PSYCHOMETRIKA—VOL. 89, NO. 2, 626–657
JUNE 2024
https://doi.org/10.1007/s11336-024-09951-y

A SPECTRAL METHOD FOR IDENTIFIABLE GRADE OF MEMBERSHIP ANALYSIS WITH BINARY RESPONSES

LING CHEN AND YUQI GU

COLUMBIA UNIVERSITY

Grade of membership (GoM) models are popular individual-level mixture models for multivariate categorical data. GoM allows each subject to have mixed memberships in multiple extreme latent profiles. Therefore, GoM models have a richer modeling capacity than latent class models that restrict each subject to belong to a single profile. The flexibility of GoM comes at the cost of more challenging identifiability and estimation problems. In this work, we propose a singular value decomposition (SVD)-based spectral approach to GoM analysis with multivariate binary responses. Our approach hinges on the observation that the expectation of the data matrix has a low-rank decomposition under a GoM model. For *identifiability*, we develop sufficient and almost necessary conditions for a notion of expectation identifiability. For *estimation*, we extract only a few leading singular vectors of the observed data matrix and exploit the simplex geometry of these vectors to estimate the mixed membership scores and other parameters. We also establish the consistency of our estimator in the double-asymptotic regime where both the number of subjects and the number of items grow to infinity. Our spectral method has a huge computational advantage over Bayesian or likelihood-based methods and is scalable to large-scale and high-dimensional data. Extensive simulation studies demonstrate the superior efficiency and accuracy of our method. We also illustrate our method by applying it to a personality test dataset.

Key words: grade of membership model, identifiability, latent variable model, mixed membership model, successive projection algorithm, singular value decomposition, spectral method.

Multivariate categorical data are routinely collected in various social and behavioral sciences, such as psychological tests (Chen et al., 2019), educational assessments (Shang et al., 2021), and political surveys (Chen et al., 2021b). In these applications, it is often of great interest to use latent variables to model the unobserved constructs such as personalities, abilities, political ideologies, etc. Popular latent variable models for multivariate categorical data include the item response theory models (IRT; Embretson & Reise 2013) and latent class models (LCM; Hagenaars & McCutcheon 2002), which employ continuous and discrete latent variables, respectively. Different from these modeling approaches, the grade of membership (GoM) models (Woodbury et al., 1978; Erosheva, 2002; Erosheva, 2005) allow each observation to have mixed memberships in multiple *extreme latent profiles*. GoM models assume that each observation has a latent membership vector with K continuous membership scores that sum up to one. Each membership score quantifies the extent to which this observation belongs to each of K extreme profiles. So GoM can be viewed as incorporating both the continuous aspect (via the membership scores) and discrete aspect (via the K extreme latent profiles) of latent variables. More generally, GoM belongs to the broad family of mixed membership models for individual-level mixtures (Airoldi et al., 2014). Thanks to their nice interpretability and rich expressive power, variants of mixed membership models including GoM are widely used in many applications such as survey data modeling (Erosheva et al., 2007), response time modeling (Pokropek, 2016), topic modeling (Blei et al., 2003), social networks (Airoldi et al., 2008), and data privacy (Manrique-Vallier & Reiter, 2012).

The flexibility of GoM models comes at the cost of more challenging identifiability and estimation problems. In the existing literature on GoM model estimation, Bayesian inference using

Correspondence should be made to Yuqi Gu, Department of Statistics, Columbia University, New York, NY10027, USA. Email: yuqi.gu@columbia.edu

© 2024 The Author(s), under exclusive licence to The Psychometric Society

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

Markov Chain Monte Carlo (MCMC) is perhaps the most prevailing approach (Erosheva, 2002; Erosheva et al., 2007; Gormley & Murphy, 2009; Gu et al., 2023). However, the posterior distributions of the GoM model parameters are complicated, after integrating out the individual-level membership scores. Many studies developed advanced MCMC algorithms for approximate posterior computation. Yet MCMC sampling is time-consuming and typically not computationally efficient. On the other hand, the frequentist estimation approach of marginal maximum likelihood (MML) would bring a similar challenge, because the marginal likelihood still involves the intractable integrals of those latent membership scores. Actually, it was pointed out in Borsboom et al. (2016) that GoM models are very useful in identifying meaningful profiles in applications including depression, personality disorders, etc., but they are temporarily not widely used in psychometrics due to the lack of readily accessible and efficient statistical software.

Recently, the development of the R package `sirt` (Robitzsch & Robitzsch, 2022) provides a joint maximum likelihood (JML) algorithm for GoMs based on the iterative estimation method proposed in Erosheva (2002). In contrast with MML, the JML approach treats the subjects' latent membership scores as fixed unknown parameters rather than random quantities. This approach hence circumvents the need of evaluating the intractable integrals during estimation. JML is currently considered the most efficient tool for estimating GoM models. However, due to its iterative manner, JML's efficiency is still unsatisfactory when applied to very large-scale data with many observations and many items. Therefore, it is desirable to develop more scalable and non-iterative estimation methods and aid psychometric researchers and practitioners to perform GoM analysis of modern item response data.

In addition to the difficulty of estimation, model identifiability is also a challenging issue for GoM models. A model is identifiable if the model parameters can be reliably recovered from the observed data. Identifiability is crucial to ensuring valid statistical estimation as well as meaningful interpretation of the inferred latent structures. The handbook of Airoldi et al. (2014) emphasizes theoretical difficulties of identifiability in mixed membership models, including GoM models. Recently, recognizing the difficulty of establishing identifiability of GoM models, Gu et al. (2023) proposed to incorporate a dimension-grouping modeling component to GoM and established the population identifiability for this new model. However, their identifiability results do not apply to the original GoM. In addition, their identifiability notion only concerns the population parameters in the model, but excludes the individual-level latent membership scores.

To address the aforementioned issues, we propose a novel singular value decomposition (SVD)-based spectral approach to GoM analysis with multivariate binary data. Our approach hinges on the observation that the expectation of the response matrix admits a low-rank decomposition under GoM. Our contributions are three-fold. *First*, we consider a notion of *expectation identifiability* and establish identifiability for GoM models with binary responses. Under this new notion, the identifiable quantities include not only the population parameters, but also the individual membership scores that indicate the grades of memberships. Specifically, we derive sufficient conditions that are almost necessary for identifiability. *Second*, based on our new identifiability results, we propose an SVD-based spectral estimation method scalable to large-scale and high-dimensional data. *Third*, we establish the consistency of our spectral estimator in the double-asymptotic regime where both the number of subjects N and the number of items J grow to infinity. Both the population parameters and the individual membership scores can be consistently estimated on average. In the simulation studies, we empirically verify the identifiability results and also demonstrate the superior efficiency and accuracy of our algorithm. A real data example also illustrates that meaningful interpretation can be drawn after applying our proposed method.

The rest of the paper is structured as follows. Section 1 introduces the model setup and lays out the motivation for this work. Section 2 presents the identifiability results. Section 3 proposes a spectral estimation algorithm and establishes its consistency. Section 4 conducts simulation

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

studies to assess the performance of the proposed method and empirically verify the identifiability results. Section 5 illustrates the proposed method using a real data example in a psychological test. Finally, Sect. 6 concludes the paper and discusses future research directions. The proofs of the identifiability results are included in Appendix.

1. Model Setup and Motivation

GoM models can be used to model multivariate categorical data with a mixed membership structure. In this work, we focus on multivariate binary responses, which are very commonly encountered in social, behavioral, and biomedical applications, including yes/no responses in social science surveys (Erosheva et al., 2007; Chen et al., 2021b), correct/wrong answers in educational assessments (Shang et al., 2021), and presence/absence of symptoms in medical diagnosis (Woodbury et al., 1978). We point out that our identifiability results and spectral estimation method in the binary case will illuminate the key structure of GoM and pave the way for generalizing to the general categorical response case. We will briefly discuss the possibility of such extensions in Sect. 6. In our binary response setting, denote the number of items by J. For a random subject i, denote his or her observed response to the j-th item by $R_{ij} \in \{0, 1\}$ for $j = 1, \ldots, J$.

A GoM model is characterized by two levels of modeling: the population level and the individual level. On the population level, K *extreme latent profiles* are defined to capture a finite number of prototypical response patterns. For $k \in 1, \ldots, K$, the k-th extreme latent profile is characterized by the item parameter vector $\boldsymbol{\theta}_k = (\theta_{1k}, \ldots, \theta_{JK})$ with $\theta_{jk} \in [0, 1]$ collecting the Bernoulli parameters of conditional response probabilities. Specifically,

$$\theta_{jk} = \mathbb{P}(R_{ij} = 1 \mid \text{subject } i \text{ solely belongs to the } k\text{-th extreme latent profile}). \tag{1}$$

We collect all the item-level Bernoulli parameters in a $J \times K$ matrix $\boldsymbol{\Theta} = (\theta_{jk}) \in \mathbb{R}^{J \times K}$. On the individual level, each subject i has a latent membership vector $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$, satisfying $\pi_{ik} \geq 0$ and $\sum_{k=1}^{K} \pi_{ik} = 1$. Here π_{ik} indicates the extent to which subject i partially belongs to the k-th extreme profile, and they are called membership scores. It is now instrumental to compare the assumption of the GoM model with that of the latent class model (LCM; Goodman 1974; Hagenaars & McCutcheon 2002). Both GoM and LCM share a similar formulation of the item parameters $\boldsymbol{\Theta}$ as defined in (1). However, under an LCM, each subject i is associated with a categorical variable $z_i \in [K]$ instead of a membership vector. This means LCM restricts each subject to solely belonging to a single profile, as opposed to partially belonging to multiple profiles in the GoM. Further, the fundamental representation theorem in Erosheva et al. (2007) shows that a GoM model can be reformulated with an LCM but with K^J latent classes instead of K latent classes. In summary, GoM is a more general and flexible tool than LCM for modeling multivariate data, but also exhibits a more complicated model structure. It is therefore more challenging to establish identifiability and perform estimation under the GoM setting.

Given the membership score vector $\boldsymbol{\pi}_i$ and the item parameters $\boldsymbol{\Theta}$, the conditional probability of the i-th subject providing a positive response to the j-th item is

$$\mathbb{P}(R_{ij} = 1 \mid \boldsymbol{\pi}_i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_{ik}\theta_{jk}. \tag{2}$$

In other words, the positive response probability for a subject to an item is a convex combination of the extreme profile response probabilities θ_{jk} weighted by the subject's membership scores π_{ik}.

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

The GoM model assumes that a subject's responses $R_{i,1}, \ldots, R_{i,J}$ are conditionally independent given the membership scores $\boldsymbol{\pi}_i$. We consider a sample of N i.i.d. subjects and collect all the membership scores in a $N \times K$ matrix $\boldsymbol{\Pi} = (\pi_{ik}) \in \mathbb{R}^{N \times K}$.

In the GoM modeling literature (e.g., Erosheva 2002), there are two perspectives of dealing with the membership scores $\boldsymbol{\pi}_i$: the random-effect and the fixed-effect perspectives. The random effect perspective treats $\boldsymbol{\pi}_i$ as random and assumes that they follow some distribution parameterized by $\boldsymbol{\alpha}$: $\boldsymbol{\pi}_i \sim D_{\boldsymbol{\alpha}}(\cdot)$. Note that $\boldsymbol{\pi}_i \in \Delta_{K-1} = \{\mathbf{x} = (x_1, \ldots, x_K) : x_i \geq 0, \sum_{i=1}^{K} x_i = 1\}$ where Δ_{K-1} denotes the probability simplex. A common choice of the distribution $D_{\boldsymbol{\alpha}}$ for $\boldsymbol{\pi}_i$ is the Dirichlet distribution (Blei et al., 2003).

From the above random-effect perspective, the *marginal likelihood* function for a GoM model is

$$L(\boldsymbol{\Theta}, \boldsymbol{\alpha} \mid \mathbf{R}) = \prod_{i=1}^{N} \int_{\Delta_{K-1}} \prod_{j=1}^{J} \left(\sum_{k=1}^{K} \pi_{ik}\theta_{jk} \right)^{R_{ij}} \left(1 - \sum_{k=1}^{K} \pi_{ik}\theta_{jk} \right)^{1-R_{ij}} dD_{\boldsymbol{\alpha}}(\boldsymbol{\pi}_i), \tag{3}$$

where the membership scores $\boldsymbol{\pi}_i$'s are marginalized out with respect to their distribution $D_{\boldsymbol{\alpha}}$. The integration of the products of sums in (3) poses challenges to establishing identifiability. This difficulty motivated Gu et al. (2023) to introduce a dimension-grouping component to simplify the integrals and then prove identifiability for that new model. In terms of estimation, the marginal maximum likelihood (MML) approach maximizes (3) to estimate the population parameters $(\boldsymbol{\Theta}, \boldsymbol{\alpha})$ rather than the individual membership scores $\boldsymbol{\Pi}$. Bayesian inference with MCMC is also often used for estimation (Erosheva et al., 2007; Manrique-Vallier & Reiter, 2012; Gu et al., 2023), where inferring parameters like $\boldsymbol{\alpha}$ in the Dirichlet distribution $D_{\boldsymbol{\alpha}}(\cdot)$ typically require the Metropolis–Hastings sampling.

On the other hand, the fixed-effect perspective of GoM treats the membership scores $\boldsymbol{\pi}_i$ as fixed unknown parameters and aims to directly estimate them. This approach does not model the distribution of the membership scores and hence circumvents the need to evaluate the intractable integrals during estimation. In this case, if still adopting the likelihood framework, the *joint likelihood* function of both $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$ for a GoM model is

$$L(\boldsymbol{\Pi}, \boldsymbol{\Theta} \mid \mathbf{R}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left(\sum_{k=1}^{K} \pi_{ik}\theta_{jk} \right)^{R_{ij}} \left(1 - \sum_{k=1}^{K} \pi_{ik}\theta_{jk} \right)^{1-R_{ij}}. \tag{4}$$

The joint maximum likelihood (JML) approach maximizes (4) to estimate $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$. Based on an iterative algorithm proposed by Erosheva (2002), the R package `sirt` (Robitzsch & Robitzsch, 2022) provides a function `JML` to solve this optimization problem under GoM. JML methods are typically known as inconsistent for many traditional models (Neyman & Scott, 1948) when the sample size goes to infinity (large N), but the number of observed variables is finite (fixed J). Nonetheless, in modern large-scale assessments or surveys where the data collection scope is unprecedentedly big and high-dimensional, both N and J can be quite large. JML is currently considered the most efficient tool for estimating GoM models. However, due to its iterative manner, JML's efficiency is still unsatisfactory when applied to modern big datasets with many observations and many items. Therefore, it is desirable to develop more scalable and non-iterative estimation methods and aid psychometric researchers and practitioners in performing GoM analysis of item response data.

To address the above issues in the GoM analysis, in this work, we propose a novel singular value decomposition (SVD) based spectral approach. Our approach hinges on the observation that the expectation of the response matrix under a GoM model admits a low-rank decomposition. To

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

see this, it is useful to summarize (2) in matrix form

$$\mathbf{R}_0 := \mathbb{E}[\mathbf{R}] = \underbrace{\mathbf{\Pi}}_{N\times K}\underbrace{\mathbf{\Theta}^\top}_{K\times J}, \tag{5}$$

where $\mathbf{R} = (R_{ij}) \in \{0, 1\}^{N\times J}$ denotes the binary response data matrix, and $\mathbf{R}_0 = (R_{0,ij}) \in \mathbb{R}^{N\times J}$ is the element-wise expectation of $\mathbf{R}$. Note that the factorization in (5) implies that the $N \times J$ matrix $\mathbf{R}_0$ has rank at most K, which is the number of extreme latent profiles. Since K is typically (much) smaller than N and J, the decomposition (5) exhibits a low-rank structure. Therefore, we can consider the singular value decomposition (SVD) of $\mathbf{R}_0$:

$$\mathbf{R}_0 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \tag{6}$$

where $\mathbf{\Sigma}$ is a $K \times K$ diagonal matrix collecting the K singular values of $\mathbf{R}_0$; denote these singular values by $\sigma_1 \geq \cdots \geq \sigma_K \geq 0$ and write $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_K)$. Matrices $\mathbf{U}_{N\times K}$, $\mathbf{V}_{J\times K}$ collect the corresponding left and right singular vectors and satisfy $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_K$. Our high-level idea is to utilize the top K left singular vectors of the data matrix $\mathbf{R}$ to identify and estimate $\mathbf{\Pi}$ and subsequently $\mathbf{\Theta}$. In the following two sections, we will present new identifiability results and develop a spectral estimation algorithm for GoM models based on the SVD in (6).

