Psychometrika (2026), **91**, 695–723
doi:10.1017/psy.2026.10095

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Spectral Clustering with Likelihood Refinement for High-Dimensional Latent Class Recovery

Zhongyuan Lyu[1] and Yuqi Gu[2]

[1]Columbia University, USA; [2]Department of Statistics, Columbia University, USA

Corresponding author: Yuqi Gu; Email: yuqi.gu@columbia.edu

(Received 25 May 2025; revised 2 February 2026; accepted 3 February 2026)

Abstract

Latent class models (LCMs) are widely used for identifying unobserved subgroups from multivariate categorical data in social sciences, with binary data as a particularly popular example. However, accurately recovering individual latent class memberships remains challenging, especially when handling high-dimensional datasets with many items. This work proposes a novel two-stage algorithm for LCMs suited for high-dimensional binary responses. Our method first initializes latent class assignments by an easy-to-implement spectral clustering algorithm, and then refines these assignments with a one-step likelihood-based update. This approach combines the computational efficiency of spectral clustering with the improved statistical accuracy of likelihood-based estimation. We establish theoretical guarantees showing that this method is minimax-optimal for latent class recovery in the statistical decision theory sense. The method also leads to exact clustering of subjects with high probability under mild conditions. As a byproduct, we propose a computationally efficient consistent estimator for the number of latent classes. Extensive experiments on both simulated data and real data validate our theoretical results and demonstrate our method's superior performance over alternative methods.

Keywords: high-dimensional data; joint maximum likelihood estimation; latent class model; spectral clustering

1. Introduction

Latent class models (LCMs; Goodman, 1974; Lazarsfeld, 1968) are popular tools for uncovering unobserved subgroups from multivariate categorical responses, with broad applications in social and behavioral sciences (Berzofsky et al., 2014; Korpershoek et al., 2015; Wang & Hanges, 2011; Zeng et al., 2023). At the core of an LCM, we aim to identify K latent classes of N individuals based on their categorical responses to J items. In this article, we focus on the widely used LCMs with binary responses, which are prevailing in educational assessments (correct/wrong responses to test questions) and social science surveys (yes/no responses to survey items). Accurately and efficiently recovering latent classes in large N and large J settings remains a significant statistical and computational challenge.

Traditional approaches to LCM are typically based on maximum likelihood estimation. In the classical random-effect formulation, the latent class labels of subjects are treated as random variables, and the model is estimated by the marginal maximum likelihood (MML) approach, in which the latent labels are marginalized out. The expectation–maximization (EM) algorithm (Dempster et al., 1977) is often used for this purpose, but it becomes computationally expensive and lacks theoretical guarantees in the modern regime, where both N and J are large. In contrast, the joint maximum likelihood (JML)

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

approach adopts a fixed-effect perspective, treating the latent class labels as unknown parameters to be estimated. However, due to the inherent non-convexity of the joint likelihood function, JML is still highly sensitive to initialization and prone to be trapped in local optima. Recently, Zeng et al. (2023) showed that JML is consistent under the large N and large J regime. Nonetheless, they do not provide any algorithmic guarantee, and the convergence rate established therein is polynomial in J, whose optimality remains unclear.

In this article, we consider a different approach related to spectral clustering, which leverages the intrinsic approximate low-rank structure of the data matrix and is computationally efficient. In psychometrics and related fields, the common heuristic practice for spectral clustering is to first construct a similarity matrix among individuals or items, apply eigenvalue decomposition to it, and then perform K-means clustering of the eigenspace embeddings (Chen et al., 2017; Von Luxburg, 2007). More recently, Chen and Li (2019) directly perform the singular value decomposition (SVD) on a regularized version of the data matrix itself. Similar ideas have also been developed for nonlinear item factor analysis (Zhang et al., 2020) and the grade-of-membership model (Chen & Gu, 2024).

We next explain our idea of directly adapting the SVD to perform spectral clustering for binary-response LCMs. We collect the binary responses in an $N \times J$ matrix $\mathbf{R}$ with binary entries, then under the typical LCM assumption, the $\mathbf{R}$ can be written as $\mathbf{R} = \mathbb{E}[\mathbf{R}] + \mathbf{E}$, where $\mathbb{E}[\mathbf{R}]$ is the low-rank "signal" matrix due to the existence of K latent classes, and $\mathbf{E}$ is a mean-zero "noise" matrix. Denote by $\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top$ the truncated rank-K SVD of $\mathbf{R}$, then $\mathbf{R} \approx \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top$ approximately captures the "signal" part. In particular, we first extract $\widehat{\mathbf{U}}$, the leading K left singular vectors of the data matrix to perform dimensionality reduction. Then we apply K-means clustering algorithm on the N rows of $\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}$, the left singular vectors weighted by their corresponding singular values $\widehat{\mathbf{\Sigma}}$, to obtain the latent class label estimates. While variants of this simple yet efficient algorithm have attracted significant attention in network analysis (Zhang, 2024) and sub-Gaussian mixture models (Zhang & Zhou, 2024), it has not been widely explored or studied in the context of LCMs. The only exception is the recent work by Lyu et al. (2025), which extended spectral clustering to LCMs with individual-level degree heterogeneity and showed that spectral methods can lead to exact recovery of the latent class labels in certain regime. However, whether the spectral method gives an optimal way for estimating latent class labels in LCMs remains unknown. In this article, we resolve this question by proposing a novel two-stage method:

1. *First step:* Obtain initial latent class labels of individuals via the SVD-based spectral clustering, which leverages the approximate low-rankness of the data matrix.
2. *Second step:* Refine the latent class labels through a single step of likelihood maximization, which "sharpens" the coarse initial labels using the likelihood information.

This hybrid approach simultaneously takes advantage of the scalability of spectral methods and leverages the Bernoulli likelihood information to achieve a vanishing exponential error rate for clustering. We provide rigorous theoretical guarantees for our method. In summary, our method is both computationally fast and statistically optimal in the regime with a large number of items. Therefore, this work provides a useful tool to aid psychometric researchers and practitioners to perform latent class analysis of modern high-dimensional data. We emphasize that our method has no requirement on the magnitude of J compared to N, and even allows J to be much larger than N.

The rest of this article is organized as follows. After reviewing the setup of LCMs for binary responses in Section 2, we discuss the performance of spectral clustering and propose our main algorithm in Section 3. In Section 4, we establish theoretical guarantees and statistical optimality of our algorithm. We evaluate our method's performance in numerical experiments in Section 5. The proofs of the theoretical results are included in the Appendix.

Notation: We use bold capital letters, such as $\mathbf{A}, \mathbf{B}, \ldots$ to denote matrices and bold lowercase $\boldsymbol{a}, \boldsymbol{b}, \ldots$ to denote vectors. For any positive integer m, denote $[m] := \{1, \ldots, m\}$. For any matrix $\mathbf{A} \in \mathbb{R}^{N \times J}$ and

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

any $i \in [N]$ and $j \in [J]$, we use $A_{i,j}$ to denote its entry on the i-th row and j-th column, and use $\mathbf{A}_{i,:}$ (or $\mathbf{A}_{:,j}$) to denote its i-th row (or j-th column) vector. Let $\lambda_k(\mathbf{A})$ denote the k-th largest singular value of $\mathbf{A}$ for $k = 1, \ldots, \min\{N, J\}$. Let $\|\cdot\|$ denote the spectral norm (operator norm) for matrices and ℓ_2 norm for vectors, and $\|\cdot\|_{\mathrm{F}}$ denote the Frobenius norm for matrices. For two non-negative sequences $\{a_N\}$ and $\{b_N\}$, we write $a_N \lesssim b_N$ (or $a_N \gtrsim b_N$) if there exists some constant $C > 0$ independent of N such that $a_N \le C b_N$ (or $b_N \le C a_N$); $a_N \asymp b_N$ if $a_N \lesssim b_N$ and $b_N \lesssim a_N$ hold simultaneously; $a_N = o(b_N)$ if $b_N > 0$ and $a_N / b_N \to 0$.

2. Latent class model with binary responses

We consider a dataset of N individuals (e.g., survey or assessment respondents and voters) and J items (e.g., questions in a survey or an assessment and policy items). The observed data are represented as a binary response matrix $\mathbf{R} \in \{0,1\}^{N \times J}$, where $R_{i,j} = 1$ indicates a positive response from individual i to item j, and $R_{i,j} = 0$ otherwise. Each individual i belongs to one of K latent classes, and we encode the class membership by a latent vector $\boldsymbol{s}^\star = (s_1^\star, \ldots, s_N^\star)$, where $s_i^\star \in [K]$. The interpretation of the latent classes is characterized by an item parameter matrix $\boldsymbol{\Theta} = (\theta_{j,k}) \in [0,1]^{J \times K}$, where $\theta_{j,k}$ denotes the probability that an individual in class k gives a positive response to item j:

$$\mathbb{P}(R_{i,j} = 1 \mid s_i^\star = k) = \theta_{j,k}, \quad \forall i \in [N], j \in [J]. \tag{1}$$

Under the LCM, a subject's J responses are usually assumed to be conditionally independent given his or her latent class membership.

