Modeling Longitudinal Data with Application to Educational and Psychological Measurement

Francesco Bartolucci*¹

* Department of Economics, Finance and Statistics - University of Perugia - IT

Conference JCS-CLADAG 2012

¹bart@stat.unipg.it
The main aim is to show how latent variable models for longitudinal data based on a *hidden Markov chain* may be used to for the analysis of item response data observed at the same or at different occasions.

These models are here referred to as *latent Markov models* (Wiggins, 1973; Bartolucci, Farcomeni & Pennoni, 2012).

*Main applications* are in:

- **education**: the outcomes are responses to test items about a specific subject (e.g., sample of students responding items in Mathematics).
- **psychology**: the outcomes indicate the capability in performing a specific task (e.g., sample of children asked to recognize certain symbols).

The outcomes are typically binary (*dichotomously-scored items*) even if the case of categorical/ordinal outcomes (*polytomously-scored items*) may be of interest.
Example data in education

- Data provided by the *Educational Testing Service* (Bartolucci, 2006)

- Data concern responses of 1,510 students to 12 items on Mathematics within *National Assessment of Educational Progress* 1996 project:

  1. Round to thousand place
  2. Write fraction that represents shaded region
  3. Multiply two negative integers
  4. Reason about sample space (number correct)
  5. Find amount of restaurant tip
  6. Identify representative sample
  7. Read dials on a meter
  8. Find \((x, y)\) solution of linear equation
  9. Translate words to symbols
  10. Find number of diagonals in polygon from a vertex
  11. Find perimeter (quadrilateral)
  12. Reason about betweenness
The items are **dichotomously-scored** (1 for correct response, 0 for wrong response), so that the data have the following structure:

<table>
<thead>
<tr>
<th>examinee</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1509</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1510</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Scientific questions

In analyzing item responses, certain issues of interest may be dealt with by *standard models* of Item Response Theory (IRT) or Latent Class (LC) type:

1. estimation of the **difficulty level** of each item
2. assessment of **other characteristics** of each item (e.g., discriminating power)
3. prediction of the **ability level** of every examinee and/or classification of the examinees according to the ability levels

These models assume that, for every subject, the **ability level is constant** and then **learning-through-training phenomena** or **tiring effects** are ruled out.

Through a *latent Markov model* we can also model the evolution of the ability in time and then test an IRT or LC model for the data at hand (against violations of the constant ability assumption)
Latent Markov (LM) model

- **Basic notation**:
  - \( n \): sample size
  - \( T \): number of test items
  - \( Y^{(t)}_i \): response of subject \( i \) to item \( t \) (with categories \( 0, \ldots, c - 1 \))
  - \( Y_i \): vector of responses associated to subject \( i \), \( Y_i = (Y^{(1)}_i, \ldots, Y^{(T)}_i) \)
  - \( y^{(t)}_i \): specific realization of \( Y^{(t)}_i \)
  - \( y_i \): specific realization of \( Y_i \), \( y_i = (y^{(1)}_i, \ldots, y^{(T)}_i) \)

In applying the LM model, the responses to the items are considered on the same footing as *longitudinal responses*, with every time occasion corresponding to a specific item.

The connection with longitudinal structures motivates the *notation* \( Y^{(t)}_i \) for the response to an item (this notation may be simply generalized to the multivariate context).
Model assumptions

- The model is based on the following **assumptions**:
  - **local independence**: for $i = 1, \ldots, n$, the response variables in $Y_i$ are conditionally independent given a latent process $U_i = (U_i^{(1)}, \ldots, U_i^{(T)})$
  - the latent process $U_i$ follows a **first-order Markov chain** with state space \{1, \ldots, $k$\}, initial probabilities $\pi_u$, $u = 1, \ldots, k$, and transition probabilities $\pi_u^{(t)}|\bar{u}, \bar{u}, u = 1, \ldots, k$
  - the latent processes $U_1, \ldots, U_n$ are **mutually independent** so that also the response vectors $Y_1, \ldots, Y_n$ are mutually independent

- The latent states $u$ correspond to different **latent classes** in the population from which the sample comes

- Under suitable constraints (to be introduced), these latent classes are interpreted in terms of **different levels of ability**
The LM model may then be seen as a *generalization of the LC model* (Lazarsfeld & Henry, 1968), in which subjects are allowed to move between latent classes (corresponding to different ability levels).
Model parameters

- **Parameters of the LM model:**
  - initial probabilities: \( \pi_u = p(U_i^{(1)} = u), \quad u = 1, \ldots, k \)
  - transition probabilities:
    \[
    \pi^{(t)}_{u|\bar{u}} = p(U_i^{(t)} = u | U_i^{(t-1)} = \bar{u}), \quad t = 2, \ldots, T, \quad \bar{u}, u = 1, \ldots, k
    \]
  - conditional response probabilities:
    \[
    \phi^{(t)}_{y|u} = p(Y_i^{(t)} = y | U_i^{(t)} = u), \quad t = 1, \ldots, T, \quad u = 1, \ldots, k, \quad y = 0, \ldots, c - 1
    \]

- **Number of free parameters:**
  \[
  \# \text{par} = Tk(c - 1) + k - 1 + (T - 1)k(k - 1)
  \]

- In practice, **constraints on the parameters** \( \phi^{(t)}_{y|u} \) and \( \pi^{(t)}_{u|\bar{u}} \) are also necessary to make the model more parsimonious and to ensure identifiability.
Constraints on the parameters

