Bayesian unconstrained estimation for Exploratory Multidimensional IRT models

Abstract

In modern validity theory, a major concern is the construct validity of a test, which is commonly assessed through confirmatory or exploratory factor analysis. In the framework of Bayesian exploratory Multidimensional Item Response Theory (MIRT) models, we discuss two methods aimed at investigating the underlying structure of a test, in order to verify if the latent model adheres to a chosen simple factorial structure. This purpose is achieved by addressing in different ways the problem of the rotational indeterminacy of the final solution. The first approach prescribes a 2-steps procedure. The parameter estimates are obtained through an unconstrained MCMC sampler. The rotational invariance is, then, addressed with a post-processing step based on the Consensus Simple Target Rotation technique. In the second approach, both rotational invariance and simple structure retrieval are addressed **within the MCMC sampling scheme** by introducing a sparsity-inducing prior on the discrimination parameters. Through simulation as well as real-world studies, we prove that the proposed methods are able to correctly infer the underlying sparse structure and to retrieve interpretable solutions.

Keywords: IRT, construct validity, sparse modelling, rotational invariance.

Bayesian unconstrained estimation for Exploratory Multidimensional IRT models

1 Introduction

In the field of test construction, the concept of validity is central and crucial to the development of reliable measurement instruments. In the classical model of test validity, there are three main types of validity evidence: *construct validity*, *content validity* and *criterion validity*. Modern validity theory defines construct validity as the encompassing concern of validity research, incorporating all other types, and essential to the validation of a test (Messick, 1995). As pointed out by Floyd & Widaman (1995), construct validity is supported if the instrument factorial structure is consistent with the constructs the instrument intends to measure. Therefore, factor analysis has long been associated with this type of validity (Goodwin, 1999) and investigation of the factorial structure of the test has been carried out by specifying exploratory (EFA) or confirmatory (CFA) factor analytic measurement models (Di Stefano & Hess, 2005; Henson & Roberts, 2006). In the confirmatory approach, if the theoretical factorial structure is not supported, a sequence of modifications of the model can be carried out to improve the fit. However, in order to discover misspecified loadings, it is more convenient to apply a rotation to the factor loading matrix obtained with an EFA approach. Comprehensive reviews of rotation methods, aimed at detecting simpler or sparse structures, can be found in Browne (2001) and Trendafilov (2014).

Within the social and behavioural sciences, item-level data are often categorical in nature and, in such a case, item factor analysis (IFA) offers an appropriate alternative to common linear factor analysis. Various IFA models can be found in both Structural Equation Models (SEM) and Item Response Theory (IRT) literatures.

In this paper, in the framework of Bayesian Multidimensional Item Response Theory (MIRT) models (Béguin & Glas, 2001), we consider two alternative approaches to exploratory IFA which aim to recover a simple factorial structure for a given test instrument, without imposing ex-ante hard constraints on the factor loading matrix. The

first approach does not constrain the parameter space a priori, but fixes the rotation problem ex-post. In particular, the test structure is explored by transforming the output of an unconstrained MCMC sampler using the Consensus Simple Target Rotation (Lorenzo-Seva et al., 2002) technique. This procedure allows to perform an oblique rotation of two or more loading matrices achieving a compromise between agreement and simplicity. In the second approach, soft constraints are imposed on the discrimination parameters by introducing a sparsity-inducing prior (Mitchell & Beauchamp, 1988; West, 2003; Ishwaran & Rao, 2005) that favours shrinkage, enforcing in this way the sparsity of the factorial solution.

The paper is organised as follows. Section 2 briefly reviews IFA models, introducing the adopted MIRT formulation for polytomous items, whose identifiability issues are discussed in Section 2.1. Section 3 describes the estimation procedure involving an unconstrained MCMC sampler followed by the post-processing step based on the Consensus Simple Target Rotation technique. In Sections 4, we discuss the Bayesian formulation of the MIRT model which includes sparsity inducing priors. The proposed methodologies are evaluated on both simulated and real-world datasets, and the results are discussed in Sections 5 and 6, respectively. *Further simulation studies are presented in the Supplementary Materials.* Section 7 concludes the paper with a discussion of the findings.

2 Multidimensional IRT model for Likert type data

In the IRT literature, several IFA models have been proposed for ordered polytomous data (Ostini & Nering, 2005), differing in the item response function and the number of parameters included in the formulation. In our study, we focus on the two parameter formulation which is more appropriate for the type of data most often encountered in psychological and sociological research. In fact, for rating data it has been argued that there is no guessing or any similar phenomenon that requires lower or upper asymptote parameters (see Maydeu-Olivares et al., 2011, and references therein) . As for the

specification of the response function, we consider the multidimensional generalisation of the Samejima's (1969) graded model. Maydeu-Olivares (2005), comparing the fit of different parametric IRT models to some personality scales, showed how Samejima's model provided the best fit to each of the five scales considered. In addition, Samejima's model is formally equivalent to the ordinal FA model (Takane & de Leeuw, 1987).

Given a test consisting of K ordered categorical variables and assuming M latent traits, the two-parameter normal ogive (2PNO) formulation of the multidimensional graded response model is given by (Béguin & Glas, 2001):

$$P(X_{i,k} = c|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_k, \boldsymbol{\gamma}_k) = \Phi\left(\boldsymbol{\alpha}_k'\boldsymbol{\theta}_i - \gamma_{k,c-1}\right) - \Phi\left(\boldsymbol{\alpha}_k'\boldsymbol{\theta}_i - \gamma_{k,c}\right). \tag{1}$$

In this equation, $X_{i,k}$ is the observed response of person i (i=1,. . . ,N) to item k; c denotes the category of the ordered response scale ($c = 1, ..., C$), and Φ is the standard normal cumulative distribution function. The probability of responding a certain category c depends on the M-dimensional vector $\boldsymbol{\theta}_i = (\theta_{i,1}, \ldots, \theta_{i,M})'$ of the unobserved latent trait scores for subject i, on the M-dimensional vector $\boldsymbol{\alpha}_k = (\alpha_{k,1} \ldots, \alpha_{k,M})'$ of item discrimination parameters and on the $(C-1)$-dimensional vector $\boldsymbol{\gamma}_k = (\gamma_{k,1} \ldots \gamma_{k,C-1})'$ of ordered category thresholds for item k. In the IRT literature, the latent traits are known as person parameters, while the discriminations and thresholds are referred to as item parameters. The factorial structure of the model is represented by the $(K \times M)$ matrix $\boldsymbol{A}$ containing the discrimination parameters.

In a Bayesian estimation framework, we exploit data augmentation technique (Tanner & Wong, 1987) to draw samples from the **joint** posterior distribution of the parameters. We assume that a continuous variable Z_k underlies the observed ordinal measure X_k, and that there is a linear relationship between item and person parameters and the underlying variable, such that $Z_{i,k} = \boldsymbol{\alpha}_k'\boldsymbol{\theta}_i + \epsilon_{i,k}$, with $\epsilon_{i,k} \sim \mathcal{N}(0, 1)$, $\forall\, i, k$. The relation between the

observed items and the underlying variables is given by the threshold model

$$X_{i,k} = c \quad if\ \gamma_{k,c-1} \leq Z_{i,k} \leq \gamma_{k,c},\ c = 1, \ldots, C;\ \gamma_{k,0} = -\infty, \gamma_{k,C} = \infty \tag{2}$$

The full conditional of most parameters can be specified in closed form, which allows for a Gibbs sampler, although Metropolis-Hastings steps are required to sample the ordered threshold parameters and the latent trait correlation matrix.

