Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities

Francesco Bartolucci*

October 21, 2005

Abstract

For a class of latent Markov (LM) models for discrete variables having a longitudinal structure, we introduce an approach for formulating and testing linear hypotheses on the transition probabilities of the latent process. For the maximum likelihood estimation of a LM model under hypotheses of this type, we outline an EM algorithm based on well-known recursions in the hidden Markov literature. We also show that, under certain assumptions, the asymptotic null distribution of the likelihood ratio (LR) statistic for testing a linear hypothesis on the transition probabilities of a LM model, against a less stringent linear hypothesis on the transition probabilities of the same model, is of chi-bar-squared type. As a particular case, we derive the asymptotic distribution of the LR statistic between a latent class model and its LM version, which may be used to test the hypothesis of absence of transition between latent states. The approach is illustrated through a series of simulations and two applications, the first of which is based on educational testing data collected within the National Assessment of Educational Progress 1996, and the second on data, concerning the use of marijuana, which have been collected within the National Youth Survey 1976-1980.

KEY WORDS: Boundary problem; Chi-bar-squared distribution; Constrained statistical inference; EM algorithm; Item Response Theory; Latent class model; Longitudinal data.

*Istituto di Scienze Economiche, Università di Urbino, Via Saffi, 42, 61029 Urbino, Italy, *email:* Francesco.Bartolucci@uniurb.it

1 Introduction

The latent Markov (LM) model was introduced by Wiggins (1973) for the analysis of longitudinal data and has been successfully applied in several fields, such as psychological and educational measurement (Langeheine, *et al.*, 1994, Vermunt *et al.*, 1999), sociology (Van de Pol and Langeheine, 1990, Humphreys, 1998, Mannan and Koval, 2003), medicine (Auranen *et al.*, 2000, Cook *et al.*, 2000), criminology (Bijleveld and Mooijaart, 2003) and the analysis of customer behaviour (Poulsen, 1990); for a review see Langeheine (2002). The basic assumption of this model is that any occasion-specific response variable depends only on a discrete latent variable, which in turn depends on the latent variables corresponding to the previous occasions according to a first-order Markov chain. Therefore, the LM model may be seen as an extension of a Markov chain model which allows for measurement errors, but also as an extension of the latent class (LC) model (Lazarsfeld and Henry, 1968, Goodman, 1974), in which the assumption that any subject belongs to the same latent class throughout the survey is suitably relaxed. As such, this model enables several aspects to be taken into account which characterize longitudinal studies, i.e. serial dependence between observations, measurement errors, and unobservable heterogeneity.

The LM model may be constrained in several ways in order to make it more parsimonious and easier to interpret. In particular, we deal with a class of LM models in which: (i) the conditional distribution of the response variables given the latent process may be formulated as in a generalized linear model; (ii) the latent process is time-homogeneous with transition probabilities that may be formulated as in a linear probability model. For the maximum likelihood (ML) estimation of these models, we outline an EM algorithm (Dempster *et al.*, 1977), in which the E-step is based on well-known recursions in the hidden Markov literature (see MacDonald and Zucchini, 1997) and the M-step is based on Newton-Raphson type iterative algorithms. In order to simplify the implementation of the algorithm, we often make use of the matrix notation and we give some simple rules to avoid numerical instability when long sequences of response variables are analysed. We also deal with the asymptotic distribution of the likelihood ratio (LR) statistic for testing linear hypotheses on the parameters of the model assumed on the transition probabilities of the latent process. Among these hypotheses, those formulated by constraining certain transition probabilities to be equal to 0 are of particular interest. An example is the hypothesis of no transition between latent states, which may be tested by comparing a LC model with a LM model based on a transition matrix with at least one off-diagonal element allowed to be positive. Note that hypotheses of this type cannot be tested within more conventional approaches, in which the transition probabilities are modelled through a *link function* based, for instance, on multinomial logits (Vermunt *et al.*, 1999). On the other hand, the parameters may

be on the boundary of the parameter space under these hypotheses and therefore we are dealing with an inferential problem under non-standard conditions (Self ad Liang, 1987), which typically occurs in constrained statistical inference (Silvapulle and Sen, 2004). By applying some results known in this literature, we show that the LR statistic in question has an asymptotic chi-bar-squared distribution under the null hypothesis. The finite sample accuracy of the inference based on this asymptotic distribution is investigated through a series of simulations.

In short, the main contribution of the present paper is the formulation of a general framework for likelihood inference under linear hypotheses on the transition probabilities of a class of LM models. Within this framework, we also allow for hypotheses expressed by constraining one or more probabilities to be equal to 0. Hypotheses of this type are of interest in many situations. To our knowledge, a framework like this has not been previously considered in the literature concerning LM models. It has not been explicitly considered either within other inferential approaches or within the hidden Markov literature, which is strongly related to the literature concerning LM models (for a review see Archer and Titterington, 2002, and Scott, 2002). In contrast, ML estimation of LM models has been considered since long time (Poulsen, 1982) and, under standard conditions, the use of the LR statistic for testing hypotheses on the parameters of a LM model has already been suggested (see, for instance, Visser, 2002).

The paper is organized as follows. The class of LM models is illustrated in Section 2 and ML estimation of these models in Section 3. In the latter we also deal with the Fisher information matrix and the problem of classifying subjects on the basis of the latent states. The asymptotic distribution of the LR statistic under linear hypotheses on the transition probabilities is dealt with in Section 4. Finally, two applications involving real datasets are illustrated in Section 5 and the main conclusions are drawn in Section 6.

2 A class of homogeneous latent Markov models

Let $\boldsymbol{Y} = \{Y_t,\ t = 1, \ldots, s\}$ be a sequence of s discrete random variables with support $\{1, \ldots, d\}$. The basic assumption of the LM model is that these variables are conditionally independent given a latent process $\boldsymbol{X} = \{X_t,\ t = 1, \ldots, s\}$ which follows a first-order Markov chain with state space $\{1, \ldots, c\}$. This assumption obviously makes sense only if the elements of $\boldsymbol{Y}$ correspond to measurements repeated at different occasions on the same subject. This is the case, not only of longitudinal data, but also of data derived from the administration of a set of test items to a group of subjects, which frequently arise in psychological and educational measurement (see also Example 1 in Section 2.1). In the latter case, a LM model may be validly applied only if the items are administered to all the subjects in the same order. Note also that assuming that the latent

process follows a first order Markov chain is equivalent to assuming that any latent variable X_t is conditionally independent of $X_1, \ldots, X_{t-2}$ given X_{t-1}. This assumption is seldom considered restrictive and, also because of its easy interpretation, is preferred to more complex assumptions on the dependence structure of the latent variables. In this paper, we also assume that the latent process is time-homogeneous, so that the transition probabilities $\pi_{x|w} = p(X_t = x|X_{t-1} = w)$, $w, x = 1, \ldots, c$, do not depend on t, whereas the initial probabilities $\lambda_x = p(X_1 = x)$, $x = 1, \ldots, c$, are completely unconstrained.

The assumptions above imply that the distribution of $\boldsymbol{X}$ may be expressed as

$$p(\boldsymbol{x}) = p(\boldsymbol{X} = \boldsymbol{x}) = \lambda_{x_1} \prod_{t>1} \pi_{x_t|x_{t-1}}, \tag{1}$$

while the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ may be expressed as

$$p(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}) = \prod_{t} \phi_{t,y_t|x_t}, \tag{2}$$

where $\phi_{t,y|x} = p(Y_t = y|X_t = x)$. Consequently, for the *manifest distribution* of $\boldsymbol{Y}$ we have the following expression

$$p(\boldsymbol{y}) = p(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{x_1} \phi_{1,y_1|x_1} \lambda_{x_1} \left(\sum_{x_2} \phi_{2,y_2|x_2} \pi_{x_2|x_1} \cdots \left(\sum_{x_s} \phi_{s,y_s|x_s} \pi_{x_s|x_{s-1}} \right) \right).$$

Using the matrix notation, this probability may also be expressed as $p(\boldsymbol{y}) = \boldsymbol{q}(\boldsymbol{y}_{\leqslant s})' \mathbf{1}_s$, where $\mathbf{1}_j$ is a column vector of j ones and, for $t = 1, \ldots, s$, $\boldsymbol{y}_{\leqslant t} = \{y_j,\ j = 1, \ldots, t\}$ and $\boldsymbol{q}(\boldsymbol{y}_{\leqslant t})$ is a column vector with elements $p(X_t = x, Y_1 = y_1, \ldots, Y_t = y_t)$, $x = 1, \ldots, c$. This vector may be computed through the following recursion

$$\boldsymbol{q}(\boldsymbol{y}_{\leqslant t}) = \begin{cases} \operatorname{diag}(\boldsymbol{\phi}_{1,y_1}) \boldsymbol{\lambda} & \text{if } t = 1 \\ \operatorname{diag}(\boldsymbol{\phi}_{t,y_t}) \boldsymbol{\Pi}' \boldsymbol{q}(\boldsymbol{y}_{\leqslant t-1}) & \text{otherwise} \end{cases}, \tag{3}$$

with $\boldsymbol{\lambda} = \{\lambda_x,\ x = 1, \ldots, c\}$ denoting the initial probability vector, $\boldsymbol{\Pi} = \{\pi_{x|w},\ w, x = 1, \ldots, c\}$ denoting the transition probability matrix and $\boldsymbol{\phi}_{t,y} = \{\phi_{t,y|x},\ x = 1, \ldots, c\}$ denoting the vector of the conditional probabilities of the response variables given the latent process. When s is very large, the elements of $\boldsymbol{q}(\boldsymbol{y}_{\leqslant t})$ could become so small as to vanish during recursion (3). To avoid this, we can multiply the quantity obtained at any step of this recursion by a suitable constant a. Thus, the vector computed at the t-th step of the recursion, denoted by $\boldsymbol{q}^*(\boldsymbol{y}_{\leqslant t})$, will be equal to $a^t \boldsymbol{q}(\boldsymbol{y}_{\leqslant t})$, while $p^*(\boldsymbol{y}) = \boldsymbol{q}^*(\boldsymbol{y}_{\leqslant s})' \mathbf{1}_s$ will be equal to $a^s p(\boldsymbol{y})$. As will be clear in the following, this dummy renormalization does not affect the estimation process.

In the *basic* LM model illustrated above, the conditional probabilities of the response variables given the latent process and the transition probabilities between the latent states are completely unconstrained. In many situations, however, it makes sense to parameterize these probabilities in order to give a more parsimonious and easily interpretable LM model.

2.1 Modelling conditional probabilities of the response variables

Let $\boldsymbol{\phi}$ be a column vector with elements $\phi_{t,y|x}$, $t = 1, \ldots, s$, $x = 1, \ldots, c$, $y = 1, \ldots, d$, arranged by letting the index y run faster than x and the latter run faster than t and let $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\phi})$ be a link function that maps $\boldsymbol{\phi}$ onto $\mathbb{R}^{sc(d-1)}$. We assume that

$$\boldsymbol{\eta} = \boldsymbol{Z}\boldsymbol{\gamma}, \tag{4}$$

where $\boldsymbol{Z}$ is a full rank design matrix and $\boldsymbol{\gamma}$ is a vector of parameters. Obviously, regardless of the specific choice of the link function, letting $\boldsymbol{Z} = \boldsymbol{I}_{sc(d-1)}$, with $\boldsymbol{I}_j$ denoting an identity matrix of dimension j, is equivalent to assuming that the conditional distribution of any Y_t given X_t is unconstrained. More interesting cases are illustrated below.