- **Constraints of time homogeneity:**
  - of the transition probabilities (meaning that the evolution of the ability is always the same):
    \[ \pi_{u|\bar{u}}^{(t)} = \pi_{u|\bar{u}}, \quad t = 2, \ldots, k, \quad \bar{u}, u = 1, \ldots, k \]
  - of the conditional response probabilities (justified only in the presence of repeated measurements):
    \[ \phi_{y|u}^{(t)} = \phi_{y|u}, \quad t = 1, \ldots, T, \quad u = 1, \ldots, k, \quad y = 0, \ldots, c - 1 \]
  
- **More general constraints** may be expressed by letting:
  - \( \phi_{u}^{(t)} \): vector with elements \( \phi_{y|u}^{(t)}, y = 0, \ldots, c - 1 \)
  - \( \rho_{\bar{u}}^{(t)} \): vector of the off-diagonal elements \( \pi_{u|\bar{u}}^{(t)} \) in the \( \bar{u} \)-th row of the transition matrix (in red)

\[
\Pi^{(t)} = \begin{pmatrix}
\pi_{1|1}^{(t)} & \pi_{1|2}^{(t)} & \pi_{1|3}^{(t)} \\
\pi_{2|1}^{(t)} & \pi_{2|2}^{(t)} & \pi_{2|3}^{(t)} \\
\pi_{3|1}^{(t)} & \pi_{3|2}^{(t)} & \pi_{3|3}^{(t)}
\end{pmatrix}
\]
Constraints on the measurement model component (conditional distribution of the responses given the latent process):

\[ g(\phi_u^{(t)}) = W_u^{(t)} \beta \]

- \( g(\cdot) \): link function based on standard logits with binary responses or generalized logits (local, global, etc.) with categorical responses with more than two categories

Constraints on the latent model component (distribution of the latent process):

\[ \rho_{\bar{u}}^{(t)} = Z_{\bar{u}}^{(t)} \delta \]

By requiring that \( W_{\bar{u}}^{(t)} = W_{\bar{u}} \) and/or \( Z_{\bar{u}}^{(t)} = Z_{\bar{u}} \) we include the constraint of time homogeneity in the measurement and/or latent model component

A linear model is formulated for the transition probabilities in order to allow the constraint that certain probabilities are equal to 0 (however inequality constraints on \( \delta \) need to be considered)
Constraints on the transition probabilities

- The simplest constraints is that the *transition matrix is diagonal* (transitions between latent states are not allowed and then LM model specializes into LC model):

\[
\Pi^{(t)} = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
\end{pmatrix}
\]

- A more sophisticated constraint is that the *transition matrix is symmetric*:

\[
\Pi^{(t)} = \begin{pmatrix}
1 - (\delta_1 + \delta_2) & \delta_1 & \delta_2 \\
\delta_1 & 1 - (\delta_1 + \delta_3) & \delta_3 \\
\delta_2 & \delta_3 & 1 - (\delta_2 + \delta_3) \\
\end{pmatrix}
\]

- Provided that latent states are ordered, triangular transition matrices may be used to formulate a certain type of *evolution of the ability across time* (e.g., non decreasing in time)
Latent Markov Rasch (LMR)

- The LMR model is formulated by assuming that the transition probabilities are time homogenous and that

$$\log \frac{\phi_{1|u}^{(t)}}{\phi_{0|u}^{(t)}} = \log \frac{p(Y_i^{(t)} = 1|U_i^{(t)} = u)}{p(Y_i^{(t)} = 0|U_i^{(t)} = u)} = \alpha_u - \psi^{(t)}$$

- $\alpha_u$: ability of subjects in latent state $u$
- $\psi^{(t)}$: difficulty level of item $t$

- Number of **free parameters** (considering one identifiability constraint on the parameters $\alpha_u$ or $\psi^{(t)}$):

$$\#\text{par} = (k + T - 1) + (k - 1) + k(k - 1)$$

- The model makes sense only if the test items are administered in the same order to all examinees

- It may be seen as a generalization of the LC Rasch model, which holds when the transition matrices are diagonal
Manifest distribution of the response variables

- **Conditional distribution of the response variables given the latent variables** (hypothesis of local independence):

  \[ p(y_i|U_i = u) = \prod_{t=1}^{T} \phi_{y_i(t)}^{(t)}|u^{(t)}}, \quad u = (u^{(1)}, \ldots, u^{(T)}) \]

- **Distribution of the latent variables** (first-order Markov chain):

  \[ p(U_i = u) = \pi_{u^{(1)}} \prod_{t=2}^{T} \pi_{u^{(t)}}|u^{(t-1)} \]

- **Manifest distribution of the response variables** (via marginalization):

  \[
p(y_i) = \sum_{u^{(1)}=1}^{k} \cdots \sum_{u^{(T)}=1}^{k} p(y_i|U_i = u)p(U_i = u) \]

  \[
  = \sum_{u^{(1)}=1}^{k} \cdots \sum_{u^{(T)}=1}^{k} \pi_{u^{(1)}} \phi_{y_i^{(1)}}^{(1)}|u^{(1)} \prod_{t=2}^{T} \pi_{u^{(t)}}|u^{(t-1)} \phi_{y_i^{(t)}}^{(t)}|u^{(t)} \]
Computing the manifest distribution involves a sum over the $k^T$ vectors $u$ (this number may be huge).

