Introduction to latent variable models

Lecture 3

Francesco Bartolucci
Department of Economics, Finance and Statistics
University of Perugia, IT
bart@stat.unipg.it
Outline

- Latent Markov model
- Maximum likelihood estimation via EM algorithm
- Constrained formulations of the model
- Likelihood ratio testing of linear hypotheses on the parameters
Latent Markov (LM) model (Wiggins, 1973)

- This is a model for the analysis of *longitudinal categorical* data which is used in many contexts, e.g. psychological and educational measurement, criminology and educational measurement.

- Let $\mathbf{Y} = \{Y_t, \; t = 1, \ldots, T\}$ denote the vector of categorical response variables. The LM model assumes that:
  
  ▶ (local independence, LI) the response variables are conditionally independent given a latent process $\mathbf{U} = \{U_t, \; t = 1, \ldots, T\}$

  ▶ the latent process $\mathbf{U}$ follows a *first-order Markov chain* with state space $\{1, \ldots, k\}$, initial probabilities $\pi_u$ and transition probabilities $\pi_{v|u}$, with $u, v = 1, \ldots, k$. 

• Each latent state $u$ corresponds to a class of subjects in the population, and is characterized by:

- **Initial probability**
  \[ \pi_u = p(U_1 = u) \]

- **Transition probabilities** (which may also be time-specific in the non-homogenous case)
  \[ \pi_{v|u} = p(U_t = v|U_{t-1} = u), \quad t = 2, \ldots, T, \quad v = 1, \ldots, k \]

- **Distribution of the response variables**
  \[ \phi_{t,y|u} = p(Y_t = y|U_t = u), \quad t = 1, \ldots, T, \quad y = 0, \ldots, l - 1 \]

  In the basic version we have time homogeneity \( (\phi_{t,y|u} = \phi_y|u) \)
Manifest distribution

- Because of LI, the **conditional distribution** of $Y$ given $U$ is:

$$p(y|u) = p(Y = y|U = u) = \prod_t \phi_{t,y_t|u_t}$$

- **Distribution** of $U$:

$$p(u) = p(U = u) = \pi_u \prod_{t>1} \pi_{u_t|u_{t-1}}$$

- **Manifest distribution** of $Y$:

$$p(y) = p(Y = y) = \sum_u p(y|u)p(u)$$

- This may be *efficiently computed* through suitable recursions known in the hidden Markov literature (MacDonald & Zucchini, 1997)
Comparison with Latent Class (LC) model

- The LM model may then be seen as a **generalization of the LC model** (Lazarsfeld and Henry, 1968) in which subjects are allowed to move between latent classes.
Maximum likelihood (ML) estimation

- **Log-likelihood** of the model

\[ \ell(\theta) = \sum_y n(y) \log[p(y)] \]

▷ \( \theta \): vector of all model parameters \((\pi_u, \pi_v|u, \phi_{t,y}|u)\)

- \( \ell(\theta) \) may be maximized with respect to \( \theta \) by an *Expectation-Maximization (EM) algorithm* (Dempster et al., 1977)

- Here the *complete data* correspond to the frequencies \(m(u, y)\) of any latent process configuration \(u\) and any response configuration \(y\)
The algorithm \textit{alternates two steps} until convergence in $\ell(\theta)$:

\textbf{E}: for any $u$ and $y$ compute $\hat{m}(u, y)$, the \textit{conditional expected value} of $m(u, y)$ given $n(y)$ and the current value of $\theta$

\textbf{M}: update $\theta$ by \textit{maximizing the log-likelihood of the complete data}

$$
\ell^*(\theta) = \sum_u \sum_y m(u, y) \log[p(y|u)p(u)]
$$

with any frequency $m(u, y)$ substituted by the corresponding expected value $\hat{m}(u, y)$ computed during the E-step

The E-step is performed by means of \textit{certain recursions} which may be easily implemented through matrix notation (Bartolucci, 2006)
An application to marijuana consumption dataset

- 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 $s = 5$ ordinal variables, one for each annual wave, with 3 categories:
  - 1: never in the past year
  - 2: no more than once a month in the past year
  - 3: more than once a month in the past year

- Exercise: fit the basic LM model using the LMest package in R
Constrained LM models

- Several constraints may be formulated on the LM model. We consider in particular:

- **Constrains on the conditional distribution** of response variables of type

  \[ \eta(\phi) = Z\gamma, \quad K\gamma \geq 0 \]

  with \( \eta(\phi) \) denoting a suitable link function of the probabilities \( \phi_{t,y_t|u_t} \). We can have for instance a LM version of the Rasch model