Example 1. In the case of binary variables, by parameterizing the conditional distribution of Y_t given $X_t = x$ through the logit $\eta_{t|x} = \log(\phi_{t,2|x}/\phi_{t,1|x})$ and assuming that

$$\eta_{t|x} = \psi_x - \delta_t, \qquad t = 1, \ldots, s, \quad x = 1, \ldots, c, \tag{5}$$

we can formulate a LM version of the Rasch (1961) model, the most well-known Item Response Theory (IRT) model (see, among others, Hambleton and Swaminathan, 1985, and De Boeck and Wilson, 2004). This model finds its natural application in educational assessment, when a group of examinees has to be assessed on the basis of the responses they provide to a series of test items. In this setting, the parameter ψ_x may be interpreted as the ability of the subjects in the x-th latent class and δ_t as the difficulty of the t-th item. In a similar way, we can formulate a multidimensional version of the Rasch model based on one of the parameterizations considered by Kelderman and Rijkes (1994). Note that, by assuming that the latent process follows a Markov chain, we allow a subject to move from one latent class to another; this is not allowed within the LC formulation of the Rasch model studied by Lindsay *et al.* (1991). Thus, we take into account the possibility of an implicit learning phenomenon or of an item which may provide clues for responding to other items. In these cases, the assumption of local independence (LI), which is crucial in IRT, is violated; for a discussion see Hambleton and Swaminathan (1985, Sec. 2.3). By means of the proposed approach, we can obviously test for the presence of phenomena of this type. □

Example 2. When the response variables have more than two levels, we can parameterize the conditional distribution of Y_t given $X_t = x$ through logits with the first category as baseline, i.e. $\eta_{t,y|x} = \log(\phi_{t,y+1|x}/\phi_{t,1|x})$, and assume that

$$\eta_{t,y|x} = \psi_x - \delta_{t,y}, \qquad t = 1, \ldots, s, \quad x = 1, \ldots, c, \quad y = 1, \ldots, d-1.$$

However, when the response variables have an ordinal nature, global logits are more suitable; these logits are based on the comparison between the survival function and the distribution function for any possible cut-point y, i.e.

$$\eta_{t,y|x} = \log \frac{\phi_{t,y+1|x} + \cdots + \phi_{t,d|x}}{\phi_{t,1|x} + \cdots + \phi_{t,y|x}}, \quad y = 1, \ldots, d-1;$$

see also Samejima (1996). □

In the present approach, we also allow for inequality constraints of the type $\boldsymbol{K}\boldsymbol{\gamma} \geqslant \mathbf{0}_k$ on the parameters $\boldsymbol{\gamma}$, where $\boldsymbol{K}$ is a full rank matrix with k rows and $\mathbf{0}_j$ denotes a vector of j zeros. The main use of these constraints is for making the latent states uniquely identifiable.

Example 3. If the same logit link function considered in Example 1 is used to parametrize the distribution of a set of binary response variables, we can require that

$$\eta_{t|1} \leqslant \cdots \leqslant \eta_{t|c}, \quad t = 1, \ldots, s, \tag{6}$$

so that the latent states are ordered from that corresponding to the smallest probability of success to that corresponding to the largest probability of success. We may also combine this requirement with a Rasch parametrization of type (5). In this case, we need a reduced set of inequality constraints, $\psi_1 \leqslant \cdots \leqslant \psi_c$, to ensure that (6) holds. □

2.2 Modelling transition probabilities

Let $\boldsymbol{\pi} = \text{vec}(\boldsymbol{\Pi})$, where $\text{vec}(\cdot)$ is the row vector operator, and $\boldsymbol{\rho}$ denote the sub-vector of $\boldsymbol{\pi}$ in which the diagonal elements of $\boldsymbol{\Pi}$ are omitted because redundant and note that $\boldsymbol{\pi} = \boldsymbol{a} + \boldsymbol{A}\boldsymbol{\rho}$, with $\boldsymbol{a}$ and $\boldsymbol{A}$ defined in Appendix A1. We assume that

$$\boldsymbol{\rho} = \boldsymbol{W}\boldsymbol{\beta}, \tag{7}$$

where $\boldsymbol{W}$ is a full rank matrix of size $c(c-1) \times b$ with at most one positive element in any row and all the other elements equal to 0. Note that we are formulating a linear model directly on the transition probabilities and not on a vector of parameters related to such probabilities through a link function that maps them onto $\mathbb{R}^{c(c-1)}$. However, as already mentioned in Section 1, our approach allows the formulation of the hypothesis that one or more elements of $\boldsymbol{\Pi}$ are equal to 0. On the other hand, in order to ensure that all the transition probabilities are non-negative, we have to impose suitable restrictions on the parameter vector $\boldsymbol{\beta}$. Due to these restrictions, we are not in a standard inferential problem in order to derive the asymptotic distribution of the LR statistic for testing hypotheses on $\boldsymbol{\beta}$; this will be discussed in detail in Section 4. In particular, it may be easily verified that the constraint that all the diagonal elements of $\boldsymbol{\Pi}$ are non-negative

is equivalent to the constraint $\boldsymbol{T}\boldsymbol{W}\boldsymbol{\beta} \leqslant \mathbf{1}_c$, with $\boldsymbol{T} = \boldsymbol{I}_c \otimes \mathbf{1}'_{c-1}$. Moreover, since we require $\boldsymbol{W}$ to have at most one positive element in any row, the constraint that all the off-diagonal elements of $\boldsymbol{\Pi}$ are non-negative may simply be expressed as $\boldsymbol{\beta} \geqslant \mathbf{0}_b$. As shown by the examples below, this requirement of the design matrix $\boldsymbol{W}$ does not limit the applicability of the present class of models. Note that the model in which the transition probabilities are unconstrained may be formulated by letting $\boldsymbol{W} = \boldsymbol{I}_{c(c-1)}$.

Example 4. A significant reduction of the number of parameters of a LM model may be achieved by constraining all the off-diagonal elements of the transition matrix $\boldsymbol{\Pi}$ to be equal to each other; with $c = 3$, for instance, we have

$$\boldsymbol{\Pi} = \begin{pmatrix} 1-2\beta & \beta & \beta \\ \beta & 1-2\beta & \beta \\ \beta & \beta & 1-2\beta \end{pmatrix}. \tag{8}$$

This constraint may be formulated by letting $\boldsymbol{W} = \mathbf{1}_{c(c-1)}$. A less stringent constraint is that $\boldsymbol{\Pi}$ is symmetric; this is equivalent to assuming that the probability of transition from latent state w to latent state x is the same as that of the reverse transition:

$$\boldsymbol{\Pi} = \begin{pmatrix} 1-(\beta_1+\beta_2) & \beta_1 & \beta_2 \\ \beta_1 & 1-(\beta_1+\beta_3) & \beta_3 \\ \beta_2 & \beta_3 & 1-(\beta_2+\beta_3) \end{pmatrix}.$$

This hypothesis may be formulated by letting $\boldsymbol{W} = \boldsymbol{U} + \boldsymbol{L}$, where $\boldsymbol{U}$ and $\boldsymbol{L}$ consist, respectively, of the subset of columns of $\boldsymbol{I}_{c(c-1)}$ corresponding to the elements of $\boldsymbol{\rho}$ which are upper triangular and lower triangular in $\boldsymbol{\Pi}$ (excluding those in the diagonal). □

Example 5. When the latent states are ordered in a meaningful way by assuming, for instance, that (6) holds, it may be interesting to formulate the hypothesis that a subject in latent state w may move only to latent state $x = w+1, \ldots, c$ or to $x = w+1$. With $c = 3$, for instance, we have respectively

$$\boldsymbol{\Pi} = \begin{pmatrix} 1-(\beta_1+\beta_2) & \beta_1 & \beta_2 \\ 0 & 1-\beta_3 & \beta_3 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Pi} = \begin{pmatrix} 1-\beta_1 & \beta_1 & 0 \\ 0 & 1-\beta_2 & \beta_2 \\ 0 & 0 & 1 \end{pmatrix}.$$

The first hypothesis may be formulated by letting $\boldsymbol{W} = \boldsymbol{U}$ and the second one by letting $\boldsymbol{W}$ equal to a suitable subset of columns of $\boldsymbol{U}$. Another formulation which is appropriate with ordered states is based on the assumption that, for any w and x, the probability $\pi_{x|w}$ is proportional to $2^{-|x-w|}$ and therefore decreases when the distance from w and x increases, i.e.

$$\boldsymbol{\Pi} = \frac{1}{4}\begin{pmatrix} 4-3\beta & 2\beta & \beta \\ 2\beta & 4-4\beta & 2\beta \\ \beta & 2\beta & 4-3\beta \end{pmatrix}.$$

In this case $\boldsymbol{W} = (\, 2^{-1}\boldsymbol{i}_1 \quad \cdots \quad 2^{-(c-1)}\boldsymbol{i}_{c-1}\,)$, where $\boldsymbol{i}_j$ is obtained by summing the columns of $\boldsymbol{I}_{c(c-1)}$ corresponding to the elements $\pi_{x|w}$ of $\boldsymbol{\rho}$ with $|x-w| = j$. □

2.3 Parameter space

By combining different choices of the parametrization of the conditional distribution of the response variables (given the latent process) and of the transition probabilities we give rise to a class of LM models which obviously includes the basic LM model illustrated at the beginning of this section. We recall that these parametrizations are formulated as $\boldsymbol{\eta} = \boldsymbol{Z}\boldsymbol{\gamma}$, $\boldsymbol{K}\boldsymbol{\gamma} \geqslant \boldsymbol{0}_k$, for a suitable chosen link function $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\phi})$, and $\boldsymbol{\rho} = \boldsymbol{W}\boldsymbol{\beta}$. Therefore, the vector of the non-redundant parameters of any of these models may be expressed as $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')'$, where $\boldsymbol{\alpha} = \{\log(\lambda_{x+1}/\lambda_1),\, x = 1, \ldots, c-1\}$ is a vector of logits for the initial probabilities of the latent states.

An important point is the shape of the *parameter space*. We have that

$$\boldsymbol{\theta} \in \boldsymbol{\Theta} = \mathcal{A} \times \mathcal{B} \times \mathcal{C},$$

where $\mathcal{A} = \mathbb{R}^{c-1}$, $\mathcal{B} = \{\boldsymbol{\beta} : \boldsymbol{\beta} \geqslant \boldsymbol{0}_b,\, \boldsymbol{T}\boldsymbol{W}\boldsymbol{\beta} \leqslant \boldsymbol{1}_c\}$ and $\mathcal{C} = \{\boldsymbol{\gamma} : \boldsymbol{K}\boldsymbol{\gamma} \geqslant \boldsymbol{0}_k\}$. This notation will be useful in Section 4 to illustrate the derivation of the asymptotic distribution of the LR statistic for testing linear hypotheses on $\boldsymbol{\beta}$.

3 Maximum likelihood estimation

Let $n_{\boldsymbol{y}}$ be the frequency of the response configuration $\boldsymbol{y}$ in a sample of n subjects, let $\mathcal{Y}$ be the set of the distinct response configurations observed at least once and let $\boldsymbol{n} = \{n_{\boldsymbol{y}}, \boldsymbol{y} \in \mathcal{Y}\}$ be the vector of the corresponding frequencies. By assuming that these subjects are independent of each other, the log-likelihood of any model in the class of LM models outlined in the previous section may be expressed as

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

where $p(\boldsymbol{y})$ is computed as a function of $\boldsymbol{\theta}$ by using recursion (3). Note that, since the cardinality of $\mathcal{Y}$ is always less than or equal to n, $\ell(\boldsymbol{\theta})$ may even be computed for large values of s, provided that n is reasonable. Note also that, if a dummy renormalization is used in recursion (3), the log-likelihood of the model may be computed on the basis of the renormalized probabilities $p^*(\boldsymbol{y})$ as $\sum_{\boldsymbol{y} \in \mathcal{Y}} n_{\boldsymbol{y}} \log[p^*(\boldsymbol{y})] - ns\log(a)$, where the last term does not depend on $\boldsymbol{\theta}$ and therefore may be omitted.

