A nonparametric multidimensional latent class IRT model in a Bayesian framework

Francesco Bartolucci, Alessio Farcomeni and Luisa Scaccia

Abstract We propose a nonparametric Item Response Theory model for dichotomously scored items in a Bayesian framework. Partitions of the items are defined on the basis of inequality constraints among the latent class success probabilities. A Reversible Jump type algorithm is described for sampling from the posterior distribution. A consequence is the possibility to make inference on the number of dimensions (i.e., number of groups of items measuring the same latent trait) and to cluster items when unidimensionality is violated.

Key words: Item response theory, unidimensionality, stochastic partition.

1 Introduction

Educational and psychological tests are often based on a set of items which measure a *unidimensional* latent trait, that is, a single personal aspect which is not directly observable (e.g., ability in a certain subject, tendency toward a certain behavior). When the test is unidimensional, the responses to the items may be validly summarized by a single indicator (e.g., the sum of the correct responses at individual level) and respondents may be globally ranked according to such an indicator and the distance between any two respondents in terms of the single latent trait may be

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

Alessio Farcomeni
Dipartimento di Sanitá Pubblica e Malattie Infettive, Sapienza - Università di Roma, Piazzale Aldo Moro, 5, 00186 Roma, Italy, e-mail: alessio.farcomeni@uniroma1.it

Luisa Scaccia
Dipartimento di Economia e Diritto, Università di Macerata, Via Crescimbeni 20, 62100 Macerata, Italy, e-mail: scaccia@unimc.it

simply measured. A consequent important aspect is how to test the unidimensionality assumption and, in case it is violated, how to group items in a sensible way so that items in the same group measure the same latent trait. Bartolucci (2007) introduced a multidimensional parametric Item Response Theory (IRT) model for dichotomously-scored items, which is based on the assumption that respondents are grouped into k latent classes of ability, and found the number of dimensions, s, and clusters of items through a hierarchical agglomerative clustering algorithm based on the model likelihood. However, this approach is based on certain parametric assumptions which may affect the selected number of dimensions.

In this work, we propose to select s relying on a completely nonparametric model formulated along the lines of Forcina and Bartolucci (2004). This formulation is based on a set of inequalities on the conditional probabilities of success in each item given the level of the ability. The distribution of the ability is still assumed to be discrete, therefore having k latent classes. Consequently, two items measure the same dimension if their success probabilities have the same ordering with respect to the latent classes. Any specific model depends on the number of latent classes and the set of inequalities on success probabilities, which, in turn, determines a certain partition of the items into s groups. Inference on the nonparametric IRT models proposed is based on the Bayesian paradigm, allowing us to work with unknown k and s. Relying on the *encompassing approach* of Klugkist et al (2005), we formulate the priors on the parameters of a model that includes any other model of interest. See also Bartolucci et al (2012). Such *encompassing model* is the latent class model (Lazarsfeld and Henry, 1968) with k classes. This automatically defines the priors on any nested model. For estimation purposes, we use the Reversible Jump (RJ) algorithm (Green, 1995; Green and Richardson, 2001) applied to the latent class model. The output is then suitably post-processed to estimate the posterior probability of any nonparametric IRT model. An alternative algorithm, expected to be more efficient, is also outlined.

The paper is organized as follows. Section 2 formalizes the nonparametric IRT model and deals with Bayesian estimation. Section 3 illustrates the approach through an application on the Mathematics test data used in Bartolucci (2007).

2 Model Formulation and Bayesian Inference

Let Y_{ij}, $i = 1,\ldots,n$, $j = 1,\ldots,r$ denote the binary outcome measured on the i-th subject for the j-th item. We assume that the sample of respondents is drawn from a population divided into k latent classes, with individuals in the same class sharing the same ability level. Thus the ability is represented by a discrete latent variable $C =$ having k support points denoted, without loss of generality, by $1,\ldots,k$. Let $\pi_1,\ldots,\pi_k$ be the class weights and $\lambda_{cj} = p(Y_{ij} = 1|C = c)$ denote the probability of success at the j-th item for any subject i in class c. Given two items, j_1 and j_2 say, these are said to measure the same dimension if there exists a permutation of $1,\ldots,k$, denoted by $c_1,\ldots,c_k$, such that

$$\lambda_{c_1 j} \leq \ldots \leq \lambda_{c_k j}, \quad j = j_1, j_2. \tag{1}$$

In other words, the success probabilities of the two items are ordered in the same way. Such a characterization of items measuring the same dimension is completely nonparametric, in contrast with the one in Bartolucci (2007) which is based on a parametric formulation of λ_{cj}. For the full set of items, the nonparametric IRT model is specified by fixing k and a certain permutation $c_1^{(j)}, \ldots, c_k^{(j)}$, of the type (1), for every item $j = 1, \ldots, r$. If there are s different permutations, there are s groups of items measuring distinct dimension, which are are denoted by $\mathscr{J}_1, \ldots, \mathscr{J}_s$, collected in $\mathscr{J}$.

