Modeling nonignorable missingness in multidimensional latent class IRT models

Silvia Bacci, Francesco Bartolucci and Bruno Bertaccini

Abstract A relevant problem in applications of Item Response Theory (IRT) models is the presence of nonignorable missing responses. We propose a multidimensional latent class IRT model in which the missingness mechanism is driven by a latent variable (propensity to answer) correlated with the latent variable for the ability (or abilities) measured by the test items. These two latent variables are assumed to have a joint discrete distribution. This assumption is convenient both from the computational point of view and for the decisional process, since individuals are classified in homogeneous latent classes which may be associated to the same treatment. Moreover, this assumption avoids parametric formulations for the distribution of the latent variables, giving rise to a semiparametric model. The proposed approach is illustrated through an application to data coming from a Students' Entry Test for the admission to the courses in Economics in an Italian University.

Key words: EM algorithm, Semiparametric inference, Students' Entry Test.

1 Introduction

A relevant problem in applications of Item Response Theory (IRT) models is due to missing responses to some items. Indeed, ignorable missing responses do not rep-

Silvia Bacci
Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via A. Pascoli 20, 06123 Perugia, e-mail: silvia.bacci@stat.unipg.it

Francesco Bartolucci
Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via A. Pascoli 20, 06123 Perugia, e-mail: bart@stat.unipg.it

Bruno Bertaccini
Dipartimento di Statistica "G. Parenti", Università di Firenze, Viale Morgagni 59, 50134 Firenze, e-mail: bertaccini@ds.unifi.it

resent any particular problem, whereas nonignorable missing response need special attention to avoid wrong inferential conclusions [5]. A typical example of nonignorable missing responses is observed in the context of ability tests which, in order to avoid guessing, penalize a wrong item response by a greater extent with respect to a missing response.

The main literature [4] treats the problem of nonignorable missing responses by assuming that the observed item responses depend both on the latent ability (or abilities) intended to be measured by the test and on another latent variable which is identified as the propensity to answer. Usually, a parametric class of multidimensional IRT models is adopted, which is based on the multivariate Normal distribution for the two (or more) latent variables. Recently, an alternative nonparametric approach based on the conditional maximum likelihood estimation has been proposed by specifying the multidimensional IRT model according to the Rasch assumptions [2]. The main drawback is that the conditional approach does not allow us to measure the correlation between latent variables; moreover, its use is limited to data coherent with the Rasch paradigm.

Our aim is to propose the use of an alternative multidimensional IRT model based on the assumption of discreteness of the latent variables [1], so that the missing process may be modeled in a semiparametric way. Our proposal presents several advantages with respect to the parametric approach based on Normal latent variables. Firstly, it is more flexible because it does not introduce any restrictive assumption about the distribution of latent variables. Secondly, it allows to skip the well-known problem of the intractability of multidimensional integrals which characterizes the marginal log-likelihood function of a continuous multidimensional IRT model. Finally, detecting homogenous classes of individuals is convenient for certain decisional processes, because individuals in the same class may be associated to the same decision (e.g., admitted, admitted with reserve, not admitted).

The remainder of the paper is organized as follows. In Section 2 we describe the class of multidimensional latent class (LC) IRT models adopted to allow for noningorale missing responses. In Section 3 we illustrate the proposed class of models through an application to data arising from the Students' Entry Test for the admission to the Economics courses of the University of Florence (Italy).

2 The proposed model

Given a set of J binary items for the measurement of s distinct student's abilities, let $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_{s+1})'$ be the vector of latent variables that drives the response process and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{s+1})'$ denote one of its possible realizations. In particular, Θ_1 is the latent variable for the propensity to answer and $\Theta_2, \ldots, \Theta_{s+1}$ are those for the student's abilities measured by the test. The random vector $\boldsymbol{\Theta}$ is assumed to have a discrete distribution with k support points, denoted by $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k$, and probabilities $\pi_1, \ldots, \pi_k$, with $\pi_c = p(\boldsymbol{\Theta} = \boldsymbol{\xi}_c)$, $c = 1, \ldots, k$.

In order to model the response process, for a generic examinee we denote by R_j the binary variable equal to 1 if this examinee provides a response to item j and to 0 otherwise, with $j = 1,\dots,J$. Moreover, we introduce the symbol X_j^* to denote the "true" binary response to item j that is observable only if $R_j = 1$, and in this case equal to the manifest binary variable X_j, and unobservable if $R_j = 0$. Then, we formulate a local-independence assumption by requiring that the pairs of variables (R_j, X_j^*), $j = 1,\dots,J$, are conditionally independent given the latent vector $\boldsymbol{\Theta}$. Moreover, we assume that R_j and X_j^* are conditionally independent given Θ_1 and Θ_{d_j+1}, where d_j indicates the ability measured by j-th item. Finally, for the conditional distribution of every R_j and X_j^*, given the corresponding latent variables, we assume the following two-parameter logistic parametrization. Let $p_j(\theta_1) = p(R_j = 1|\Theta_1 = \theta_1)$ and $p_j^*(\theta_{d_j+1}) = p(X_j^* = 1|\Theta_{d_j+1} = \theta_{d_j+1})$. We have that:

