PSYCHOMETRIKA—VOL. 84, NO. 1, 124–146
MARCH 2019
https://doi.org/10.1007/s11336-018-9646-5

JOINT MAXIMUM LIKELIHOOD ESTIMATION FOR HIGH-DIMENSIONAL EXPLORATORY ITEM FACTOR ANALYSIS

YUNXIAO CHEN

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

XIAOOU LI

UNIVERSITY OF MINNESOTA

SILIANG ZHANG

FUDAN UNIVERSITY

Joint maximum likelihood (JML) estimation is one of the earliest approaches to fitting item response theory (IRT) models. This procedure treats both the item and person parameters as unknown but fixed model parameters and estimates them simultaneously by solving an optimization problem. However, the JML estimator is known to be asymptotically inconsistent for many IRT models, when the sample size goes to infinity and the number of items keeps fixed. Consequently, in the psychometrics literature, this estimator is less preferred to the marginal maximum likelihood (MML) estimator. In this paper, we re-investigate the JML estimator for high-dimensional exploratory item factor analysis, from both statistical and computational perspectives. In particular, we establish a notion of statistical consistency for a constrained JML estimator, under an asymptotic setting that both the numbers of items and people grow to infinity and that many responses may be missing. A parallel computing algorithm is proposed for this estimator that can scale to very large datasets. Via simulation studies, we show that when the dimensionality is high, the proposed estimator yields similar or even better results than those from the MML estimator, but can be obtained computationally much more efficiently. An illustrative real data example is provided based on the revised version of Eysenck's Personality Questionnaire (EPQ-R).

Key words: joint maximum likelihood estimator, item response theory, IRT, high-dimensional data, alternating minimization, projected gradient descent, personality assessment.

1. Introduction

Exploratory Item Factor Analysis (IFA; Bock et al. 1988) has been widely used as an analytic approach to analyzing item-level data within social and behavioral sciences (Bartholomew et al. 2008). Such data are typically either dichotomous (e.g., disagree vs. agree) or polytomous (e.g., strongly disagree, disagree, neither, agree, and strongly agree), for which the standard linear factor models may not be suitable (Wirth and Edwards 2007). Exploratory IFA uncovers and interprets the underlying structure of data by learning the association between the items and the latent factors based on the estimated factor loadings. It has received many applications in social and behavioral sciences, including but not limited to personality, quality-of-life, and clinical research (e.g., Edelen and Reeve 2007; Lee and Ashton, 2009; Reise and Waller 2009).

There are a wide range of psychometric models for exploratory item factor analysis. For the purpose of exploratory analysis, all these models handle multiple latent factors, including

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11336-018-9646-5) contains supplementary material, which is available to authorized users.

Correspondence should be made to Yunxiao Chen, London School of Economics and Political Science, London, UK. Email: y.chen186@lse.ac.uk

multidimensional two-parameter logistic model (M2PL; Reckase 1972, 2009) for dichotomous responses, the multidimensional graded response model (e.g., Cai 2010a) and multidimensional partial credit model (Yao and Schwarz 2006) for polytomous responses, and normal ogive (i.e., probit) models for dichotomous and polytomous responses (Bock et al. 1988). The readers are referred to Wirth and Edwards (2007) for a comprehensive review of the IFA literature. For ease of exposition, we focus on IFA models for dichotomous responses, while point out that our developments can be extended to polytomous data.

The most commonly used method for parameter estimation in exploratory IFA is marginal maximum likelihood (MML) estimation based on an expectation–maximization (EM) algorithm (Bock and Aitkin, 1981; Bock et al. 1988). In this approach, the item parameters are estimated by maximizing the marginal likelihood function, in which the person parameters (i.e., latent factors) have been integrated out. This approach typically involves evaluating a K-dimensional integral, where K is the number of latent factors. The computational complexity of evaluating this integral grows exponentially with the latent dimension K, and the computation becomes infeasible when the latent dimension is too high. In fact, the Gauss–Hermite quadrature-based integration used by Bock and Aitkin (1981) is not recommended for more than five factors (Wirth and Edwards, 2007), which limits the use of MML estimation in large-scale data analysis where many latent factors may be present. In filling this gap, many approaches have been proposed to approximate the integral, including adaptive Gaussian quadrature methods (e.g., Schilling and Bock 2005), Monte Carlo integration (e.g., Meng and Schilling 1996), fully Bayesian estimation methods (e.g., Béguin and Glas 2001; Bolt and Lall, 2003; Edwards 2010), and data augmented stochastic approximation algorithms (e.g., Cai, 2010a; 2010b). However, even with these state-of-the-art algorithms, the computation is time-consuming with the presence of many latent factors.

Alternative approaches have been proposed for parameter estimation in IFA that avoid evaluating high-dimensional integrals. These approaches are computationally more efficient and thus may be more suitable for the analysis of large-scale data. In particular, Lee et al. (1990) propose to first estimate the inter-item polychoric correlation matrix using pairwise response data and then to estimate the loadings by conducting factor analysis based on the estimated polychoric correlation matrix. However, this approach relies heavily on the assumptions of normal ogive models and can hardly be generalized when other link functions are used. Jöreskog and Moustaki (2001) propose a composite likelihood approach that maximizes the sum of all univariate and bivariate marginal likelihoods. In this approach, only one- and two-dimensional numerical integrals need to be evaluated, which is computationally more affordable than that of the MML approach. However, this approach still relies heavily on the assumption that the latent factors follow a multivariate normal distribution which may not always be satisfied in applications.

Joint maximum likelihood (JML) estimator is one of the earliest approaches to parameter estimation for IFA models that is known to be computationally efficient (see Chapter 8, Embretson and Reise, 2000). This approach is first suggested in Birnbaum (1968) when the basic forms of item response theory models are proposed and has been used in item response analysis for many years (Lord, 1980; Mislevy and Stocking, 1987) until the MML approach becomes dominant. The key difference between the MML and the JML methods is that in the MML approach the person parameters are treated as random effects and are integrated out from the likelihood function, while in the JML approach the person parameters are treated as fixed effect parameters and kept in the likelihood function. As a result, the evaluation of numerical integrals in the MML approach is replaced by maximizing with respect to the person parameters in the JML approach. Under a latent factor model with a high latent dimension, the computational complexity of the latter is much lower than that of the former. However, in the IFA literature, JML estimation is less preferred to MML estimation. This is because, under the classical asymptotic setting where the number of respondents grows to infinity and the number of items is fixed, the number of parameters in the joint likelihood function also grows to infinity, for which the standard theory for maximum likelihood

estimation does not apply. Consequently, the point estimation of every single item parameter is inconsistent (Neyman and Scott 1948; Andersen 1973; Haberman 1977; Ghosh 1995) even for simple IRT models, let alone the validity of the standard errors for the item parameter estimates.

Despite its statistical inconsistency in the classical sense, the JML approach is computationally efficient, easily programmable, and generally applicable to many IRT models (Embretson and Reise, 2000). Though possibly biased, the empirical performance of JML estimator for point estimation is usually reasonable, especially when constraints are placed on the JML solution. Given the unique strength of JML-based estimation, its properties are worth investigating from a theoretical perspective. In this paper, we provide statistical theory to IFA based on the joint likelihood for analyzing large-scale data where both the number of people and the number of items are large. Our asymptotic setting differs from the standard one by letting both the numbers of people and items grow to infinity. This setting seems reasonable for analyzing large-scale item response data. Similar asymptotic settings have been considered in psychometric research, including the analysis of unidimensional IRT models (Haberman 1977, 2004) and diagnostic classification models (Chiu et al. 2016). Under this asymptotic setting, we propose a constrained joint maximum likelihood estimator (CJMLE) that has certain notion of statistical consistency in recovering factor loadings. Since the number of loading parameters grows to infinity under this asymptotic setting, this notion of consistency is different from that in the classical sense for maximum likelihood estimation. Specifically, we show that, up to a rotation, the proportion of inconsistently estimated loading parameters converges to zero in probability.

The major advantage of the proposed CJMLE over the MML-based approaches is its low computational cost. An alternating minimization (AM) algorithm with projected gradient decent update is proposed, which can be parallelled efficiently. Specifically, we implement this parallel computing algorithm in R with core functions written in C++ through Open Multi-Processing (OpenMP, Dagum and Menon 1998) that can scale to very large data. For example, the algorithm can fit a dataset with 125,000 respondents, 500 items, and 10 latent traits within 3 min on a single Intel® machine[1] with four cores. Compared with Lee et al. (1990) and Jöreskog and Moustaki (2001), our method is not only more flexible for its ability to handle almost all IFA models, but also computationally more efficient. Specifically, the computational complexity of our method is linear in the number of items in each iteration, while that of Lee et al. (1990) and Jöreskog and Moustaki (2001) is quadratic.

As an illustration, we apply the proposed estimator to a personality assessment dataset based on a revised version of the Eysenck's personality questionnaire (Eysenck et al. 1985). This dataset contains 79 items, which are designed to measure three personality factors, Extraversion (E), Neuroticism (N), and Psychoticism (P). It is found that a three-factor model fits the data best, according to a cross-validation procedure. In addition, the three factors identified by the Geomin rotation (Yates 1988) correspond well to the three factors in Eysenck's model of personality.

The remainder of the paper is organized as follows. In Sect. 2, we propose the constrained joint maximum likelihood estimator under a general form of IFA models and establish its asymptotic properties. Then in Sect. 3, a computational algorithm is proposed. Simulation studies and real data analysis are presented in Sects. 4 and 5, respectively. Finally, discussions are provided in Sect. 6. Proofs of our theoretical results are provided in supplementary material.

2. Constrained Joint Maximum Likelihood Estimation

2.1. IFA Models for Dichotomous Responses

We focus on a class of IFA models for dichotomous responses, which includes the M2PL model and the normal ogive model as special cases. Let $i = 1, \ldots, N$ indicate respondents and

[1] Core(TM) i7CPU@5650U@2.2 GHz.

