Likelihood inference for a latent Markov Rasch model with application to educational assessment

F. Bartolucci

Università di Perugia

bart@stat.unipg.it

Outline

- Latent class Rasch model
- Latent Markov Rasch model
- Likelihood inference for the latent Markov Rasch model:
  - ▷ maximum likelihood estimation
  - ▷ likelihood ratio testing of hypotheses on the parameters
- Application to educational testing data
- Multivariate extension for dealing with longitudinal data

Latent class Rasch (LCR) model

- Version of the Rasch (1961) model in which the ability of a subject is represented by a latent variable Θ having a discrete distribution with k support points (De Leeuw & Verhelst, 1986, Lindsay *et al.*, 1991):

$$\begin{pmatrix} \xi_1 & \cdots & \xi_k \\ \pi_1 & \cdots & \pi_k \end{pmatrix}$$

- Equivalently, the observed sample of subjects is assumed to come from a population made of k groups (or latent classes) with all the subjects in the same group sharing the same ability level

- The conditional distribution of any binary response variable X_j $(j = 1, \ldots, J)$, given the ability, is formulated through a logistic link function, and local independence (LI) is still assumed

- Probability that a subject in the cth latent class (for that $\Theta = \xi_c$) responds correctly to the jth item ($j = 1, \ldots, J$):

$$\lambda_{j|c} = p(X_j = 1|\Theta = \xi_c) = \frac{e^{\xi_c - \beta_j}}{1 + e^{\xi_c - \beta_j}}$$

  - ▷ β_j: difficulty parameter for the item

- Conditional distribution of $\boldsymbol{X} = (\, X_1 \quad \cdots \quad X_J \,)$ given Θ:

$$p(\boldsymbol{x}|c) = p(\boldsymbol{X} = \boldsymbol{x}|\Theta = \xi_c) = \prod_j \lambda_{j|c}^{x_j}(1 - \lambda_{j|c})^{1-x_j}$$

- Manifest distribution of $\boldsymbol{X}$ (marginal with respect to Θ):

$$p(\boldsymbol{x}) = p(\boldsymbol{X} = \boldsymbol{x}) = \sum_c p(\boldsymbol{x}|c)\pi_c$$

Latent Markov Rasch (LMR) model

- It may be seen as a generalization of the LCR model in which the distribution of any X_j depends on a specific latent variable Θ_j

- The latent variables $\Theta_1, \ldots, \Theta_J$ are assumed to follow a homogeneous first-order Markov chain (Wiggins, 1973) with initial probability vector $\boldsymbol{\pi}$ and transition probability matrix $\mathbf{\Pi}$, with

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_k \end{pmatrix} \qquad \mathbf{\Pi} = \begin{pmatrix} \pi_{11} & \cdots & \pi_{1k} \\ \vdots & \ddots & \vdots \\ \pi_{k1} & \cdots & \pi_{kk} \end{pmatrix}$$

- The model makes sense only if the test items are administered in the same order to all the subjects, the same order with that the response variables X_j are arranged in the vector $\boldsymbol{X}$

- The vector of latent variables $\boldsymbol{\Theta} = \begin{pmatrix} \Theta_1 & \cdots & \Theta_J \end{pmatrix}$ may assume k^J possible configurations $\boldsymbol{\xi_c} = \begin{pmatrix} \xi_{c_1} & \cdots & \xi_{c_J} \end{pmatrix}$ indexed by the vector $\boldsymbol{c} = \begin{pmatrix} c_1 & \cdots & c_J \end{pmatrix}$

- Because of LI, the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{\Theta}$ is:

$$p(\boldsymbol{x}|\boldsymbol{c}) = p(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\Theta} = \boldsymbol{\xi_c}) = \prod_j \lambda_{j|c_j}^{x_j}(1 - \lambda_{j|c_j})^{1-x_j}$$

- Distribution of $\boldsymbol{\Theta}$: $p(\boldsymbol{c}) = p(\boldsymbol{\Theta} = \boldsymbol{\xi_c}) = \pi_{c_1} \prod_{j>1} \pi_{c_{j-1}c_j}$

- Manifest distribution of $\boldsymbol{X}$: $p(\boldsymbol{x}) = p(\boldsymbol{X} = \boldsymbol{x}) = \sum_{\boldsymbol{c}} p(\boldsymbol{x}|\boldsymbol{c})p(\boldsymbol{c})$

