A NOTE ON EXPLORATORY ITEM FACTOR ANALYSIS BY SINGULAR VALUE DECOMPOSITION

HAORAN ZHANG

FUDAN UNIVERSITY

YUNXIAO CHEN

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

XIAOOU LI

UNIVERSITY OF MINNESOTA

We revisit a singular value decomposition (SVD) algorithm given in Chen et al. (Psychometrika 84:124–146, 2019b) for exploratory item factor analysis (IFA). This algorithm estimates a multidimensional IFA model by SVD and was used to obtain a starting point for joint maximum likelihood estimation in Chen et al. (2019b). Thanks to the analytic and computational properties of SVD, this algorithm guarantees a unique solution and has computational advantage over other exploratory IFA methods. Its computational advantage becomes significant when the numbers of respondents, items, and factors are all large. This algorithm can be viewed as a generalization of principal component analysis to binary data. In this note, we provide the statistical underpinning of the algorithm. In particular, we show its statistical consistency under the same double asymptotic setting as in Chen et al. (2019b). We also demonstrate how this algorithm provides a scree plot for investigating the number of factors and provide its asymptotic theory. Further extensions of the algorithm are discussed. Finally, simulation studies suggest that the algorithm has good finite sample performance.

Key words: exploratory item factor analysis, IFA, singular value decomposition, double asymptotics, generalized PCA for binary data.

1. Background

Exploratory IFA (Bock et al. 1988) has been widely used for analyzing item-level data in social and behavioral sciences (Bartholomew et al. 2008). We consider a standard exploratory IFA setting for binary item response data. Let $Y_{ij} \in \{0, 1\}$ be a random variable, denoting individual i's response to item j, where $i = 1, \ldots, N$, and $j = 1, \ldots, J$. Moreover, IFA assumes that an individual i's responses are driven by K latent factors, denoted by $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})^\top$. We consider a general family of multidimensional IFA models (Reckase 2009), which assumes that

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

where $\mathbf{a}_j = (a_{j1}, \ldots, a_{jK})^\top$ is typically known as the loading parameters, d_j is an intercept parameter, and $f : \mathbb{R} \mapsto (0, 1)$ is a pre-specified monotone increasing function which guarantees

(1) to be a valid probability. Using the terminology from generalized linear models, f is called the inverse link function. Note that (1) includes the widely used multidimensional two-parameter logistic (M2PL) model and multidimensional normal ogive model as special cases, for which $f(x) = \exp(x)/(1 + \exp(x))$ and $f(x) = \int_{-\infty}^{x} \exp(-t^2/2)/(2\pi)dt$, respectively. Moreover, we assume local independence; that is, $Y_{i1},..., Y_{iJ}$ are conditionally independent given $\boldsymbol{\theta}_i$. Finally, $\boldsymbol{\theta}_i, i = 1, \ldots, N$, are independent and identically distributed, following an unknown distribution F.

A major focus of exploratory IFA is to estimate the loading matrix $A = (a_{jk})_{J \times K}$, which helps to understand the latent structure underlying the set of items. It is worth noting that the loading matrix can only be recovered up to an oblique rotation (Browne 2001).[1] That is, model (1) will remain unchanged, with a rotated loading vector $\tilde{\mathbf{a}}_j = O^{\top}\mathbf{a}_j$ and $\tilde{\boldsymbol{\theta}}_i = O^{-1}\boldsymbol{\theta}_i$, where O is an $K \times K$ invertible matrix that is also known as an oblique rotation. Recognizing the rotational indeterminacy issue, exploratory IFA typically proceeds in two steps. In the first step, an estimate $\hat{A}$ is obtained, using an arbitrary way to fix the rotation. Then in the second step, analytic rotational methods are applied to $\hat{A}$ to obtain a more sparse loading matrix for better interpretability.

An analytic rotation finds a rotation matrix O such that $\hat{A}O$ minimizes a certain "complexity function," where a lower value of the complexity function indicates more sparsity in the loading matrix (see Browne 2001, for a review of analytic rotations). It implicitly assumes that the true loading matrix has a sparse pattern; i.e., each item is only directly associated with a small number of factors.

In this note, we focus on the first step of exploratory IFA. In particular, we study an estimator given in Chen et al. (2019b) that is based on SVD. Compared to other estimators, this estimator is computationally much faster and does not suffer from convergence issues. It was used to obtain a starting point for a constrained joint maximum likelihood estimator (CJMLE). Simulation studies showed that the convergence of CJMLE can be improved by using the SVD-based estimator as a starting point. Moreover, this SVD-based estimator itself is reasonably accurate when both N and J are large. Thus, it can be used not only as a starting point for the CJMLE, but also as a quick and high-quality solution to large-scale exploratory IFA problems. In what follows, we investigate the statistical properties of this estimator.

2. Main Results

SVD-Based Estimator We restate this SVD-based algorithm below.[2]

Algorithm 1. (SVD-based estimator for exploratory IFA)

1. *Input response $Y = (y_{ij})_{N \times J}$, the number of factors K, inverse link function f, and truncation parameter $\epsilon_{N,J} > 0$.*
2. *Apply the singular value decomposition to Y and obtain $Y = \sum_{j=1}^{J} \sigma_j \mathbf{u}_j \mathbf{v}_j^{\top}$, where $\sigma_1 \geq \cdots \geq \sigma_J \geq 0$ are the singular values, and $\mathbf{u}_j$s and $\mathbf{v}_j$s are left and right singular vectors, respectively.*
3. *Let $X = (x_{ij})_{N \times J} = \sum_{k=1}^{\tilde{K}} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\top}$, where $\tilde{K} = \max\left\{K + 1, \underset{k}{\operatorname{argmax}}\{\sigma_k \geq 1.01\sqrt{N}\}\right\}$.*

---

[1]We only discuss oblique rotation here, as our general exploratory IFA model does not require the factors to be uncorrelated. If the factors are further required to be uncorrelated, then the loading matrix can be recovered up to an orthogonal rotation, for which the rotation matrix O is an orthogonal matrix (e.g., Kaiser 1958).

[2]The original algorithm was described in the supplementary material of Chen et al. (2019b). The algorithm here is a slightly modified version. The major modification is in Step 3 of the algorithm that requires at least $K + 1$ singular values to be retained. This modification can improve the finite-sample performance of the algorithm; see Remark 4 for more discussions. The other modifications are mainly to simplify the exposition of the algorithm.