In the following, we show how, in order to estimate $\boldsymbol{\theta}$, we can maximize $\ell(\boldsymbol{\theta})$ by means of an EM algorithm (Dempster *et al.*, 1977). We also deal with the Fisher information matrix and the

estimation of class membership probabilities which may be used to classify subjects on the basis of the response configurations they provide.

3.1 The EM algorithm

Suppose that we knew the frequencies $m_{\boldsymbol{x},\boldsymbol{y}}$'s of the contingency table in which the subjects are cross-classified according to the latent configuration $\boldsymbol{x}$ and the response configuration $\boldsymbol{y}$. We could then compute the *complete data log-likelihood*

$$\ell^{\dagger}(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{\boldsymbol{y}\in\mathcal{Y}} m_{\boldsymbol{x},\boldsymbol{y}} \log[p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})], \tag{9}$$

where $\mathcal{X}$ is the support of $\boldsymbol{X}$ and $p(\boldsymbol{y}|\boldsymbol{x})$ and $p(\boldsymbol{x})$ are defined, respectively, in (2) and (1). The following decomposition holds

$$\ell^{\dagger}(\boldsymbol{\theta}) = \ell_1^{\dagger}(\boldsymbol{\alpha}) + \ell_2^{\dagger}(\boldsymbol{\beta}) + \ell_3^{\dagger}(\boldsymbol{\gamma}), \tag{10}$$

where

$$\begin{aligned}
\ell_1^{\dagger}(\boldsymbol{\alpha}) &= \sum_x f_x \log(\lambda_x) \\
\ell_2^{\dagger}(\boldsymbol{\beta}) &= \sum_w \sum_x g_{w,x} \log(\pi_{x|w}) \\
\ell_3^{\dagger}(\boldsymbol{\gamma}) &= \sum_t \sum_x \sum_y h_{t,x,y} \log(\phi_{t,y|x}),
\end{aligned}$$

with f_x denoting the number of subjects which at the first occasion are in the latent state x, $g_{w,x}$ denoting the number of transitions from latent state w to latent state x and $h_{t,x,y}$ denoting the number of subjects which at occasion t are in latent state x and provide response y.

Remark 1 *Since we assume that all the probabilities λ_x and $\phi_{t,y|x}$ are strictly positive, we may always compute $\ell_1^{\dagger}(\boldsymbol{\alpha})$ and $\ell_3^{\dagger}(\boldsymbol{\gamma})$. In contrast, since we may have one or more transition probabilities $\pi_{x|w}$ equal to 0, computing $\ell_2^{\dagger}(\boldsymbol{\beta})$ requires special care. Therefore, we use the convention that, if $\pi_{x|w} = 0$ by assumption, $g_{w,x}\log(\pi_{x|w}) \equiv 0$ for all possible values of $g_{w,x}$, whereas, if $\pi_{x|w} = 0$, but not by assumption, $g_{w,x}\log(\pi_{x|w})$ is set equal to 0 when $g_{w,x} = 0$ and to $-\infty$ when $g_{w,x} > 0$. Thus, a necessary condition for a vector $\tilde{\boldsymbol{\beta}}$ to maximize $\ell_2^{\dagger}(\boldsymbol{\beta})$ is that the corresponding probabilities $\tilde{\pi}_{x|w}$ not constrained to be 0 by assumption are greater than 0 for any w and x such that $g_{w,x} > 0$.*

Because of decomposition (10), the vector $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}', \tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\gamma}})'$ which maximizes $\ell^{\dagger}(\boldsymbol{\theta})$ may be found by maximizing its components separately as follows:

- **Maximization of** $\ell_1^{\dagger}(\boldsymbol{\alpha})$: Simply set $\tilde{\boldsymbol{\alpha}}$ equal to the vector $\{\log(f_{x+1}/f_1),\, x = 1, \ldots, c-1\}$;

- **Maximization of $\ell_2^\dagger(\boldsymbol{\beta})$**: Let $\bar{\boldsymbol{\pi}}$ be the sub-vector of $\boldsymbol{\pi}$ whose elements are not constrained to 0 under (7) and let $\bar{\boldsymbol{g}}$ be the corresponding vector with elements $g_{w,x}$ and note that $\bar{\boldsymbol{\pi}} = \boldsymbol{B}\boldsymbol{\pi}$, where $\boldsymbol{B}$ is obtained by selecting a suitable subset of rows from the matrix $\boldsymbol{I}_{c^2}$. Consequently, because of Remark 1, we can express the log-likelihood at issue as

$$\ell_2^\dagger(\boldsymbol{\beta}) = \bar{\boldsymbol{g}}' \log(\bar{\boldsymbol{\pi}}) = \bar{\boldsymbol{g}}' \log[\boldsymbol{B}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{W}\boldsymbol{\beta})] + constant.$$

Provided all the elements of $\bar{\boldsymbol{g}}$ are strictly positive, we also have that $\ell_2^\dagger(\boldsymbol{\beta})$ tends to $-\infty$ when one or more elements of $\bar{\boldsymbol{\pi}}$ approach 0; therefore, the vector $\tilde{\boldsymbol{\beta}}$ that maximizes this function must be an interior point of the space $\mathcal{B}$ defined in Section 2.3. Moreover, for any interior point $\boldsymbol{\beta}$ of $\mathcal{B}$, the second derivative of $\ell_2^\dagger(\boldsymbol{\beta})$, equal to $-\boldsymbol{W}'\boldsymbol{A}'\boldsymbol{B}'\mathrm{diag}(\bar{\boldsymbol{\pi}})^{-2}\mathrm{diag}(\bar{\boldsymbol{g}})\boldsymbol{B}\boldsymbol{A}\boldsymbol{W}$, is negative definite and therefore this function has a unique maximum. This maximum may be found by means of a Fisher-scoring algorithm that at any step updates the estimate of $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}^{(0)} + [\boldsymbol{F}_2^\dagger(\boldsymbol{\beta}^{(0)})]^{-1}\boldsymbol{s}_2^\dagger(\boldsymbol{\beta}^{(0)}),$$

where $\boldsymbol{\beta}^{(0)}$ denotes the estimate of $\boldsymbol{\beta}$ at the end of the previous step and

$$\begin{aligned} \boldsymbol{s}_2^\dagger(\boldsymbol{\beta}) &= \boldsymbol{W}'\boldsymbol{A}'\boldsymbol{B}'\mathrm{diag}(\bar{\boldsymbol{\pi}})^{-1}\bar{\boldsymbol{g}} \\ \boldsymbol{F}_2^\dagger(\boldsymbol{\beta}) &= \boldsymbol{W}'\boldsymbol{A}'\boldsymbol{B}'\mathrm{diag}[\boldsymbol{B}(\dot{\boldsymbol{g}} \otimes \boldsymbol{1}_{c-1})]\mathrm{diag}(\bar{\boldsymbol{\pi}})^{-1}\boldsymbol{B}\boldsymbol{A}\boldsymbol{W}, \end{aligned}$$

with $\dot{\boldsymbol{g}} = \{g_{w,\cdot},\ w = 1, \ldots, c\}$, $g_{w,\cdot} = \sum_x g_{w,x}$, are, respectively, the score vector and the Fisher information matrix for $\ell_2^\dagger(\boldsymbol{\beta})$. When one or more elements of $\bar{\boldsymbol{g}}$ are equal to 0, the maximum of $\ell_2^\dagger(\boldsymbol{\beta})$ might be attained at a $\tilde{\boldsymbol{\beta}}$ having one or more elements equal to 0 and therefore the algorithm above must be appropriately modified. A simpler solution consists in adding a small positive number, e.g. 10^{-8}, to all the elements of $\bar{\boldsymbol{g}}$, so that the same algorithm may be used without further adjustments. We observed that the effect on the solution is usually negligible and no extra time is required for the implementation.

- **Maximization of $\ell_3^\dagger(\boldsymbol{\gamma})$**: To take into account the presence of inequality constraints of the type $\boldsymbol{K}\boldsymbol{\gamma} \geqslant \boldsymbol{0}_k$ on the parameter vector $\boldsymbol{\gamma}$, this maximization is performed by means of an algorithm based on the solution of a series of constrained least squares problems. At any step of this algorithm, the estimate of $\boldsymbol{\gamma}$ is updated by solving

$$\min_{\boldsymbol{K}\boldsymbol{\gamma} \geqslant \boldsymbol{0}_k} (\boldsymbol{\gamma} - \boldsymbol{w}^{(0)})'\boldsymbol{F}_3^\dagger(\boldsymbol{\gamma}^{(0)})(\boldsymbol{\gamma} - \boldsymbol{w}^{(0)}),$$

where $\boldsymbol{w}^{(0)} = \boldsymbol{\gamma}^{(0)} + [\boldsymbol{F}_3^\dagger(\boldsymbol{\gamma}^{(0)})]^{-1}\boldsymbol{s}_3^\dagger(\boldsymbol{\gamma}^{(0)})$ is the *working dependent variable*, $\boldsymbol{\gamma}^{(0)}$ is the estimate of $\boldsymbol{\gamma}$ at the end of the previous step and $\boldsymbol{s}_3^\dagger(\boldsymbol{\gamma})$ and $\boldsymbol{F}_3^\dagger(\boldsymbol{\gamma})$ are, respectively, the

score vector and the Fisher information matrix for $\ell_3^\dagger(\boldsymbol{\beta})$. The latter two are given by

$$\begin{aligned} \boldsymbol{s}_3^\dagger(\boldsymbol{\gamma}) &= \boldsymbol{Z}'\boldsymbol{D}(\boldsymbol{\eta})'\mathrm{diag}(\boldsymbol{\phi})^{-1}\boldsymbol{h} \\ \boldsymbol{F}_3^\dagger(\boldsymbol{\gamma}) &= \boldsymbol{Z}'\boldsymbol{D}(\boldsymbol{\eta})'\mathrm{diag}(\dot{\boldsymbol{h}} \otimes \mathbf{1}_d)\mathrm{diag}(\boldsymbol{\phi})\boldsymbol{D}(\boldsymbol{\eta})\boldsymbol{Z}, \end{aligned}$$

where the vector $\boldsymbol{h}$ has elements $h_{t,x,y}$ arranged as in $\boldsymbol{\phi}$, $\boldsymbol{D}(\boldsymbol{\eta})$ is the derivative matrix of $\boldsymbol{\phi}$ with respect to $\boldsymbol{\eta}'$ and $\dot{\boldsymbol{h}}$ is a column vector with elements $h_{t,x,\cdot} = \sum_y h_{t,x,y}$, $t = 1, \ldots, s$, $x = 1, \ldots, c$, arranged by letting the index x run faster than t. For a detailed description of this constrained maximization algorithm in similar contexts, see Dardanoni and Forcina (1998) and Bartolucci and Forcina (2000). Note that, when there are no inequality constraints on $\boldsymbol{\gamma}$, the algorithm reduces to the usual Fisher-scoring algorithm. Moreover, when we have no restriction on the conditional distribution of any Y_t given X_t, i.e. $\boldsymbol{Z} = \boldsymbol{I}_{sc(d-1)}$, we have an explicit solution for $\tilde{\boldsymbol{\gamma}}$ or, equivalently, for $\tilde{\boldsymbol{\eta}}$.

