Modeling Longitudinal Data by Latent Markov Models with Application to Educational and Psychological Measurement

Francesco Bartolucci

Abstract I review a class of models for longitudinal data, showing how it may be applied in a meaningful way for the analysis of data collected by the administration of a series of items finalized to educational or psychological measurement. In this class of models, the unobserved individual characteristics of interest are represented by a sequence of discrete latent variables, which follows a Markov chain. Inferential problems involved in the application of these models are discussed considering, in particular, maximum likelihood estimation based on the Expectation-Maximization algorithm, model selection, and hypothesis testing. Most of these problems are common to hidden Markov models for time-series data. The approach is illustrated by different applications in education and psychology.

Keywords Forward and Backward recursions • Expectation-Maximization algorithm • Hidden Markov models • Rasch model

1 Introduction

Among the statistical models for the analysis of longitudinal data which are available in the literature (Diggle et al. 2002; Fitzmaurice et al. 2009; Frees 2004), those based on a latent Markov chain have a special role when the main interest is on individual characteristics which are not directly observable. Such models are based on assumptions similar to those of hidden Markov models; for a recent review see Zucchini and MacDonald (2009). Applications focusing on individual characteristics which are not directly observable usually arise in education and psychology, when these characteristics are measured through the responses to a

F. Bartolucci (✉)
Department of Economics, University of Perugia, Perugia, Italy
e-mail: bart@stat.unipg.it

D. Vicari et al. (eds.), *Analysis and Modeling of Complex Data in Behavioral and Social Sciences*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-06692-9_2,

series of items, also corresponding to specific tasks, administered at consecutive occasions.

One of the first authors who dealt with models for longitudinal data which are based on a latent Markov chain, LM models for short, was Wiggins (1973); see Bartolucci et al. (2013) for a review. The basic assumption of these models is that the response variables, corresponding to test items in the present context, are conditionally independent given the Markov chain. Due to the adoption of suitable parametrizations, such as that characterizing the Rasch model (Rasch 1961) or that characterizing the graded response model (Samejima 1969, 1996), the states of the chain may be interpreted as different levels of ability or tendency towards a certain behavior; these states identify different latent classes in the population from that the observed sample comes. Of particular interest is the possibility of estimating probabilities of transition between the classes; these probabilities may be allowed to depend on individual covariates or experimental factors.

Aim of the present paper is to review LM models in the context of educational and psychological measurement, also discussing likelihood based inference. In particular, maximum likelihood estimation may be performed by an Expectation-Maximization (EM) algorithm implemented by the Baum-Welch forward and backward recursions (Baum et al. 1970), which have been developed in the literature on hidden Markov models for time-series data. Moreover, model selection, regarding in particular the number of states (or latent classes), is based on information criteria, such as the Akaike Information Criterion (AIC) (Akaike 1973) or the Bayesian Information Criterion (BIC) (Schwarz 1978). Finally, hypothesis testing may be based on the likelihood ratio statistic which, under certain regularity conditions (Bartolucci 2006), has a null asymptotic distribution of chi-bar-squared type, that is, a finite mixture of chi-squared distributions.

The remainder of this paper is organized as follows. Basic assumptions of LM models for longitudinal data are illustrated in Sect. 2 with focus on educational and psychological contexts. Likelihood inference, regarding in particular parameter estimation, model selection, and testing, is dealt with in Sect. 3. Finally, some applications of the discussed statistical models in educational and psychological fields are briefly described in Sect. 4.

2 Model Assumptions and Likelihood Inference

Suppose that a set of J items is administered at T consecutive occasions to a sample of n subjects and let $Y_{ij}^{(t)}$, $i = 1, \ldots, n$, $j = 1, \ldots, J$, $t = 1, \ldots, T$, denote the random variable for the response to item j at occasion t by subject i. This random variable is binary in the case of dichotomously-scored items and categorical, with more than two categories, in the case of polytomously-scored items. In the second case the response categories are typically ordered. In any case, the number of response categories is denoted by c and they are labelled from 0 to $c - 1$.