4. *Let $\hat{X} = (\hat{x}_{ij})_{N \times J}$ be defined as*

$$\hat{x}_{ij} = \begin{cases} \epsilon_{N,J}, & \text{if } x_{ij} < \epsilon_{N,J}, \\ x_{ij}, & \text{if } \epsilon_{N,J} \leq x_{ij} \leq 1 - \epsilon_{N,J}, \\ 1 - \epsilon_{N,J}, & \text{if } x_{ij} > 1 - \epsilon_{N,J}. \end{cases}$$

5. *Let $\tilde{M} = (\tilde{m}_{ij})_{N \times J}$, where $\tilde{m}_{ij} = f^{-1}(\hat{x}_{ij})$.*
6. *Let $\hat{\mathbf{d}} = (\hat{d}_1, \ldots, \hat{d}_J)$, where $\hat{d}_j = (\sum_{i=1}^{N} \tilde{m}_{ij})/N$.*
7. *Apply singular value decomposition to $\hat{M} = (\tilde{m}_{ij} - \hat{d}_j)_{N \times J}$ to have $\hat{M} = \sum_{j=1}^{J} \hat{\sigma}_j \hat{\mathbf{u}}_j \hat{\mathbf{v}}_j^\top$, where $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_J \geq 0$ are the singular values, and $\hat{\mathbf{u}}_j$s and $\hat{\mathbf{v}}_j$s are the left and right singular vectors, respectively.*
8. *Output $\hat{A} = \frac{1}{\sqrt{N}}(\hat{\sigma}_1 \hat{\mathbf{v}}_1, \ldots, \hat{\sigma}_K \hat{\mathbf{v}}_K)$, $\hat{\Theta} = \sqrt{N}(\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_K)$.*

Remark 1. SVD is a powerful tool for the factorization of rectangular matrices that has been widely used in multivariate statistics for the dimension reduction in data (Wall et al. 2003). Thanks to the mathematical properties of SVD, the estimator given by Algorithm 1 is analytic that does not suffer from convergence issues. On the other hand, as the objective functions of the CJMLE and the marginal maximum likelihood estimator (MMLE; Bock and Aitkin 1981) are nonconvex, there is no guarantee for finding their global optima. In addition, this SVD approach is also much faster than the other estimators, including the CJMLE and MMLE. In particular, the computation of the MMLE based on the vanilla expectation maximization algorithm is not affordable when the latent dimension K is of a moderate size (e.g., $K \geq 5$). Even the stochastic algorithms for the MMLE (Cai 2010a; 2010b; Zhang et al. 2020) and the alternating minimization algorithm for the CJMLE (Chen et al. 2019b; 2019c) are much slower than the SVD algorithm, as these algorithms typically need a large number of iterations to converge. A speed comparison is provided in the simulation study between the SVD method and the CJMLE.

Remark 2. Algorithm 1 can be viewed as a generalization of PCA to binary data. PCA is an SVD-based algorithm (e.g., Chapter 14, Friedman et al. 2001) that is fast and commonly used for exploratory linear factor analysis. Unfortunately, PCA cannot be applied to exploratory IFA, due to the nonlinear link function in IFA models. Unlike PCA which applies SVD only once, Algorithm 1 applies SVD twice. The first application of SVD and the inverse transformation (Steps 2–5) denoise and linearize the data. Then, the second application of SVD (Steps 6–7) is essentially doing PCA to the linearized data.

Remark 3. Similar as the CJMLE (Chen et al. 2019b; 2019c), this SVD-based estimator does not require the latent distribution F to be known or to take a parametric form as is required in the MMLE approach. Moreover, exploratory IFA based on tetrachoric/polychoric correlations (Muthén 1984; Lee et al. 1990; Lee et al. 1992; Jöreskog 1994) or composite-likelihood-based estimator (Katsikatsou et al. 2012) requires F to be multivariate normal, with the former approach further requiring the inverse link f to be probit. In this sense, the SVD-based estimator and the CJMLE require less model assumptions than the other estimators. As a price, their consistency requires stronger conditions, specifically, a double asymptotic regime where both N and J diverge.

Remark 4. Steps 2–4 of the algorithm essentially follow the same procedure of Chatterjee (2015) for matrix estimation. We thus refer the readers to Chatterjee (2015) for the details. A small difference is that we require $\tilde{K} \geq K + 1$ in Step 3 of the algorithm. This modification does not affect the asymptotic behavior of the estimator. However, it can improve the finite-sample performance when N and J are not large enough. Intuitively, we need $\tilde{K}$ to be at least $K + 1$, in

order to recover the matrix $(d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i)_{N \times J}$ which is of rank $K + 1$. The constant 1.01 in Step 3 of the algorithm follows Theorem 1.1 of Chatterjee (2015), which makes use of the fact that $Var(Y_{ij}) \leq 1/4$. This constant can be replaced by any fixed constant in the open interval (1, 1.5), without affecting its consistency given in Theorem 1. We set it to be 1.01, because according to Theorem 1.1 of Chatterjee (2015) this constant should be chosen close to 1 for better accuracy.

Remark 5. The truncation step (Step 4) is necessary, as it guarantees the existence of a solution. This is because, even though x_{ij} in Step 3 is approximating the true probability $\Pr(Y_{ij} = 1)$, it is not guaranteed to be in the interval (0, 1). As a consequence, $f^{-1}(x_{ij})$ may not be well defined. The pre-specified truncation parameter $\epsilon_{N,J} > 0$ determines the truncation level. As shown in the sequel, the choice of $\epsilon_{N,J}$ affects the statistical consistency of the proposed algorithm. Under certain circumstances, we will need the truncation parameter $\epsilon_{N,J}$ to decay to zero as N and J grow to infinity, which is why we attach subscripts N and J to the truncation parameter. In practice, the performance of the proposed method tends to be insensitive to the choice of $\epsilon_{N,J}$ when it is chosen sufficiently small, which is justified theoretically by Propositions 1 and 2, under two specific settings. In the numerical analysis of this paper, we use $\epsilon_{N,J} = 10^{-4}$ as a default value.