2.1 Indeterminacy and prior specification

The multidimensional 2PNO model needs identification restrictions given the over-parameterisation. This would require fixing a minimal number of constraints for the model parameters. For location and scale indeterminacy, we assume zero mean and unit variance latent traits. More specifically, in the prior specifications, we assume that all person parameters are independent and identically distributed samples from a multivariate normal distribution, that is $\forall\, i : \boldsymbol{\theta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\theta)$ (Béguin and Glas, 2001), where $\boldsymbol{\Sigma}_\theta$ is a correlation matrix.

The prior distribution for the threshold parameters is usually chosen to be noninformative (Fox, 2010). Here, we define a uniform prior distribution for parameter $\gamma_{k,c}$ truncated to the region $\{\gamma_{k,c} \in \mathcal{R}, \gamma_{k,c-1} \leq \gamma_{k,c} \leq \gamma_{k,c+1}\}$, $c = 1, \ldots, C-1,\ \forall\, k$, to take account of the order constraints (Albert & Chib, 1993).

The remaining issue to address is the rotational indeterminacy of the model, meaning that the latent trait space does not have a unique orientation and the latent axes may be rotated without affecting the probability of a certain response. Thus, for any $M \times M$ invertible matrix $\boldsymbol{T}$, we have $P(X_{i,k} = c|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_k, \boldsymbol{\gamma}_k) = P(X_{i,k} = c|\boldsymbol{T}^{-1}\boldsymbol{\theta}_i, \boldsymbol{T}'\boldsymbol{\alpha}_k, \boldsymbol{\gamma}_k)$ (de Ayala, 2009).

In Bayesian EFA, a common approach to fix the rotational indeterminacy is to impose a positive lower triangular (PLT) structure on the upper $(M \times M)$ block of the factor loadings matrix (Geweke & Zhou, 1996; Lopes & West, 2004). As the constraints are

imposed on particular elements of the discrimination matrix, inference results may depend on the ordering of the variables (Lopes & West, 2004; Pape, 2015).

To avoid this shortcoming, we propose two different estimation approaches aiming at retrieving simple factorial structure without imposing any hard constraints on the discrimination parameters. The first method, discussed in Section 3, performs a post-hoc identification procedure based on oblique rotations of the unconstrained MCMC posterior draws. The second estimation method imposes soft constraints by introducing the sparsity promoting priors, illustrated in Section 4, for the discrimination parameters.

3 An ex-post identification approach to investigate simple structures

For exploratory item factor analysis the latent trait covariance matrix is usually assumed to be an identity matrix, and the initially factor solution is afterwards rotated to a simpler structure to better facilitate interpretation (Chalmers, 2012).

On the assumption that the priors for the person parameters and the discrimination coefficients are $\boldsymbol{\theta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $\forall i$, and $\alpha_{k,m} \sim \mathcal{N}(0, \sigma^2_\alpha)$, $\forall k, m$, the invariance of the likelihood with respect to orthogonal rotations transfers to the posterior distribution, since also the prior distributions are invariant under this kind of rotation. Therefore, when no constraints are imposed on the discrimination parameter matrix, the Gibbs sampler generates an orthogonally mixed sample, meaning that the factor space is orthogonally transformed during the process of sampling. The Gibbs sequences for the discrimination parameters and the latent trait scores are obtained not from the posterior distribution of interest, but from an orthogonal mixture of this distribution (Pape, 2015). In other words, the obtained draws lack orientation and thus do not allow for meaningful inference via the calculation of arithmetic means (Aßmann et al., 2012).

In the EFA framework, Aßmann et al. (2012, 2016) propose an ex-post approach which fixes the rotation problem via a Procrustean postprocessing of the output of an unconstrained Gibbs sampler. Here, following a similar approach, we consider a two step

procedure. In the first step, assuming uncorrelated latent factors and a normal prior for each discrimination parameter, we implement an unconstrained MCMC sampling scheme. In the second step, the Consensus Simple Target Rotation (CSTR; Lorenzo-Seva et al., 2002), is applied to obtain an oblique rotation of the MCMC draws with respect to a common sparse target matrix. The CSTR method allows to both solve the orthogonally mixing problem of the unconstrained MCMC sampling scheme and to investigate the presence of a simple structure. In fact, this technique first derives a semi-specified sparse target matrix from the output of the generalised Procrustes analysis (GPA; Gower, 1975) of the loading matrix draws, then performs a joint simplicity rotation of the MCMC output.

Denoting with $\{\boldsymbol{A}_r\}_{r=1}^{R}$ a sequence of R draws from the orthogonally mixed posterior distribution of $\boldsymbol{A}$, the GPA yields a "consensus matrix" $\boldsymbol{A}^*$ by minimising the quadratic loss function:

$$\underset{\{\{\boldsymbol{U}_r\}_{r=1}^{R},\boldsymbol{A}^*\}}{\arg\min} \sum_{r=1}^{R} L_2\left(\boldsymbol{A}^*, \boldsymbol{A}_r\boldsymbol{U}_r\right) \quad s.t.\ \boldsymbol{U}_r'\boldsymbol{U}_r = \boldsymbol{I} \tag{3}$$

where $L_2\left(\boldsymbol{A}^*, \boldsymbol{A}_r\boldsymbol{U}_r\right) = tr\left[(\boldsymbol{A}_r\boldsymbol{U}_r - \boldsymbol{A}^*)'(\boldsymbol{A}_r\boldsymbol{U}_r - \boldsymbol{A}^*)\right]$, and $\{\mathbf{U}_r\}_{r=1}^{R}$ is a set of orthonormal matrices. This optimisation problem can be solved by a two-step iterative procedure, based on an initial choice of the consensus matrix. Aßmann et al. (2012) suggest to use the last draw of the unconstrained MCMC algorithm to set an initial value for $\boldsymbol{A}^*$, but they state that the algorithm is robust with respect to this initial choice, as long as it stems from the orthogonally mixing posterior distribution. The weighted varimax rotation (Cureton & Mulaik, 1975) of the consensus matrix, $\mathbf{C} = \boldsymbol{A}^*\mathbf{W}$, normalised row-wise to unit sums of square, allows the derivation of a semi-specified sparse target matrix $\mathbf{B}$. Denoting by m_j and s_j the column mean and standard deviation of the squared elements of $\mathbf{C}$, the entry $b_{i,j}$ of $\mathbf{B}$ is set to zero if $c_{i,j}^2 < (m_j + s_j/4)$, and left unspecified otherwise.

Finally, each original MCMC draw $\boldsymbol{A}_r$, $r = 1, \ldots, R$, is independently rotated towards the partly specified target $\mathbf{B}$ using the transformation matrix $\mathbf{T}_r$ which minimises

$$L_2\left(\boldsymbol{B}, \boldsymbol{A}_r\boldsymbol{T}_r\right) \quad s.t.\ diag\left(\boldsymbol{T}_r^{-1}(\boldsymbol{T}_r^{-1})'\right) = \boldsymbol{I} \tag{4}$$

The latent trait score draws $\{\boldsymbol{\Theta}_r\}_{r=1}^R$ are rotated accordingly, $\boldsymbol{\Theta}_r(\boldsymbol{T}_r^{-1})'$, $r = 1, \ldots, R$, and the correlation matrices are given by $\boldsymbol{\Sigma}_r = \boldsymbol{T}_r^{-1}(\boldsymbol{T}_r^{-1})'$.

