

A multidimensional finite mixture SEM for non-ignorable missing responses to test items

Modelli mistura ad equazioni strutturali per il trattamento di dati mancanti non casualmente

Silvia Bacci and Francesco Bartolucci

Abstract A structural equation model is proposed to deal with dichotomously-scored items in the presence of missing not at random responses. Two types of latent variables are introduced: one refers to the abilities measured by the questionnaire and the other describes the propensity to answer. A semi-parametric approach is adopted, where both types of latent variable are discretely distributed. In such a way, the proposed structural equation model reduces to a multidimensional latent class item response theory model. Individual covariates may also be included in order to explain the probabilities of belonging to each latent class. For this aim, a multinomial logistic parametrization is introduced. A simulation study is then performed to evaluate the finite-sample properties of the parameter estimates.

Abstract *Per spiegare la presenza di risposte mancanti non casualmente a domande di tipo binario si propone un modello ad equazioni strutturali, caratterizzato da due tipi di variabili latenti: un primo tipo corrisponde alle abilità misurate dalle domande del questionario, mentre l'altro tipo viene introdotto per descrivere la propensione dell'individuo a non rispondere. Si assume che entrambi i tipi di variabile latente abbiano una distribuzione discreta, così da adottare un approccio semi-parametrico. In tal modo, il modello proposto si riduce ad un modello per item di risposta di tipo multidimensionale e a classi latenti. Inoltre, si prevede la possibilità di far dipendere le probabilità di appartenenza alle classi latenti da covariate individuali, tramite una parametrizzazione logistica multinomiale. Il metodo proposto viene valutato attraverso uno studio di simulazione.*

Key words: Finite mixture models, Item response theory, Latent class models, Semiparametric inference

Silvia Bacci

Department of Economics, Via A. Pascoli 20, 06123 Perugia (IT), e-mail: silvia.bacci@stat.unipg.it

Francesco Bartolucci

Department of Economics, Via A. Pascoli 20, 06123 Perugia (IT) e-mail: bart@stat.unipg.it

1 Introduction

A relevant problem in applications of Item Response Theory (IRT) models is related to non-ignorable missing responses (or missing not at random, MNAR; Little and Rubin, 2002) to certain items. A typical case of MNAR responses is observed with educational tests where, in order to avoid guessing, a wrong item response is penalized to a greater extent in comparison with a missing response. In such a context, it is natural to suppose that the choice of not answering to a given item is related to the ability (or abilities) measured by the test. To avoid wrong inferential conclusions and loss of relevant information for the assessment of the latent trait level, the missingness mechanism must be modeled in a suitable way.

In the statistical literature, different approaches exist to model a non-ignorable missing mechanism. Here we rely on the interesting contribution by Holman and Glas (1995), who, elaborating an original idea of Lord (1983), suggested to treat the problem of MNAR responses by assuming that the observed item responses depend on the latent ability (or abilities) intended to be measured by the test and another latent variable which represents the propensity to answer by the individuals. The result consists in a multidimensional IRT model (Reckase, 2010), which assumes the normality of the latent traits. Along the same guidelines, Bertoli-Barsotti and Punzo (2013) proposed a non-parametric approach based on the Rasch's paradigm (Rasch, 1960).

In this contribution, we introduce a Structural Equation Model (SEM, Duncan, 1975) characterized by: (i) at least two latent traits, similarly to what above discussed, (ii) discreteness of these latent traits, so that homogenous unobservable classes of individuals are detected, and (iii) presence of individual covariates which explain the probability of belonging to a given latent class. In such a way, a semi-parametric alternative to the proposals of Holman and Glas (1995) and Bertoli-Barsotti and Punzo (2013) is formulated in the setting of finite mixture SEMs (Dolan and van der Maas, 1998), which reduces to a multidimensional latent class IRT model (Bartolucci, 2007; Bartolucci et al., 2014).

The proposed model may be estimated through the discrete marginal maximum likelihood method, based on an Expectation-Maximization (EM) algorithm (Dempster et al., 1977). In order to assess the finite-sample properties of the parameter estimates obtained from the estimation process, we performed a simulation study under different scenarios corresponding to different structures of missing data.