Statistical Consistency In what follows, we establish the theoretical consistency of this method. In particular, we show that this SVD-based algorithm is consistent under similar asymptotic setting and notion of consistency as in Chen et al. (2019b) and Chen et al. (2019c). The proofs of our theoretical results are given in the supplementary material. More precisely, we consider a loss function on the recovery of the true loading matrix $A^* = (a_{jk}^*)_{J \times K}$ up to an oblique rotation

$$L_{N,J}(A^*, \hat{A}) = \min_{O \in \mathbb{R}^{K \times K}} \left\{ \frac{\|A^* - \hat{A}O\|_F^2}{JK} \right\}, \tag{2}$$

where the subscripts N and J are used to emphasize that the loss function depends on the sample size N and the number of items J, and $\|X\|_F = \sqrt{\sum_i \sum_j x_{ij}^2}$ denotes the Frobenius norm of a matrix $X = (x_{ij})$. Under mild technical conditions and a double asymptotic setting where both N and J grow to infinity, we show that the loss function $L_{N,J}(A^*, \hat{A})$ converges to zero in probability. The regularity conditions and the consistency result are formally described in Theorem 1, with two special cases discussed in the sequel. Similar double asymptotic settings have been considered in psychometric research, including the analyses of unidimensional IRT models (Haberman 1977; 2004) and diagnostic classification models (Chiu et al. 2016). The following regularity conditions are needed for our main result in Theorem 1. As will be discussed in the sequel, these conditions are mild.

A1. There exists a constant C such that $\sqrt{(d_j^*)^2 + \|\mathbf{a}_j^*\|^2} \leq C$, for $j = 1, \dots, J$, where d_j^* and $\mathbf{a}_j^*$ are the true item parameters.

A2. The true person parameters $\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_N^*$ are independent and identically distributed (i.i.d.) following a distribution F which has mean $\mathbf{0}$ and positive definite covariance matrix Σ.

A3. The inverse link function f is strictly monotone increasing, continuously differentiable, and Lipschitz continuous with Lipschitz constant L. We further assume that

$$\lim_{x \to -\infty} f(x) = 0, \quad \text{and} \quad \lim_{x \to \infty} f(x) = 1.$$

A4. There exists a constant C_1, such that the Kth singular value of A^*, denoted by $\sigma_K(A^*)$, satisfies $\sigma_K(A^*) \geq C_1\sqrt{J}$ for all J.

A5. The sample size N is no less than the number of items J, i.e., $N \geq J$.

Theorem 1. *Suppose that conditions A1–A5 are satisfied. Further suppose that $\epsilon_{N,J} \leq \frac{1}{5}$ and satisfies*

$$\Pr\left(\|\boldsymbol{\theta}_1^*\| \geq h(2\epsilon_{N,J})/C\right) = o(N^{-1}), \tag{3}$$

$$\frac{(h(2\epsilon_{N,J}))^{\frac{K+1}{K+3}}}{(\epsilon_{N,J}g(\epsilon_{N,J}))^2} = o(J^{\frac{1}{K+3}}), \tag{4}$$

where

$$h(y) = \max\{|f^{-1}(y)|, |f^{-1}(1-y)|\}, \quad y \in (0, 0.5), \tag{5}$$

$$g(y) = \inf\{f'(x) : x \in [f^{-1}(y), f^{-1}(1-y)]\}, \quad y \in (0, 0.5). \tag{6}$$

Then, the estimate $\hat{A}$ given by Algorithm 1 satisfies $L_{N,J}(A^*, \hat{A}) \xrightarrow{pr} 0$, as $N, J \to \infty$.

Remark 6. We remark that the notion of consistency for the estimation of the loading matrix is weaker than that in the traditional sense, since the loss function (2) is an average of the entrywise losses when J grows. Let $\tilde{O}$ minimize the right-hand side of (2), and let $\tilde{A} := (\tilde{a}_{jk})_{J \times K} = \hat{A}\tilde{O}$. Then, (2) converges to 0 means that for any $\epsilon > 0$, $(\sum_{j=1}^{J}\sum_{k=1}^{K} 1_{\{|a_{jk}^* - \tilde{a}_{jk}| > \epsilon\}})/JK$ also converges to 0. That is, the proportion of inaccurately estimated loading parameters converges to zero in probability under the optimal rotation. Due to the double asymptotic setting, our theoretical result only suggests the sensible use of the SVD-based algorithm when the sample size N and the number of items J are both large.

Remark 7. It has been well understood that PCA can consistently estimate a linear factor model under a similar double asymptotic setting (Stock and Watson 2002), which provides the theoretical justification for the use of PCA in exploratory linear factor analysis. Theorem 1 can be viewed as a similar result for exploratory item factor analysis.

Remark 8. We provide some discussions on the regularity conditions required in Theorem 1. Assumption A1 requires that the parameters of each item, including the intercept and slope parameters, should not be too large. That is, the presence of an extreme item is likely to distort the analysis. Assumption A2 is a very standard assumption in exploratory IFA. It is more flexible than many exploratory IFA settings, as it does not require the distribution F to be multivariate normal.

Assumption A3 is satisfied by the logistic and probit link functions, two most commonly used link functions in exploratory IFA, but it excludes, for example, the multidimensional version of the three-parameter logistic model, as a special case. Assumption A4 requires that there is sufficient variability in the items. The same assumption is also required in Chen et al. (2019b) and Chen et al. (2019c). In fact, this assumption is satisfied with probability tending to one, when the true loadings $\mathbf{a}_j^*$ are i.i.d. samples from a K-variate distribution whose covariance matrix is non-degenerate. Finally, assumption A5 is practically reasonable, as in large-scale measurement, the sample size is usually larger than the number of items. Since people and items are almost mathematically symmetric in the IFA model, similar asymptotic results can be derived when $J \geq N$.

Remark 9. We further provide some intuitions on the reason why the algorithm works. Steps 2–4 essentially follow the same procedure of Chatterjee (2015) for matrix estimation. The procedure guarantees the loss $\sum_{i,j}(f(d_j^* + (\mathbf{a}_j^*)^\top \boldsymbol{\theta}_i^*) - \hat{x}_{ij})^2/(NJ)$ to be small with high probability, where d_j^* and $\mathbf{a}_j^*$ denote the true item-specific parameters and $\boldsymbol{\theta}_i^*$ denotes the true person parameters sampled from distribution F. Further with conditions A1 and A3, Steps 5 and 6 guarantee the average loss $\sum_{i=1}^{N}\sum_{j=1}^{J}((\mathbf{a}_j^*)^\top \boldsymbol{\theta}_i^* - \hat{\mathbf{a}}_j^\top \hat{\boldsymbol{\theta}}_i)^2/(NJ)$ to be small with high probability. Finally, under conditions A2 and A4, the famous Davis–Kahan–Wedin theorem from matrix perturbation theory (see e.g., Stewart and Sun 1990; O'Rourke et al. 2018) guarantees that $L_{N,J}(A^*, \hat{A})$ is small with high probability.

