PSYCHOMETRIKA—VOL. 81, NO. 4, 921–939
DECEMBER 2016
DOI: 10.1007/s11336-016-9529-6

# LATENT VARIABLE SELECTION FOR MULTIDIMENSIONAL ITEM RESPONSE THEORY MODELS VIA $L_1$ REGULARIZATION

JIANAN SUN

BEIJING FORESTRY UNIVERSITY

YUNXIAO CHEN

EMORY UNIVERSITY

JINGCHEN LIU AND ZHILIANG YING

COLUMBIA UNIVERSITY

TAO XIN

BEIJING NORMAL UNIVERSITY

We develop a latent variable selection method for multidimensional item response theory models. The proposed method identifies latent traits probed by items of a multidimensional test. Its basic strategy is to impose an $L_1$ penalty term to the log-likelihood. The computation is carried out by the expectation–maximization algorithm combined with the coordinate descent algorithm. Simulation studies show that the resulting estimator provides an effective way in correctly identifying the latent structures. The method is applied to a real dataset involving the Eysenck Personality Questionnaire.

Key words: latent variable selection, multidimensional item response theory model, $L_1$ regularization, expectation–maximization, BIC.

## 1. Introduction

Psychological and educational tests are often conducted to investigate multiple latent traits or skills by making use of dichotomous-response or polytomous-response items. A key element in such tests is the relationship between the items and the latent traits. It is conventional to pre-specify the relationship by experts' prior knowledge of the items and of the latent traits. The correct specification of latent traits associated with each item is crucial both for the model parameter calibration and for the assessment of each individual. Misspecification of the item–trait relationship may lead to serious model lack of fit and, consequently, erroneous assessment. An interesting question is whether this can be estimated empirically based on the data. In this paper, this question is cast as a variable selection problem under the setting of multidimensional item response theory (IRT) models.

The concept of the multidimensional IRT can be traced back to McDonald (1967), Lord and Novick (1968), and Reckase (1972). As summarized in Reckase (1997, 2009) and Embretson and Reise (2000), the multidimensional IRT models contain two or more parameters to describe the interaction between the latent traits and the responses to items. Some additional references are Sympson (1978), Embretson (1984), Ansley and Forsyth (1985), Way, Ansley, and Forsyth (1988), Ackerman (1989), and Reckase (1997).

Correspondence should be made to Jingchen Liu, Columbia University, New York, USA. Email: jcliu@stat.columbia.edu

To be more precise, we associate each subject (examinee) to an unobserved multivariate trait vector denoted by $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^K)^{\mathrm{T}}$, each component of which represents one latent trait. Throughout this paper, we assume that the responses are binary. For other types of responses, the proposed method can be adapted. Each subject responds to $J$ items. The response function of item $j$ takes the form $P(Y_j = 1|\boldsymbol{\theta}) = F(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta} + b_j)$ where $\mathbf{a}_j = (a_{j1}, \ldots, a_{jK})^{\mathrm{T}}$ and $F$ is some cumulative distribution function. We say that item $j$ is associated with latent trait $k$ if $a_{jk} \neq 0$. Of interest in this paper is the set of traits that are associated with each item. We formulate it as a latent variable selection problem, that is, for each item, we select the set of latent variables influencing the distribution of its responses. We employ variable selection techniques developed for regular regression models for our analysis. An equivalent approach would be defining the incidence matrix $\Lambda = (\lambda_{jk})$ where $\lambda_{jk} = I(a_{jk} \neq 0)$ and consider $\Lambda$ as part of the model parameters. This, however, could look more intimidating due to the discreteness of $\Lambda$. We adopt the latent variable formulation throughout our analysis.

In the literature, various methods have been introduced for the item parameter estimation including the marginal likelihood method (Bock, Gibbons, & Muraki, 1988), Bayesian estimation via Markov chain Monte Carlo (Béguin & Glas, 2001; Bolt & Lall, 2003), least squares method (McDonald, 1982), etc. There are also discussions of the dimension estimation of $K$ (Kang, 2006; Svetina & Levy, 2012).

Another related literature is the confirmatory analysis (Mckinley, 1989), for which each item is known to be associated with a subset of the available latent traits. Under this context, the item–trait association is specified by the matrix $\Lambda$. Confirmatory analysis based on a multidimensional IRT model is one of the nonlinear versions of the confirmatory factor analysis that is initially proposed by Jöreskog (1969). Typical confirmatory analysis assumes that $\Lambda$ is known and that the item parameters are estimated given a pre-specified $\Lambda$.

In the current analysis, each item is associated with only a subset of the latent traits, which implicitly requires certain practical interpretability of each latent trait. Unlike the usual exploratory analysis, for which a nondegenerate rotation on $\boldsymbol{\theta}$ yields a mathematically equivalent model, we will impose certain constraints during the parameter estimation and the estimates are not rotation invariant. These constraints are based on empirical knowledge of the items and may affect the final result. We recommend researchers to consider different constraints and to make comparisons among them. There is some early work attempting to address similar issues. For instance, Ackerman (1994) suggested that one could understand how the latent traits were measured by test items by fitting the test data with a multidimensional IRT model from the graphical perspective. However, to the authors' best knowledge, estimation of the $\Lambda$-matrix via a latent variable selection framework has not yet been formally discussed in the literature. The study in this paper fills in this void.

The dependence of the responses on the latent traits falls into the form of a generalized linear model, where $\boldsymbol{\theta}$ plays the role of covariates while $\mathbf{a}_j$ is the vector of regression coefficients. Of interest is the estimation of the nonzero elements of the $a$-vector, which corresponds to a latent variable selection problem. There is a rich statistics literature on variable and model selection. Various information criteria have been proposed under the regression context including AIC, BIC, DIC, $C_{\mathrm{p}}$, etc. (Akaike, 1974; Schwarz, 1978; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002; Mallows, 1973). One issue in the application of these information criteria lies in computation. Suppose that there are $J$ item and $K$ attributes. Then, there are in total $J \times K$ $a$-coefficients. In order to minimize the information criteria, one has to evaluate them over $2^{J\times K}$ models, which typically induces a substantial computation overhead. Due to this concern, we consider a different approach that is the $L_1$ regularization. The $L_1$-regularized regression is originally introduced for linear models. It is also known as the least absolute shrinkage and selection operator (Tibshirani, 1996). It receives much attention for solving the variable selection problems for both linear and generalized linear models (for instance, Friedman, Hastie, Hofling, & Tibshirani, 2007; Friedman, Hastie, & Tibshirani, 2010).

One advantage of the $L_1$-regularized regression over the information criterion approach is that its computation is much more tractable than that of the latter. From the methodological point of view, the current problem is different from the traditional variable selection for generalized linear models. In the regression formulation, the latent traits play the role of covariates. Under the regular variable selection setting, covariates are fully observed. For the current problem, the latent traits are not directly observed and therefore the $L_1$ penalty is applied to the marginal log-likelihood of the observed data with $\boldsymbol{\theta}$ being integrated out. As for the corresponding computation, we apply the expectation–maximization (EM) algorithm treating $\boldsymbol{\theta}$ as the missing data.

The rest of the paper is organized as follows. Section 2 includes the specification of the multidimensional IRT models for dichotomous responses, the estimation of the $\Lambda$-matrix via a regularized estimator, and the corresponding computation. Simulation studies are included in Sect. 3 illustrating the performance of the proposed method. Real data analysis is provided in Sect. 4. A concluding remark is provided in Sect. 5.

## 2. Latent Variable Selection via $L_1$-Regularized Regression

### 2.1. Compensatory Multidimensional IRT Models

Consider a test containing $J$ items and $K$ latent traits. The traits are represented by a $K$-dimensional vector $\boldsymbol{\theta} \in R^K$, the $K$-dimensional Euclidean space. Each subject responds to all $J$ items. For the present discussion, all responses are dichotomous. Denote the response to item $j$ by $y_j$. A compensatory two-parameter multidimensional model is

$$P\left(y_j = 1|\boldsymbol{\theta},\, \mathbf{a}_j,\, b_j\right) = F\left(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta} + b_j\right), \tag{1}$$

where $F\colon R \to [0,\, 1]$ is a pre-specified nondecreasing function, $\mathbf{a}_j = (a_{j1}, \ldots, a_{jK})^{\mathrm{T}}$ and $b_j$ are the item-specific parameters. In particular, $\mathbf{a}_j$ is the discrimination parameter vector and $b_j$ is the difficulty parameter. We define

$$\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_J), \quad \mathbf{b} = (b_1, \ldots, b_J)^{\mathrm{T}}.$$

Furthermore, the local independence among the responses is assumed, that is, conditional on $\boldsymbol{\theta}$, $y_1, \ldots, y_J$ are jointly independent.