The traditional likelihood-based approach takes a random-effect perspective to LCMs and maximizes the marginal likelihood. This approach treats the latent class labels $\{s_i\}$ as random by assuming $p_k = \mathbb{P}(s_i = k)$ for each k and marginalizing s_i out:

$$L(\boldsymbol{\Theta} \mid \mathbf{R}) = \sum_{i \in [N]} \log \left(\sum_{k \in [K]} p_k \prod_{j \in [J]} \left(\theta_{j,k}^{R_{i,j}} \left(1 - \theta_{j,k}\right)^{1 - R_{i,j}} \right) \right). \tag{2}$$

A standard approach for maximizing the marginal likelihood is the EM algorithm (Dempster et al., 1977). However, EM is highly sensitive to initialization. Poor initial values often lead to convergence at local maxima, so practitioners resort to multiple random restarts at substantial computational cost. Moreover, when both J and K grow large, the computation for evaluating the objective in (2) becomes numerically unstable due to its log-sum-of-products structure.

Another strategy is to consider the fixed-effect LCM, where the JML naturally comes into play. In particular, the joint log-likelihood function of latent label vector $\boldsymbol{s}$ and item parameters $\boldsymbol{\Theta}$ can be written as

$$L(\boldsymbol{s}, \boldsymbol{\Theta} \mid \mathbf{R}) = \sum_{i \in [N]} \sum_{j \in [J]} \Big(R_{i,j} \log(\theta_{j,s_i}) + (1 - R_{i,j}) \log(1 - \theta_{j,s_i}) \Big). \tag{3}$$

The simple structure of (3) implies that the optimization of JML can be decoupled across $[N]$ and $[J]$. While JML is known to be inconsistent in low-dimensional settings (Neyman & Scott, 1948), the recent study by Zeng et al. (2023) has shown that JML can achieve consistent estimation of the latent labels in the high-dimensional regime. This property makes the JML approach appealing for large-scale modern datasets. It is worth noting that JML can be regarded as a version of classification EM (CEM; Celeux & Govaert, 1992), but with the difference of omitting the class proportion parameters. See Section 3.3 for more details.

Throughout the article, we consider the fixed-effect LCM and our goal is to recover the true latent labels $\boldsymbol{s}^\star$ from large-scale and high-dimensional data with $N, J \to \infty$. The performance of any clustering

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

method is evaluated via the Hamming error up to label permutations:

$$\ell\left(\boldsymbol{s},\boldsymbol{s}^{\star}\right) := \min_{\pi\in\mathfrak{S}_K} \frac{1}{N}\sum_{i\in[N]} \mathbb{I}\left(s_i \neq \pi\left(s_i^{\star}\right)\right), \tag{4}$$

where $\mathfrak{S}_K$ denotes the set of all permutation maps π from $[K] \to [K]$. This metric accounts for the inherent symmetry in clustering structure.

3. Two-stage clustering algorithm

3.1. Spectral clustering for the latent class model

Our primary goal is to estimate the latent label vector $\boldsymbol{s}^{\star}$. However, maximizing the joint likelihood (3) directly still suffers from computationally intractable for large N and J. We first consider spectral clustering, which is a computationally efficient method that leverages the approximate low-rank structure of the data matrix. Let $\mathbf{Z} = (Z_{i,j}) \in \{0,1\}^{N\times K}$ be a latent class membership matrix with $Z_{i,k} = 1$ if $s_i^{\star} = k$ and $Z_{i,l} = 0$ otherwise. Under the fixed-effect LCM, we have $R_{i,j} \sim \text{Bern}\left(\sum_{k=1}^{K} Z_{i,k}\theta_{j,k}\right)$ and we can write $R_{i,j} = \sum_{k=1}^{K} Z_{i,k}\theta_{j,k} + E_{i,j}$, where $\sum_{k=1}^{K} Z_{i,k}\theta_{j,k} = \mathbb{E}[R_{i,j}]$ and $E_{i,j}$ is mean-zero random noise. We have the following crucial decomposition of the data matrix $\mathbf{R}$ into two additive components using matrix notation:

$$\mathbf{R} = \underbrace{\mathbf{Z}}_{N\times K}\ \underbrace{\boldsymbol{\Theta}^{\top}}_{K\times J} + \underbrace{\mathbf{E}}_{\text{noise}}. \tag{5}$$

Here, $\mathbf{Z}\boldsymbol{\Theta}^{\top} = \mathbb{E}[\mathbf{R}]$ is the "signal" part containing the crucial latent class information (and thus is low-rank as $K \ll \min\{N,J\}$), and $\mathbf{E} = (E_{i,j}) \in \mathbb{R}^{N\times J}$ is a mean-zero "noise" matrix with $\mathbb{E}[\mathbf{E}] = \mathbb{E}[\mathbf{R} - \mathbb{E}[\mathbf{R}]] = \mathbf{0}_{N\times J}$.

Given the above observation (5), we first briefly outline the rationale of spectral clustering under the oracle noiseless setting (i.e., when $\mathbf{E} = \mathbf{0}$). Let $\mathbf{Z}\boldsymbol{\Theta}^{\top} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\top}$ be the rank-K SVD of $\mathbf{Z}\boldsymbol{\Theta}^{\top}$. Notably, the ith row vector in $\mathbf{U}\boldsymbol{\Sigma}$ can be written as

$$\boldsymbol{e}_i^{\top}\mathbf{U}\boldsymbol{\Sigma} = \boldsymbol{e}_i^{\top}\mathbf{Z}\boldsymbol{\Theta}^{\top}\mathbf{V} = \sum_{k=1}^{K} Z_{i,k}\boldsymbol{\Theta}_{:,k}^{\top}\mathbf{V} = \boldsymbol{\Theta}_{:,s_i}^{\top}\mathbf{V}, \quad \forall i \in [N]. \tag{6}$$

Thus, data points in the same latent class have identical row vectors in $\mathbf{U}\boldsymbol{\Sigma}$, namely, one of the K distinct vectors $\left\{\boldsymbol{\Theta}_{:,k}^{\top}\mathbf{V}\right\}_{k=1}^{K}$. Hence, the true latent class label information can be readily read off from rows of $\mathbf{U}\boldsymbol{\Sigma}$. In the realistic noisy data setting, denote by $\widehat{\mathbf{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{V}}^{\top}$ the top-K SVD of the data matrix $\mathbf{R}$. Then intuitively, $\mathbf{U}\boldsymbol{\Sigma}$ would be well-approximated by $\widehat{\mathbf{U}}\widehat{\boldsymbol{\Sigma}}$ provided that $\mathbf{E}$ is relatively small in magnitude, and the row vectors of $\widehat{\mathbf{U}}\widehat{\boldsymbol{\Sigma}}$ are noisy point clouds distributed around the K points. Therefore, K-means clustering can deliver reasonable estimates of class labels. We illustrate this point using an numerical example in Figure 1, where the row vectors of $\mathbf{U}\boldsymbol{\Sigma}$ (left) and $\widehat{\mathbf{U}}\widehat{\boldsymbol{\Sigma}}$ (right) are plotted as points in $\mathbb{R}^2$, respectively.

The above high-level idea can be formalized as the spectral clustering algorithm (Algorithm 1), investigated in Zhang and Zhou (2024) for sub-Gaussian mixture models. Algorithm 1 first obtains the rank-K SVD of $\mathbf{R}$ denoted by $\widehat{\mathbf{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{V}}^{\top}$, and then performs K-means clustering on the rows of $\widehat{\mathbf{U}}\widehat{\boldsymbol{\Sigma}}$, the left singular vectors weighted by the singular values.

Remark 1. An equivalent formulation of (3) aggregates the 2^J unique response patterns (weighted by their frequencies). We do not adopt this perspective because it becomes less suitable, both computationally and theoretically, for high-dimensional data with very large J. On the computational side, our approach does not rely on evaluating the likelihood through these unique response patterns. On the theoretical side, our spectral method is motivated by the approximate low-rank structure of the $N \times J$ binary response matrix $\mathbf{R} \in \mathbb{R}^{N\times J}$ itself, which does not utilize the unique response patterns.

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

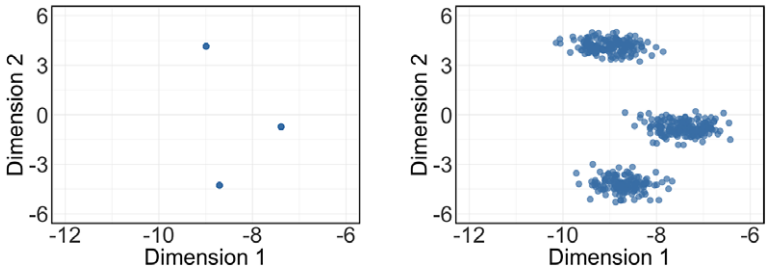


Figure 1. An illustration for spectral clustering: row vectors of $\mathbf{U}\Sigma$ (left) and $\widehat{\mathbf{U}}\widehat{\Sigma}$ (right). Setting: $N = 500$, $J = 250$, and $K = 3$.

Algorithm 1: Spectral clustering (Spec).