Remark 10. Equations (3) and (4) are requirements on the truncation parameter $\epsilon_{N,J}$, which depends on both the tail of distribution F and the properties of the inverse link function. Roughly speaking, Equation (3) is saying that $\epsilon_{N,J}$ cannot be too large. This is because, given F and f, the probability in (3) is increasing in $\epsilon_{N,J}$. Requiring the probability being $o(N^{-1})$ implies that $\epsilon_{N,J}$ cannot be large. This requirement is intuitive, because $\tilde{M}$ can be a poor approximation to $M^* = (m_{ij}^*)_{N\times J} := (d_j^* + (\mathbf{a}_j^*)^\top \boldsymbol{\theta}_i^*)_{N\times J}$, when many entries of M^* are larger than $h(\epsilon_{N,J})$. The function $h(\cdot)$ transforms the truncation on x_{ij} to a truncation on $\tilde{m}_{ij}$. Using $h(2\epsilon_{N,J})$ instead of $h(\epsilon_{N,J})$ is for technical reasons.

Equation (4) requires that $\epsilon_{N,J}$ cannot be too small, as the left-hand side of (4) is decreasing in $\epsilon_{N,J}$. This requirement is also intuitive. Note that $|\tilde{m}_{ij}| \leq h(\epsilon_{N,J})$, where $h(\epsilon_{N,J})$ is decreasing in $\epsilon_{N,J}$. Therefore, a sufficiently large choice of $\epsilon_{N,J}$ avoids the approximation error $\|\tilde{M} - M^*\|_F$ being too large when there exist some extreme estimates $\tilde{m}_{ij}$. Function $g(\cdot)$ measures the local flatness of the inverse link f. The true matrix M^* is more difficult to estimate when $g(\epsilon_{N,J})$ is smaller. This is because $|\tilde{m}_{ij} - m_{ij}^*|$ can be large, even when $|\hat{x}_{ij} - f(m_{ij}^*)|$ is small, due to the local flatness of the inverse link function.

Remark 11. We take a stochastic design for the true person parameters and a fixed design for the true item parameters, following the convention of item factor analysis (e.g., Bartholomew et al. 2008). It is worth pointing out that whether taking a stochastic or fixed design is not essential under our double asymptotic regime. For example, the consistent result of Theorem 1 still holds, if we can replace condition A2 by a corresponding fixed design as in Chen et al. (2019b).

Following the discussion on $\epsilon_{N,J}$ in Remark 10, we consider two concrete settings under which the requirement on $\epsilon_{N,J}$ becomes more specific. These results are given in Propositions 1 and 2.

Proposition 1. *Suppose that F has a compact support. More precisely, there exists a constant C_0, satisfying*

$$\Pr(\|\boldsymbol{\theta}_1^*\| \geq C_0) = 0,$$

under the law of F. If we fix $\epsilon_{N,J}$ to be a constant ϵ independent of N and J, satisfying

$$0 < \epsilon \leq \frac{1}{2}\min\left\{1 - f\left(C\sqrt{C_0^2 + 1}\right), f\left(-C\sqrt{C_0^2 + 1}\right), \frac{2}{5}\right\}, \tag{7}$$

then (3) and (4) are satisfied. This choice of $\epsilon_{N,J}$, together with the regularity conditions in Theorem 1, guarantees $L_{N,J}(A^*, \hat{A})$ to converge to zero in probability.

Proposition 2. *Consider exploratory IFA based on the M2PL model, where F is a multivariate sub-Gaussian distribution[3] and f is the logistic link. Suppose that there exists a constant $\beta \geq 1$ such that*

$$J \leq N \leq J^\beta. \tag{8}$$

Then,

(3) *and* (4) *hold, for any $\epsilon_{N,J}$ taking the form*

$$\epsilon_{N,J} = \gamma_0 J^{-\gamma_1}, \tag{9}$$

where γ_0 and γ_1 are any constants satisfying $\gamma_0 > 0$ and $\gamma_1 \in (0, (4(K+3))^{-1})$. The choice of $\epsilon_{N,J}$ following (9), *together with the regularity conditions in Theorem* 1, *guarantees $L_{N,J}(A^*, \hat{A})$ to converge to zero in probability.*

According to the result of Proposition 1, it suffices to choose $\epsilon_{N,J}$ as a sufficiently small positive constant, when F has a bounded support. Under the setting of Proposition 2, to ensure consistency, one has to let $\epsilon_{N,J}$ decay to zero at an appropriate rate. Note that even in the second setting where the support of F is unbounded, $\epsilon_{N,J}$ is almost like a constant, as it decays to zero very slowly when J grows. These results suggest that we may choose $\epsilon_{N,J}$ to be a sufficiently small constant in practice.

On the Choice of K In the previous discussion, the number of factors K is assumed to be known. In practice, however, this information is often unknown and an important task in exploratory IFA is to determine the number of factors based on data. When conducting exploratory linear factor analysis, one typically gains the first idea by examining the scree plot from principal component analysis. Thanks to the connection between Algorithm 1 and PCA as discussed in Remark 2, a similar scree plot is available from the current method.

The scree plot is produced as follows. We first run Algorithm 1, but replace the unknown K in Step 1 of the algorithm by a reasonably large number $K^\dagger$. Then, a scree plot can be obtained by plotting $\hat{\sigma}_k$ in a descending order, for $\hat{\sigma}_k$s produced by Step 7 of Algorithm 1. Figure 1 shows such a scree plot, for which the data are generated from a five-factor model ($K = 5$) with $J = 200$ and $N = 4000$, and the input number of factors is set to be $K^\dagger = 10$ in Step 1 of the algorithm. Unsurprisingly, an obvious gap is observed between $\hat{\sigma}_5$ and $\hat{\sigma}_6$. In fact, when data follow an IFA model, such a gap in the singular values is guaranteed to exist asymptotically, no matter what the input dimension is. In practice, the latent dimension K can be chosen by identifying the singular value gap from the scree plot.

Theorem 2. *Under the same conditions as Theorem* 1 *and when the input dimension $K^\dagger$ in Algorithm* 1 *is set fixed (i.e., independent of N and J) but not necessarily equal to the true number of factors, there exists a constant $\delta > 0$ such that for the true number of factors K,*

$$\lim_{N,J \to \infty} \Pr\left(\frac{\hat{\sigma}_K}{\sqrt{NJ}} > \delta\right) = 1, \text{ and } \frac{\hat{\sigma}_{K+1}}{\sqrt{NJ}} \xrightarrow{pr} 0,$$

as N and J grow to infinity simultaneously.