There are two popular choices for $F$: the normal ogive model and the logistic model. For the normal ogive model, $F$ is chosen to be the cumulative distribution function of the standard normal distribution (Bock et al., 1988), that is, $P(y_j = 1|\boldsymbol{\theta},\, \mathbf{a}_j,\, b_j) = \int_{-\infty}^{\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta}+b_j} (2\pi)^{-1/2}\mathrm{e}^{-\frac{u^2}{2}}\,\mathrm{d}u$. For the logistic model, $F$ is chosen to be the logistic function (McKinley & Reckase, 1982), that is,

$$P\left(y_j = 1|\boldsymbol{\theta},\, \mathbf{a}_j,\, b_j\right) = \frac{\exp\left(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta} + b_j\right)}{1 + \exp\left(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta} + b_j\right)}. \tag{2}$$

We also consider the three-parameter model that further includes a guessing probability $c_j$, that is,

$$P\left(y_j = 1|\boldsymbol{\theta},\, \mathbf{a}_j,\, b_j,\, c_j\right) = c_j + \left(1 - c_j\right) F\left(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{a}_j + b_j\right). \tag{3}$$

By setting $c_j = 0$, the above model recovers (1). In addition, the latent trait vector $\boldsymbol{\theta}$ follows the multivariate normal prior distribution with zero-mean vector and covariance matrix $\Sigma$ that is assumed to be known in most discussions. Further discussion on the case of unknown $\Sigma$ will also be provided.

### 2.2. Latent Variable Selection via $L_1$ Regularization

*2.2.1. Estimation via $L_1$ Regularization for Two-Parameter Models* As stated in the Sect. 1, we consider the matrix $\Lambda = (\lambda_{jk})_{J\times K}$ that is defined as

$$\lambda_{jk} = I\left(a_{jk} \neq 0\right). \tag{4}$$

Suppose that the responses of $N$ examinees have been collected. Let $\boldsymbol{\theta}_i$ be the latent trait of examinee $i$, $\mathbf{Y} = (y_{ij})_{N\times J}$ denote the data, $y_{ij}$ be the response of the examinee $i$ to item $j$, and $Y_i = (y_{i1}, \ldots, y_{iJ})$ be the vector of responses of examinee $i$. The latent traits $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ are independently and identically distributed following the prior distribution $N(0,\ \Sigma)$ whose density is denoted by $\varphi(\boldsymbol{\theta})$. Conditional on $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$, $Y_1, \ldots, Y_N$ follow the two-parameter model (1) and admit the local independence assumption among items. Then, the complete data [i.e., observed data $\mathbf{Y}$ and missing data $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$] likelihood for the two-parameter IRT model is

$$L(\mathbf{A},\ \mathbf{b};\ \mathbf{Y},\ \boldsymbol{\Theta}) = \prod_{i=1}^{N} \varphi\left(\boldsymbol{\theta}_i\right) \prod_{j=1}^{J} \left[F\left(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i + b_j\right)^{y_{ij}} \left[1 - F\left(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i + b_j\right)\right]^{1-y_{ij}}\right]. \tag{5}$$

Furthermore, the log-likelihood of the observe data $\mathbf{Y}$ is given by

$$l(\mathbf{A},\ \mathbf{b};\ \mathbf{Y}) = \log\left[\int_{\boldsymbol{\Theta}\in\mathbf{R}^{K\times N}} L(\mathbf{A},\ \mathbf{b};\ \mathbf{Y},\ \boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}\right]. \tag{6}$$

In the exploratory factor analysis, one maximizes the log-likelihood function and obtains the maximum likelihood estimator

$$(\tilde{\mathbf{A}},\ \tilde{\mathbf{b}}) = \arg\max_{\mathbf{A},\,\mathbf{b}} l(\mathbf{A},\ \mathbf{b};\ \mathbf{Y}).$$

The maximum likelihood estimator does not directly serve the purpose of variable selection. We further consider the $L_1$-regularized estimator

$$\left(\hat{\mathbf{A}}_\eta,\ \hat{\mathbf{b}}_\eta\right) = \arg\max_{\mathbf{A},\,\mathbf{b}} \left\{l(\mathbf{A},\ \mathbf{b};\ \mathbf{Y}) - \eta\|\mathbf{A}\|_1\right\}, \tag{7}$$

where $\eta > 0$ and

$$\|\mathbf{A}\|_1 = \sum_{j=1}^{J}\sum_{k=1}^{K} \left|a_{jk}\right|.$$

The regularization parameter $\eta$ controls sparsity. By choosing $\eta = 0$, the $L_1$-regularized estimator $(\hat{\mathbf{A}}_\eta,\ \hat{\mathbf{b}}_\eta)$ recovers the maximum likelihood estimator that almost surely contains all nonzero estimates of the $a$-coefficients and it corresponds to no sparsity. On the other hand, by choosing $\eta$ sufficiently large (for instance, $\eta = \infty$), the corresponding estimate of the discrimination parameters are $\hat{\mathbf{A}}_\infty = 0$. In this case, any nonzero discrimination parameter $a_{jk}$ would make the penalized log-likelihood negative infinity. Thus, $\eta = \infty$ corresponds to complete sparsity. Generally speaking, the regularization parameter $\eta$ controls the sparsity and large values of $\eta$ lead to more sparse estimates of $\mathbf{A}$. We hope to find an appropriate $\eta \in (0,\ +\infty)$, under which the zero patterns of $\hat{A}_\eta$ are consistent with the true loading structure.

*2.2.2. Choice of Regularization Parameter $\eta$* We apply the Bayesian information criterion (Schwarz, 1978) to choose the sparsity parameter $\eta$. In particular, each $\eta$ results in an estimated matrix $\Lambda$ that further corresponds to a BIC value. Then, we choose the parameter $\eta$ that leads to the smallest BIC value. More precisely, we let $\Lambda(\mathbf{A})$ be the incidence matrix corresponding to the nonzero pattern of the coefficient matrix $\mathbf{A}$ according to (4). For each matrix $\Lambda_*$, the Bayesian information criterion is defined as

$$\text{BIC}_{\Lambda_*} = -2 \max_{\Lambda(\mathbf{A})=\Lambda_*, \mathbf{b}} l(\mathbf{A}, \mathbf{b}; \mathbf{Y}) + \|\mathbf{A}\|_0 \log N. \tag{8}$$

The above maximized likelihood is subject to the constraint that $\mathbf{A}$ is consistent with the matrix $\Lambda_*$, that is, $\Lambda(\mathbf{A}) = \Lambda_*$. Furthermore, the notation $\|\mathbf{A}\|_0 = \sum_{j,k} I(a_{jk} \neq 0)$ is the $L_0$ norm.

The regularization parameter is chosen as follows. For each $\eta$, we first obtain the estimate $(\hat{\mathbf{A}}_\eta, \hat{\mathbf{b}}_\eta)$ via (7). Next, we obtain from $\hat{\mathbf{A}}_\eta$ an estimated matrix $\Lambda_\eta = \Lambda(\hat{\mathbf{A}}_\eta)$. We fit the multidimensional two-parameter IRT model based on $\Lambda_\eta$ without penalty and compute the BIC value as in (8), denoted by $\text{BIC}_{\Lambda_\eta}$. The regularization parameter $\eta$ is chosen to be the one admitting the smallest BIC value, that is,

$$\eta_* = \arg\min_{\eta} \text{BIC}_{\Lambda_\eta}.$$

*Remark 1.* To guarantee parameter identifiability, some constraints need to be imposed on the item parameters. As summarized by Béguin and Glas (2001), there are typically two ways to impose constraints. One is to set $a_{jk} = 0$ for $j = 1, \ldots, K-1$ and $k = j+1, \ldots, K$ (Fraser & McDonald, 1988), which is similar to the constraint of Jöreskog (1969). The other is to set $a_{jj} = 1$ and $a_{jk} = 0$ for $j = 1, \ldots, K$, $k = 1, \ldots, K$, and $j \neq k$. Note that for the former constraint, rotating the parameter space is usually necessary for the interpretation of the factor patterns (Bolt & Lall, 2003; Cai, 2010).

In this paper, we adopt a similar approach as the second. In particular, each of the first $K$ items is associated with only one trait, that is $a_{ii} \neq 0$ and $a_{ij} = 0$, for $1 \leq i \neq j \leq K$. This corresponds to the fact that a sub-matrix of $\Lambda$ is known to be the identity matrix (after appropriate reordering of the rows and the columns), but the coefficients $a_{ii}$'s are not necessarily unity. We further restrict the variances of $\boldsymbol{\theta}_i$ to be unity.

In practice, one may impose different constraints on $\mathbf{A}$ or $\Lambda$ to ensure identifiability. In the simulation study and the real data analysis, we experimented two different sets of constraints and found that the results are similar. The second constraint is as follows. We identify $K$ items (e.g., the first $K$ items) and let $a_{ii} \neq 0$ for $i = 1, \ldots, K$. Unlike the first constraint, we do not force $a_{ij}$ $(i \neq j)$ to be zero. Rather, we impose $L_1$ penalties on them. Thus, the penalty includes all elements in $\mathbf{A}$ except for $a_{ii}$ for $i = 1, \ldots, K$.