The manifest distribution is efficiently computed by the Baum and Welch forward recursion (Baum et al., 1970; Welch, 2003) based on:

$$q^{(t)}(u, y_i) = p(U^{(t)} = u, y_i^{(1)}, \ldots, y_i^{(t)})$$

- for $t=1$, we compute:
  $$q^{(1)}(u, y_i) = \pi_u \phi^{(1)}_{y_i}$$

- for $t>1$, we compute:
  $$q^{(t)}(u, y_i) = \sum_{\bar{u}} q^{(t-1)}(\bar{u}, y_i) \pi^{(t)}_{u|\bar{u}} \phi^{(t)}_{y_i}$$

At the end of the recursion, the manifest distribution is obtained as (the number of operations increases linearly in $T$):

$$p(y_i) = \sum_u q^{(T)}(u, y_i)$$
Maximum likelihood estimation

- Given a sample of $n$ independent units that provided response vectors $y_1, \ldots, y_n$, the model log-likelihood is:

$$\ell(\theta) = \sum_{i=1}^{n} \log p(y_i) = \sum_{y} n_y \log p(y)$$

- $\theta$: vector of all free parameters affecting $p(y_i)$
- $y$: generic response configuration $y = (y^{(1)}, \ldots, y^{(T)})$
- $n_y$: frequency of the response configuration $y$ in the observed sample

- $\ell(\theta)$ may be maximized $\theta$ by the Expectation-Maximization (EM) algorithm (Baum et al., 1970; Dempster et al., 1977)

- The EM algorithm alternates two steps (E-step, M-step) until convergence in $\ell(\theta)$ and is based on the complete data log-likelihood (that we could compute if we knew the latent state of every subject at each occasion)
**Complete data log-likelihood:**

\[ \ell^* (\theta) = \sum_u \sum_y m_{uy} \log [p(y|u)p(u)] \]

- \( m_{uy} \): (unobservable) frequency of the latent process configuration \( u \) and response configuration \( y \)

**A more explicit version** of the complete data log-likelihood:

\[ \ell^* (\theta) = \sum_{t=1}^{T} \sum_{u=1}^{k} \sum_{y=0}^{c-1} a^{(t)}_{uy} \log \phi_{y|u}^{(t)} \]

\[ + \sum_{u=1}^{k} b_u^{(1)} \log \pi_u + \sum_{t=2}^{T} \sum_{\bar{u}=1}^{k} \sum_{u=1}^{k} b_{\bar{u}u}^{(t)} \log \pi_{u|\bar{u}}^{(t)} \]

- \( a^{(t)}_{uy} \): frequency of subjects responding by \( y \) at occasion \( t \) and belonging to latent state \( u \) at the same occasion
- \( b_u^{(t)} \): frequency of subjects in latent state \( u \) at occasion \( t \)
- \( b_{\bar{u}u}^{(t)} \): number of transitions from latent state \( \bar{u} \) to state \( u \) at occasion \( t \)
E-step

The E-step consists of computing the *conditional expected value* of every frequency $a_{uy}^{(t)}$, $b_{u}^{(t)}$, and $b_{uu}^{(t)}$ given the observed data and the current parameter vector:

- $\hat{a}_{uy}^{(t)} = \mathbb{E}(a_{uy}^{(t)} | \text{obs. data}) = \sum_{i=1}^{n} p(U_{i}^{(t)} = u | y_i) I(y_i^{(t)} = y)$
- $\hat{b}_{u}^{(t)} = \mathbb{E}(b_{u}^{(t)} | \text{obs. data}) = \sum_{i=1}^{n} p(U_{i}^{(t)} = u | y_i)$
- $\hat{b}_{uu}^{(t)} = \mathbb{E}(b_{uu}^{(t)} | \text{obs. data}) = \sum_{i=1}^{n} p(U_{i}^{(t-1)} = \bar{u}, U_{i}^{(t)} = u | y_i)$

The posterior probabilities are computed by a suitable *forward-backward recursion* (Baum *et al.*, 1970)

Formulae based on the *aggregated data* (i.e., frequencies $n_y$) may be directly used for the expected values
M-step

- The M-step consists of updating the model parameters by maximizing \( \mathbb{E}[\ell^*(\theta)] \) (defined as \( \ell^*(\theta) \) with every frequency substituted by its expected value):

  - for the initial probabilities \( \pi_u \) there is an explicit solution:
    \[
    \pi_u = \frac{\hat{b}_u^{(1)}}{n}, \quad u = 1, \ldots, k
    \]

  - for the transition probabilities \( \pi_u(t) \) there are explicit solutions for many models; under the constraint of time homogeneity (\( \pi_u(t) = \pi_u | \bar{u} \)) for instance:
    \[
    \pi_{u|\bar{u}} = \frac{\hat{b}_{u|u}^{(t)}}{\hat{b}_{u}^{(t-1)}}, \quad \bar{u}, u = 1, \ldots, k
    \]

  otherwise an iterative algorithm is necessary

  - for the conditional response probabilities \( \phi_{y|u}(t) \) a simple iterative algorithm is necessary; if these are unconstrained:
    \[
    \phi_{y|u}^{(t)} = \frac{\hat{a}_{uy}^{(t)}}{\hat{b}_{u}^{(t)}}, \quad t = 1, \ldots, T, \quad u = 1, \ldots, k, \quad y = 0, \ldots, c - 1
    \]
Model selection and hypothesis testing

In applying LM model, the *choice of the number of latent states* \( k \) is necessary, when this number is not *a priori* fixed.