Since the frequencies f_x's, $g_{w,x}$'s and $h_{t,x,y}$'s in (9) are unknown, the EM algorithm maximizes $\ell^\dagger(\boldsymbol{\theta})$ as above (M step), once these frequencies have been substituted with the corresponding conditional expected values given the observed data and the current value of the parameters (E step). This process is iterated until convergence in $\ell(\boldsymbol{\theta})$. At the E step, in particular, the conditional expected values at issue are obtained as:

$$\begin{aligned} \mathrm{E}(F_x|\boldsymbol{n}) &= \sum_{\boldsymbol{y}\in\mathcal{Y}} n_{\boldsymbol{y}} r_1(x|\boldsymbol{y}) \\ \mathrm{E}(G_{w,x}|\boldsymbol{n}) &= \sum_{\boldsymbol{y}\in\mathcal{Y}} n_{\boldsymbol{y}} \sum_{t>1} r_t(w,x|\boldsymbol{y}) \\ \mathrm{E}(H_{t,x,y}|\boldsymbol{n}) &= \sum_{\boldsymbol{y}\in\mathcal{Y}} n_{\boldsymbol{y}} I(y_t = y) r_t(x|\boldsymbol{y}), \end{aligned}$$

where $I(\cdot)$ is the indicator function, $r_t(x|\boldsymbol{y}) = p(X_t = x|\boldsymbol{Y} = \boldsymbol{y})$ and $r_t(w,x|\boldsymbol{y}) = p(X_{t-1} = w, X_t = x|\boldsymbol{Y} = \boldsymbol{y})$. Let $\boldsymbol{R}_t(\boldsymbol{y})$, $t = 2, \ldots, s$, be the $c \times c$ matrix with elements $r_t(w,x|\boldsymbol{y})$ arranged by letting w run by row and x by column. This matrix may be easily computed through the following backward recursion which is well known in the hidden Markov chain literature (see Baum *et al.*, 1970, Levinson *et al.*, 1983, and MacDonald and Zucchini, 1997, Sec. 2.2):

$$\boldsymbol{R}_t(\boldsymbol{y}) = \mathrm{diag}[\boldsymbol{q}(\boldsymbol{y}_{\leqslant t-1})]\boldsymbol{\Pi}\mathrm{diag}(\boldsymbol{\phi}_{t,y_t})\mathrm{diag}[\boldsymbol{r}(\boldsymbol{y}_{>t})]/p(\boldsymbol{y}), \quad t = 2, \ldots, s, \tag{11}$$

where $\boldsymbol{y}_{>t} = \{y_j,\ j = t+1, \ldots, s\}$ and

$$\boldsymbol{r}(\boldsymbol{y}_{>t}) = \begin{cases} \mathbf{1}_c & \text{if } t = s \\ \boldsymbol{\Pi}\mathrm{diag}(\boldsymbol{\phi}_{t+1,y_{t+1}})\boldsymbol{r}(\boldsymbol{y}_{>t+1}) & \text{otherwise} \end{cases}. \tag{12}$$

When s is very large, this recursion may require a dummy renormalization. This may be performed again by multiplying the quantity computed at any step of (12) by a suitable constant a.

The resulting vector computed at the t-th step, $\boldsymbol{r}^*(\boldsymbol{y}_{>t})$, may be used instead of $\boldsymbol{r}(\boldsymbol{y}_{>t})$ in (11), provided that $\boldsymbol{q}(\boldsymbol{y}_{\leqslant t-1})$ and $p(\boldsymbol{y})$ are substituted with the corresponding renormalized quantities $\boldsymbol{q}^*(\boldsymbol{y}_{\leqslant t-1})$ and $p^*(\boldsymbol{y})$.

Finally, note that any M step consists in solving a series of simple maximization problems which have a unique solution. Therefore, along the same lines as Shi *et al.* (2005), it is possible to prove that the observed log-likelihood $\ell(\boldsymbol{\theta})$ has increased after any EM step and that the above algorithm converges to a local maximum of this function. However, this local maximum cannot be guaranteed to correspond to the global maximum since, as for any other latent variable model, the likelihood may be multimodal. As usual, this problem may be addressed by trying different initializations of the algorithm and then choosing $\hat{\boldsymbol{\theta}}$, the parameter value which at convergence gives the highest value of $\ell(\boldsymbol{\theta})$, as the ML estimate of $\boldsymbol{\theta}$. On the other hand, a small simulation study, the results of which are not reported here, showed us that the chance of there being more than one local maximum is usually low when the number of observations is large in comparison to the number of parameters and the assumed model holds.

3.1.1 Initialization of the algorithm

We observed that an effective strategy to initialize the above EM algorithm is by performing the first E step with probabilities $\phi_{t,y|x}$'s, λ_x's and $\pi_{x|w}$'s chosen as follows:

- $\phi_{t,y|x} = e_{x,y}\, n_{t,y} / \sum_j e_{x,j}\, n_{t,j}$, where $n_{t,y}$ is the observed frequency of $Y_t = y$ and the coefficient $e_{x,y}$ increases with y when $x > (c+1)/2$, decreases with y when $x < (c+1)/2$ and is constant in y when $x = (c+1)/2$.
- $\lambda_x = 1/c$ for any x;
- $\pi_{x|w} = 1 - \tau$ when $w = x$ and $\pi_{x|w} = \tau/(c-1)$ otherwise, where τ is a suitable constant between 0 and 1.

If it is necessary to try different initializations of the EM algorithm, these probabilities may be generated at random by drawing any block of them summing to 1 from a Dirichlet distribution with parameters defined according to the rules given above. For instance, the starting value of $\boldsymbol{\lambda}$ could be drawn from a Dirichlet distribution with parameter vector proportional to $\mathbf{1}_c$, the first row of $\boldsymbol{\Pi}$ could be drawn from a Dirichlet distribution with parameter vector proportional to $((1-\tau), \tau/(c-1), \ldots, \tau/(c-1))'$ and so on.

3.2 Fisher information matrix, standard errors and local identifiability

The Fisher information matrix of the observed data may be expressed as

$$\boldsymbol{F}(\boldsymbol{\theta}) = n\boldsymbol{Q}(\boldsymbol{\theta})'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{Q}(\boldsymbol{\theta}), \tag{13}$$

where $\boldsymbol{p}$ is a vector with elements $p(\boldsymbol{y})$ for any $\boldsymbol{y}$ arranged in lexicographical order and $\boldsymbol{Q}(\boldsymbol{\theta})$ is the derivative matrix of this vector with respect to $\boldsymbol{\theta}'$ (see Appendix A2). In particular, we are usually interested in $\boldsymbol{F}(\hat{\boldsymbol{\theta}})$, i.e. the Fisher information computed at the ML estimate of $\boldsymbol{\theta}$, which may be used for computing standard errors and checking identifiability; this matrix is also used within the testing procedure illustrated in Section 4. Note that when s is large, $\boldsymbol{F}(\hat{\boldsymbol{\theta}})$ cannot be computed as in (13). In this case, we can validly use the empirical variance-covariance matrix of the score, instead of $\boldsymbol{F}(\hat{\boldsymbol{\theta}})$; see also McLachlan and Peel (2000, Sec. 2.15).

The standard error for a parameter estimate may be computed as the square root of the corresponding diagonal element of $\boldsymbol{F}(\hat{\boldsymbol{\theta}})^{-1}$. However, if the true value of $\boldsymbol{\theta}$ is close to the boundary of the parameter space $\boldsymbol{\Theta}$, in this way we can considerably overestimate the true standard error. On the other hand, knowledge of standard errors for the parameter estimates is relatively less important in this situation since the distribution of the estimators may depart significantly from normality, and therefore even though these standard errors are known, they cannot be validly used to construct confidence intervals and testing hypotheses in the usual way; for a discussion on this issue see Silvapulle and Sen (2004, Sec. 4.9).

The information matrix can also be used for checking local identifiability of the model at $\hat{\boldsymbol{\theta}}$; this is a weaker condition than global identifiability on which literature on latent variable models has focused since long time (see, for example, McHugh, 1956, and Goodman, 1974). In particular, a commonly accepted procedure for checking that the model is locally identifiable at $\hat{\boldsymbol{\theta}}$ consists in checking that $\boldsymbol{F}(\hat{\boldsymbol{\theta}})$ is of full rank. This is essentially equivalent to checking that the Jacobian $\boldsymbol{Q}(\hat{\boldsymbol{\theta}})$ is of full rank; see also Rothenberg (1971).

3.3 Estimation of class membership probabilities and classification

Once a certain LM model has been fitted, the element $\hat{\lambda}_{t,x}$, $x = 1, \ldots, c$, of the vector $\hat{\boldsymbol{\lambda}}_t = (\hat{\boldsymbol{\Pi}}^{t-1})'\hat{\boldsymbol{\lambda}}$ is an estimate of the proportion of subjects which are in latent class x at the t-th occasion. Note that, when $t = 1$, this vector equals $\hat{\boldsymbol{\lambda}}$; the same occurs, regardless of t, when the transition matrix is diagonal. Moreover, when we are dealing with a LM model based on parametrization (5), we can estimate the average ability of the population at occasion t as $\hat{\overline{\psi}}_t = \sum_x \hat{\psi}_x \hat{\lambda}_{t,x}$.

Note that from the EM algorithm we also obtain, for any x and t and any response configuration $\boldsymbol{y}$ observed at least once, the estimate $\hat{r}_t(x|\boldsymbol{y})$ of the conditional probability of being in latent class x at occasion t given $\boldsymbol{y}$. By using these *posterior estimates* we can assign a subject to a given latent class on the basis of response configuration he/she provided or, when the model is based on a Rasch parametrization, estimate his/her ability as $\hat{\psi}_t(\boldsymbol{y}) = \sum_x \hat{\psi}_x \hat{r}_t(x|\boldsymbol{y})$.

4 Testing linear hypotheses on the transition probabilities

In this section, we deal with the asymptotic distribution of the LR statistic for testing a hypothesis of the type

$$H_0 : \boldsymbol{M}\boldsymbol{\beta} = \mathbf{0}_m,$$

with $\boldsymbol{M}$ denoting a full rank matrix of dimension $m \times b$, on the latent process parameters. Note that we are not formulating a hypothesis directly on the transition probabilities, but on a reduced vector of parameters on which they depend through a linear model of type (7). This gives more flexibility to the approach. We can test, for instance, the hypothesis that the transition matrix $\boldsymbol{\Pi}$ is diagonal, and therefore a LC formulation holds, against the alternative that $\boldsymbol{\Pi}$ is symmetric or with all the off-diagonal elements equal to each other (see Example 4). We also recall that the parameter space of $\boldsymbol{\beta}$, $\mathcal{B}$, is specified by means of inequality constraints (see Section 2.3) which, under H_0, may imply additional equality constraints on $\boldsymbol{\beta}$ with respect to those formulated as $\boldsymbol{M}\boldsymbol{\beta} = \mathbf{0}_m$. This happens, for instance, when $\boldsymbol{M} = \mathbf{1}_b'$. In this case, only $\boldsymbol{\beta} = \mathbf{0}_b$ jointly satisfies $\boldsymbol{\beta} \geqslant \mathbf{0}_b$ and $\boldsymbol{M}\boldsymbol{\beta} = \mathbf{0}_m$, even if $\boldsymbol{M}$ has only one row. In order to avoid these situations, and without loss of generality, we require the dimension of the space $\mathcal{B}_0 = \mathcal{B} \cap \{\boldsymbol{\beta} : \boldsymbol{M}\boldsymbol{\beta} = \mathbf{0}_m\}$ to be equal to $b - m$. It is also convenient to reformulate more precisely the testing problem we are dealing with as $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ against $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta} \backslash \boldsymbol{\Theta}_0$, where $\boldsymbol{\Theta}$ is the parameter space of the assumed model and $\boldsymbol{\Theta}_0 = \mathcal{A} \times \mathcal{B}_0 \times \mathcal{C}$.