$j = 1, \ldots, J$ indicate items. Each respondent i is represented by a K-dimensional latent vector $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})^\top$, and each item is represented by $K + 1$ parameters including an intercept parameter d_j and K loading parameters $\mathbf{a}_j = (a_{j1}, \ldots, a_{jK})^\top$. Let Y_{ij} be the response from respondent i to item j, which is assumed to follow distribution

$$P(Y_{ij} = 1|\boldsymbol{\theta}_i, d_j, \mathbf{a}_j) = f(d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i), \tag{1}$$

where $f(x)$ is a pre-specified link function. Given the latent vector $\boldsymbol{\theta}_i$, respondent i's responses $Y_{i1}, \ldots, Y_{iJ}$ are assumed to be conditionally independent. This assumption is known as the local independence assumption, a standard assumption for item factor analysis. We denote the observed value of Y_{ij} by y_{ij}.

The framework (1) includes the M2PL model and the normal ogive model as special cases. Specifically, for the M2PL model, the link function takes the logistic form

$$f(x) = \frac{\exp(x)}{1 + \exp(x)},$$

and for the normal ogive model, the link function becomes

$$f(x) = \int_{-\infty}^{x} \phi(t)\mathrm{d}t,$$

where $\phi(x)$ is the probability density function of a standard normal distribution. Besides these two widely used models, other link functions may also be used, such as a complimentary log–log link or a link function with pre-specified lower and/or upper asymptotes.

Given a model, the MML-based IFA further requires the specification of a prior distribution on the latent factors $\boldsymbol{\theta}_i$. In fact, the consistency of MML estimation relies on the correct specification of the prior distribution, under the classical asymptotic setting. For exploratory IFA, a commonly adopted assumption is that $\boldsymbol{\theta}_i$ follows a K-dimensional standard normal distribution. In the implementation of the Gauss–Hermite quadrature-based EM algorithm, this distribution is further approximated by a discrete distribution supported on a ball. In contrast, as will be described in the sequel, the JML-based IFA does not require the specification of a prior.

2.2. *Constrained Joint Maximum Likelihood Estimation*

Under the general model form (1), the joint likelihood function is a function of both the item parameters $\mathbf{a}_j$ and d_j and the person parameters $\boldsymbol{\theta}_i$, specified as

$$\begin{aligned} &L(\boldsymbol{\theta}_i, \mathbf{a}_j, d_j : i = 1, \ldots, N, j = 1, \ldots, J) \\ &\quad = \prod_{i=1}^{N} \prod_{j=1}^{J} f(d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i)^{y_{ij}} (1 - f(d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i))^{1-y_{ij}}. \end{aligned} \tag{2}$$

The classical JML estimator is defined as the maximizer of the joint likelihood function

$$\begin{aligned} &(\hat{\boldsymbol{\theta}}_i, \hat{\mathbf{a}}_j, \hat{d}_j : i = 1, \ldots, N, j = 1, \ldots, J) \\ &\quad = \underset{\boldsymbol{\theta}_i, \mathbf{a}_j, d_j}{\arg\max} \log L(\boldsymbol{\theta}_i, \mathbf{a}_j, d_j : i = 1, \ldots, N, j = 1, \ldots, J). \end{aligned} \tag{3}$$

One issue with the JML estimator is that estimates are not available for items or persons with perfect scores (all 1 s or all 0 s), when no constraints are placed. To avoid this issue, we propose a constrained joint maximum likelihood estimator (CJMLE), defined as

$$\begin{aligned}&(\hat{\boldsymbol{\theta}}_i, \hat{\mathbf{a}}_j, \hat{d}_j : i = 1, \ldots, N, j = 1, \ldots, J)\\ &\quad = \underset{\boldsymbol{\theta}_i, \mathbf{a}_j, d_j}{\arg\max} \log L(\boldsymbol{\theta}_i, \mathbf{a}_j, d_j : i = 1, \ldots, N, j = 1, \ldots, J)\\ &s.t. \sqrt{1 + \|\boldsymbol{\theta}_i\|^2} \le C, \sqrt{d_j^2 + \|\mathbf{a}_j\|^2} \le C, i = 1, \ldots, N, j = 1, \ldots, J.\end{aligned} \tag{4}$$

Throughout this paper, $\|\mathbf{x}\|$ denotes the Euclidian norm of a vector $\mathbf{x} = (x_1, \ldots, x_K)$, defined as $\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots x_K^2}$. In (4), C is a pre-specified positive constant that imposes regularization on the magnitudes of the person-wise parameters and the item-wise parameters. Since the feasible set given by the constraints in (4) is compact and the objective function is continuous, the optimization problem is guaranteed to have a solution. Therefore, estimates exist even for items and persons with perfect scores. It is also worth pointing out that the solution to (4) is not unique, due to rotational indeterminacy (Browne 2001), to be further discussed in Sect. 2.4. As will also be shown in Sect. 2.4, the CJMLE has statistical guarantees for any sufficiently large value of C, under the asymptotic regime where both N and J grow to infinity. In the rest of the paper, we use $C = 5\sqrt{K}$ as a default value under the M2PL model.

2.3. Theoretical Properties: Recovery of Response Probabilities

We establish the asymptotic properties of the CJMLE defined in (4). We denote $\boldsymbol{\theta}_i^*$, $\mathbf{a}_j^*$, and d_j^* the true model parameters, where $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, J$. In this analysis, the dimension K of the latent space is known, while in practice one may choose a dimension K either via cross-validation or by using an information criterion. We introduce the following notations.

1. $\Theta = (\theta_{ik})_{N\times K}$ denotes the matrix of person parameters.
2. $A = (a_{jk})_{J\times K}$ denotes the matrix of factor loadings.
3. $\mathbf{d} = (d_1, \ldots, d_J)$ denotes the vector of intercept parameters.
4. $\Theta^* = (\theta_{ik}^*)_{N\times K}$, $A^* = (a_{jk}^*)_{J\times K}$ and $\mathbf{d}^* = (d_1^*, ..., d_J^*)$ denote the true parameters.
5. $\mathbf{1}_N = (1, \ldots, 1)$ denotes a vector with all N entries being 1.
6. $X_{[k]}$ denotes the kth column vector of a matrix X.
7. $\hat{\Theta} = (\hat{\theta}_{ik})_{N\times K}$, $\hat{\mathbf{d}} = (\hat{d}_1, \ldots, \hat{d}_J)$, and $\hat{A} = (\hat{a}_{jk})_{J\times K}$ denote the CJMLE given in (4).
8. $\|X\|_F = \sqrt{\sum_{i=1}^N \sum_{j=1}^J x_{ij}^2}$ denotes the Frobenius norm of a matrix $X = (x_{ij})_{N\times J}$.

In addition, we require the following regularity conditions.

A1. $\sqrt{1 + \|\boldsymbol{\theta}_i^*\|^2} \le C$ and $\sqrt{(d_j^*)^2 + \|\mathbf{a}_j^*\|^2} \le C, i = 1, \ldots, N, j = 1, \ldots, J$.

A2. The link function f is differentiable, satisfying

$$\sup_{|x|\le C^2} \frac{|f'(x)|}{f(x)(1-f(x))} < \infty \text{ and } \sup_{|x|\le C^2} \frac{f(x)(1-f(x))}{(f'(x))^2} < \infty.$$

These two conditions are reasonable and easy to understand. Condition A1 requires that the true person parameters and the true item parameters satisfy the constraints used in the CJMLE defined in (4). Condition A2 requires that the link function f has a certain level of smoothness. In particular, the commonly used link functions, including the logit, probit, and the complimentary log–log links, satisfy A2.

Theorem 1. *Suppose that assumptions A1 and A2 are satisfied. Then there exist constants C_1 and C_2 that depend on the value of C (but independent of N and J), such that*

$$\frac{1}{NJ}\left\|\hat{\Theta}\hat{A}^\top + \mathbf{1}_N\hat{\mathbf{d}}^\top - \Theta^*(A^*)^\top - \mathbf{1}_N\mathbf{d}^{*\top}\right\|_F^2 \le C_2\sqrt{\frac{J+N}{NJ}} \tag{5}$$

is satisfied with probability at least $1 - C_1/(N+J)$, where $\|\cdot\|_F$ denotes the matrix Frobenius norm defined above.

The proof of Theorem 1 is given in the supplementary material that makes use of a concentration inequality proved in Davenport et al. (2014). The bound (5) is satisfied for all N and J, without requiring N and J to grow to infinity. When both N and J grow to infinity, Theorem 1 implies that the left side of (5) converges to 0 in probability.

Theorem 1 is essentially about the accuracy of estimating the true response probabilities. This is because the conditional distribution of Y_{ij} depends on $\boldsymbol{\theta}_i$, $\mathbf{a}_j$, and d_j only through $d_j + \mathbf{a}_j^\top\boldsymbol{\theta}_i$, the (i, j)th entry of the matrix $\Theta A^\top + \mathbf{1}_N\mathbf{d}^\top$. Consequently, the left side of (5) quantifies an averaged discrepancy between the true values $\Theta^*(A^*)^\top + \mathbf{1}_N\mathbf{d}^*$ and their estimates $\hat{\Theta}\hat{A}^\top + \mathbf{1}_N\hat{\mathbf{d}}^\top$. Moreover, Theorem 1 implies the consistent recovery of the response probabilities in an average sense, as described in Corollary 1.

Corollary 1. (Recovery of Response Probabilities) *Under the same conditions as Theorem 1, when N and J grow to infinity,*

$$\frac{\sum_{i=1}^N\sum_{j=1}^J\left(f(\hat{d}_j + \hat{\mathbf{a}}_j^\top\hat{\boldsymbol{\theta}}_i) - f(d_j^* + (\mathbf{a}_j^*)^\top\boldsymbol{\theta}_i^*)\right)^2}{NJ} \tag{6}$$

converges to zero in probability.