- **Constrains on the transition probabilities** of type

  \[ \rho = W\delta \]

  with \( \rho \) denoting the vector of the off-diagonals elements of \( \Pi \), i.e. \( \pi_{u|u}, \ u, v = 1, \ldots, k, \ u \neq v \)
Latent Markov Rasch (LMR) model

- It may be seen as a generalization of the LC Rasch model in which the distribution of any $Y_t$ depends on a specific latent variable $U_t$, with

  $$\log \frac{\phi_{t,1|u}}{\phi_{t,0|u}} = \log \frac{p(Y_t = 1|U_t = u)}{p(Y_t = 0|U_t = u)} = \xi_u - \beta_t$$

- The latent variables $U_1, \ldots, U_T$ are assumed to follow a homogeneous first-order Markov chain with initial probabilities $\pi_u$ and transition probabilities $\pi_v|u$

- The model makes sense only if the test items are administered in the same order to all subjects, the same order with that the response variables $Y_t$ are arranged in the vector $Y$. 
Constraints on the transition probabilities

• By the linear form $\rho = W\delta$ we can also formulate the constraint that two or more transition probabilities are equal to 0, e.g.

$$
\Pi = \begin{pmatrix}
1 - (\delta_1 + \delta_2) & \delta_1 & \delta_2 \\
0 & 1 - \delta_3 & \delta_3 \\
0 & 0 & 1
\end{pmatrix}
$$

in the second case the LM model specializes into the LC model

• These constraints cannot be expressed when a linear form is assumed on a link function of the transition probabilities

• Suitable constraints have to be put on $\delta$ in order to ensure that all the transition probabilities are non-negative
Likelihood ratio (LR) testing of linear hypotheses

• 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 ML estimate of $\theta$ under $H_0$

▷ $\hat{\theta}$: unconstrained ML estimate of $\theta$

• For a hypothesis of type

$$H_0: L\gamma = 0, \quad \eta(\phi) = Z\gamma,$$

we are in a *regular inferential problem* and the standard theory applies for deriving the null asymptotic distribution of $D$
Hypotheses on the transition probabilities

- For a hypothesis of type

\[ H_0 : M\delta = 0, \quad \rho = W\delta, \]

we are not in a regular inferential problem because of the non-negativity constraint on the transition probabilities \( \pi_v|_u \)

- The non-negativity constraint may be directly formulated on the parameters \( \delta \) as

\[ \delta \geq 0, \quad TW\delta \leq 1_k, \quad \text{with} \quad T = I_k \otimes 1'_{k-1} \]

- We are not in a regular inferential problem since it may happen that some elements of \( \delta \) are equal to 0 under \( H_0 \), and so the true value of the parameters is on the boundary of the parameter space.
• 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 and Liang, 1987, Silvapulle and Sen, 2004)

• Under *suitable regularity conditions*, we have:

\[
D \xrightarrow{d} \chi^2_{m-g} + \bar{\chi}^2(\Sigma_0, \mathcal{O}^g)
\]

▷ \( m \) : number of constraints on \( \delta \)

▷ \( g \) : number of elements of \( \delta \) constrained to be 0 under \( H_0 \)

▷ \( \bar{\chi}^2(\Sigma_0, \mathcal{O}^g) \): *chi-bar squared* distribution

▷ \( \Sigma_0 \) : asymptotic variance-covariance matrix of the MLE of the elements \( \delta \) constrained to be 0 under \( H_0 \)

▷ \( \mathcal{O}^g \) : orthant of dimension \( g \)
• The asymptotic distribution is a *mixture of chi-squared* distributions, so that a \( p \)-value for \( D \) may be computed as

\[
p\text{-value} = \sum_{i=0}^{g} w_i(\Sigma_0, O^g) p(\chi^2_{i+m-g} \geq d)
\]

\( \triangleright \) \( w_i(\Sigma_0, C) \): weights which may be estimated (with the required precision) through a simple Monte Carlo algorithm

• When the *transition matrix only depends on one parameter* \( \delta \), e.g.

\[
\Pi = \begin{pmatrix}
1 - 2\delta & \delta \\
\delta & 1 - 2\delta
\end{pmatrix},
\]

the asymptotic distribution of the LR statistic \( D \) for testing \( H_0 : \delta = 0 \), equivalent to \( \Pi = I_k \) (LC model), is:

\[
\frac{1}{2} \chi_0^2 + \frac{1}{2} \chi_1^2 \quad \Longrightarrow \quad \text{\( p \)-value} = \frac{1}{2} p(\chi_1^2 \geq d)
\]
Chi-bar squared distribution $\bar{\chi}^2(\Sigma, C)$

- This is the *distribution of the random variable* (e.g. Shapiro, 1988)

  $$Q = V'\Sigma^{-1}V - \min_{\hat{V} \in C}(\hat{V} - V)'\Sigma^{-1}(\hat{V} - V)$$

