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
- Multivariate extension
- Dealing with individual covariates

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 $\boldsymbol{Y} = \{Y_t,\, t = 1, \ldots, T\}$ 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 $\boldsymbol{U} = \{U_t,\, t = 1, \ldots, T\}$
  - the latent process $\boldsymbol{U}$ follows a *first-order Markov chain* with state space $\{1, \ldots, k\}$, initial probabilities π_u and transition probabilities $\pi_{v|u}$, with $u, v = 1, \ldots, k$

- Each latent states 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)

$$\pi_{v|u} = p(U_t = v|U_{t-1} = u), \quad t = 2, \ldots, T, \;\; v = 1, \ldots, k$$

  - ▷ *distribution of the response variables*

$$\phi_{t,y|u} = p(Y_t = y|U_t = u), \quad t = 1, \ldots, T, \;\; y = 0, \ldots, l-1$$

Manifest distribution

- Because of LI, the *conditional distribution* of $\boldsymbol{Y}$ given $\boldsymbol{U}$ is:

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

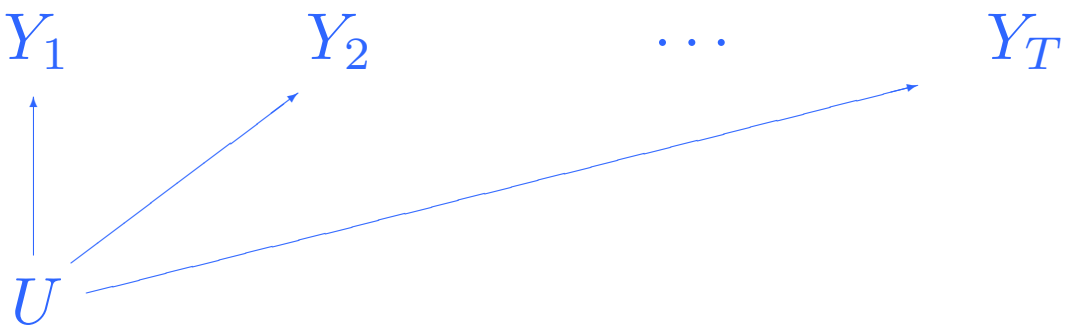
- *Distribution* of $\boldsymbol{U}$: $p(\boldsymbol{u}) = p(\boldsymbol{U}=\boldsymbol{u}) = \pi_{u_1} \prod_{t>1} \pi_{u_t|u_{t-1}}$

- *Manifest distribution* of $\boldsymbol{Y}$: $p(\boldsymbol{y}) = p(\boldsymbol{Y}=\boldsymbol{y}) = \sum_{\boldsymbol{u}} p(\boldsymbol{y}|\boldsymbol{u})p(\boldsymbol{u})$

- This may be *efficiently computed* through suitable recursions known in the hidden Markov literature (MacDonald & Zucchini, 1997)

Comparison with Latent Class (LC) model

LC:



LM:

- 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

$$\ell(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}} n(\boldsymbol{y}) \log[p(\boldsymbol{y})]$$

  - ▷ $\boldsymbol{\theta}$: vector of all model parameters $(\pi_u,\ \pi_{v|u},\ \phi_{t,y|u})$

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

- Here the *complete data* correspond to the frequencies $m(\boldsymbol{u}, \boldsymbol{y})$ of any latent process configuration $\boldsymbol{u}$ and any response configuration $\boldsymbol{y}$

EM algorithm

- The algorithm *alternates two steps* until convergence in $\ell(\boldsymbol{\theta})$:

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

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

$$\ell^*(\boldsymbol{\theta}) = \sum_{\boldsymbol{u}} \sum_{\boldsymbol{y}} m(\boldsymbol{u}, \boldsymbol{y}) \log[p(\boldsymbol{y}|\boldsymbol{u}) p(\boldsymbol{u})]$$

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

- The E-step is performed by means of *certain recursions* which may be easily implemented through matrix notation (Bartolucci, 2006)

Constrained LM models

- *Several constraints* may formulated on the LM model. We consider in particular:
  - ▷ *constraints on the conditional distribution* of response variables of type

$$\boldsymbol{\eta}(\boldsymbol{\phi}) = \boldsymbol{Z}\boldsymbol{\gamma}, \quad \boldsymbol{K}\boldsymbol{\gamma} \geq \mathbf{0}$$

    with $\boldsymbol{\eta}(\boldsymbol{\phi})$ 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
  - ▷ *constraints on the transition probabilities* of type

$$\boldsymbol{\rho} = \boldsymbol{W}\boldsymbol{\delta}$$

    with $\boldsymbol{\rho}$ denoting the vector of the off-diagonals elements of $\boldsymbol{\Pi}$, i.e. $\pi_{v|u},\ u, v = 1, \ldots, k,\ u \neq v$