The two most used criteria are the *Akaike Information Criterion* (AIC; Akaike, 1973) and the *Bayesian Information Criterion* (BIC; Schwarz, 1978) which are based on the minimum of:

\[
AIC = -2\ell(\hat{\theta}) + 2 \times \#\text{par}
\]

\[
BIC = -2\ell(\hat{\theta}) + \log(n) \times \#\text{par}
\]

\( \hat{\theta} \): maximum likelihood estimate of the model of interest

*Other criteria* which take into account the quality of the classification may also be used (e.g., Normalized Entropy Criterion, NEC; Celeux & Soromenho, 1996)

AIC, BIC and related criteria may also be used to *select a constrained LM model*, once \( k \) has been fixed.
Likelihood ratio testing of hypotheses

- For a fixed $k$, to test a linear hypotheses $H_0$ on the parameters of the LM model we can use the *likelihood ratio* (LR) statistic

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

- $\hat{\theta}_0$: constrained maximum likelihood estimate of $\theta$ under $H_0$
- $\hat{\theta}$: unconstrained maximum likelihood estimate of $\theta$

- For a hypothesis of type

$$H_0 : M\delta = 0 \quad \text{(under the linear model } \rho_{\bar{u} \bar{u}}^{(t)} = Z_{\bar{u}\bar{u}}^{(t)} \delta)$$

we are not in a *regular inferential problem* and the standard theory applies for deriving the null asymptotic distribution of $D$ because of the non-negativity constraint on the transition probabilities $\pi_{u|\bar{u}}^{(t)}$

- *These constraints* may be directly expressed as:

$$\delta \geq 0, \quad TZ_{\bar{u}\bar{u}}^{(t)} \delta \leq 1_k, \quad \text{with } T = I_k \otimes 1_{k-1}'$$
The asymptotic distribution of $D$ under a linear hypothesis of type $H_0: M\delta = 0$ has been derived by Bartolucci (2006) by using certain results known in constrained statistical inference (Self & Liang, 1987; Silvapulle & Sen, 2004)

Under suitable regularity conditions, we have:

$$D \xrightarrow{d} \chi^2_{m-g} + \chi^2(\Sigma_0, O^g)$$

- $m$: number of constraints on $\delta$
- $g$: number of elements of $\delta$ constrained to be 0 under $H_0$
- $\chi^2(\Sigma_0, O^g)$: chi-bar squared distribution (mixture of standard chi-squared distributions with suitable weights)
- $\Sigma_0$: asymptotic variance-covariance matrix of the maximum likelihood estimator of the elements $\delta$ constrained to be 0 under $H_0$
- $O^g$: orthant of dimension $g$
The asymptotic distribution is a \textit{mixture of chi-squared} distributions, so that a \textit{p}-value for \( D \) may be computed as:

\[
p\text{-value} = \sum_{j=0}^{g} w_j(\Sigma_0, O^g) p(\chi_{j+m-g}^2 \geq d)
\]

\( w_j(\Sigma_0, O^g) \): weights which may be estimated (with the required precision) through a simple Monte Carlo algorithm

When the \textit{transition matrix only depends on one parameter} \( \delta \), the asymptotic distribution of the LR statistic \( D \) for testing \( H_0 : \delta = 0 \) is:

\[
\frac{1}{2} \chi_0^2 + \frac{1}{2} \chi_1^2 \implies p\text{-value} = \frac{1}{2} p(\chi_1^2 \geq d)
\]

\textbf{Example} in which the transition matrix depends on only one parameter:

\[
\Pi^{(t)} = \Pi = \begin{pmatrix}
1 - 2\delta & \delta & \delta \\
\delta & 1 - 2\delta & \delta \\
\delta & \delta & 1 - 2\delta
\end{pmatrix}, \quad t = 1, \ldots, T
\]

so that \( H_0 \) is equivalent to \( \Pi^{(t)} = I_k \) for all \( t \) (LC model)
Path prediction

- The posterior probabilities $p(U_{i}^{(t)} = u | y_{i})$ may be used to predict the latent state of a subject at a given time occasion; latent state assigned to subject $i$ at occasion $t$:

$$\hat{u}_{i}^{(t)} : p(U_{i}^{(t)} = \hat{u}_{i}^{(t)} | y_{i}) = \max_{u} p(U_{i}^{(t)} = u | y_{i})$$

- Under an LMR (or similar) model, the ability of subject $i$ may be predicted by:

$$\sum_{u=1}^{k} \hat{\alpha}_{u} p(U_{i}^{(t)} = u | y_{i})$$

- More sophisticated is the problem of path prediction, that is, finding the most likely sequence $\hat{u}_{i} = (\hat{u}_{i}^{(1)}, \ldots, \hat{u}_{i}^{(T)})$ for subject $i$:

$$\hat{u}_{i} : p(U_{i}^{(1)} = \hat{u}_{i}^{(1)}, \ldots, U_{i}^{(T)} = \hat{u}_{i}^{(T)} | y_{i}) = \max_{u} p(U_{i} = u | y_{i})$$

- For this aim we need an iterative algorithm due to Viterbi (1967) and further elaborated by Juan and Rabiner (1991)
Application to the NAEP data

- Data analyzed by the LMR model with $k = 3$ latent states (chosen by BIC)

- Estimates of item parameters:
  \[
  \hat{\psi} = (0.000, 0.040, -0.704, 1.013, -1.560, -0.043, \\
  -0.705, -1.250, -0.387, -0.587, -2.532, -2.587)'
  \]

- Estimates of ability parameters and initial and transition probabilities:
  \[
  \hat{\alpha} = \begin{pmatrix} -0.619 \\ 0.967 \\ 2.561 \end{pmatrix}, \quad \hat{\pi} = \begin{pmatrix} 0.163 \\ 0.483 \\ 0.354 \end{pmatrix}, \quad \hat{\Pi} = \begin{pmatrix} 1.000 & 0.000 & 0.000 \\ 0.000 & 0.982 & 0.018 \\ 0.000 & 0.011 & 0.989 \end{pmatrix}
  \]