  ▶ $V$: random vector of dimension $v$ with distribution $\mathcal{N}(0, \Sigma)$

  ▶ $C$: convex cone in $\mathbb{R}^v$

- It corresponds to a *mixture of chi-squared distributions*, so that

  $$p(Q \geq q) = \sum_{i=0}^{v} w_i(\Sigma, C)p(\chi^2_i \geq q)$$

  Through a mixture we can also express the distribution $\chi^2_h + \bar{\chi}^2(\Sigma, C)$

- The *weights* $w_i(\Sigma, C)$ may be computed explicitly only in particular cases; these weights may always be estimated (with the required precision) through a simple Monte Carlo algorithm
An application to educational testing data

- Application to a dataset concerning the responses of a group of $n = 1,510$ examinees to a set of $J = 12$ test items on Mathematics.

- The dataset has been extrapolated from a larger dataset collected in 1996 by the Educational Testing Service (USA) within a project called the National Assessment of Educational Progress (NAEP).

- The items were administered to all examinees in the same order and therefore the use of the LMR model is appropriate for studying possible violations of the LI assumption.

- For this dataset we chose $k = 3$ latent classes; it corresponds to the number of classes for which the LCR model has the smallest BIC (Schwarz, 1978).
Parameter estimates under the LMR model

- Estimates of item and latent process parameters:

\[ \hat{\beta} = \begin{pmatrix} 0.000 & 0.040 & -0.704 & 1.013 & -1.560 & -0.043 \\ -0.705 & -1.250 & -0.387 & -0.587 & -2.532 & -2.587 \end{pmatrix} \]

\[ \hat{\xi} = \begin{pmatrix} -0.619 \\ 0.967 \\ 2.561 \end{pmatrix} \quad \hat{\pi} = \begin{pmatrix} 0.163 \\ 0.483 \\ 0.354 \end{pmatrix} \quad \hat{\Pi} = \begin{pmatrix} 1.000 & 0.000 & 0.000 \\ 0.000 & 0.982 & 0.018 \\ 0.000 & 0.011 & 0.989 \end{pmatrix} \]

- 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.
Goodness-of-fit and comparison with LCR model

- The maximum log-likelihood of the LMR model is \( \ell(\hat{\theta}) = -10,163.6 \) with 22 (non-redundant) parameters and its deviance with respect to the saturated model is 2,014.2 with 4,073 degrees of freedom.

- For the LCR model, we have \( \ell(\hat{\theta}_0) = -10,166.3 \) with 16 parameters.

- The LR 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 either the LI assumption or 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

- Assumed parametrization:

\[ \eta_{t,y|u} = \log \frac{p(Y_t > y|U_t = u)}{p(Y_t \leq y|U_t = u)} = \xi_u + \beta_y, \quad y = 1, 2 \]

- \( \eta_{t,y|u} \): \( y \)-th conditional global logit for \( Y_t \) given \( U_t = u \)
- \( \xi_u \): tendency to use marijuana for the subjects in latent class \( u \)
- \( \beta_y \): tendency to use marijuana common to all subjects

- The LR statistic of the resulting LM model with respect to the initial LM model is \( D = 23.58 \) (\( p \)-value=0.600); therefore this parametrization cannot be rejected
Restrictions on the latent process parameters

- For the hypothesis $\pi_{3|1} = \pi_{1|3} = 0$ (tridiagonal transition matrix) the LR statistic with respect to the previous model is $D = 2.02$ ($p$-value=0.172)

- For the hypothesis $\pi_{v|u} = 0$, $v < u$ (triangular transition matrix), we have $D = 4.67$ ($p$-value=0.059)

- For the hypothesis $\pi_{v|u}$, $u \neq v$ (diagonal transition matrix), we have $D = 233.73$ ($p$-value < $10^{-4}$)

- We chose as final model the one based on a tridiagonal transition matrix. This means that a transition from latent state $u$ to latent state $v$ is only possible when $v = u - 1$ or $v = u + 1$
### Parameter estimates

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\xi}_u$</th>
<th></th>
<th>$y$</th>
<th>$\hat{\beta}_y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.165</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5.751</td>
<td>2</td>
<td>0.686</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10.876</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1:** Estimates of the parameters $\xi_u$’s and $\beta_y$’s for the final LM model

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\pi}_{vu}$</th>
<th>$\hat{\pi}_{vu}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.896</td>
<td>0.835</td>
</tr>
<tr>
<td>2</td>
<td>0.089</td>
<td>0.070</td>
</tr>
<tr>
<td>3</td>
<td>0.015</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Table 2:** Estimated initial probabilities $\lambda_u$’s and transition probabilities $\pi_{vu}$’s for the final LM model