2. Identifiability Results

The study of identifiability in statistics dates back to Koopmans & Reiersol (1950). A model is identifiable if the model parameters can be reliably recovered from the observed data. Identifiability is a crucial property of a statistical model as it is a prerequisite for valid and reproducible statistical inference. In latent variable modeling, identifiability is especially essential since it is a foundation for meaningful interpretation of the latent constructs.

Traditionally, identifiability of a statistical model means that the population parameters can be uniquely determined from the marginal distribution of the observed variables (Koopmans & Reiersol, 1950; Goodman, 1974). In the context of GoM models, this notion of *population identifiability* is equivalent to identifying parameters $(\mathbf{\Theta}, \boldsymbol{\alpha})$ from the marginal distribution in (3). The complicated integrals in (3) make it difficult to establish population identifiability, which motivated Gu et al. (2023) to propose a dimension-grouping modeling component to simplify GoM and prove identifiability for that new model. However, it remains unknown whether the original GoM models can be identified.

In this work, we consider a new notion of identifiability, which we term as *expectation identifiability*. This notion concerns not only the item parameters, but also the individual membership scores. Similar identifiability notions are widely adopted and studied in the network modeling and topic modeling literature, e.g., Jin et al. (2023), Ke & Jin (2023), Mao et al. (2021), and Ke & Wang (2022). Specifically, recall from (5) that the expectation of the data matrix $\mathbf{R}$ has a low-rank decomposition $\mathbf{R}_0 = \mathbf{\Pi}\mathbf{\Theta}^\top$, we seek to understand under what conditions this decomposition is unique. Note that both the Bernoulli probabilities $\mathbf{\Theta}$ and the membership scores $\mathbf{\Pi}$ are treated as parameters to be identified. We call a parameter set $(\mathbf{\Pi}, \mathbf{\Theta})$ valid if $\boldsymbol{\pi}_i \in \Delta_{K-1}$ and $\theta_{jk} \in [0, 1]$ for all i, j, and k. We formally define expectation identifiability below.

Definition 1. (Expectation identifiability) A GoM model with parameter set $(\mathbf{\Pi}, \mathbf{\Theta})$ is said to be identifiable, if for any other valid parameter set $(\widetilde{\mathbf{\Pi}}, \widetilde{\mathbf{\Theta}})$, $\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{\Theta}}^\top = \mathbf{\Pi}\mathbf{\Theta}^\top$ holds if and only if $(\mathbf{\Pi}, \mathbf{\Theta})$ and $(\widetilde{\mathbf{\Pi}}, \widetilde{\mathbf{\Theta}})$ are identical up to a permutation of the K extreme profiles.

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

One might ask whether identifying $\mathbf{\Pi}$ and $\mathbf{\Theta}$ from the expectation $\mathbf{R}_0$ has any implications on identifying $\mathbf{\Pi}$ and $\mathbf{\Theta}$ from the observed data $\mathbf{R}$. In fact, when both N and J are large with respect to K, it is known that the difference between the low-rank decompositions of $\mathbf{R}_0$ and $\mathbf{R}$ is small in a certain sense (Chen et al., 2021a). We will revisit and elaborate on this subtlety when describing our estimation method and presenting the simulation results. Briefly speaking, studying the expectation identifiability problem from $\mathbf{R}_0$ is very meaningful in modern large-scale and high-dimensional data settings with large N and J.

We next present our new identifiability conditions for GoM models. We first define the important concept of pure subjects.

Definition 2. (Pure subject) Subject i is a pure subject for extreme profile k if the only positive entry of $\boldsymbol{\pi}_i$ is located at index k, that is,

$$\boldsymbol{\pi}_i = (0, \ldots, 0, \underbrace{1}_{k\text{-th entry}}, 0, \ldots, 0).$$

In words, subject i is a pure subject for profile k if it solely belongs to this profile and has no membership in any other profile. We consider the following condition for $\mathbf{\Pi}$.

Condition 1. $\mathbf{\Pi}$ *satisfies that every extreme latent profile has at least one pure subject.*

Condition 1 is a quite mild assumption on $\mathbf{\Pi}$, because it only requires that each of the K extreme profiles has at least one representative subject among all N subjects. Intuitively, this condition is reasonable because the existence of these representative subjects indeed helps pinpoint the meaning and interpretation of the extreme profiles. In real data applications of the GoM model, each pure subject is characterized by a prototypical response pattern indicating a particular classification or diagnosis. Specifically, each column of the $J \times K$ item parameter matrix $\mathbf{\Theta}$ is examined to coin the interpretation of each extreme profile. So, the existence of a pure subject in the kth ($1 \leq k \leq K$) extreme profile means that there indeed exists a prototypical subject characterized by the parameters in the kth column of the $\mathbf{\Theta}$ matrix. As a concrete applied example, Woodbury et al. (1978) fitted the GoM model to a clinical dataset, where the item parameters for four extreme latent profiles were estimated. Then each of the extreme profiles was interpreted according to its response characteristics revealed via the item parameters. There, the four extreme profiles were interpreted as "Asymptomatic," "Moderate," "Acyanotic Severe," "Cyanotic Severe" in Woodbury et al. (1978). In this context, having pure subjects in each of these extreme profiles is a practically meaningful assumption, because it just means that in the sample with a large number of N subjects, there exist an "Asymptomatic" subject, a "Moderate" subject, an "Acyanotic Severe" subject, and a "Cyanotic Severe" subject.

Under Condition 1, $\mathbf{\Pi}$ contains one identity submatrix $\mathbf{I}_K$ after some row permutation. For any matrix $\mathbf{A}$, we use $\mathbf{A}_{\mathbf{S},:}$ to denote the rows with indices in $\mathbf{S}$. Under Condition 1, denote $\mathbf{S} = (S_1, \ldots, S_K)$ as the index vector of one set of K pure subjects such that $\mathbf{\Pi}_{\mathbf{S},:} = \mathbf{I}_K$. So $S_1, \ldots, S_K$ are distinct integers ranging in $\{1, \ldots, N\}$. For example, if the first K rows of $\mathbf{\Pi}$ is equal to $\mathbf{I}_K$, then $\mathbf{S} = (1, 2, \ldots, K)$. Recall $\mathbf{R}_0 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the SVD for $\mathbf{R}_0$, where $\mathbf{U}$ is a $N \times K$ matrix collecting the K left singular vectors as columns. Interestingly, Condition 1 induces a *simplex geometry* on the row vectors of $\mathbf{U}$. We have the following important proposition, which serves as a foundation for both our identifiability results and estimation procedure.

Proposition 1. *Under Condition 1, the left singular matrix* $\mathbf{U}$ *satisfies*

$$\mathbf{U} = \mathbf{\Pi}\mathbf{U}_{\mathbf{S},:}. \tag{7}$$

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

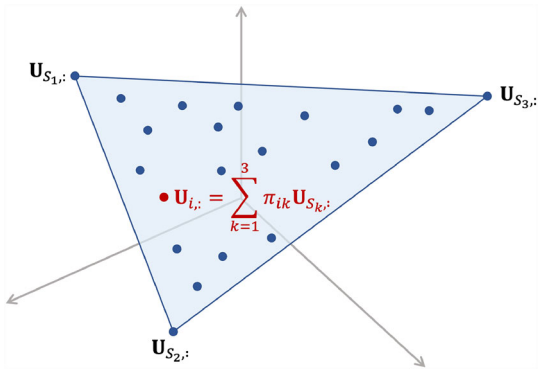

FIGURE 1.
Illustration of the simplex geometry of the $N \times K$ left singular matrix $\mathbf{U}$ with $K = 3$. The solid dots represent the row vectors of $\mathbf{U}$ in $\mathbb{R}^3$, and the three simplex vertices (i.e, vertices of the triangle) correspond to the three types of pure subjects. All the dots lie in this triangle.

Furthermore, $\mathbf{\Pi}$ *and* $\mathbf{\Theta}$ *can be written as*

$$\mathbf{\Pi} = \mathbf{U}\mathbf{U}_{\mathbf{S},:}^{-1}, \tag{8}$$

$$\mathbf{\Theta} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{\Pi}(\mathbf{\Pi}^\top\mathbf{\Pi})^{-1} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}_{\mathbf{S},:}^\top. \tag{9}$$

We elaborate more on Proposition 1. Equation (7) in Proposition 1 implies that the left singular matrix $\mathbf{U}$ and the membership score matrix $\mathbf{\Pi}$ differ by a linear transformation, which is the $K \times K$ matrix $\mathbf{U}_{\mathbf{S},:}$. Condition 1 and the properties of the singular value decomposition (such as the columns of $\mathbf{V}$ being orthogonal to each other and $\mathbf{\Sigma}$ being invertible) are used to prove (7). Equations (8) and (9) imply that if an index set of pure subjects $\mathbf{S}$ is known, then the parameters of interest $\mathbf{\Pi}$ and $\mathbf{\Theta}$ can be written in closed forms in terms of the SVD and $\mathbf{S}$. More specifically, (7) is equivalent to

$$\mathbf{U}_{i,:} = \sum_{k=1}^{K} \pi_{ik}\mathbf{U}_{S_k,:}, \ i = 1, \ldots, N. \tag{10}$$

Each $\mathbf{U}_{i,:}$ is the embedding of the i-th subject into the top-K left singular subspace of $\mathbf{R}_0$, and all the rows $\mathbf{U}_{1,:}, \ldots, \mathbf{U}_{N,:}$ can be plotted as points in the K-dimensional Euclidean space $\mathbb{R}^K$. Geometrically, since $\sum_{k=1}^{K} \pi_{ik} = 1$ with $\pi_{ik} \geq 0$, we know that each $\mathbf{U}_{i,:}$ is a convex combination of $\mathbf{U}_{S_1,:}, \ldots, \mathbf{U}_{S_K,:}$, which are the embeddings of the K types of pure subjects. This means in $\mathbb{R}^K$, all the subjects lie in a simplex (i.e., the generalization of a triangle or tetrahedron to higher dimensions) whose vertices are these K types of pure subjects. Note that $\mathbf{U}_i$ and $\mathbf{U}_{i'}$ overlap if they have the same membership score vectors $\boldsymbol{\pi}_i = \boldsymbol{\pi}_{i'}$. Figure 1 gives an illustration of this simplex geometry on such embeddings in $\mathbb{R}^3$ with $K = 3$.

Similar simplex structure in the spectral domain was first discovered and used for estimation under the degree-corrected mixed membership network model (Jin et al. 2023, first posted on arXiv in 2017). Later, spectral approaches to estimating mixed memberships via exploiting the simplex structure are also used for related network models (Mao et al., 2021) and topic models (Ke & Wang, 2022). Compared to network models where the data matrix is symmetric, the GoM model has an $N \times J$ asymmetric data matrix. Compared to topic models, the entries of the perturbation matrix $\mathbf{R} - \mathbf{R}_0$ in the GoM model independently follow Bernoulli distributions, whereas in topic models, the entries of the perturbation matrix follow multinomial distributions.

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

It is worth noting that the expectation identifiability in Definition 1 is closely related to the uniqueness of non-negative matrix factorization (NMF, Donoho & Stodden 2003; Hoyer 2004; Berry et al. 2007). NMF seeks to decompose a nonnegative matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ into $\mathbf{M} = \mathbf{WH}$, where both $\mathbf{W} \in \mathbb{R}^{m\times r}$ and $\mathbf{H} \in \mathbb{R}^{r\times n}$ are non-negative matrices. An NMF is called separable if each column of $\mathbf{W}$ appears as a column of $\mathbf{M}$ (Donoho & Stodden, 2003). If we write $\mathbf{M} = \mathbf{R}_0^\top, \mathbf{W} = \mathbf{\Theta}, \mathbf{H} = \mathbf{\Pi}^\top$, then the separability condition for this NMF aligns with Condition 1. It is also generally assumed that $\mathbf{W}$ is of full rank, otherwise $\mathbf{H}$ typically cannot be uniquely determined (Gillis & Vavasis, 2013). Theorem 2 shows that when Condition 1 holds, $\mathbf{\Theta}$ being full-rank suffices for GoM model identifiability. We will also show that model identifiability still holds with certain relaxations on the rank of $\mathbf{\Theta}$. On another note, estimating an NMF usually involves direct manipulation of the original data matrix, which can be computationally inefficient when dealing with large datasets. In our approach, we employ an NMF algorithm in Gillis & Vavasis (2013) on the SVD of the data matrix instead of the data matrix itself to estimate GoM parameters. Since our procedure operates on the singular subspace with significantly lower dimension than the original data space, it yields lower computational cost compared to conventional NMF procedures.

We next present our identifiability results for GoM models. We first show that the pure-subject Condition 1 is almost necessary for the identifiability of a GoM model.

Theorem 1. *Suppose $\theta_{jk} \in (0, 1)$ for all $j = 1, \ldots, J$ and $k = 1, \ldots, K$. If there is one extreme profile that does not have any pure subject, then the GoM model is not identifiable.*

The proofs of the theorems are all deferred to Appendix. Theorem 1 reveals the importance of Condition 1 for identifiability. In fact, later we will use this condition as a foundation for our estimation algorithm. Our next theorem presents sufficient and almost necessary conditions for GoM models to be identifiable.

Theorem 2. *Suppose $\mathbf{\Pi}$ satisfies Condition 1.*

(a) If $rank(\mathbf{\Theta}) = K$, then the GoM model is identifiable.
(b) If $rank(\mathbf{\Theta}) = K - 1$ and no column of $\mathbf{\Theta}$ is an affine combination of the other columns of $\mathbf{\Theta}$, then the GoM model is identifiable. (An affine combination of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is defined as $\sum_{i=1}^n a_i\mathbf{x}_i$ with $\sum_{i=1}^n a_i = 1$.)
(c) In any other case, if there exists a subject i such that $\pi_{ik} > 0$ for every $k = 1, \ldots, K$, then the GoM model is not identifiable.

The high-level proof idea of Theorem 2 shares a similar spirit to Theorem 2.1 in Mao et al. (2021). We next explain and interpret the three settings in Theorem 2. According to part (a), if the K item parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ (i.e., the K columns of $\mathbf{\Theta}$) are linearly independent, then the GoM model is identifiable under Condition 1. In part (b), for identifiability to hold when $rank(\mathbf{\Theta}) = K - 1$, any item parameter vector $\boldsymbol{\theta}_k$ cannot be written as an affine combination of the remaining vectors $\{\boldsymbol{\theta}_{k'} : k' \neq k\}$. This is a weaker requirement on $\mathbf{\Theta}$ compared to part (a). Part (c) states that if the conditions in parts (a) or (b) do not hold and if there exists a *completely mixed subject* that partially belongs to all profiles (i.e., $\pi_{ik} > 0$ for all k), then the model is not identifiable. Part (c) also shows that the sufficient identifiability conditions in (a) and (b) are close to being necessary, because the existence of a completely mixed subject is a very mild assumption. We next further give three toy examples with $K = 3$ and $J = 4$ to illustrate the conditions in Theorem 2.

Example 1. Consider

$$\mathbf{\Theta} = \begin{pmatrix} 0.2 & 0.8 & 0.8 \\ 0.2 & 0.8 & 0.2 \\ 0.8 & 0.2 & 0.8 \\ 0.8 & 0.2 & 0.2 \end{pmatrix}.$$

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

It is easy to verify that $rank(\mathbf{\Theta}) = K = 3$. This case falls into scenario (a) in Theorem 2, so a GoM model parameterized by $(\mathbf{\Pi}, \mathbf{\Theta})$ is identifiable if $\mathbf{\Pi}$ satisfies Condition 1.

Example 2. Consider

$$\mathbf{\Theta} = \begin{pmatrix} 0.2 & 0.8 & 0.8 \\ 0.2 & 0.8 & 0.8 \\ 0.8 & 0.2 & 0.8 \\ 0.8 & 0.2 & 0.8 \end{pmatrix}.$$

Now $rank(\mathbf{\Theta}) = K - 1 = 2$ since the third column of $\mathbf{\Theta}$ is a linear combination of the first two columns. However, there is no column of $\mathbf{\Theta}$ that is an affine combination of the other columns of $\mathbf{\Theta}$. This case falls into scenario (b) in Theorem 2, so a GoM model parameterized by $(\mathbf{\Pi}, \mathbf{\Theta})$ is identifiable if $\mathbf{\Pi}$ satisfies Condition 1.

Example 3. Consider

$$\mathbf{\Theta} = \begin{pmatrix} 0.2 & 0.8 & 0.5 \\ 0.2 & 0.8 & 0.5 \\ 0.8 & 0.2 & 0.5 \\ 0.8 & 0.2 & 0.5 \end{pmatrix}.$$

In this case, $rank(\mathbf{\Theta}) = K - 1 = 2$, and the third column of $\mathbf{\Theta}$ is an affine combination of the first two columns. This case falls into scenario (c) in Theorem 2, so if there exists a subject that partially belongs to all K profiles, then the GoM model is not identifiable.