Data: Data matrix $\mathbf{R} \in \mathbb{R}^{N\times J}$. Number of latent classes K

Result: Estimated latent class labels $\widetilde{\boldsymbol{s}} \in [K]^N$

1 Perform rank-K SVD on $\mathbf{R}$ to obtain $\sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\boldsymbol{u}}_i \widehat{\boldsymbol{v}}_i^{\top}$, where $\widehat{\lambda_1} \geq \widehat{\lambda_2} \geq \cdots \geq \widehat{\lambda_K} \geq 0$, $\widehat{\mathbf{U}} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K) \in \mathbb{R}^{N\times K}$ and $\widehat{\mathbf{V}} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K) \in \mathbb{R}^{J\times K}$.

2 Perform K-means on the rows of $\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}$ (or equivalently, $\mathbf{R}\widehat{\mathbf{V}}$), i.e.,

$$\left(\widetilde{\boldsymbol{s}}, \{\widetilde{\boldsymbol{c}}_k\}_{k\in[K]}\right) = \underset{\boldsymbol{s}\in[K]^N, \{\boldsymbol{c}_k\}_{k\in[K]}\in\mathbb{R}^K}{\arg\min} \sum_{i\in[N]} \left\|\widehat{\mathbf{U}}_{i,:}\widehat{\mathbf{\Sigma}} - \mathbf{c}_{s_i}\right\|^2.$$

3.2. One-step likelihood refinement with spectral initialization

Despite computationally efficient, spectral clustering itself ignores the likelihood information in the LCM that might improve statistical accuracy. This motivates our proposed Algorithm 2 named spectral clustering with one-step likelihood refinement algorithm, or SOLA for short.

In particular, the first stage is to perform spectral clustering to obtain an initial class label estimate $\widetilde{\boldsymbol{s}}$. This stage utilizes the inherent approximate low-rank structure of data but does not use the likelihood information. In the second stage, recall the joint likelihood function defined in (3) as

$$L(\boldsymbol{s}, \mathbf{\Theta} \mid \mathbf{R}) = \sum_{i\in[N]} \sum_{j\in[J]} \Big(R_{i,j} \log(\theta_{j,s_i}) + (1 - R_{i,j}) \log(1 - \theta_{j,s_i}) \Big).$$

We then perform the following two steps:

1. (Maximization) For fixed $\widetilde{\boldsymbol{s}}$, find the $\mathbf{\Theta}$ maximizing the objective $L(\widetilde{\boldsymbol{s}}, \mathbf{\Theta} \mid \mathbf{R})$, which turns out to have a explicit solution as

$$\widehat{\theta}_{j,k} = \frac{\sum_{i\in[N]} R_{i,j} \mathbb{I}(\widetilde{s}_i = k)}{\sum_{i\in[N]} \mathbb{I}(\widetilde{s}_i = k)}, \quad \forall j \in [J],\ k \in [K].$$

   In other words, this is the sample average of the j-th column of $\mathbf{R}$ based on estimated class labels $\{i \in [N] : \widetilde{s}_i = k\}$.
2. (Assignment) Given $\widehat{\mathbf{\Theta}}$, the objective function $L(\boldsymbol{s}, \widehat{\mathbf{\Theta}} \mid \mathbf{R})$ can be decoupled into N terms, each corresponding to an individual data point, and it is equivalent to assign label $k \in [K]$ to sample i which maximize the log-likelihood independently for each $i \in [N]$:

$$\widehat{s}_i = \underset{k\in[K]}{\arg\max} \sum_{j\in[J]} \left[R_{i,j} \log\left(\widehat{\theta}_{j,k}\right) + (1 - R_{i,j}) \log\left(1 - \widehat{\theta}_{j,k}\right) \right], \quad i \in [N].$$

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

It is readily seen that these two steps increase the joint likelihood monotonically. However, convergence to true $\boldsymbol{s}^\star$ depends on the performance of the initial estimator $\widetilde{\boldsymbol{s}}$ in the first stage, which we obtained using spectral clustering. The above details are summarized in Algorithm 2, and we provide its algorithmic guarantee in Section 4.

Algorithm 2: SOLA.

Data: Matrix $\mathbf{R} \in \mathbb{R}^{N\times J}$, rank K
Result: Estimated labels $\widehat{\boldsymbol{s}} \in [K]^N$
1 Initial label estimates: $\widetilde{\boldsymbol{s}} = \mathsf{Spec}(\mathbf{R})$.
2 (Maximization) Initial center estimates :

$$\widehat{\theta}_{j,k} = \frac{\sum_{i\in[N]} R_{i,j}\mathbb{I}(\widetilde{s}_i = k)}{\sum_{i\in[N]} \mathbb{I}(\widetilde{s}_i = k)}, \quad j \in [J],\ k \in [K]. \tag{7}$$

3 (Assignment) One-step likelihood-based refinement:

$$\widehat{s}_i = \arg\max_{k\in[K]} \sum_{j\in[J]} \left[R_{i,j}\log\left(\widehat{\theta}_{j,k}\right) + (1-R_{i,j})\log\left(1-\widehat{\theta}_{j,k}\right)\right], \quad i \in [N]. \tag{8}$$

Interestingly, our two-stage procedure resonates with the spirit of the "initialize-and-refine" scheme that has been used in network analysis for stochastic block models (SBMs) (Chen et al., 2022; Gao et al., 2017; Lyu et al., 2023), while the model considered here differs fundamentally in that LCMs handle high-dimensional response data without relying on pairwise interaction structures inherent to SBMs.

In practice, we observe that performing multiple likelihood-based refinement steps may lead to improved convergence and performance compared to the one-step refinement procedure. Thus, we also consider a multiple-step refinement variant of Algorithm 2, which we denote by SOLA+, in the later simulation studies. We set the number of "Maximization + Assignment" refinement steps for SOLA+ to be 10 in all our simulations. In principle, one could also adapt the algorithm by monitoring the change in the joint log-likelihood values and stopping the algorithm once the absolute difference between successive iterations falls below a small threshold (Di Mari et al., 2023).

3.3. Extension: Classification EM refinement

Our proposed framework can be readily extended to one-step CEM refinement, where CEM is another widely used approach to estimate LCMs (Celeux & Govaert, 1992; Zeng et al., 2023). It can be regarded as a variant of the standard EM algorithm designed specifically for clustering. Unlike traditional EM, which calculates "soft" assignments by estimating the probability that each observation belongs to every latent class, CEM uses "hard" assignments: each observation is directly assigned to the class with the highest posterior probability. By explicitly modeling the mixture proportion parameters for the latent classes, CEM adjusts the likelihood to account for varying cluster sizes. Concretely, in our setting, one can adapt the one-step refinement in Algorithm 2 to a CEM step by including class proportions. Besides updating $\widehat{\theta}_{j,k}$ as in (7), we also estimate the proportion parameters $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \ldots, \widehat{p}_K) \in [0,1]^K$ by

$$\widehat{p}_k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{\widetilde{s}_i = k\}, \quad k \in [K]. \tag{9}$$

Then, for each individual i, we update their latent class label based on both the item parameters and the estimated class proportions:

$$\widehat{s}_i = \arg\max_{k\in[K]}\left[\log\widehat{p}_k + \sum_{j=1}^{J}\left(R_{i,j}\log\widehat{\theta}_{j,k} + (1-R_{i,j})\log\left(1-\widehat{\theta}_{j,k}\right)\right)\right]. \tag{10}$$

This one-step CEM update replaces the likelihood-refinement stage of our two-stage estimator: start with spectral clustering, then perform the CEM step above. We provide its theoretical guarantee

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

in Section 4. In practice, one could easily generalize the above CEM procedure to a multiple-step refinement variant, which is analogous to SOLA+.

4. Theoretical guarantees

To deliver the theoretical results, we start by making the following mild model assumptions.

Assumption 1. $\theta_{j,k} \in [c_\theta, C_\theta]$ for some universal constants $0 < c_\theta < C_\theta < 1$, and $j \in [J]$, $k \in [K]$.

Assumption 2. $\min_{k\in[K]} \left|\left\{i \in [N] : s_i^\star = k\right\}\right| \geq \alpha N/K$ for some universal constant $\alpha \in (0,1]$.

Assumption 1 is a standard assumption for considering a compact parameter space for $\boldsymbol{\Theta}$. Assumption 2 is imposed for both practical consideration and technical reason. Without this, rare classes could be statistically indistinguishable and computationally unstable to estimate.

4.1. Theoretical guarantee of spectral clustering

In this section, we introduce the theoretical guarantee of spectral clustering (Algorithm 1). To that end, we define some necessary notations. The difficulty of recovering $\mathbf{s}^\star$ depends on how distinct the latent classes are, which, in the context of spectral clustering, is characterized by the following quantity:

$$\Delta := \min_{k_1 \neq k_2 \in [K]} \left\| \boldsymbol{\Theta}_{:,k_1} - \boldsymbol{\Theta}_{:,k_2} \right\| = \min_{k_1 \neq k_2 \in [K]} \left[\sum_{j=1}^{J} \left(\theta_{j,k_1} - \theta_{j,k_2}\right)^2 \right]^{1/2}. \tag{11}$$

This quantity measures the minimal Euclidean distance between any two class-specific item parameter vectors. Intuitively, a larger Δ yields a larger gap between the columns of $\boldsymbol{\Theta}$ and a bigger distinction among the latent classes, and therefore produces clearer separation in the spectral embeddings (referring to (6)), making recovery of the true class labels easier. In addition, let $\lambda_1 \geq \cdots \geq \lambda_K > 0$ denote the singular values of $\mathbb{E}[\mathbf{R}] = \mathbf{Z}\boldsymbol{\Theta}^\top$ in decreasing order. The following result characterizes the clustering error of Algorithm 1.