- The easiest item is the 12th, whereas the most difficult is the 4th
- The 1st class is that of the least capable subjects and the 3rd is that of the most capable subjects
- The 2nd class is the largest in the population and there is a small chance of transition only between the last two classes
Comparison with the LCR model

- The LMR model may be used to test the hypothesis that the ability level remains constant \((H_0 : \Pi = I_k)\), which is crucial in IRT models.
- For the LMR model, we have \(\ell(\hat{\theta}_0) = -10,163.6\) with 22 free parameters.
- For the LC Rasch (LCR) model (LMR model under \(H_0\)), we have \(\ell(\hat{\theta}_0) = -10,166.3\) with 16 free parameters.
- The LR test statistic between the LMR model and the LCR model is
  \[
  D = -2(-10,166.3 + 10,163.6) = 5.5
  \]
  with a \(p\)-value of 0.08 and therefore there is not enough evidence against the LCR model.
- The estimates of the difficulty and ability parameters under the LCR model are very close to those under the LMR model.
Application to marijuana consumption dataset

- Application with ordinal variables in which the hypothesis of time-constant latent trait is rejected (Bartolucci, 2006)

- Dataset taken from five annual waves (1976-80) of the National Youth Survey (Elliot et al., 1989)

- The dataset is based on \( n = 237 \) respondents aged 13 years in 1976. The use of marijuana is measured through of \( T = 5 \) ordinal variables, one for each annual wave, with \( c = 3 \) categories:
  - 0: never in the past year
  - 1: no more than once a month in the past year
  - 2: more than once a month in the past year

- Different LM models with \( k = 3 \) latent states are used, since 3 is the number of observable categories and then the response variables may be seen as measurements with error of the true marijuana consumption (latent state)
Assumed parametrization of the conditional response probabilities (based on global logits):

\[
\log \frac{p(Y_i^{(t)} \geq y|U_i^{(t)} = u)}{p(Y_i^{(t)} < y|U_i^{(t)} = u)} = \log \frac{\phi_y^{(t)} + \cdots + \phi_{c-1|u}^{(t)}}{\phi_{0|u}^{(t)} + \cdots + \phi_{y-1|u}^{(t)}} = \alpha_u + \psi_y, \quad y = 1, \ldots, c - 1
\]

- \(\alpha_u\): tendency to use marijuana for the subjects in latent class \(u\)
- \(\psi_y\): threshold which is time homogenous (repeated measurement of the same variable)

The transition probabilities are time homogenous (\(\pi_{u|\bar{u}}^{(t)} = \pi_{u|\bar{u}}\))

The model may be seen as an extension of the Graded Response Model (with constant difficulty levels) in which the latent construct may have a certain evolution
Hypotheses on the transition probabilities

- Starting from the LM model with time-homogenous transition probabilities, **different constraints** on these probabilities are considered.

- For the hypothesis \( \pi_{3|1} = \pi_{1|3} = 0 \) (**tridiagonal transition matrix**):
  \[
  D = 2.02, \quad p\text{-value} = 0.172
  \]

- For the hypothesis \( \pi_{u|\bar{u}} = 0, u < \bar{u} \) (**upper triangular transition matrix**):
  \[
  D = 4.67, \quad p\text{-value} = 0.059
  \]

- For the hypothesis \( \pi_{u|\bar{u}} = 0, u \neq \bar{u} \) (**diagonal transition matrix**):
  \[
  D = 233.73, \quad p\text{-value} < 10^{-4}
  \]

- The **selected LM final model** is based on a tridiagonal transition matrix (transition from latent state \( \bar{u} \) to latent state \( u \) is only possible when \( u = \bar{u} - 1 \) or \( u = \bar{u} + 1 \)).
Parameter estimates

Estimates of the parameters in the conditional response probabilities:

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\alpha}_u )</th>
<th>y</th>
<th>( \hat{\psi}_y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.165</td>
</tr>
<tr>
<td>2</td>
<td>5.751</td>
<td>2</td>
<td>0.686</td>
</tr>
<tr>
<td>3</td>
<td>10.876</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Estimates of the initial and transition probabilities:

|   | \( \hat{\pi}_u \) | \( \hat{\pi}_{u|\bar{u}} \) |
|---|-------------------|----------------------------|
|   |                   |   | \( u = 1 \) | \( u = 2 \) | \( u = 3 \) |
| 1 | 0.896             | 1 | 0.835       | 0.165       | 0.000       |
| 2 | 0.089             | 2 | 0.070       | 0.686       | 0.244       |
| 3 | 0.015             | 3 | 0.000       | 0.082       | 0.918       |

There is an overall tendency towards an increase in the use of marijuana (comparison of the probabilities above and below the diagonal)
Marginal probabilities of the latent states

- Representation of the marginal probabilities of the latent stats and predicted of the overall tendency of marijuana consumption (corresponding average of $\hat{\alpha}_u$):
Application to psychological test dataset

- Application with **binary variables** collected by a **complex design** (Bartolucci & Solis-Trapala, 2010)

- Dataset collected by a **psychological experiment on children** which is aimed at measuring two psychological constructs (Shimmon, 2004):
  - **Attentional Flexibility (AF):** ability to shift the attention from one set of rules to another
  - **Inhibitory Control (IC):** ability to suppress responses to an irrelevant stimulus

- The data are collected from the administration of a battery of tests to 115 children (aged 34-55 months at the beginning) during two **single testing sessions** one week apart which are replicated over two 6-month periods