The *observed log-likelihood* of the model defined above may be easily computed as

$$\ell(\boldsymbol{\Lambda}, \boldsymbol{\pi}) = \sum_i \log \left[\sum_c \pi_c \prod_j \lambda_{cj}^{y_{ij}} (1 - \lambda_{cj})^{1 - y_{ij}} \right], \tag{2}$$

where $\boldsymbol{\Lambda}$ is the $k \times r$ dimensional matrix of probabilities λ_{cj}, $\boldsymbol{\pi}$ is the vector of class weights π_c, and y_{ij} is the observed value of Y_{ij}. To make estimation easier it is convenient to introduce the latent class indicators z_{ic}, $i = 1, \ldots, n$, $c = 1, \ldots,$, where $z_{ic} = 1$ if the i-th subject is in latent class c; see for instance Diebolt and Robert (1994). The *complete* or *augmented* data log-likelihood, after augmenting the data with z_{ic}, is then

$$\ell_c(\boldsymbol{\Lambda}, \boldsymbol{\pi}) = \sum_c z_{ic} \log(\pi_c) + \sum_c \sum_i \sum_j z_{ic} [y_{ij} \log(\lambda_{cj}) + (1 - y_{ij}) \log(1 - \lambda_{cj})]. \tag{3}$$

2.1 Prior Distributions

Any model of the type above is nested in a latent class model in which the probabilities λ_{cj} are left unconstrained (Lazarsfeld and Henry, 1968). Then, once the priors have been specified for this model, we can automatically specify those of any nested model by the encompassing approach (Klugkist et al, 2005): prior distributions for nested models are automatically derived by truncating the parameter space according to the constraints of interest.

For the encompassing model we adopt Bayes-Laplace priors for the success probabilities and class weights (Tuyl et al, 2009). This choice reduces to an (unconditional) uniform prior for λ_{cj}, $c = 1, \ldots, k$. For the class weights this choice corresponds to a Dirichlet distribution with vector of parameters having all elements equal to 1. Finally, we use a uniform prior for k in the discrete set $1, \ldots, k_{\max}$.

2.2 Estimation strategy based on the Reversible Jump algorithm

Our estimation strategy makes use of the RJ algorithm, which samples from the posterior distribution of all the parameters of the latent class model, including k. The RJ output is then post-processed for identifiability (Frühwirth-Schnatter, 2001) and to deliver all the different partitions of items visited by the algorithm.

The algorithm performs the following steps:

1. Sample indicators of latent class z_{ic} from their full conditional distribution:

$$\Pr(z_{ic}=1|Y,\lambda,\pi)=\frac{\pi_c\prod_j\lambda_{cj}^{y_{ij}}(1-\lambda_{cj})^{1-y_{ij}}}{\sum_h\pi_h\prod_j\lambda_{hj}^{y_{ij}}(1-\lambda_{hj})^{1-y_{ij}}}.$$

2. Update λ_{cj}. For each $j=1,\ldots,r$, we propose simultaneous independent zero-centered normal increments of the current logit ($\boldsymbol{\lambda}_j$), where $\boldsymbol{\lambda}_j=(\lambda_{1j},\ldots,\lambda_{kj})$. The candidate $\boldsymbol{\lambda}_j^\star$ is accepted with probability equal to $\min(1,p_{\boldsymbol{\lambda}_j^\star})$, where

$$\begin{aligned}\log(p_{\boldsymbol{\lambda}_j^\star})=&\sum_c\sum_i z_{ic}\{y_{ij}\log(\lambda_{cj}^\star/\lambda_{cj})+(1-y_{ij})\log[(1-\lambda_{cj}^\star)/(1-\lambda_{cj})]\}+\\&+\sum_c\left[\log(\lambda_{cj}^\star)+\log(1-\lambda_{cj}^\star)-\log(\lambda_{cj})-\log(1-\lambda_{cj})\right].\end{aligned}\tag{4}$$