Proposition 1 (Adapted from Theorem 3.1 in Zhang & Zhou, 2024). *Define* $\sigma_\theta^2 := (1-2\theta)/\left[2\log\left((1-\theta)\theta^{-1}\right)\right]$ *for* $\theta \in (0,1)\backslash\{1/2\}$ *and* $\sigma_\theta^2 := 1/4$ *for* $\theta = 1/2$, *and let* $\bar{\sigma} := \max_{j,k} \sigma_{\theta_{j,k}}$. *Assume* $\mathrm{rank}(\boldsymbol{\Theta}) = K$, $\alpha N/K^2 \geq 10$ *and*

$$\min\left\{ \frac{\Delta}{\alpha^{-1/2} K \left(1 + \sqrt{\frac{J}{N}}\right)\bar{\sigma}}, \frac{\lambda_K}{\left(\sqrt{N} + \sqrt{J}\right)\bar{\sigma}} \right\} \to \infty. \tag{12}$$

Then, for the spectral clustering estimator $\widetilde{\boldsymbol{s}}$ *in Algorithm 1, we have*

$$\mathbb{E}\ell\left(\widetilde{\boldsymbol{s}}, \boldsymbol{s}^\star\right) \leq \exp\left(-\frac{\Delta^2}{8\bar{\sigma}^2}\left(1 - o(1)\right)\right) + \exp\left(-\frac{N}{2}\right), \tag{13}$$

where $\ell(\cdot,\cdot)$ *is defined in (4). Moreover, if further assume that* $\Delta^2 \geq 8(1+c)\bar{\sigma}^2 \log N$ *for any constant* $c > 0$, *then we have* $\ell\left(\widetilde{\boldsymbol{s}}, \boldsymbol{s}^\star\right) = 0$ *with probability* $1 - \exp(-N/2) - \exp(-\Delta/\bar{\sigma})$.

Proposition 1 delivers that the simple and computationally efficient spectral clustering can essentially lead to exponentially small error rate with respect to the latent class separation Δ. Notably, we can have exact recovery of all latent class labels (i.e., $\ell\left(\widetilde{\boldsymbol{s}}, \boldsymbol{s}^\star\right)$ with $\widetilde{s}_i = \pi(s_i)$ for all $i = 1, \ldots, N$) with high probability as long as $\Delta^2/\bar{\sigma}^2 \geq 8(1+c)\log N$. In the LCM we study, we have $\bar{\sigma} \leq 1/4$ by definition, and hence it suffices to have $\Delta^2 \geq 2(1+c)\log N$ to achieve exact recovery. This is not hard to achieve even for slowly growing number of items, because if we have constant separation of item parameters for each item with $\delta = \min_{k_1 \neq k_2 \in [K], j \in [J]} |\theta_{j,k_1} - \theta_{j,k_2}|$, then $\Delta^2 = J\delta^2$. In this case, having $J > 2(1+c)\delta^{-2}\log N$ for exact

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

recovery corresponds to a moderately growing dimension. In sharp contrast, the JML estimator analyzed in Zeng et al. (2023) only achieves a polynomial convergence rate of order $O(J^{-1/2})$. Moreover, they cannot achieve exact recovery according to the technical condition and conclusion therein. Therefore, the new spectral estimator $\widetilde{\boldsymbol{s}}$ represents a substantial improvement with exponential error decay in (13).

We also note that the condition (12) on Δ and λ_K is mild and can be easily satisfied in the modern high-dimensional data setting. We next give a concrete example. Let us consider a common generative setting for item parameters $\boldsymbol{\Theta}$, where its entries are i.i.d. generated from the Beta distribution.

Proposition 2. *Consider* $\theta_{j,k} \overset{i.i.d.}{\sim} Beta(a,b)$ *for all* $j \in [J]$ *and* $k \in [K]$ *with* $a,b > 0$. *Define*

$$B := \frac{a\left(a+b+a\left(a+b\right)^{-1}\right)}{\left(a+b\right)\left(a+b+1\right)},$$

then, under Assumption 2, we have $\lambda_K \gtrsim \sqrt{\alpha BNJ}$ *and* $\Delta \gtrsim \sqrt{BJK}$ *with probability at least* $1 - K(e^{-c_1 B^2 J} + e^{-c_2 BJ})$ *for some universal constants* $c_1, c_2 \in (0,1)$.

Proposition 2 entails that the condition on Δ and λ_K for the success of spectral clustering in Proposition 1 can be easily satisfied when $\boldsymbol{\Theta}$ consists of Beta distributed entries, a common assumption in Bayesian inference for the LCM. Specifically, combined with Assumption 1, a sufficient condition on B for spectral clustering to have the desirable error control (13) would be $B \gg K(\frac{1}{N} + \frac{1}{J})$, which is easily satisfied if the Beta parameters a and b are of the constant order.

Despite the above nice theoretical guarantee for spectral clustering, the clustering error's reliance on the Euclidean metric Δ in Proposition 1 is inherent from the general framework analyzed in Zhang and Zhou (2024), and it turns out to be not statistically optimal due to ignorance of the likelihood information. We will discuss and address this issue next.

4.2. Fundamental statistical limit for LCMs

In this section, we provide the understanding of the information-theoretic limit (fundamental statistical limit) of clustering in LCMs. To compare different statistical procedures, we consider the minimax framework that had been widely considered in both statistical decision theory and information theory in the past decades (Le Cam, 2012; Takezawa, 2005). The minimax risk is defined as the infimum, over all possible estimators, of the maximum loss (here, the mis-clustering error) taken over the entire parameter space. In other words, it is the rate of the best estimator one can have among all statistical procedures in the worst-case scenario.

Formally, we establish the minimax lower bound for mis-clustering error under the fixed-effect LCM in the following result, whose proof is adapted from Lyu et al. (2025) and can be found in Section B.1.

Theorem 1 (Minimax lower bound). *Consider the following parameter space for* $\mathbb{E}[\boldsymbol{R}]$:

$$\mathcal{P}_K(\boldsymbol{s},\boldsymbol{\Theta}) := \left\{\bar{\boldsymbol{R}} : \bar{R}_{i,j} = \theta_{j,s_i}, \boldsymbol{s} \in [K]^N, \theta_{j,k} \in [0,1], \forall i \in [N], j \in [J], k \in [K]\right\}.$$

Define $I^\star := \min_{k_1 \neq k_2 \in [K]} \sum_{j \in [J]} I(\theta_{j,k_1}, \theta_{j,k_2})$ *with*

$$I(p,q) := -2\log\left(\sqrt{pq} + \sqrt{(1-p)(1-q)}\right), \quad \forall p,q \in (0,1),$$

then we have

$$\inf_{\widehat{\boldsymbol{s}}} \sup_{\mathcal{P}_K(\boldsymbol{s},\boldsymbol{\Theta})} \mathbb{E}\ell\left(\widehat{\boldsymbol{s}},\boldsymbol{s}^\star\right) \geq \exp\left(-\frac{I^\star}{2}\left(1+o(1)\right)\right), \tag{14}$$

where the infimum is taken over all possible estimators of the latent class labels.

Roughly speaking, Theorem 1 informs us that under the fixed-effect LCM, the error rate $\exp(-I^\star/2)$ is the lowest possible clustering error rate that no estimator can uniformly surpass. Notably, $I(p,q)$ is

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

exactly the Renyi divergence (Rényi, 1961) of order 1/2 between two Bernoulli distributions, $\text{Bern}(p)$ and $\text{Bern}(q)$, and hence $\sum_{j\in[J]} I(\theta_{j,k_1},\theta_{j,k_2})$ can be interpreted as the Renyi divergence between two Bernoulli random vectors with parameters $\boldsymbol{\Theta}_{:,k_1}$ and $\boldsymbol{\Theta}_{:,k_2}$, respectively. $I^\star$ can be regarded as the overall signal-to-noise ratio (SNR) which further takes account into all possible pairs of $(k_1,k_2)\in[K]^2$ with $k_1\neq k_2$. Although the form of $I^\star$ might seem complicated at its first glance, it actually exactly reflects the likelihood information in the LCMs. By definition of the minimax lower bound, we know that $\exp(-I^\star/2)$ cannot be larger than the spectral clustering error rate $\exp(-\Delta^2/(8\bar{\sigma}^2))$ obtained in Proposition 1. We will discuss this in more detail in Section 4.4.