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 π_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 $\boldsymbol{Y}$

Constraints on the transition probabilities

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

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

  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 $\boldsymbol{\delta}$ 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

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

  - ▷ $\widehat{\boldsymbol{\theta}}_0$: constrained ML estimate of $\boldsymbol{\theta}$ under H_0
  - ▷ $\widehat{\boldsymbol{\theta}}$: unconstrained ML estimate of $\boldsymbol{\theta}$

- For a hypothesis of type

$$H_0 : \boldsymbol{L}\boldsymbol{\gamma} = \mathbf{0}, \quad \boldsymbol{\eta}(\boldsymbol{\phi}) = \boldsymbol{Z}\boldsymbol{\gamma},$$

  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

$$H_0 : \boldsymbol{M}\boldsymbol{\delta} = \mathbf{0}, \quad \boldsymbol{\rho} = \boldsymbol{W}\boldsymbol{\delta},$$

  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 $\boldsymbol{\delta}$ as

$$\boldsymbol{\delta} \geq \mathbf{0}, \quad \boldsymbol{T}\boldsymbol{W}\boldsymbol{\delta} \leq \mathbf{1}_k, \qquad \text{with} \quad \boldsymbol{T} = \boldsymbol{I}_k \otimes \mathbf{1}'_{k-1}$$

- We *are not in a regular inferential problem* since it may happen that some elements of $\boldsymbol{\delta}$ 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 $H_0 : \boldsymbol{M}\boldsymbol{\delta} = \mathbf{0}$ 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:

$$D \xrightarrow{d} \chi^2_{m-g} + \overline{\chi}^2(\boldsymbol{\Sigma}_0, \mathcal{O}^g)$$

  - m : number of constraints on $\boldsymbol{\delta}$
  - g : number of elements of $\boldsymbol{\delta}$ constrained to be 0 under H_0
  - $\overline{\chi}^2(\boldsymbol{\Sigma}_0, \mathcal{O}^g)$: *chi-bar squared* distribution
  - $\boldsymbol{\Sigma}_0$: asymptotic variance-covariance matrix of the MLE of the elements $\boldsymbol{\delta}$ 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

$$p\text{-value} = \sum_{i=0}^{g} w_i(\boldsymbol{\Sigma}_0, \mathcal{O}^g) p(\chi^2_{i+m-g} \geq d)$$

  - ▷ $w_i(\boldsymbol{\Sigma}_0, \mathcal{C})$: weights which may be estimated (with the required precision) through a simple Monte Carlo algorithm

- When the *transition matrix only depends on one parameter* δ, e.g.

$$\boldsymbol{\Pi} = \begin{pmatrix} 1-2\delta & \delta \\ \delta & 1-2\delta \end{pmatrix},$$

the asymptotic distribution of the LR statistic D for testing $H_0 : \delta = 0$, equivalent to $\boldsymbol{\Pi} = \boldsymbol{I}_k$ (LC model), is:

$$\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1 \qquad \Longrightarrow \qquad p\text{-value} = \frac{1}{2}p(\chi^2_1 \geq d)$$

Chi-bar squared distribution $\overline{\chi}^2(\boldsymbol{\Sigma}, \mathcal{C})$

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

$$Q = \boldsymbol{V}'\boldsymbol{\Sigma}^{-1}\boldsymbol{V} - \min_{\widehat{\boldsymbol{V}} \in c}(\widehat{\boldsymbol{V}} - \boldsymbol{V})'\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{V}} - \boldsymbol{V})$$

  - ▷ $\boldsymbol{V}$: random vector of dimension v with distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$
  - ▷ $\mathcal{C}$: convex cone in $\mathcal{R}^v$

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

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

  Through a mixture we can also express the distribution $\chi_h^2 + \overline{\chi}^2(\boldsymbol{\Sigma}, \mathcal{C})$

- The *weights* $w_i(\boldsymbol{\Sigma}, \mathcal{C})$ may 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:

$$\widehat{\boldsymbol{\beta}} = (\ 0.000 \quad 0.040 \quad -0.704 \quad 1.013 \quad -1.560 \quad -0.043$$
$$-0.705 \quad -1.250 \quad -0.387 \quad -0.587 \quad -2.532 \quad -2.587)'$$

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

- 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 $\ell(\widehat{\boldsymbol{\theta}}) = -10,163.6$ 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 $\ell(\widehat{\boldsymbol{\theta}}_0) = -10,166.3$ 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