In practice, the constraint on $\Lambda$ relies on a priori knowledge of the items and the entire study. It is usually formulated to meet specific needs. For instance, if we want to define a factor (a skill or a mental disorder) by an item or multiple items, then these items are naturally included in the constraints. We would like to raise a warning for readers that inappropriate constraints on $\Lambda$ (equivalent, identifying wrong items for each trait) may lead to misleading or noninterpretable results. We recommend trying different constraints, checking if the results are consistent, and selecting the most sensible one.

*2.2.3. Three-Parameter Models* We now consider the multidimensional three-parameter model that is often employed to account for the possibility of guessing. The latent variable selection method proposed for the two-parameter model (1) can be generalized to the model in (3). An $L_1$-regularized estimator can be obtained as

$$(\hat{\mathbf{A}},\ \hat{\mathbf{b}},\ \hat{\mathbf{c}}) = \arg\max_{\mathbf{A},\,\mathbf{b},\,\mathbf{c}} \{l(\mathbf{A},\ \mathbf{b},\ \mathbf{c};\ \mathbf{Y}) - \eta\|\mathbf{A}\|_1\}, \tag{9}$$

where $l(\mathbf{A},\ \mathbf{b},\ \mathbf{c};\ \mathbf{Y})$ is the marginal log-likelihood based on the observed data $\mathbf{Y}$.

Our empirical findings show that the estimator in (9) does not perform stably. This is mostly due to the introduction of the guessing parameter $\mathbf{c}$ that is difficult to estimate accurately. In the simulation studies in Sect. 3, we assume that the guessing parameters for all items are known. From the practical point of view, if the items are multiple-choice exam problems, the guessing parameter may be set to one over the number of choices. We emphasize that the estimation problem of the guessing parameter is not peculiar to our method. It turns out to be a general issue for the three-parameter multidimensional IRT models. For example, the softwares Noharm (Fraser & McDonald, 1988) and Testfact (Bock et al., 2003) both require specifying the guessing parameters for the estimation of the multidimensional three-parameter IRT models. Therefore, future investigations are needed along this line and we will thus further improve our regularized estimator piggybacking on the improvement of the estimation of the guessing parameter $\mathbf{c}$.

*2.2.4. On the Correlation Among the Traits $\boldsymbol{\theta}$* The regularized estimator is introduced assuming that the covariance matrix of $\boldsymbol{\theta}$ is known. In case that $\Sigma$ is unknown, we suggest two treatments. One is to consider $\Sigma$ as an additional parameter in the specification of the log-likelihood function (6) and maximize it together with $\mathbf{A}$. This approach typically induces additional computation. In the subsequent analysis, we consider a second approach that estimates $\Sigma$ through an exploratory analysis (without regularization on the parameter $\mathbf{A}$) under the constraints in Remark 1, with which $\Sigma$ can be uniquely identified. Then, we rescale the variances in $\hat{\Sigma}$ to be unity, treat it as true, and proceed to the regularized estimator (7).

### *2.3. Computation via Expectation–Maximization and Coordinate Descent Algorithm*

In this section, we proceed to the computation of the estimators in (7) for a given sparsity parameter $\eta$. Notice that implementation in maximizing the regularized likelihood is not straightforward. We apply the EM algorithm (Dempster, Laird, & Rubin, 1977) that is developed to compute the maximum likelihood estimator or posterior mode in the presence of missing data. The EM algorithm is an iterative algorithm. Each iteration consists of two steps. The $E$-step computes the expected log-likelihood with respect to the posterior distribution of the missing data and the $M$-step maximizes the expected log-likelihood computed from the $E$-step. Adapted to our particular problem, the $E$-step is not in a closed form and we compute the expected log-likelihood via numerical approximation. For the $M$-step, we use coordinate descent that is developed for the computation of $L_1$-regularized estimators for generalized linear models (Friedman et al., 2010). More detailed computation scheme is described as follows.

*2.3.1. E-Step* Let $(\mathbf{A}^{(t)},\ \mathbf{b}^{(t)})$ be the parameter values at the $t$th iteration. In order to evolve to the $(t+1)$th iteration, one first computes the expected complete data log-likelihood with respect to the posterior distribution

$$Q\left(\mathbf{A},\ \mathbf{b}|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)}\right) = E\left[\log\{L(\mathbf{A},\ \mathbf{b};\ \mathbf{Y},\ \boldsymbol{\Theta})\}|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)},\ \mathbf{Y}\right],$$

where $L(\mathbf{A},\ \mathbf{b};\ \mathbf{Y},\ \boldsymbol{\Theta})$ is defined as in (5). The above expectation $E\{\cdot|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)},\ \mathbf{Y}\}$ is taken with respect to $\boldsymbol{\Theta}$ under the posterior distribution

$$p\left(\boldsymbol{\Theta}|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)},\ \mathbf{Y}\right) \propto L\left(\mathbf{A}^{(t)},\ \mathbf{b}^{(t)};\ \mathbf{Y},\ \boldsymbol{\Theta}\right). \tag{10}$$

The posterior expectation in the definition of the $Q$-function is not in a closed form. We evaluate $Q$ numerically as follows. First, we write

$$Q\left(\mathbf{A},\ \mathbf{b}|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)}\right)=\sum_{j=1}^{J}Q_j\left(\mathbf{a}_j,\ b_j|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)}\right),$$

where

$$\begin{aligned}&Q_j\left(\mathbf{a}_j,\ b_j|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)}\right)\\&\quad=\sum_{i=1}^{N}E\left[y_{ij}\log\left\{F\left(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i+b_j\right)\right\}+\left(1-y_{ij}\right)\log\left\{1-F\left(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i+b_j\right)\right\}\Big|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)},\ Y_i\right].\end{aligned}$$

The $\boldsymbol{\theta}_i$'s are independent under the posterior distribution that is given by

$$p\left(\boldsymbol{\theta}_i|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)},\ Y_i\right)\propto\prod_{j=1}^{J}F\left(\left(\mathbf{a}_j^{(t)}\right)^{\mathrm{T}}\boldsymbol{\theta}_i+b_j\right)^{y_{ij}}\left[1-F\left(\left(\mathbf{a}_j^{(t)}\right)^{\mathrm{T}}\boldsymbol{\theta}_i+b_j\right)\right]^{1-y_{ij}}\varphi\left(\boldsymbol{\theta}_i\right).$$

We approximate this integration by a summation. More precisely, we consider grid points $\mathcal{G}\subseteq[-4,\ 4]^K$ and approximate the posterior distribution by

$$\begin{aligned}&\hat{p}\left(\boldsymbol{\theta}_i|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)},\ Y_i\right)\\&\quad\propto\begin{cases}\prod_{j=1}^{J}F((\mathbf{a}_j^{(t)})^{\mathrm{T}}\boldsymbol{\theta}_i+b_j)^{y_{ij}}[1-F((\mathbf{a}_j^{(t)})^{\mathrm{T}}\boldsymbol{\theta}_i+b_j)]^{1-y_{ij}}g(\boldsymbol{\theta}_i) & \text{if }\boldsymbol{\theta}_i\in\mathcal{G},\\ 0 & \text{otherwise,}\end{cases}\end{aligned}$$

and $\sum_{\boldsymbol{\theta}\in\mathcal{G}}\hat{p}(\boldsymbol{\theta}|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)},\ Y_i)=1$. Thus, $Q_j$ is approximated by

$$\begin{aligned}&\hat{Q}_j\left(\mathbf{a}_j,\ b_j|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)}\right)\\&\quad=\sum_{i=1}^{N}\sum_{\boldsymbol{\theta}_i\in\mathcal{G}}\left[y_{ij}\log\left\{F\left(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i+b_j\right)\right\}+\left(1-y_{ij}\right)\log\left(1-F\left(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i+b_j\right)\right)\right]\\&\quad\hat{p}\left(\boldsymbol{\theta}_i|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)},\ Y_i\right).\end{aligned}$$

Thus, the $Q$-function is computed by

$$\hat{Q}\left(\mathbf{A},\ \mathbf{b}|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)}\right)=\sum_{j=1}^{J}\hat{Q}_j\left(\mathbf{a}_j,\ b_j|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)}\right).$$

We choose $\mathcal{G}$ to be $\mathcal{S}\times\cdots\times\mathcal{S}$, where $\mathcal{S}$ is the set of $M=21$ (for $K=3$) and 11 (for $K=4$) grid points on the interval $[-4,\ 4]$.