In the following, we take a closer look at why the quantity $I^\star$ shows up as the fundamental statistical limit. We analyze an idealized scenario where the true item parameters $\boldsymbol{\Theta}$ are known, which we term as the *oracle* setting. For simplicity, we consider $K=2$ latent classes and $I^\star=\sum_{j\in[J]} I(\theta_{j,1},\theta_{j,2})$. In this regime, we can relate the clustering problem for the i-th sample to the following binary hypothesis testing problem:

$$H_0: s_i^\star = 1, \quad \text{v.s.} \quad H_0: s_i^\star = 2.$$

By the Neyman–Pearson lemma, the test that gives the optimal Type-I plus Type-II error is the likelihood ratio test. Consequently, the oracle classifier assigns an observation to the class that maximizes the log-likelihood:

$$\bar{s}_i = \arg\max_{k\in[2]} \sum_{j\in[J]} \Big[R_{i,j}\log\theta_{j,k} + (1-R_{i,j})\log\big(1-\theta_{j,k}\big)\Big]. \tag{15}$$

Assuming without loss of generality that $s_i^\star=1$, one can show that the mis-clustering probability satisfies

$$\mathbb{P}\big(\bar{s}_i\neq s_i^\star\big) = \mathbb{P}\left(\sum_{j\in[J]} R_{i,j}\log\frac{\theta_{j,2}(1-\theta_{j,1})}{\theta_{j,1}(1-\theta_{j,2})} \geq \sum_{j\in[J]}\log\frac{1-\theta_{j,1}}{1-\theta_{j,2}}\right) \leq \exp\left(-\frac{\sum_{j\in[J]} I(\theta_{j,1},\theta_{j,2})}{2}\right).$$

In other words, the above error rate in the oracle setting directly informs the information-theoretic limit of any clustering procedure in the LCM with binary responses, as indicated by Theorem 1. Note that the oracle estimator (15) that achieves this optimal rate is constructed using the true $\boldsymbol{\Theta}$ and cannot be applied in practice. Later, we will construct a computationally feasible estimator (i.e., our SOLA estimator) that achieves this optimal rate without relying on prior knowledge of $\boldsymbol{\Theta}$.

4.3. Theoretical guarantee of SOLA

In this section, we give the theoretical guarantee of SOLA. To facilitate the theoretical analysis, we consider a sample-splitting variant of SOLA (Algorithm 3), where the sample data points are randomly partitioned into two subsets $\mathcal{S}_1,\mathcal{S}_2$ of size $\lceil N/2\rceil$ and $N-\lceil N/2\rceil$. The key idea is to estimate $\boldsymbol{\Theta}$ using one subset of samples, and then update labels for the remaining ones. The procedure can be repeated by switching the roles of the two subsets, and we can obtain labels for all samples. Finally, an alignment step is needed due to the permutation ambiguity of clustering in the two subsets. We have the following result, whose proof is deferred to Section B.2.

Theorem 2. *Suppose Assumptions 1 and 2 and the assumptions of Proposition 1 hold. In addition, assume that* $N\gtrsim K\log N$, $JK\lesssim\min\{\exp(-c\Delta^2/\bar{\sigma}^2), N^{\gamma_0}\}$, *for some universal constants* $0<c<1/8$ *and* $\gamma_0>0$. *If the SNR condition satisfies*

$$\frac{\Delta/\bar{\sigma}}{\sqrt{J}K^{3/4}(\log N)^{1/4}/N^{1/4}}\to\infty,$$

then, for $\widehat{\boldsymbol{s}}$ *from Algorithm 3, we have*

$$\mathbb{E}\ell\big(\widehat{\boldsymbol{s}},\boldsymbol{s}^\star\big) \leq \exp\left(-\frac{I^\star}{2}(1-o(1))\right).$$

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

Algorithm 3: SOLA with sample splitting.

Data: Data matrix $\mathbf{R} \in \{0,1\}^{N\times J}$, index sets $\mathcal{S}_1$ and $\mathcal{S}_2$, number of clusters K
Result: estimated labels $\widehat{\boldsymbol{s}} \in [K]^N$

1 Construct $\mathbf{R}_{\mathcal{S}_1}, \mathbf{R}_{\mathcal{S}_2} \in \{0,1\}^{N/2\times J}$, which are sub-matrices of $\mathbf{R}$ according to $\mathcal{S}_1, \mathcal{S}_2$.

2 Initial label estimates: $\widetilde{\mathbf{s}}^{(m)} = \mathsf{Spec}\left(\mathbf{R}_{\mathcal{S}_m}\right) \in \{0,1\}^{N/2}$ for $m = 1,2$.

3 Initial center estimates: for define

$$\widetilde{\theta}_{j,k}^{(m)} = \frac{\sum_{i\in\mathcal{S}_m} R_{i,j}\mathbb{I}\left(\widetilde{s}_i^{(m)} = k\right)}{\sum_{i\in\mathcal{S}_m} \mathbb{I}\left(\widetilde{s}_i^{(m)} = k\right)}, \quad \forall j \in [J],\ k \in [K],\ m = 1,2.$$

4 One-step likelihood-based refinement:

$$\widehat{s}_i^{(m)} = \arg\max_{k\in[K]} \sum_{j\in[J]} \left[R_{i,j}\log\left(\widetilde{\theta}_{j,k}^{(m)}\right) + (1-R_{i,j})\log\left(1-\widetilde{\theta}_{j,k}^{(m)}\right)\right], \quad \forall i \in \mathcal{S}_l, \quad l \neq m \in \{1,2\}. \tag{16}$$

5 Label alignment: Set $\widehat{s}_i = \widehat{s}_i^{(1)}$ for $\forall i \in \mathcal{S}_1$. Let $\widehat{\pi} := \arg\min_{\pi\in\mathfrak{S}_K} \left\|\widehat{\boldsymbol{\Theta}}^{(1)} - \widehat{\boldsymbol{\Theta}}^{(2)}\mathbf{G}_\pi\right\|_{\mathrm{F}}$ and set $\widehat{s}_i = \widehat{\pi}\left(\widehat{s}_i^{(2)}\right)$ for $\forall i \in \mathcal{S}_2$, where $\mathfrak{S}_K$ is the set of all permutations of $[K]$ and $\mathbf{G}_\pi$ is the column permutation matrix corresponding to π.

Combining Theorems 1 and 2, we can see that the estimator of Algorithm 3 achieves the fundamental statistical limit and is optimal for estimating the latent class labels. We have several remarks on Theorem 2. First, it is interesting to consider the regime when $J = o(\sqrt{\frac{N\log N}{K}})$, since otherwise the SNR condition would already implies exact recovery (perfect estimation of latent class labels with high probability) for spectral clustering, as shown by Theorem 2 and Proposition 1. Second, the sample splitting step is crucial in the current proof of Theorem 2 to decouple the dependence between $R_{i,j}$ and $\widehat{\boldsymbol{\Theta}}$ in (8). This step serves purely as a theoretical device to facilitate the theoretical analysis of clustering error rate, as commonly adopted in theoretical statistics (Abbe, 2018; Lei & Zhu, 2017; Rinaldo et al., 2019; Vu, 2018). In practice, halving the sample only increases the estimation error of $\widehat{\boldsymbol{\Theta}}$ by a constant factor, which is negligible in theory and immaterial in practice when the sample size N is large. Alternatively, it may be possible in the future to show that the same theoretical guarantee holds for the original version of SOLA, that is, Algorithm 2, by developing a leave-one-out analysis similar to that in Zhang and Zhou (2024). However, the analysis will be highly technical and much more involved than that in Zhang and Zhou (2024) and beyond the scope of our current focus. Last, although sample splitting simplifies theoretical analysis, it may not be necessary in practice; we use the original SOLA (Algorithm 2) for all numerical experiments and observe satisfactory empirical results.

Remark 2. Our work is related to the Tensor-EM method introduced by Zeng et al. (2023), which also follows an initialization-and-refinement strategy. In the high level, both our methods require moment-based methods in the initialization. However, there are two important distinctions. Zeng et al. (2023) employ a tensor decomposition that leverages higher-order moments and can be computationally intensive and unstable. In contrast, our spectral initialization exploits only the second moment of data (since SVD of the data matrix is equivalent to eigendecomposition of the Gram matrix), making it both simple and efficient. Second, the error rate established in Zeng et al. (2023) is of polynomial order, that is, $O\left(J^{-1/2}\right)$, whereas our SOLA achieves the statistically optimal exponential error rate with respect to $I^\star$.

As an extension of Theorem 2, we have a similar theoretical guarantee for the one-step CEM refinement proposed in Section 3.3, whose proof is given in Section B.4.

Corollary 1. *Suppose the conditions in Corollary 2 hold. Consider the output of Algorithm 3 with (16) replaced by (9) and sample splitting version of (10), we have*

$$\mathbb{E}\ell\left(\widehat{\boldsymbol{s}}, \boldsymbol{s}^\star\right) \leq \exp\left(-\frac{I^\star}{2}\left(1-o(1)\right)\right).$$

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

Theorem 1 indicates that one-step CEM refinement achieves the same exponential error rate $\exp(-I^*/2)$ as Algorithm 3. There might be a gap between these two methods when Assumption 2 does not hold, which is an interesting future direction to consider.

4.4. Comparison of spectral clustering and SOLA

In this section, we compare the convergence rate of spectral clustering and SOLA under LCM. While spectral clustering is widely employed due to its computational efficiency and solid theoretical guarantees in general clustering settings, we demonstrate that SOLA attains minimax-optimal clustering error rates in LCMs that spectral clustering fails to achieve.