4 Sparsity inducing prior to investigate simple structure

The rotational invariance issue can be dealt with imposing non invariant priors on the discrimination parameter matrix. To ensure identifiability, one can take independent normal priors for each discrimination parameters and fix certain loadings to constant values. As pointed out in Section 2.1, this may be accomplished by imposing a PLT structure on the discrimination parameter matrix. Further restrictions may also be imposed to allow for dedicated measurements. These restrictions impose patterns of zeroes in the discrimination parameter matrix and postulates a priori the relationship between the observed variables and their underlying latent traits. In practice, one can think of such *hard constraints* as the outcome of a Bayesian analysis with spiked priors. However, one may want to impose *soft constraints*, shrinking estimates in certain directions without forcing them. That is, one may want to perform Bayesian analysis incorporating qualitative prior information, so that likelihood information is blended with prior information rather than simply discarded. By following this approach, which is in between posing no restrictions and forcing restrictions, some structures may be favoured probabilistically instead of being imposed. In particular, assuming that the discrimination parameter matrix contains coefficients close to zero, on the one hand, and larger discrimination coefficients, on the other hand, a Spike and Slab prior can be specified for each discrimination parameter. This kind of priors can be constructed as a two-point mixture distribution made up of a *spike* component, which concentrates its mass at values close to zero, to provide strong shrinkage of small effects to zero, and a *slab* component, which has its mass spread over a wide range of plausible values, to allow parameters to escape strong shrinkage. In our study, we consider the spike and slab prior defined by a

two component normal mixture model (George & Mcculloch, 1993)

$$\alpha_{k,m}|\zeta_{k,m} \sim (1-\zeta_{k,m})\mathcal{N}(0,\tau^2) + \zeta_{k,m}\mathcal{N}(0,g^2\tau^2) \tag{5}$$

where τ is positive but small, such that $\alpha_{k,m}$ is close to zero when $\zeta_{k,m}=0$, and g is large enough to allow reasonable deviations from zero when $\zeta_{k,m}=1$. In addition, the prior probability that factor θ_m has a nonzero effect on item k is $P(\zeta_{k,m}=1)=1-P(\zeta_{k,m}=0)=p_{k,m}$. Hereafter, we refer to this prior specification as the Stochastic Search Variable Selection (SSVS) prior. The classification between zero and non zero coefficients can be based on the posterior probability of inclusion (ppi), given by $P(\zeta_{k,m}=1|\mathbf{X})$ (Frühwirth-Schnatter & Wagner, 2010). In particular, the median probability model (Barbieri & Berger, 2004; Ishwaran & Rao, 2005) can be defined as the model consisting of those discrimination parameters whose posterior probability of inclusion is higher than 50%.

5 Simulation study

In order to evaluate the performance of the proposed procedures, we perform a simulation study. We consider a multidimensional structure which represents a generalisation of the unidimensional approach, since the data matrix contains more than one latent variables, but each item loads only onto a specific factor. In other terms, there is an independent-cluster (IC) latent structure (McDonald, 2000). In the IRT literature, this simple structure has been referred to as a multi-unidimensional schema (Sheng & Wikle, 2007) or a between-item multidimensional IRT model (Adams et al., 1997). Assuming $M=4$ latent constructs, each measured by 5 four-point Likert items, such that $K=20$ is the total number of observed categorical variables, we simulated 100 datasets, setting the sample size to 250, 500 and 1000 and considering uncorrelated, weakly correlated and strongly correlated latent traits. For the weakly and the strongly correlated factors, the determinants of the simulation correlation matrices are 0.697 and 0.367, respectively.

Estimation results for the MCMC algorithm exploiting the spike and slab prior for the discrimination parameters (hereafter denoted as SSVS-MIRT) are compared with the ones obtained through the post processing algorithm based on the CSTR procedure (hereafter denoted as CSTR-MIRT). Since both procedures do not rely on hard constraints to solve the identifiability problem, the posterior estimates of the discrimination matrix can be affected by column and sign switching issues. Therefore, to assess the properties of the estimators over all the 100 simulated datasets, for each dataset we apply shift and sign change to the columns of the posterior estimates $\hat{\mathbf{A}}$, by choosing the signed-permutation matrix $\mathbf{J}^*$ such that the quadratic loss function $L_2\left(\mathbf{A}, \mathbf{J}^*\hat{\mathbf{A}}\right)$ is minimised. Here, $\mathbf{A}$ denotes the discrimination matrix used in the simulation. The posterior estimates of the latent trait scores and correlation matrix are transformed accordingly.

With regard to the identification of a simpler structure, in the traditional exploratory framework, a common criterion for the selection of relevant discrimination parameters is to fix a cutoff value, such that only the items whose absolute discrimination value is larger than the cutoff will load onto the corresponding latent trait. In contrast, in the sparse Bayesian model the relevance of a discrimination parameter is assessed on the basis of its posterior probability of inclusion. In the analysis of the simulation results, in order to evaluate the capacity of the algorithm to recover the true structure, we take into account the following measures: *retrieved level of sparsity*, defined as the proportion of zero discrimination parameters in the estimated sparse solution; *sensitivity*, measured as the proportion of correctly identified nonzero discrimination coefficients; *specificity*, given by the proportion of correctly identified zero coefficients; *accuracy*, representing the proportion of identified zero and non-zero loadings. In particular, for the SSVS-MIRT approach we focus on the median probability model ($ppi > 0.5$), while in the CSTR-MIRT procedure we fix the cutoff value to 0.3. In addition, we show also how those indexes vary with respect to different thresholds for the posterior probability of inclusion and the discrimination value.

In the following Sections, we discuss the discrimination parameter estimates and the

factorial structure retrieval. More details on simulation parameters, estimation procedure and results can be found in the Supplementary Materials, along with simulation results for different numbers of observed items and latent traits. In addition, in the Supplementary Materials, we present also simulation results for different within-item multidimensional model (Adams et al., 1997) allowing for an increasing number of cross-loadings. **The simulation results confirm the good performances of the proposed estimation algorithms for different level of sparsity of the latent structure.**

5.1 Discrimination parameter estimates and factorial structure retrieval

In the Bayesian estimation procedure for the CSTR-MIRT approach we assume a zero mean unit variance Gaussian prior for all the factor loadings, while in the SSVS-MIRT Spike and Slab prior in equation (5) we set $\tau = 0.01$, $g = 100$ and $p_{k,m} = 0.5$ for all $k = 1, \ldots, K$ and $m = 1, \ldots, M$. In Section 5.2, we illustrate how modifying those hyperparameters affects the posterior estimates of the discrimination parameter matrix and the factorial structure retrieval.

Table 1 shows the average absolute bias (AAB) and the average mean squared error (AMSE) over all the discrimination parameter posterior estimates, where the posterior estimates are obtained as the means of the corresponding posterior distributions. As it can be highlighted, both procedures yield comparable results, though the SSVS estimates are less biased and more efficient; moreover we can notice how results improve with the increase of the sample size. The latent trait correlation structure does not seem to affect the estimation bias and efficiency for the SSVS procedure, while for the CSTR estimates the bias increase with the increasing of the correlation.

For a sample size equal to $N = 500$, Figure 1 represents the simulation discrimination parameters and the means of the posterior estimates across the 100 simulated datasets, along with the 2−standard deviation intervals around the means. For the the coefficients fixed to zero in the simulation settings, results show a better performance of the procedure

based on the sparsity inducing prior.

Figure 2 shows the boxplots of the distributions of the numbers of relevant discrimination parameters selected by fixing different thresholds for their posterior probabilities of inclusion (upper panel), and their absolute values (lower panel). It can be highlighted how in the SSVS approach, for $ppi \geq 0.5$ the number of selected relevant coefficients shows a high stability. In particular, for uncorrelated latent traits, fixing $ppi = 1$ leads to select 20 coefficients in all the simulated datasets, while for the median model the number of relevant parameters varies between a minimum of 20 and a maximum of 24, independently of the correlation structure. For correlated latent traits, setting $0.2 \leq |\alpha_{k,m}| \leq 0.5$ in the CSTR procedure leads to an adequate stability in the number of selected relevant parameters.