The LR statistic for testing H_0 against H_1 may be expressed as

$$D = -2[\ell(\hat{\boldsymbol{\theta}}_0) - \ell(\hat{\boldsymbol{\theta}})], \tag{14}$$

where $\hat{\boldsymbol{\theta}}_0$ is the ML estimate of $\boldsymbol{\theta}$ under the constraint $\boldsymbol{M}\boldsymbol{\beta} = \mathbf{0}_m$ and $\hat{\boldsymbol{\theta}}$ is the unconstrained estimate. Note that $\hat{\boldsymbol{\theta}}_0$ may be computed through the same EM algorithm illustrated in the previous section for computing $\hat{\boldsymbol{\theta}}$; the only difference with respect to the original formulation is that we have to use $\boldsymbol{W}\boldsymbol{M}^{\perp}$ instead of $\boldsymbol{W}$ as a design matrix for the linear model on the transition probabilities, where $\boldsymbol{M}^{\perp}$ is a $b \times (b-m)$ full rank matrix spanning the null space of $\boldsymbol{M}$. Now let $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0', \boldsymbol{\beta}_0', \boldsymbol{\gamma}_0')'$ be the true value of $\boldsymbol{\theta}$ under H_0 and assume that $\boldsymbol{T}\boldsymbol{W}\boldsymbol{\beta}_0 < \mathbf{1}_c$ and $\boldsymbol{K}\boldsymbol{\gamma}_0 > \mathbf{0}_k$. If we also have that $\boldsymbol{\beta}_0 > \mathbf{0}_b$, $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$ and therefore it follows from standard

inferential results that D has asymptotic χ^2_m distribution. This occurs, for instance, when we are testing a LM model based on a transition matrix of type (8) against the same LM model in which this matrix is unconstrained; in this case, D has asymptotic $\chi^2_{c(c-1)-1}$ distribution under the null hypothesis. In contrast, if some elements of $\boldsymbol{\beta}_0$ are equal to 0, $\boldsymbol{\theta}_0$ is on the boundary of $\boldsymbol{\Theta}$ and therefore we are in a non-standard case. This occurs, for instance, when we are testing a LC model against a LM model.

The asymptotic distribution of a LR test statistic when the parameters are on the boundary of the parameter space has been studied by many authors, such as Chernoff (1954), Shapiro (1985) and Self and Liang (1987), for instance; for a review see also Silvapulle and Sen (2004). In our case let $\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{F}(\boldsymbol{\theta})/n$ be the average information matrix of the assumed LM model and $\boldsymbol{G}$ be the block of g rows of the matrix $\boldsymbol{I}_b$ such that the vector $\boldsymbol{G\beta}$ contains the elements of $\boldsymbol{\beta}$ which are constrained to 0 under H_0; note that $\boldsymbol{GM}^{\perp} = \boldsymbol{0}_g$. Also let $\boldsymbol{J}$ be the block of the remaining $b-g$ rows of $\boldsymbol{I}_b$ and assume that the elements of $\boldsymbol{J\beta}_0$ are strictly positive under H_0. Thus, among all the inequality constraints involved in the specification of the parameter space, only those formulated as $\boldsymbol{G\beta} \geqslant \boldsymbol{0}_g$ are *active*, in the sense that only these constraints have some chance of being violated when n grows indefinitely. Consequently, following Silvapulle and Sen (2004, Proposition 4.8.2; see also Self and Liang, 1987, Th. 3), we find that D is asymptotically distributed as

$$Q = \min_{\boldsymbol{\theta}\in\mathcal{L}} (\boldsymbol{\theta}-\boldsymbol{V})'\boldsymbol{H}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{V}) - \min_{\boldsymbol{\theta}\in\mathcal{P}} (\boldsymbol{\theta}-\boldsymbol{V})'\boldsymbol{H}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{V}),$$

where $\mathcal{L} = \{\boldsymbol{\theta} : \boldsymbol{M\beta} = \boldsymbol{0}_m\}$ is a linear space nested into the polyhedral convex cone $\mathcal{P} = \{\boldsymbol{\theta} : \boldsymbol{G\beta} \geqslant \boldsymbol{0}_g\}$ and $\boldsymbol{V} \sim N(\boldsymbol{0}, \boldsymbol{H}(\boldsymbol{\theta}_0)^{-1})$. It turns out that the distribution of Q, and therefore the asymptotic distribution of D, is of chi-bar-square type, denoted by $\overline{\chi}^2$, which is well-known in constrained statistical inference (for a review see Shapiro, 1988, and Silvapulle and Sen, 2004, Ch. 3). More precisely, we have:

Theorem 1 *Provided that $\boldsymbol{H}(\boldsymbol{\theta}_0)$ is of full rank and that H_0 holds with $\boldsymbol{J\beta}_0 > \boldsymbol{0}_{b-g}$, $\boldsymbol{TW\beta}_0 < \boldsymbol{1}_c$ and $\boldsymbol{K\gamma}_0 > \boldsymbol{0}_k$, the asymptotic distribution of the LR statistic (14) is the same as that of $Z_1 + Z_2$, where Z_1 and Z_2 are two independent random variables with*

$$Z_1 \sim \chi^2_{m-g} \quad \textit{and} \quad Z_2 \sim \overline{\chi}^2(\boldsymbol{\Sigma}_0, \mathcal{O}^g),$$

where $\boldsymbol{\Sigma}_0 = \boldsymbol{GH}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{G}'$ and $\mathcal{O}^g$ denotes the non-negative orthant of dimension g.

Note that the distribution function of $Z_1 + Z_2$ may be expressed as

$$p(Z_1 + Z_2 \leqslant z) = \sum_{j=0}^{g} w_j(g, \boldsymbol{\Sigma}_0) p(\chi^2_{m-g+j} \leqslant z), \tag{15}$$

where $w_j(g, \boldsymbol{\Sigma}_0)$, $j = 0, \ldots, g$, are suitable weights which may be computed through a simple Monte Carlo simulation procedure, such as the one outlined in Silvapulle and Sen (2004, Sec. 3.5). By exploiting this expression, we can easily compute an asymptotic p-value for D, once we have substituted $\boldsymbol{\theta}_0$ in $\boldsymbol{\Sigma}_0$ with its consistent estimate $\hat{\boldsymbol{\theta}}_0$. Note also that, even though we need to compute the weights in (15) by simulation, the resulting procedure for computing a p-value for D is much faster than a parametric bootstrap which requires estimating the same models several times and is therefore impractical in many real applications.

As a particular case of Theorem 1, we have that in which there are no elements of $\boldsymbol{\beta}$ constrained to 0 under H_0 ($g = 0$) and so the LR statistic D has χ^2_m asymptotic distribution. Two other interesting cases are dealt with in the following Corollary.

Corollary 1 *Under the same regularity conditions as for Theorem 1, the LR statistic D for testing a LC model based on parametrization (4) against the corresponding LM version with unconstrained transition matrix $\boldsymbol{\Pi}$ has asymptotic distribution*

$$\overline{\chi}^2(\boldsymbol{\Sigma}_0, \mathcal{O}^{c(c-1)}), \tag{16}$$

with $\boldsymbol{\Sigma}_0$ equal to the block of $\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}$ corresponding to the parameters in $\boldsymbol{\beta}$. When instead the transition matrix of the LM model is parameterized through a single parameter ($b = 1$), D has asymptotic distribution

$$0.5\chi^2_0 + 0.5\chi^2_1. \tag{17}$$

The latter case, in which the weights of the $\overline{\chi}^2$ distribution are explicitly given, occurs, for instance, when the larger model involved in the comparison is a LM model based on a transition matrix of type (8).

4.1 A simulation study

In order to assess the accuracy of the finite sample inference based on the asymptotic results illustrated above, we carried out a simulation study of the LR statistic between:

- (D_1) an unconstrained LC model with $c = 2$ classes and its LM version;
- (D_2) a LC Rasch model with $c = 3$ classes and its LM version with transition matrix constrained as in (8);
- (D_3) a LM model with $c = 3$ states and transition matrix of the type

$$\boldsymbol{\Pi} = \begin{pmatrix} 1-\beta & \beta & 0 \\ 0 & 1-\beta & \beta \\ 0 & 0 & 1 \end{pmatrix}$$

and the same model with transition matrix of the type

$$\mathbf{\Pi} = \begin{pmatrix} 1-\beta_1 & \beta_1 & 0 \\ \beta_2 & 1-(\beta_2+\beta_3) & \beta_3 \\ 0 & \beta_4 & 1-\beta_4 \end{pmatrix},$$

under the assumption that, conditionally on the latent process, the response variables have the same distribution.

In particular, for various values of s and n and in the case of binary response variables, we generated 2000 random samples from a set of suitably chosen LM models. For each of these samples, we computed the value of each of the LR statistics above, together with the corresponding p-value based on the asymptotic null distribution which is of type (16) for D_1, (17) for D_2 and $\chi^2_1 + \overline{\chi}^2(\mathbf{\Sigma}_0, \mathcal{O}^2)$, with $\mathbf{\Sigma}_0$ suitably defined, for D_3.

In a first series of these simulations, random samples were generated from the smaller model considered within any LR statistic (e.g. the LC model with $c = 2$ classes for D_1), with parameters chosen as follows:

- for D_1: $\phi_{t,1|1} = 1/3, \phi_{t,2|1} = 2/3, \forall t$, and $\pi_x = 1/2, \forall x$;
- for D_2: $\phi_{t,1|1} = 1/4, \phi_{t,2|1} = 1/2, \phi_{t,3|1} = 3/4, \forall t$, and $\pi_x = 1/3, \forall x$;
- for D_3: $\phi_{t,1|1} = 1/4, \phi_{t,2|1} = 1/2, \phi_{t,3|1} = 3/4, \forall t$, $\pi_x = 1/3, \forall x$, and $\beta = 0.1$.

Thus, it was possible to compare the actual significance level of the test with the nominal level. For nominal levels below 0.25, this comparison is shown in Figure 1 for $s = 6, 12, 24$ and $n = 100, 200, 400$ and in Figure 2 for $s = 100$ and $n = 250, 500, 1000$. In the latter case we use $\beta = 0.02$ for the model under which the null distribution of D_3 has been simulated.

On the basis of the results from these simulations, we can conclude that, provided s is moderate (Figure 1), the asymptotic null distribution of D_1, D_2 and D_3 is a reasonable approximation of the finite sample distribution, even when the sample size is small in comparison to the number of possible configurations of the response variables. With s large (Figure 2), the quality of the approximation of the asymptotic null distribution is still good for the statistic D_3, but gets worse for D_1 and D_2. This difference is due to the fact that the models fitted for computing D_1 and D_2 have a number of parameters that increase with s, whereas this does not occur for D_3. Note however, that even though the quality of the approximation of the asymptotic distribution is not as good for D_1 and D_2, the actual significance level is smaller than the nominal level and therefore the asymptotic distribution may safely be used in this case also.