3. SVD-Based Spectral Estimation Method and Its Consistency

3.1. Estimation Algorithm

In the literature of GoM model estimation, the most prevailing approaches are perhaps Bayesian inferences using Markov chain Monte Carlo (MCMC) algorithms such as Gibbs and Metropolis–Hastings sampling (Erosheva, 2002; Erosheva et al., 2007; Gu et al., 2023). However, MCMC is time-consuming and typically not computationally efficient. As Borsboom et al. (2016) points out, despite their usefulness, GoM models are somewhat underrepresented in psychometric applications due to the lack of readily accessible statistical software. Recently, the R package `sirt` (Robitzsch & Robitzsch, 2022) provides a joint maximum likelihood (JML) algorithm to fit GoM models. This algorithm implements the Lagrange multiplier method proposed in Erosheva (2002) and solves the optimization problem in a gradient descent fashion. Although this JML algorithm is computationally more efficient compared to MCMC algorithms, it is still not scalable to very large-scale response data due to its iterative manner. Therefore, it is of interest to develop a non-iterative estimation method suitable to analyze modern datasets with a large number of items and subjects.

We next propose a fast SVD-based spectral method to estimate GoM models. Recall that Proposition 1 establishes the expressions for $\mathbf{\Pi}, \mathbf{\Theta}$ in (8) and (9). In practice, since $\mathbf{S}$ is not known, we propose to estimate it using a vertex-hunting technique called *successive projection algorithm* (SPA; Araújo et al. 2001; Gillis & Vavasis 2013). As stated in Proposition 1, Condition 1 induces a simplex geometry on the row vectors of $\mathbf{U}$ and the simplex vertices correspond to the pure subjects $\mathbf{S}$. To locate K vertices for any input matrix $\mathbf{U} \in \mathbb{R}^{N \times K}$ that has such a simplex

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

structure, SPA first finds the subject with the maximum row norm in $\mathbf{U}$. That is, the first vertex index is

$$\widehat{S}_1 = \underset{1 \le i \le N}{\operatorname{argmax}} \|\mathbf{U}_{i,:}\|_2.$$

Here and after we use $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}}$ to denote the ℓ_2 norm of any vector $\mathbf{x}$. Since the l_2 norm of any convex combination of the vertices is at most the maximum l_2 norm of the vertices, this step is guaranteed to return one of the vertices of the simplex. SPA then projects all the remaining row vectors $\{\mathbf{U}_{i,:} : i \neq \widehat{S}_1\}$ onto the subspace that is orthogonal to $\mathbf{U}_{\widehat{S}_1,:}$. Mathematically, denote $\mathbf{v}_1 := \mathbf{U}_{\widehat{S}_1,:}/\|\mathbf{U}_{\widehat{S}_1,:}\|_2$ as the scaled vector with unit norm, then the projected vectors are the rows of the matrix $\mathbf{U}(\mathbf{I}_K - \mathbf{v}_1\mathbf{v}_1^\top)$, where $\mathbf{I}_K - \mathbf{v}_1\mathbf{v}_1^\top$ is a $K \times K$ projection matrix. In the second step, SPA finds the second vertex index as the one that has the maximum norm among the projected row vectors,

$$\widehat{S}_2 = \underset{i \neq \widehat{S}_1}{\operatorname{argmax}} \|\mathbf{U}_{i,:}(\mathbf{I}_K - \mathbf{v}_1\mathbf{v}_1^\top)\|_2.$$

The above procedure of first finding the row with the maximum projected norm and then projecting the remaining rows onto the orthogonal space is iteratively employed until finding all K vertex indices. Sequentially, for each $k = 1, \ldots, K-1$, define a unit-norm vector $\mathbf{v}_k = \mathbf{U}_{\widehat{S}_k,:}/\|\mathbf{U}_{\widehat{S}_k,:}\|_2$, then the $(k+1)$-th vertex index is estimated as

$$\widehat{S}_{k+1} = \underset{i \notin \{\widehat{S}_1, \ldots, \widehat{S}_k\}}{\operatorname{argmax}} \|\mathbf{U}_{i,:}(\mathbf{I}_K - \mathbf{v}_1\mathbf{v}_1^\top) \cdots (\mathbf{I}_K - \mathbf{v}_k\mathbf{v}_k^\top)\|_2.$$

Here the projection matrices $(\mathbf{I}_K - \mathbf{v}_1\mathbf{v}_1^\top), \ldots, (\mathbf{I}_K - \mathbf{v}_k\mathbf{v}_k^\top)$ sequentially project the rows of $\mathbf{U}$ to the orthogonal spaces of those already found vertices $\mathbf{U}_{\widehat{S}_1,:}, \ldots, \mathbf{U}_{\widehat{S}_k,:}$. This SPA procedure can be intuitively understood by visually inspecting the toy example in Fig. 1. Since $\mathbf{U}_{1,:}, \ldots, \mathbf{U}_{N,:}$ lie in a triangle in Fig. 1, it is not hard to see that the vector with the largest norm should be one of the three vertices $\mathbf{U}_{S_1,:}, \mathbf{U}_{S_2,:}, \mathbf{U}_{S_3,:}$, say $\mathbf{U}_{S_3,:}$. Furthermore, after projecting the remaining $\mathbf{U}_{i,:}$ onto the space orthogonal to $\mathbf{U}_{S_3,:}$, the maximum norm of the projected vectors should correspond to $i = S_1$ or S_2. This observation intuitively justifies that SPA can find the correct set of pure subjects given a simplex structure. With the estimated pure subjects $\widehat{\mathbf{S}}$, $\mathbf{\Pi}$ and $\mathbf{\Theta}$ can be subsequently obtained via (8) and (9).

The above estimation procedure is based on the assumption that $\mathbf{R}_0$ is known. In practice, we only have access to the binary random data matrix $\mathbf{R}$ whose expectation is $\mathbf{R}_0$. Fortunately, it is known that for a large-dimensional random matrix with a low-rank expectation, the top-K SVD of the random matrix and the SVD of its expectation are close (e.g., see Chen et al. 2021a). This nontrivial theoretical result is our key insight and motivation to consider the top-K SVD of $\mathbf{R}$ as a surrogate for that of $\mathbf{R}_0$:

$$\mathbf{R} \approx \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top, \tag{11}$$

where $\widehat{\mathbf{\Sigma}}$ is a $K \times K$ diagonal matrix collecting the K largest singular values of $\mathbf{R}$, and $\widehat{\mathbf{U}}_{N\times K}$, $\widehat{\mathbf{V}}_{J\times K}$ collect the corresponding left and right singular vectors with $\widehat{\mathbf{U}}^\top\widehat{\mathbf{U}} = \widehat{\mathbf{V}}^\top\widehat{\mathbf{V}} = \mathbf{I}_K$. Specifically, Chen et al. (2021a) proved that the difference between $\mathbf{U}$ and $\widehat{\mathbf{U}}$ up to a rotation is small when N and J are large with respect to K. Therefore, since Proposition 1 shows that the population row vectors $\mathbf{U}_{1,:}, \ldots, \mathbf{U}_{N,:}$ form a simplex structure, the empirical row vectors $\{\widehat{\mathbf{U}}_{i,:}\}_{i=1}^N$ are expected to form a noisy simplex cloud distributed around the population simplex. We call this noisy cloud

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

Algorithm 1
Prune

Input: Empirical top-K left singular matrix $\widehat{\mathbf{U}}$, the number of nearest neighbors r, two quantiles $q, e \in (0, 1)$.
Output: Set of subject indices $\widehat{\mathbf{P}}$ to be pruned (i.e., removed)
1: **for** $i \in 1, \ldots, N$ **do**
2: $\quad l_i = \|\widehat{\mathbf{U}}_{i,:}\|_2$
3: **end for**
4: $\widehat{\mathbf{P}}_0 = \{i : l_i \geq$ upper-q quantile of $\{l_i : 1 \leq i \leq N\}\}$
5: **for** $i \in \widehat{\mathbf{P}}_0$ **do**
6: $\quad \mathbf{d}_i = \{$the distances from $\widehat{\mathbf{U}}_i$ to its r nearest neighbors$\}$
7: $\quad x_i = \text{average}(\mathbf{d}_i)$
8: **end for**
9: $\widehat{\mathbf{P}} = \{i : x_i \geq$ upper-e quantile of $\{x_i : i \in \widehat{\mathbf{P}}_0\}\}$

the empirical simplex. For an illustration of the population and the empirical simplex, see Fig. 2 in Sect. 4.

In order to recover the population simplex from the empirical one, we use a pruning step similar to that in Mao et al. (2021) to reduce noise. We summarize this pruning procedure in Algorithm 1. The high-level idea behind pruning is that if $\widehat{\mathbf{U}}_{i,:}$ has a large norm but very few close neighbors, then it is likely to be outside of the population simplex and hence should be pruned (i.e., removed) before performing SPA to achieve higher accuracy in vertex hunting. More specifically, Algorithm 1 first calculates the norm of each row of $\widehat{\mathbf{U}}$ (line 1- 3) and identifies the vectors with norms in the upper q-quantile (line 4). The larger q is, the more such points are found. Then for each vector found, its average distance x_i to its r nearest neighbors is calculated (line 5- 8). Finally, the subjects to be pruned are those whose x_i belong to the upper e-quantile of all the x_i's (line 9). A larger value of e indicates that a larger proportion of the points will be pruned. According to our preliminary simulations, we observe that the estimation results are not very sensitive to these tuning parameters r, q, e. After pruning, we can use SPA to hunt for the K vertices of the pruned empirical simplex to obtain $\widehat{\mathbf{S}}$ and then estimate $\mathbf{\Pi}$ and $\mathbf{\Theta}$.

Algorithm 2 summarizes our proposed method of estimating parameters $(\mathbf{\Pi}, \mathbf{\Theta})$ based on SPA with the pruning step. We first introduce some notation. The index of the element in $\mathbf{x}$ that has the maximum value is denoted by $\text{argmax}(\mathbf{x})$. For any matrix $\mathbf{A} = (a_{ij})$, denote by $\mathbf{A}_+ = (\max\{a_{ij}, 0\})$ the matrix that retains the nonnegative values of $\mathbf{A}$ and sets any negative values to zero. Denote by $diag(\mathbf{x})$ the diagonal matrix whose diagonals are the entries of the vector $\mathbf{x}$. If $\mathbf{S}_2$ is a subvector of $\mathbf{S}_1$, denote by $\mathbf{S}_1 \backslash \mathbf{S}_2$ the complement of vector $\mathbf{S}_2$ in vector $\mathbf{S}_1$. For a positive integer M, denote $[M] = \{1, \ldots, M\}$. After obtaining the index vector of the pruned subjects $\widehat{\mathbf{P}}$ (which is a subvector of $(1, 2, \ldots, N)$) from Algorithm 1 (line 2), we use SPA on the pruned matrix $\widehat{\mathbf{U}}_{[N]\backslash\widehat{\mathbf{P}},:}$ to obtain the estimated pure subject index vector $\widehat{\mathbf{S}}$ (line 3- 8). Once this is achieved, $\mathbf{\Pi}$ and $\mathbf{\Theta}$ can be estimated by modifying (8) and (9) in Proposition 1. We first calculate

$$\widetilde{\mathbf{\Pi}} = \widehat{\mathbf{U}} \left(\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}\right)^{-1} \tag{12}$$

based on (8). The $\widetilde{\mathbf{\Pi}}$ obtained above does not necessarily fall into the parameter domain for $\mathbf{\Pi}$. Therefore, we first truncate all the entries of $\widetilde{\mathbf{\Pi}}$ to be nonnegative and then re-normalize each row to sum to one (line 10). Based on $\widehat{\mathbf{\Pi}}$, we can also estimate $\mathbf{\Theta}$ by

$$\widetilde{\mathbf{\Theta}} = \widehat{\mathbf{V}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{U}}^\top \widehat{\mathbf{\Pi}}(\widehat{\mathbf{\Pi}}^\top \widehat{\mathbf{\Pi}})^{-1} \tag{13}$$

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

Algorithm 2
GoM Estimation by Successive Projection Algorithm with Pruning

Input: Binary response matrix $\mathbf{R} \in \{0,1\}^{N\times J}$, number of extreme profiles K, threshold parameter ϵ, pruning parameters r, q, e.
Output: Estimated $\widehat{\mathbf{S}}, \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Pi}}$

1: Get the top K singular value decomposition of $\mathbf{R}$ as $\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top$
2: $\widehat{\mathbf{P}} = \text{Prune}(\widehat{\mathbf{U}}, r, q, e)$
3: $\mathbf{Y} = \widehat{\mathbf{U}}$
4: **for** $k \in \{1, \ldots, K\}$ **do**
5: $\quad \widehat{S}_k = \text{argmax}(\{\|\mathbf{Y}_{i,:}\|_2 : i \in [N]\backslash\widehat{\mathbf{P}}\})$
6: $\quad \mathbf{u} = \mathbf{Y}_{\widehat{S}_k,:}/\|\mathbf{Y}_{\widehat{S}_k,:}\|_2$
7: $\quad \mathbf{Y} = \mathbf{Y}(\mathbf{I}_K - \mathbf{u}\mathbf{u}^\top)$
8: **end for**
9: $\widetilde{\mathbf{\Pi}} = \widehat{\mathbf{U}}(\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:})^{-1}$
10: $\widehat{\mathbf{\Pi}} = \text{diag}(\widetilde{\mathbf{\Pi}}_+\mathbf{1}_K)^{-1}\widetilde{\mathbf{\Pi}}_+$
11: $\widetilde{\mathbf{\Theta}} = \widehat{\mathbf{V}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{U}}^\top\widehat{\mathbf{\Pi}}(\widehat{\mathbf{\Pi}}^\top\widehat{\mathbf{\Pi}})^{-1}$
12: Write $\widehat{\mathbf{\Theta}}$ as $(\widehat{\theta}_{j,k})$, where $\widehat{\theta}_{j,k} = \begin{cases} \widetilde{\theta}_{j,k} & \text{if } \epsilon \le \widetilde{\theta}_{j,k} \le 1-\epsilon \\ \epsilon & \text{if } \widetilde{\theta}_{j,k} < \epsilon \\ 1-\epsilon & \text{if } \widetilde{\theta}_{j,k} > 1-\epsilon \end{cases}$
($\epsilon \ge 0$ can be set to zero. In the numerical studies, we choose $\epsilon = 0.001$ to be consistent and comparable with the default setting in the JML function in the R package `sirt`.)

according to (9). Our proposed method can be viewed as a method of moments. Equations (12) and (13) are based on the first moment of the response matrix $\mathbf{R}$, where we equate the low-rank structure of the population first-moment matrix $\mathbf{R}_0$ with the observed first-moment matrix $\mathbf{R}$. Lastly, we truncate $\widetilde{\mathbf{\Theta}}$ to be between $[\epsilon, 1-\epsilon]$ and obtain the final estimator $\widehat{\mathbf{\Theta}}$ (line 12). When $\epsilon = 0$, this truncation ensures that entries of $\widehat{\mathbf{\Theta}}$ lie in the parameter domain $[0, 1]$. In the numerical studies, we choose $\epsilon = 0.001$ to be consistent and comparable with the default setting in the JML function in the R package `sirt`.

Remark 1. There are two possible ways to estimate $\mathbf{\Theta}$ according to (9). The first one is defining $\widehat{\mathbf{\Theta}}$ as the truncated version of

$$\widetilde{\mathbf{\Theta}} = \widehat{\mathbf{V}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}^\top,$$

which only uses the information of $\widehat{\mathbf{U}}$ corresponding to the K pure subjects indexed by $\widehat{\mathbf{S}}$. Another approach is estimating $\mathbf{\Theta}$ via (13), which uses information from all of the N subjects. The latter approach is expected to give more stable estimates. Our preliminary simulations also justify that using (13) indeed gives higher estimation accuracy, so we choose to estimate $\mathbf{\Theta}$ that way.

3.2. Estimation Consistency

In this subsection, we prove that our spectral method guarantees estimation consistency of both the individual membership scores $\mathbf{\Pi}$ and the item parameters $\mathbf{\Theta}$. We consider the double-asymptotic regime where both the number of subjects N and the number of items J grow to infinity and K is fixed. At the high level, since our estimators are functions of the empirical SVD, we will prove consistency by leveraging singular subspace perturbation theory (Chen et al., 2021a) that quantifies the discrepancy between the SVD of $\mathbf{R}$ and that of the low-rank expectation $\mathbf{R}_0$.