In view of Theorem 1, the minimax-optimal clustering error rate for LCM is of order $\exp(-I^\star/2)$. On the other hand, spectral clustering is shown to achieve the error rate scaling as $\exp(-\Delta^2/(8\bar{\sigma}^2))$ in Proposition 1. To understand the difference between these two quantities, we have the following result, whose proof can be found in Section B.5.

Proposition 3. *Assume that (a)* $\min_{k_1\neq k_2\in[K]}|\Omega_0(k_1,k_2)|\gtrsim J$, *where* $\Omega_0(k_1,k_2):=\{j\in[J]:|\theta_{j,k_1}-\theta_{j,k_2}|=o(1)\}$ *and (b)* $\min_{k\in[K]}\big|\{j\in[J]:|\theta_{j,k}-1/2|>c_0\}\big|\gtrsim J$ *for some universal constant* $c_0>0$. *Then we have*

$$\exp\left(-\frac{I^\star}{2}\right)\lesssim\exp\left(-\frac{\Delta^2}{8\bar{\sigma}^2}(1+c)\right),$$

for some universal constant $c>0$.

A few comments are in order regarding Proposition 3. First, under the conditions (a) and (b), the rate of spectral clustering becomes sub-optimal (in order) due to a loose constant in the exponent as the latent class separation $\Delta\to\infty$. Second, it is important to note that the mis-clustering error $\ell(\widehat{\boldsymbol{s}},\boldsymbol{s}^*)$ takes values in the finite set $\{0,1/N,\ldots,1\}$ according to the definition (4). Given this, it is only meaningful to compare error rates when the exponent $\Delta^2/\bar{\sigma}^2$ (or $I^\star$) does not exceed a constant factor of $\log N$, since otherwise exact clustering of all subjects is achieved with high probability. To see this, we use Markov inequality to get

$$\mathbb{P}(\ell(\widehat{\boldsymbol{s}},\boldsymbol{s}^*)\neq 0)=\mathbb{P}(\ell(\widehat{\boldsymbol{s}},\boldsymbol{s}^*)\geq 1/N)\leq N\mathbb{E}\ell(\widehat{\boldsymbol{s}},\boldsymbol{s}^*)\leq N\exp(-C\log N)\to 0,$$

for a constant $C>1$. In view of this, the condition (a) helps us exclude the trivial case where $I^\star\asymp\Delta^2\asymp J$, which would only allow $J\lesssim\log N$. Third, the essential order difference occurs when most of $\theta_{j,k}$'s are bounded away from $1/2$. To see this, consider the scenario when $K=2$, $\theta_{j,1}=0.5$, and $\theta_{j,2}=0.5-\delta$ for some $\delta=o(1)$. Then we have $\Delta=\frac{1}{2}J\delta^2$ and $I^\star=-J\log\left(\sqrt{\frac{1}{2}\left(\frac{1}{2}-\delta\right)}+\sqrt{\frac{1}{2}\left(\frac{1}{2}+\delta\right)}\right)=\frac{1}{2}J\delta^2(1+o(1))$. This implies that without condition (b), the exponents Δ and $I^\star$ are of the same order, resulting in at most a constant-factor difference in clustering error for spectral clustering and SOLA. We also conduct numerical experiments in Section 5 to verify the necessity of (b).

4.5. Estimation of the number of latent classes K

In practice, the true number of latent classes K may be unknown. To see how we can estimate K from data, recall that $\mathbf{R}=\mathbf{Z}\boldsymbol{\Theta}^\top+\mathbf{E}$. By Weyl's inequality and classical random-matrix theory (Vershynin, 2018), the singular values of $\mathbf{R}$ satisfy

$$\left|\lambda_i(\mathbf{R})-\lambda_i\left(\mathbf{Z}\boldsymbol{\Theta}^\top\right)\right|\leq\|\mathbf{E}\|=O_p\left(\sqrt{J}+\sqrt{N}\right),\quad i\in[N\wedge J].$$

Since $\mathbf{Z}\boldsymbol{\Theta}^\top$ has exactly K nonzero singular values $\lambda_1\geq\cdots\geq\lambda_K>0$, and Proposition 2 shows that $\lambda_K\gtrsim\sqrt{\alpha BNJ}$ under a common generative setting, the top K singular values of $\mathbf{R}$ lie well above the noise level. Accordingly, we estimate K by counting how many singular values of $\mathbf{R}$ exceed a threshold slightly above

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

Table 1. Clustering error across different methods and settings under 100 replicates

Setting/Method	SOLA	SOLA+	EM by `polca`	EM by LatentGold
$N = 190, J = 95$ ($N = 2J$)	0.0815	0.0772	0.3022	0.0645
$N = 95, J = 95$ ($N = J$)	0.1760	0.1745	0.4642	0.2609
$N = 75, J = 150$ ($N = J/2$)	0.0411	0.0411	0.5002	0.2795

the noise level:

$$\widehat{K} = \left|\left\{i \in [N \wedge J] : \lambda_i(\mathbf{R}) > 2.01\left(\sqrt{J} + \sqrt{N}\right)\right\}\right|. \tag{17}$$

Here, the factor 2.01 ensures we count only those singular values contributed by the low-rank signal. We have the following lemma certifying the consistency of $\widehat{K}$, whose proof is deferred to Section B.6.

Lemma 1. *Instate the assumptions of Proposition 1. For $\widehat{K}$ defined in (17), we have*

$$\mathbb{P}\left(\widehat{K} = K\right) = 1 - o(1), \qquad \text{as } N, J \to \infty.$$

This thresholding rule provides a formal analog of the classical scree-plot method (Cattell, 1966; Zhang et al., 2020), and its proof shares a similar spirit of Theorem 2 in Zhang et al. (2020) but with a sharper characterization of the constant factor in the noise magnitude.

5. Numerical experiments

5.1. Simulation studies

We compare our proposed methods SOLA (Algorithm 2) and SOLA+ (Algorithm 2 with multiple steps refinement), with spectral method (Zhang & Zhou, 2024), EM (Linzer & Lewis, 2011), and Tensor-CEM (Zeng et al., 2023). For the EM algorithm, we employ the popular `polca` package (Linzer & Lewis, 2011) in R, and for Tensor-CEM, we use the original MATLAB code (Zeng et al., 2023) for tensor initialization combined with our R-based implementation of the CEM refinement. Our evaluation focuses on three aspects: (a) clustering accuracy, (b) stability, and (c) computational efficiency. Throughout the simulations, we generate the true latent label $\{s_i^\star\}_{i=1}^N$ independently uniformly from $[K]$ and set $K = 3$, $N = 2J$. The parameters $\theta_{j,k}$ are independently sampled from $\text{Beta}(\beta_1, \beta_2)$. We generate 200 independent replicates and report the average mis-clustering error and computation time, excluding any replicates in which the algorithm fails (Tensor-CEM and EM). Notably, our algorithm is numerically stable without any failures in all simulations.

Remark 3 (EM implementations: `polca` vs. LatentGOLD). Beyond `polca`, we also evaluated LatentGOLD (Vermunt & Magidson, 2005), which implements a robust EM routine with safeguards against numerical underflow. Table 1 reports a representative comparison across three regimes ($N = 2J$, $N = J$, and $N = J/2$), showing that LatentGOLD is substantially more stable and accurate than `polca` in these settings. Our overall message remains unchanged: EM-type methods can be competitive in the classical regime with relatively large N and smaller J, whereas SOLA/SOLA+ become increasingly advantageous as J grows and the problem enters the high-dimensional regime (with J comparable to or larger than N). While retaining `polca` for comparison purposes in our large-scale simulation studies, we emphasize that when analyzing data with a small-to-moderate J, it is recommended to use LatentGOLD for their better numerical stability and performance.

Simulation 1: Latent class estimation accuracy for dense $\boldsymbol{\Theta}$ close to $1/2$. In the first simulation, we set $(\beta_1, \beta_2) = (5, 5)$. Under these parameters, the entries of $\boldsymbol{\Theta}$ tend to be close to 0.5. As shown in Figure 2, our proposed methods achieve lower mis-clustering errors compared to the alternative approaches. In this setting, the improvement of SOLA over the plain spectral clustering method is negligible, which

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

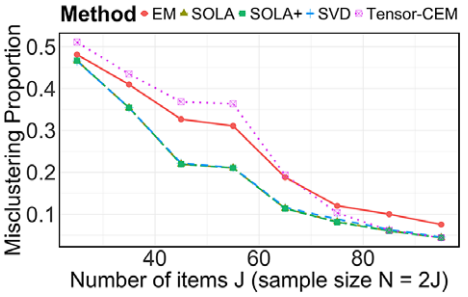


Figure 2. Simulation 1: Mis-clustering proportions versus number of items J. Entries of Θ are independently generated from Beta$(5,5)$.