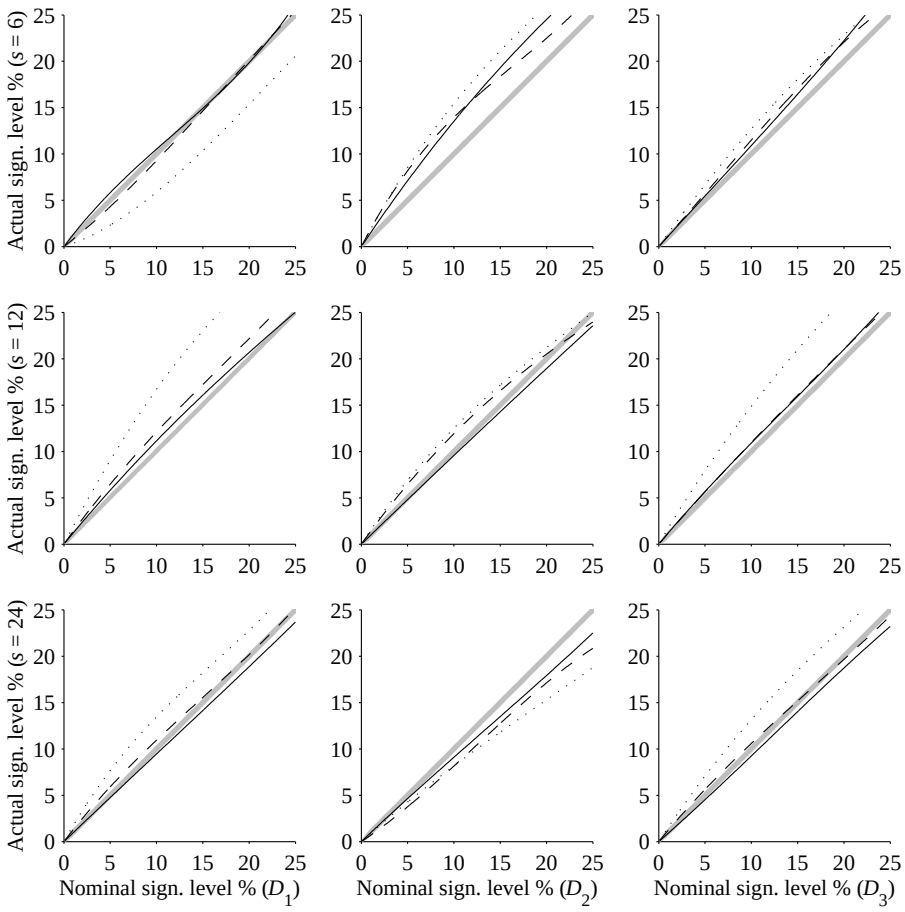

Figure 1: *Comparison between actual and nominal significance level for the tests based on the LR statistics* D_1*,* D_2 *and* D_3 *with* $s = 6, 12, 24$ *and* $n = 100$ *(dotted curve), 200 (dashed curve), 400 (solid curve).*

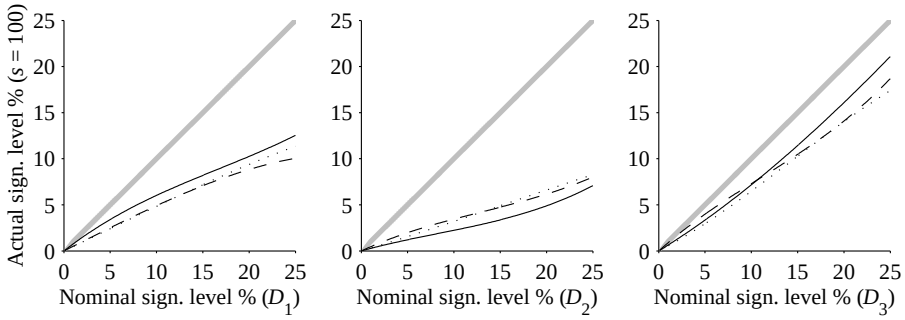

Figure 2: *Comparison between actual and nominal significance level for the tests based on the LR statistics* D_1*,* D_2 *and* D_3 *with* $s = 100$ *and* $n = 250$ *(dotted curve), 500 (dashed curve), 1000 (solid curve).*

In order to investigate the power of the LR tests in question, we also considered the following models:

- for D_1: a LM model with $c = 2$ states and transition matrix of type (8) with parameters $\phi_{t,1|1} = 1/3, \phi_{t,2|1} = 2/3, \forall t$, $\pi_x = 1/2, \forall x$, and β between 0 and 0.05 (for $s = 6, 12, 24$) and between 0 and 0.0005 (for $s = 100$);

- for D_2: a LM model with $c = 3$ states and transition matrix of type (8) with parameters $\phi_{t,1|1} = 1/4, \phi_{t,2|1} = 1/2, \phi_{t,3|1} = 3/4, \forall t$, $\pi_x = 1/3, \forall x$, and β between 0 and 0.05 (for $s = 6, 12, 24$) and between 0 and 0.0005 (for $s = 100$);

- for D_3: a LM model with $c = 3$ states and transition matrix

$$\mathbf{\Pi} = \begin{pmatrix} 1-\beta_1 & \beta_1 & 0 \\ \beta_2 & 1-\beta_1 & \beta_1-\beta_2 \\ 0 & \beta_2 & 1-\beta_2 \end{pmatrix}$$

  with parameters $\phi_{t,1|1} = 1/4, \phi_{t,2|1} = 1/2, \phi_{t,3|1} = 3/4, \forall t$, $\pi_x = 1/3, \forall x$, and $\beta_1 = 0.1$ and β_2 between 0 and 0.05 (for $s = 6, 12, 24$) and $\beta_1 = 0.02$ and β_2 between 0 and 0.0005 (for $s = 100$).

The estimated rejection rate of any test at the 5% nominal level is shown in Figure 3 for $s = 6, 12, 24$ and in Figure 4 for $s = 100$. As we may expect, the power of any test increases with the number of response variables (s) and the sample size (n). With $s = 6$ and $n = 100$, for instance, the test based on the LR statistic D_2 has a probability of about 1/4 of rejecting the null hypothesis when $\beta = 0.05$ (Figure 3). This probability becomes equal to about 3/4 when $s = 12$ and is equal to 1 with $s = 24$. Moreover, when $s = 100$ (Figure 4), this test has a very large probability of detecting also very small deviations from the null hypothesis, such as $\beta = 0.0001$.

5 Empirical illustration

To illustrate the approach proposed in this paper, we describe the analysis of two datasets, the first of which concerns the responses of a group of students to an educational test, and the second is concerned with the use of marijuana among young people.

5.1 Educational testing dataset

This dataset concerns a sample of $n = 1510$ examinees, who responded to a set of $s = 12$ items on Mathematics, extrapolated from a larger dataset collected in 1996 by the Educational Testing Service within a US project called the National Assessment of Educational Progress (NAEP); for

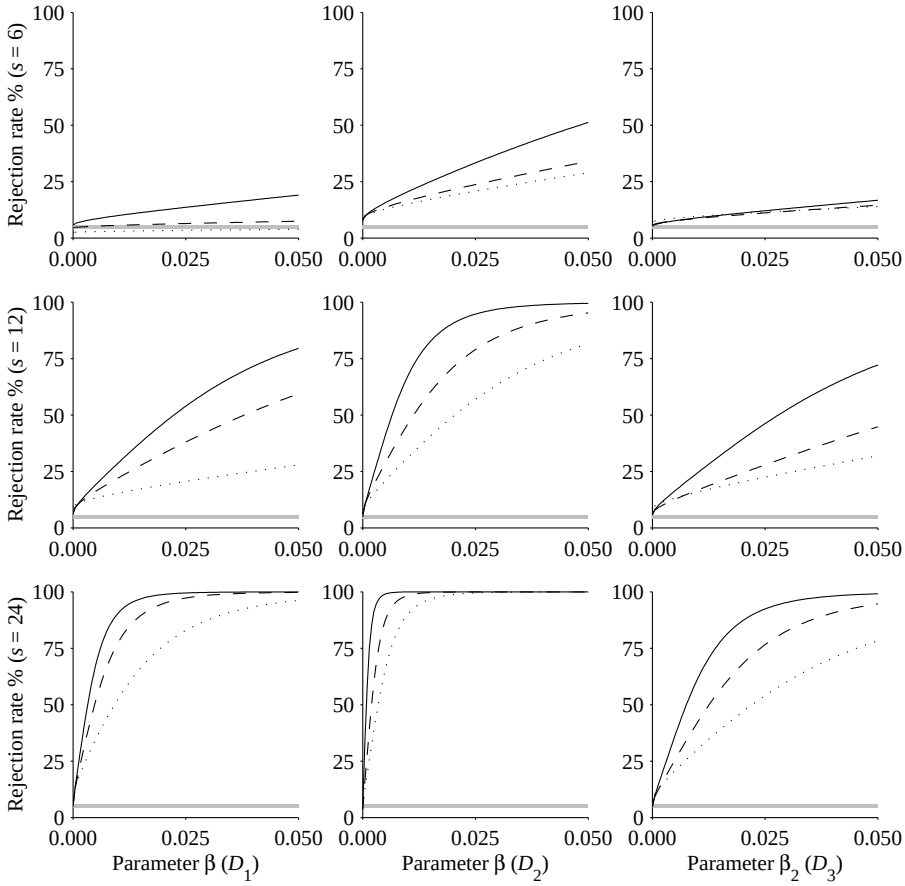

Figure 3: *Rejection rate of the tests based on the LR statistics* D_1, D_2 *and* D_3 *for* $s = 6, 12, 24$ *and* $n = 100$ *(dotted curve), 200 (dashed curve), 400 (solid curve).*

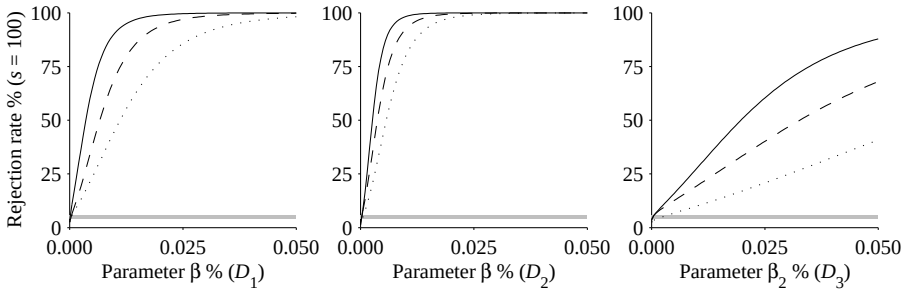

Figure 4: *Rejection rate of the tests based on the LR statistics* D_1, D_2 *and* D_3 *for* $s = 100$ *and* $n = 250$ *(dotted curve), 500 (dashed curve), 1000 (solid curve).*

a more detailed description, see Bartolucci and Forcina (2005). The interest here is in testing for the presence of phenomena, such as those described in Example 1 of Section 2.1, which may cause a violation of the basic assumptions of IRT models. Since the items were administered to all the examinees in the same order, it is appropriate to use a LM model for studying phenomena of this type. Note also that, since the contingency table on which the data may be displayed is very sparse, a direct comparison of a LC model based on a parametrization of type (4) with the saturated model is not as reliable as a comparison with a LM model based on the same parametrization. Thus, the latter may be seen as a valid alternative to the saturated model in assessing the goodness of fit of a LC model.

For this dataset we chose $c = 3$ latent states, which corresponds to the number of classes for which the basic LC model has the smallest BIC (Schwarz, 1978). With this number of states, the basic LM model has a deviance of 1899.16 with respect to the saturated model. This deviance may be considered adequate if compared with the number of degrees of freedom, equal to 4051. The estimated logits $\eta_{t|x}$'s under this model are shown in Table 1, while the estimated initial probabilities λ_x's and transition probabilities $\pi_{x|w}$'s are shown in Table 2.