---

[3] We say the distribution of a K-variate random vector $\boldsymbol{\theta}$ is sub-Gaussian, if there exist constants $b_1, b_2 > 0$ such that for any $u \in \mathbb{R}^K$, $\|\mathbf{u}\| = 1$ and $t > 0$, $\Pr(|\mathbf{u}^\top \boldsymbol{\theta}| > t) \leq b_1 e^{-b_2 t^2}$. In particular, the multivariate normal distribution is sub-Gaussian.
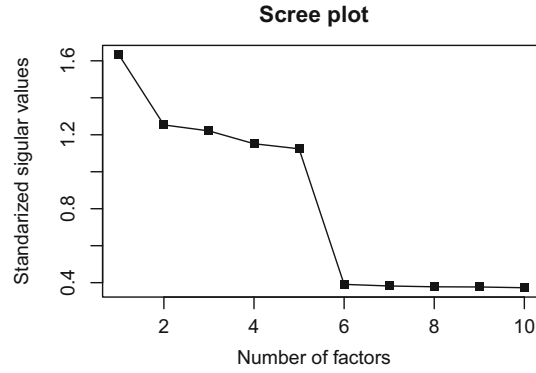
FIGURE 1.
A scree plot for choosing the number of factors. The y-axis shows the standardized singular values $\hat{\sigma}_k/\sqrt{NJ}$, where $\hat{\sigma}_k$s are obtained from Step 7 of Algorithm 1. The data are simulated from an IFA model with $K = 5$, $J = 200$, and $N = 4000$. The input dimension is set to be 10 in Algorithm 1. A singular value gap can be found between the 5th and 6th singular values

Remark 12. As shown in the proof, the input dimension $K^{\dagger}$ does not affect the asymptotics, as long as it does not grow with N and J. However, for relatively small N and J, X obtained in Step 3 of the algorithm may not reserve enough information when the input dimension is smaller than $K + 1$, which may lead to an underestimation of the number of factors. Thus, in practical applications, we recommend to choose the input dimension to be slightly larger than the maximum number of factors one suspects to exist in the data.

Statistical Efficiency We further point out that a price is paid for the computational advantage of the SVD-based estimator. To elaborate on this point, we compare it with the CJMLE (Chen et al. 2019b; 2019c). The CJMLE treats both item parameters and latent factors as fixed parameters and maximizes a joint likelihood function with respect to all the fixed parameters. The SVD-based estimator is statistically less efficient than the CJMLE, in the sense that the SVD-based estimator converges to the true parameters in a much slower rate. To make this comparison, we consider the same setting as in Proposition 1. The following proposition establishes the convergence rate for $\|X^* - \hat{X}\|_F^2/NJ$, which determines the convergence of $\hat{A}$. Here, $X^* = (f(d_j^* + \mathbf{a}_j^*(\boldsymbol{\theta}_i^*)^{\top}))_{N \times J}$ is the true item response probability matrix.

Proposition 3. *Suppose that the same assumptions as in Proposition 1 hold and choose $\epsilon_{N,J}$ as in Proposition 1. Then, we have*

$$\frac{1}{NJ}\|X^* - \hat{X}\|_F^2 = O_p(J^{-\frac{1}{K+2}}). \tag{10}$$

On the other hand, as shown in Chen et al. (2019c), the CJMLE achieves the optimal rate (in minimax sense) for estimating X^*, that is, $\|X^* - \hat{X}_{JML}\|_F^2/(NJ) = O_p(J^{-1})$, where $\hat{X}_{JML}$ denotes the CJMLE. This result suggests that the SVD-based estimator converges in a much slower rate than the CJMLE.

3. Extensions

Dealing with Missing Data With slight modification, Algorithm 1 can handle item response data with missing values. We use matrix $W = (w_{ij})_{N \times J}$ to indicate the data nonmissingness, where

$w_{ij} = 1$ indicates the response Y_{ij} is not missing and $w_{ij} = 0$ otherwise. The modified algorithm is described as follows.

Algorithm 2. (SVD-based estimator for exploratory IFA with missing data)

1. *Input nonmissing indicator $W = (w_{ij})_{N \times J}$, nonmissing responses $\{y_{ij} : w_{ij} = 1, i = 1, \ldots, N, j = 1, \ldots, J\}$, the number of factors K, inverse link function f, and truncation parameter $\epsilon_{N,J} > 0$.*
2. *Compute $\hat{p} = (\sum_{i=1}^{N} \sum_{j=1}^{J} w_{ij})/(NJ)$ as the proportion of observed responses.*
3. *For each i and j, let $z_{ij} = y_{ij}$, if $w_{ij} = 1$, and $z_{ij} = 0$ if $w_{ij} = 0$.*
4. *Apply the singular value decomposition to Z to obtain $Z = \sum_{j=1}^{J} \sigma_j \mathbf{u}_j \mathbf{v}_j^{\top}$, where $\sigma_1 \geq \cdots \geq \sigma_J \geq 0$ are the singular values and $\mathbf{u}_j$s and $\mathbf{v}_j$s are left and right singular vectors, respectively.*
5. *Let*

$$X = (x_{ij})_{N \times J} = \frac{1}{\hat{p}} \sum_{k=1}^{\tilde{K}} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\top},$$

   where $\tilde{K} = \max \left\{ K + 1, \underset{k}{\operatorname{argmax}}\{\sigma_k \geq 1.01\sqrt{N(\hat{p} + 3\hat{p}(1 - \hat{p}))}\} \right\}$.
6. *Let $\hat{X} = (\hat{x}_{ij})_{N \times J}$ be defined as*

$$\hat{x}_{ij} = \begin{cases} \epsilon_{N,J}, & \text{if } x_{ij} < \epsilon_{N,J}, \\ x_{ij}, & \text{if } \epsilon_{N,J} \leq x_{ij} \leq 1 - \epsilon_{N,J}, \\ 1 - \epsilon_{N,J}, & \text{if } x_{ij} > 1 - \epsilon_{N,J}. \end{cases}$$