Those findings are confirmed when we analyse the performance of both estimation methods in terms of accuracy considering different sample sizes. In fact, Figure 3 shows how for the CSTR procedure the maximum accuracy for all the simulation settings is reached setting the cutoff value to 0.3 or 0.4 , while for the the SSVS approach for $ppi \geq 0.5$ the accuracy reaches an approximately steady level. For both approaches, the accuracy index increases as the sample size increases and, for any given sample size, is higher for uncorrelated latent traits. For a sample size equal to 500 (see Table 2), using the median probability model ($ppi \geq 0.5$), in the SSVS approach, and the traditional cut-off value $|\alpha| \geq 0.3$, in the CSTR procedure, leads to satisfactorily retrieve the level of sparsity of the simulation, equal to 75%. Furthermore, the accuracy of the estimated sparse structure is very high for both approaches. It is worth noticing how the median model for the SSVS procedure always leads to correctly identifying the relevant discrimination parameters. The CSTR procedures shows high levels of sensitivity and specificity, with some drawbacks in case of uncorrelated latent traits.

5.2 Sensitivity of the retrieved sparsity with respect to the spike and slab prior hyperparameters

In this Section, we analyse how the posterior estimates of the discrimination parameters and the retrieval of the factorial structure are affected by the choice of the SSVS prior hyperparameter values. In particular, in order to investigate how the choice of the prior probability that factor θ_m has a nonzero effect on item k, in addition to $p_{k,m} = 0.50$, we set $p_{k,m}$ to 0.25 and 0.75, considering $\tau = 0.01, g = 100$. Then, fixing $p_{k,m} = 0.50$, we put $\tau = 0.001, g = 1000$, and $\tau = 0.1, g = 10$. For each combination of these hyperparameters, Tables 3 and 4 present the absolute average bias and the average mean squared error, distinct for the discrimination parameters set to zero or not zero in the simulation setting. The bias and the efficiency of the discrimination parameter posterior estimates are quite similar with respect to the choice of the SSVS hyperparameters. However, for uncorrelated latent traits, setting $p_{k,m} = 0.25$, $\tau = 0.01$ and $g = 100$ yields higher AAB and AMSE in the estimates of the non zero factor loadings. For this combination of hyperparameters, the lower capacity to correctly identified non zero discrimination coefficients is confirmed by the sensitivity index, depicted in Figure 4, which decreases as the posterior probability of inclusion increases. On the other hand, for $p_{k,m} = 0.75$, $\tau = 0.01$ and $g = 100$, the sensitivity index is close to 1, independently of the choice of the posterior probability of inclusion, while the capacity to correctly identifying zero coefficients, measured by the specificity index, is inadequate for $ppi < 0.5$. With regard to the variances of the two component normal mixture distribution in equation (5), Figure 5 shows how the accuracy is stably high for a $ppi > 0.1$, though setting $p_{k,m} = 0.5$, $\tau = 0.1$ and $g = 10$ leads to a lower capacity of correctly identifying relevant coefficients for a posterior probability of inclusion equal to 1. It is worth noticing that, for all the SSVS prior hyperparameters combinations, the median probability model leads to an accuracy index close to 1.

6 The Human Styles Questionnaire structure

In this Section, the proposed procedures are applied to estimate the factorial structure of data collected through a validated test instrument. Martin et al. (2003) discuss the development and initial validation of the Humor Styles Questionnaire (HSQ). This instrument aims to assess four dimensions relating to different uses or functions of humor in everyday life: one's relationships with others (*Affiliative*); the self (*Self-enhancing*); the self at the expense of others (*Aggressive*); relationships at the expense of self (*Self-defeating*). The analysed dataset[1] contains the responses to 32 items (8 items per dimension), rated on a 5-point Likert scale. Records with missing values have been removed from the analysis and the final sample is composed by 993 respondents. Since, the authors hypothesise that some positive correlations can exist among the four humour dimensions, due to the overlap between the various functions of humour, in our exploratory MIRT model we consider a 4-dimensional solution with correlated factors. **We also considered 1 to 5-factor solutions and compared those models by using the following index (Spiegelhalter et al., 2002; Gelman et al., 2003; Celeux et al., 2006)**

$$DIC_V = \overline{D(\boldsymbol{\xi}, \boldsymbol{\theta})} + p_V = -2E_{\boldsymbol{\xi},\boldsymbol{\theta}}\left[\log f(\mathbf{X}|\boldsymbol{\xi}, \boldsymbol{\theta})\right] + 2Var_{\boldsymbol{\xi},\boldsymbol{\theta}}\left[\log f(\mathbf{X}|\boldsymbol{\xi}, \boldsymbol{\theta})\right] \tag{6}$$

where $\boldsymbol{\xi}$ denotes the item parameters and $f(\mathbf{X}|\boldsymbol{\xi}, \boldsymbol{\theta})$ is the likelihood function. The deviance information criterion comparison, shown in Table 5, supports the choice of the 4-factor solution.

In what follows, we compare the estimates of the discrimination parameter matrix obtained employing the SSVS algorithm and the CSTR post-processing procedure. For the Spike and Slab priors we set $p_{k,m} = 0.25$, $\tau = 0.1$ and $g = 10$. Full estimation results are provided in the Supplementary Materials.

Figure 6 shows the rank order of the SSVS estimates of the discrimination

[1] Data are available at `<http://personality-testing.info/_rawdata/>`

parameters, sorted respect to their posterior probabilities of inclusion (panel a-MIRT-SSVS), and of the CSTR estimates, sorted according to their absolute values (panel b- CSTR-MIRT). It can be noticed how the posterior probabilities of inclusion allow for a sharper distinction between relevant and non relevant discrimination coefficients. Table 6 shows how choosing the median model in the SSVS procedure allows to achieve a retrieved level of sparsity equal to 73.4%. On the other hand, choosing a cut-off value $|\alpha| \geq 3$ in the CSTR post-processing procedure leads to a sparsity level of 74.2%. The hypothesised and the retrieved sparse structures are represented in Figure 7. We can highlight how for both estimation procedures, the retrieved sparse structures adhere in an adequate manner to the hypothesised model. More in details, for the SSVS-MIRT procedure there are only items items showing crossloadings. On the other hand, for the CSTR procedure setting the cut-off value to $|\alpha| = 0.3$ leads to a solution with only one item loading into two traits.

7 Conclusion

Investigating the number and the nature of the dimensions underlying a test is an important aspect of construct validation. In this paper, we have discussed two different Bayesian estimation procedures for exploratory MIRT models aimed to assess the factorial structure of a test. The novelty of our proposals relies in the fact that, in contrast with classical approaches adopted in EFA framework, we do not impose any hard constraints on the factorial structure. The first proposed procedure derives parameter estimates by means of an unconstrained Gibbs sampler followed by a post-processing step aimed to address simultaneously the rotational invariance issue and the investigation of a simple structure. In the second approach, instead, the normal ogive formulation of MIRT polytomous models is modified by introducing a sparsity-inducing prior on the discrimination parameters. Both estimation procedures allow to evaluate the coherence of the estimated discrimination parameter matrix with the hypothesised test structure and favour the interpretability of

the final solution. As shown in the simulation study of Section 5, both methods correctly retrieve the underlying factorial structure. Comparing the results, it can be noticed that SSVS-MIRT estimation approach tends to be less biased and more efficient than CSTR-MIRT procedure. One interesting aspect arises when testing the sensitivity of the results to different choices for the cut-off values that regulate the sparsity of the factor loading matrix. The simulation results show how SSVS-MIRT is less sensitive to different thresholds and the use of the posterior probability of inclusion provides a sharper distinction between relevant and non relevant loadings. Moreover, the use of *ppi* has the advantage of reflecting the degree of confidence in a discrimination parameter being effectively zero. Indeed, this allows for a straightforward definition of a threshold value. On the contrary, in CSTR-MIRT model, the idea is simply to ignore the discrimination parameters with lower magnitudes and the cut-off value does not carry any intrinsic meaning.

