

Modeling nonignorable missingness in multidimensional latent class IRT models

Silvia Bacci*¹, Francesco Bartolucci*, Bruno Bertaccini**

*Dipartimento di Economia, Finanza e Statistica - Università di Perugia

**Dipartimento di Statistica "G. Parenti" - Università di Firenze

Università La Sapienza, Roma, 20-22 June 2012

¹silvia.bacci@stat.unipg.it

Outline

- 1 Introduction
 - Motivation
- 2 Multidimensional LC IRT models
 - Preliminaries
 - The general formulation
 - Maximum log-likelihood estimation
- 3 Modeling nonignorable missingness
- 4 Application to Students' Entry Test
- 5 Conclusions
- 6 References

Introduction

- **Motivation:**

Measurement of **ability** in presence of a **penalty** factor for missing responses

- **Aim:**

We aim to measure the ability by modeling in a suitable way the **nonignorable missingness** due to the penalty factor

- **Method:**

We propose a semi-parametric approach based on the class of **Multidimensional Latent Class (LC) Item Response Theory (IRT) models**

Motivation

- In educational tests in order to avoid guessing, a wrong item response may often be penalized by a greater extent with respect to a missing response
- In this context missing responses are **not missing at random** (NMAR - Little and Rubin, 1987)
- We may model the nonignorable missingness by assuming that the observed item responses depend both on latent ability (or abilities) measured by the test and on **another latent variable** which is identified as the **propensity to answer**.

Problem: Is it possible to use standard IRT models?

Limits of standard IRT models

Main assumptions of standard IRT models

- Unidimensionality of latent traits: all the set of items contribute to measure **the same latent trait**
Therefore, nonignorable missingness cannot be treated as a specific latent trait
- Often, **normality** of latent trait is assumed

However, ...

- A same questionnaire is usually used to measure **several** latent traits
- We are interested in assessing and testing the **correlation between latent traits**
- Often, normality of latent trait is not a realistic assumption
- In some contexts (e.g., educational setting) can be useful to assume that population is composed by **homogeneous classes of individuals** with very similar latent characteristics (Lazarsfeld and Henry, 1968), so that individuals in the same class will receive the same kind of decision (e.g., admitted/not admitted)

Multidimensional LC IRT models

The class of multidimensional LC IRT models (Bartolucci, 2007; Von Davier, 2008) is characterized by these main features:

- More latent traits are simultaneously considered (**multidimensionality**)
- These latent traits are represented by a random vector with a **discrete distribution** common to all subjects (each support point of such a distribution identifies a different latent class of individuals)
- **Different item parameterisations** may be adopted for the probability of a given response to each item (e.g., Rasch and 2-PL for binary items; global logit or local logit for ordinal items with free or constrained item discrimination and difficulty parameters)

More in detail . . .

Basic notation:

- s : number of latent variables corresponding to the different traits measured by the items
- $\Theta = (\Theta_1, \dots, \Theta_s)$: vector of latent variables
- $\theta = (\theta_1, \dots, \theta_s)$: one of the possible realizations of Θ
- δ_{id} : dummy variable equal to 1 if item i measures latent trait of type d , $d = 1, \dots, s$
- k : number of latent classes of individuals

Assumptions

- Items are binary or ordinal polytomously-scored
- The set of items measures s **different latent traits**
- Each item measures only one latent trait
- The random vector Θ has a **discrete distribution** with support points $\{\xi_1, \dots, \xi_k\}$ and weights $\{\pi_1, \dots, \pi_k\}$
- The number k of latent classes is the same for each latent trait
- Manifest distribution of the full response vector $\mathbf{Y} = (Y_1, \dots, Y_k)'$:

$$p(\mathbf{Y} = \mathbf{y}) = \sum_{c=1}^C p(\mathbf{Y} = \mathbf{y} | \Theta = \xi_c) \pi_c$$

where $\pi_c = p(\Theta = \xi_c)$ and (assumption of *local independence*)

$$p(\mathbf{Y} = \mathbf{y} | \Theta = \xi_c) = \prod_{i=1}^I p(Y_i = y_i | \Theta = \xi_c)$$