- Each session is based on a **series of tasks** which are scored by a binary response variable:
  - “day/night" or “abstract pattern" which are aimed at measuring IC and are repeated 16 times
  - “DCCS face-up" or “DCCS face-down" which are aimed at measuring AF and are repeated 6 times
The length of the response sequence is 44 (16+6+16+6) for each pair of testing sessions; the number of observations for the same subject is at most $3 \times 44 = 132$ (there is a drop out for certain subjects)

a) experimental design 1

b) experimental design 2
Model formulation

- An **LM model is selected** (by BIC) which is based on $k = 3$ latent states

- Two different levels of AF and IC are associated with each latent state (**multidimensionality**)

- The probability of success to each task depends on AF or IC as in the **two-parameter logistic model** (including a discriminant index parameter)

- The **initial probabilities** of the latent Markov chain depend on age through a multinomial logistic parameterization

- There are **four distinct transition matrices**:
  - $\Pi_1 = \Pi_2$: transition within the sequence of IC tasks ("day/night" or "abstract pattern")
  - $\Pi_3 = \Pi_4$: transition within the sequence of AF tasks ("DCCS face-up" or "DCCS face-down")
  - $\Pi_5 = \Pi_6$: transition from AF and IC (a week later)
  - $\Pi_7 = \Pi_8$: transition from AF and IC (six months later)
Estimates of the most interesting parameters

- Estimates of the **conditional response probabilities**:

<table>
<thead>
<tr>
<th>task</th>
<th>latent state</th>
</tr>
</thead>
<tbody>
<tr>
<td>day/night (IC)</td>
<td>0.004 0.510 0.994</td>
</tr>
<tr>
<td>abstract pattern (IC)</td>
<td>0.413 0.953 0.998</td>
</tr>
<tr>
<td>DCCS face-down (AF)</td>
<td>0.536 0.000 0.998</td>
</tr>
<tr>
<td>DCCS face-up (AF)</td>
<td>0.526 0.000 1.000</td>
</tr>
</tbody>
</table>

- The latent states are ordered according to IC ability, but not according to the AF ability (multidimensionality); the hypothesis of unidimensionality is rejected.

- For the second state there are zero probabilities due to the fact some participants scored all zeros during the full 6-item sequences (typical of these experiments).

- It is confirmed that “day/night" task is more difficult than “abstract pattern"
Estimates of the transition probabilities:

<table>
<thead>
<tr>
<th>Latent state</th>
<th>Within IC items</th>
<th></th>
<th>Within AF items</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.981</td>
<td>0.019</td>
<td>0.000</td>
<td>0.588</td>
</tr>
<tr>
<td>2</td>
<td>0.023</td>
<td>0.915</td>
<td>0.063</td>
<td>0.005</td>
</tr>
<tr>
<td>3</td>
<td>0.011</td>
<td>0.037</td>
<td>0.951</td>
<td>0.023</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Latent state</th>
<th>Between AF and IC a week later</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.576</td>
</tr>
<tr>
<td>2</td>
<td>0.061</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Latent state</th>
<th>Between AF and IC 6 months later</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>0.300</td>
<td>0.700</td>
</tr>
<tr>
<td>2</td>
<td>0.180</td>
<td>0.117</td>
</tr>
<tr>
<td>3</td>
<td>0.122</td>
<td>0.136</td>
</tr>
</tbody>
</table>

The hypothesis that any of the transition matrices is diagonal is rejected leading to the conclusion that there is an an evolution in the ability level.

The improvement in the ability level (transition to the 3rd state) is relevant after one week.
Path Prediction

- Predicted sequence of latent state for certain subjects (the response is represented by a dot at 1 when wrong and at 3 when correct):
In certain contexts we want to associate more item responses to the same time occasion, so that the latent trait remains constant for these items \( \Rightarrow \) **multivariate LM model**

Applications in which sample units are collected in clusters (e.g., students collected in classes) may also be of interest \( \Rightarrow \) **multilevel LM model**

**Basic notation** (multivariate case):
- \( n \): sample size
- \( T \): number of time occasions (batteries of tests)
- \( r \): number of response variables observed at each occasion \( t \)
- \( Y_{ij}^{(t)} \): response of subject \( i \) to item \( j \) administered at occasion \( t \)
- \( Y_i^{(t)} \): vector of responses observed at occasion \( t \), \( Y_i^{(t)} = (Y_{i1}^{(t)}, \ldots, Y_{ir}^{(t)}) \)
- \( y_{ij}^{(t)} \): realization of \( Y_{ij}^{(t)} \)
- \( y_i^{(t)} \): realization of \( Y_i^{(t)} \), \( y_i^{(t)} = (y_{i1}^{(t)}, \ldots, y_{ir}^{(t)}) \)
Multivariate LM model

The model is based on an extension of the assumption of local independence according to which:

- the response vectors $Y_i^{(1)}, \ldots, Y_i^{(T)}$ are conditionally independent given the latent process $U_i = (U_i^{(1)}, \ldots, U_i^{(T)})$
- the response variables $(Y_{i1}^{(t)}, \ldots, Y_{ir}^{(t)})$ in every vector $Y_i^{(t)}$ are conditionally independent given the latent variable $U_i^{(t)}$

New parameters of the model are the conditional response probabilities:

$$
\phi_{jy|u}^{(t)} = p(Y_{ij}^{(t)} = y | U_i^{(t)} = u)
$$

The assumption of local independence implies that the conditional distribution of the responses given the latent process is:

$$
p(y_i^{(t)} | U_i^{(t)} = u) = \prod_{j=1}^{r} \phi_{jy|u}^{(t)}
$$

$$
p(y_i^{(1)}, \ldots, y_i^{(T)} | U_i = u) = \prod_{t=1}^{T} p(y_i^{(t)} | U_i^{(t)} = u^{(t)}), \quad u = (u^{(1)}, \ldots, u^{(T)})
$$
The latent process $U_i$ is still assumed to follow a **first-order Markov chain**

The manifest distribution of $Y_{i}^{(1)}, \ldots, Y_{i}^{(T)}$, that is, $p(y_{i}^{(1)}, \ldots, y_{i}^{(T)})$, is computed through recursions similar to those used in the univariate case.