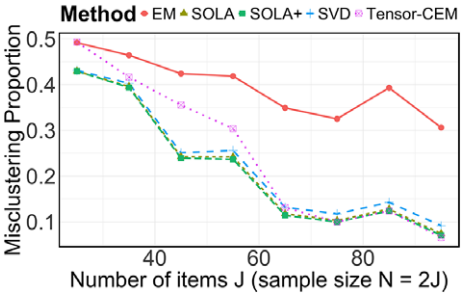


Figure 3. Simulation 2: Mis-clustering proportions versus number of items J. Entries of Θ are independently generated from Beta$(1,8)$.

is expected since $\theta_{j,k}$'s values near 0.5 reduce the additional gain from incorporating the likelihood information. This observation supports the necessity of having $\theta_{j,k}$'s well separated from 0.5 (see Proposition A.1) to have likelihood-based refinement improve upon spectral clustering. It is also worth noting that Tensor-CEM performs poorly when both the number of items J and the sample size N are small, likely due to instability introduced by its tensor-based initialization step.

Simulation 2: Latent class estimation accuracy for sparse item parameters $\mathbf{\Theta}$. In the second simulation, we choose $(\beta_1,\beta_2) = (1,8)$, a sparse scenario, where the $\theta_{j,k}$'s tend to be close to 0. Figure 3 illustrates that our proposed methods exhibit superior clustering accuracy when N and J are small, while maintaining competitive performance as these dimensions increase. In this sparse setting, the traditional EM algorithm performs the worst because $\theta_{j,k}$'s are near the boundary (close to 0), making it difficult for random initializations to converge to optimal solutions. Furthermore, the gap between spectral clustering and SOLA becomes more pronounced when the $\theta_{j,k}$'s are bounded away from 0.5.

Simulation 3: Stability analysis. We next assess the stability of the different methods under the sparse setting $(\beta_1,\beta_2) = (1,8)$. In particular, we record a "failure" point if either (a) the estimated number of classes degenerates from the pre-specified K to a smaller value or (b) the algorithm produces an error during execution. As shown in Figure 4, our proposed methods are robust across a wide range of N and J without any failures in any simulation trial. In particular, the failure rate for Tensor-CEM exceeds 15% for small N and J, likely due to its sensitive initialization, while the EM algorithm becomes increasingly unstable as the dimensions grow. These results underscore the robustness of our SOLA approaches.

Simulation 4: Computational efficiency. Finally, we compare the computation time required by each method. In our experiments, the SVD step and our SOLA and SOLA+ algorithms are implemented in R. The total computation time for Tensor-CEM is the sum of the MATLAB initialization and the R refinement stages. As reported in Table 2 and illustrated in Figure 5, our proposed methods are significantly faster than both the EM and Tensor-CEM approaches, with computation time comparable to that of SVD.

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

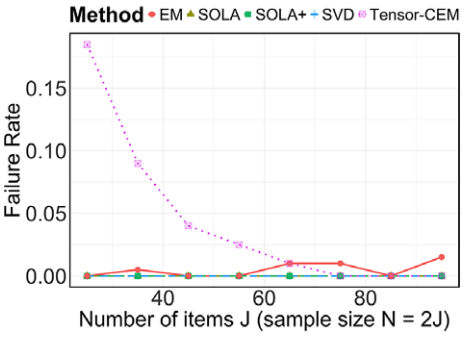


Figure 4. Simulation 3: Failure rate versus the number of items J. Entries of Θ are independently generated from $\text{Beta}(1,8)$.

Table 2. Simulation 4: Running time (seconds) of different methods

	SVD	SOLA	SOLA+	EM	Tensor-CEM
$J = 25$	4.15×10^{-4}	8.62×10^{-4}	4.19×10^{-3}	1.42×10^{-2}	5.12×10^{-1}
$J = 95$	2.33×10^{-3}	4.66×10^{-3}	2.51×10^{-2}	1.80×10^{-1}	6.95×10^{-1}

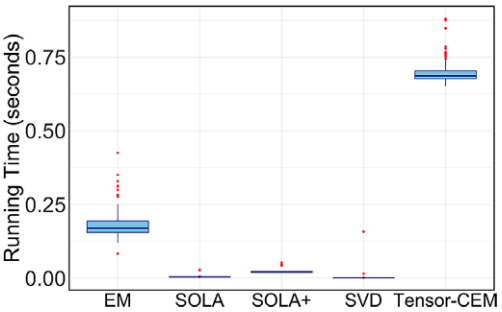


Figure 5. Simulation 4: Running time (seconds) of different methods.

Simulation 5: Latent class estimation accuracy under different regimes. In the fifth simulation study, we also employ Spec-EM (EM algorithm initialized by spectral clustering), motivated by the fact that when the number of items is moderately large, spectral clustering provides an effective initialization for the classical EM algorithm for LCMs. In particular, we consider the following regimes: (i) $N = J$ (Figure 6), (ii) $N = J/2$ (Figure 7), (iii) fix $N = 100$ and vary J (Figure 8), (iv) fix $J = 100$ and vary N (Figure 9), (v) fix $J = 30$ and vary N (Figure 10), and (vi) $N = J$ with $\{s_i^\star\}_{i=1}^N$ being independently generated from $[3]$ with probability $(p_1, p_2, p_3) = (1/6, 1/3, 1/2)$ (Figure 11). In all these settings, we set $(\beta_1, \beta_2) = (1, 8)$. In addition, we consider (vii) $N = J$ under the fixed design of $\boldsymbol{\Theta}$ (Figure 12), where the fixed item parameter matrix $\boldsymbol{\Theta}$ is constructed by vertically concatenating copies of a 5×3 block matrix $\boldsymbol{\Theta}_0$ as follows:

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_0 \\ \vdots \\ \boldsymbol{\Theta}_0 \end{bmatrix} \Bigg\} b \text{ copies}, \qquad \boldsymbol{\Theta}_0 := \begin{bmatrix} 0.3 & 0.7 & 0.7 \\ 0.3 & 0.7 & 0.3 \\ 0.7 & 0.3 & 0.7 \\ 0.7 & 0.3 & 0.3 \\ 0.7 & 0.7 & 0.3 \end{bmatrix}.$$

Here, the block matrix $\boldsymbol{\Theta}_0$ is vertically stacked b times to yield $J = 5b$ items with $K = 3$ latent classes. It is clear that SOLA and SOLA+ perform significantly better in most scenarios except the most traditional

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

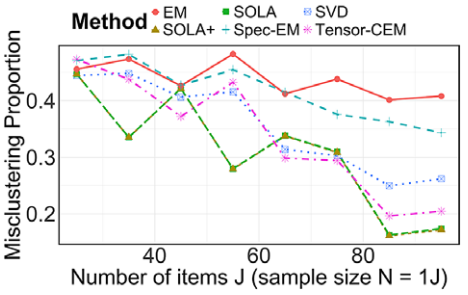


Figure 6. Simulation 5-1: Mis-clustering proportions versus number of items J under $N = J$. Entries of Θ are independently generated from Beta$(1,8)$.

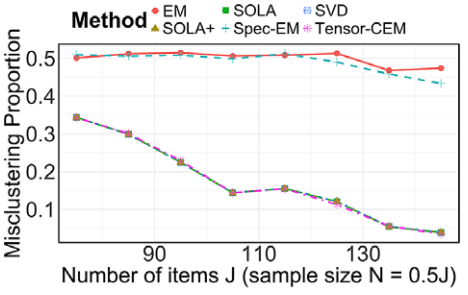


Figure 7. Simulation 5-2: Mis-clustering proportions versus number of items J under $N = 0.5J$. Entries of Θ are independently generated from Beta$(1,8)$.

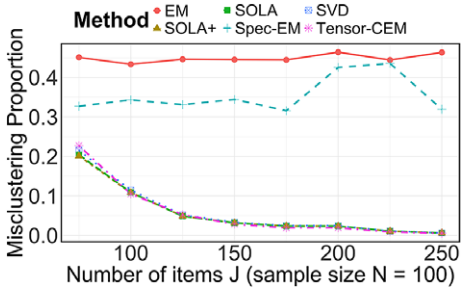


Figure 8. Simulation 5-3: Mis-clustering proportions versus number of items J under $N = 100$. Entries of Θ are independently generated from Beta$(1,8)$.

large-N, small-J regime, where Spec-EM performs better. This underpins both the validity of SOLA and spectral clustering.

Simulation 6: Estimation of number of latent classes. Our theoretical guarantee for the selection criterion of K relies on random matrix theory under the asymptotic regime $N,J \to \infty$. This means relatively large N and J are needed for the theoretical guarantee to hold, which is consistent with our focus on high-dimensional data. The constant "2" in the threshold reflects a worst-case bound and can indeed be conservative, especially when entries of $\mathbf{\Theta}$ are mostly concentrated away from $1/2$. To illustrate its effectiveness in practice, we have performed simulation studies with $N = 2J$ and $K = 3$, under several generative mechanisms for $\mathbf{\Theta}$ (i.i.d. Beta$(1,1)$, Beta$(1,4)$, and Beta$(1,5)$, which reflect different sparsity levels, and also the fixed design as described in Simulation 5). As shown in Table 3, the criterion successfully recovers K in high-dimensional regimes. Our real-data analysis in Section 5.2 also demonstrates that the method works well when the data are sufficiently large. For smaller N and J,

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

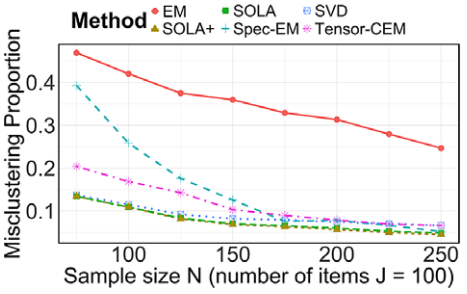


Figure 9. Simulation 5-4: Mis-clustering proportions versus sample size N under $J = 100$. Entries of Θ are independently generated from Beta$(1, 8)$.