Some examples

- Multidimensional LC 2PL model:

$$\log \frac{p(Y_i = 1|\boldsymbol{\theta})}{p(Y_i = 0|\boldsymbol{\theta})} = \lambda_i \left(\sum_{d=1}^s \delta_{id} \theta_d - \beta_i \right)$$

- Multidimensional LC GRM model:

$$\log \frac{p(Y_i \geq h|\boldsymbol{\theta})}{p(Y_i < h|\boldsymbol{\theta})} = \lambda_i \left(\sum_{d=1}^s \delta_{id} \theta_d - \beta_{ih} \right), \quad h = 1, \dots, H_i - 1$$

- Multidimensional LC GPCM model:

$$\log \frac{p(Y_i = h|\boldsymbol{\theta})}{p(Y_i = h - 1|\boldsymbol{\theta})} = \lambda_i \left(\sum_{d=1}^s \delta_{id} \theta_d - \beta_{ih} \right), \quad h = 1, \dots, H_i - 1$$

- Multidimensional LC RSM model:

$$\log \frac{p(Y_i = h|\boldsymbol{\theta})}{p(Y_i = h - 1|\boldsymbol{\theta})} = \sum_{d=1}^s \delta_{id} \theta_d - (\beta_i + \tau_h), \quad h = 1, \dots, H - 1$$

Maximum log-likelihood estimation

Let j denote a generic subject and let η the vector containing all the free parameters. The log-likelihood may be expressed as

$$\ell(\eta) = \sum_j \log(p(\mathbf{Y}_j = \mathbf{y}_j))$$

- Estimation of η may be obtained by the **discrete (or LC) MML approach** (Bartolucci, 2007)
- $\ell(\eta)$ may be efficiently maximize by the **EM algorithm** (Dempster et al., 1977)
- The software for the model estimation has been implemented in R
- Number of free parameters is given by:

$$\#par = (k - 1) + sk + \left[\sum_{i=1}^I (H_i - 1) - s \right] + a(r - s), \quad a = 0, 1,$$

where $a = 0$ when $\lambda_i = 1, \forall i = 1, \dots, I$, and $a = 1$ otherwise

Approaches to model nonignorable missingness

The class of Multidimensional LC IRT models may be used as a **semi-parametric approach** to treat with nonignorable missingness, as an alternative to:

- **Parametric approach** (Holman and Glas, 2005): multidimensional IRT models based on the multivariate Normality for the latent variables
 - Cons: intractability of multidimensional integral which characterizes the marginal log-likelihood function of a multidimensional IRT model based on Normality assumption
- **Non-parametric approach** (Bertoli-Barsotti and Punzo, 2012): multidimensional Rasch-type models (based on conditional maximum likelihood)
 - Cons: the use of this approach is limited to Rasch-type models and it does not allow the correlation between latent variables

The model

- Let $\Theta = (\Theta_1, \dots, \Theta_s)$ be the vector of latent variables, where Θ_1 denotes the propensity to answer and $\Theta_2, \dots, \Theta_s$ are the latent abilities measured by the test
- Let R_i be the binary variable equal to 1 if individual j provides a response to item i and to 0 otherwise, with $i = 1, \dots, I$
- Let Y_i^* denote the “true” binary response to item i that is observable only if $R_i = 1$, and in this case equal to the manifest binary variable Y_i , and unobservable if $R_i = 0$
- We require that the pairs of variables (R_i, Y_i^*) , $i = 1, \dots, I$, are conditionally independent given the latent variables in Θ