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

- A LM model with $k = 3$ latent states has initially been used. The model has a deviance of 85.80 with 204 degrees of freedom

Restrictions on the conditional distribution of the response variables

- 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 $U_t = u$
  - ξ_u : tendency to use marijuana for the subjects in latent class u
  - β_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

u	$\hat{\xi}_u$	y	$\hat{\beta}_y$
1	0.000	1	0.165
2	5.751	2	0.686
3	10.876		

Table 1: *Estimates of the parameters ξ_u's and β_y's for the final LM model*

			$\hat{\pi}_{v\|u}$		
x	$\hat{\pi}_u$	u	$v=1$	$v=2$	$v=3$
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

Table 2: *Estimated initial probabilities λ_u's and transition probabilities $\pi_{v|u}$'s for the final LM model*

Multivariate extension

- We have a *vector of response variables* $\boldsymbol{Y}_t = (\, Y_{t1} \quad \cdots \quad Y_{tJ} \,)$ for any time occasion t $(t = 1, \ldots, T)$
- The elements of $\boldsymbol{Y}_t$ are assumed to be *conditionally independent* given a time-specific latent variable U_t (LI), so that

$$p_t(\boldsymbol{y}_t|u) = p(\boldsymbol{Y}_t = \boldsymbol{y}_t|U_t = u) = \prod_j \phi_{tj,y_{tj}|u}$$

$$\phi_{tj,y|u} = p(Y_{tj} = y, U_t = u)$$

- The latent variables $U_1, \ldots, U_T$ are assumed to follow a *first-order Markov chain* (possibly non-homogeneous)
- The *joint distribution* of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ may be computed through recursions similar to those used for the univariate model

Likelihood inference

- The *log-likelihood* of the multivariate LMR model,

$$\ell(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}_1} \cdots \sum_{\boldsymbol{y}_T} n(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T) \log[p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)],$$

  may be maximized through an EM algorithm similar to that described for the univariate model (Bartolucci, Pennoni & Francis, 2006)

- A *hypothesis* H_0 on the parameters may be tested through the LR statistic $D = -2[\ell(\widehat{\boldsymbol{\theta}}_0) - \ell(\widehat{\boldsymbol{\theta}})]$

- One of the *hypotheses of main interest* is $H_0 : \boldsymbol{\Pi} = \boldsymbol{I}_k$ (no transition between latent classes)

- The *null asymptotic distribution* of D under linear hypotheses on the transition probabilities is still of chi-bar squared type

Extensions to include covariates

- Two possible *choices to include individual covariates*:

1. on the *measurement model* so that we have random intercepts.

   With binary response variables we could assume:

$$\phi_{it,y|u} = p(y_{it} = y | U_{it} = u, \boldsymbol{X}_i),$$

$$\log \frac{\phi_{it,1|u}}{\phi_{it,0|u}} = \xi_u + \boldsymbol{x}'_{it}\boldsymbol{\beta}, \quad i = 1, \ldots, n, \;\; t = 1, \ldots, T, \;\; u = 1, \ldots,$$

   Alternative parameterizations are possible with categorical/ordinal response variables even in the multivariate case (Bartolucci & Farcomeni, 2009, *Jasa*)

2. On the model for the *distribution of the latent variables* (via a multinomial logit parametrization):
   - Initial probabilities:

$$\pi_{i,u} = p(U_{i1} = u|\boldsymbol{x}_{i1}), \qquad \log\frac{\pi_{i,u}}{\pi_{i,1}} = \boldsymbol{x}_{i1}'\boldsymbol{\beta}_u, \quad u = 2, \ldots, k$$

   - Transition probabilities:

$$\begin{aligned} \pi_{i,v|u} &= p(U_{it} = v|U_{i,t-1} = u, \boldsymbol{x}_{it}), \\ \log\frac{\pi_{i,v|u}}{\pi_{i,u|u}} &= \boldsymbol{x}_{it}'\boldsymbol{\gamma}_{uv}, \quad u, v = 2, \ldots, k,\ u \neq v \end{aligned}$$

Alternative parameterizations are possible with ordered latent classes ($\lambda_{t1} \leq \cdots \leq \lambda_{tk}$, $t = 1, \ldots, T$)

- The models based on the two extensions have *different interpretations*:
  1. the latent variables are used to account for the *unobserved heterogeneity* and then the model may be seen as a discrete version of the logistic model with one random effect
  2. the *main interest is on a latent variable* which is measured through the observable response variables (e.g. health status) and on how this latent variable depends on the covariates

- Only the *M-step of the EM algorithm* must be modified