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

Before stating the theorem, we introduce some notations. Write $f(n) \lesssim g(n)$ if there exists a constant $c > 0$ such that $|f(n)| \leq c|g(n)|$ holds for all sufficiently large n, and write $f(n) \gtrsim g(n)$ if there exists a constant $c > 0$ such that $|f(n)| \geq c|g(n)|$ holds for all sufficiently large n. For any matrix $\mathbf{A}$, denote its kth largest singular value as $\sigma_k(\mathbf{A})$, and define its condition number $\kappa(\mathbf{A})$ as the ratio of its largest singular value to its smallest singular value. For any $m \times n$ matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{m\times n}$, denote its Frobenius norm by $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$.

Condition 2. *$\kappa(\mathbf{\Pi}) \lesssim 1$, $\kappa(\mathbf{\Theta}) \lesssim 1$, $\sigma_K(\mathbf{\Pi}) \gtrsim \sqrt{N}$, and $\sigma_K(\mathbf{\Theta}) \gtrsim \sqrt{J}$.*

Condition 2 is a reasonable and mild assumption on the parameter matrices. As a simple example, it is not hard to verify that if $\mathbf{\Pi}$ and $\mathbf{\Theta}$ both consist of many identity submatrices $\mathbf{I}_K$ vertically stacked together, then Condition 2 is satisfied. The following theorem establishes the consistency of estimating both $\mathbf{\Pi}$ and $\mathbf{\Theta}$ using our spectral estimator.

Theorem 3. *Consider $\widetilde{\mathbf{\Pi}} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}^{-1}$, $\widetilde{\mathbf{\Theta}} = \widehat{\mathbf{V}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}^{\top}$. Assume that Conditions 1 and 2 hold. If $N, J \to \infty$ with $N/J^2 \to 0$ and $J/N^2 \to 0$, then we have*

$$\frac{1}{\sqrt{NK}}\|\widetilde{\mathbf{\Pi}} - \mathbf{\Pi}\mathbf{P}\|_F \xrightarrow{P} 0, \quad \frac{1}{\sqrt{JK}}\|\widetilde{\mathbf{\Theta}}\mathbf{P} - \mathbf{\Theta}\|_F \xrightarrow{P} 0, \tag{14}$$

where the notation $\xrightarrow{P}$ means convergence in probability, and $\mathbf{P}$ is a $K \times K$ permutation matrix which has exactly one entry of "1" in each row/column and zero entries elsewhere.

Theorem 3 implies that

$$\frac{1}{\sqrt{NK}}\|\widetilde{\mathbf{\Pi}} - \mathbf{\Pi}\mathbf{P}\|_F = \sqrt{\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}(\widetilde{\pi}_{i,k} - \pi_{i,\phi(k)})^2} \xrightarrow{P} 0;$$
$$\frac{1}{\sqrt{JK}}\|\widetilde{\mathbf{\Theta}}\mathbf{P} - \mathbf{\Theta}\|_F = \sqrt{\frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}(\widetilde{\theta}_{j,\phi(k)} - \theta_{j,k})^2} \xrightarrow{P} 0,$$

where $\phi : \{1, \ldots, K\} \to \{1, \ldots, K\}$ is a permutation map determined by the $K \times K$ permutation matrix $\mathbf{P}$. These results mean that as $N, J \to \infty$, the average squared estimation error across all entries in the mixed membership score matrix $\mathbf{\Pi}$ and that across all entries in the item parameter matrix $\mathbf{\Theta}$ both converge to zero in probability. This double-asymptotic regime with both N and J going to infinity and this consistency notion in scaled Frobenius norm are similar to those considered in the joint MLE approach to item factor analysis in Chen et al. (2019) and Chen et al. (2020). To the best of our knowledge, this is the first time that consistency results are established for GoM models in this modern regime.

4. Simulation Studies

4.1. Evaluating the Proposed Method

We carry out simulation studies to evaluate the accuracy and computational efficiency of our new method. We consider $K \in \{3, 8\}$, $N \in \{200, 1000, 2000, 3000, 4000, 5000\}$, and $J = N/5$, which correspond to large-scale and high-dimensional data scenarios. Such simulation regimes

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

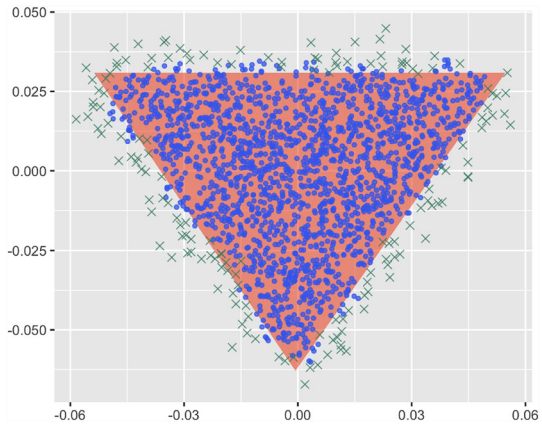

FIGURE 2.
Row vectors of $\mathbf{U}$ and $\widehat{\mathbf{U}}$ projected to $\mathbb{R}^2$ in the simulation setting with $N = 2000$ and $K = 3$. The red-shaded area is the population simplex, the green crosses are the removed subjects in pruning, and the blue dots form the empirical simplex retained after pruning.

share a similar spirit with those in Zhang et al. (2020), which proposed and evaluated an SVD-based approach for item factor analysis. In each simulation setting, we generate 100 independent replicates. Within each setting and each replication, the rows of $\mathbf{\Pi}$ are independently simulated from the Dirichlet($\boldsymbol{\alpha}$) distribution with $\boldsymbol{\alpha}$ equal to the K-dimensional all-one vector, and the first K rows of $\mathbf{\Pi}$ are set to the identity matrix $\mathbf{I}_K$ in order to satisfy the pure-subject Condition 1. The entries in $\mathbf{\Theta}$ are independently simulated from the uniform distribution on [0, 1]. Our preliminary simulations suggest setting the tuning parameters in the pruning Algorithm 1 to $r = 10$, $q = 0.4$, $e = 0.2$ with 8% subjects removed before SPA, so that the pruned empirical simplex is adequately close to the population simplex. We use the above setting throughout all the simulations and the real data analysis unless otherwise specified. It turns out the performance of our method is robust to these tuning parameter values and not much tuning is needed in practice.

To illustrate the simplex geometry of the population and empirical left singular matrices $\mathbf{U}$ (from the SVD of $\mathbf{R}_0$) and $\widehat{\mathbf{U}}$ (from the top-K SVD of $\mathbf{R}$), we plot $\mathbf{U}_{i,:}$ and $\widehat{\mathbf{U}}_{i,:}$ with $N = 2000$ and $K = 3$ in Fig. 2. All vectors are projected to two dimensions for better visualization of the simplex (i.e., triangle) structure. In Fig. 2, the red-shaded area corresponds to the population simplex, the green crosses are the removed subjects selected by the pruning Algorithm 1, and the blue dots form the empirical simplex obtained after pruning. As we can see, the empirical row vectors $\widehat{\mathbf{U}}_{i,:}$ approximately form a simplex cloud around the population simplex. After pruning out the noisy vectors, the resulting empirical simplex is close to the population one. This fact not only illustrates the effectiveness of the pruning procedure in Algorithm 1, but also confirms the usefulness of our notion of expectation identifiability by showing the close proximity of $\mathbf{U}$ and $\widehat{\mathbf{U}}$. The resemblance between the scatter plots of the rows of $\mathbf{U}$ and $\widehat{\mathbf{U}}$ implies that the simplex geometry of $\mathbf{U}$ holds approximately for $\widehat{\mathbf{U}}$, and hence justifies that SPA can be applied to $\widehat{\mathbf{U}}$ to estimate $\mathbf{\Pi}$ and $\mathbf{\Theta}$.

We compare the performance of our proposed method with the `JML` algorithm (Joint Maximum Likelihood) in the R package `sirt`, because the latter is currently considered as the most efficient estimation method for GoM models. We follow the default settings of `JML` with the maximum iteration number set as 600, the global parameter convergence criterion as 0.001, the maximum change in relative deviance as 0.001, and the minimum value of π_{ik} and θ_{jk} as 0.001. We measure the parameter estimation error by the mean absolute error (MAE). That is, the error between the estimated $\widehat{\mathbf{\Pi}}$ and the ground truth $\mathbf{\Pi}$ for each replicate is quantified by the mean

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

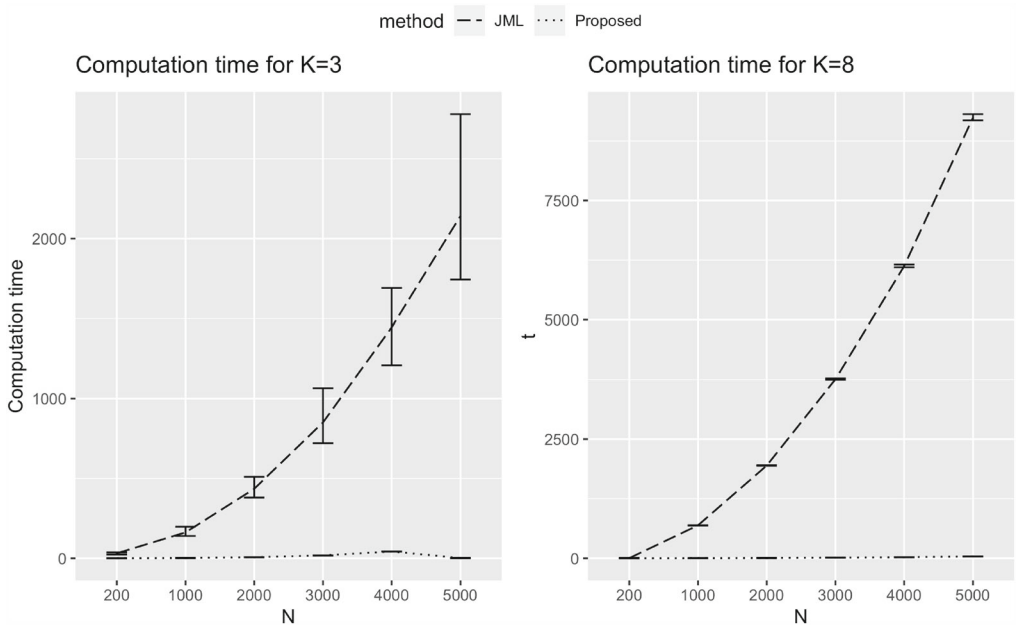

FIGURE 3.
Computation time for $K = 3$ (left) and $K = 8$ (right) in simulations. For each simulation setting, we show the median, 25% quantile, and 75% quantile of the computation time for the 100 replications.

absolute bias (and similarly for $\boldsymbol{\Theta}$):

$$l(\boldsymbol{\Pi}, \widehat{\boldsymbol{\Pi}}) := \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} |\pi_{ik} - \widehat{\pi}_{ik}|, \quad l(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}) := \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} |\theta_{jk} - \widehat{\theta}_{jk}|.$$

We present the comparisons between our spectral method and JML in terms of computation time and estimation error in Table 1 and Figs. 3, 4, 5. As shown in Fig. 3, the computation time for both methods increases as the sample size N and the number of items J increase. Notably, JML takes significantly more time than our proposed method, especially when N and J are large. For example, when $N = 5000$, $J = 1000$, $K = 8$, it takes about 3 h for JML to reach the maximum iteration number for each replication, while it takes less than 40 s on average for our proposed method. Table 1 further records the mean computational time in seconds across replications for JML and the proposed method. Moreover, we observe that JML is not able to converge in almost all replications when $K = 8$, including when the sample size is as small as $N = 200$. In summary, when the number of extreme latent profiles K is large enough, JML not only takes a long time to run but also is unable to reach convergence given the default convergence criterion.

The huge computational advantage of the proposed method does not come at the cost of degraded estimation accuracy. Figures 4 and 5 show that the estimation error decreases as the sample size N and the number of items J increase for both methods. When the sample size is large enough ($N \geq 2000$) for $K = 3$, our proposed method gives more accurate estimation on average compared to JML. For $K = 8$, our proposed method yields higher estimation accuracy for $\boldsymbol{\Theta}$ for all sample sizes on average. When $N \geq 2000$, the estimation accuracy of $\boldsymbol{\Pi}$ is slightly worse but comparable to JML. We point out that due to their non-iterative nature, SVD or eigendecomposition-based methods typically tend to give worse estimation compared to iterative

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

TABLE 1.
Table of average computational time in seconds across replications for JML and the proposed spectral method for $K = 3$ and $K = 8$

K	Method	$N = 200$	$N = 1000$	$N = 2000$	$N = 3000$	$N = 4000$	$N = 5000$
$K = 3$	JML	30.3	170.9	473.9	923.4	1528.4	2363.9
	Proposed	7.5e-2	1.5	6.8	17.5	42.3	10.6
$K = 8$	JML	1.7	691.5	1957.5	3761.7	6134.3	9251.7
	Proposed	6.4e-2	1.3	5.0	11.6	23.0	37.2

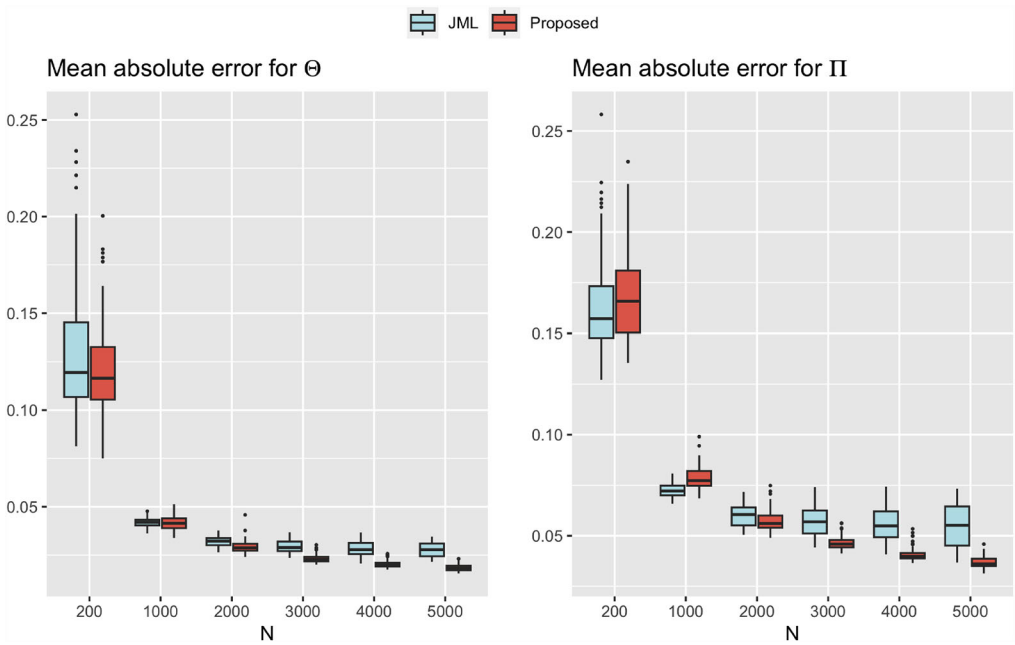

FIGURE 4.
Simulation results of estimation error for $K = 3$. The boxplots represent the mean absolute error for $\mathbf{\Theta}$ (left) and $\mathbf{\Pi}$ (right) versus the sample size N.

methods that aim to find the MLE; for example, see the comparison of an SVD-based method and a joint MLE method for item factor analysis in Zhang et al. (2020). However, it turns out that given a fixed computational resource (i.e., the default maximum iteration number in the `sirt` package), the iterative method JML can give worse estimation accuracy than our spectral method.