- In the following we assume that $p(R_i)$ depends only on Θ_1 , whereas $p(Y_i^*)$ depends only on the corresponding Θ_{d_i+1} ($d_i + 1 = 2, \dots, s$)
- We also assume that Θ_1 and Θ_{d_i+1} are correlated, so that Θ_{d_i+1} has an indirect effect on $p(R_i)$
- The magnitude of correlation between Θ_1 and Θ_{d_i+1} may be interpreted as an indication of the extent to which ignorability of missingness is violated: a correlation equal to 0 implicates that the missing data are Missing At Random
- We outline that other assumptions are theoretically possible (Holman and Glas, 2005), as follows:
 - $p(R_i)$ depends on both Θ_1 and Θ_{d_i+1} , whereas $p(Y_i^*)$ depends only on Θ_{d_i+1}
 - $p(R_i)$ depends only on Θ_1 , whereas $p(Y_i^*)$ depends on both Θ_1 and Θ_{d_i+1}
 - both $p(R_i)$ and $p(Y_i^*)$ depend on Θ_1 and Θ_{d_i+1}

The response process is described by two 2-PL models:

$$\log \frac{p(R_i = 1 | \Theta_1 = \theta_1)}{p(R_i = 0 | \Theta_1 = \theta_1)} = \lambda_i(\theta_1 - \beta_i) \quad (1)$$

$$\log \frac{p(Y_i^* = 1 | \Theta_{d_i+1} = \theta_{d_i+1}, R_j = 1)}{p(Y_i^* = 0 | \Theta_{d_i+1} = \theta_{d_i+1}, R_i = 1)} = \lambda_i^*(\theta_{d_i+1} - \beta_i^*) \quad (2)$$

Equations (1) and (2) define an **s-dimensional LC IRT model** having the following **manifest distribution**

$$p(\mathbf{r}_j, \mathbf{y}_j) = \sum_c \pi_c \prod_i p_i(\xi_{c1})^{r_{ji}} [1 - p_i(\xi_{c1})]^{1-r_{ji}} \times \\ \times \prod_{i:r_{ji}=1} p_i^*(\xi_{c,d_i+1})^{y_{ji}} [1 - p_i^*(\xi_{c,d_i+1})]^{1-y_{ji}}$$

where $\mathbf{r}_j = (r_{j1}, \dots, r_{jI})$, where r_{ji} is the generic value of R_i , and $\mathbf{y}_j = (y_{j1}, \dots, y_{jI})$, where $y_{ji} = 0, 1$ is the realization of Y_j^* when $r_{ji} = 1$ (the response is provided) and it is let equal to an arbitrary value otherwise.

Data

- Student's Entry Test for the admission to the Economics courses of the University of Florence (Italy)
- 1264 students
- three latent abilities: Logic (Θ_2 , 13 items), Mathematics (Θ_3 , 13 items), and Verbal Comprehension (Θ_4 , 10 items)
- all items are of multiple choice type, with one correct answer and four distractors, and they are polytomously scored, being 1 for correct response, -0.25 for wrong response and 0 for missing response
- the scoring system is communicated to the candidates before the test starting
- we estimate a constrained version of the proposed model, having $\lambda_i = \lambda_i^* = 1$

Choice of the number of latent classes

A crucial point with latent class models concerns the choice of the number k of components

- coherently with the main literature we suggest to use an information criterion, such as AIC or BIC indices
- the selected number of classes is the one corresponding to the minimum value of AIC or BIC
- The model is fitted for increasing values of k until AIC or BIC does not start to increase; then, the previous value of k is taken as the optimal one
- We outline that, in some practical situations, the number of latent classes is known or it is suggested by considerations of convenience
- In the context of the Students' Entry Test, we need to classify students in **at least $k = 3$ latent classes**, so as to discern among students that are: admitted, not admitted, and one or more groups of admitted with reserve

Main results

Estimated support points ($\hat{\xi}_c$), weights ($\hat{\pi}_c$), and average probabilities to answer given the class ($\bar{p}(\hat{\xi}_c)$) for $k = 3$ and $k = 4$