	$\hat{\eta}_{t\|x}$ (LM)			$\hat{\eta}_{t\|x}$ (LC)		
t	$x=1$	$x=2$	$x=3$	$x=1$	$x=2$	$x=3$
1	-0.728	1.133	2.250	-0.711	1.114	2.241
2	-1.092	1.180	2.794	-1.087	1.168	2.783
3	-0.995	0.149	1.817	-1.012	0.153	1.801
4	0.078	2.297	3.509	0.085	2.292	3.508
5	-1.346	-0.508	0.594	-1.344	-0.500	0.587
6	-0.378	0.907	2.084	-0.361	0.911	2.081
7	-0.816	0.177	1.578	-0.807	0.190	1.577
8	-2.410	-0.571	1.772	-2.303	-0.536	1.774
9	-1.533	0.431	2.844	-1.415	0.453	2.846
10	-1.440	0.384	1.854	-1.306	0.410	1.854
11	-2.804	-1.784	0.027	-2.743	-1.656	0.033
12	-3.217	-1.411	-0.248	-3.010	-1.309	-0.251

Table 1: *Estimated logits $\eta_{t|x}$'s under the basic LM model and the corresponding LC model*

				$\hat{\pi}_{x\|w}$ (LM)			$\hat{\pi}_{x\|w}$ (LC)		
x	$\hat{\lambda}_x$ (LM)	$\hat{\lambda}_x$ (LC)	w	$x=1$	$x=2$	$x=3$	$x=1$	$x=2$	$x=3$
1	0.178	0.166	1	0.982	0.018	0.000	1.000	0.000	0.000
2	0.444	0.435	2	0.000	0.987	0.013	0.000	1.000	0.000
3	0.378	0.400	3	0.000	0.003	0.997	0.000	0.000	1.000

Table 2: *Estimated initial probabilities λ_x's and transition probabilities $\pi_{x|w}$'s under the basic LM model and the corresponding LC model*

Note that, according to the estimates in Table 1, the 1st latent class may be interpreted as that of the less capable subjects, whereas the 3rd class may be interpreted as that of the most capable subjects. Moreover, according to the results in Table 2, the second class contains the largest number of subjects and there is a high persistence, i.e. the probability of transition from one latent state to another is always very low. The largest off-diagonal element of the estimated transition probability, $\hat{\pi}_{2|1}$, is in fact equal to 0.018.

In order to simplify this model, we first tried to impose an additive structure of type (5) on the logits $\eta_{t|x}$'s. In this way, however, the deviance increases by 115.02 with a p-value, computed on the basis of the χ^2_{22} distribution, very close to 0 and therefore this restriction must be rejected. We then tried to impose some restrictions on the latent structure. In particular, we first tried to impose the constraint that all the off-diagonal elements of $\mathbf{\Pi}$ are equal to each other as in (8). The estimate of the common transition probability is $\hat{\beta} = 0.001$, while the LR statistic D with respect to the initial LM model is equal to 2.24 with a p-value, computed on the basis of the χ^2_5 distribution, equal to 0.814. Consequently, this restriction cannot be rejected. Finally, we tried to impose the restriction of absence transition between latent states. For the LC model, resulting from this restriction, we have a LR statistic of 0.61 with respect to the LM model with constant transition probabilities. According to Corollary 1, this statistic has null asymptotic distribution of type (17), so that the corresponding p-value is 0.216; consequently, the LC model cannot be rejected and therefore there is no evidence of violations of the LI assumption. A similar conclusion may be reached by directly comparing this model with the basic LM model in which the transition matrix is unconstrained. In this case, $D = 2.86$ with a p-value of 0.261 computed on the basis of distribution (16). The estimates of the parameters of the LC model are shown in Tables 1 and 2 together with those of the parameters of the basic LM model.

5.2 Marijuana consumption dataset

This dataset has been taken from five annual waves (1976-80) of the National Youth Survey (Elliot *et al.*, 1989) and is based on $n = 237$ respondents aged 13 years in 1976. The use of marijuana is measured through of $s = 5$ ordinal variables, one for each annual wave, with $d = 3$ levels equal to: 1 for never in the past year; 2 for no more than once a month in the past year; 3 for more than once a month in the past year. A variety of models which may be used for the analysis of this dataset has been illustrated by Vermunt and Hagenaars (2004). In particular, they mention: (i) marginal models, which represent an extension of log-linear models; (ii) transitional models, which include Markov chain models; (iii) random effect models, which include the LC

model. However, as mentioned by the same authors, a LM approach is desirable since it combines many characteristics of the approaches above and may be appreciated for its flexibility and easy interpretation. An analysis of this type is illustrated below.

Also for this dataset, we used LM models with $c = 3$ latent states. With this number of states, the basic LM model has a deviance of 85.80 with 204 degrees of freedom, with respect to the saturated model. The corresponding parameter estimates are shown in Tables 3 and 4. In order to simplify this model, we assumed the following parametrization for the conditional global logits of any response variable Y_t given the latent state $X_t = x$:

$$\eta_{t,y|x} = \psi_x + \delta_y, \qquad t = 1, \ldots, s, \quad x = 1, \ldots, c, \quad y = 1, \ldots, d-1, \tag{18}$$

where ψ_x may be interpreted as the tendency to use marijuana for the subject in the latent class x and the parameter δ_y is common to all the response variables Y_t, because these variables are of the same nature, since they are repeated measurements of the same phenomenon under the same circumstances.

	$\hat{\eta}_{t,y\|x}$					
	$x=1$		$x=2$		$x=3$	
t	$y=1$	$y=2$	$y=1$	$y=2$	$y=1$	$y=2$
1	-3.548	-5.236	-3.672	-13.816	1.128	-1.187
2	-4.066	-13.816	-0.393	-13.816	3.082	0.445
3	-13.122	-13.816	2.694	-2.650	2.752	1.805
4	-2.942	-4.092	1.025	-13.816	13.816	3.090
5	-2.018	-3.580	1.140	-13.816	13.816	4.989

Table 3: *Estimated global logits $\eta_{t,y|x}$'s under the basic LM model*

			$\hat{\pi}_{x\|w}$		
x	$\hat{\lambda}_x$	w	$x=1$	$x=2$	$x=3$
1	0.791	1	0.911	0.068	0.021
2	0.137	2	0.090	0.746	0.163
3	0.072	3	0.000	0.128	0.872

Table 4: *Estimated initial probabilities λ_x's and transition probabilities $\pi_{x|w}$'s under the basic LM model*

The deviance of the resulting model with respect to the initial model is 23.58 with a p-value equal to 0.600, and therefore this restriction cannot be rejected. We then tried to add some restrictions on the latent structure. In particular, for the hypothesis $\pi_{3|1} = \pi_{1|3} = 0$, which implies that the transition matrix is tridiagonal, the LR statistic with respect to the previous model is equal to 2.02 with a p-value of 0.172, whereas this statistic is equal to 4.67 with a p-value of 0.059 for the hypothesis that the transition matrix is upper triangular. Finally, for the null hypothesis

of absence of transition, the LR statistic is equal to 233.73 with a p-value less than 10^{-4}. In the light of these results, we chose as our final model the one based on parametrization (18) and the hypothesis that the transition matrix is tridiagonal; the latter implies that a transition from latent state w to latent state x is possible only when $x = w - 1$ or $x = w + 1$. The parameter estimates of this model are shown in Tables 5 and 6.

x	$\hat{\psi}_x$	y	$\hat{\delta}_y$
1	0.000	1	0.165
2	5.751	2	0.686
3	10.876		

Table 5: *Estimates of the parameters ψ_x's and δ_y's for the final LM model*

				$\hat{\pi}_{x\|w}$		
x	$\hat{\psi}_x$	$\hat{\lambda}_x$	w	$x = 1$	$x = 2$	$x = 3$
1	0.000	0.896	1	0.835	0.165	0.000
2	5.751	0.089	2	0.070	0.686	0.244
3	10.876	0.015	3	0.000	0.082	0.918

Table 6: *Estimated initial probabilities λ_x's and transition probabilities $\pi_{x|w}$'s for the final LM model*

In Figure 5 we also show, for $t = 1, \ldots, s$, the estimates $\hat{\lambda}_{t,x}$'s of the marginal probabilities of the latent states and the estimate $\widehat{\overline{\psi}}_t$ of the average tendency to use marijuana, computed as described in Section 3.3. Finally, in Figure 5 we show, again for $t = 1, \ldots, s$, the posterior estimates $\hat{r}_t(x|\boldsymbol{y})$'s and $\hat{\psi}_t(\boldsymbol{y})$, given the response configuration $\boldsymbol{y} = (\,0 \quad 0 \quad 1 \quad 2 \quad 2\,)$. Note that $\hat{\psi}_t(\boldsymbol{y})$ increases much faster in time than $\widehat{\overline{\psi}}_t$ and therefore we can conclude that, for a subject with this response configuration, the tendency to use marijuana increases much faster than the average. At occasion t, the distance from the average could be simply measured by $\widehat{\psi}_t(\boldsymbol{y}) - \widehat{\overline{\psi}}_t$.

6 Conclusions

A general framework has been proposed for the formulation and testing of linear hypotheses on the transition probabilities of a class of LM models for discrete variables having a longitudinal structure. These hypotheses may be combined with restrictions on the conditional distribution of the response variables given the latent process, which may be formulated as in a generalized linear model. In order to test linear hypotheses on the transition probabilities, we recommend the use of the LR statistic and we prove that the null asymptotic distribution of this statistic is of chi-bar-squared type, i.e. a mixture of chi-squared distributions with weights that may be

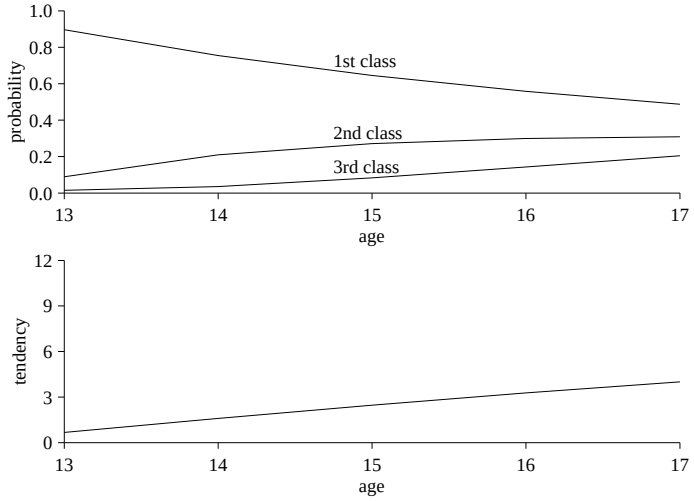

Figure 5: *Estimated probabilities* $\hat{\lambda}_{t,x}$*'s and estimated average tendency to use marijuana* $\hat{\bar{\psi}}_t$

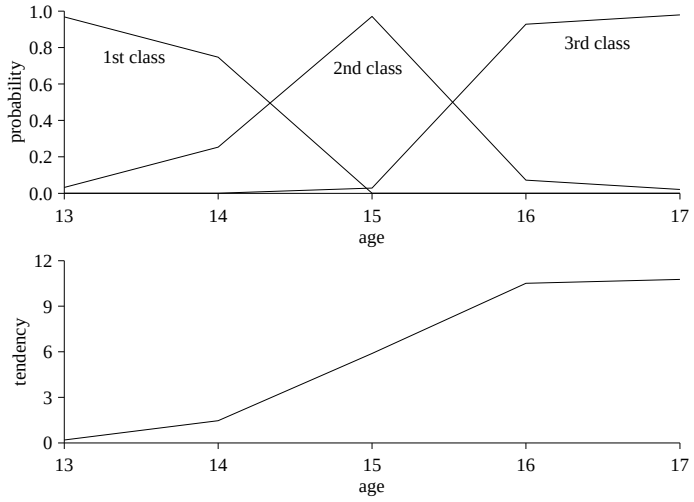

Figure 6: *Estimated probabilities* $\hat{r}_t(x|\boldsymbol{y})$*'s and estimated tendency to use marijuana* $\hat{\psi}_t(\boldsymbol{y})$ *for a subject with response configuration* $\boldsymbol{y} = (0 \quad 0 \quad 1 \quad 2 \quad 2)$

simply computed by simulation. This is one of the main findings of the paper. Note that we have only dealt explicitly with the case of hypotheses on the latent process since hypotheses on the conditional distribution of the response variables given the latent process may be tested on the basis of standard inferential results when this distribution is parametrized as within our approach.