Note that $f(\hat{d}_j + \hat{\mathbf{a}}_j^\top\hat{\boldsymbol{\theta}}_i)$ is the predicted probability of $Y_{ij} = 1$ given by the CJMLE and $f(d_j^* + (\mathbf{a}_j^*)^\top\boldsymbol{\theta}_i^*)$ is the corresponding true probability. Therefore, the result of Corollary 1 implies that the predicted probabilities and their true values are close in an average sense. It further implies that only a small proportion of true item response probabilities are not estimated well; that is, for any small constant $\epsilon > 0$, the proportion

$$\frac{\sum_{i=1}^N\sum_{j=1}^J 1_{\{|f(\hat{d}_j+\hat{\mathbf{a}}_j^\top\hat{\boldsymbol{\theta}}_i) - f(d_j^*+(\mathbf{a}_j^*)^\top\boldsymbol{\theta}_i^*)|>\epsilon\}}}{NJ}$$

converges to zero in probability. This property may be important to psychological measurement, as the item response probabilities completely characterize the respondents' behavior on the items.

To our knowledge, the type of asymptotic result established in Corollary 1 is not considered in the classical asymptotic theory based on the marginal maximum likelihood. In fact, under the classical asymptotic setting, the quantity (6) does not converge to zero in probability if the number of items J is fixed, no matter how the parameters are estimated.

2.4. Theoretical Properties: Recovery of Loadings

We now study the recovery of the loading structure A^*, which is of particular interest in exploratory IFA. Specifically, we will show that $\hat{A}$ given by the CJMLE approximates A^* well in a sense to be clarified.

We start the discussion with the identifiability of the model parameters. Given all the true response probabilities, or equivalently, the matrix $\Theta^*(A^*)^\top + \mathbf{1}_N \mathbf{d}^{*\top}$, the parameters Θ^*, A^*, and $\mathbf{d}^*$ cannot be uniquely determined. To avoid this indeterminacy issue, we impose the following regularity condition on the true person parameters.

A3. The true person parameters satisfy

$$\mathbf{1}_N^\top \Theta^*_{[k]} = 0, \tag{7}$$

$$\frac{1}{N}(\Theta^*_{[k]})^\top \Theta^*_{[k]} = 1, \tag{8}$$

$$(\Theta^*_{[k]})^\top \Theta^*_{[k']} = 0, \quad k, k' = 1, \ldots, K, k \neq k'. \tag{9}$$

The constraints in A3 are similar to assuming the means and covariance matrix of $\boldsymbol{\theta}_i$ are 0s and identity matrix, respectively, when analyzing data using an MML approach. Even under these constraints, Θ^* and A^* are only determined up to a rotation, known as rotational indeterminacy. A summary of the phenomenon of rotational indeterminacy is given in the supplementary material.

Taking the constraints (7)–(9) into account, we standardize the CJMLE solution $(\hat{\Theta}, \hat{A}, \hat{\mathbf{d}})$, so that the same constraints are satisfied. The standardized solution is denoted by $(\tilde{\Theta}, \tilde{A}, \tilde{\mathbf{d}})$, where the standardization procedure is given in the supplementary material. We then show that $\tilde{A}$ accurately estimates A^* up to a rotation, when the following regularity condition also holds.

A4. There exists a positive constant $C_3 > 0$, such that the Kth (i.e., the smallest) singular value of A^*, denoted by σ_K^*, satisfying $\sigma_K^* \geq C_3\sqrt{J}$, for all J.

Theorem 2. *Suppose that assumptions A1–A4 are satisfied. Then the following scaled Frobenius loss*

$$\min_{Q} \left\{ \frac{1}{JK} \|A^* - \tilde{A}Q\|_F^2 : Q^\top Q = I_{K\times K} \right\} \tag{10}$$

converges to zero in probability as $N, J \to \infty$, where $\tilde{A}$ is the standardized version of $\hat{A}$.

We remark on the result of Theorem 2. Suppose that $\tilde{Q}$ minimizes the optimization problem (10). In addition, we denote $\bar{A} = (\bar{a}_{jk})_{J\times K} = \tilde{A}\tilde{Q}$. Then (10) converging to zero implies that for any $\epsilon > 0$,

$$\lim_{N,J\to\infty} \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} 1_{\{|a^*_{jk} - \bar{a}_{jk}| > \epsilon\}}}{JK} = 0.$$

That is, the proportion of inaccurately estimated loading parameters converges to zero in probability under the optimal rotation.

In practice, the optimal rotation $\tilde{Q}$ is not available, since A^* is unknown. A suitable rotation may be obtained by using analytic rotation methods (see, e.g., Browne 2001) to yield a simple pattern of factor loadings that is easy to interpret, where a simple loading pattern refers to a loading matrix with many entries close to 0, so that each item is mainly associated with a small number of latent factors and each latent factor is mainly associated with a small number of items. When the true loading matrix A^* has a simple pattern, we believe that a certain notion of consistency can be established for analytic rotation methods.

Finally, we remark that condition A4 is mild. In fact, when $\mathbf{a}_j^*$s are i.i.d. random vectors from a distribution and the covariance matrix of $\mathbf{a}_j^*$ is strictly positive definite, $\sigma_K^* \geq C_3\sqrt{J}$ is satisfied with probability close to 1 for sufficiently large J, when taking C_3 to be $0.5\sqrt{\lambda_K}$, where λ_K is the smallest eigenvalue of the covariance matrix of $\mathbf{a}_j^*$.

2.5. *Extension: Analyzing Missing Data*

In practice, each respondent may only respond to a small proportion of items, possibly due to the data collection design. The proposed CJMLE also handles missing data. More precisely, let W_{ij} indicate whether or not the (i, j)th entry of the response matrix is missing, where $W_{ij} = 0$ if the corresponding response is missing and $W_{ij} = 1$ otherwise. We say the missingness is ignorable when the following equation holds

$$\begin{aligned} &P(Y_{i1} = y_1, \ldots, Y_{iJ} = y_J, W_{i1} = \omega_1, \ldots, W_{iJ} = \omega_J | \boldsymbol{\theta}_i, \mathbf{a}_j, d_j) \\ &\quad = P(Y_{i1} = y_1, \ldots, Y_{iJ} = y_J | \boldsymbol{\theta}_i, \mathbf{a}_j, d_j) \times P(W_{i1} = \omega_1, \ldots, W_{iJ} = \omega_J | \boldsymbol{\theta}_i, \mathbf{a}_j, d_j) \\ &\quad = \left(\prod_{j=1}^{J} P(Y_{ij} = y_j | \boldsymbol{\theta}_i, \mathbf{a}_j, d_j) \right) \times \left(\prod_{j=1}^{J} P(W_{ij} = \omega_j | \boldsymbol{\theta}_i, \mathbf{a}_j, d_j) \right). \end{aligned}$$

Let ω_{ij} be a realization of W_{ij}. Then the responses y_{ij} are only observed for the entries with $\omega_{ij} = 1$. Under ignorable missingness, the joint likelihood function becomes

$$\begin{aligned} &L(\boldsymbol{\theta}_i, \mathbf{a}_j, d_j : i = 1, \ldots, N, j = 1, \ldots, J) \\ &\quad = \prod_{i,j:\omega_{ij}=1} f(d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i)^{y_{ij}} (1 - f(d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i))^{1-y_{ij}}. \end{aligned} \tag{11}$$

When $\omega_{ij} = 1$ for all i and j, no response is missing and (11) becomes the same as (2).

The statistical guarantee established earlier for complete data can be extended to data with massive missingness. For technical simplicity, we assume that the data are missing completely at random.

A5. W_{ij}s are i.i.d. Bernoulli random variables with

$$P(W_{ij} = 1) = \frac{n}{NJ},$$

for some $n > 0$.

Under this assumption, Theorems 3 and 4 extend Theorems 1 and 2 by allowing for missing data. In fact, Theorems 1 and 2 can be viewed as special cases of Theorems 3 and 4 when $n = NJ$. The proofs of Theorems 3 and 4 are given in the supplementary material.

Theorem 3. *Suppose that assumptions A1, A2, and A5 are satisfied. Further assume that* $n \geq (N + J) \log(JN)$. *Then there exist constants* C_1 *and* C_2 *that depend on the value of* C *(but independent of* N *and* J*), such that*

$$\frac{1}{NJ} \| \hat{\Theta} \hat{A}^\top + \mathbf{1}_N \hat{\mathbf{d}}^\top - \Theta^* (A^*)^\top - \mathbf{1}_N \mathbf{d}^{*\top} \|_F^2 \leq C_2 \sqrt{\frac{J + N}{n}} \tag{12}$$

is satisfied with probability at least $1 - C_1/(N + J)$.

Theorem 4. *Suppose that assumptions A1–A5 are satisfied. Further assume that* $n \geq (N + J) \log(JN)$. *Then the following scaled Frobenius loss*

$$\min_{Q} \left\{ \frac{1}{JK} \| A^* - \tilde{A} Q \|_F^2 : Q^\top Q = I_{K \times K} \right\} \tag{13}$$

converges to zero in probability as $N, J \to \infty$*, where* $\tilde{A}$ *is the standardized version of* $\hat{A}$*.*

Noting that when $n \geq (N + J)\log(JN)$, the right side of equations (12) converges to zero when N and J grow to infinity. Consequently, Corollary 1 can be extended to this missing data setting. This asymptotic validity of the CJMLE for missing data suggests its potential in applications of test equating and linking, which can be formulated into missing data analysis problems (see, e.g., von Davier 2010).

We provide a discussion on condition A5. Under certain data collection designs, data can be regarded as missing completely at random (MCAR). However, it is often the case in practice that the MCAR assumption may be too strong. Instead, it may be more reasonable to assume missing at random (MAR), under which the probability of observing a response $P(W_{ij} = 1)$ depends on the corresponding parameter values, including $\boldsymbol{\theta}_i$, $\mathbf{a}_j$, and d_j. Our theoretical results may be extended to the MAR setting (e.g., using techniques from Cai and Zhou 2013).