In the above context, the basic assumption of LM models is that, for every sample unit i, the random variables $Y_{ij}^{(t)}$, $j = 1, \ldots, J$, $t = 1, \ldots, T$, are conditionally independent given a sequence of latent variables $U_i^{(1)}, \ldots, U_i^{(T)}$. These latent variables identify latent classes of subjects sharing common characteristics. Moreover, the distribution of these variables is assumed to follow a first-order Markov chain with k states and homogeneous or non-homogenous transition probabilities. In particular, the initial probabilities are denoted by $\pi_u = p(U_i^{(1)} = u)$, $u = 1, \ldots, k$, whereas the transition probabilities are denoted by $\pi_{v|u}^{(t)} = p(U_i^{(t)} = v | U_i^{(t-1)} = u)$, $u, v = 1, \ldots, k$, $t = 2, \ldots, T$, in the time non-homogenous case. These probabilities are collected in the $k \times k$ *transition matrix* $\boldsymbol{\Pi}^{(t)}$, with the index u running by row and the index v running by column. In the time-homogenous case we have $\pi_{v|u}^{(t)} = \pi_{v|u}$ for $t = 2, \ldots, T$.

In educational and psychological measurement, the interpretation of the model is enforced by letting every manifest variable $Y_{ij}^{(t)}$ to depend only on the corresponding latent variable $U_i^{(t)}$ according to a suitable parametrization for the conditional distribution of the former given the latter. For instance, with dichotomously-scored test items ($c = 2$), we may adopt a Rasch parametrization (Bartolucci et al. 2008; Rasch 1961) by requiring that

$$\log \frac{p(Y_{ij}^{(t)} = 1 | U_i^{(t)} = u)}{p(Y_{ij}^{(t)} = 0 | U_i^{(t)} = u)} = \psi_u - \beta_j, \quad j = 1, \ldots, J, \ u = 1, \ldots, k,$$

for all t, where ψ_u is a parameter interpreted as the ability level of the examinees in latent class u. Moreover, β_j is the difficulty level of item j. The constraint that this difficulty level does not vary with t makes sense only if the same battery of items is administered at all occasions and then, in certain contexts, this constraint must be relaxed in a suitable way.

The above parametrization may be extended in a natural way to the case of $c > 2$ response categories. If these categories are ordered, we may assume a parametrization which is also adopted in the graded response model (Samejima 1969, 1996), that is,

$$\log \frac{p(Y_{ij}^{(t)} \geq y | U_i^{(t)} = u)}{p(Y_{ij}^{(t)} < y | U_i^{(t)} = u)} = \psi_u - \beta_{jy},$$
$$j = 1, \ldots, J, \ u = 1, \ldots, k, \ y = 1, \ldots, c-1,$$

where the parameters have an interpretation similar to the previous one. In particular, ψ_u is still interpreted as the ability level of subjects in latent class u. Note that, in order to avoid a wrong model specification, the β_{jy} parameters are increasing ordered in y for all j. Under such a parametrization, the initial probabilities of the latent Markov chain allow us to study the distribution of the ability (or another latent trait of interest) among the examinees at the beginning of the period of observation.

Moreover, the probabilities of transition between the latent classes allow us to study the evolution of the ability, even depending on particular individual covariates or factors (e.g., teaching method).

It has to be clear that we can also use the probabilities

$$\phi_{jy|u} = p(Y_{ij}^{(t)} = y|U_i^{(t)} = u), \quad t = 1, \ldots, T, \ u = 1, \ldots, k, \ y = 0, \ldots, c-1,$$

as free parameters, without assuming any specific parametrization. In this case, if covariates are ruled out and the transition probabilities are not constrained to be homogeneous, then a multivariate version of the so-called *basic LM model* (Bartolucci et al. 2013) results. However, the interpretation of the latent classes may be more difficult in terms, for instance, of different levels of ability, since it is not ensured that these classes are monotonically ordered according to the conditional distribution of the response variables.