7. *Let $\tilde{M} = (\tilde{m}_{ij})_{N \times J}$, where $\tilde{m}_{ij} = f^{-1}(\hat{x}_{ij})$.*
8. *Let $\hat{\mathbf{d}} = (\hat{d}_1, \ldots, \hat{d}_J)$, where $\hat{d}_j = (\sum_{i=1}^{N} \tilde{m}_{ij})/N$.*
9. *Apply singular value decomposition to $\hat{M} = (\tilde{m}_{ij} - \hat{d}_j)_{N \times J}$ to have $\hat{M} = \sum_{j=1}^{J} \hat{\sigma}_j \hat{\mathbf{u}}_j \hat{\mathbf{v}}_j^{\top}$, where $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_J \geq 0$ are the singular values and $\hat{\mathbf{u}}_j$s and $\hat{\mathbf{v}}_j$ are the left and right singular vectors, respectively.*
10. *Output $\hat{A} = \frac{1}{\sqrt{N}}(\hat{\sigma}_1 \hat{\mathbf{v}}_1, \ldots, \hat{\sigma}_K \hat{\mathbf{v}}_K)$, $\hat{\Theta} = \sqrt{N}(\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_K)$.*

Remark 13. It is easy to see that $\hat{p} = 1$ when there is no missing data. In that case, Algorithm 2 becomes exactly the same as Algorithm 1. Steps 2–5 essentially follow the same procedure of Chatterjee (2015) for matrix completion, and the rest of the steps are the same as those in Algorithm 1. Specifically, missing data are first imputed by zero in Step 3 of the algorithm. The bias brought by the simple imputation procedure is corrected in Step 5, by multiplying the factor $1/\hat{p}$. Similar to Algorithm 1, the choice of $\tilde{K}$ in Step 5 is determined by the procedure of Chatterjee (2015) with a small modification which guarantees $\tilde{K} \geq K + 1$.

In fact, when the entries of the item response matrix are missing completely at random, using a similar proof, one can show that $\hat{A}$ given by Algorithm 2 is still consistent, under some mild condition on the missing data mechanism and the same conditions as in Theorem 1. Specifically, the following condition is needed, in addition to conditions A1–A5.

A6. The w_{ij}s are independent and identically distributed from a Bernoulli distribution with $\Pr(w_{ij} = 1) = p$, where $0 < p \leq 1$ is a constant which does not depend on N and J.

Under conditions A1–A6, the following proposition holds that guarantees the consistency of the proposed SVD estimator.

Proposition 4. *Under the same conditions as Theorem* 1 *plus condition A6, the estimate $\hat{A}$ given by Algorithm* 2 *satisfies $L_{N,J}(A^*, \hat{A}) \xrightarrow{pr} 0$, as $N, J \to \infty$.*

Dealing with Ordinal Data In exploratory IFA, ordinal data are also commonly encountered, due to the wide use of Likert-scale items. With slight modification, the SVD method can also be used to analyze ordinal data. This is achieved by applying Algorithm 1 to multiple dichotomized versions of data.

More precisely, consider data $Y = (Y_{ij})_{N \times J}$, where $Y_{ij} \in \{0, 1, \ldots, T\}$. We consider a general family of graded response-type models:

$$\Pr(Y_{ij} \geq t | \boldsymbol{\theta}_i) = f(d_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i), \tag{11}$$

where d_{jt} is an item- and category-specific intercept parameter, and the rest of the notations are the same as that of model (1). Note that the linear combination of the factors $\mathbf{a}_j^\top \boldsymbol{\theta}_i$ does not depend on the response category and appears in all the submodels $\Pr(Y_{ij} \geq t | \boldsymbol{\theta}_i)$ for $t = 1, \ldots, T$. When $f(x) = \exp(x)/(1 + \exp(x))$ takes the logistic form, model (11) becomes the multidimensional graded response model (Muraki and Carlson 1995).

Model (11) is closely related to the general model (1) for binary data. In fact, if we dichotomize data at response category t, i.e., $Y_{ij}^{(t)} = 1_{\{Y_{ij} \geq t\}}$, then binary data $Y_{ij}^{(t)}$ follows model (1) with the same loading parameters. Therefore, the loading matrix A can be estimated by applying Algorithm 1 to dichotomized data $Y^{(t)} = (1_{\{y_{ij} \geq t\}})_{N \times J}$, for some $t = 1, \ldots, T$. The estimation accuracy may be further improved by aggregating the results from multiple dichotomized versions of data. This aggregation method is summarized by Algorithm 3.

Algorithm 3. (SVD-based estimator for exploratory IFA with ordinal data)

1. *Input response $Y = (y_{ij})_{N \times J}$, the number of categories T, the number of factors K, inverse link function f, and truncation parameter $\epsilon_{N,J} > 0$.*
2. *For $t = 1, \ldots, T$, apply Algorithm 1 to dichotomized data $Y^{(t)} = (1_{\{y_{ij} \geq t\}})_{N \times J}$ and obtain $\hat{M}^{(t)}$ from Step 7 of Algorithm 1.*
3. *Let $\hat{M} = (\sum_{t=1}^T \hat{M}^{(t)})/T$. Apply singular value decomposition to $\hat{M}$ and obtain $\hat{M} = \sum_{j=1}^J \hat{\sigma}_j \hat{\mathbf{u}}_j \hat{\mathbf{v}}_j^\top$, where $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_J \geq 0$ are the singular values and $\hat{\mathbf{u}}_j$s and $\hat{\mathbf{v}}_j$ are left and right singular vectors, respectively.*
4. *Output $\hat{A} = \frac{1}{\sqrt{N}}(\hat{\sigma}_1 \hat{\mathbf{v}}_1, \ldots, \hat{\sigma}_K \hat{\mathbf{v}}_K)$, $\hat{\Theta} = \sqrt{N}(\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_K)$.*

4. Simulation

Simulation Setting We consider $K = 4$ and 8, $J = 200, 400, 600, 800, 1000$, and 1200, and $N = 20J$. For each combination of N, J, and K, two different latent distributions F are considered, one is a K-variate standard normal distribution, and the other is a K-variate normal distribution $N(\mathbf{0}, (\sigma_{ij})_{K \times K})$, where $\sigma_{ij} = 1$ if $i = j$ and $\sigma_{ij} = 0.3$ if $i \neq j$. The inverse link f is chosen to be logistic, i.e., $f(x) = \exp(x)/(1 + \exp(x))$. This leads to 24 different simulation settings, for all possible combinations of N, J, K, and F.