These main findings are confirmed also in the real-world data application. Here, both procedures show good performance. The estimated sparse factorial structures exhibit good coherence with the test design and the interpretability of the solution is greatly enhanced.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, *21*(1), 1–23. doi: 10.1177/0146621697211001

Albert, J. H., & Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, *88*(422), 669–679. doi: 10.1080/01621459.1993.10476321

Aßmann, C., Boysen-Hogrefe, J., & Pape, M. (2012). *The Directional Identification Problem in Bayesian Factor Analysis: An Ex-Post Approach.* (Vol. 1799; Working Papers). Kiel Institute for the World Economy.

Aßmann, C., Boysen-Hogrefe, J., & Pape, M. (2016). Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *Journal of Econometrics*, *192*(1), 190-206. doi: 10.1016/j.jeconom.2015.10.010

Barbieri, M. M., & Berger, O. B. (2004). Optimal Predictive Model Selection. *The Annals of Statistics*, *32*(3), 870–897. doi: 10.1214/009053604000000238

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541–562. doi: 10.1007/BF02296195

Browne, M. W. (2001). An Overview of Analytic Rotation in Exploratory Factor Analysis. *Multivariate Behavioral Research*, *36*(1), 111–150. doi: 10.1207/S15327906MBR3601_05

Celeux, G., Forbes, F., Robert, C., & Titterington, D. (2006). Deviance Information Criteria for Missing Data Models. *Bayesian Analysis*, *1*(4), 651–674. doi: 10.1214/06-BA122

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. doi: 10.18637/jss.v048.i06

Cureton, E. E., & Mulaik, S. (1975). The weighted varimax rotation and the promax rotation. *Psychometrika*, *40*, 183–195. doi: 10.1007/BF02291565

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press. doi: 10.1111/j.1745-3984.2010.00124.x

Di Stefano, C., & Hess, B. (2005). Using Confirmatory Factor Analysis for Construct Validation: An Empirical Review. *Journal of Psychoeducational Assessment*, *23*(3), 225–241. doi: 10.1177/073428290502300303

Floyd, F. J., & Widaman, K. F. (1995). Factor Analysis in the Development and Refinement of Clinical Assessment Instruments. *Psychological Assessment*, *7*(3), 286–299. doi: 10.1037/1040-3590.7.3.286

Fox, J. P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York, NY: Springer New York. doi: 10.1007/978-1-4419-0742-4

Frühwirth-Schnatter, S., & Wagner, H. (2010). Bayesian Variable Selection for Random Intercept Modeling of Gaussian and non-Gaussian Data. In *Bayesian statistics 9*. Oxford University Press. doi: 10.1093/acprof:oso/9780199694587.003.0006

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2003). *Bayesian Data Analysis, Second Edition*. Chapman and Hall/CRC.

George, E. I., & Mcculloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. doi: 10.1080/01621459.1993.10476353

Geweke, J., & Zhou, G. (1996). Measuring the Pricing Error of the Arbitrage Pricing Theory. *The Review of Financial Studies*, *9*(2), 557–587. doi: 10.1093/rfs/9.2.557

Goodwin, D. L. (1999). The Role of Factor Analysis in the Estimation of Construct Validity. *Measurement in Physical Education and Exercise Science*, *3*(2), 85–100. doi: 10.1207/s15327841mpee0302_2

Gower, J. C. (1975). Generalised Procrustes Analysis. *Psychometrika*, *1*(40), 33–51. doi: 10.1007/BF02291478

Henson, R. K., & Roberts, J. K. (2006). Use of Exploratory Factor Analysis in Published Research. *Educational and Psychological Measurement*, *66*(3), 393–416. doi: 10.1177/0013164405282485

Ishwaran, H., & Rao, J. S. (2005). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics*, *33*(2), 730–773. doi: 10.1214/009053604000001147

Lopes, H. F., & West, M. (2004). Bayesian Model Assessment in Factor Analysis. *Statistica Sinica*, *14*, 41–67.

Lorenzo-Seva, U., Kiers, H. A. L., & ter Berge, J. M. F. (2002). Techniques for oblique factor rotation of two or more loading matrices to a mixture of simple structure and optimal agreement. *British Journal of Mathematical and Statistical Psychology*, *55*, 337–360. doi: 10.1348/000711002760554624

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality*, *37*(1), 48–75. doi: 10.1016/S0092-6566(02)00534-2

Maydeu-Olivares, A. (2005). Further Empirical Results on Parametric Versus Non-Parametric IRT Modeling of Likert-Type Personality Data. *Multivariate Behavioral Research*, *2*(40), 261-279. doi: 10.1207/s15327906mbr4002_5

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the Fit of Item Response Theory and Factor Analysis Models. *Structural Equation Modeling*, *3*(18), 333-356. doi: 10.1080/10705511.2011.581993

McDonald, R. P. (2000). A Basis for Multidimensional Item Response Theory. *Applied Psychological Measurement*, *24*(2), 99–114. doi: 10.1177/01466210022031552

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi: 10.1037/0003-066X.50.9.74

Mitchell, T., & Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032. doi: 10.1080/01621459.1988.10478694

Ostini, R., & Nering, L. (2005). *Polytomous Item Response Theory Models*. Thousand Oaks, CA: SAGE. doi: 10.4135/9781412985413

Pape, M. (2015). *Model Identification in Bayesian Analysis of Static and Dynamic Factor Models* (Unpublished doctoral dissertation). Christian-Albrechts-Universität zu Kiel.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*(17).

Sheng, Y., & Wikle, C. (2007). Comparing Multiunidimensional and Unidimensional Item Response Theory Models. *Educational and Psychological Measurement*, *67*(6), 899–919. doi: 10.1177/0013164406296977

Spiegelhalter, D. J., Best, N., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, *64*(4), 583–639. doi: 10.1111/1467-9868

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *3*(52), 393-408. doi: 10.1007/BF02294363

Tanner, M. A., & Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540. doi: 10.2307/2289457

Trendafilov, N. T. (2014). From simple structure to sparse components: a review. *Computational Statistics*, *29*(3), 431-454. doi: 10.1007/s00180-013-0434-5

West, M. (2003). Bayesian Factor Regression Models in the "Large p, Small n" Paradigm. In J. M. Bernardo et al. (Eds.), *Bayesian statistics* (Vol. 7, pp. 733–742). Oxford: Oxford University Press.