In the following we first describe the model assumptions (Section 2) and then the results of the simulation study (Section 3).

2 The model

For a random subject drawn from the population of interest, let Y_j denote the response provided by the subject to binary item j , with $j = 1, \dots, m$. In order to model the response process, we have to consider that the subject may answer correctly

($Y_j = 1$) or incorrectly ($Y_j = 0$) or he/she may skip the question, so that Y_j can be observed or not. Therefore, for $j = 1, \dots, m$, we also introduce the binary indicator R_j equal to 1 if the individual provides a response to item j and to 0 otherwise (i.e., Y_j is missing). Moreover, we consider a set of c exogenous individual covariates denoted by $X_1, \dots, X_c$.

In order to explain the association between the exogenous variables $X_1, \dots, X_c$ and the endogenous variables $Y_1, \dots, Y_m$, we introduce $s + 1$ latent variables. The first s latent variables, denoted by $U_1, \dots, U_s$, represent the latent traits that are measured by the test items. The remaining latent variable, denoted by V , is interpreted as the propensity to answer. The latent variables $U_1, \dots, U_s$ and V are assumed conditionally independent given the covariates $X_1, \dots, X_c$. For an illustration of the model see Figure 1.

Considering Y_j and R_j as deriving from a discretization of continuous variables denoted by Y_j^* and R_j^* , we formulate the following equations entering the measurement component of the proposed SEM:

$$Y_j = I\{Y_j^* \geq 0\}, \quad (1)$$

$$R_j = I\{R_j^* \geq 0\}, \quad (2)$$

$$Y_j^* = \alpha_j \sum_{d=1}^s z_{dj} U_d - \beta_j + \varepsilon_{1j}, \quad (3)$$

$$R_j^* = \gamma_j \sum_{d=1}^s z_{dj} U_d + \gamma_{2j} V - \delta_j + \varepsilon_{2j}, \quad (4)$$

for $j = 1, \dots, m$, where $I\{\cdot\}$ is the indicator function equal to 1 if its argument is true and to 0 otherwise and ε_{1j} and ε_{2j} are independent error terms. Moreover, z_{dj} is a dummy assuming value 1 if item j measures latent trait of type d and 0 otherwise.

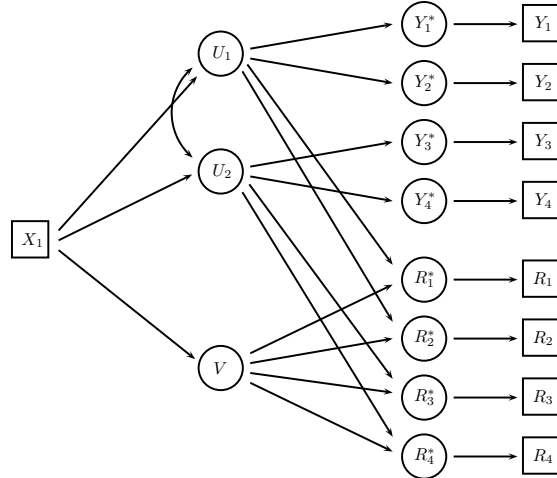


Fig. 1 Proposed SEM for only one covariate ($c = 1$), four items ($m = 4$), and two dimensions ($s = 2$), with the first two items measuring the first dimension and the other two items measuring the second dimension.

Parameters β_j and δ_j are interpreted as difficulty parameters because higher values of them correspond to smaller values of Y_j^* and R_j^* . The slope α_j measures the effect of the latent variable U_{d_j} , where d_j is the dimension measured by item j , on Y_j^* and, similarly, γ_j and γ_{2j} measure the effect on R_j^* of U_{d_j} and V , respectively.

According to the proposed model, the observed response to a given item j depends only on the ability U_{d_j} measured by the test, whereas the event of answering to item j depends both on U_{d_j} and on the propensity to respond V . The idea behind this assumption is that better students are more willing to respond due to their confidence on the correctness of the response.