For each simulation setting, 100 independent replications are generated, with the item parameters keeping fixed across replications. When $J = 200$ and given K, the item parameters are generated as follows.
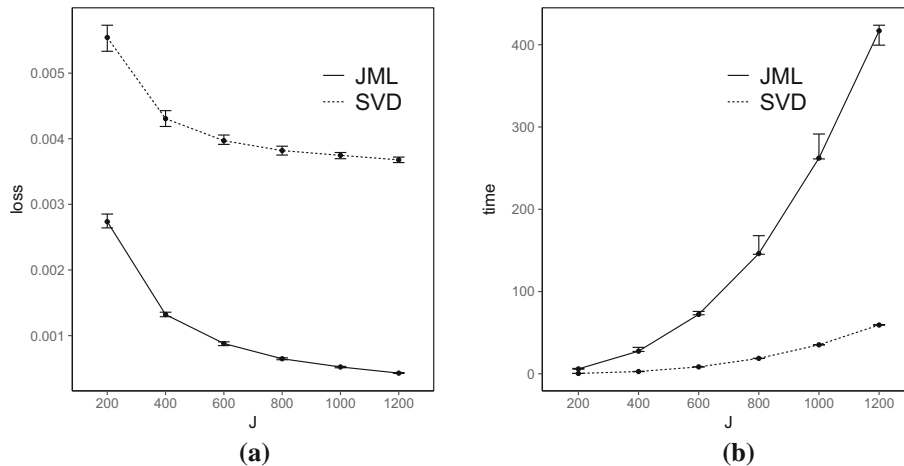
FIGURE 2.
Simulation results when $K = 4$ and the true factors are independent. Panel **a** shows the number of items J in x-axis versus the loss (2) in y-axis, and Panel **b** shows the number of items J in x-axis versus the computation time (in seconds) in y-axis. For each metric and each method, we show the median, 25% quantile, and 75% quantile based on the 100 independent replications

1. $d_1^*,\ldots, d_{200}^*$ are i.i.d. from a uniform distribution over interval $[-1, 1]$.
2. $\mathbf{a}_1^*,\ldots, \mathbf{a}_{200}^*$ are i.i.d., with $\mathbf{a}_j^* = (a_{j1}^{\dagger} q_{j1}, \ldots, a_{jK}^{\dagger} q_{jK})^{\top}$. Here, $a_{jk}^{\dagger}$s are i.i.d. from a uniform distribution over interval $[1, 2]$, and $\mathbf{q}_j = (q_{j1}, \ldots, q_{jK})^{\top}$ are i.i.d. from a uniform distribution over $\mathcal{Q}_K$. Specifically,

$$\mathcal{Q}_4 = \left\{ (q_1, \ldots, q_4)^{\top} : q_k \in \{0, 1\}, \sum_{k=1}^{4} q_k \geq 1, \text{ and } \sum_{k=1}^{4} q_k \leq 3 \right\},$$

and

$$\mathcal{Q}_8 = \left\{ (q_1, \ldots, q_8)^{\top} : q_k \in \{0, 1\}, \sum_{k=1}^{8} q_k \geq 1, \text{ and } \sum_{k=1}^{8} q_k \leq 3 \right\}.$$

The $\mathbf{q}_j$s lead to sparse loading vectors.

When $J > 200$, we set the item parameters by repeating multiple times the parameters under $J = 200$ and the same K. For example, when $J = 400$, we set parameters for items 1–200 and those for items 201–400 to be the same as the parameters generated under the setting $J = 200$.
Results Each simulated dataset is analyzed using the SVD-based estimator, with the truncation parameter $\epsilon_{N,J}$ set to be 10^{-4}. The performance of the SVD-based estimator is compared with that of the CJMLE.[4] The results are shown in Figs. 2, 3, 4, and 5.

The loss for the SVD-based estimator decreases when N and J simultaneously grow, under all settings. Reasonable accuracy can be achieved when N and J are reasonably large, in which case the SVD-based estimator may be directly used for data analysis. For example, under the setting that $K = 4$ and F is multivariate standard normal, the loss function is already around

---

[4]The CJMLE is implemented using R package *mirtjml* (Zhang et al. 2018). All the computation is conducted on a single Intel®Gold 6130 core.
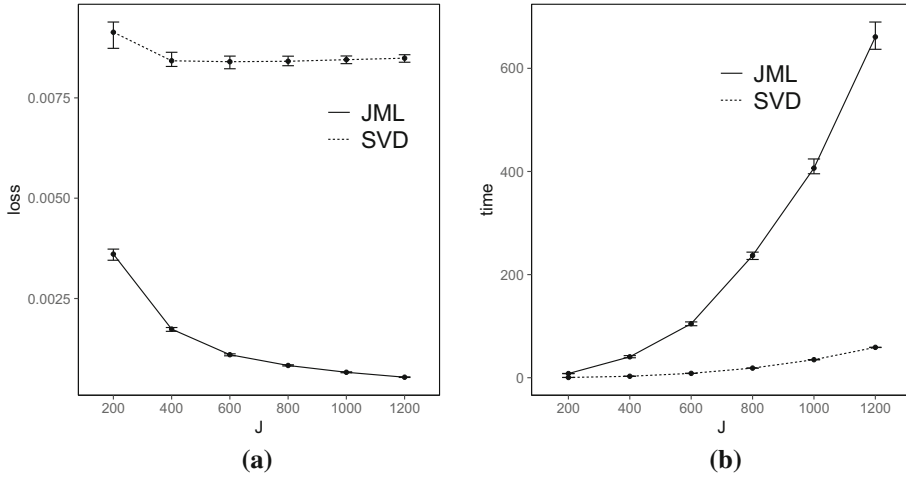
FIGURE 3.
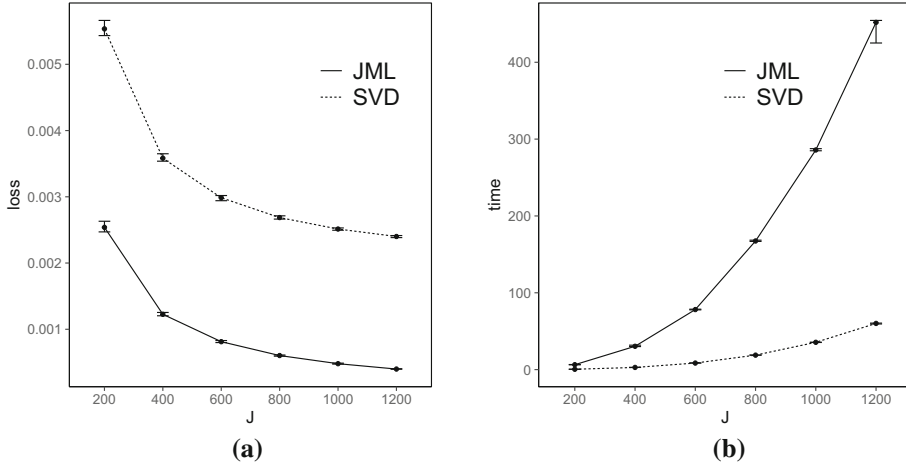Simulation results when $K = 4$ and the true factors are correlated. The two panels show the same metrics as in Fig. 2

FIGURE 4.
Simulation results when $K = 8$ and the true factors are independent. The two panels show the same metrics as in Fig. 2

0.006 when J is 200. It suggests that the average entrywise error is around 0.08. In addition, the loss for the SVD-based estimator tends to be smaller when the factors are independent than that when they are correlated, for the same N, J, and K. This is because, the signal in the data is weaker in the latter case, due to the redundant information in correlated factors.