- $p(\boldsymbol{x})$ may be efficiently computed through suitable recursions known in the hidden Markov literature (MacDonald & Zucchini, 1997)

- The greater flexibility of the LMR model with respect to the LCR model is due to the fact that, in the first model, a subject is allowed to move between latent classes

LCR:

$$\begin{tikzcd}
X_1 & X_2 & \cdots & X_J \\
\Theta \arrow[u] \arrow[ur] \arrow[urrr] & & &
\end{tikzcd}$$

LMR:

$$\begin{tikzcd}
X_1 & X_2 & \cdots & X_J \\
\Theta_1 \arrow[u] \arrow[r] & \Theta_2 \arrow[u] \arrow[r] & \cdots \arrow[r] & \Theta_J \arrow[u]
\end{tikzcd}$$

- The way in which a subject may move between latent classes depends on the transition probabilities $\pi_{c_{j-1}c_j}$

Some examples (with $k = 2$ latent classes)

- A subject may move from the 1st to the 2nd latent class (prob. 0.2) and from the 2nd to the 1st latent class (prob. 0.1):

$$\mathbf{\Pi} = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix}$$

- Only transition from the 1st to the 2nd latent class is allowed (prob. 0.3):

$$\mathbf{\Pi} = \begin{pmatrix} 0.7 & 0.3 \\ 0 & 1 \end{pmatrix}$$

- Transition between latent classes is not allowed (the LMR model specializes into the LCR model):

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

Maximum likelihood (ML) estimation

- Log-likelihood of the model

$$\ell(\boldsymbol{\phi}) = \sum_{\boldsymbol{x}} n(\boldsymbol{x}) \log[p(\boldsymbol{x})]$$

  - ▷ $\boldsymbol{\phi}$: vector of all model parameters $(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\Pi})$
  - ▷ $n(\boldsymbol{x})$: observed frequency of the response configuration $\boldsymbol{x}$

- $\ell(\boldsymbol{\phi})$ may be maximized with respect to $\boldsymbol{\phi}$ by an Expectation-Maximization (EM) algorithm (Dempster *et al*., 1977), which is based on the concept of *complete data*

- In our context, the complete data correspond to the frequencies $m(\boldsymbol{c}, \boldsymbol{x})$ of any latent process configuration $\boldsymbol{c}$ and any response configuration $\boldsymbol{x}$

EM algorithm

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

  E: for any $\boldsymbol{c}$ and $\boldsymbol{x}$ compute $\widehat{m}(\boldsymbol{c},\boldsymbol{x})$, the conditional expected value of $m(\boldsymbol{c},\boldsymbol{x})$ given the observed data $n(\boldsymbol{x})$ and the current value of $\boldsymbol{\phi}$

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

$$\ell^{\star}(\boldsymbol{\phi}) = \sum_{\boldsymbol{c}} \sum_{\boldsymbol{x}} m(\boldsymbol{c},\boldsymbol{x}) \log[p(\boldsymbol{x}|\boldsymbol{c})p(\boldsymbol{c})],$$

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

- The E-step is performed by means of certain recursions which may be easily implemented through matrix notation

Likelihood ratio (LR) testing of hypotheses

- To test a hypotheses H_0 on the parameters of the LMR model we can use the LR statistic

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

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

- The asymptotic distribution of D under hypotheses formulated on $\mathbf{\Pi}$ is of chi-bar squared ($\bar{\chi}^2$) type (Bartolucci, 2006), a distribution known in constrained statistical inference (Silvapulle & Sen, 2004)

- By using the $\bar{\chi}^2$ distribution, we can compute a p-value for D on the basis of which we can decide whether to reject or not H_0

The chi-bar squared distribution

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

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

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

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

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

  On the basis of this formula we can to compute a p-value for D

- The weights $w_i(\mathcal{C}, \boldsymbol{\Sigma})$ may computed explicitly only in particular cases; these weights may always be estimated (with the required precision) through a simple Monte Carlo algorithm

Testing the LCR model

- One of the most interesting hypothesis on $\mathbf{\Pi}$ is

$$H_0 : \mathbf{\Pi} = \boldsymbol{I}_k,$$

i.e. absence of transition between latent classes, so that a subject has always the same ability level

- Violations of this hypothesis typically arise when there exists an implicit learning phenomenon or certain items provide clues for responding to other items (Hambleton & Swaminathan, 1985)