$$\log\frac{p_j(\theta_1)}{1-p_j(\theta_1)} = \gamma_j(\theta_1 - \beta_j), \tag{1}$$

$$\log\frac{p_j^*(\theta_{d_j+1})}{1-p_j^*(\theta_{d_j+1})} = \gamma_j^*(\theta_{d_j+1} - \beta_j^*), \tag{2}$$

where γ_j and γ_j^* are discrimination parameters and β_j and β_j^* are difficulty parameters. Equations (1) and (2) define an $(s+1)$-dimensional latent class IRT model as described in [1].

In order to estimate the model on the basis of the observed responses provided by a sample of n examinees, we need to obtain the manifest distribution of these data. For every subject i we observe the vector $\mathbf{r}_i = (r_{i1},\dots,r_{iJ})$, where r_{ij} is the value of R_j, and $\mathbf{x}_i = (x_{i1},\dots,x_{iJ})$, where $x_{ij} = 0,1$ is the realization of X_j^* when $r_{ij} = 1$ (the response is provided) and it is let equal to an arbitrary value otherwise. The above assumptions implies that the manifest distribution may be expressed as

$$\begin{aligned} p(\mathbf{r}_i,\mathbf{x}_i) = \sum_c \pi_c \prod_j p_j(\xi_{c1})^{r_{ij}}[1-p_j(\xi_{c1})]^{1-r_{ij}} \times \\ \times \prod_{j:r_{ij}=1} p_j^*(\xi_{c,d_j+1})^{x_{ij}}[1-p_j^*(\xi_{c,d_j+1})]^{1-x_{ij}}. \end{aligned}$$

In order to estimate the vector $\boldsymbol{\eta}$ of the model parameters, the log-likelihood $\ell(\boldsymbol{\eta}) = \sum_i \log p(\mathbf{r}_i,\mathbf{x}_i)$ is maximized through the EM algorithm [3].

3 Application to Students' Entry Test: main results

A constrained version of the proposed model, with $\gamma_j = \gamma_j^* = 1$, $j = 1,\dots,J$, was applied for the analysis of data arising from the Student's Entry Test to the courses in Economics administered to 1264 students in September 2011 at the University of Florence. The test measures three latent variables: 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 a correct response, -0.25 for a wrong response, and 0 for a missing response. The scoring system is communicated to the candidates before the test starting.

The estimated support points $\hat{\boldsymbol{\xi}}_c$ (centered at $\mathbf{0}$) and probabilities $\hat{\pi}_c$ are shown in Table 1 for the four-dimensional LC IRT model with $k = 3$ and $k = 4$ classes.

Table 1 Estimated support points ($\hat{\boldsymbol{\xi}}_c$), weights ($\hat{\pi}_c$), and average probabilities to answer given the class ($\bar{p}(\hat{\boldsymbol{\xi}}_c)$) 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
$\bar{p}(\hat{\boldsymbol{\xi}}_c)$	0.8298	0.8360	0.6484	0.8131	0.8074	0.6377	0.8507

With reference to the model with $k = 3$ classes, we observe that class 1 includes students with the highest ability in Logic and Mathematics and a good level of Verbal Comprehension (33.81%), whereas students in class 2 present the worst levels for all the three abilities (38.24%). Moreover, class 3 collects students with the highest performance in Verbal Comprehension, but with some deficiencies in Mathematics (27.95%). Some differences turn up with $k = 4$, being identified a further class (class 4) that represents students with scores under the average for all the three abilities (36.57%). Finally, the propensity to answer ($\hat{\xi}_{c1}$) is quite high for all classes except than for class 3 (both for $k = 3$ and $k = 4$). In fact, for these classes the average probability to answer, $\bar{p}(\hat{\boldsymbol{\xi}}_c)$, is greater than 80%.

As concerns further developments of the proposed approach, we intend to extend the model so as to allow the latent class weights (and then the ability levels) to depend on individual covariates (e.g., type of high school diploma).

References

1. Bartolucci, F.: A class of multidimensional IRT models for testing unidimensionality and clustering items. Psychometrika, **72**, 141–157 (2007)
2. Bertoli-Barsotti, L., Punzo, A.: Modelling missingness with a Rasch-type model. Psicológica, (in press)
3. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Stat. Soc. B Met., **39**, 1–38 (1977)
4. Holman, R., Glas, A.W.: Modelling non-ignorable missing-data mechanisms with item response theory models. Brit. J. Math. Stat. Psy., **58**, 1 – 17 (2005)
5. Little, R.J.A., Rubin, D.B.: Statistical analysis with missing data, Boston: Wiley (1987)