*2.3.2. M-Step* With the $Q$-function computed in the $E$-step, we further perform the $M$-step, that is,

$$\left(\mathbf{A}^{(t+1)},\ \mathbf{b}^{(t+1)}\right) = \arg\max_{\mathbf{A},\,\mathbf{b}} \left\{\hat{Q}\left(\mathbf{A},\ \mathbf{b}|\mathbf{A}^{(t)},\ \mathbf{b}^{(t)}\right) - \eta\|\mathbf{A}\|_1\right\}. \tag{11}$$

Notice that the function $\hat{Q}$ factorizes to the sum of $\hat{Q}_j$'s. Each $\hat{Q}_j$ is a function only of $\mathbf{a}_j$ and $b_j$. Then, the above maximization can be reduced to maximizing each $\hat{Q}_j$ separately, that is,

$$\left(\mathbf{a}_j^{(t+1)},\ b_j^{(t+1)}\right) = \arg\max_{\mathbf{a}_j,\,b_j} \left\{\hat{Q}_j\left(\mathbf{a}_j,\ b_j|\mathbf{a}_j^{(t)},\ b_j^{(t)}\right) - \eta\left\|\mathbf{a}_j\right\|_1\right\}. \tag{12}$$

The above maximization is of a much lower dimension than that of (11). It is straightforward to verify that $\mathbf{A}^{(t+1)} = (\mathbf{a}_j^{(t+1)}: 1 \le j \le J)$ and $\mathbf{b}^{(t+1)} = (b_j^{(t+1)}: 1 \le j \le J)$. For the optimization of the parameters for each item in (12), we use the coordinate descent algorithm that is developed by Friedman et al. (2010). The detailed algorithm is described in the appendix. The EM algorithm evolves according to (11) until convergence that is monitored by certain criterion. For the three-parameter model, the regularized estimators can be computed in a similar way via EM.

## 3. Simulation

In this section, we perform simulations to illustrate the performance of the proposed method under various settings. As the main objective of the study is the $\Lambda$-matrix, we mainly consider the correct estimation rate of the $\Lambda$-matrix that is defined as

$$\text{CR} = \frac{1}{K(J-K)} \sum_{K+1\le j<J,\ 1\le k\le K} I\left(\hat{\lambda}_{jk} = \lambda_{jk}\right), \tag{13}$$

where $\hat{\Lambda} = (\hat{\lambda}_{jk})$ is an estimate and $\Lambda$ is the true matrix. In what follows, we investigate the correct estimation rates of the $L_1$-regularized estimator with the regularization parameter $\eta$ chosen according to the Bayesian information criterion under various model settings. For the estimate of the $\mathbf{A}$-matrix, we consider the mean squared error for each entry.

### *3.1. Two-Parameter Logistic Model*

We generate samples from model (2) for $K = 3$ and 4, respectively. For $K = 3$, we consider two different $\mathbf{A}$-matrices given as in Tables 1 and 2, denoted by $\mathbf{A}_1$ and $\mathbf{A}_2$. We chose the $\mathbf{A}$-matrices so that they contain some single-, double-, and triple-attribute items. The difference between these two matrices is that the coefficients in $\mathbf{A}_1$ are larger and there are more single-trait items. Thus, $\mathbf{A}_1$ is considered as easier to estimate. For $K = 4$, the matrix $\mathbf{A}_3$ is given in Table 3. Furthermore, the latent traits $\boldsymbol{\theta}$ have variance one and a common correlation $\rho = 0.1$ and $\Sigma$ is considered as unknown. The difficulty parameters $b_j$ are all zero. For each $\mathbf{A}$-matrix, we generate 50 independent datasets of sample size $N = 2000$ to evaluate the frequentist properties of our estimator.

The parameters are estimated via the algorithm described as in Sect. 2.3 with the sparsity parameter $\eta$ chosen according to BIC as in Sect. 2.2. To ensure identifiability, we consider the following two sets of constraints on the parameters.

(1) We designate one item for each factor and this item is associated with only that factor. That is, we set sub-$\Lambda$-matrix corresponding to the $K$ items to be identity. This is the first constraint specified in Remark 1.

TABLE 1.
$\mathbf{A}_1$.

| Latent traits | Items | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 1.9 | 1.7 | 1.5 | 1.3 | 1.1 | 0.9 | 0.7 | 0.5 | 0.3 | 0.0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.9 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Latent traits | Items | | | | | | | | | |
| | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 1.7 | 1.5 | 1.3 | 1.1 | 0.9 | 0.7 | 0.5 | 0.3 | 0.0 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.9 | 1.7 |
| Latent traits | Items | | | | | | | | | |
| | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.5 | 0.7 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.1 | 1.3 | 1.5 |
| 3 | 1.5 | 1.3 | 1.1 | 0.9 | 0.7 | 0.5 | 0.3 | 0.0 | 0.0 | 0.0 |
| Latent traits | Items | | | | | | | | | |
| | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
| 1 | 0.9 | 1.1 | 1.3 | 0.0 | 0.0 | 0.0 | 0.3 | 0.5 | 0.7 | 0.9 |
| 2 | 0.0 | 0.0 | 0.0 | 0.5 | 0.7 | 0.9 | 1.1 | 1.3 | 1.5 | 1.7 |
| 3 | 0.7 | 0.9 | 1.1 | 1.3 | 1.5 | 1.7 | 1.9 | 1.9 | 1.1 | 0.5 |

(2) We designate one item for each factor. This item is associated with that factor for sure and may also be associated with others. That is, we set the diagonal elements of the sub-$\Lambda$-matrix corresponding to the $K$ items to be ones and off diagonal elements have no constraint. Technically, the $L_1$ penalty includes all coefficients except for $(a_{1,1},\ a_{10,2},\ a_{19,3})$ in the case of $\mathbf{A}_1$. Notice that this constraint is much weaker than the first one, nevertheless still ensures identifiability as long as it is correctly specified (due to the regularization on other coefficients).

We treat the covariance $\Sigma$ as unknown and estimate it via a constrained exploratory analysis as mentioned previously. The computation time of the estimator for the first $\eta$ is around 30 min. Once an estimate of $\mathbf{A}$ for some $\eta$ is obtained, it is used as starting points for the computation of other $\eta$'s. Each additional $\eta$ requires about 10 min.

To illustrate the performance, we investigate the correct estimation rates in (13) from different aspects. First, Fig. 1 shows the histograms of the correct estimation rates over the 50 independent datasets for $\mathbf{A}_1$ under constraints 1 and 2. The overall rates are well over 95 %. We also consider the mean squared error for each $a_{ij}$ and there are $40 \times 3 = 120$ MSE's in total whose histograms are also shown in Fig. 1.

As we mentioned, the regularization parameter is chosen to minimize the BIC value, denoted by $\eta_*$. Its correct estimation rate is denoted by $\mathrm{MR}_*$. As the true $\Lambda$-matrix is known, we can further choose $\eta$ to maximize the correct estimation rates that is denoted by $\mathrm{MR}_0$. The first plot in Fig. 2 shows the scatter plot of the pair $(\mathrm{MR}_*,\ \mathrm{MR}_0)$ for all 50 datasets. BIC is a reasonable criterion to select $\eta$ in terms of maximizing the correct estimation rate.

Furthermore, we investigate a dataset that is randomly chosen from the 50 simulated datasets to illustrate the performance of BIC in selecting the regularization parameters. We standardize the BIC values as a function of $\eta$ by some linear transformations such that it sits well in the same

TABLE 2.
$\mathbf{A}_2$.

| Latent traits | Items | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 1.2 | 1.0 | 0.8 | 0.6 | 0.4 | 0.2 | 0.1 | 0.0 | 0.0 | 0.0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.2 | 1.0 | 0.8 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Latent traits | Items | | | | | | | | | |
| | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 0.6 | 0.4 | 0.2 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 1.2 | 1.0 | 0.8 | 0.6 | 0.4 | 0.2 |
| Latent traits | Items | | | | | | | | | |
| | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| 1 | 0.0 | 0.8 | 0.6 | 0.4 | 0.2 | 0.8 | 0.6 | 0.4 | 0.2 | 0.8 |
| 2 | 0.0 | 0.2 | 0.8 | 0.6 | 0.4 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.8 | 0.6 | 0.4 |
| Latent traits | Items | | | | | | | | | |
| | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
| 1 | 0.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.8 | 0.6 | 0.4 | 0.2 |
| 2 | 0.0 | 0.6 | 0.4 | 0.2 | 0.8 | 0.6 | 0.2 | 0.4 | 0.6 | 0.8 |
| 3 | 0.2 | 0.8 | 0.6 | 0.4 | 0.2 | 0.1 | 0.8 | 0.6 | 0.4 | 0.2 |

TABLE 3.
$\mathbf{A}_3$.

| Latent traits | Items | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 1.5 | 1 | 0.5 | 0.0 | 0 | 0.0 | 0.0 | 0 | 0.0 | 0.0 |
| 2 | 0.0 | 0 | 0.0 | 1.5 | 1 | 0.5 | 0.0 | 0 | 0.0 | 0.0 |
| 3 | 0.0 | 0 | 0.0 | 0.0 | 0 | 0.0 | 1.5 | 1 | 0.5 | 0.0 |
| 4 | 0.0 | 0 | 0.0 | 0.0 | 0 | 0.0 | 0.0 | 0 | 0.0 | 1.5 |
| Latent traits | Items | | | | | | | | | |
| | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 1 | 0 | 0.0 | 0.5 | 0.5 | 0.5 | 0.0 | 0 | 0.0 | 0.5 | 0.0 |
| 2 | 0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1 | 0.0 | 1.0 | 1.5 |
| 3 | 0 | 0.0 | 0.0 | 1.5 | 0.0 | 1.5 | 0 | 1.5 | 1.5 | 1.0 |
| 4 | 1 | 0.5 | 0.0 | 0.0 | 0.5 | 0.0 | 1 | 1.5 | 0.0 | 0.5 |

plot as the mis-estimation rate that is the compliment of the correct estimation rate. The second plot in Fig. 2 shows BIC and the mis-estimation rate as a function of $\eta$ in the same plot. The BIC and mis-estimation curves both decrease first and then increase. The decreasing slope of the BIC curve is induced by the $\log N$ penalty. The minima of both curves coincide suggesting that BIC is a good criterion for selecting $\eta$.