The proposed approach covers the case of i.i.d. observations and therefore it is directly applicable in a contingency table setting. However, it may be easily extended to the case in which individual covariates are present. For instance, we could parameterize the conditional distribution of any response variable given the corresponding latent variable as in a generalized linear model with coefficients depending on the latter. In practice, this results in a model with occasion-specific random effects which follow a first-order Markov chain. The results concerning

the distribution of the LR statistic for testing hypotheses on the transition matrix of the latent process hold for these models as well, but with minor adjustments.

Another possible extension of the proposed approach is to the case of linear inequality constraints on the transition probabilities. Estimating a LM model under constraints of this type is not a difficult task and may be performed by means of a modified version of the EM algorithm illustrated in Section 3. In particular, in order to take these constraints into account, the maximization of $\ell_2^\dagger(\boldsymbol{\beta})$ during the M step of this algorithm must be performed by solving a series of constrained least square problems similar to those solved for the maximization of $\ell_3^\dagger(\boldsymbol{\gamma})$. On the other hand, the asymptotic distribution of the LR statistic for testing hypotheses of this type is not in general of chi-bar squared type and may be difficult to use in practice for computing p-values. For this reason, and since it does not seem that hypotheses of particular interest may be expressed through inequality constraints on the transition probabilities, we explicitly considered only the case of equality constraints.

Acknowledgments

I would like to thank the Editor, an Associate Editor and three anonymous Referees for their stimulating comments.

Appendix

A1: relation between $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$

We have that $\boldsymbol{\pi} = \boldsymbol{a} + \boldsymbol{A}\boldsymbol{\rho}$, where $\boldsymbol{a}$ is obtained by stacking the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_c$ one below the other, with $\boldsymbol{a}_x$ denoting a c-dimensional column vector of zeros apart from the x-th element equal to 1, and $\boldsymbol{A}$ is a block diagonal matrix with blocks $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_c$, where $\boldsymbol{A}_x = \overline{\boldsymbol{I}}_{c,x} - \boldsymbol{a}_x \boldsymbol{1}'_{c-1}$ and $\overline{\boldsymbol{I}}_{c,x}$ is obtained by removing the x-th column from a $\boldsymbol{I}_c$ matrix.

A2: derivative of $\boldsymbol{p}$ with respect to $\boldsymbol{\theta}'$

The derivative matrix $\boldsymbol{Q}(\boldsymbol{\theta})$ has elements

$$p^{(\alpha_j)}(\boldsymbol{y}) = \frac{\partial p(\boldsymbol{y})}{\partial \alpha_j}, \quad p^{(\beta_j)}(\boldsymbol{y}) = \frac{\partial p(\boldsymbol{y})}{\partial \beta_j}, \quad p^{(\gamma_j)}(\boldsymbol{y}) = \frac{\partial p(\boldsymbol{y})}{\partial \gamma_j},$$

arranged in the appropriate order, which may be computed on the basis of the same recursion defined in Section 2 to compute $p(\boldsymbol{y})$. Let $\boldsymbol{\lambda}^{(\alpha_j)}$ denote the derivative of the vector $\boldsymbol{\lambda}$ with respect to α_j, $\boldsymbol{\Pi}^{(\beta_j)}$ that of $\boldsymbol{\Pi}$ with respect to β_j and $\boldsymbol{\phi}_{t,y}^{(\gamma_j)}$ that of $\boldsymbol{\phi}_{t,y}$ with respect to γ_j. Consequently,

$p^{(\theta)}(\boldsymbol{y}) = \boldsymbol{q}^{(\theta)}(\boldsymbol{y}_{\leqslant s})'\mathbf{1}_s$, with θ that may correspond to α_j, β_j or γ_j, where

$$\boldsymbol{q}^{(\alpha_j)}(\boldsymbol{y}_{\leqslant t}) = \begin{cases} \text{diag}(\boldsymbol{\phi}_{1,y_1})\boldsymbol{\lambda}^{(\alpha_j)} & \text{if } t = 1 \\ \text{diag}(\boldsymbol{\phi}_{t,y_t})\boldsymbol{\Pi}'\boldsymbol{q}^{(\alpha_j)}(\boldsymbol{y}_{\leqslant t-1}) & \text{otherwise} \end{cases}$$

$$\boldsymbol{q}^{(\beta_j)}(\boldsymbol{y}_{\leqslant t}) = \begin{cases} \mathbf{0}_c & \text{if } t = 1 \\ \text{diag}(\boldsymbol{\phi}_{t,y_t})(\boldsymbol{\Pi}^{(\beta_j)})'\boldsymbol{q}(\boldsymbol{y}_{\leqslant t-1}) + \text{diag}(\boldsymbol{\phi}_{t,y_t})\boldsymbol{\Pi}'\boldsymbol{q}^{(\beta_j)}(\boldsymbol{y}_{\leqslant t-1}) & \text{otherwise} \end{cases}$$

$$\boldsymbol{q}^{(\gamma_j)}(\boldsymbol{y}_{\leqslant t}) = \begin{cases} \text{diag}(\boldsymbol{\phi}^{(\gamma_j)}_{1,y_1})\boldsymbol{\lambda} & \text{if } t = 1 \\ \text{diag}(\boldsymbol{\phi}^{(\gamma_j)}_{t,y_t})\boldsymbol{\Pi}'\boldsymbol{q}(\boldsymbol{y}_{\leqslant t-1}) + \text{diag}(\boldsymbol{\phi}_{t,y_t})\boldsymbol{\Pi}'\boldsymbol{q}^{(\gamma_j)}(\boldsymbol{y}_{\leqslant t-1}) & \text{otherwise} \end{cases}.$$

References

Archer, G. E. B. and Titterington, D. M. (2002), Parameter estimation for hidden Markov chains, *Journal of Statistical Planning and Inference*, **108** , pp. 365-390.

Auranen, K., Arjas, E., Leino, T. and Takala, A. K. (2000), Transmission of pneumococcal carriage in families: A latent Markov process model for binary longitudinal data, *Journal of the American Statistical Association*, **95**, pp. 1044-1053.

Bartolucci, F. and Forcina, A. (2000), A Likelihood Ratio Test For MTP_2 Within Binary Variables, *Annals of Statistics*, **28**, pp. 1206-1218.

Bartolucci, F. and Forcina, A. (2005), Likelihood inference on the underlying structure of IRT models, *Psychometrika*, **70**, pp. 31-43.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics*, **41**, pp. 164-171.

Bijleveld, C. C. J. H. and Mooijaart, A. (2003), Latent Markov Modelling of Recidivism Data, *Statistica Neerlandica*, **57**, pp. 305-320.

Chernoff, H. (1954), On the distribution of the likelihood ratio, *Annals of Mathematical Statistics*, **25**, pp. 573-578.

Cook, R. J., Ng, E. T. M. and Meade, M. O. (2000), Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models, *Biometrics*, **56**, pp. 1109-1117.

Dardanoni, V. and Forcina, A. (1998), A Unified approach to likelihood inference on stochastic orderings in a nonparametric context, *Journal of the American Statistical Association*, **93**, 1112-1123.

De Boeck, P. and Wilson, M. (2004), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, Springer, Berlin.

Dempster A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, 1-22.

Elliot, D. S., Huizinga, D. and Menard, S. (1989), *Multiple Problem Youth: Delinquence, Substance Use and Mental Health Problems.* Springer-Verlag, New York.

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

Hambleton, R. K. and Swaminathan, H. (1985), *Item Response Theory: Principles and Applications*, Kluwer Nijhoff Publishing, Boston.

Humphreys, K. (1998), The latent Markov chain with multivariate random effects – An evaluation of instruments measuring labor market status in the British Household Panel Study, *Sociological Methods and Research*, **26**, pp. 269-299.

Kelderman, H. and Rijkes, C. P. M. (1994), Loglinear multidimensional IRT models for polytomously scored items, *Psychometrika*, *59*, 147-177.

Langeheine, R. (2002), Latent Markov Chains, in *Advances in Latent Class Analysis*, A. L. McCutcheon and J. A. Hagenaars editors, University Press, Cambridge.

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

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

Levinson, S. E., Rabiner, L. R. and Sondhi, M. M. (1983), An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition, *Bell System Technical Journal*, **62**, pp. 1035-1074.

Lindsay, B., Clogg, C. and Grego, J. (1991), Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis, *Journal of the American Statistical Association*, **86**, 96-107.

MacDonald, I. L. and Zucchini, W. (1997), *Hidden Markov and other Models for Discrete-Valued Time Series*, Chapman and Hall, New York.

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

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

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

Poulsen, C. S. (1982), *Latent structure analysis with choice modeling applications*, Aarhus School of Business Administration and Economics, Aarhus.

Poulsen, C. S. (1990), Mixed Markov and latent Markov modelling applied to brand choice behaviour, *International Journal of Research in Marketing*, **7**, pp. 5-19.

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**, pp 321-333.

Rothenberg, T. (1971), Identification in parametric models, *Econometrica*, **39**, pp. 577-591.

Samejima, F. (1996), Evaluation of mathematical models for ordered polytomous responses, *Behaviormetrika*, **23**, pp. 17-35.

Scott, S. L. (2002), Bayesian methods for hidden Markov models: Recursive computing in the 21st century, *Journal of the American Statistical Association*, **97**, pp. 337-351.

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

Self, S. G. and Liang, K.-Y. (1987), Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association*, **82**, pp. 605-610.

Shapiro, A. (1985), Asymptotic Distribution of Test Statistics in the Analysis of Moment Structures under Inequality Constraints, *Biometrika*, **72**, pp. 133-144.

Shapiro, A. (1988), Towards a Unified Theory of Inequality Constrained Testing in Multivariate Analysis, *International Statistical Review*, **56**, pp. 49-62.

Shi, N.-Z., Zheng, S.-R. and Guo, J. (2005), The restricted EM algorithm under inequality restrictions on the parameters, *Journal of Multivariate Analysis*, **92**, pp. 53-76.

Silvapulle, M. J. and Sen, P. K. (2004), *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*, Wiley, New York.

Van de Pol, F. J. R. and Langeheine, R. (1990), Mixed Markov latent class models, *Sociological Methodology*, in C.C. Clogg editor, Blackwell, Oxford, pp. 213-247.

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

Vermunt, J. K. and Hagenaars, J. A. (2004), Ordinal longitudinal data analysis, *Methods in Human Growth Research* R.C. Hauspie, N. Cameron and L. Molinari editors, pp. 374-393. Cambridge, Cambridge University Press, Cambridge.

Visser, I., Raijmakers, M. E. J. and Molenaar, P. C. M. (2002), Fitting hidden Markov models to psychological data, *Scientific Programming*, **10**, pp. 185-199.

Wiggins, L. M. (1973), *Panel Analysis*, Elsevier, Amsterdam.