Moreover, we compare the performance of the two estimators. The CJMLE is always more accurate than the SVD-based estimator. This is consistent with the asymptotic theory that the CJMLE is statistically more efficient. However, if we compare the computation time of the two approaches, the SVD-based estimator is substantially faster. Under the most time-consuming setting where $J = 1200$, $K = 8$ and the factors are correlated, the SVD approach only takes about 60 seconds, while the CJMLE takes about 17 minutes. Note that as shown in Chen et al. (2019b), CJMLE is already substantially faster than the marginal maximum likelihood estimator.
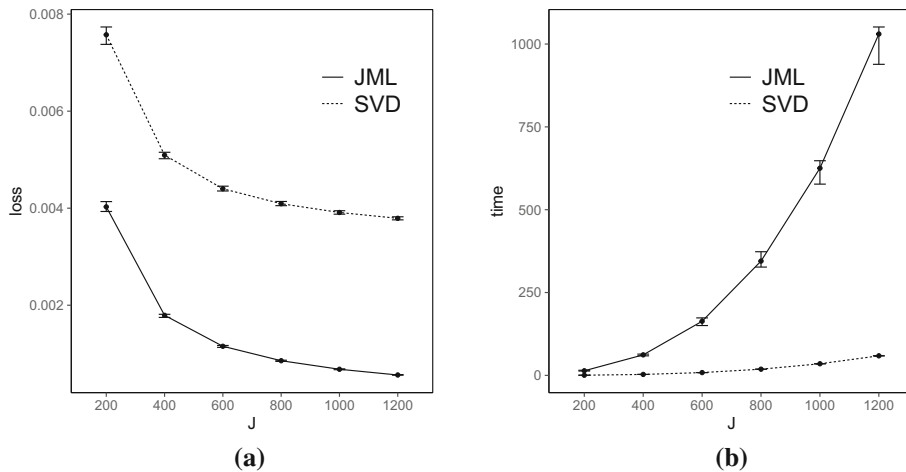
FIGURE 5.
Simulation results when $K = 8$ and the true factors are correlated. The two panels show the same metrics as in Fig. 2

Given its reasonable accuracy and computational advantage, the SVD-based estimator may be a good alternative to the CJMLE and the MMLE in large-scale exploratory IFA problems.

5. Concluding Remarks

As shown in this note, the proposed SVD-based algorithm is statistically consistent and has good finite sample performance in large-scale exploratory IFA problems. Although not statistically most efficient, the algorithm has its unique strengths over other exploratory IFA methods. In particular, it is computationally much faster. In addition, it guarantees a unique solution, while most of the other estimators can suffer from convergence issues for involving nonconvex optimization, including the CJMLE and MMLE.

Given its computational advantages and good finite sample performance, the SVD-based estimator can be used, not only as a starting point for other estimators to improve their numerical convergence, but also as an alternative estimator for data analysis. Specifically, in large-scale exploratory IFA applications, we suggest to start data exploration with the SVD-based estimator. Using this estimator, we can quickly gain some understanding about the number of factors underlying the data, and the loading structures of IFA models assuming different numbers of factors. Such initial knowledge helps us to focus on a smaller set of latent dimension K. For these latent dimensions, we tend to further investigate their loading structures by the CJMLE, using the corresponding SVD solutions as starting points. When sample and item sizes are relatively smaller, the traditional methods may be more suitable, such as the MMLE and the composite-likelihood-based estimator.

One limitation of the SVD-based estimator is that it is not easy to make statistical inference on the estimated loading matrix, such as constructing a confidence interval for an estimated loading parameter. This type of inference problem is not an issue for estimators based on the marginal likelihood, for which the asymptotic regime let N diverge and keep J fixed. However, it is a general challenge for both the SVD-based estimator and the CJMLE, whose consistency relies on a double asymptotic regime and the notion of consistency is weaker than that in the traditional sense. In recent years, this type of inference problems has received much attention in statistics

(Chen et al. 2019a; Xia and Yuan 2019). However, to the best of our knowledge, no results have been obtained under an IFA model. We leave this problem for future investigation.

Acknowledgments

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

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

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

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

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

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

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, *43*, 177–214.

Chen, Y., Fan, J., Ma, C., & Yan, Y. (2019a). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, *116*, 22931–22937.

Chen, Y., Li, X., & Zhang, S. (2019b). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, *84*, 124–146.

Chen, Y., Li, X., & Zhang, S. (2019c). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2019.1635485.

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

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York, NY: Springer.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, *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. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381–389.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis*, *56*, 4243–4258.

Lee, S.-Y., Poon, W.-Y., & Bentler, P. (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics and Probability Letters*, *9*, 91–97.

Lee, S.-Y., Poon, W.-Y., & Bentler, P. M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika*, *57*, 89–105.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*, 73–90.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

O'Rourke, S., Vu, V., & Wang, K. (2018). Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications*, *540*, 26–59.

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

Stewart, G., & Sun, J. (1990). *Matrix perturbation theory*. Cambridge, MA: Academic Press.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*, 1167–1179.

Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In D. P. Berrar, W. Dubitzky, & M. Granzow (Eds.), *A practical approach to microarray data analysis* (pp. 91–109). New York, NY: Springer.

Xia, D., & Yuan, M. (2019). Statistical inferences of linear forms for noisy matrix completion. arXiv preprint arXiv:1909.00116.

Zhang, S., Chen, Y., & Li, X. (2018). *mirtjml: Joint maximum likelihood estimation for high-dimensional item factor analysis*. R package version, 1.2.

Zhang, S., Chen, Y., & Liu, Y. (2020). An improved stochastic EM algorithm for large-scale full-information item factor analysis. *British Journal of Mathematical and Statistical Psychology*, *73*, 44–71.