   The first line on the right side is the log-likelihood ratio. The ratio between the prior densities cancels out when using uniform priors for λ_{cj}, as suggested. Also the ratio between the proposal densities cancels out, apart from logarithm of Jacobian of the logit transformation, given in the second line of 4.
3. Sample the weights $\pi_1,\ldots,\pi_k$ from the full conditional distribution, which is a Dirichlet with parameters $(1+\sum_i z_{i1},\ldots,1+\sum_i z_{ik})$.
4. Update k. We follow the approach consisting on a random choice between splitting an existing latent class into two and merging two existing classes into one. The probabilities of these alternatives are b_k and $1-b_k$, respectively. Of course $b_1=1$ and $b_{k_{\max}}=0$, and otherwise we choose $b_k=0.5$ for $k=2,\ldots,k_{\max}-1$. For the combine proposal we randomly choose a pair of classes (c_1,c_2), with $\pi_{c_1}<\pi_{c_2}$, not necessarily adjacent in terms of the current value of their weights. These two classes are merged into a new one, labeled $c^\star=c_2-1$, reducing k by 1. We then reallocate all those observations y_{ij}, $j=1,\ldots,r$, with $z_{ic_1}=1$ and $z_{ic_2}=1$ to the new class $c^\star$ and create values for $\lambda_{c^\star j}$ and $\pi_{c^\star}$ in such a way that:

$$\lambda_{c^\star j}=\lambda_{c_2 j}\qquad\text{and}\qquad\pi_{c^\star}=\pi_{c_1}+\pi_{c_2}.$$

   In the split proposal, a class $c^\star$ is chosen at random and split into two new ones labeled c_1 and c_2, augmenting k by 1. The place assigned to the class c_1 is randomly chosen between 1 and $c^\star$, while the class c_2 takes the place $c^\star+1$. Values for $\pi_{c_1},\pi_{c_2},\lambda_{c_1 j},\lambda_{c_2 j}$, for $j=1,\ldots,r$, are created by generating a scalar u_1 and a vector $u_2=(u_{2j})_{j=1}^r$, respectively as $u_1\sim U[0;0.5]$ and $u_{2j}\sim U[0;1]$ and setting:

$$\pi_{c_1} = u_1 \pi_{c^\star} \quad , \quad \pi_{c_2} = (1-u_1)\pi_{c^\star},$$
$$\lambda_{c_1 j} = u_{2j} \text{ and } \lambda_{c_2 j} = \lambda_{c^\star j} \qquad \text{for } j = 1,\ldots,r. \tag{5}$$

Finally we reallocate all those observations y_{ij}, $j = 1,\ldots,r$, with $z_{ic^\star} = 1$ between the two new classes, in a way analogous to the standard Gibbs allocation move, used in step 1. We accept the split move with probability $\min(1, p_k)$, where

$$p_k = (\text{likelihood ratio}) \times \frac{Pr(k+1)}{Pr(k)} \times \frac{\mathscr{D}(\pi_1,\ldots,\pi_{k+1})}{\mathscr{D}(\pi_1^\star,\ldots,\pi_k^\star)}$$
$$\times \frac{(\pi_{c_1})^{\sum_i z_{ic_1}} (\pi_{c_2})^{\sum_i z_{ic_2}}}{(\pi_{c^\star}^\star)^{\sum_i z_{ic^\star}^\star}} \times \frac{2(1-b_{k+1})}{b_k P_{\text{alloc}}} \times \pi_{c_2-1}^\star, \tag{6}$$

where P_{alloc} is the probability of this particular allocation and $\mathscr{D}$ is the Dirichlet density with all parameters equal to 1. The first four terms in the product are the ratio of the likelihood and the priors for the new parameter set to those for the old one. The fifth term is the proposal ratio. The last term is the Jacobian of the transformation from $(\pi_{c^\star}, \lambda_{c^\star 1},\ldots,\lambda_{c^\star r}, u_1, u_{21},\ldots,u_{2r})$ to $(\pi_{c_1}, \lambda_{c_1 1},\ldots,\lambda_{c_1 r}, \pi_{c_2}, \lambda_{c_2 1},\ldots,\lambda_{c_2 r})$. The combine move is accepted with probability $\min(1, p_k^{-1})$, with some obvious substitutions in the expression for p_k.