We also compare our spectral method with the Gibbs sampling method for the GoM model. The parameters in the Gibbs sampler are initialized from their prior distributions. The number of burn-in samples is set to 5000. We take the average of 2000 samples after the burn-in phase as the Gibbs sampling estimates. Since Gibbs sampling is an MCMC algorithm, the computation can take a long time when the sample size N and number of items J are large. Therefore, we only consider $N = 200, 1000, 2000$, and $K = 3$. We compare the computation time and estimation accuracy of our proposed method, JML, and Gibbs sampling. The results are summarized in Tables 2 and 3. Compared to Gibbs sampling, our proposed method is approximately 10,000 times faster when $N = 2000$ and $K = 3$. In terms of the estimation accuracy, when the sample

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

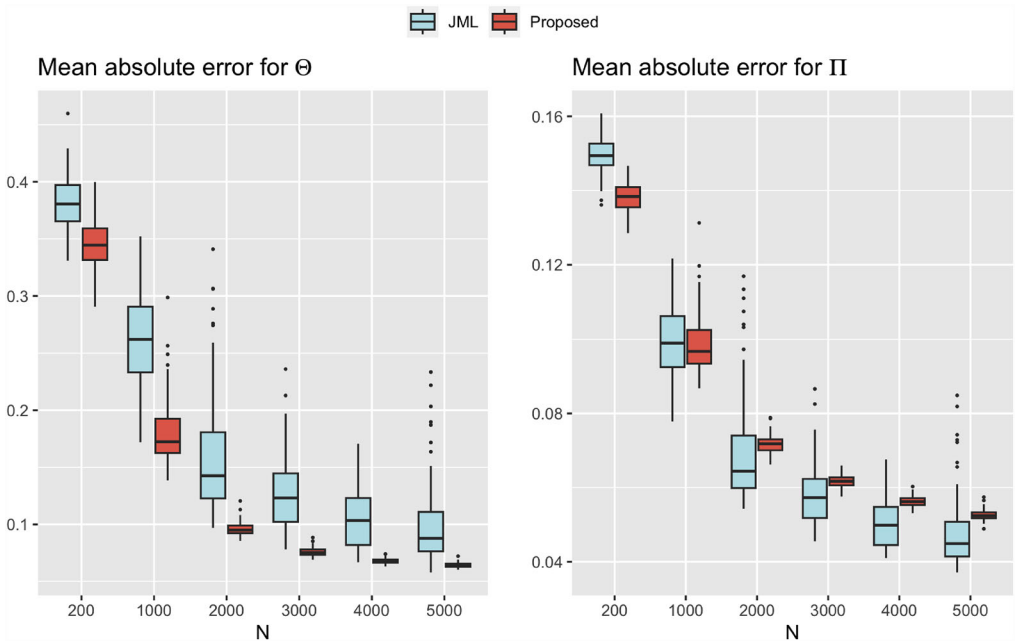

FIGURE 5.
Simulation results of estimation error for $K = 8$. The boxplots represent the mean absolute error for $\boldsymbol{\Theta}$ (left) and $\boldsymbol{\Pi}$ (right) versus the sample size N.

TABLE 2.
Average computation time in seconds for each method and sample size in simulations when $K = 3$

	$N = 200$	$N = 1000$	$N = 2000$
Proposed	0.1	1.5	6.8
JML	30.3	170.9	473.9
Gibbs	566.7	13119.2	52255.8

TABLE 3.
Average mean absolute error for $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$ for each method and sample size in simulations when $K = 3$

$\boldsymbol{\Theta}$	$N = 200$	$N = 1000$	$N = 2000$	$\boldsymbol{\Pi}$	$N = 200$	$N = 1000$	$N = 2000$
Proposed	0.12	0.04	0.03	Proposed	0.17	0.08	0.06
JML	0.13	0.04	0.03	JML	0.16	0.07	0.06
Gibbs	0.09	0.03	0.02	Gibbs	0.13	0.07	0.05

size is small, the Gibbs sampling has better accuracy. When the sample size is large enough, the three methods are comparable in terms of estimation accuracy. These simulation results justify that our method is more suitable for large-scale and high-dimensional datasets.

We also note that without the pure-subject Condition 1, one can still obtain estimates from the JML or the Gibbs sampler by directly running their corresponding estimation algorithms. However, simply running those algorithms for arbitrary data generated under the GoM model may

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

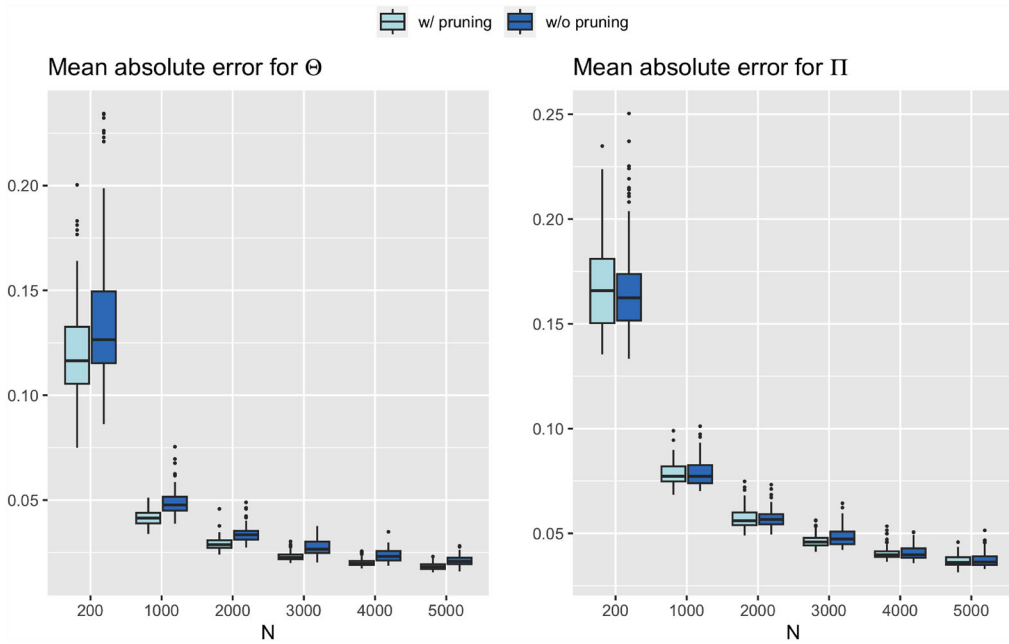

FIGURE 6.
Comparison of estimation results with and without the pruning procedure when $K = 3$.

give misleading results if the model is not identifiable, because identifiability is the prerequisite for any valid statistical inference. In contrast, for our proposed spectral estimator, under the pure subject condition, we have established both identifiability (Theorem 1) and estimation consistency (Theorem 3) for both the mixed membership scores $\mathbf{\Pi}$ and the item parameters $\mathbf{\Theta}$.

We also compare the simulation results of our proposed method with and without the pruning step when $K = 3$ to examine the effectiveness of pruning. The comparison of the estimation accuracy for the two approaches is summarized in Fig. 6. For the estimation of $\mathbf{\Theta}$, estimation with pruning is consistently better for all sample sizes. For the estimation of $\mathbf{\Pi}$, estimation results are comparable for the two approaches. When sample size is large, pruning gives slightly better results.

To summarize, the above simulation results demonstrate the superior efficiency and accuracy of our proposed method to estimate GoM models. Our method provides comparable estimation results compared with JML and Gibbs sampling and is much more scalable to large datasets with many subjects and many items.

4.2. *Verifying Identifiability*

We also conduct another simulation study to verify the identifiability results. We consider three different cases with $K = 3$, $N \in \{200, 1000, 2000, 3000\}$, and $J = N/5$. In Case 1, we set the ground truth $\mathbf{\Theta}$ by vertically concatenating copies of the identifiable 4×3 $\mathbf{\Theta}$-matrix in Example 1, and the generation mechanism for $\mathbf{\Pi}$ remains the same as in Sect. 4.1. In Case 2, we use the same $\mathbf{\Theta}$ as in Case 1, while for $\mathbf{\Pi}$, after generating rows of $\mathbf{\Pi}' = (\pi'_{ik})$ from Dirichlet($\mathbf{1}$), we truncate π'_{ik} to be no less than $1/3$ and then re-normalize each row of it; after this operation the minimum entry in the resulting $\mathbf{\Pi}$ is 0.2. Such a generated $\mathbf{\Pi} = (\pi_{ik})$ does not satisfy the pure-subject Condition 1. In Case 3, we generate $\mathbf{\Pi}$ using the same mechanism as in Sect. 4.1, but generate $\mathbf{\Theta}$ whose rows are replicates of the vector $(0.8,\ 0.5,\ 0.2)$ so that $rank(\mathbf{\Theta}) = 1$.

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

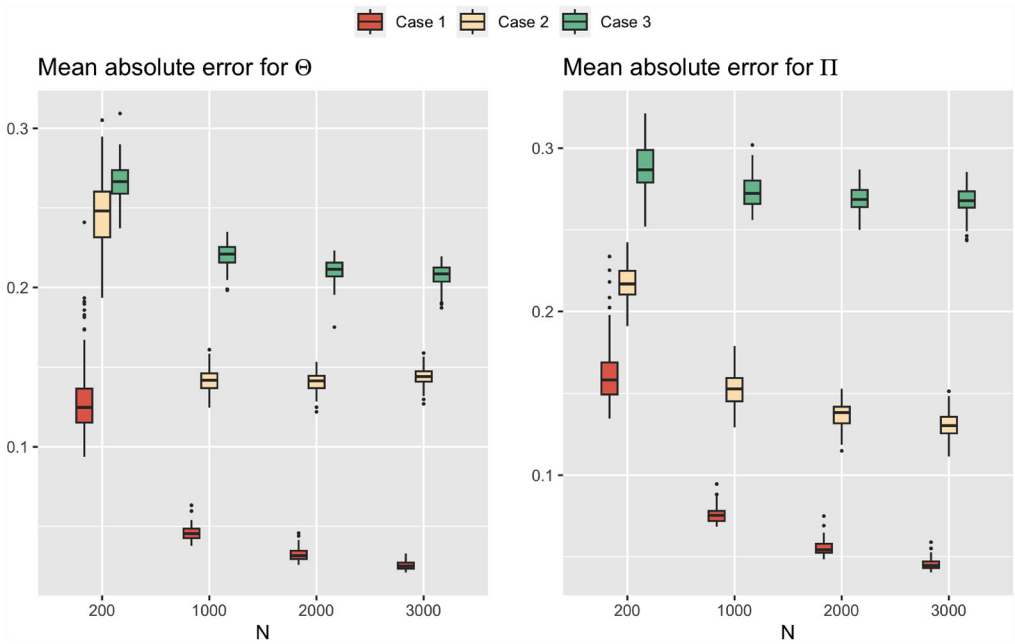

FIGURE 7.
A simulation study verifying identifiability. Estimation errors for three different cases; see the concrete settings of Cases 1, 2, and 3 in the main text. The box plots represent the mean absolute error for $\boldsymbol{\Theta}$ (left) and $\boldsymbol{\Pi}$ (right) versus the sample size N.

Case 1 falls into part (a) in Theorem 2 and is identifiable, whereas Cases 2 and 3 correspond to part (c) and are not identifiable. Figure 7 shows that the estimation errors in Cases 2 and 3 are significantly larger than those in Case 1. These results empirically verify our identifiability conclusions in Theorem 2.

5. Real Data Example

We illustrate our proposed method by applying it to a real-world personality test dataset, the Woodworth Psychoneurotic Inventory (WPI) dataset. WPI was designed by the United States Army during World War I to identify soldiers at risk for shell shock and is often credited as the first personality test. The WPI dataset can be downloaded from the Open Psychometrics Project website: http://openpsychometrics.org/_rawdata/. The dataset consists of binary yes/no responses to $J = 116$ items from 6019 subjects. We remove subjects with missing responses and only keep the subjects who are at least ten years old. This screening process leaves us with $N = 3842$ subjects.

We apply both the JML method and our new spectral method to the WPI dataset to compare computation time. Figure 8 shows the computation time in seconds versus the number of extreme profiles K for the two methods. For $K \geq 4$, the number of iterations for JML reaches the default maximum iteration number in the `sirt` package and does not converge. Similar to the simulations, our spectral method takes significantly less computation time compared to JML. This observation again confirms that the proposed method is scalable to real-world datasets with a large sample size and a relatively large number of items.

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

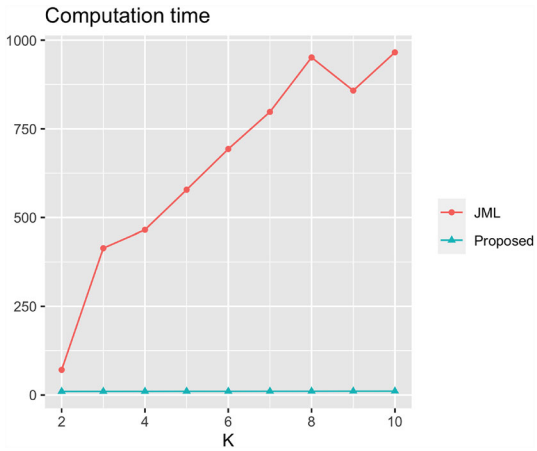

FIGURE 8.
Computation time for the WPI dataset. The lines indicate the run time in seconds versus K for JML and our spectral method. Note that for $K \geq 4$, the number of iterations in JML reaches the default maximum iteration number.

Choosing the number of extreme latent profiles K is a nontrivial problem for GoM models in practice. Available model selection techniques include the Akaike information criterion (AIC; Akaike 1998) and the Bayesian information criterion (BIC; Schwarz 1978) for likelihood-based methods, and the deviance information criteria (DIC; Spiegelhalter et al. 2002) for Bayesian MCMC methods. For factor analysis and principal component analysis, parallel analysis (Horn, 1965; Dobriban & Owen, 2019) is one popular eigenvalue-based method to select the latent dimension. For the WPI dataset, when choosing $K = 3$, we observe that the estimated $\widehat{\mathbf{\Theta}}$ matrix has three well-separated column vectors, which imply a meaningful interpretation of the extreme profiles. When increasing K to 4, the columns of the estimated $\widehat{\mathbf{\Theta}}$ are no longer that well separated and interpretable. Moreover, choosing $K = 3$ produces the least reconstruction error compared to $K = 2$ or 4. That is, $K = 3$ leads to the smallest mean absolute error between $\widehat{\mathbf{\Pi}}\widehat{\mathbf{\Theta}}^\top$ and the observed data matrix $\mathbf{R}$. Since the goal of the current data analysis is mainly to illustrate the proposed spectral method, we next present and discuss the estimation results for $K = 3$. The important problem of how to select K in GoM models in a principled manner is left as a future direction.

We next take a closer look at the estimation result given by the proposed method for $K = 3$. In order to interpret each extreme latent profile, we present the heatmap of part of the estimated item parameter matrix $\widehat{\mathbf{\Theta}}$ in Fig. 9. Since the number of items $J = 116$ is large, we only display a subset of the items for better visualization. More specifically, 30 out of 116 items are chosen in Fig. 9 based on the criterion that the chosen items have the largest variability in $(\widehat{\theta}_{j,1}, \widehat{\theta}_{j,2}, \widehat{\theta}_{j,3})$. Based on Fig. 9, we interpret profile 1 as people who are physically unhealthy since they have higher probabilities of fainting, dyspepsia, and asthma or hay fever. People belonging to profile 2 tend to be socially passive since they are worrying, do not find their way easily, and get tired of things or people easily. Profile 3 on the other hand is identified as the healthy group.

Figure 10 shows the ternary diagram of the estimated membership scores $\widehat{\mathbf{\Pi}}$ made with the R package `ggtern`. The WPI dataset comes with the age information of each subject, and we color code the subjects in Fig. 10 according to their ages. The darker color represents older people while the lighter color represents younger people. Each dot represents a subject, and the location of the dot in the equilateral triangle depicts the three membership scores of this subject. Specifically, dots with a large membership score on a profile are closer to the vertex of this profile. One can see that the pure-subject Condition 1 is satisfied here since there are dots located almost exactly at each

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

	profile 1	profile 2	profile 3
Is it easy to get you cross or grouchy?	0.458	0.999	0.324
Do you get tired of people quickly?	0.404	0.999	0.367
Does your mind wander badly so that you lose track of what you are doing?	0.335	0.999	0.32
Did you ever have the habit of biting your finger nails?	0.493	0.782	0.385
Do you have the sensation of falling when going to sleep?	0.498	0.897	0.241
Do your interests change frequently?	0.44	0.895	0.275
Have you ever had fits of dizziness?	0.51	0.999	0.214
Do you have nightmares?	0.516	0.87	0.132
Do you get tired of work?	0.478	0.999	0.573
Do you think you worry too much when you have an unfinished job on your hands?	0.483	0.999	0.529
Do ideas run through your head so that you cannot sleep?	0.68	0.999	0.626
As a child did you like to play alone better than to play with other children?	0.322	0.967	0.453
Do you worry too much about little things?	0.286	0.999	0.405
Were you shy with other boys [or girls]?	0.21	0.999	0.575
Did you ever have asthma or hay fever [allergies]?	0.559	0.578	0.262
Did you ever have dyspepsia [indigestion]?	0.631	0.562	0.119
Have you ever fainted away?	0.448	0.324	0.088
Did you ever have the habit of wetting the bed?	0.237	0.22	0.034
Do you find your way about easily?	0.999	0.205	0.796
Do you like outdoor life?	0.999	0.227	0.737
Do you get used to new places quickly?	0.999	0.132	0.802
Can you stand disgusting smells?	0.727	0.218	0.666
Have you a good appetite?	0.819	0.427	0.999
Did the other children let you play with them?	0.767	0.409	0.999
Can you do the little chores of the day without worrying over them?	0.891	0.172	0.996
Can you stand the sight of blood?	0.933	0.712	0.888
Can you stand pain quietly?	0.969	0.625	0.824
Is it easy to make you laugh?	0.838	0.577	0.925
Did the teachers in school generally treat you right?	0.629	0.578	0.999
Have your employers generally treated you right?	0.584	0.577	0.999

FIGURE 9.
Heatmap of $\widehat{\boldsymbol{\Theta}}$ of a subset of 30 WPI items. The values are the estimated probability of responding "yes" for each item given each extreme profile.

of the three vertices in Fig. 10. This figure also reveals that darker dots are more gathered around the vertex of profile 1, which means that older people are more likely to belong to the extreme profile identified as physically unhealthy. Correspondingly, younger people are closer to profile 2 or 3. Recalling our earlier interpretation of the extreme profiles from Fig. 9, these results are intuitively meaningful since older people tend to be less healthy compared with younger people. It is worth emphasizing that the age information is *not used* in our estimation of the GoM model, but our method is able to generate interpretable results with respect to age.