The correct estimation rates of $\Lambda$ and MSE of $\mathbf{A}_2$ are given in Fig. 3. The correct estimation rates are lower (still mostly over 90 %) because the magnitude of the coefficients are smaller. The corresponding results for the four-dimensional case $\mathbf{A}_3$ is given in Fig. 4. The results are similar.

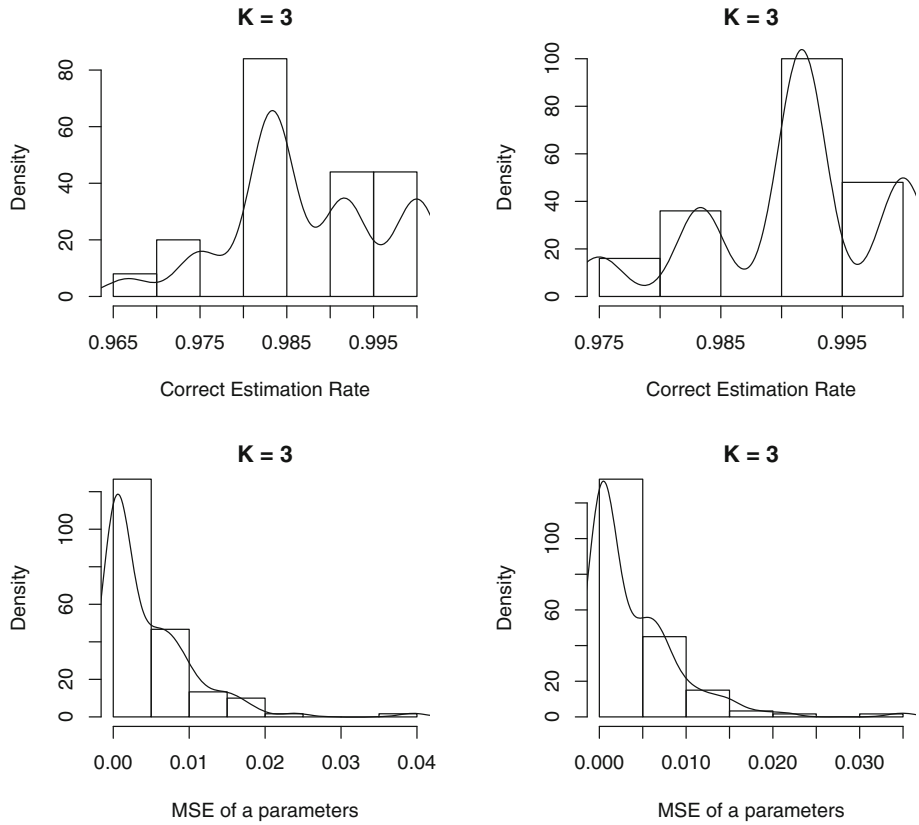


FIGURE 1.
Histograms of the correct rates for $\Lambda$ (*row 1*) and MSE of the estimate of the $a$ parameters for $\mathbf{A}_1$ (*row 2*) under constraint 1 (*left column*) and constraint 2 (*right column*).

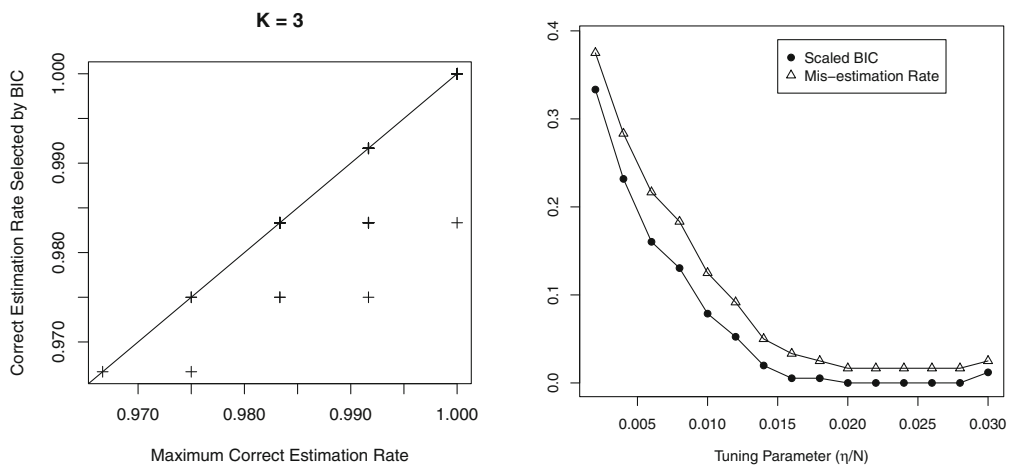


FIGURE 2.
*Left* comparing the correct estimation rates selected by BIC and the optimal rates. *Right* mis-estimation rates and BIC against $\eta$.

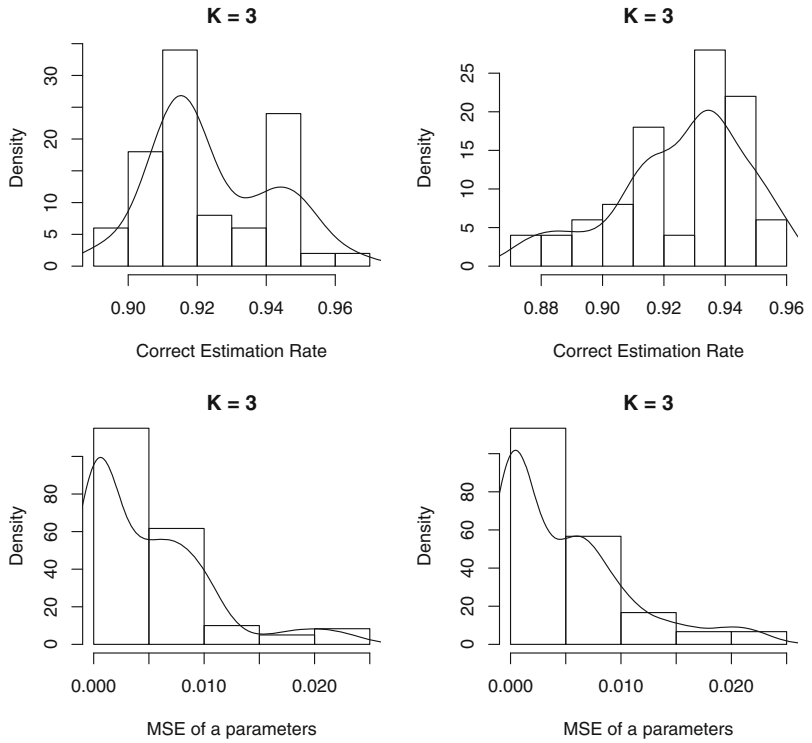


FIGURE 3.
Histograms of the correct rates for $\Lambda$ (*row 1*) and MSE of the estimate of the $a$ parameters for $\mathbf{A}_2$ (*row 2*) under constraint 1 (*left column*) and constraint 2 (*right column*).

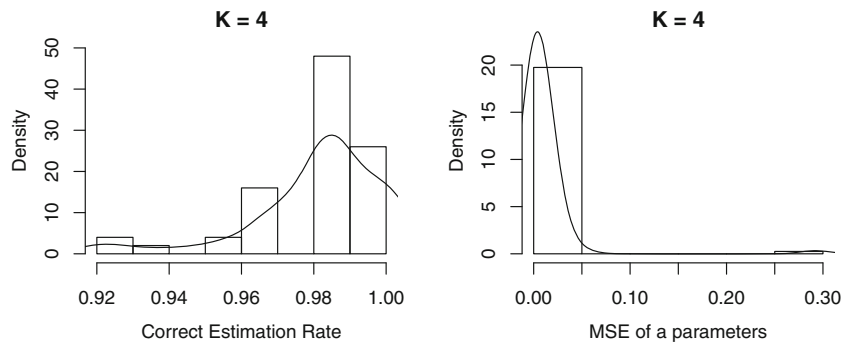


FIGURE 4.
Histograms of the correct rates for $\Lambda$ (*row 1*) and MSE of the estimate of the $a$ parameters for $\mathbf{A}_3$ (*row 2*) under constraint 1 (*left column*) and constraint 2 (*right column*).