From the RJ output, we estimate the posterior probability of any nonparametric IRT model visited at least once and the posterior distribution of its parameters. Let $k^{(t)}$ be the number of classes of the model visited at sweep t of the algorithm and $\boldsymbol{\Lambda}^{(t)}$ and $\boldsymbol{\pi}^{(t)}$ be the parameters of this model, with $t = 1,\ldots,T$. Then, we examine every matrix $\boldsymbol{\Lambda}^{(t)}$ and, for $j = 1,\ldots,r$, we obtain the permutations $c_1^{(j)},\ldots,c_{k^{(t)}}^{(j)}$ such that the probabilities in the j-th column of this matrix satisfy inequality (1). As clarified before, these permutations define a partition of the items in groups corresponding to different dimensions. In particular, the permutation at step t is denoted $\mathscr{J}_1^{(t)},\ldots,\mathscr{J}_{s^{(t)}}^{(t)}$, where $s^{(t)}$ is the number of dimensions that is found. To avoid a sort of label-switching problem, the groups are ordered so that $\mathscr{J}_1^{(t)}$ includes the first item, $\mathscr{J}_2^{(t)}$ includes the item with the smallest index among those excluded from $\mathscr{J}_1^{(t)}$, and so on. Finally, the posterior probability of the model with a certain k and a certain partition of items $\mathscr{J}_1,\ldots,\mathscr{J}_s$ based on s dimensions is estimated as:

$$Pr(k, \mathscr{J}_1,\ldots,\mathscr{J}_s) = \frac{1}{T} \sum_{t:s^{(t)}=s} I\left\{ \mathscr{J}_1^{(t)} = \mathscr{J}_1,\ldots,\mathscr{J}_s^{(t)} = \mathscr{J}_s \right\}, \tag{7}$$

where the sum is over all sweeps for which $s^{(t)} = s$ and $I\{\cdot\}$ is the indicator function.

On the basis of posterior probabilities in (7), different strategies may be adopted for model selection. We suggest selecting first the value of k on the basis of the largest number of visits. Then, conditionally on the value of k, we take the partition with the highest value of the probability in (7). This strategy is similar to that in Bartolucci (2007). Alternatively, k and the partition $\mathscr{J}_1,\ldots,\mathscr{J}_s$ can be chosen jointly as those with the highest posterior probability in (7). This method may lead

to a value of k which is suboptimal in terms of posterior probability, and then we prefer the first strategy in order to avoid such an incoherence. However, this point deserves to be investigated in more detail.

2.3 An alternative strategy

The algorithm in Section 2.2 is based on post-processing the RJ output. A different algorithm might be based on in-line processing as briefly illustrated in the following. This alternative algorithm might be more efficient than the previous one, given that the parameter space is reduced. However, it is more difficult to find a reversible move to update k. Thus, we propose to use this algorithm for fixed k.

Under this second approach, we need to specify a prior on the partition $\mathscr{J}$. A simple solution is the uniform on the space of partitions, which arises under an urn model (Corander et al, 2007; Van Cutsem, 1996). Note that the prior for λ_{cj} conditional on the order constraints is rather cumbersome, as it is proportional to $\prod_d \prod_{j \in \mathscr{J}_d} I\{\lambda_{c_{1d}j} \leq \cdots \leq \lambda_{c_{kd}j}\}$. However constraints can be simply imposed during the estimation algorithm, and thus we can ignore the factor above.

The algorithm proceeds along the following steps:

1. Sample indicators of latent class z_{ic} from their full conditional, as in Section 2.2.
2. Update the partition $\mathscr{J}$. We propose two possible moves, which can be sequentially performed. In the *re-allocation of items* move we pick $\mathscr{J}_{d_1}$, with $|\mathscr{J}_{d_1}| > 1$, $j \in \mathscr{J}_{d_1}$, and $\mathscr{J}_{d_2}$ uniformly at random, and we build a proposal $\mathscr{J}^\star$ by adding j to $\mathscr{J}_{d_2}$. In the *exchange of items* move one element from $\mathscr{J}_{d_2}$ is moved to $\mathscr{J}_{d_1}$ and we have no cardinality constraints. These transition mechanisms define an aperiodic and irreducible finite Markov chain, which is guaranteed to converge. For a similar strategy, see Corander et al (2009). The moves are accepted with probability calculated according to the Metropolis-Hastings rule.
3. Update $\boldsymbol{\lambda}_j$ for $j = 1, \ldots, r$, as in Section 2.2. After having sampled the candidate $\boldsymbol{\lambda}_j^\star$, we impose the ordering constraints as implied by the current $\mathscr{J}$. It is straightforward to check that, in the acceptance probability of the re-ordered $\boldsymbol{\lambda}_j^\star$, the terms involved in the ordering cancel out (Green and Richardson, 2001), so that the acceptance probability is the same as the one in Section 2.2.
4. Sample a permutation $d_1, \ldots, d_s$ of $1, \ldots, s$ (this is done to improve the algorithm mixing). For $d = d_1, \ldots, d_s$ a new permutation $c_{1d}, \ldots, c_{kd}$, is drawn uniformly at random in the set of all possible permutations of $\{1, \ldots, k\}$ excluding those active in any other latent class. The move is accepted with a probability which preserves the detailed balance condition.
5. Sample $\pi_1, \ldots, \pi_k$ from their full conditional, as in Section 2.2.
6. Update the number of groups s. This is a relatively new problem, which might be solved by adding a split/merge move to vary the number of item clusters. This is in parallel with popular RJ algorithms, but in a different perspective.