6. Discussion

In this paper, we have adopted a spectral approach to GoM analysis of multivariate binary responses. Under the notion of expectation identifiability, we have proposed sufficient conditions

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

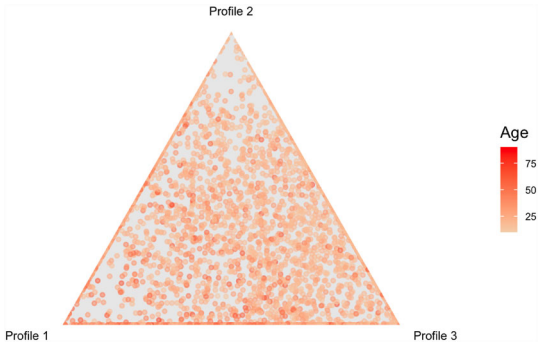

FIGURE 10.
Barycentric plot of the estimated membership scores $\widehat{\mathbf{\Pi}}$ for WPI data, color coded with the age covariate.

that are close to being necessary for GoM models to be identifiable. For estimation, we have proposed an efficient SVD-based spectral algorithm to estimate the subject-level and population-level parameters in the GoM model. Our spectral method has a huge computational advantage over Bayesian or likelihood-based methods and is scalable to large-scale and high-dimensional data. Simulation results demonstrate the superior efficiency and accuracy of our method, and also empirically corroborate our identifiability conclusions. We hope this work provides a useful tool for psychometric researchers and practitioners by making GoM analysis less computationally daunting and statistically mysterious.

The expectation identifiability considered in this work is a suitable identifiability notion for large-scale and high-dimensional GoM models. A recent paper Gu et al. (2023) studied the population identifiability of a variant of the GoM model called the dimension-grouped mixed membership model. Generally speaking, population identifiability is a traditional notion of identifiability which aims at identifying the population parameters in the model but not the individual latent variables. In the context of the dimension-grouped GoM model, population identifiability in Gu et al. (2023) means identifying the item parameters $\mathbf{\Theta}$ and the distribution parameters for $\mathbf{\Pi}$ (i.e., the Dirichlet distribution parameters $(\alpha_1, \ldots, \alpha_K)$, where each row in $\mathbf{\Pi}$ is assumed to follow Dirichlet$(\alpha_1, \ldots, \alpha_K)$), but not identifying the entries in $\mathbf{\Pi}$ directly. Such a traditional identifiability notion is more suitable for the low-dimensional case with small J and large N, as considered in Gu et al. (2023). In contrast, in this work we are motivated by the large-scale and high-dimensional setting with both N and J going to infinity. In this setting, expectation identifiability that concerns both $\mathbf{\Theta}$ and $\mathbf{\Pi}$ is a suitable notion to study, and we have also established the corresponding consistency result for $\mathbf{\Theta}$ and $\mathbf{\Pi}$ when the pure subject condition for expectation identifiability is satisfied.

The pure subject Condition 1 is a mild condition that is crucial to both our identifiability result and estimation procedure. It may be of interest to test whether this condition holds in practice. Testing Condition 1 is equivalent to testing whether a data cloud in a general K-dimensional space has a simplex structure, which is a non-trivial problem. When $K = 3$, a visual inspection is plausible by plotting the row vectors of the left singular matrix and checking if the point cloud forms a triangle in the 3-dimensional space. However, when K is larger than 3, visual inspection becomes infeasible. Recently, a formal statistical procedure has been proposed by a working paper Freyaldenhoven et al. (2023) to test the *anchor word assumption* for topic models, which is another type of mixed membership models. The anchor word assumption requires that there exists at least one anchor word for each topic, which is analogous to our pure subject Condition 1. Freyaldenhoven et al. (2023) considers the hypothesis testing of the existence of anchor words. They first characterizes that for a matrix $\mathbf{P}$ that admits a low-rank factorization under the anchor

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

word assumption, one can write $\mathbf{P} = \mathbf{CP}$ for some matrix $\mathbf{C}$ that belongs to a certain set $\mathcal{C}_K$. Based on this property, Freyaldenhoven et al. (2023) then construct a test statistic $T = \inf_{\mathbf{C}\in\mathcal{C}_K} \|\mathbf{P}-\mathbf{CP}\|$ to test the null hypothesis that the anchor-word assumption holds. To achieve a level-α test, the null hypothesis will be rejected if T is larger than the $(1-\alpha)\%$ quantile of the distribution of the test statistic under the null. We conjecture that it might be possible to generalize this procedure to the GoM setting and leave this direction as future work.

There are several additional research directions worth exploring in the future. First, this work has focused on binary responses, while in practice it is also of interest to perform GoM analysis of polytomous responses, such as Likert-scale responses in psychological questionnaires (see, e.g., Gu et al. 2023). It would be desirable to extend our method to such multivariate polytomous data. Under a GoM model for polytomous data, a similar low-rank structure as (5) should still exist for each category of the responses. Since our method is built upon the low-rank structure and the pure subject condition, we conjecture that exploiting such structures in polytomous data could still lead to nice spectral algorithms. Second, it is worth considering developing a model that directly incorporates additional covariates into the GoM analysis. Our current method does not use additional covariates, such as the age information in the WPI dataset. Proposing a covariate-assisted GoM model and a corresponding spectral method may be methodologically interesting and practically useful.

Acknowledgments

This work is partially supported by National Science Foundation Grant DMS-2210796. The authors thank the editor Prof. Matthias von Davier, an associate editor, and three anonymous reviewers for their helpful and constructive comments.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code Availability The MATLAB code implementing the proposed method is available at this link: https://github.com/lscientific/spectral_GoM.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Appendix A: Proofs of the Identifiability Results

Proof of Proposition 1. If we take the rows that correspond to $\mathbf{S}$ of both sides in the SVD (6) and use the fact that $\mathbf{\Pi}_{\mathbf{S},:} = \mathbf{I}_K$, then

$$\mathbf{U}_{\mathbf{S},:}\mathbf{\Sigma}\mathbf{V}^\top = [\mathbf{R}_0]_{\mathbf{S},:} = \mathbf{\Pi}_{\mathbf{S},:}\mathbf{\Theta}^\top = \mathbf{\Theta}^\top. \tag{15}$$

This gives an expression of $\mathbf{\Theta}$

$$\mathbf{\Theta} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}_{\mathbf{S},:}^\top. \tag{16}$$

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

Further note that

$$\mathbf{U} = \mathbf{R}_0 \mathbf{V} \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Pi} \boldsymbol{\Theta}^\top \mathbf{V} \boldsymbol{\Sigma}^{-1}. \tag{17}$$

If we plug (15) into (17) and note that $\mathbf{V}$ have orthogonal columns, we have

$$\mathbf{U} = \boldsymbol{\Pi} \mathbf{U}_{\mathbf{S},:} \boldsymbol{\Sigma} \mathbf{V}^\top \mathbf{V} \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Pi} \mathbf{U}_{\mathbf{S},:}. \tag{18}$$

Equation (18) also tells us that $\mathbf{U}_{\mathbf{S},:}$ must be full rank since both $\mathbf{U}$ and $\boldsymbol{\Pi}$ have rank K. Therefore, we have an expression of $\boldsymbol{\Pi}$:

$$\boldsymbol{\Pi} = \mathbf{U}(\mathbf{U}_{\mathbf{S},:})^{-1}.$$

On the other hand, based on the singular value decomposition $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top = \boldsymbol{\Pi}\boldsymbol{\Theta}^\top$, we can left multiply $(\boldsymbol{\Pi}^\top \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}^\top$ with both hand sides to obtain

$$\begin{aligned} &(\boldsymbol{\Pi}^\top \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}^\top \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^\top = \boldsymbol{\Theta}^\top \\ \implies \quad &\boldsymbol{\Theta} = \mathbf{V} \boldsymbol{\Sigma} \mathbf{U}^\top \boldsymbol{\Pi} (\boldsymbol{\Pi}^\top \boldsymbol{\Pi})^{-1}. \end{aligned} \tag{19}$$

We next show that (16) and (19) are equivalent. Since $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_K$, (18) leads to

$$\mathbf{U}_{\mathbf{S},:}^\top \boldsymbol{\Pi}^\top \boldsymbol{\Pi} \mathbf{U}_{\mathbf{S},:} = \mathbf{I}_K,$$

which yields

$$(\boldsymbol{\Pi}^\top \boldsymbol{\Pi})^{-1} = \mathbf{U}_{\mathbf{S},:} \mathbf{U}_{\mathbf{S},:}^\top.$$

Plugging this equation into (19), we have

$$\begin{aligned} \boldsymbol{\Theta} &= \mathbf{V} \boldsymbol{\Sigma} \mathbf{U}^\top \boldsymbol{\Pi} \mathbf{U}_{\mathbf{S},:} \mathbf{U}_{\mathbf{S},:}^\top \\ &= \mathbf{V} \boldsymbol{\Sigma} \mathbf{U}^\top \mathbf{U} (\mathbf{U}_{\mathbf{S},:})^{-1} \mathbf{U}_{\mathbf{S},:} \mathbf{U}_{\mathbf{S},:}^\top \\ &= \mathbf{V} \boldsymbol{\Sigma} \mathbf{U}_{\mathbf{S},:}^\top \end{aligned}$$

This shows the equivalence of (16) and (19) and completes the proof of the proposition. □

Proof of Theorem 1. Without loss of generality, assume that the first extreme latent profile does not have a pure subject. Then $\pi_{i1} \leq 1 - \delta$, $\forall i = 1, \ldots, N$ for some $\delta > 0$. For each $0 < \epsilon < \delta$, define a $K \times K$ matrix

$$\mathbf{M}_\epsilon = \begin{bmatrix} 1 + (K-1)\epsilon^2 & -\epsilon^2 \mathbf{1}_{K-1}^\top \\ \mathbf{0}_{K-1} & \epsilon \mathbf{1}_{K-1} \mathbf{1}_{K-1}^\top + (1 - (K-1)\epsilon) \mathbf{I}_{K-1} \end{bmatrix}.$$

We will show that $\widetilde{\boldsymbol{\Pi}}_\epsilon = \boldsymbol{\Pi} \mathbf{M}_\epsilon$ and $\widetilde{\boldsymbol{\Theta}}_\epsilon = \boldsymbol{\Theta} \mathbf{M}_\epsilon^{-1}$ form a valid parameter set. That is, each element of $\widetilde{\boldsymbol{\Pi}}_\epsilon$ and $\widetilde{\boldsymbol{\Theta}}_\epsilon$ lies in $[0, 1]$ and the rows of $\widetilde{\boldsymbol{\Pi}}_\epsilon$ sum to one. Since $\mathbf{M}_0 = \mathbf{I}_K$ and by the continuity of matrix determinant, $\mathbf{M}_\epsilon$ is full rank when ϵ is small enough. Also notice that $\mathbf{M}_\epsilon \mathbf{1}_K = \mathbf{1}_K$. Therefore, $\widetilde{\boldsymbol{\Pi}}_\epsilon \mathbf{1}_K = \boldsymbol{\Pi} \mathbf{M}_\epsilon \mathbf{1}_K = \boldsymbol{\Pi} \mathbf{1}_K = \mathbf{1}_N$. For each $i = 1, \ldots, N$, $\widetilde{\pi}_{i1} = \pi_{i1}(1 + (K-1)\epsilon^2) \geq 0$. For any fixed $k = 2, \ldots, K$, $(\mathbf{M}_\epsilon)_{kk} = 1 - (K-2)\epsilon$ and $(\mathbf{M}_\epsilon)_{mk} = \epsilon$ for $m \neq k$. Thus when $\epsilon \leq 1/(K-1)$, we have $(\mathbf{M}_\epsilon)_{mk} \geq \epsilon$ for any $m = 1, \ldots, K$. Therefore, the following inequalities hold for each $i = 1, \ldots, N$ and $k = 2, \ldots, K$:

$$\widetilde{\pi}_{ik} = -\epsilon^2 \pi_{i1} + \sum_{m=2}^{K} \pi_{im} (\mathbf{M}_\epsilon)_{mk} \geq -\epsilon^2 \pi_{i1} + \epsilon \sum_{m=2}^{K} \pi_{im}$$

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

$$\geq -\epsilon^2(1-\delta) + \epsilon(1-\pi_{i1}) \geq -\epsilon^2(1-\delta) + \epsilon\delta \geq \epsilon\delta^2 > 0.$$

Here we also used $\sum_{k=1}^{K} \pi_{ik} = 1$, $\pi_{i1} \leq 1-\delta$, and $\epsilon < \delta$.
Further notice that

$$\mathbf{M}_\epsilon - \mathbf{I}_K = \begin{bmatrix} \epsilon^2(K-1) & -\epsilon^2 \mathbf{1}_{K-1}^\top \\ \mathbf{0}_{K-1} & \epsilon \mathbf{1}_{K-1}\mathbf{1}_{K-1}^\top - \epsilon(K-1)\mathbf{I}_{K-1} \end{bmatrix},$$

which leads to $\|\mathbf{M}_\epsilon - \mathbf{I}_K\|_F \xrightarrow{\epsilon \to 0} 0$. Here $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$ is the Frobenius norm of any matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{m \times n}$. By the continuity of matrix inverse and Frobenius norm,

$$\|\mathbf{M}_\epsilon^{-1} - \mathbf{I}_K\|_F \xrightarrow{\epsilon \to 0} 0.$$

Therefore,

$$\|\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}\|_F = \|\mathbf{\Theta}(\mathbf{I}_K - \mathbf{M}_\epsilon^{-1})\|_F \leq \|\mathbf{\Theta}\|_2 \|\mathbf{I}_K - \mathbf{M}_\epsilon^{-1}\|_F \xrightarrow{\epsilon \to 0} 0.$$

Since all the elements of $\mathbf{\Theta}$ are strictly in $(0, 1)$, the elements of $\widetilde{\mathbf{\Theta}}$ must be in $[0, 1]$ when ϵ is small enough. Also note that $\mathbf{M}_\epsilon$ is not a permutation matrix when $\epsilon > 0$; thus, the GoM model is not identifiable up to a permutation. This completes the proof of the theorem. □

Proof of Theorem 2. Suppose $rank(\mathbf{\Theta}) = r \leq K$. Now, consider SVD $\mathbf{R}_0 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{N \times r}$, $\mathbf{V} \in \mathbb{R}^{J \times r}$, $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$. For simplicity, we continue to use the same notations $\mathbf{U}$, $\mathbf{\Sigma}$, $\mathbf{V}$ here even though the matrix dimensions have changed. Without loss of generality, we can reorder the subjects and memberships so that $\mathbf{\Pi}_{1:K,:} = \mathbf{I}_K$ from Assumption 1. According to Proposition 1,

$$\mathbf{U} = \mathbf{\Pi}\mathbf{U}_{1:K,:}. \tag{20}$$

Since $rank(\mathbf{\Pi}) = K$, $rank(\mathbf{U}) = r$, we must have $\text{rank}(\mathbf{U}_{1:K,:}) = \text{rank}(\mathbf{U}) = r$.
Suppose another set of parameters $(\widetilde{\mathbf{\Pi}}, \widetilde{\mathbf{\Theta}})$ yields the same $\mathbf{R}_0$ and we denote its corresponding pure subject index vector as $\widetilde{\mathbf{S}}$ so that $\widetilde{\mathbf{\Pi}}_{\widetilde{\mathbf{S}},:} = \mathbf{I}_K$. Similarly, we have

$$\mathbf{U} = \widetilde{\mathbf{\Pi}}\mathbf{U}_{\widetilde{\mathbf{S}},:}. \tag{21}$$