Table 1
Average absolute bias and average mean squared error for the discrimination parameters over the 100 *simulated datasets (*SSVS: $\tau = 0.01$, $g = 100$, $p_{k,m} = 0.5$*).*

		SSVS-MIRT		CSTR-MIRT	
Sample Size	Latent trait correlation	AAB	AMSE	AAB	AMSE
250	*uncorrelated*	0.010	0.008	0.012	0.016
	weakly correlated	0.012	0.008	0.027	0.016
	strongly correlated	0.010	0.012	0.053	0.022
500	*uncorrelated*	0.008	0.004	0.010	0.010
	weakly correlated	0.007	0.004	0.018	0.007
	strongly correlated	0.008	0.005	0.046	0.012
1000	*uncorrelated*	0.004	0.002	0.007	0.004
	weakly correlated	0.004	0.002	0.014	0.004
	strongly correlated	0.005	0.002	0.042	0.007

Table 2
Mean and standard deviation over the 100 *simulated datasets of the sparsity level, and the sensitivity, specificity and accuracy indexes. Cut-off values: SSVS-MIRT ppi* ≥ 0.5*; CSTR-MIRT* $|\alpha| \geq 0.3$ *(*$N = 500$; SSVS: $\tau = 0.01$, $g = 100$, $p_{k,m} = 0.5$*)*

	Latent trait correlation		sparsity level	sensitivity	specificity	accuracy
SSV MIRT	uncorrelated	mean	0.739	1.000	0.985	0.989
		std	*0.010*	*0.000*	*0.013*	*0.010*
	weakly correlated	mean	0.737	1.000	0.982	0.987
		std	*0.012*	*0.000*	*0.016*	*0.012*
	strongly correlated	mean	0.735	1.000	0.980	0.985
		std	*0.012*	*0.000*	*0.017*	*0.012*
CSTR MIRT	uncorrelated	mean	0.749	0.988	0.994	0.993
		std	*0.008*	*0.055*	*0.021*	*0.029*
	weakly correlated	mean	0.750	1.000	1.000	1.000
		std	*0.001*	*0.005*	*0.000*	*0.001*
	strongly correlated	mean	0.748	0.999	0.997	0.997
		std	*0.006*	*0.009*	*0.007*	*0.006*

Table 3
Average absolute bias and average mean squared error for the discrimination parameters over the 100 *simulated datasets for different prior probabilities that latent trait* θ_m *has a nonzero effect on item* k *(*$N = 500; \tau = 0.01$, $g = 100$*).*

Discrimination coefficients	Latent trait correlation	$p_{k,m} = 0.50$ AAB	$p_{k,m} = 0.50$ AMSE	$p_{k,m} = 0.25$ AAB	$p_{k,m} = 0.25$ AMSE	$p_{k,m} = 0.75$ AAB	$p_{k,m} = 0.75$ AMSE
zero	*uncorrelated*	0.003	0.001	0.004	0.002	0.004	0.001
	weak	0.003	0.001	0.003	0.001	0.004	0.002
	strong	0.004	0.001	0.004	0.002	0.007	0.003
non zero	*uncorrelated*	0.024	0.015	0.056	0.070	0.026	0.016
	weak	0.019	0.013	0.013	0.020	0.021	0.014
	strong	0.020	0.016	0.014	0.024	0.030	0.019

Table 4
Average absolute bias and average mean squared error for the discrimination parameters over the 100 *simulated datasets for different priors on the variances of the two component normal mixture distribution (*$N = 500; p_{k,m} = 0.5$*).*

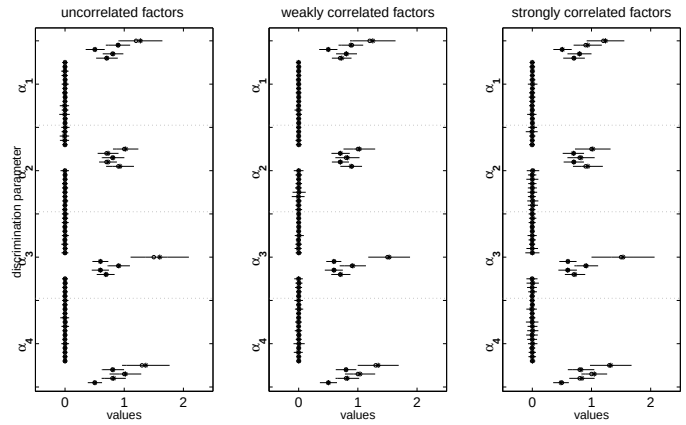
Discrimination coefficients	Latent trait correlation	$\tau = 0.001$ $g = 1000$		$\tau = 0.1$ $g = 10$	
		AAB	AMSE	AAB	AMSE
zero	*uncorrelated*	0.003	0.001	0.005	0.002
	weak	0.003	0.001	0.005	0.002
	strong	0.004	0.001	0.006	0.003
non zero	*uncorrelated*	0.020	0.016	0.027	0.015
	weak	0.017	0.013	0.021	0.014
	strong	0.021	0.015	0.025	0.018

Table 5
DIC values for the HSQ dataset.

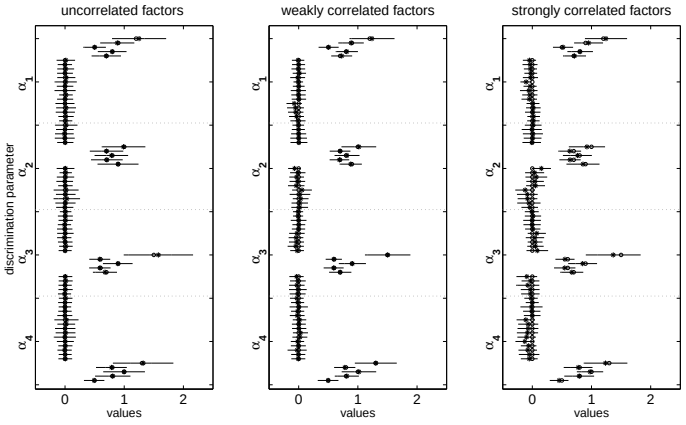
Number of factors	CSTR-MIRT			SSVS-MIRT		
	$\overline{D(\boldsymbol{\xi},\boldsymbol{\theta})}$	p_V	DIC_V	$\overline{D(\boldsymbol{\xi},\boldsymbol{\theta})}$	p_V	DIC_V
1	84742	225	84967	84746	229	84975
2	80591	701	81292	80598	578	81176
3	77529	859	78388	77547	861	78408
4	74489	1273	75761	74518	1216	75734
5	73409	5331	78741	73471	7602	81074

Table 6
HSQ scale: retrieved levels of sparsity for different cut-off values for the posterior probability of inclusion, in the SSVS-MIRT approach, and for the discrimination parameters, in the CSTR-MIRT post-processing procedure

cut-off value ν	SSVS-MIRT percentage of $P(\zeta_{k,m} = 1\|\boldsymbol{X}) < \nu$	CSTR-MIRT percentage of $\|\alpha_{k,m}\| < \nu$
0.1	60.1 %	48.4 %
0.2	66.4 %	66.4 %
0.3	68.7 %	74.2 %
0.4	70.3 %	75.8 %
0.5	73.4 %	76.6 %
0.6	74.2 %	78.9 %
0.7	75.8 %	82.0 %
0.8	75.8 %	85.2 %
0.9	75.8 %	88.3 %
1.00	77.3 %	89.1 %



(a) SSVS-MIRT



(b) CSTR-MIRT

Figure 1. Discrimination parameter estimates: simulation values ($*$); mean of the posterior estimates ($\circ$) over all the 100 simulated datasets; lines represents 2−standard deviation intervals around the means ($N = 500$. *SSVS:* $\tau = 0.01$, $g = 100$, $p_{k,m} = 0.5$).