	$k = 3$			$k = 4$			
	$c = 1$	$c = 2$	$c = 3$	$c = 1$	$c = 2$	$c = 3$	$c = 4$
$\hat{\xi}_{c1}$	0.2845	0.3335	-0.8004	0.1564	0.1162	-0.8585	0.4495
$\hat{\xi}_{c2}$	1.1107	-1.1095	0.1743	1.6900	-1.9835	0.0707	-0.1881
$\hat{\xi}_{c3}$	1.0611	-0.7073	-0.3159	1.5907	-1.0928	-0.3217	-0.2498
$\hat{\xi}_{c4}$	0.6158	-1.3336	1.0796	1.3921	-1.9542	1.0163	-0.6772
$\hat{\pi}_c$	0.3381	0.3824	0.2795	0.2196	0.1614	0.2533	0.3657
$\bar{p}(\hat{\xi}_c)$	0.8298	0.8360	0.6484	0.8131	0.8074	0.6377	0.8507

Correlations

Correlations between item difficulties of Θ_1 and Θ_s , $s = 2, 3, 4$ ($\rho(\beta_{.1}, \beta_{.l})$) for $k = 3$ and $k = 4$

	$\rho(\beta_{.1}, \beta_{.2})$	$\rho(\beta_{.1}, \beta_{.3})$	$\rho(\beta_{.1}, \beta_{.4})$
$k = 3$	0.7270	0.4700	0.6092
$k = 4$	0.7384	0.4659	0.6169

Correlations between latent variables, for $k = 3$ (in red) and $k = 4$ (in blue)

	Θ_1	Θ_2	Θ_3	Θ_4
Θ_1	1.0000	-0.1559	0.2136	-0.6631
Θ_2	-0.0435	1.0000	0.9317	0.8427
Θ_3	0.1364	0.9432	1.0000	0.5896
Θ_4	-0.5113	0.8808	0.7478	1.0000

Conclusions

- We described a class of IRT models based on (i) the multidimensionality and (ii) the discreteness of latent traits, which allows to overcome the main drawbacks of standard IRT models
- We illustrated how the Multidimensional LC IRT models may be used to treat with nonignorable missingness
- The proposed approach was illustrated through an application to the educational setting in presence of penalty

What's next?

- Allowing for free discrimination parameters
- Extension to **latent regression**, by introducing covariates that explain the latent traits

$$\log \frac{p(R_i = 1 | \Theta_1 = \theta_1)}{p(R_i = 0 | \Theta_1 = \theta_1)} = \lambda_i \left(\sum_{h=1}^p \phi_{h1} Z_{hj} + \alpha_{c1} - \beta_i \right)$$

$$\log \frac{p(Y_i^* = 1 | \Theta_{d_i+1} = \theta_{d_i+1}, R_i = 1)}{p(Y_i^* = 0 | \Theta_{d_i+1} = \theta_{d_i+1}, R_i = 1)} = \lambda_i^* \left(\sum_{h=1}^p \phi_{h,d_i+1} Z_{hj} + \alpha_{c,d_i+1} - \beta_i^* \right)$$

- Z_1, \dots, Z_p are the observed covariates (e.g., type of high school)
- $\phi'_h = (\phi_{h1}, \dots, \phi_{hs})$ is the vector of regression coefficients of Z_h on the s -th latent trait
- $\alpha'_c = (\alpha_{c1}, \dots, \alpha_{cs})$ is the vector of residuals

Main references

Bartolucci F. (2007), A class of multidimensional IRT models for testing unidimensionality and clustering items, *Psychometrika*, 72, 141-157.

Bertoli-Barsotti L. and Punzo, A. (in press), Modelling missingness with a Rasch-type model, *Psicológica*.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

Holman, R. and Glas, A.W. (2005), Modelling non-ignorable missing-data mechanisms with item response theory models, *Brit. J. Math. Stat. Psy.*, **58**, 1 – 17.

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

Little, R.J.A. and Rubin, D.B. (1987), *Statistical analysis with missing data*, Boston: Wiley.

von Davier, M. (2008), A general diagnostic model applied to language testing data. *Brit. J. Math. Stat. Psy.*, 61(2), 287- 307.