We also observe that equations (1)-(4) rule out a direct effect of the covariates $X_1, \dots, X_c$ on the response variables. Indeed, covariates enter in the conditional distributions of $U_1, \dots, U_s$ and V , as it is clarified in the following.

We assume that the vector of latent variables $(U_1, \dots, U_s)'$ and V are independent, given the covariates $X_1, \dots, X_c$, and have a discrete distribution, with k_1 support points $(u_{1h_1}, \dots, u_{sh_1})'$, $h_1 = 1, \dots, k_1$, and k_2 support points v_{h_2} , $h_2 = 1, \dots, k_2$, respectively. The corresponding weights are denoted by

$$\begin{aligned}\lambda_{h_1}(\mathbf{x}) &= p(U_1 = u_{1h_1}, \dots, U_s = u_{sh_1} | X_1 = x_1, \dots, X_c = x_c), \quad h_1 = 1, \dots, k_1, \\ \pi_{h_2}(\mathbf{x}) &= p(V = v_{h_2} | X_1 = x_1, \dots, X_c = x_c), \quad h_2 = 1, \dots, k_2,\end{aligned}$$

which depend on the observed individual covariates $\mathbf{x} = (x_1, \dots, x_c)'$. In particular, we assume the following multinomial logistic parametrization:

$$\log \frac{\lambda_{h_1+1}(\mathbf{x})}{\lambda_1(\mathbf{x})} = \phi_{0h_1} + \mathbf{x}' \boldsymbol{\phi}_{1h_1}, \quad h_1 = 1, \dots, k_1 - 1,$$

where ϕ_{0h_1} and $\boldsymbol{\phi}_{1h_1}$ are regression parameters to be estimated. A similar parametrization is assumed for the conditional probabilities $\pi_{h_2}(\mathbf{x})$, $h_2 = 1, \dots, k_2 - 1$.

Regarding the error terms ε_{1j} and ε_{2j} , we assume that they are independent and have a standard logistic distribution. This is a convenient assumption implying that the distribution of Y_j and R_j given the corresponding latent variables satisfies a logistic model. In particular, for $j = 1, \dots, m$ and with

$$\begin{aligned}p_{h_1j} &= p(Y_j = 1 | U_1 = u_{1h_1}, \dots, U_s = u_{sh_1}), \quad h_1 = 1, \dots, k_1 \\ q_{h_1h_2j} &= p(R_j = 1 | U_1 = u_{1h_1}, \dots, U_s = u_{sh_1}, V = v_{h_2}), \quad h_1 = 1, \dots, k_1, h_2 = 1, \dots, k_2,\end{aligned}$$

we have the following 2PL parametrization (Birnbbaum, 1968):

$$\begin{aligned}\log \frac{p_{h_1j}}{1 - p_{h_1j}} &= \alpha_j \sum_{d=1}^s z_{dj} u_{dh_1} - \beta_j, \\ \log \frac{q_{h_1h_2j}}{1 - q_{h_1h_2j}} &= \gamma_{1j} \sum_{d=1}^s z_{dj} u_{dh_1} + \gamma_{2j} v_{h_2} - \delta_j.\end{aligned}$$

The above assumptions imply that the proposed SEM reduces to a multidimensional latent class IRT model similar to the model proposed by Bartolucci (2007)

for the augmented set of response variables $Y_1, \dots, Y_m, R_1, \dots, R_m$. In particular, this model is characterized by the following features: (i) certain response variable Y_j may be missing at random (given also $R_1, \dots, R_m$), whereas the variables R_j are always observed; (ii) every Y_j depends on a specific latent variable U_{d_j} through a 2PL parametrization, whereas R_j depends on both U_{d_j} and another latent variable V ; (iii) the latent variables $U_1, \dots, U_s$ and V have a discrete distribution depending on the individual covariates $X_1, \dots, X_c$ through a multinomial logistic model.