2.6. Selection of Number of Factors

We provide a cross-validation method for the selection of the number K of latent factors when it is unknown. Let $\Omega = (\omega_{ij})_{N\times J}$ be the indicator matrix of non-missing responses. We randomly split the non-missing responses into B non-overlapping sets that are of equal sizes, indicated by $\Omega^{(b)} = (\omega_{ij}^{(b)})_{N\times J}$, $b = 1, 2, \ldots, B$, satisfying $\sum_{b=1}^{B} \Omega^{(b)} = \Omega$. Moreover, we denote $\Omega^{(-b)} = \sum_{b'\neq b} \Omega^{(b')}$, indicating the data excluding set b.

For a given latent dimension K, we find the CJMLE based on the non-missing responses indicated by $\Omega^{(-b)}$. The CJMLE solution is denoted by $(\hat{\Theta}^{(b)}, \hat{A}^{(b)}, \hat{\mathbf{d}}^{(b)})$. As defined below, the cross-validation error for fold b is computed based on the accuracy of predicting the responses in the set $\Omega^{(b)}$ using $(\hat{\Theta}^{(b)}, \hat{A}^{(b)}, \hat{\mathbf{d}}^{(b)})$.

$$err^{(b)}(K) = \sum_{i,j:\, \omega_{ij}^{(b)}=1} \left(y_{ij} - f\left(\hat{d}_j^{(b)} + \left(\hat{\mathbf{a}}_j^{(b)}\right)^{\top} \hat{\boldsymbol{\theta}}_i^{(b)} \right) \right)^2.$$

The overall cross-validation error is defined as

$$err(K) = \sum_{b=1}^{B} err^{(b)}(K).$$

The latent dimension K that yields the smallest cross-validation error is selected. In the analysis of this paper, we set $B = 5$ (i.e., fivefold cross-validation).

3. Computation

We develop an alternating minimization algorithm for solving the optimization problem (4). In fact, the first JML estimation paradigm employed in Birnbaum (1968) can be regarded as an alternating minimization algorithm. This paradigm is the basis for JML estimation for many IRT computer programs in general use (Baker 1987). As indicated by its name, this algorithm decomposes the parameters into two sets, the person parameters and the item parameters, and alternates between minimizing one set of parameters given the other. It is worth noting that given the person parameters, the optimization with respect to item parameters can be split into J

independent optimization problems, each containing $(\mathbf{a}_j, d_j)$, $j = 1, \ldots, J$. Similarly, the person parameters can also be updated independently for $\boldsymbol{\theta}_i$, $i = 1, \ldots, N$, given the item parameters.

To handle the constraints in (4), a projected gradient descent update is used in each iteration, defined as follows. We first define projection operator

$$\text{Prox}_C(\mathbf{y}) = \underset{\mathbf{x}:\|\mathbf{x}\|\le C}{\arg\min} \|\mathbf{y} - \mathbf{x}\|^2 = \begin{cases} \mathbf{y} & \text{if } \|\mathbf{y}\| \le C; \\ \frac{C}{\|\mathbf{y}\|}\mathbf{y} & \text{if } \|\mathbf{y}\| > C. \end{cases} \tag{14}$$

Here, $\text{Prox}_C(\mathbf{y})$ returns the projection of $\mathbf{y}$ onto the feasible set. Consider optimization problem

$$\begin{aligned} &\min_{\mathbf{x}} \ f(\mathbf{x}) \\ &s.t. \ \|\mathbf{x}\| \le C, \end{aligned} \tag{15}$$

where f is a differentiable convex function. Denote the gradient of f by g. Then a projected gradient descent update at $\mathbf{x}^{(0)}$ is defined as

$$\mathbf{x}^{(1)} = \text{Prox}_C(\mathbf{x}^{(0)} - \eta g(\mathbf{x}^{(0)})),$$

where $\eta > 0$ is a step size decided by line search. Due to the projection, $\|\mathbf{x}^{(1)}\| \le C$. Furthermore, it can be shown that for sufficiently small η, $f(\mathbf{x}^{(1)}) < f(\mathbf{x}^{(0)})$, when f satisfies mild regularity conditions and $\|g(\mathbf{x}^{(0)})\| \ne 0$; see Parikh and Boyd (2014) for further details.

Algorithm 1 (Alternating Minimization Algorithm for CJMLE)

1 *(Initialization) Input responses y_{ij}, non-missing response indicator ω_{ij}, dimension K of latent space, constraint parameter C, iteration number $m = 0$, and initial value $\Theta^{(0)}$, $A^{(0)}$, $\mathbf{d}^{(0)}$.*

2 *(Alternating minimization) At the $m + 1$th iteration,*

(a) *Perform parallel computation for $i = 1, \ldots, N$. For each respondent i, update $\boldsymbol{\theta}_i^{(m+1)} = Prox_{\sqrt{C^2-1}}\left(\boldsymbol{\theta}_i^{(m)} - \eta \mathbf{g}_i^{(m)}\right)$, where $\mathbf{g}_i^{(m)}$ is the gradient of*

$$l_i^{(m)}(\boldsymbol{\theta}) = -\sum_{j:\omega_{ij}=1} \left\{ y_{ij} \log f(d_j^{(m)} + \boldsymbol{\theta}^\top \mathbf{a}_j^{(m)}) + (1 - y_{ij}) \log(1 - f(d_j^{(m)} + \boldsymbol{\theta}^\top \mathbf{a}_j^{(m)})) \right\}$$

at $\boldsymbol{\theta}_i^{(m)}$. $\eta > 0$ is a step size chosen by line search.

(b) *Given $\boldsymbol{\theta}_i^{(m+1)}$, $i = 1, \ldots, N$ from (a), perform parallel computation for $j = 1, \ldots, J$. For each item j, update $(d_j^{(m+1)}, \mathbf{a}_j^{(m+1)}) = Prox_C\left((d_j^{(m)}, \mathbf{a}_j^{(m)}) - \eta \tilde{\mathbf{g}}_j^{(m)}\right)$, where $\tilde{\mathbf{g}}_j^{(m)}$ is the gradient of*

$$\tilde{l}_j^{(m)}(d, \mathbf{a}) = -\sum_{i:\omega_{ij}=1} \left\{ y_{ij} \log f(d + \mathbf{a}^\top \boldsymbol{\theta}_i^{(m+1)}) + (1 - y_{ij}) \log(1 - f(d + \mathbf{a}^\top \boldsymbol{\theta}_i^{(m+1)})) \right\}$$

at $(d_j^{(m)}, \mathbf{a}_j^{(m)})$. $\eta > 0$ is a step size chosen by line search.

Iteratively perform (a) and (b) until convergence.

3 *(Output) Output $\hat{\Theta} = \Theta^{(M)}$, $\hat{A} = A^{(M)}$, and $\hat{\mathbf{d}} = \mathbf{d}^{(M)}$, where M is the last iteration number.*

The algorithm guarantees the joint likelihood function to increase in each iteration, when the step size η in each iteration is properly chosen by line search. The parallel computing in step 2 of the algorithm is implemented through OpenMP (Dagum and Menon 1998), which greatly speeds up the computation even on a single machine with multiple cores. The efficiency of this parallel algorithm is further amplified, when running on a computer cluster with many machines. We also develop a singular value decomposition based algorithm for generating a good starting point for Algorithm 1. The details of this algorithm are given in the supplementary material.

4. Simulation Study

4.1. Simulation Study I

Simulation setting In this study, we evaluate the proposed method by Monte Carlo simulation under a variety of settings, listed as follows.

1. A growing sequence of number of items is considered: $J = 100, 200, \ldots, 500$.
2. Let the number of people $N = \tau J$, where $\tau = 10$ and 25.
3. Two choices of K are considered, $K = 3$ and 10.

This leads to 20 different settings. Under each setting, 100 replications are generated. For each setting, the true model parameters are generated as follows. We first generate $\boldsymbol{\theta}_i^0 = (\theta_{i1}^0, \ldots, \theta_{iK}^0)$ i.i.d. from a K-variate truncated normal distribution, for $i = 1, \ldots, N$. More precisely, the probability density function of $\boldsymbol{\theta}_i^0$ is given by

$$h_1(\mathbf{x}) \propto 1_{\{\|\mathbf{x}\| \le 4\sqrt{K}\}} \prod_{k=1}^{K} \phi(x_k),$$

where $\mathbf{x} = (x_1, \ldots, x_K)$ and $\phi(\cdot)$ denotes the probability density function of a standard normal distribution. This truncated normal distribution is very close to a K-variate standard normal distribution, since the probability $P(\|\mathbf{X}\| \ge 4\sqrt{K})$ is almost 0 when $\mathbf{X}$ follows a K-variate standard normal distribution. We then generate d_j^0 i.i.d. from uniform distribution over the interval $[-2, 2]$, for $j = 1, \ldots, J$. We finally generate $\mathbf{a}_j^0$s, $j = 1, \ldots, J$, so that many of them are sparse. Specifically, we let $\mathbf{q}_j = (q_{j1}, \ldots, q_{jK})$ be a random vector, satisfying

$$P(\mathbf{q}_j = \mathbf{q}) = \frac{1}{2^K - 1},$$

where $\mathbf{q} \in \{0, 1\}^K$ and $\mathbf{q} \ne (0, \ldots, 0)$. Also let γ_{jk} be i.i.d. uniformly distributed over the interval $[0.5, 2.5]$. Then we obtain $\mathbf{a}_j^0 = (q_{j1}\gamma_{j1}, \ldots, q_{jK}\gamma_{jK})$. We obtain Θ^*, A^*, and $\mathbf{d}^*$ by standardizing $\Theta^0 = (\theta_{ik}^0)_{N\times K}$, $A^0 = (a_{jk}^0)_{N\times K}$ and $\mathbf{d}^0 = (d_1^0, \ldots, d_J^0)$.