- Since the LCR model is a particular case of the LMR model in which $\mathbf{\Pi} = \boldsymbol{I}_k$, testing H_0 is equivalent to testing the LCR model (in particular the LI assumption of this model) against the LMR model

A particular case

- Suppose that, under the larger model, the transition matrix depends only on one parameter, e.g.

$$\mathbf{\Pi} = \begin{pmatrix} 1-2\alpha & \alpha & \alpha \\ \alpha & 1-2\alpha & \alpha \\ \alpha & \alpha & 1-2\alpha \end{pmatrix}$$

- The asymptotic distribution of the LR statistic D for testing $H_0 : \mathbf{\Pi} = \boldsymbol{I}_k$ is simply

$$0.5\chi_0^2 + 0.5\chi_1^2$$

- A p-value for D may be explicitly computed as

$$\frac{1}{2}p(\chi_1^2 \geqslant d)$$

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 the 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 class; 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:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 0.000 & 0.040 & -0.704 & 1.013 & -1.560 & -0.043 \\ -0.705 & -1.250 & -0.387 & -0.587 & -2.532 & -2.587 \end{pmatrix}$$

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

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

Analysis of longitudinal data

- The LMR model may be also used to analyze longitudinal data collected by administering certain sets of items to the same subjects at different time occasions

- Here the interest is typically on the evolution of the latent characteristics (e.g. ability, quality of life, propensity to commits crimes) of the subjects involved in the study

- Within the LMR framework, these characteristics are represented through a latent process whose dynamics depends on the transition matrix $\mathbf{\Pi}$

- For the analysis of longitudinal data, we have to use a multivariate version of the LMR model (Langeheine & van de Pol, 2002)

Multivariate LMR model

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

$$\begin{aligned} p_t(\boldsymbol{x}|c) &= p(\boldsymbol{X}_t = \boldsymbol{x}|\Theta_t = \xi_c) = \prod_j \lambda_{tj|c}^{x_j}(1-\lambda_{tj|c})^{1-x_j} \\ \lambda_{tj|c} &= p(X_{tj} = 1|\Theta_t = \xi_c) = \frac{e^{\xi_c - \beta_{tj}}}{1 + e^{\xi_c - \beta_{tj}}} \end{aligned}$$

- The latent variables $\Theta_1, \ldots, \Theta_T$ are assumed to follow a first-order Markov chain (possibly non-homogeneous)
- The joint distribution of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ may be computed through recursions similar to those used for the univariate model

Likelihood inference

- The log-likelihood of the multivariate LMR model,

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

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

- One of the main interesting hypotheses to test is $H_0 : \boldsymbol{\Pi} = \boldsymbol{I}_k$, i.e. absence of educational progress during the longitudinal study

- The shape of the asymptotic distribution of D under hypotheses on $\boldsymbol{\Pi}$ is currently under investigation

An application to criminal data

- We are currently studying a dataset extracted from the *England and Wales Offenders Index*, a court based record of criminal histories of all the offenders born in England and Wales
- The dataset concerns a sample of $n = 11,402$ subjects ($9,232$ males and $2,170$ females) born in 1953, and followed through to 1993
- Offences are combined into $J = 10$ major groups described in the literature and $T = 6$ age bands (10-15, 16-20, 21-25, 26-30, 31-35, 36-40) have been considered
- Preliminary results have shown that the multivariate LMR model may describe adequately the phenomenon of interest, even though there is evidence of multidimensionality which should be taken in due accout

References

Bartolucci F. (2006), Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities, *JRSS-B, in press* (`http://www.stat.unipg.it/~bart`).

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

Bartolucci, F. Pennoni, F. & Francis, B. (2006), A latent Markov model for detecting patterns of criminal activity, *manuscript under review*

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

De Leeuw, J. & Verhelst, N. (1986), Maximum likelihood estimation in generalized Rasch models, *Journal of Educational Statistics*, **11**, 183-196.

Dempster A. P., Laird, N. M. & Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion), *JRSS-B*, **39**, 1-38.

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

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

Langeheine R. & van de Pol F. (2002), Latent Markov chains, *Applied latent Class Analysis*, J. A. Hagenaars & A. L. McCutcheon editors, Cambridge University Press, 304-344.

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

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

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

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.

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

Self, S. G. & 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.

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

Vermunt, J. K., Langeheine, R. and Böckenholt, 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.

Wiggins, L. M. (1973), *Panel Analysis: Latent Probability Models for Attitude and Behavior Process*, Elsevier, Amsterdam.

Subject of test items

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