More complex formulations of the LM models, with respect to those described above, are available in the literature; see Bartolucci et al. (2013) for a complete review about these models. A typical extension of interest is for the inclusion of individual covariates, possibly time varying, that affect the initial and the transition probabilities of the latent Markov chain. A more complex extension is to deal with multilevel longitudinal data in which subjects are collected in clusters, such as students in school. In this case, further latent variables are used to account for the effect of each cluster on the response variables; see Bartolucci and Lupparelli (2012) and the references therein.

Finally, it is worth noting that the modeling framework illustrated in this section may be also applied to the case of unbalanced panel settings in which the number of time occasions is not the same for all subjects, due to some forms of ignorable drop-out. In this case, the number of time occasions for subject i is indicated by T_i and must be substituted to T in the expressions above.

3 Likelihood Inference

Estimation of an LM model is usually performed by maximizing its likelihood. In the case of independent sample units and when individual covariates are ruled out, this likelihood has logarithm

$$\ell(\boldsymbol{\theta}) = \sum_i \log p(\mathbf{y}_i), \tag{1}$$

where $\boldsymbol{\theta}$ is the vector of all model parameters and $p(\mathbf{y}_i)$ is the manifest probability of the sequence of item responses provided by subject i, which are collected in the vector $\mathbf{y}_i$. This probability is in practice computed by a suitable recursion which is well known in the hidden Markov literature and has been set up by Baum and Welch (Baum et al. 1970; Welch 2003; Zucchini and MacDonald 2009).

In order to maximize the log-likelihood in (1), the main tool is the Expectation-Maximization (EM) algorithm (Baum et al. 1970; Dempster et al. 1977), which is based on the so-called *complete data log-likelihood*, that is, the log-likelihood that we could compute if we knew the value of $U_i^{(t)}$ for every subject i and time occasion t. When the conditional response probabilities do not depend on t, as in the parametrizations illustrated in Sect. 2, this function may be expressed as:

$$\ell^*(\boldsymbol{\theta}) = \sum_j \sum_t \sum_u \sum_y a_{juy}^{(t)} \log \phi_{jy|u} + \sum_u b_u^{(1)} \log \pi_u + \sum_{t>1} \sum_u \sum_v b_{uv}^{(t)} \log \pi_{v|u}^{(t)}, \tag{2}$$

where $a_{juy}^{(t)}$ is the frequency of subjects responding by y to item j at occasion t and belonging to latent state u at the same occasion, $b_u^{(t)}$ is the frequency of subjects in latent state u at occasion t, and $b_{uv}^{(t)}$ is the number of transitions from latent state u at occasion $t-1$ to state v at occasion t.

The EM algorithm alternates two steps until convergence in the model log-likelihood $\ell(\boldsymbol{\theta})$. The E-step consists of computing the conditional expected value of every frequency in (2) given the observed data and the current value of the parameters. The M-step consists of maximizing the function $\ell^*(\boldsymbol{\theta})$ in which these frequencies have been substituted by the corresponding expected values. An implementation of this algorithm, and of a bootstrap algorithm to compute standard errors for the parameter estimates, is available in the package `LMest` for `R` (Bartolucci 2012).

With more complex formulations of the LM model, the EM algorithm is still used for parameter estimation. In particular, in the presence of individual covariates affecting the initial and transition probabilities of the Markov chain, the log-likelihood to be maximized is expressed as in (1) with $p(\mathbf{y}_i)$ substituted by $p(\mathbf{y}_i|\mathbf{x}_i)$. The latter may be computed by the same recursion mentioned above (Baum et al. 1970), while an extended version of the complete log-likelihood in (2) is used within the algorithm; then the EM algorithm is not much more complex than the one used for the model without covariates. On the other hand, estimating a multilevel LM model requires a more complex implementation of the EM algorithm and then we refer the reader to specific articles, see in particular (Bartolucci and Lupparelli 2012; Bartolucci et al. 2011), for a detailed description.