Taking the $\widetilde{\mathbf{S}}$ rows of both sides of (20) and the first K rows of both sides of (21) yields

$$\mathbf{\Pi}_{\widetilde{\mathbf{S}},:}\mathbf{U}_{1:K,:} = \mathbf{U}_{\widetilde{\mathbf{S}},:}, \ \mathbf{U}_{1:K,:} = \widetilde{\mathbf{\Pi}}_{1:K,:}\mathbf{U}_{\widetilde{\mathbf{S}},:}.$$

The above equation shows that $\mathbf{U}_{\widetilde{\mathbf{S}},:}$ is in the convex hull created by the rows of $\mathbf{U}_{1:K,:}$, and $\mathbf{U}_{1:K,:}$ is in the convex hull created by the rows of $\mathbf{U}_{\widetilde{\mathbf{S}},:}$. Therefore, there must exist a permutation matrix $\mathbf{P}$ such that $\mathbf{U}_{\widetilde{\mathbf{S}},:} = \mathbf{P}\mathbf{U}_{1:K,:}$. Combining this fact with (20) and (21) leads to

$$(\mathbf{\Pi} - \widetilde{\mathbf{\Pi}}\mathbf{P})\mathbf{U}_{1:K,:} = 0. \tag{22}$$

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

Proof of part (a). For part (a), $r = K$ and $\mathbf{U}_{1:K,:}$ is full rank according to (20). In this case, (22) directly leads to $\mathbf{\Pi} = \widetilde{\mathbf{\Pi}}\mathbf{P}$ and thus $\widetilde{\mathbf{\Theta}} = \mathbf{\Theta}\mathbf{P}^\top$.
Now generally consider $r < K$. By permuting the rows and columns of $\mathbf{\Theta}$, we can write

$$\mathbf{\Theta} = \begin{bmatrix} \mathbf{C} & \mathbf{C}\mathbf{W}_1 \\ \mathbf{W}_2^\top \mathbf{C} & \mathbf{W}_2^\top \mathbf{C}\mathbf{W}_1 \end{bmatrix}, \tag{23}$$

where $\mathbf{C} \in \mathbb{R}^{r\times r}$ is full rank, $\mathbf{W}_1 \in \mathbb{R}^{r\times(K-r)}$ and $\mathbf{W}_2 \in \mathbb{R}^{r\times(J-r)}$. Now comparing the block columns of (23) and $\mathbf{\Theta} = \mathbf{V}\mathbf{\Sigma}(\mathbf{U}_{1:K,:})^\top$ gives

$$\begin{aligned} \begin{bmatrix} \mathbf{I}_r \\ \mathbf{W}_2^\top \end{bmatrix} \mathbf{C} &= \mathbf{V}\mathbf{\Sigma}(\mathbf{U}_{1:r,:})^\top, \\ \begin{bmatrix} \mathbf{I}_r \\ \mathbf{W}_2^\top \end{bmatrix} \mathbf{C}\mathbf{W}_1 &= \mathbf{V}\mathbf{\Sigma}(\mathbf{U}_{(r+1):N,:})^\top. \end{aligned} \tag{24}$$

Since $\mathbf{C}$ is full rank, $\mathbf{U}_{1:r,:}$ has to also be full rank and (24) can be translated into

$$\mathbf{U}_{(r+1):N,:} = \mathbf{W}_1^\top \mathbf{U}_{1:r,:}.$$

Therefore,

$$\mathbf{U}_{1:K,:} = \begin{bmatrix} \mathbf{I}_r \\ \mathbf{W}_1^\top \end{bmatrix} \mathbf{U}_{1:r,:}. \tag{25}$$

By plugging the (25) into (22) and again using the fact that $\mathbf{U}_{1:r,:}$ is full rank, we have

$$(\mathbf{\Pi} - \widetilde{\mathbf{\Pi}}\mathbf{P}) \begin{bmatrix} \mathbf{I}_r \\ \mathbf{W}_1^\top \end{bmatrix} = \mathbf{0}. \tag{26}$$

Proof of part (b). Denote $\mathbf{A} := \mathbf{\Pi} - \widetilde{\mathbf{\Pi}}\mathbf{P}$. If $r = K - 1$, $\mathbf{W}_1 = (W_{1,1}, \ldots, W_{1,K-1})$ is a $(K-1)$-dimensional vector and (26) gives us

$$\mathbf{A}_{:,j} + W_{1,j}\mathbf{A}_{:,k} = 0, \quad \forall j = 1, \ldots, K-1. \tag{27}$$

Denote an r-dimensional column vector with all entries equal to one by $\mathbf{1}_r$. Right multiplying $\mathbf{1}_r$ to both sides of (26) yields

$$\mathbf{A} \begin{bmatrix} \mathbf{1}_r \\ \mathbf{W}_1^\top \mathbf{1}_r \end{bmatrix} = 0.$$

Also, note that both $\mathbf{\Pi}$ and $\widetilde{\mathbf{\Pi}}\mathbf{P}$ have row sums of 1. Hence,

$$\begin{aligned} &\sum_{j=1}^{K} \mathbf{A}_{:,j} = \mathbf{0}_N, \\ &\sum_{j=1}^{K-1} \mathbf{A}_{:,j} + \mathbf{W}_1^\top \mathbf{1}_r \mathbf{A}_{:,K} = \mathbf{0}_N. \end{aligned}$$

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

Taking the difference of the two equations above gives $(1 - \mathbf{W}_1^\top \mathbf{1}_r)\mathbf{A}_{:,K} = \mathbf{0}_N$. If $\mathbf{W}_1^\top \mathbf{1}_r \neq 1$, then $\mathbf{A}_{:,K}$ has to be $\mathbf{0}_N$, which implies $\mathbf{A}_{:,j} = \mathbf{0}_N$ for all $j = 1 \ldots, K-1$ according to (27). Therefore, $\mathbf{A} = \mathbf{\Pi} - \widetilde{\mathbf{\Pi}}\mathbf{P} = \mathbf{0}_N$, which leads to $\widetilde{\mathbf{\Theta}} = \mathbf{\Theta}\mathbf{P}^\top$.
Note that using (25) leads to

$$\mathbf{\Theta}^\top = \mathbf{U}_{1:K,:}\mathbf{\Sigma}\mathbf{V}^\top = \begin{bmatrix} \mathbf{I}_{K-1} \\ \mathbf{W}_1^\top \end{bmatrix} \mathbf{U}_{1:(K-1),:}\mathbf{\Sigma}\mathbf{V}^\top = \begin{bmatrix} \mathbf{I}_{K-1} \\ \mathbf{W}_1^\top \end{bmatrix} (\mathbf{\Theta}_{:,1:(K-1)})^\top.$$

Hence $\mathbf{\Theta}_{:,K} = \mathbf{\Theta}_{:,1:(K-1)}\mathbf{W}_1$. Therefore, the condition $\mathbf{W}_1^\top \mathbf{1}_r \neq 1$ is equivalent to the K-th column of $\mathbf{\Theta}$ not being an affine combination of the other columns.
Proof of part (c). Now consider the case of *either* $r = K-1$ with $\mathbf{W}_1^\top \mathbf{1}_r = 1$, *or* $r < K-1$. Assume subject m is completely mixed so that $\pi_{m,k} > 0, \forall k = 1, \ldots, K$. Define

$$\widetilde{\boldsymbol{\pi}}_i^\top = \begin{cases} \boldsymbol{\pi}_i^\top & \text{if } i \neq m \\ \boldsymbol{\pi}_m^\top + \epsilon \boldsymbol{\beta}^\top [-\mathbf{W}_1^\top,\ \mathbf{I}_{K-r}] & \text{if } i = m \end{cases},$$

where $\epsilon > 0$ is small enough so that $\widetilde{\boldsymbol{\pi}}_i \in (0, 1)$, and $\boldsymbol{\beta} \in \mathbb{R}^{K-r}$ is such that $\boldsymbol{\beta}^\top(\mathbf{1}_{K-r} - \mathbf{W}_1^\top \mathbf{1}_r) = 0$. Note that such $\boldsymbol{\beta} \neq \mathbf{0}$ always exists under the assumption in part (c), because if $r = K-1$ with $\mathbf{W}_1^\top \mathbf{1}_r = 1$, then $\boldsymbol{\beta}^\top(\mathbf{1}_{K-r} - \mathbf{W}_1^\top \mathbf{1}_r) = \beta(1-1) = 0$ holds for any $\beta \in \mathbb{R}$; if $r < K-1$, then $K - r \geq 2$ and $\boldsymbol{\beta}$ has dimension at least two, so the inner product equation $\boldsymbol{\beta}^\top(\mathbf{1}_{K-r} - \mathbf{W}_1^\top \mathbf{1}_r) = 0$ must have a nonzero solution $\boldsymbol{\beta}$. The constructed $\widetilde{\mathbf{\Pi}}$ have row sums of 1 by the construction of $\boldsymbol{\beta}$. Furthermore, $\widetilde{\mathbf{\Pi}}\mathbf{U}_{1:K,:}$ and $\mathbf{\Pi}\mathbf{U}_{1:K,:}$ can only be different on the m-th row, and

$$\widetilde{\boldsymbol{\pi}}_m^\top \mathbf{U}_{1:K,:} = \boldsymbol{\pi}_m^\top \mathbf{U}_{1:K,:} + \epsilon \boldsymbol{\beta}^\top [-\mathbf{W}_1^\top,\ \mathbf{I}_{K-r}] \begin{bmatrix} \mathbf{I}_r \\ \mathbf{W}_1^\top \end{bmatrix} \mathbf{U}_{1:r,:} = \boldsymbol{\pi}_m^\top \mathbf{U}_{1:K,:}.$$

Hence, $\widetilde{\mathbf{\Pi}}\mathbf{U}_{1:K,:} = \mathbf{\Pi}\mathbf{U}_{1:K,:}$. This gives us

$$\mathbf{\Pi}\mathbf{\Theta}^\top = \mathbf{\Pi}\mathbf{U}_{1:K,:}\mathbf{\Sigma}\mathbf{V}^\top = \widetilde{\mathbf{\Pi}}\mathbf{U}_{1:K,:}\mathbf{\Sigma}\mathbf{V}^\top = \widetilde{\mathbf{\Pi}}\mathbf{\Theta}^\top.$$

We can see that $(\mathbf{\Pi}, \mathbf{\Theta})$ and $(\widetilde{\mathbf{\Pi}}, \mathbf{\Theta})$ yield the same model but $\mathbf{\Pi} \neq \widetilde{\mathbf{\Pi}}$. This completes the proof for part (c). □

Appendix B: Proof of the Consistency Theorem 3

For any matrix $\mathbf{A}$ with SVD $\mathbf{A} = \mathbf{U}_\mathbf{A}\mathbf{\Sigma}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top$, define

$$\text{sgn}(\mathbf{A}) := \mathbf{U}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top.$$

According to Remark 4.1 in Chen et al. (2021a), for any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times r}$, $r \leq n$:

$$\text{sgn}(\mathbf{A}^\top \mathbf{B}) = \arg \min_{\mathbf{O} \in \mathcal{O}^{r \times r}} \|\mathbf{A}\mathbf{O} - \mathbf{B}\|,$$

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

where $\mathcal{O}^{r\times r}$ is the set of all orthonormal matrices of size $r \times r$. The 2-to-∞ norm of matrix $\mathbf{A}$ is defined as the maximum row l_2 norm, i.e., $\|\mathbf{A}\|_{2,\infty} = \max_i \|\mathbf{e}_i^\top \mathbf{A}\|$. Define

$$r = \frac{\max\{N, J\}}{\min\{N, J\}}.$$

Under Condition 2, we have $\kappa(\mathbf{R}_0) = \frac{\sigma_1(\mathbf{\Pi}\mathbf{\Theta}^\top)}{\sigma_K(\mathbf{\Pi}\mathbf{\Theta}^\top)} \leq \frac{\sigma_1(\mathbf{\Pi})\sigma_1(\mathbf{\Theta})}{\sigma_K(\mathbf{\Pi})\sigma_K(\mathbf{\Theta})} = \kappa(\mathbf{\Pi})\kappa(\mathbf{\Theta}) \lesssim 1$ and $\sigma_K(\mathbf{R}_0) \geq \sigma_K(\mathbf{\Pi})\sigma_K(\mathbf{\Theta}) \gtrsim \sqrt{NJ}$.

Lemma 1. *Under Condition 2, if $N/J^2 \to 0$ and $J/N^2 \to 0$, then with probability at least $1 - O((N+J)^{-5})$, one has*

$$\|\widehat{\mathbf{U}} - \mathbf{U} \cdot sgn(\mathbf{U}^\top \widehat{\mathbf{U}})\|_{2,\infty} \lesssim \frac{\sqrt{r} + \sqrt{\log(N+J)}}{\sqrt{NJ}} \tag{28}$$

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\|_\infty \lesssim \sqrt{\frac{r\log(N+J)}{\min\{N, J\}}}. \tag{29}$$

Here the infinity norm $\|\mathbf{A}\|_\infty$ for any matrix $\mathbf{A}$ is defined as the maximum absolute entry value. We write the RHS of (28) as ε and the RHS of (29) as η.

Proof of Lemma 1. We will use Theorem 4.4 in Chen et al. (2021a) to prove the lemma and will verify the conditions of that theorem are satisfied.

Define the incoherence parameter $\mu := \max\left\{\frac{N\|\mathbf{U}\|_{2,\infty}^2}{K}, \frac{J\|\mathbf{V}\|_{2,\infty}^2}{K}\right\}$. Note that

$$\|\mathbf{U}\|_{2,\infty} \leq \|\mathbf{U}_{\mathbf{S},:}\|_{2,\infty} \leq \|\mathbf{U}_{\mathbf{S},:}\| = \frac{1}{\sigma_K(\mathbf{\Pi})} \lesssim \frac{1}{\sqrt{N}},$$

since all rows of $\mathbf{U}$ are convex combinations of $\mathbf{U}_{\mathbf{S},:}$. On the other hand,

$$\begin{aligned}
\|\mathbf{V}\|_{2,\infty} &= \|\mathbf{\Theta}\mathbf{U}_{\mathbf{S},:}^{-\top}\mathbf{\Sigma}^{-1}\|_{2,\infty} \leq \|\mathbf{\Theta}\|_{2,\infty}\|\mathbf{U}_{\mathbf{S},:}^{-\top}\mathbf{\Sigma}^{-1}\| \\
&\leq \|\mathbf{\Theta}\|_{2,\infty}\|\mathbf{U}_{\mathbf{S},:}^{-1}\| \cdot \frac{1}{\sigma_K(\mathbf{\Pi}\mathbf{\Theta}^\top)} = \frac{\|\mathbf{\Theta}\|_{2,\infty}\sigma_1(\mathbf{\Pi})}{\sigma_K(\mathbf{\Pi}\mathbf{\Theta}^\top)} \\
&\leq \frac{\|\mathbf{\Theta}\|_{2,\infty}\kappa(\mathbf{\Pi})}{\sigma_K(\mathbf{\Theta})} \leq \frac{\sqrt{K}\kappa(\mathbf{\Pi})}{\sigma_K(\mathbf{\Theta})} \lesssim \frac{1}{\sqrt{J}}.
\end{aligned}$$

Therefore, $\mu \lesssim 1$.

