & PENNONI, F. 2014. Longitudinal Analysis of Self-reported Health Status by Mixture Latent Auto-regressive Models. *Journal of the Royal Statistical Society: Series C*, **63**, 267–288.

BARTOLUCCI, F., PANDOLFI, S., & PENNONI, F. 2017. LMest: An R Package for Latent Markov Models for Longitudinal Categorical Data. *Journal of Statistical Software*, Accepted.

BAUM, L. E., 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**, 164–171.

DEMPSTER, A. P., LAIRD, N. M., & RUBIN, D. B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society B*, **39**, 1–38.

VITERBI, A. J. 1967. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, **13**, 260–269.