Recovery of response probabilities We first show results on the recovery of the response probabilities $f(d_j^* + (\mathbf{a}_j^*)^\top \boldsymbol{\theta}_i^*)$. Specifically, Fig. 1 shows the value of the scaled Frobenius loss

$$\frac{1}{NJ}\|\hat{\Theta}\hat{A}^\top + \mathbf{1}_N\hat{\mathbf{d}}^\top - \Theta^*(A^*)^\top - \mathbf{1}_N\mathbf{d}^{*\top}\|_F^2$$

on the y-axis versus the number of items J on the x-axis, under different settings on the ratios between N and J and on the latent dimension K. To provide information on the Monte Carlo

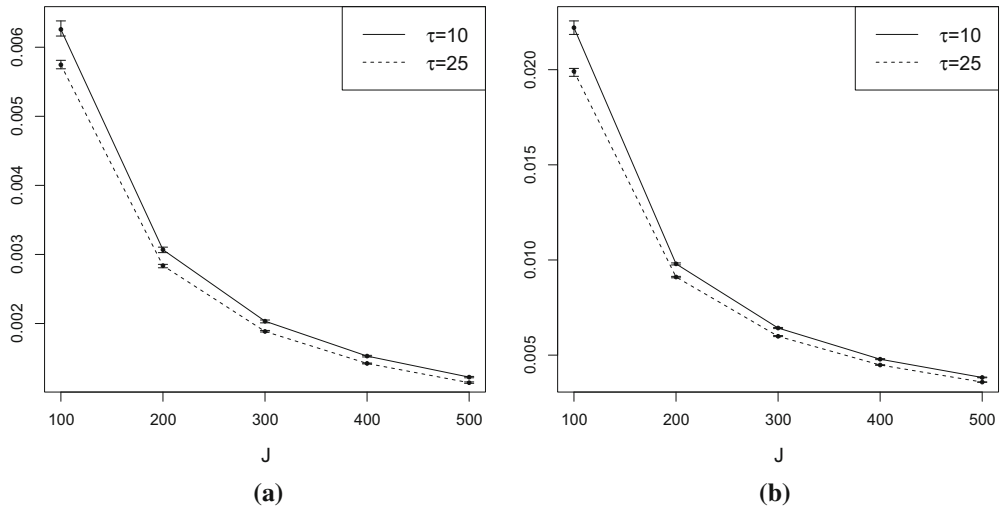

FIGURE 1.
The scaled Frobenius loss for the recovery of response probabilities when $K = 3$ (left panel) and $K = 10$ (right panel). **a** K = 3 and **b** K = 10.

error, the upper and lower quartiles of the scaled Frobenius loss over 100 replications for each setting are provided. From these figures, it can be seen that the scaled Frobenius loss decreases as N and J simultaneously increase.

Recovery of factor loading matrix up to an orthogonal rotation We then show results on the recovery of the factor loading parameters (up to an orthogonal rotation). In particular, Fig. 2 shows the scaled Frobenius loss on the recovery of the loading matrix

$$\min_{Q} \left\{ \frac{1}{JK} \|A^* - \tilde{A}Q\|_F^2 : Q^\top Q = I_{K\times K} \right\},$$

on the y-axis versus the number of items J on the x-axis. These plots are similar to those above for the recovery of response probabilities. Under each setting, the loss decreases toward zero when N and J simultaneously increase.

Selection of latent dimension by cross-validation The performance of the cross-validation method for selecting the latent dimension K is evaluated. When the true latent dimension $K = 3$, we consider a candidate set $\{2, 3, 4\}$, and when the true latent dimension $K = 10$, we choose from $\{9, 10, 11\}$. Fivefold cross-validation is used to choose the latent dimension from the candidate set. According to the simulation result, when the true latent dimension $K = 3$, the cross-validation approach always correctly selects K. When $K = 10$, 100% accuracy is achieved except when J is relatively small (68% for $\tau = 10$, $J = 100$ and 86% for $\tau = 25$, $J = 100$).

4.2. Simulation Study II

Simulation setting In this study, we compare the proposed CJMLE with MMLE, where the latter is obtained via an EM algorithm with fixed quadrature points. We compare under a setting where $K = 2$, since the EM algorithm for MMLE is computationally very intensive when K is larger. We consider a growing sequence of number of items $J = 100, 200, \ldots, 500$ and the number of people $N = \tau J$, where $\tau = 10$ and 25. Each setting is replicated 100 times.

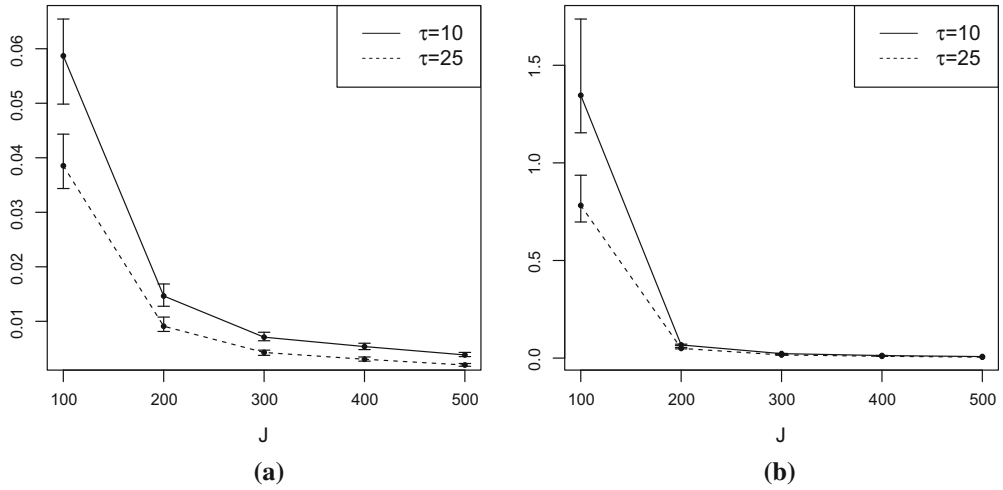

FIGURE 2.
The scaled Frobenius loss for the recovery of loading matrix up to an orthogonal rotation when $K = 3$ (left panel) and $K = 10$ (right panel). **a** K = 3 and **b** K = 10.

Two settings are considered for the generation of $\boldsymbol{\theta}_i^0$. In the first setting, we generate $\boldsymbol{\theta}_i^0$s i.i.d. from a bivariate standard normal distribution, for $i = 1, \ldots, N$. In the second setting, we generate $\boldsymbol{\theta}_i^0$s i.i.d. from a more skewed distribution, by generating θ_{i1}^0 and θ_{i2}^0 independently from a scaled and shifted Beta distribution. More precisely, we let

$$\theta_{ik}^0 = \frac{\zeta_{ik} - \frac{2}{7}}{\sqrt{\frac{5}{196}}}, k = 1, 2, i = 1, \ldots, N,$$

where ζ_{ik}s are i.i.d. random variables that follow a Beta(2,5) distribution. The scaling and shifting standardize θ_{ik}^0 to have mean zero and variance one. The distribution of $\boldsymbol{\theta}_i^0$ is visualized in Fig. 3 through a contour plot of its density function. Given $\boldsymbol{\theta}_i^0$s, the item parameters $\mathbf{a}_j^0$ and d_j^0 are generated in the same way as in Study I. We treat $\boldsymbol{\theta}_i^0$, $\mathbf{a}_j^0$, and d_j^0 as the true model parameters and evaluate the two estimation approaches based on (1) the recovery of the response probabilities $f((\mathbf{a}_j^0)^\top \boldsymbol{\theta}_i^0 + d_j^0)$, (2) the recovery of the factor loading matrix $A^0 = (a_{jk}^0)$ up to an orthogonal rotation, and (3) computation time.

We point out that under the above simulation setting, the assumption A1 which is required in our theory for the CJMLE is not completely satisfied due to the ways $\boldsymbol{\theta}_i^0$s are generated. Moreover, in the current implementation of MMLE, a multivariate standard normal distribution is used as the prior for the latent factors. This prior is correctly specified when $\boldsymbol{\theta}_i^0$s are generated from the bivariate standard normal distribution and is misspecified when $\boldsymbol{\theta}_i^0$s are generated from the scaled and shifted Beta distribution.

The EM algorithm for the MMLE is implemented using the mirt package (Chalmers 2012) in statistical software R. Specifically, for the numerical integral in the E-step, 31 quadrature points are used for each dimension. The comparison of computation time is fair, in the sense that both algorithms are implemented in R language with core functions written in C++, given the same starting values, and performed on computers with the same configuration.

Results The results are given in Fig. 4 through 7 and Tables 1 and 2, where the results are similar under both settings for $\boldsymbol{\theta}_i^0$. In terms of the recovery of the loading matrix up to an orthogonal

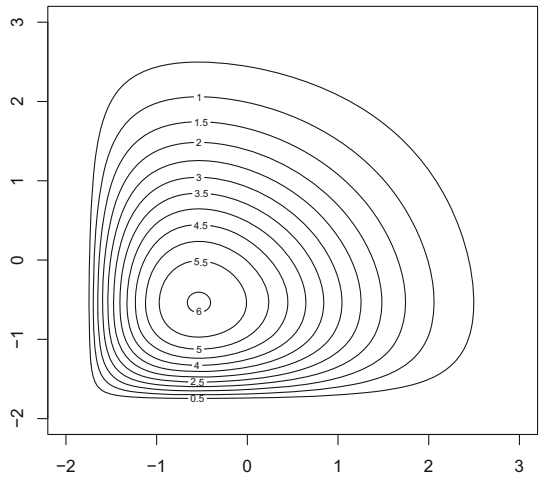

FIGURE 3.
The contour plot of the probability density function for $\boldsymbol{\theta}_i^0$, when θ_{i1}^0 and θ_{i1}^0 are independent and identically distributed, following a scaled and shifted Beta distribution.

rotation, as shown in Figs. 4 and 5, the MMLE performs better when N and J are small and the CJMLE outperforms the MMLE when both N and J are sufficiently large, regardless of the ways $\boldsymbol{\theta}_i^0$s are generated. It is also observed that the scaled Frobenius loss keeps decreasing for the CJMLE when N and J grow simultaneously, which is not the case for the MMLE. For the MMLE, even when the prior distribution is correctly specified for the latent traits, the scaled Frobenius loss for the recovery of the loading matrix first decreases and then increases when N and J simultaneously increase. This is possibly due to the approximation error brought by the fixed quadrature points and is worth future investigation from a theoretical perspective. In terms of the recovery of the item response probabilities based on the scaled Frobenius loss, which is presented in Figs. 6 and 7, the CJMLE always outperforms the MMLE throughout all the settings. Finally, according to Tables 1 and 2, the CJMLE is substantially faster than the EM algorithm for MMLE. For example, when $J = 500$, $N = 5000$, and $\boldsymbol{\theta}_i^0$ follows a bivariate standard normal distribution, the median computation time for the CJMLE is 80 s, while that for the MMLE via the EM algorithm is more than 2000 s.