Even in the multivariate case, **constraints** may be expressed on:

- **measurement model component**: distribution of the response variables given the latent variable (about the parameters $\phi_{jy|u}^{(t)}$) by a Generalized Linear parametrization:
  \[ g(\phi_{jy|u}^{(t)}) = W_{j|u}^{(t)} \beta \]
  - $\phi_{jy|u}^{(t)}$: vector with elements $\phi_{jy|u}^{(t)}$ for $y = 0, \ldots, c - 1$
  - $g(\cdot)$: link function based on standard logits or similar effects

- **latent model component**: distribution of the latent process (about the parameters $\pi_{u|\bar{u}}^{(t)}$), based on a linear model of type:
  \[ \rho_{\bar{u}}^{(t)} = Z_{\bar{u}}^{(t)} \delta \]
  - $\rho_{\bar{u}}^{(t)}$: vector of the off-diagonal elements in the $\bar{u}$-th row of the $t$-th transition matrix
Likelihood inference for the multivariate LM model

- **Log-likelihood** of the multivariate LM model:

\[
\ell(\theta) = \sum_{i=1}^{n} \log[p(y_{i}^{(1)}, \ldots, y_{i}^{(T)})];
\]

it may be maximized through an EM algorithm similar to that described for the univariate model (Bartolucci, Pennoni & Francis, 2007)

- A **hypothesis** \(H_0\) on the parameters may be tested through the LR statistic

\[
D = -2[\ell(\hat{\theta}_0) - \ell(\hat{\theta})]
\]

- One of the **hypotheses of main interest** is \(H_0: \Pi^{(t)} = I_k\) (no transition between latent states)

- The **null asymptotic distribution** of \(D\) under linear hypotheses on the transition probabilities is still of chi-bar squared type
Multilevel latent Markov model

- **Basic notation** (multivariate case):
  - $H$: number of clusters
  - $n_h$: size of cluster $h$
  - $Y_{hi}^{(t)}$: response of subject $i$ in cluster $h$ to item $j$ administered at occasion $t$
  - $Y_{hi}^{(t)}$: vector of responses observed at occasion $t$, $Y_{hi}^{(t)} = (Y_{hi1}^{(t)}, \ldots, Y_{hir}^{(t)})$
  - $y_{hi}^{(t)}$: realization of $Y_{hi}^{(t)}$
  - $y_{hi}^{(t)}$: realization of $Y_{hi}^{(t)}$, $y_{hi}^{(t)} = (y_{hi1}^{(t)}, \ldots, y_{hir}^{(t)})$

- **Assumptions of the multilevel LM model**:
  - for every unit $hi$, local independence between and within the response vectors $Y_{hi}^{(t)}$ is assumed given a latent process $V_{hi} = (V_{hi1}^{(1)}, \ldots, V_{hi}^{(T)})$ which follows a first-order Markov chain
  - the effect of each cluster $h$ on the parameters of the Markov chain (for the subjects belonging to the cluster) is accounted for by a latent variable $U_h$
  - latent/observable variables referred to different subjects in the same cluster $h$ are assumed to be conditionally independent given $U_h$
Each latent variable $U_h$ is assumed to have a *discrete distribution* with $k_1$ support points, whereas the distribution of the latent process $V_{hi}$ is based on $k_2$ states.

The model includes *possible covariates* collected in:

- $x_h$: vector of cluster-level covariates (for cluster $h$)
- $z_{hi}^{(t)}$: vector of individual-level covariates (for unit $hi$ at occasion $t$)

The *latent structure* depends on the parameters:

- **mass probabilities** for the distribution of $U_h$:
  \[
  \omega_{hu} = p(U_h = u | x_h)
  \]

- **initial probabilities** for the latent process $V_{hi}$:
  \[
  \pi_{hiv|u} = p(V_{hi}^{(1)} = v | U_h = u, z_{hi}^{(1)})
  \]

- **transition probabilities** for the latent process $V_{hi}$:
  \[
  \pi_{hiv|u\bar{v}} = p(V_{hi}^{(t)} = v | U_h = u, V_{hi}^{(t-1)} = \bar{v}, z_{hi}^{(t)})
  \]
A multinomial logit parametrization may be adopted for the distribution of $U_h$:

$$\log \frac{\omega_{hu}}{\omega_{h1}} = \tau_0 u + \mathbf{x}_h' \mathbf{\tau}_1 u, \quad u = 2, \ldots, k_1$$

With ordered latent states (as in the Rasch model) it is convenient to model initial and transition probabilities through global logits.