We start by choosing with probability 0.5 between a split and a merge move. If $s = 1$, only a split move is allowed, while if $s = \min(k!, r)$, only a merge move is allowed. In the split move, a class is chosen uniformly at random among those with cardinality strictly larger than one. Let this class be $\mathscr{J}_d$. An integer m between 1 and $|\mathscr{J}_d| - 1$ is drawn uniformly and m elements at random are eliminated from $\mathscr{J}_d$ and used to form a new group, labeled $\mathscr{J}_d + 1$ and all remaining groups are relabeled accordingly. The $\boldsymbol{\Lambda}$ ordering of the new group is randomly chosen from the permutations not currently in use. The parameters $\boldsymbol{\Lambda}$ of the m elements are reordered accordingly, but not otherwise updated.
In the merge move, a class at random is combined with the next one, delivering its $\boldsymbol{\Lambda}$ ordering to the new class. The moves defined form a reversible pair and they are accepted with a probability calculated to preserve detailed balance.

3 An Application in Education Assessment

The proposed approach is applied on a dataset concerning a sample of $n = 1510$ examinees who responded to a set of $r = 12$ items on Mathematics. This dataset is part of a larger dataset collected in 1996 by the Educational Testing Service within the NAEP project; see Bartolucci and Forcina (2005) for a deeper description.

The results correspond to runs of 100 000 sweeps after a burn-in of 10 000 sweeps. Models with a number of latent classes up to $k_{\max} = 10$ were considered. Table 1 shows the posterior distribution $Pr(k|Y)$, from which the model with $k = 4$ latent classes seems to be favored.

k	1	2	3	4	5	6	7	8	9	10
$Pr(k\|Y)$	0.0000	0.0000	0.3684	0.4621	0.1328	0.0311	0.0048	0.0006	0.0002	0.0000

Table 1 Estimated posterior probabilities of the number of latent classes.

Conditionally on $k = 4$, the partition which is most visited by the algorithm is $\mathscr{J} = (\{1,2,5,9,10,11\},\{3,8\},\{4,6,7,12\})$. This partition does not anyway receive a very large posterior mass, given that it is visited only 367 times. If unidimensionality is actually the focus of the analysis, one may also proceed *marginally* with respect to k. In our application, regardless of k, $s = 1$ occurred 36777 times. Using the encompassing approach of Klugkist et al (2005), given that the prior probability is about 0.10; the resulting Bayes factor is 3.7. We therefore actually have some evidence in favor of unidimensionality for the Mathematics data set.

References

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

Bartolucci F, Forcina A (2005) Likelihood inference on the underlying structure of IRT models. Psychometrika 70:31–43

Bartolucci F, Scaccia L, Farcomeni A (2012) Bayesian inference through encompassing priors and importance sampling for a class of marginal models for categorical data. Computational Statistics and Data Analysis 56:4067–4080

Corander J, Gyllenberg M, Koski T (2007) Random partition models and exchangeability for Bayesian identification of population structure. Bull Math Biol 69:797–815

Corander J, Gyllenberg M, Koski T (2009) Bayesian unsupervised classification framework based on stochastic partitions of data and a parallel search strategy. Advances in Data Analysis and Classification 3:3–24

Diebolt J, Robert C (1994) Estimation of finite mixture distributions through bayesian sampling. Journal of the Royal Statistical Society, Series B 56:363–375

Forcina A, Bartolucci F (2004) Modelling quality of life variables with non-parametric mixtures. Environmetrics pp 519–528

Frühwirth-Schnatter S (2001) Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96:194209

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and bayesian model determination. Biometrika 82:711–732

Green PJ, Richardson S (2001) Hidden Markov models and disease mapping. Journal of the American Statistical Association 97:1055–1070

Klugkist I, Kato B, Hoijtink H (2005) Bayesian model selection using encompassing priors. Statistica Nederlandica 59(57-69)

Lazarsfeld PF, Henry NW (1968) Latent Structure Analysis. Houghton Mifflin, Boston

Tuyl F, Gerlach R, Mengersen K (2009) Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. Bayesian Analysis 4:151–158

Van Cutsem B (1996) Combinatorial structures and structures for classification. Computational Statistics and Data Analysis 23:165–188