### *3.2. Three-Parameter Logistic Model*

For the three-parameter model (3), we only consider the cases that $K = 2$. The $\mathbf{A}$-matrix is given as in Table 4. The guessing parameters are set to be 0.1 and known. The rest of the setting is similar to that of the simulation for the two-parameter model. The simulation results are summarized in Fig. 5.

TABLE 4.
$\mathbf{A}_4$.

| Latent traits | Items | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 1.7 | 1.5 | 1.3 | 1.1 | 0.9 | 0.7 | 0.5 | 0.3 | 0.0 | 0.0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.7 | 1.5 |
| Latent traits | Items | | | | | | | | | |
| | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.5 | 0.7 | 0.9 |
| 2 | 1.3 | 1.1 | 0.9 | 0.7 | 0.5 | 0.3 | 1.7 | 1.5 | 1.3 | 1.1 |

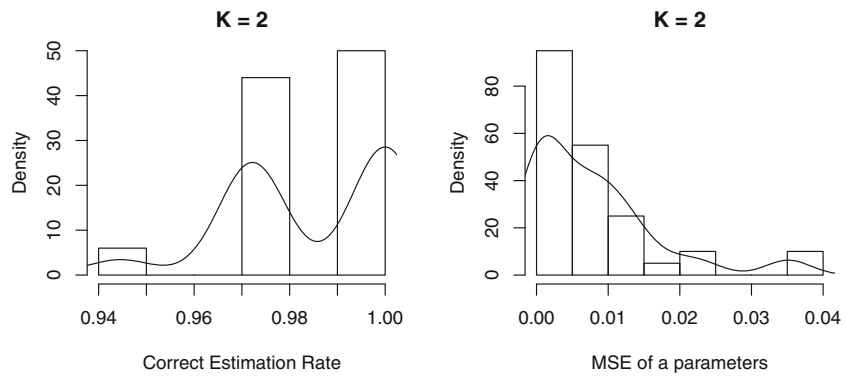


FIGURE 5.
Histograms of the correct rates for $\Lambda$ and MSE of the estimate of the $a$ parameters under constraint 1 for the three-parameter logistic model.

## 4. Real Data Analysis

The study records 824 females' responses to the revised Eysenck Personality Questionnaire short scales. There are in total 36 items. Based on Eysenck's idea, there are three factors: P (Psychoticism), E (Extraversion), and N (Neuroticism) scales. The study was originally a confirmatory analysis; see the initial work of Eysenck and Barrett (2013) and subsequent analysis by Maydeu-Olivares and Liu (2015). In the pre-specified $\Lambda$-matrix of the confirmatory analysis, each item is associated with only one factor. In particular, items 1–12 are associated with "Psychoticism," items 13–24 to "Extraversion," items 25–36 to "Neuroticism." The specific questions are presented in Table 5. Furthermore, the dataset has been preprocessed so that the negatively worded items have already been reversely scored (marked by "R" in Table 5). Thus, "yes" to a question is coded as "0" if the question has been reversed.

In the analysis, we impose two sets of different constraints on $\Lambda$ to ensure identifiability. They eventually lead to similar results.

(1) We designate two items for each factor and these two items are associated with only that factor. In particular, for "Psychoticism," we select items 1 and 2 and set rows 1 and 2 of $\Lambda$ to be (1, 0, 0), for "Extraversion," we set rows 13 and 14 of $\Lambda$ to be (0, 1, 0), for "Neuroticism," we set rows 25 and 26 of $\Lambda$ to be (0, 0, 1).

TABLE 5.
The revised Eysenck Personality Questionnaire short scales.

| | |
|---|---|
| 1 | Would you take drugs which may have strange or dangerous effects? |
| 2 | Do you prefer to go your own way rather than act by the rules? |
| 3 | Do you think marriage is old fashioned and should be done away with? |
| 4 | Do you think people spend too much time safeguarding their future with savings and insurance? |
| 5 | Would you like other people to be afraid of you? |
| 6(R) | Do you take much notice of what people think? |
| 7(R) | Would being in debt worry you? |
| 8(R) | Do good manners and cleanliness matter much to you? |
| 9(R) | Do you enjoy co-operating with others? |
| 10(R) | Does it worry you if you know there are mistakes in your work? |
| 11(R) | Do you try not to be rude to people? |
| 12(R) | Is it better to follow society's rules than go your own way? |
| 13 | Are you a talkative person? |
| 14 | Are you rather lively? |
| 15 | Can you usually let yourself go and enjoy yourself at a lively party? |
| 16 | Do you enjoy meeting new people? |
| 17 | Do you usually take the initiative in making new friends? |
| 18 | Can you easily get some life into a rather dull party? |
| 19 | Do you like mixing with people? |
| 20 | Can you get a patty going? |
| 21 | Do you like plenty of bustle and excitement around you? |
| 22 | Do other people think of you as being very lively? |
| 23(R) | Do you tend to keep in the background on social occasions? |
| 24(R) | Are you mostly quiet when you are with other people? |
| 25 | Does your mood often go up and down? |
| 26 | Do you ever feel 'just miserable' for no reason? |
| 27 | Are you an irritable person? |
| 28 | Are your feelings easily hurt? |
| 29 | Do you often feel 'fed-up'? |
| 30 | Are you often troubled about feelings of guilt? |
| 31 | Would you call yourself a nervous person? |
| 32 | Are you a worrier? |
| 33 | Would you call yourself tense or 'highly-strung'? |
| 34 | Do you worry too long after an embarrassing experience? |
| 35 | Do you suffer from 'nerves'? |
| 36 | Do you often feel lonely? |

(2) We designate two items for each factor. These two items are associated with that factor for sure but possibly with others too. More specifically, for "Psychoticism," we select items 1 and 2 and set rows 1 and 2 of $\Lambda$ to be (1, ?, ?), for "Extraversion," we set rows 13 and 14 of $\Lambda$ to be (?, 1, ?), for "Neuroticism," we set rows 25 and 26 of $\Lambda$ to be (?, ?, 1). The question mark "?" means that this entry is to be estimated. Technically, we do not penalize the coefficients ($a_{1,1}$, $a_{2,1}$, $a_{12,2}$, $a_{13,2}$, $a_{25,3}$, $a_{26,3}$) and penalize all other $a_{ij}$'s.

We have also experimented with constraints on other items and the results are similar and we only report the results of the above selection. For the covariance matrix of $\boldsymbol{\theta}$, we estimate it by fitting a confirmatory model stated at the beginning of this section and treat it as known. The rescaled estimate (to variance one) is

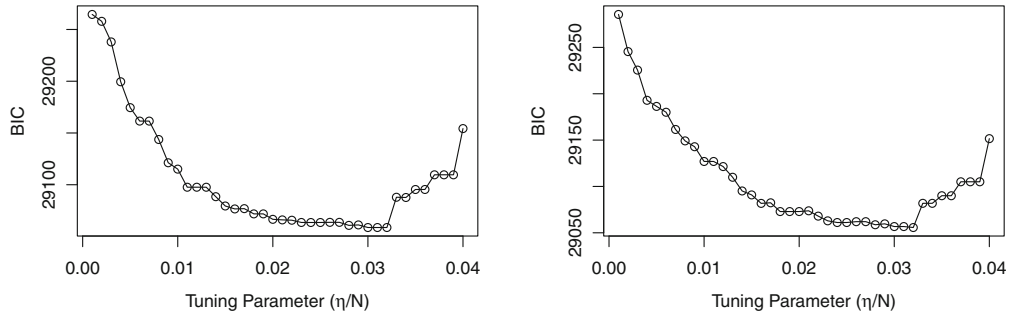


FIGURE 6.
The BIC and GIC values on the solution path for the EPQ-R data for constraint 1 (*left*) and constraint 2 (*right*).