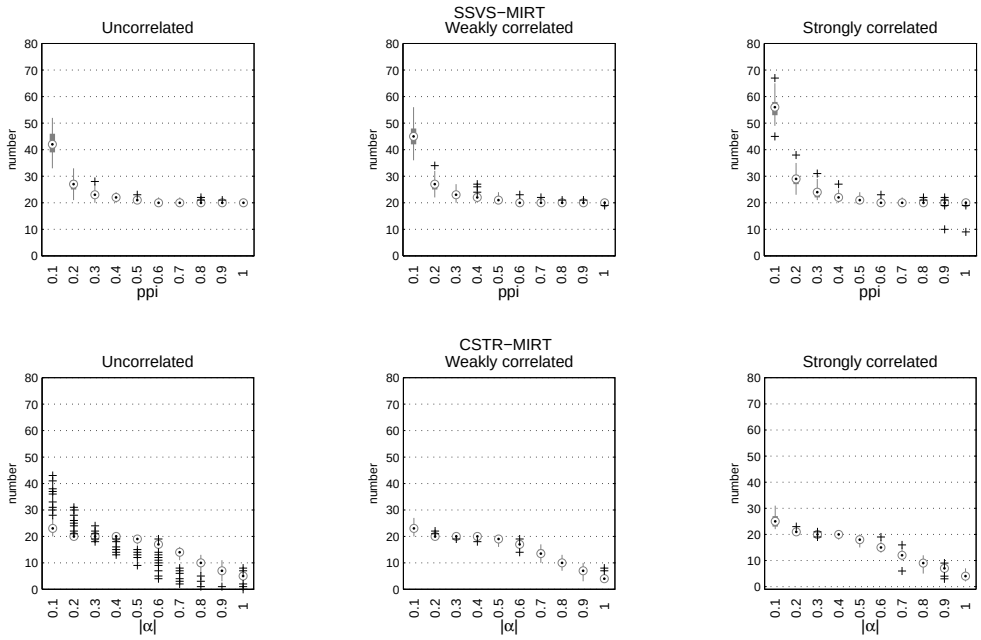


Figure 2. Distributions of the number of relevant discrimination parameters over the 100 simulated datasets according to different cut-off values for the posterior probability of inclusion (upper panel: SSVS-MIRT procedure - $\tau = 0.01$, $g = 100$, $p_{k,m} = 0.5$) and the discrimination parameter absolute value (lower panel : CSTR-MIRT procedure). $N = 500$

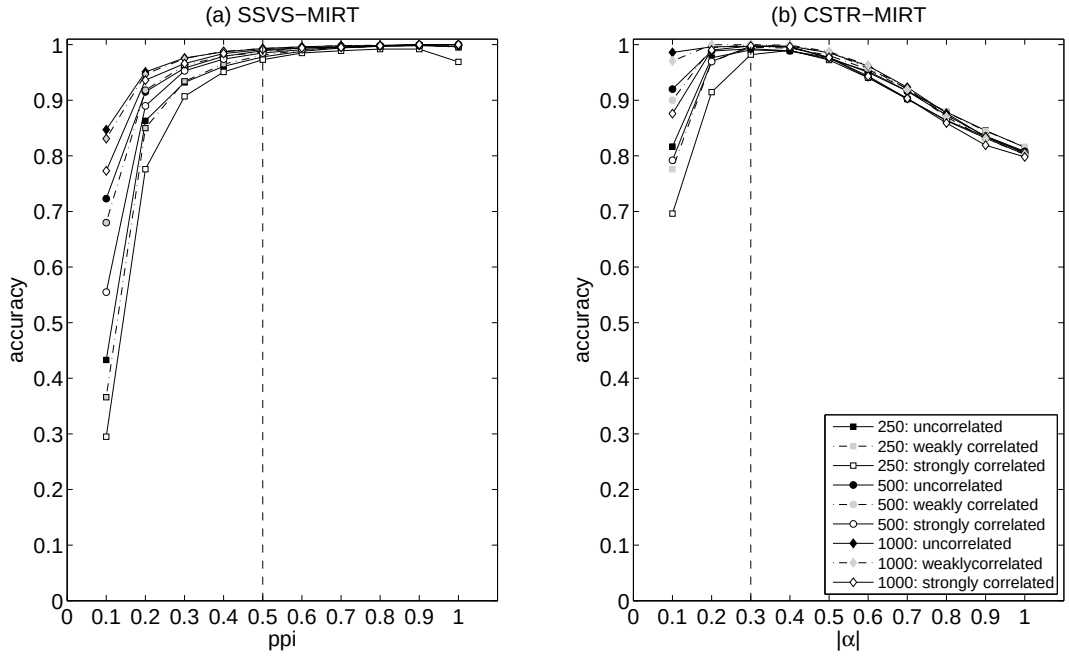


Figure 3. Accuracy performance according to different cut-off values for the posterior probability of inclusion ((a) SSVS-MIRT procedure - $\tau = 0.01$, $g = 100$, $p_{k,m} = 0.5$) and for the discrimination parameter absolute value ((b) CSTR-MIRT procedure).

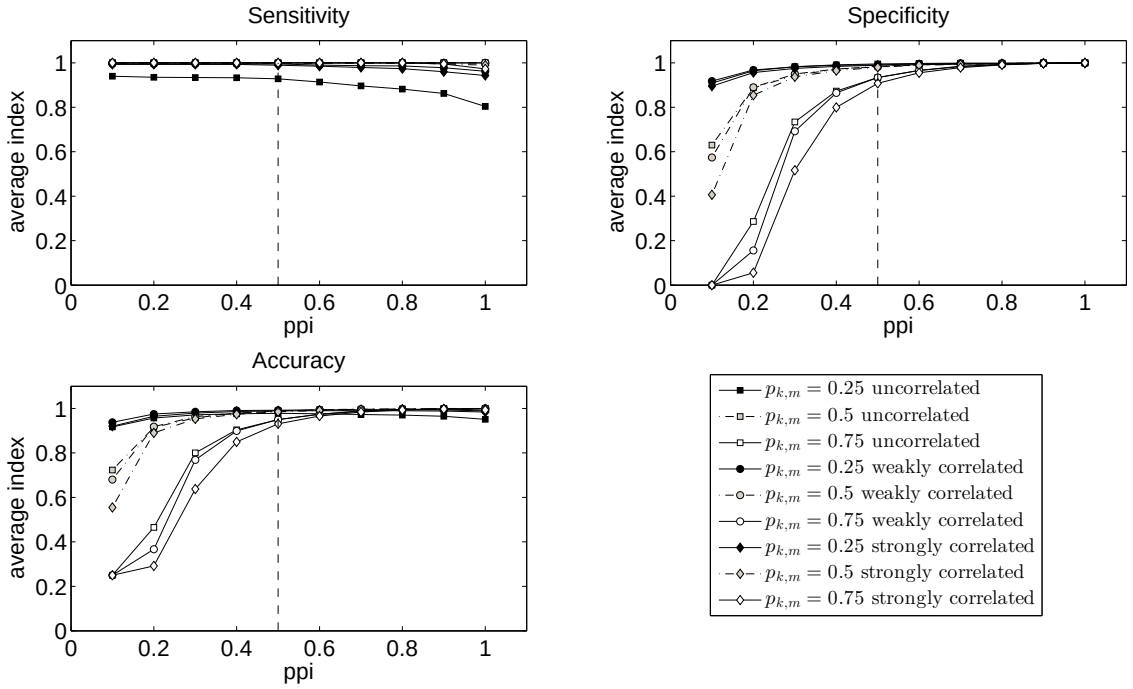


Figure 4. Sensitivity, specificity and accuracy indexes according to different cut-off values for the posterior probability of inclusion (*N=500;* $\tau = 0.01$, $g = 100$)