4.3. Simulation Study III

We further compare the proposed CJMLE algorithm with a Metropolis–Hastings Robbins–Monro (MHRM) algorithm (Cai 2010a, 2010b), which is one of the state-of-the-art algorithms for high-dimensional item factor analysis. This algorithm is implemented in IRT software **flexMIRT**®.

Simulation setting We compare under a setting where $K = 10$, the number of items $J = 100, 200, \cdots, 500$, and the number of people $N = 10J$. We generate $\boldsymbol{\theta}_i^0$s i.i.d. from a bivariate standard normal distribution, for $i = 1, \ldots, N$. The item parameters $\mathbf{a}_j^0$ and d_j^0 are generated in the same way as in Study I. Each setting is replicated 10 times.[2]

Results The two algorithms are compared under the same criteria as in Study II. The results are shown in Fig. 8 and Table 3. According to these results, under the current setting, the CJMLE is

[2]The small number of replications is due to the constraint that **flexMIRT**® needs to be run on a local Windows® machine.

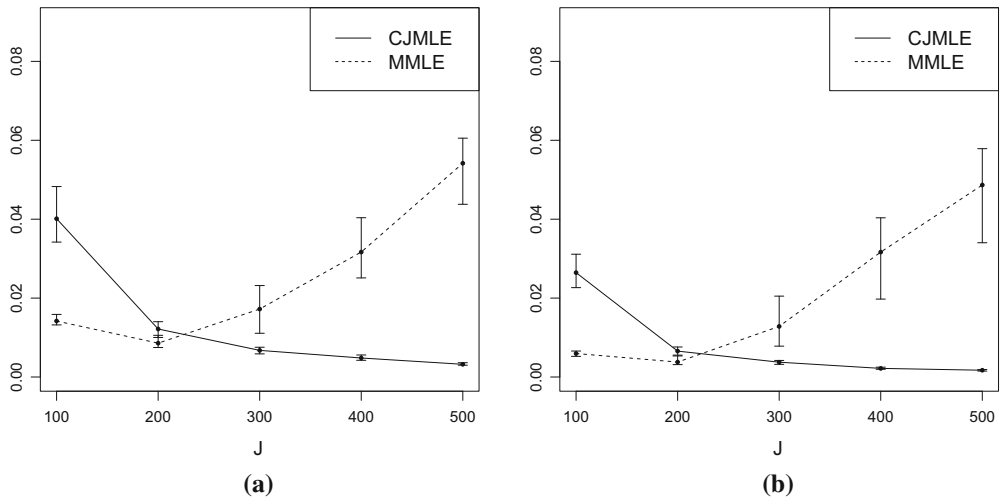

FIGURE 4.
Comparison between the CJMLE and MMLE on the recovery of loading matrix up to an orthogonal rotation, when $\boldsymbol{\theta}_i^0$ follows a standard bivariate normal distribution. **a** $\tau = 10$ and **b** $\tau = 25$.

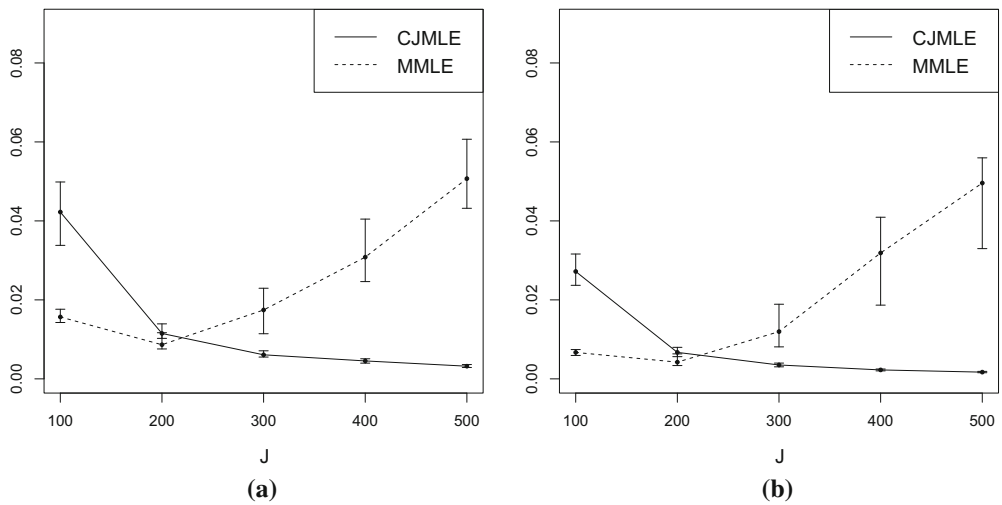

FIGURE 5.
Comparison between the CJMLE and MMLE on the recovery of loading matrix up to an orthogonal rotation, when $\boldsymbol{\theta}_i^0$ is generated based on a Beta distribution. **a** $\tau = 10$ and **b** $\tau = 25$.

not only much faster than the MHRM method, but also more accurate in terms of the recovery of factor loading parameters when $J \geq 200$ (panel (a) of Fig. 8) and in terms of the recovery of item response probabilities (panel (b) of Fig. 8). It is noticed that similar to the result of Study II, the scaled Frobenius loss for the recovery of the loading matrix keeps increasing when N and J simultaneously increase. It may be due to that the default stopping criterion in **flexMIRT**® for the MHRM algorithm does not adapt well to the simultaneous growth of N and J.

TABLE 1.
Speed comparison (in s) between CJMLE and MMLE measured in seconds on a single Intel® E5-2650v4 core, when $\boldsymbol{\theta}_i^0$ follows a standard bivariate normal distribution.

$\tau = 10$	$J = 100$	$J = 200$	$J = 300$	$J = 400$	$J = 500$
CJMLE					
(25% quantile)	1.5	6.2	16.6	35.7	78.8
(50% quantile)	1.5	7.5	16.7	36.1	80.2
(75% quantile)	1.5	7.6	20.2	36.5	80.9
MMLE					
(25% quantile)	44.9	157.1	354.9	771.6	1599.5
(50% quantile)	78.0	333.4	500.0	1079.0	2008.5
(75% quantile)	93.6	459.8	745.5	1637.8	2932.5
$\tau = 25$	$J = 100$	$J = 200$	$J = 300$	$J = 400$	$J = 500$
CJMLE					
(25% quantile)	4.0	16.2	43.6	95.0	198.6
(50% quantile)	4.0	16.3	43.8	95.9	211.0
(75% quantile)	4.0	16.4	53.2	96.4	245.0
MMLE					
(25% quantile)	75.5	511.1	1095.4	1741.2	2799.4
(50% quantile)	145.9	741.8	2227.8	2901.5	3742.4
(75% quantile)	186.1	898.0	3038.1	4785.2	6387.0

TABLE 2.
Speed comparison (in s) between CJMLE and MMLE measured in seconds on a single Intel® E5-2650v4 core, when $\boldsymbol{\theta}_i^0$ is generated based on a Beta distribution.

$\tau = 10$	$J = 100$	$J = 200$	$J = 300$	$J = 400$	$J = 500$
CJMLE					
(25% quantile)	1.4	5.4	14.1	30.9	64.7
(50% quantile)	1.4	5.5	14.2	34.5	70.8
(75% quantile)	1.4	6.7	17.6	35.3	72.7
MMLE					
(25% quantile)	59.1	154.5	344.3	783.3	1583.1
(50% quantile)	81.9	289.3	522.3	987.1	2059.3
(75% quantile)	102.9	477.2	782.8	1455.0	2958.3
$\tau = 25$	$J = 100$	$J = 200$	$J = 300$	$J = 400$	$J = 500$
CJMLE					
(25% quantile)	3.7	14.4	37.5	85.4	164.0
(50% quantile)	3.7	14.6	37.6	87.3	180.5
(75% quantile)	3.7	17.9	37.8	89.1	189.1
MMLE					
(25% quantile)	81.2	363.1	1262.1	1624.4	2651.7
(50% quantile)	145.7	685.4	2222.1	2259.0	3390.7
(75% quantile)	189.5	850.6	3033.7	4369.4	7164.7

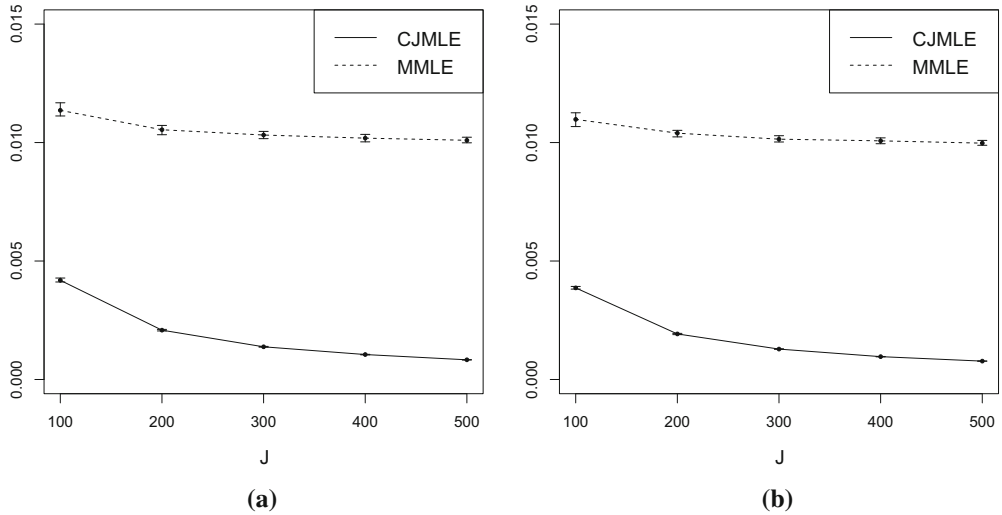

FIGURE 6.
Comparison between the CJMLE and MMLE on the recovery of the item response probabilities, when $\boldsymbol{\theta}_i^0$ follows a standard bivariate normal distribution. **a** $\tau = 10$ and **b** $\tau = 25$.