A crucial point in applying an LM model is the choice of the number of latent states, k. This choice is usually accomplished by an information criterion based on penalizing the maximum value of the log-likelihood. For instance, BIC leads to selecting the value of k which minimizes $BIC = -2\ell(\hat{\boldsymbol{\theta}}) + g \log(n)$, where g is the number of free parameters (Schwarz 1978). Alternatively, we can use AIC (Akaike 1973), which is based on an index similar to the previous one, where the penalization term is $2g$ instead of $g \log(n)$. BIC usually leads to more parsimonious models and it is typically preferred to AIC.

Another important point concerns how to test hypotheses of interest on the parameters. In many cases, the standard asymptotic theory may be employed to

test these hypotheses on the basis of a likelihood ratio statistic. In particular, the null asymptotic distribution turns out to be of chi-squared type. However, in certain cases, and in particular when the hypothesis is expressed through linear constraints on the transition probabilities, a more complex asymptotic distribution results, that is, the chi-bar-squared distribution. This distribution may be expressed as a mixture of standard chi-squared distributions with suitable weights, which may be computed analytically in certain simple cases or by a suitable Monte Carlo method in general (Bartolucci 2006; Shapiro 1988). In particular, this distribution arises when testing that transitions between latent states are not allowed and then an LM model specializes into a latent class model (Bartolucci 2006). This hypothesis is simply expressed by constraining the transition matrices to be equal to an identity matrix. A less restrictive constraint is that these matrices are triangular, so that a certain type of evolution of the latent trait is considered. For instance, with $k = 3$ states, these two constraints are expressed as follows:

$$\boldsymbol{\Pi}^{(t)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\Pi}^{(t)} = \begin{pmatrix} \pi_{1|1}^{(t)} & \pi_{2|1}^{(t)} & \pi_{3|1}^{(t)} \\ 0 & \pi_{2|2}^{(t)} & \pi_{3|2}^{(t)} \\ 0 & 0 & 1 \end{pmatrix}.$$

When the latent states correspond to increasing levels of ability in an educational application, the second matrix corresponds to the hypothesis that the ability never decreases during the period of observations. Many other constraints on the transition matrices may be of interest, such as that of symmetry.

4 Applications in Educational and Psychological Measurement

In order to illustrate how LM models may be used in the educational and psychological measurement, in the following we describe some benchmark applications.

4.1 Evolution of Ability Level in Mathematics

The first application concerns testing the hypothesis of absence of tiring or learning-through-training phenomena during the administration of a series of 12 items on Mathematics (Bartolucci 2006; Bartolucci et al. 2008). In this case, we are not properly dealing with longitudinal data since all items were administered at the same occasion; however, an LM model makes sense since these items were administered in the same order to all examinees. In particular, the adopted LM model is based on a Rasch parametrization.

The main conclusion of the study is that there is not an evolution of the ability during the administration of the test items and then there is no evidence of the existence of tiring or learning-through-training phenomena. This conclusion is reached by comparing, by a likelihood ratio test statistic, the LM Rasch model with homogenous transition probabilities with a constrained version of this model in which the transition matrices are equal to an identity matrix (null hypothesis). The corresponding p-value, which is larger than 0.05, is computed on the basis of the chi-bar-squared distribution and leads to the conclusion that there is not enough evidence against the null hypothesis.

4.2 Evolution of Psychological Traits in Children

The second application (Bartolucci and Solis-Trapala 2010) concerns data collected through a psychological experiment based on tests which were administered at different occasions to pre-school children in order to measure two types of ability: inhibitory control and attentional flexibility. In this case, the model is more complex than the one adopted for the first application since it is multidimensional and then subjects are classified in latent classes according to different abilities. Moreover, more complex parametrizations than the Rasch parametrization are adopted and transition probabilities are suitably formulated so as to account for certain experimental features. In particular, the maximum number of items administered to the same child is 132; these items were administered in three separated periods of time.