TABLE 6.
The model selected by BIC in constraint 1.

| | | $A$ | | $b$ | | | $A$ | | $b$ | | | $A$ | | $b$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.50 | 0.00 | 0.00 | −2.58 | 13 | 0.00 | 1.89 | 0.00 | 1.14 | 25 | 0.00 | 0.00 | 1.50 | 1.10 |
| 2 | 1.44 | 0.00 | 0.00 | 0.55 | 14 | 0.00 | 2.55 | 0.00 | 2.01 | 26 | 0.00 | 0.00 | 1.10 | 0.70 |
| 3 | 1.23 | 0.00 | 0.00 | −2.42 | 15 | 0.00 | 1.57 | 0.00 | 1.53 | 27 | 0.39 | 0.00 | 1.29 | −0.83 |
| 4 | 0.74 | 0.00 | 0.00 | −0.89 | 16 | 0.00 | 1.71 | 0.00 | 3.10 | 28 | 0.00 | 0.00 | 1.27 | 1.24 |
| 5 | 1.09 | 0.00 | 0.00 | −3.07 | 17 | 0.00 | 1.40 | 0.00 | 0.61 | 29 | 0.00 | 0.00 | 1.53 | 0.08 |
| 6 | 0.87 | 0.00 | −0.60 | −1.23 | 18 | 0.00 | 2.56 | 0.00 | −1.43 | 30 | 0.00 | 0.32 | 1.28 | −0.04 |
| 7 | 0.96 | 0.27 | −0.37 | −2.49 | 19 | 0.00 | 1.76 | 0.00 | 3.20 | 31 | 0.00 | −0.27 | 2.03 | −1.12 |
| 8 | 1.13 | 0.00 | 0.00 | −2.26 | 20 | 0.00 | 2.19 | 0.00 | −0.60 | 32 | −0.66 | 0.00 | 2.27 | 0.94 |
| 9 | 1.20 | −0.63 | 0.00 | −2.89 | 21 | 0.00 | 1.23 | 0.00 | 1.18 | 33 | 0.00 | 0.00 | 1.84 | −1.46 |
| 10 | 0.87 | 0.00 | 0.00 | −1.96 | 22 | 0.00 | 2.34 | 0.00 | 0.92 | 34 | −0.66 | −0.21 | 1.44 | 0.69 |
| 11 | 1.45 | 0.00 | 0.00 | −3.09 | 23 | 0.53 | 2.62 | 0.00 | 0.68 | 35 | 0.00 | 0.00 | 2.05 | −1.15 |
| 12 | 1.28 | 0.00 | 0.00 | −0.16 | 24 | 0.00 | 2.21 | 0.00 | 1.29 | 36 | 0.00 | 0.00 | 1.19 | −0.99 |

$$\hat{\Sigma} = \begin{pmatrix} 1.00 & 0.11 & -0.03 \\ 0.11 & 1.00 & -0.25 \\ -0.03 & -0.25 & 1.00 \end{pmatrix}.$$

For each set of constraints, we compute BIC for the regularization parameter for $\eta \in [0.00, 0.04]$. The plots of the BIC values against $\eta$ are shown in Fig. 6. For constraint 1, the BIC selects $\lambda = 0.030$ and the coefficients are shown in Table 6. For constraint 2, the BIC selects $\lambda = 0.032$ and the estimated coefficients are shown in Table 7.

The nonzero patterns of the $a$-coefficients in both tables are very similar and so it is with their estimated values. In fact, constraint 1 is inconsistent with Table 7 on items 1, 13, and 25 that are forced to be single-trait items. But, the results on other unconstrained items are similar. This also illustrates the robustness of the current method. According to the **A**-matrix, most items remain associated with a single trait. There are some associated with more than one trait. We examined those items and found that most of them are very sensible. For instance, item 6 "Do you take much notice of what people think?" (a reverse question) is also related "Neuroticism" that is symboled by anxiety, fear, worry, etc., and its wording is similar to those of items 32 and 34, for item 9 "Do you enjoy co-operating with others?" (a reverse question, originally designed for "Psychoticism"), there is a good reason to believe that it is associated with "Extraversion,"

TABLE 7.
The model selected by BIC in constraint 2.

| | | $A$ | | $b$ | | | $A$ | | $b$ | | | $A$ | | $b$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.54 | 0.00 | 0.51 | −2.68 | 13 | 0.00 | 2.04 | 0.37 | 1.17 | 25 | 0.00 | 0.28 | 1.63 | 1.13 |
| 2 | 1.44 | 0.00 | 0.00 | 0.56 | 14 | 0.00 | 2.54 | 0.00 | 2.01 | 26 | 0.00 | 0.00 | 1.13 | 0.71 |
| 3 | 1.21 | 0.00 | 0.00 | −2.41 | 15 | 0.00 | 1.56 | 0.00 | 1.53 | 27 | 0.39 | 0.00 | 1.32 | −0.84 |
| 4 | 0.74 | 0.00 | 0.00 | −0.89 | 16 | 0.00 | 1.71 | 0.00 | 3.11 | 28 | 0.00 | 0.00 | 1.27 | 1.24 |
| 5 | 1.06 | 0.00 | 0.00 | −3.05 | 17 | 0.00 | 1.40 | 0.00 | 0.61 | 29 | 0.00 | 0.00 | 1.54 | 0.08 |
| 6 | 0.87 | 0.00 | −0.58 | −1.23 | 18 | 0.00 | 2.51 | 0.00 | −1.42 | 30 | 0.00 | 0.32 | 1.27 | −0.04 |
| 7 | 0.96 | 0.28 | −0.32 | −2.49 | 19 | 0.00 | 1.75 | 0.00 | 3.20 | 31 | 0.00 | −0.26 | 1.98 | −1.11 |
| 8 | 1.13 | 0.00 | 0.00 | −2.26 | 20 | 0.00 | 2.16 | 0.00 | −0.60 | 32 | −0.66 | 0.00 | 2.22 | 0.93 |
| 9 | 1.20 | −0.64 | 0.00 | −2.88 | 21 | 0.00 | 1.23 | 0.00 | 1.19 | 33 | 0.00 | 0.00 | 1.84 | −1.46 |
| 10 | 0.88 | 0.00 | 0.00 | −1.96 | 22 | 0.00 | 2.32 | 0.00 | 0.92 | 34 | −0.66 | −0.20 | 1.42 | 0.69 |
| 11 | 1.45 | 0.00 | 0.00 | −3.08 | 23 | 0.54 | 2.62 | 0.00 | 0.69 | 35 | 0.00 | 0.00 | 2.01 | −1.13 |
| 12 | 1.27 | 0.00 | 0.00 | −0.16 | 24 | 0.00 | 2.21 | 0.00 | 1.30 | 36 | 0.00 | 0.00 | 1.19 | −1.00 |

TABLE 8.
The estimated **A**-matrix of the exploratory analysis.

| | | $A$ | | | | $A$ | | | | $A$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.61 | 0.00 | 0.00 | 13 | 0.00 | 2.00 | 0.00 | 25 | 0.00 | 0.00 | 1.64 |
| 2 | 1.34 | 0.00 | 0.00 | 14 | 0.00 | 2.57 | 0.00 | 26 | 0.00 | 0.00 | 1.14 |
| 3 | 1.29 | −0.37 | 0.04 | 15 | 0.32 | 1.53 | −0.16 | 27 | 0.21 | −0.19 | 1.30 |
| 4 | 0.74 | 0.04 | −0.09 | 16 | 0.08 | 1.70 | −0.47 | 28 | −0.45 | −0.12 | 1.37 |
| 5 | 1.15 | −0.14 | 0.28 | 17 | 0.14 | 1.38 | −0.14 | 29 | −0.13 | −0.15 | 1.58 |
| 6 | 1.10 | −0.24 | −0.81 | 18 | 0.72 | 2.50 | −0.32 | 30 | −0.38 | 0.23 | 1.34 |
| 7 | 1.09 | 0.25 | −0.53 | 19 | −0.34 | 1.88 | −0.14 | 31 | −0.54 | −0.47 | 2.12 |
| 8 | 1.27 | −0.45 | −0.27 | 20 | 0.48 | 2.11 | −0.27 | 32 | −1.12 | −0.22 | 2.41 |
| 9 | 1.15 | −0.68 | 0.11 | 21 | 0.02 | 1.24 | −0.16 | 33 | −0.04 | −0.30 | 1.86 |
| 10 | 1.01 | 0.10 | −0.36 | 22 | 0.23 | 2.36 | −0.07 | 34 | −1.00 | −0.30 | 1.62 |
| 11 | 1.46 | 0.15 | 0.04 | 23 | 0.83 | 2.53 | −0.38 | 35 | −0.35 | −0.28 | 2.07 |
| 12 | 1.22 | −0.06 | −0.19 | 24 | 0.20 | 2.16 | −0.23 | 36 | −0.02 | −0.25 | 1.21 |

TABLE 9.
The $\Lambda$-matrix with threshold 0.5.

| | | $\Lambda$ | | | | $\Lambda$ | | | | $\Lambda$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 13 | 0 | 1 | 0 | 25 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 14 | 0 | 1 | 0 | 26 | 0 | 0 | 1 |
| 3 | 1 | 0 | 0 | 15 | 0 | 1 | 0 | 27 | 0 | 0 | 1 |
| 4 | 1 | 0 | 0 | 16 | 0 | 1 | 0 | 28 | 0 | 0 | 1 |
| 5 | 1 | 0 | 0 | 17 | 0 | 1 | 0 | 29 | 0 | 0 | 1 |
| 6 | 1 | 0 | 1 | 18 | 1 | 1 | 0 | 30 | 0 | 0 | 1 |
| 7 | 1 | 0 | 1 | 19 | 0 | 1 | 0 | 31 | 1 | 0 | 1 |
| 8 | 1 | 0 | 0 | 20 | 0 | 1 | 0 | 32 | 1 | 0 | 1 |
| 9 | 1 | 1 | 0 | 21 | 0 | 1 | 0 | 33 | 0 | 0 | 1 |
| 10 | 1 | 0 | 0 | 22 | 0 | 1 | 0 | 34 | 1 | 0 | 1 |
| 11 | 1 | 0 | 0 | 23 | 1 | 1 | 0 | 35 | 0 | 0 | 1 |
| 12 | 1 | 0 | 0 | 24 | 0 | 1 | 0 | 36 | 0 | 0 | 1 |

item 27 "Are you an irritable person?" is associated with both "Psychoticism" (aggressiveness) and "Neuroticism."