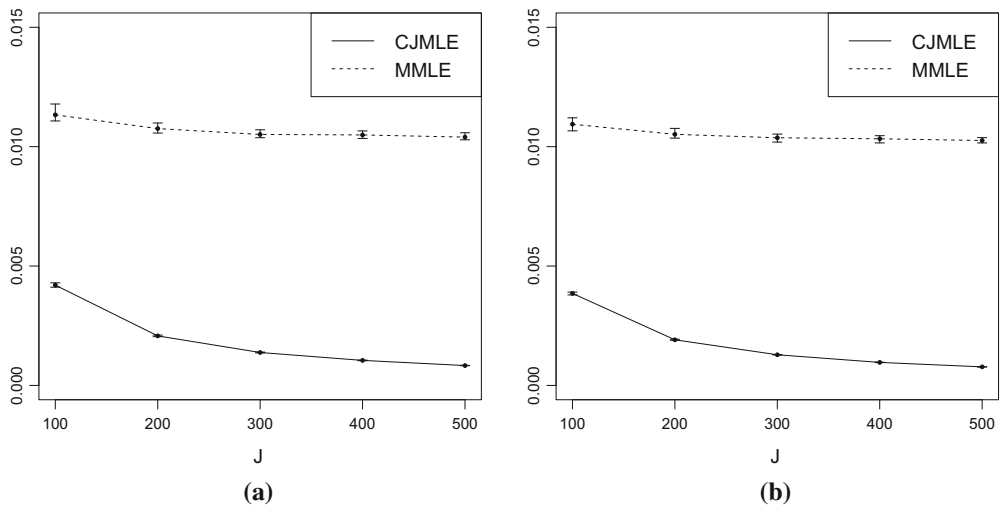

FIGURE 7.
Comparison between the CJMLE and MMLE on the recovery of the item response probabilities, when $\boldsymbol{\theta}_i^0$ is generated based on a Beta distribution. **a** $\tau = 10$ and **b** $\tau = 25$.

5. Real Data Analysis

We illustrate the use of the proposed method on the female UK normative sample data for the EPQ-R (Eysenck et al. 1985). The dataset contains the responses to 79 dichotomous items from 824 people. Among these items, items 1–32, 33–55, and 56-79 consist of the Psychoticism (items 1–32), Extraversion (items 33–55) and Neuroticism (items 56–79) scales, respectively, which are designed to measure the corresponding personality traits. The data have been pre-processed so that the negatively worded items are reversely scored. We analyze the dataset in an exploratory manner and then compare the results with the design of the items.

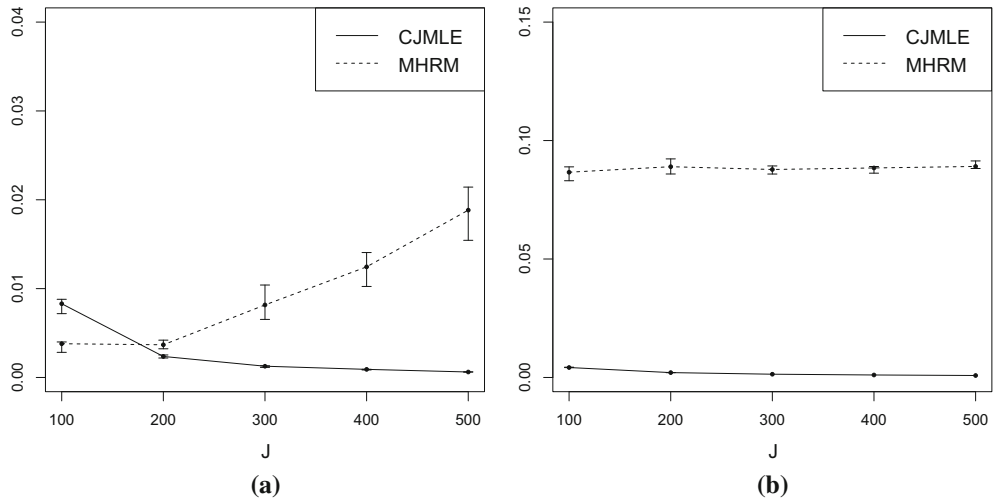

FIGURE 8.
Comparison between the CJMLE and MHRM algorithms. **a** Recovery of loading matrix and **b** recovery of item response probabilities.

TABLE 3.
Speed comparison (in s) between the CJMLE and the MR-HM measured in seconds on a single Intel® core (Xeon® CPU @2.20 GHz; RAM 3.75 GB).

$\tau = 10$	$J = 100$	$J = 200$	$J = 300$	$J = 400$	$J = 500$
CJMLE					
(25% quantile)	1.8	8.6	29.0	68.3	137.9
(50% quantile)	1.9	10.0	29.6	68.7	138.4
(75% quantile)	1.9	10.0	34.7	69.0	139.4
MHRM					
(25% quantile)	142.8	478.7	850.4	1644.8	1997.6
(50% quantile)	163.2	574.4	1077.6	2147.0	2573.4
(75% quantile)	189.7	609.7	1157.8	2403.1	2894.3

Selection of number of factors We first select the latent dimension K using a fivefold cross-validation method, as described in Sect. 2.6. The result is given in Fig. 9, where the smallest cross-validation error is achieved when $K = 3$. This result is consistent with the design of the EPQ-R. In what follows, we report the estimated parameters under the three-factor model.

Three-factor model result To anchor the latent factors, we apply an analytic rotation method, the Geomin rotation (see, e.g., Yates 1988), to the obtained three-factor solution. Geomin is an oblique rotation method that aims at finding a simple pattern of factor loadings without requiring the factors to be orthogonal to each other.

In Fig. 10, we present a heat map of the estimated factor loading matrix in absolute values. As we can see, items in the E, P, and N scales tend to have large absolute loadings on the three estimated factors, respectively. We list the top five items with the highest absolute loadings on each factor in Table 4. These items are all from the corresponding scales and are quite representative of the scales that they belong to. The correspondence between the recovered factors and the Eysenck's three personality traits is further confirmed by the high correlations between the estimated person

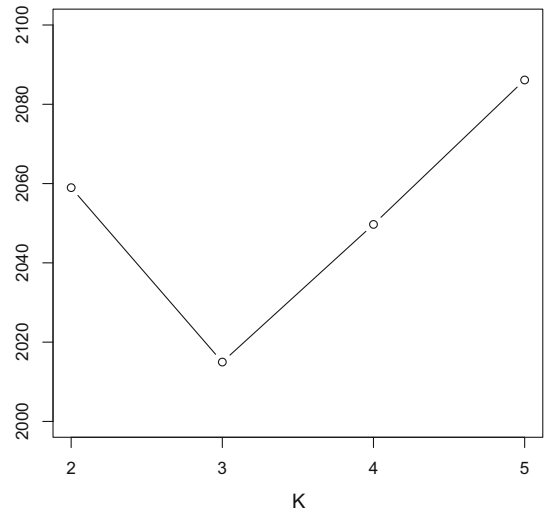

FIGURE 9.
Cross-validation errors for $K = 2, 3, 4, 5$.

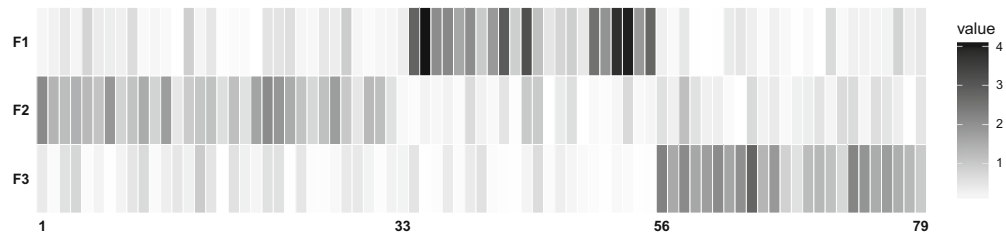

FIGURE 10.
Fitting a three-factor model to the EPQ-R data: heat map of the estimated loading matrix in absolute value under Geomin rotation.

parameters (after rotation) and the corresponding total scores on the three scales, as given in Table 5.

We further investigate the estimated person parameters. In Fig. 11, we show the histograms of the estimated person parameters of each dimension, as well as the scatter plots of the estimated person parameters for each pair of dimensions. According to the histograms, the estimated person parameters on each dimension seem to be unimodal and almost symmetric about the origin. In addition, no obvious person clusters are found according to the scatter plots. Table 6 further shows the correlations between the three estimated factors (after rotation). These correlations are relatively low, suggesting that Eysenck's three personality factors in Eysenck's model tend to be independent of each other.

Finally, a complete table of the estimated loading parameters is provided in the supplementary material.

6. Discussion

In this paper, we develop a statistical theory of joint maximum likelihood estimation under an exploratory item factor analysis framework. In particular, a constrained joint maximum likelihood estimator is proposed that differs from the traditional joint maximum likelihood estimator by adding constraints on the Euclidian norms of both the item-wise and person-wise parameters. It is

TABLE 4.
Fitting a three-factor model to the EPQ-R data: the top five items with highest absolute loadings on each factor, under the Geomin rotation.