The parameters of the proposed model may be estimated through the discrete marginal maximum likelihood approach, making use of the EM algorithm (Dempster et al., 1977) and suitable iterative algorithms of Newton-Raphson type nested in the EM algorithm; for details, see Bartolucci and Forcina (2006) and Bartolucci (2007).

3 Simulation study

We implemented a Monte Carlo simulation experiment in the R software in order to evaluate empirically the adequacy of the proposed approach in obtaining reliable estimates of the model parameters and in showing the presence of a non-ignorable missing mechanism for the item responses.

The simulation study is based on a model with two latent abilities, U_1 and U_2 , which drive the observed responses, and another latent construct, V , which corresponds to the tendency to respond. The distribution of these latent variables is based on $k_1 = k_2 = 3$ latent classes with probabilities affected by two individual covariates, X_1 and X_2 , generated from independent standard normal distributions. For the sample size we considered $n = 1000, 2000$ and for the number of items $m = 20, 40$. The items are equally distributed between U_1 and U_2 . Moreover, three types of scenario are assumed: (i) complete cases (with no missing responses), (ii) presence of missing responses depending on V , and (iii) presence of missing responses depending on V and on U_1 and U_2 . In this way, we considered a total of 12 different scenarios, under each of which 1000 samples were simulated. All the proposed scenarios assume fixed values of the population parameters, chosen in a suitable way.

A bidimensional model with $k_1 = k_2 = 3$ latent classes and a 2PL parametrization was estimated for every simulated sample generated under each of the 12 scenarios. To evaluate the goodness of the proposed model, we considered the values of the bias and of the root-mean-square-error (RMSE) for each population parameter, that is, for the support points, the regression coefficients, and the item parameters.

We observed generally satisfactory values of both bias and RMSE. We also observed that bias was approximately constant, regardless of the type of scenario. On the other hand, some differences emerged relatively to RMSE. As may be expected, the RMSE decreases when the number of items m and the sample size n increased. RMSE values also depends on the changes in the missingness structure, according to a different trend for each population parameter.

Finally, the discriminant parameters γ_j are well estimated both when the missingness process depended only on the latent trait V and when it depended also on the abilities U_d , $d = 1, 2$. These results imply that the proposed model can be used to obtain evidence about the effect of the latent abilities on the missing responses and to properly conclude about the ignorability of the missingness process. In practice, discriminant parameter estimates $\hat{\gamma}_{1j}$ not significantly different from zero for all items denote that the presence of missing responses does not depend on the abilities, whereas values of $\hat{\gamma}_{1j}$ significantly different from 0 (for at least some items) denote that the missing data provides information about the abilities.

References

- [1] Bartolucci, F.: A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, **72**, 141-157 (2007)
- [2] Bartolucci, F., Bacci, S., Gnaldi, M.: MultiLCIRT: An R package for multidimensional latent class item response models. *Comput. Statistics and Data Anal.*, **71**, 971-985 (2014)
- [3] Bartolucci, F., Forcina, A.: A class of latent marginal models for capture-recapture data with continuous covariates. *J. Am. Statistical Association*, **101**, 786-794 (2006)
- [4] Bertoli-Barsotti, L., Punzo, A.: Modelling missingness with a Rasch-type model. *Psicológica*, **34**, 97-123 (2013)
- [5] Birnbaum, A.: Some latent trait models and their use in inferring an examinees ability. In: Lord, F. M., Novick, M. R. (eds.) *Statistical Theories of Mental Test Scores*, pp. 395-479. Addison-Wesley, Reading (1968)
- [6] 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)
- [7] Dolan, C. V., van der Maas, H. L. J.: Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika*, **63**, 227-253 (1998)
- [8] Duncan, O. D.: *Introduction to structural equation models*. Academic Press, New York (1975)
- [9] 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)
- [10] Little, R.J.A., Rubin, D.B.: *Statistical analysis with missing data*. Wiley, Boston (2002)
- [11] Lord, F. M.: Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, **48**, 477-482 (1983)
- [12] Rasch, G.: *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen (1960)
- [13] Reckase, M. D.: *Multidimensional item response theory*. Springer, New York (2010)