We further compare the above results to a classical method as follows. We first fit an exploratory model on the data under the constraint 1 without the $L_1$ penalty. The estimated coefficients are given in Table 8. We then set $a_{ij}$ to zero if $|a_{ij}| < \varepsilon_0$. Table 9 shows the $\Lambda$-matrix for $\varepsilon_0 = 0.5$ that yields the closest results to ours. We see that its basic pattern is similar to that of Table 7, but the $L_1$-regularized estimator does keep some low magnitude coefficients nonzero, such as items 7, 25, 30, and 34.

## 5. Concluding Remarks

A new method based on $L_1$ regularization is proposed for the latent variable selection for compensatory IRT models. The regularization parameter $\eta$ is chosen according to the Bayesian information criterion. Simulation studies show that the resulting estimator admits a good frequentist property in identifying the underlying latent structure $\Lambda$. The result of a real data analysis is also very sensible. We would like to provide some further remarks.

First, this paper, to the authors' best knowledge, is the first work estimating the confirmatory latent structure based on the data. The proposed method is an implementable and computationally affordable procedure admitting good properties for compensatory IRT models with binary responses. For other and more general model settings, such as polytomous responses, noncompensatory IRT models, etc., a similar $\Lambda$-matrix can be defined. The current method can be adapted to those models straightforwardly. The basic idea is to add a $L_1$ penalty term to the log-likelihood and to select the regularization parameter via BIC or other appropriate criteria. Certainly, further investigations (such as simulations) on the properties of the resulted estimators are necessary.

Second, the regularization parameter $\eta$ is selected according to BIC. As the simulation study shows, this approach performs well. There are different ways to select the regularization parameter other than BIC, such as splitting the entire datasets into training and testing data and using the out-sample prediction error as a criterion to select $\eta$. We do not pursue along this line because the computation of cross-validation is intensive.

Lastly, we empirically find that the correct estimation rate of the $\Lambda$-matrix depends very much on the estimation accuracy of the item parameters, especially the discrimination parameters. This partially explains why the estimator for $\Lambda$-matrix admits better performance for the two-parameter model than the three-parameter model.

## Acknowledgments

This research was funded by Fundamental Research Funds for the Central Universities (No. BLX2014-31), NSF grant SES-1323977, NSF grant IIS-1633360, Army Research Office grant W911NF-15-1-0159, NIH grant R01GM047845, National Natural Science Foundation of China (31371047; 11171029). We also would like to thank Dr. Paul Barrett for letting us use the EPQ-R data.

## Appendix

The cyclical coordinate descent algorithm for solving the optimization (12) is introduced as follows. For each item $j$, there are one difficulty parameter $b_j$ and $K$ discrimination parameters $\mathbf{a}_j = (a_{j1}, \ldots, a_{jK})$. The algorithm update each of the $K + 1$ variables iteratively according to the following updating rule. For the difficulty parameter, there is no $L_1$ penalty and it is updated by

$$\hat{b}_j = b_j^* - \frac{\partial_{b_j} \hat{Q}_j(\mathbf{a}_j,\ b_j^* | \mathbf{a}_j^{(t)},\ b_j^{(t)})}{\partial_{b_j}^2 \hat{Q}_j(\mathbf{a}_j,\ b_j^* | \mathbf{a}_j^{(t)},\ b_j^{(t)})},$$

where $\partial \hat{Q}_j$ denotes derivative of $\hat{Q}_j(\mathbf{a}_j,\ b_j | \mathbf{a}_j^{(t)},\ b_j^{(t)})$ with respect to $b_j$ or $a_{jk}$ as labeled by the subscript and $\partial^2 \hat{Q}_j$ is the second derivative. During the above update, the discrimination vector $\mathbf{a}_j$ takes its most up-to-date value. The above update employs a local quadratic approximation of $\hat{Q}_j(\mathbf{a}_j,\ b_j^* | \mathbf{a}_j^{(t)},\ b_j^{(t)})$ as a function of $b_j$ with all the other variables fixed. For each discrimination parameter $a_{jk}$, an $L_1$ penalty is imposed and it is updated by

$$\hat{a}_{jk} = -\frac{S(-\partial^2_{a_{jk}} \hat{Q}_j(\mathbf{a}_j,\ b_j^* | \mathbf{a}_j^{(t)},\ b_j^{(t)}) \times a_{jk}^* + \partial_{a_{jk}} \hat{Q}_j(\mathbf{a}_j,\ b_j^* | \mathbf{a}_j^{(t)},\ b_j^{(t)}),\ \eta)}{\partial^2_{a_{jk}} \hat{Q}_j(\mathbf{a}_j,\ b_j^* | \mathbf{a}_j^{(t)},\ b_j^{(t)})},$$

The function $S$ is the soft threshold operator (Donoho & Johnstone, 1995):

$$S(\delta,\ \eta) = \text{sign}(\delta)(|\delta| - \eta)_+ = \begin{cases} \delta - \eta, & \text{if } \delta > 0 \text{ and } \eta < |\delta|, \\ \delta + \eta, & \text{if } \delta < 0 \text{ and } \eta < |\delta|, \\ 0, & \text{if } \eta \geq |\delta|. \end{cases}.$$

To obtain the above updating rule, we approximate a generic univariate function $f(x)$ by a quadratic function

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2,$$

where $f''(x_0)$ is negative. Furthermore, the $L_1$-penalized maximization with the approximated function

$$\sup_x \left\{ f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 - \eta|x| \right\},$$

is solved at

$$-\frac{S(-f''(x_0)x_0 + f'(x_0),\ \eta)}{f''(x_0)}.$$

## References


Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, *13*, 113–127.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*, 255–278.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*, 37–48.

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–561.

Bock, D. R., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261–280.

Bock, D. R., Gibbons, R., Schilling, S., Muraki, E., Wilson, D., & Wood, R. (2003). Testfact 4.0. In *C*omputer software and manual. Lincolnwood, IL: Scientific Software International.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*, 395–414.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*, 33–57.

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

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, *90*, 1200–1224.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, *49*, 175–186.

Embretson, S. E., & Reise, S. P. (2000). *Psychometric methods: Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Eysenck, S., & Barrett, P. (2013). Re-introduction to cross-cultural studies of the EPQ. *Personality and Individual Differences*, *54*(4), 485–489.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*, 267–269.

Friedman, J., Hastie, T., Hofling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *1*, 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202.

Kang, T. (2006). *Model selection methods for unidimensional and multidimensional IRT models*. Madison, WI: University of Wisconsin.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics*, *15*, 661–675.

Maydeu-Olivares, A., & Liu, Y. (2015). Item diagnostics in multivariate discrete data. *Psychological Methods*, *20*, 276–292.

McDonald, R. P. (1967). *Nonlinear factor analysis*. Psychometric Monographs, No. 15. Richmond, VA: Psychometric Corporation.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, *6*, 379–396.

McKinley, R. L. (1989). *Confirmatory analysis of test structure using multidimensional item response theory*. Technical Report No. RR-89-31. Princeton, NJ: Educational Testing Service.

McKinley, R. L., & Reckase, M. D. (1982). *The use of the general Rasch model with multidimensional item response data*. Technical Report No. ONR-82-1. Iowa City, IA: American College Testing Program.

Reckase, M. D. (1972). Development and application of a multivariate logistic latent trait model. Unpublished Doctoral Dissertation, Syracuse University, Syracuse, NY.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25–36.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 583–639.

Svetina, D., & Levy, R. (2012). An overview of software for conducting dimensionality assessment in multidimensional models. *Applied Psychological Measurement*, *36*, 659–669.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82–98).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267–288.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement*, *12*, 239–252.

*Manuscript Received: 8 JUN 2014*
*Final Version Received: 22 MAR 2016*
*Published Online Date: 3 OCT 2016*