Factor	Items	Content
F1	35(E+)	Are you rather lively?
	53(E−)	Do you tend to keep in the background on social occasions?
	52(E+)	Do other people think of you as being very lively?
	44(E+)	Do you like mixing with people?
	42(E+)	Can you easily get some life into a rather dull party?
F2	1(P+)	Would you take drugs which may have strange or dangerous effects?
	21(P−)	Are good manners very important?
	7(P+)	Do you think marriage is old-fashioned and should be done away with?
	22(P−)	Do good manners and cleanliness matter much to you?
	12(P+)	Would you like other people to be afraid of you?
F3	64(N+)	Are you a worrier?
	73(N+)	Do you worry too long after an embarrassing experience?
	56(N+)	Does your mood often go up and down?
	58(N+)	Do you often worry about things you should not have done or said?
	61(N+)	Do you often feel 'fed-up' ?

TABLE 5.
Fitting a three-factor model to the EPQ-R data: the correlations between the estimated person parameters and the corresponding total scores on the three scales.

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
T_1	0.89	0.06	−0.20
T_2	0.11	0.88	−0.02
T_3	−0.14	0.10	0.95

The rows of the table (T_1, T_2, T_3) correspond to the total scores on the three scales, and the columns ($\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$,) correspond to the estimated person parameters (after Geomin rotation).

shown that this estimator consistently recovers the person and item specific response probabilities and also consistently estimates the loading matrix up to a rotation, under an asymptotic regime when both the numbers of participants and items grow to infinity.

An efficient alternating minimization algorithm is proposed for the computation that is scalable to large datasets with tens of thousands of people, thousands of items, and more than ten latent traits. This algorithm iterates between two steps: updating person parameters given item parameters and updating item parameters given person parameters. In each step, the parameters can be updated in parallel for different people/items. A novel projected gradient descent update is used in each step to handle the constraints. Both our theory and computational methods are extended to analyzing data with missing responses.

The proposed method may be extended along several directions. First, the proposed theory and methods will be extended to IFA models for polytomous response data which are commonly encountered in practice. Specifically, we believe that similar theoretical results can be established for multidimensional graded models (e.g., Cai 2010a). More precisely, in a multidimensional graded model with K factors, the latent structure is still reflected by a $J \times K$ loading matrix. This

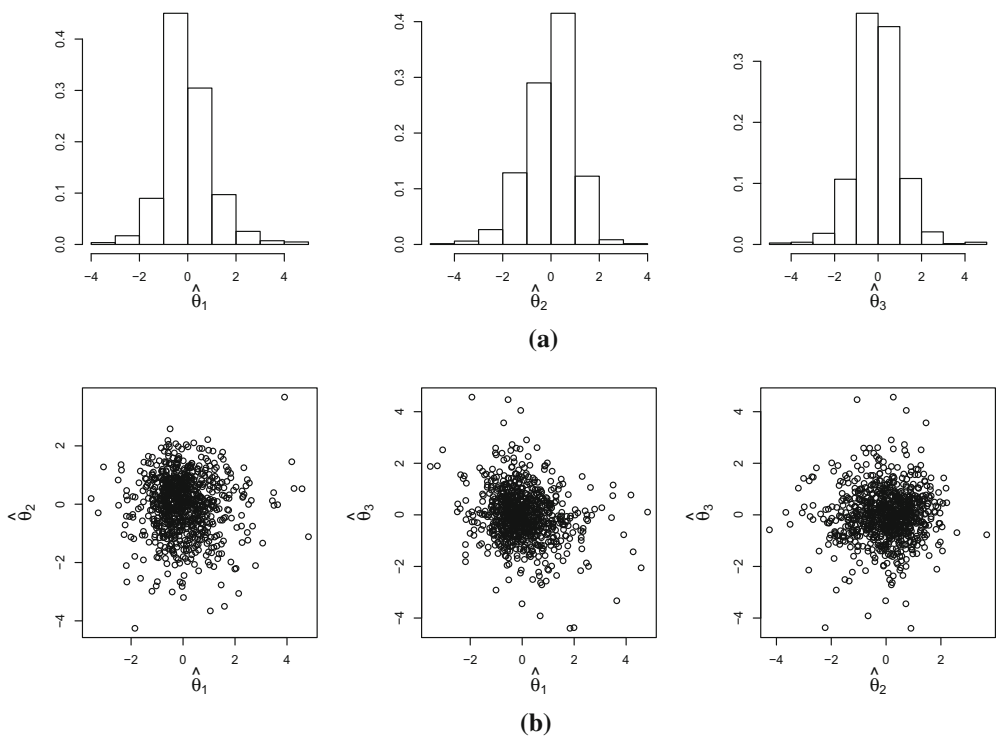

FIGURE 11.
Fitting a three-factor model to the EPQ-R data: histograms of the estimated person parameters (after Geomin rotation) of each dimension, and scatter plots of the estimated person parameters (after Geomin rotation) for each pair of dimensions. **a** Histograms of the estimated person parameters (after rotation) and **b** scatter plots of the estimated person parameters (after rotation).

TABLE 6.
Fitting a three-factor model to the EPQ-R data: the correlations between the estimated person parameters (after Geomin rotation).

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
$\hat{\theta}_1$	1.00	− 0.02	−0.21
$\hat{\theta}_2$	− 0.02	1.00	0.03
$\hat{\theta}_3$	− 0.21	0.03	1.00

loading matrix should still be consistently recovered by a CJMLE, under the same asymptotic regime.

Second, even after applying rotational methods, the obtained factor loading matrix may not be simple (i.e., sparse) enough for a good interpretation. To better pursue a simple loading structure, it may be helpful to further add L_1 regularization of factor loading parameters (Sun et al. 2016) into the current optimization program for CJMLE, under which the estimated factor loading matrix is automatically sparse, and thus, no post hoc rotation is needed. The statistical consistency of this L_1 regularized CJMLE may be further established, for which the issue of rotational indeterminacy may disappear.

Third, the missing responses are assumed to be missing completely at random in our theoretical analysis of missing data. As mentioned earlier, we believe that similar asymptotic properties still hold when relaxing this assumption to missing at random. This is left for future investigation.

Fourth, the current theoretical framework requires the number of latent factors to be known. When it is unknown, we suggest a cross-validation approach for choosing the latent dimension, which turns out to perform well according to our simulation studies and real data analysis. The statistical properties of this approach remain to be investigated. Alternatively, information criteria may be developed for determining the latent dimension.

In summary, this paper is a call to change the stereotype of joint maximum likelihood estimation as a statistically inconsistent method and a call to draw researchers' attention to the development of theory and methods for JML-based estimation. JML-based estimation is generally applicable to almost all latent variable models, easy to program, and computationally efficient. We believe that with a better theoretical understanding, JML-based estimation may become a new paradigm for the statistical analysis of latent variable models, especially for the analysis of complex and large-scale data.

Acknowledgments

We would like to thank the Editor, the Associate Editor, and the reviewers for many helpful and constructive comments. We also would like to thank Dr. Barrett for sharing the EPQ-R dataset analyzed in Sect. 5. This work was partially supported by a NAEd/Spencer Postdoctoral Fellowship [to Yunxiao Chen] and NSF grant DMS 1712657 [to Xiaoou Li].

References

Andersen, E. B. (1973). *Conditional inference and models for measuring*. Copenhagen, Denmark: Mentalhygiejnisk Forlag.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement*, *11*(2), 111–141.

Bartholomew, D. J., Moustaki, I., Galbraith, J., & Steele, F. (2008). *Analysis of multivariate social science data*. Boca Raton, FL: CRC Press.

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541–561.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261–280.

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*(6), 395–414.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33–57.

Cai, L. (2010b). Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Cai, T., & Zhou, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research*, *14*(1), 3619–3647.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Chiu, C.-Y., Köhn, H.-F., Zheng, Y., & Henson, R. (2016). Joint maximum likelihood estimation for diagnostic classification models. *Psychometrika*, *81*(4), 1069–1092.

Dagum, L., & Menon, R. (1998). OpenMP: An industry standard API for shared-memory programming. *Computational Science & Engineering, IEEE*, *5*(1), 46–55.

Davenport, M. A., Plan, Y., van den Berg, E., & Wootters, M. (2014). 1-bit matrix completion. *Information and Inference*, *3*(3), 189–223.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5–18.
Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*(3), 474–497.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
Eysenck, S. B., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, *6*(1), 21–29.
Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statistics & Probability Letters*, *23*(2), 165–170.
Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, *5*(5), 815–841.
Haberman, S. J. (2004). Joint and conditional maximum likelihood estimation for the Rasch model for binary responses. *ETS Research Report Series RR-04-20*.
Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*(3), 347–387.
Lee, K., & Ashton, M. C. (2009). Factor analysis in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology*. New York, NY: Guilford Press.
Lee, S.-Y., Poon, W.-Y., & Bentler, P. M. (1990). A three-stage estimation procedure for structural equation models with polytomous variables. *Psychometrika*, *55*(1), 45–51.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Routledge.
Meng, X.-L., & Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, *91*(435), 1254–1267.
Mislevy, R. J. & Stocking, M. L. (1987). A consumer's guide to LOGIST and BILOG. *ETS Research Report Series RR-87-43*.
Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*(1), 1–32.
Parikh, N., & Boyd, S. (2014). Proximal algorithms. Foundations and Trends. *Optimization*, *1*(3), 127–239.
Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.
Reckase, M. D. (1972). *Development and application of a multivariate logistic latent trait model*. Ph.D. thesis, Syracuse University, Syracuse NY.
Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48.
Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*(3), 533–555.
Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent variable selection for multidimensional item response theory models via L_1 regularization. *Psychometrika*, *81*(4), 921–939.
von Davier, A. (2010). *Statistical models for test equating, scaling, and linking*. New York, NY: Springer.
Wirth, R., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79.
Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, *30*(6), 469–492.
Yates, A. (1988). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. Albany, NY: State University of New York Press.

Manuscript Received: 4 JUL 2017
Published Online Date: 19 NOV 2018