This application led to the conclusion that the two abilities must be conceptualized as distinct constructs, and so a unidimensional latent variable model cannot be validly used in this context. Moreover, it was demonstrated that these abilities develop at an early age and that there are different dynamics within different sequences of task administration, with mild tiring effects within certain sequences and learning-through-training phenomena concerning other sequences.

In dealing with this applications, the authors also fitted an extended version of the LM model in which the Markov chain is of second order. This was reformulated as a first-order model with an extended state space having k^2 elements. However, the data did not provide evidence in favor of this second-order extension, meaning that the ability level at a given occasion only depends on that at the previous occasion.

4.3 Evaluation of School Effectiveness on Proficiency of Students

The third application (Bartolucci et al. 2011) involves a multilevel LM model based on a Rasch parameterization, which is used to analyze data collected by a series of test items administered to middle-school students. The overall number of items is

97; these items were administered at the end of each of the 3 years of schooling. The adopted model takes into account that students (level 1 units) are collected in schools (level 2 units) by the inclusion of further latent variables. This extension, already mentioned in Sect. 2, allows us to evaluate the effect of every school and then allows us to perform an analysis of the school effectiveness, also considering characteristics such as the type of school (e.g., public or not).

As a main result, the study found evidence of four different typologies of school which have a different effect on the ability level of their students and on the way in which this ability evolves across time. These typologies cannot be easily ordered because the effect is not in general constant across time. However, it emerges that most public schools have an intermediate positive effect on the proficiency of their students. On the other hand, a polarization is observed for non-public schools with some of them which have poor performance, while the others have very good performance.

This application may be considered as one attempt to implement an LM model having a causal perspective, which may be then used for evaluation purposes; for a similar application see Bartolucci et al. (2009). This is a promising field of application of LM models since, in a single framework, these models allow us to evaluate the effect of certain factors, not only on the distribution of a characteristic of interest, but also on its evolution, even when in a longitudinal setting this characteristic is not directly observable but it is observed through a series of time-specific response variables.

References

Akaike, H. (1973). Information theory and extension of the maximum likelihood principle. In B. N. Petrov & Csaki, F. (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society: Series B, 68*, 155–178.

Bartolucci, F. (2012). Package LMest for R, available via CRAN at http://cran.r-project.org/web/packages/LMest/index.html.

Bartolucci, F., & Lupparelli, M. (2012). Nested hidden Markov chains for modeling dynamic unobserved heterogeneity in multilevel longitudinal data. arXiv:1208.1864.

Bartolucci, F., & Solis-Trapala, I. (2010). Multidimensional latent Markov models in a developmental study of inhibitory control and attentional flexibility in early childhood. *Psychometrika, 75*, 725–743.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2013). *Latent markov models for longitudinal data*. Boca Raton: Chapman and Hall/CRC.

Bartolucci, F., Lupparelli, M., & Montanari, G. E. (2009). Latent Markov model for longitudinal binary data: an application to the performance evaluation of nursing homes. *Annals of Applied Statistics, 3*, 611–636.

Bartolucci, F., Pennoni, F., & Lupparelli, M. (2008). Likelihood inference for the latent Markov Rasch model. In C. Huber, N. Limnios, M. Mesbah, & M. Nikulin (Eds.), *Mathematical methods for survival analysis, reliability and quality of life* (pp. 239–254). London: Wiley.

Bartolucci, F., Pennoni, F., & Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics, 36*, 491–522.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *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: Series B, 39*, 1–38.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford: Oxford University Press.

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

Frees, E. W. (2004). *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge: Cambridge University Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability, 4*, 321–333.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometrika monograph, 17). Richmond, VA: Psychometric Society.

Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika, 23*, 17–35.

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

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review, 56*, 49–62.

Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter, 53*, 1–13.

Wiggins, L. M. (1973). *Panel analysis: latent probability models for attitude and behaviour processes*. Amsterdam: Elsevier.

Zucchini, W., & MacDonald, I. L. (2009). *Hidden Markov models for time series: an introduction using R*. New York: Springer.