On the other hand, we will show that $\sqrt{\log(N+J)/\min\{N, J\}} \lesssim 1$. By the symmetry of N and J, we assume $J \leq N$ without loss of generality. Thus,

$$\sqrt{\frac{\log(N+J)}{\min\{N, J\}}} = \sqrt{\frac{\log(N+J)}{J}} \lesssim \sqrt{\frac{\log(J^2+J)}{J}} \to 0.$$

Therefore, Assumption 4.2 in Chen et al. (2021a) holds and (28) and (29) can be directly obtained from Theorem 4.4 in Chen et al. (2021a). □

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

Lemma 2. *Let Conditions 1 and 2 hold. Then, there exists a permutation matrix* $\mathbf{P}$ *such that with probability at least* $1 - O((N+J)^{-5})$,

$$\|\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:} - \mathbf{P}\mathbf{U}_{\mathbf{S},:} \cdot sgn(\mathbf{U}^\top \widehat{\mathbf{U}})\| \lesssim \varepsilon. \tag{30}$$

Proof of Lemma 2. Using Proposition 1, we will apply Theorem 4 in Klopp et al. (2023) with $\widetilde{\mathbf{G}} = \widehat{\mathbf{U}}$, $\mathbf{G} = \mathbf{U} \cdot \mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})$, $\mathbf{W} = \mathbf{\Pi}$, $\mathbf{Q} = \mathbf{U}_{\mathbf{S},:} \cdot \mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})$, $\mathbf{N} = \widehat{\mathbf{U}} - \mathbf{U}\mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})$, $\mathbf{N} = \widehat{\mathbf{U}} - \mathbf{U} \cdot \mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})$. According to Lemma 1, $\|\mathbf{e}_i^\top \mathbf{N}\| \leq \varepsilon$ and $\varepsilon \lesssim \frac{\sqrt{r}+\sqrt{\log(N+J)}}{\sqrt{NJ}}$. On the other hand, $\sigma_K(\mathbf{Q}) = \sigma_K(\mathbf{U}_{\mathbf{S},:}) = \frac{1}{\sigma_1(\mathbf{\Pi})} \geq \frac{1}{\sqrt{N}}$ since $\mathbf{U} = \mathbf{\Pi}\mathbf{U}_{\mathbf{S},:}$ and $\sigma_1(\mathbf{\Pi}) \leq \|\mathbf{\Pi}\|_F \leq \sqrt{N}\max_i \left\|\mathbf{e}_i^\top \mathbf{\Pi}\right\|_2 \leq \sqrt{N}$. Therefore, $\varepsilon \leq C_* \frac{\sigma_K(\mathbf{U}_{\mathbf{S},:})}{K\sqrt{K}}$ for some $C_* > 0$ small enough. Then we can use Theorem 4 in Klopp et al. (2023) to get

$$\begin{aligned}
\|\widehat{\mathbf{U}} - \mathbf{P}\mathbf{U}_{\mathbf{S},:} \cdot \mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})\| &\leq C_0\sqrt{K}\kappa(\mathbf{U}_{\mathbf{S},:} \cdot \mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}}))\varepsilon \\
&= C_0\sqrt{K}\kappa(\mathbf{U}_{\mathbf{S},:})\varepsilon \overset{(i)}{=} C_0\sqrt{K}\kappa(\mathbf{\Pi})\varepsilon \\
&\lesssim \varepsilon \quad \text{with probability at least } 1 - O((N+J)^{-5}).
\end{aligned}$$

Here (i) is because $\mathbf{U} = \mathbf{\Pi}\mathbf{U}_{\mathbf{S},:}$. □

Proof of Theorem 3. First show that $\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}$ is not degenerate. By Weyl's inequality and Lemma 2, with probability at least $1 - O((N+J)^{-5})$, we have

$$\begin{aligned}
\sigma_K(\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}) &\geq \sigma_K(\mathbf{P}\mathbf{U}_{\mathbf{S},:}\mathbf{O}) - \|\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:} - \mathbf{P}\mathbf{U}_{\mathbf{S},:} \cdot \mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})\| \\
&\geq \sigma_K(\mathbf{U}_{\mathbf{S},:}) - \|\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:} - \mathbf{P}\mathbf{U}_{\mathbf{S},:} \cdot \mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})\|_F \\
&\gtrsim \frac{1}{\sigma_1(\mathbf{\Pi})} - \varepsilon \\
&\gtrsim \frac{1}{\sqrt{N}} - \frac{\sqrt{r} + \sqrt{\log(N+J)}}{\sqrt{NJ}} \\
&\gtrsim \frac{1}{\sqrt{N}}
\end{aligned}$$

when N, J are large enough and $\frac{N}{J^2}$ converges to zero. Therefore, $\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}$ is invertible.
For the estimation of $\mathbf{\Pi}$,

$$\begin{aligned}
\|\widetilde{\mathbf{\Pi}} - \mathbf{\Pi}\mathbf{P}\|_F &= \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}^{-1} - \mathbf{U}\mathbf{U}_{\mathbf{S},:}^{-1}\mathbf{P}\|_F \\
&\leq \underbrace{\|\widehat{\mathbf{U}}(\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}^{-1} - \mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})^\top \mathbf{U}_{\mathbf{S},:}^{-1}\mathbf{P})\|_F}_{I_1} + \underbrace{\|(\widehat{\mathbf{U}} - \mathbf{U}\mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}}))[\mathbf{P}^{-1}\mathbf{U}_{\mathbf{S},:}\mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})]^{-1}\|_F}_{I_2} \\
&=: I_1 + I_2.
\end{aligned}$$

We will look at I_1 and I_2 separately.

$$\begin{aligned}
I_1 &= \|\widehat{\mathbf{U}}(\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}^{-1} - \mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})^\top \mathbf{U}_{\mathbf{S},:}^{-1}\mathbf{P}^\top)\|_F \\
&\leq \|\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}^{-1} - \mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})^\top \mathbf{U}_{\mathbf{S},:}^{-1}\mathbf{P}^\top\|_F
\end{aligned}$$

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

$$
\begin{aligned}
&\le \|\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}^{-1}\| \|\mathrm{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})^\top \mathbf{U}_{\mathbf{S},:}^{-1}\mathbf{P}^\top\| \|\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:} - \mathbf{P}\mathbf{U}_{\mathbf{S},:}\mathrm{sgn}(\mathbf{U}^\top\widehat{\mathbf{U}})\|_F \\
&\lesssim \sqrt{N}\cdot \sigma_1(\mathbf{\Pi})\cdot \varepsilon \\
&\lesssim \sqrt{N}\cdot \frac{\sqrt{r}+\sqrt{\log(N+J)}}{\sqrt{J}} \quad \text{with probability at least } 1-O((N+J)^{-5});
\end{aligned}
$$

and

$$
\begin{aligned}
I_2 &= \|(\widehat{\mathbf{U}} - \mathbf{U}\mathrm{sgn}(\mathbf{U}^\top\widehat{\mathbf{U}}))[\mathbf{P}^{-1}\mathbf{U}_{\mathbf{S},:}\mathrm{sgn}(\mathbf{U}^\top\widehat{\mathbf{U}})]^{-1}\|_F \\
&\le \|\widehat{\mathbf{U}} - \mathbf{U}\mathrm{sgn}(\mathbf{U}^\top\widehat{\mathbf{U}})\|_F \|\mathbf{U}_{\mathbf{S},:}^{-1}\| \\
&\le \sqrt{N}\cdot\varepsilon\cdot\sigma_1(\mathbf{\Pi}) \\
&\lesssim \sqrt{N}\cdot \frac{\sqrt{r}+\sqrt{\log(N+J)}}{\sqrt{J}} \quad \text{with probability at least } 1-O((N+J)^{-5}).
\end{aligned}
$$

Therefore, with probability at least $1-O((N+J)^{-5})$,

$$
\begin{aligned}
\frac{1}{\sqrt{NK}}\|\widetilde{\mathbf{\Pi}} - \mathbf{\Pi}\mathbf{P}\|_F &\lesssim \frac{\sqrt{r}+\sqrt{\log(N+J)}}{\sqrt{J}} \\
&= \begin{cases} \frac{\sqrt{N}}{J} + \frac{\sqrt{\log(N+J)}}{\sqrt{J}} & \text{if } N > J, \\ \frac{1}{\sqrt{N}} + \frac{\sqrt{\log(N+J)}}{\sqrt{J}} & \text{if } N \le J. \end{cases}
\end{aligned}
$$

Therefore, $\frac{1}{\sqrt{NK}}\|\widetilde{\mathbf{\Pi}} - \mathbf{\Pi}\mathbf{P}\|_F$ converges to zero in probability as $N, J \to \infty$ and $\frac{N}{J^2} \to 0$. For the estimation of $\mathbf{\Theta}$,

$$
\begin{aligned}
&\|\widetilde{\mathbf{\Theta}}\mathbf{P} - \mathbf{\Theta}\|_F \\
&\quad = \|\mathbf{P}^\top\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top - \mathbf{U}_{\mathbf{S},:}\mathbf{\Sigma}\mathbf{V}^\top\|_F \\
&\quad \le \|(\mathbf{P}^\top\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:} - \mathbf{U}_{\mathbf{S},:}\mathrm{sgn}(\mathbf{U}^\top\widehat{\mathbf{U}}))\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top\|_F + \|(\mathbf{U}_{\mathbf{S},:}\mathrm{sgn}(\mathbf{U}^\top\widehat{\mathbf{U}}) - \widehat{\mathbf{U}}_{\mathbf{S},:})\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top\|_F \\
&\qquad + \|\widehat{\mathbf{U}}_{\mathbf{S},:}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top - \mathbf{U}_{\mathbf{S},:}\mathbf{\Sigma}\mathbf{V}^\top\|_F \\
&\quad \le \|\mathbf{P}^\top\widehat{\mathbf{U}}_{\widehat{\mathbf{S}},:} - \mathbf{U}_{\mathbf{S},:}\mathrm{sgn}(\mathbf{U}^\top\widehat{\mathbf{U}})\|_F\cdot\sigma_1(\mathbf{R})\cdot\|\widehat{\mathbf{V}}\| + \|\mathbf{U}_{\mathbf{S},:}\mathrm{sgn}(\mathbf{U}^\top\widehat{\mathbf{U}}) - \widehat{\mathbf{U}}_{\mathbf{S},:}\|\cdot\sigma_1(\mathbf{R})\cdot\|\widehat{\mathbf{V}}\| \\
&\qquad + \sqrt{KJ}\|\widehat{\mathbf{U}}_{\mathbf{S},:}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top - \mathbf{U}_{\mathbf{S},:}\mathbf{\Sigma}\mathbf{V}^\top\|_\infty \\
&\overset{(ii)}{\lesssim} \varepsilon\cdot\sigma_1(\mathbf{R}_0) + \varepsilon\cdot(\sigma_1(\mathbf{R}) - \sigma_1(\mathbf{R}_0)) + \sqrt{KJ}\cdot\eta \quad \text{with probability at least } 1-O((N+J)^{-5}),
\end{aligned}
$$

where (ii) results from Lemma 2. By Weyl's inequality, $|\sigma_1(\mathbf{R}) - \sigma_1(\mathbf{R}_0)| \le \|\mathbf{R} - \mathbf{R}_0\|$, where $\mathbf{R} - \mathbf{R}_0$ is a mean-zero Bernoulli matrix. According to Eq (3.9) in Chen et al. (2021a), with probability at least $1-(N+J)^{-8}$,

$$
\|\mathbf{R} - \mathbf{R}_0\| \lesssim \sqrt{N+J} + \sqrt{\log(N+J)}.
$$

Furthermore, $\sigma_1(\mathbf{R}_0) \ge \sigma_K(\mathbf{R}_0) \gtrsim \sqrt{NJ}$ by Condition 2; thus, we know that $\sigma_1(\mathbf{R}_0) \gtrsim |\sigma_1(\mathbf{R}) - \sigma_1(\mathbf{R}^*)|$ with probability at least $1-(N+J)^{-8}$. Therefore, with probability at least $1-O((N+J)^{-5})$,

$$
\|\widehat{\mathbf{\Theta}}\mathbf{P} - \mathbf{\Theta}\|_F \lesssim \varepsilon\cdot\sigma_1(\mathbf{R}_0) + \sqrt{KJ}\cdot\eta
$$

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

$$\lesssim \frac{\sqrt{r}+\sqrt{\log(N+J)}}{\sqrt{NJ}} \cdot \sqrt{N} \cdot \sqrt{J} + \sqrt{J}\sqrt{\frac{r\log(N+J)}{\min\{N,J\}}}$$
$$= \sqrt{r}+\sqrt{\log(N+J)} + \sqrt{J}\sqrt{\frac{r\log(N+J)}{\min\{N,J\}}}.$$

Thus,

$$\frac{1}{\sqrt{JK}}\|\widehat{\boldsymbol{\Theta}}\mathbf{P}-\boldsymbol{\Theta}\|_F \lesssim \frac{\sqrt{r}+\sqrt{\log(N+J)}}{\sqrt{J}} + \sqrt{\frac{r\log(N+J)}{\min\{N,J\}}}$$
$$= \begin{cases} \frac{\sqrt{N}}{J} + \frac{\sqrt{\log(N+J)}}{\sqrt{J}} + \frac{\sqrt{N\log(N+J)}}{J} & \text{if } N > J \\ \frac{1}{\sqrt{N}} + \frac{\sqrt{\log(N+J)}}{\sqrt{J}} + \frac{\sqrt{J\log(N+J)}}{N} & \text{if } N \leq J \end{cases}.$$

Therefore, $\frac{1}{\sqrt{JK}}\|\widehat{\boldsymbol{\Theta}}\mathbf{P}-\boldsymbol{\Theta}\|_F$ converges to zero in probability as $N, J \to \infty$ and $\frac{N}{J^2}, \frac{J}{N^2} \to 0$. □

References

Airoldi, E. M., Blei, D., Erosheva, E. A., & Fienberg, S. E. (2014). *Handbook of mixed membership models and their applications*. Boca Raton: CRC Press.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research, 9*, 1981–2014.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. Selected papers of Hirotugu Akaike (pp. 199–213).

Araújo, M. C. U., Saldanha, T. C. B., Galvao, R. K. H., Yoneyama, T., Chame, H. C., & Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems, 57*(2), 65–73.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., & Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis, 52*(1), 155–173.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Borsboom, D., Rhemtulla, M., Cramer, A. O., van der Maas, H. L., Scheffer, M., & Dolan, C. V. (2016). Kinds versus continua: A review of psychometric approaches to uncover the structure of psychiatric constructs. *Psychological Medicine, 46*(8), 1567–1579.

Chen, Y., Chi, Y., Fan, J., & Ma, C. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning, 14*(5), 566–806.

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

Chen, Y., Li, X., & Zhang, S. (2020). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association, 115*(532), 1756–1770.

Chen, Y., Ying, Z., & Zhang, H. (2021). Unfolding-model-based visualization: Theory, method and applications. *Journal of Machine Learning Research, 22*, 11.

Dobriban, E., & Owen, A. B. (2019). Deterministic parallel analysis: An improved method for selecting factors and principal components. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 81*(1), 163–183.

Donoho, D., & Stodden, V. (2003). When does non-negative matrix factorization give a correct decomposition into parts? *Advances in Neural Information Processing Systems, 16*.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. New York: Psychology Press.

Erosheva, E. A. (2002). *Grade of membership and latent structure models with application to disability survey data*. PhD thesis, Carnegie Mellon University.

Erosheva, E. A. (2005). Comparing latent structures of the grade of membership, Rasch, and latent class models. *Psychometrika, 70*(4), 619–628.

Erosheva, E. A., Fienberg, S. E., & Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics, 1*(2), 346.

Freyaldenhoven, S., Ke, S., Li, D., & Olea, J. L. M. (2023). *On the testability of the anchor words assumption in topic models*. Technical report, working paper, Cornell University.

Gillis, N., & Vavasis, S. A. (2013). Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(4), 698–714.

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*(2), 215–231.
Gormley, I. C., & Murphy, T. B. (2009). A grade of membership model for rank data. *Bayesian Analysis, 4*(2), 265–295.
Gu, Y., Erosheva, E. E., Xu, G., & Dunson, D. B. (2023). Dimension-grouped mixed membership models for multivariate categorical data. *Journal of Machine Learning Research, 24*(88), 1–49.
Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.
Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179–185.
Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research, 5*(9), 1457–1469.
Jin, J., Ke, Z. T., & Luo, S. (2023). Mixed membership estimation for social networks. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2022.12.003
Ke, Z. T., & Jin, J. (2023). Special invited paper: The score normalization, especially for heterogeneous network and text data. *Stat, 12*(1), e545.
Ke, Z. T., & Wang, M. (2022). Using SVD for topic modeling. *Journal of the American Statistical Association, 2022*, 1–16.
Klopp, O., Panov, M., Sigalla, S., & Tsybakov, A. (2023). Assigning topics to documents by successive projections. *Annals of Statistics* (to appear).
Koopmans, T. C., & Reiersol, O. (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics, 21*(2), 165–181.
Manrique-Vallier, D., & Reiter, J. P. (2012). Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association, 107*(500), 1385–1394.
Mao, X., Sarkar, P., & Chakrabarti, D. (2021). Estimating mixed memberships with sharp eigenvector deviations. *Journal of the American Statistical Association, 116*(536), 1928–1940.
Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society, 16*, 1–32.
Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics, 41*(3), 300–325.
Robitzsch, A., & Robitzsch, M. A. (2022). Packag 'sirt': Supplementary item response theory models.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.
Shang, Z., Erosheva, E. A., & Xu, G. (2021). Partial-mastery cognitive diagnosis models. *Annals of Applied Statistics, 15*(3), 1529–1555.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583–639.
Woodbury, M. A., Clive, J., & Garson, A., Jr. (1978). Mathematical typology: A grade of membership technique for obtaining disease definition. *Computers and Biomedical Research, 11*(3), 277–298.
Zhang, H., Chen, Y., & Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika, 85*, 358–372.

Manuscript Received: 4 MAY 2023
Accepted: 30 DEC 2023
Published Online Date: 15 FEB 2024

Downloaded from https://www.cambridge.org/core. 07 Jan 2025 at 13:38:35, subject to the Cambridge Core terms of use.