For the initial probabilities:

$$\log \frac{p(V_h^{(1)} \geq v | U_h = u, z_h^{(1)})}{p(V_h^{(1)} < v | U_h = u, z_h^{(1)})} = \gamma_0 u + \gamma_1 v + (z_h^{(1)})' \gamma_2, \quad v = 2, \ldots, k_2$$

For the transition probabilities:

$$\log \frac{p(V_h^{(t)} \geq v | U_h = u, V_h^{(t-1)} = \bar{v}, z_h^{(t)})}{p(V_h^{(t)} < v | U_h = u, V_h^{(t-1)} = \bar{v}, z_h^{(t)})} = \delta_0^{(t)} + \delta_{1v}^{(t)} + (z_h^{(t)})' \delta_2^{(t)}, \quad \bar{v} = 1, \ldots, k_2, \quad v = 2, \ldots, k_2$$
Manifest distribution of the response variables

- The model assumptions imply that the \textit{manifest distribution} of the response variables for all subjects in the same cluster \(h\) is:

\[
p(y_h) = \sum_{u=1}^{k_1} \omega_{hu} \prod_{i=1}^{n_h} p(y_{hi}^{(1)}, \ldots, y_{hi}^{(T)}|U_h = u)
\]

- \(y_h\): vector of all responses provided by the units in cluster \(h\), which is made of the subvectors \(y_{hi}^{(t)}\), \(i = 1, \ldots, n_h\), \(t = 1, \ldots, T\)

- \(p(y_{hi}^{(1)}, \ldots, y_{hi}^{(T)}|U_h = u)\) may be efficiently computed by the usual \textit{forward recursion} of Baum and Welch

- Given the independence between clusters, the \textit{manifest distribution} of all response variables is:

\[
p(y_1, \ldots, y_H) = \prod_{h=1}^{H} p(y_h)
\]
Maximum likelihood estimation

- The **log-likelihood of the multilevel LM model** is

\[
\ell(\theta) = \sum_{h=1}^{H} \log[p(y_h)]
\]

- As usual, \(\ell(\theta)\) may be maximized by the **EM algorithm** which alternates two steps until convergence:
  - **E-step**: compute the posterior distribution of each \(U_h, V_{hi}^{(t)}|U_h\), and \((V_{hi}^{(t-1)}, V_{hi}^{(t)})|U_h\) by the usual recursions
  - **M-step**: update the parameters in \(\theta\) by maximizing the expected value of the complete-data log-likelihood by iterative algorithms

- The same criteria as for standard LM models may be used for **model selection** and **hypothesis testing**

- Both \(k_1\) and \(k_2\) need to be selected for this model; the preferred criterion is the **Bayesian Information Criterion** which is based on the minimization of the usual **BIC** index
Application to educational dataset

- Application of **multilevel LM model** (Bartolucci, Pennoni & Vittadini, 2011) based on:
  - Rasch parametrization of the conditional probabilities of the responses $Y_{hij}^{(t)}$
  - multinomial logit parametrization for the distribution of $U_h$
  - global logit parametrization for the initial and transition probabilities of $V_{hi}^{(t)}$

- Data collected in Lombardy by the **Italian Institute for the Evaluation of the Educational System** (INVALSI)

- The sample concerns a cohort of 1,246 students who progressed from Grade 6 to Grade 8 (middle school), so that there are $T = 3$ waves

- The students are 994 from 13 public schools and 252 from 7 nonpublic schools; the total number of classes (clusters) is 77

- At the three occasions, 28, 30, and 39 items on Mathematics were administered to all students; some items were replicated across waves
Selection of the number of support points ($k_1$) for each $U_h$ and latent states ($k_2$) for each $V_{hi}^{(t)}$

<table>
<thead>
<tr>
<th>$k_1$</th>
<th>$k_2$</th>
<th>$\ell(\hat{\theta})$</th>
<th>BIC</th>
<th>#par</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-77656.77</td>
<td>153737.40</td>
<td>85</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>-71266.31</td>
<td>143316.67</td>
<td>110</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>-70187.30</td>
<td>141229.93</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>-69410.43</td>
<td>140110.98</td>
<td>181</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td><strong>-69374.80</strong></td>
<td><strong>140089.61</strong></td>
<td><strong>188</strong></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>-69356.58</td>
<td>140103.06</td>
<td>195</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>-69315.71</td>
<td>140156.75</td>
<td>214</td>
</tr>
</tbody>
</table>

$k_1 = 4$ and $k_2 = 6$ levels are chosen for the two latent variables
Estimated conditional probabilities for each latent state (items are decreasingly ordered according to the difficulty level)
On the basis of the estimates of the parameters for the initial and transition probabilities, we characterized *four different groups of class of students* (A, B, C, D) which have different impact on the improvement from grade 6 to 7 and from grade 7 to 8.

Classes in group A are the *worst*, whereas classes in group D are the *best* in terms of improvement of the students; classes in groups B and C have an *intermediate impact* on the ability level of students.

The effect of the *covariates for the classes* may be understood on the basis of the conditional distribution of $U_h$ given these covariates:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>type of school</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>public</td>
<td>0.325</td>
<td>0.233</td>
<td>0.376</td>
<td>0.066</td>
</tr>
<tr>
<td>nonpublic</td>
<td>0.786</td>
<td>0.000</td>
<td>0.000</td>
<td>0.214</td>
</tr>
<tr>
<td><strong>years of activity of the school</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\geq 17.5$</td>
<td>0.425</td>
<td>0.189</td>
<td>0.314</td>
<td>0.072</td>
</tr>
<tr>
<td>$&lt; 17.5$</td>
<td>0.363</td>
<td>0.196</td>
<td>0.289</td>
<td>0.152</td>
</tr>
<tr>
<td><strong>ratio students/teacher</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$&lt; 8$</td>
<td>0.541</td>
<td>0.090</td>
<td>0.205</td>
<td>0.164</td>
</tr>
<tr>
<td>$\geq 8$</td>
<td>0.337</td>
<td>0.245</td>
<td>0.363</td>
<td>0.054</td>
</tr>
</tbody>
</table>

The *type of school* has a strong effect on the ability of students.
Main references


Latent Markov Models for Longitudinal Data

Francesco Bartolucci
Alessio Farcomeni
Fulvia Pennoni