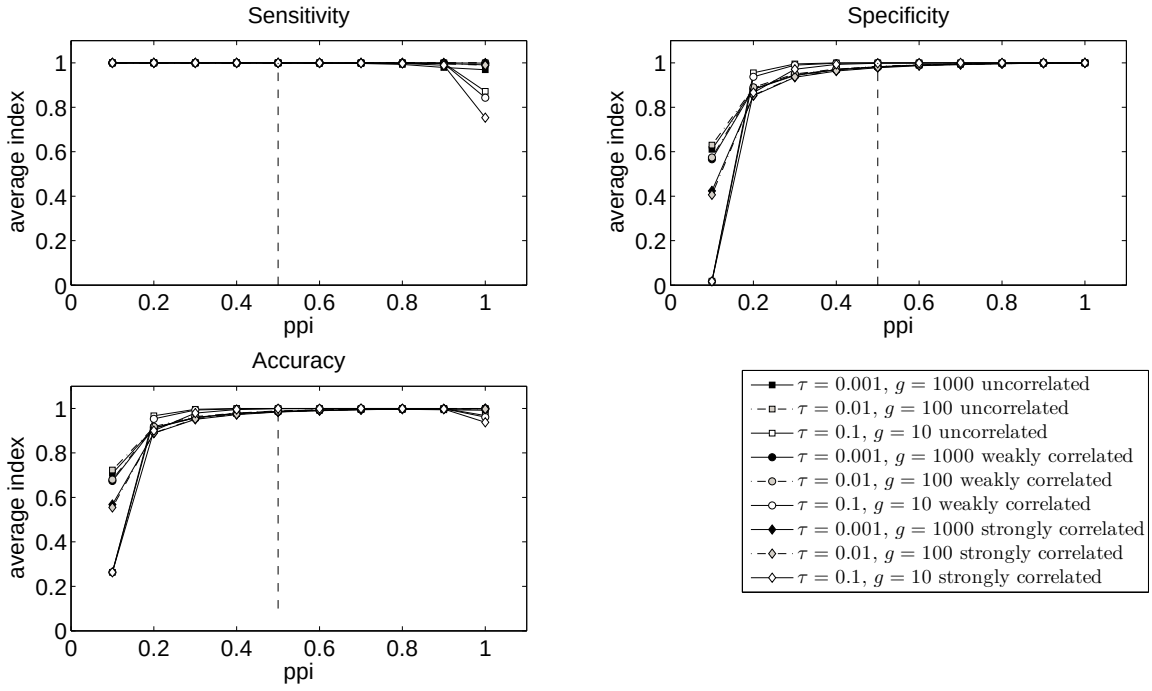


Figure 5. Sensitivity, specificity and accuracy indexes according to different cut-off values for the posterior probability of inclusion (*N=500;* $p_{k,m} = 0.5$)

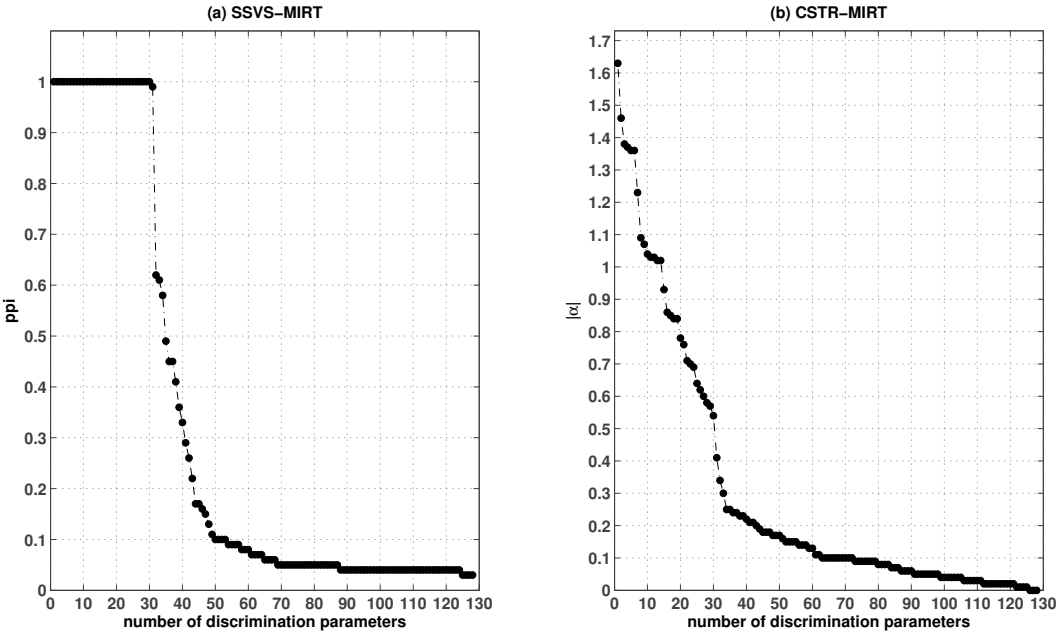


Figure 6. HSQ scale: rank order of the discrimination parameters according to their posterior probability of inclusion (panel a: SSVS-MIRT) and to their absolute values (panel b: CSTR-MIRT).

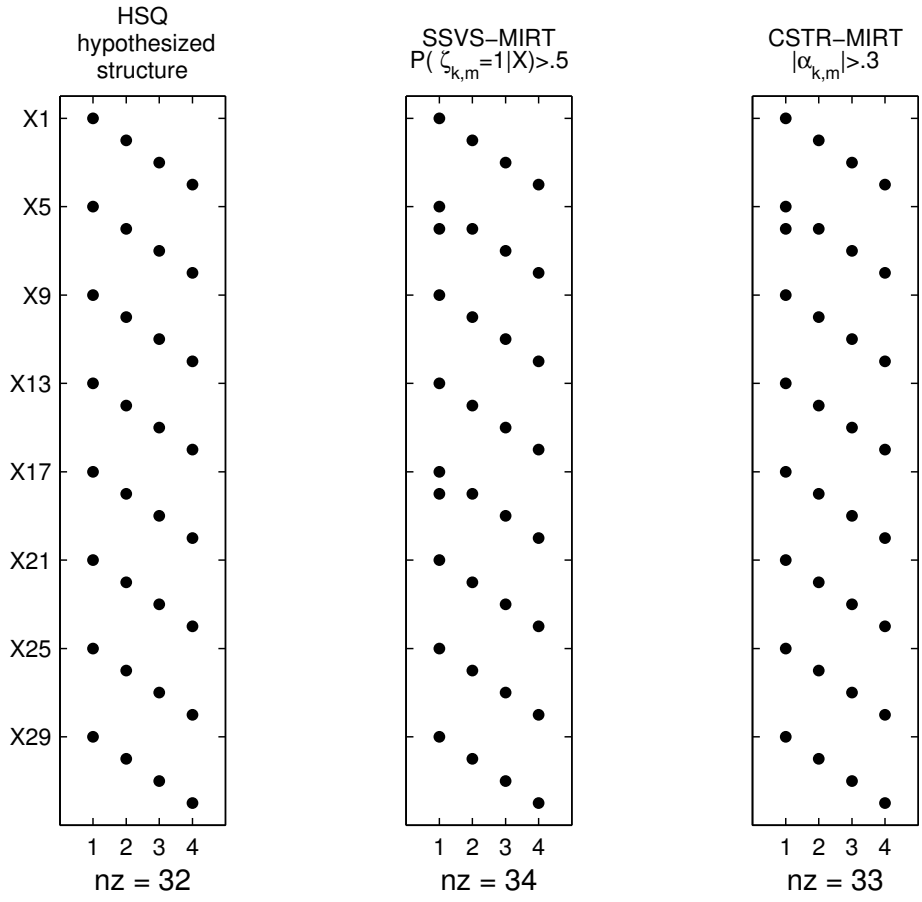


Figure 7. Sparse structures of the discrimination parameter matrix for the HSQ scale - *nz (non-zero) indicates the number of relevant discrimination parameters.*