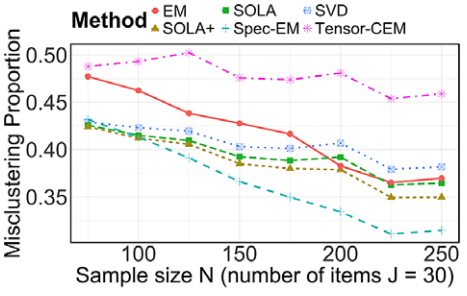


Figure 10. Simulation 5-5: Mis-clustering proportions versus sample size N under $J = 30$. Entries of Θ are independently generated from Beta$(1, 8)$.

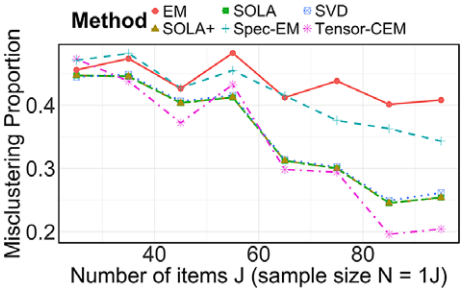


Figure 11. Simulation 5-6: Mis-clustering proportions versus number of items J under imbalanced latent classes with $N = J$. Entries of Θ are independently generated from Beta$(1, 8)$.

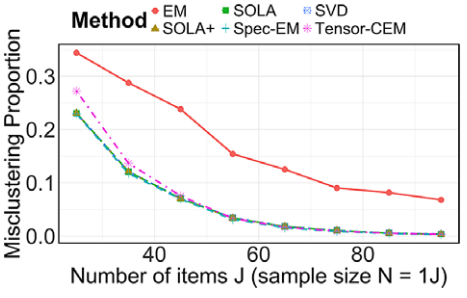


Figure 12. Simulation 5-7: Mis-clustering proportions versus number of items J under $N = J$. The Θ matrix is set to a fixed design matrix.

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

Table 3. Proportion of successfully selecting $\widehat{K}$ to be the true K based on 200 simulation replicates, under different generative mechanisms of the item parameters $\mathbf{\Theta} = (\theta_{j,k})_{J\times K}$

J	1,000	1,200	1,400	1,600	1,800	2,000
Fixed design of item parameters	100%	100%	100%	100%	100%	100%
$\theta_{j,k} \overset{i.i.d.}{\sim} \text{Beta}(1,1)$	100%	100%	100%	100%	100%	100%
$\theta_{j,k} \overset{i.i.d.}{\sim} \text{Beta}(1,4)$	0%	0%	100%	100%	100%	100%
$\theta_{j,k} \overset{i.i.d.}{\sim} \text{Beta}(1,5)$	0%	0%	0%	0%	70%	100%

Table 4. Mis-clustering error and running time (seconds) of different methods on Senate Roll Call Votes data

	SVD	SOLA	SOLA+	EM	Tensor-CEM
Error	0.0212	0.0212	0.0212	0.117	0.0212
Time	0.00179	0.0344	0.0503	3.125	5.154

we recommend that practitioners use classical information criteria, such as BIC, as a more reliable and interpretable tool for selecting K.

5.2. Real data example

We further evaluate the performance of our methods on the United States 112th Senate Roll Call Votes data, which is publicly available at https://legacy.voteview.com/senate112.htm. The dataset contains voting records for $J = 486$ roll calls by 102 U.S. senators. Following the preprocessing procedure in Lyu et al. (2025), we convert the response matrix to binary and remove senators who are neither Democrats nor Republicans, as well as those with more than 10% missing votes. This results in a final sample of $N = 94$ senators. For the remaining missing entries, we impute values by randomly assigning 0 or 1 with probability equal to the positive response rate observed from each senator's non-missing votes.

We compare the performance of the methods considered in our simulation studies. Table 4 summarizes the mis-clustering error and computation time (in seconds) for each method. Interestingly, the SVD approach yields an error comparable to those obtained with SOLA and SOLA+, suggesting that the refinement step may be unnecessary for this dataset. This observation can be partially explained by our remark following Theorem 2: when the number of items J is substantially larger than the number of individuals N, spectral clustering alone can perform very well. In contrast, the traditional EM method exhibits significantly higher mis-clustering error, and EM and Tensor-CEM suffer from high computational cost in this large-J regime.

We also validate the performance of our method for estimating K on this real dataset. In particular, the threshold defined in (17) is $2.01 \times (\sqrt{486} + \sqrt{94}) \approx 63.8$. The first three largest singular values of $\mathbf{R}$ are $\{148.1, 64.4, 16.6\}$, which leads to an estimated number of clusters $\widehat{K} = 2$. Notably, this estimated value exactly matches the known number of clusters in the dataset (i.e., $K = 2$).

6. Discussion

In this work, we have proposed SOLA, a simple yet powerful two-stage algorithm, under binary-response LCMs. Our method efficiently exploits the low-rank structure of the response matrix and further leverages the likelihood information. We have proved that SOLA attains the optimal mis-clustering rate and achieves the fundamental statistical limit. We have also empirically demonstrated its superior accuracy, stability, and speed compared to other methods.

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

Several interesting directions remain for future research. Many applications involve multivariate polytomous responses with more than two categories for each item. Extending SOLA to handle polytomous-response LCMs is an important future direction. It remains intriguing to investigate how to modify the algorithm accordingly and obtain similar theoretical guarantees. In addition, beyond estimation, practitioners often also want to conduct statistical inference on item parameters $\boldsymbol{\Theta}$. It is worth exploring that whether we can build on SOLA to develop an estimator of item parameters $\boldsymbol{\Theta}$ together with confidence intervals.

Funding statement. This research is partially supported by the NSF Grant DMS-2210796.

Competing interests. The authors declare none.

References

Abbe, E. (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, *18*(177), 1–86.

Bandeira, A. S., & Van Handel, R. (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Annals of probability: An official journal of the Institute of Mathematical Statistics*, *44*(4), 2479–2506.

Berzofsky, M. E., Biemer, P. P., & Kalsbeek, W. D. (2014). Local dependence in latent class analysis of rare and sensitive events. *Sociological Methods & Research*, *43*(1), 137–170.

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of Independence*. Oxford University Press.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Celeux, G., & Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, *14*(3), 315–332.

Chen, L., & Gu, Y. (2024). A spectral method for identifiable grade of membership analysis with binary responses. *Psychometrika*, *89*(2), 626–657.

Chen, S., Liu, S., & Ma, Z. (2022). Global and individualized community detection in inhomogeneous multilayer networks. *The Annals of Statistics*, *50*(5), 2664–2693.

Chen, Y., & Li, X. (2019). Exploratory data analysis for cognitive diagnosis: Stochastic co-blockmodel and spectral co-clustering. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 287–306). Springer International Publishing.

Chen, Y., Li, X., Liu, J., Xu, G., & Ying, Z. (2017). Exploratory item classification via spectral graph clustering. *Applied Psychological Measurement*, *41*(8), 579–599.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Di Mari, R., Bakk, Z., Oser, J., & Kuha, J. (2023). A two-step estimator for multilevel latent class analysis with covariates. *Psychometrika*, *88*(4), 1144–1170.

Gao, C., Ma, Z., Zhang, A. Y., & Zhou, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, *18*(1), 1980–2024.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.

Korpershoek, H., Kuyper, H., & van der Werf, G. (2015). Differences in students' school motivation: A latent class modelling approach. *Social Psychology of Education*, *18*, 137–163.

Lazarsfeld, P. F. & Henry, N. W. (1968). Latent structure analysis. Boston: Houghton Mifflin.

Le Cam, L. (2012). *Asymptotic methods in statistical decision theory*. Springer Science & Business Media.

Lei, J., & Zhu, L. (2017). Generic sample splitting for refined community recovery in degree corrected stochastic block models. *Statistica Sinica*, 1639–1659.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*, 1–29.

Lyu, Z., Chen, L., & Gu, Y. (2025). Degree-heterogeneous latent class analysis for high-dimensional discrete data. *Journal of the American Statistical Association*, *120*(552), 2435–2448.

Lyu, Z., Li, T., & Xia, D. (2023). *Optimal clustering of discrete mixtures: Binomial, poisson, block models, and multi-layer networks*. Preprint, arXiv:2311.15598.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, 1–32.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: Contributions to the theory of statistics* (Vol. 4, pp. 547–562). University of California Press.

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

Rinaldo, A., Wasserman, L., & G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, *47*(6), 3438–3469.
Takezawa, K. (2005). *Introduction to nonparametric regression*. John Wiley & Sons.
Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.0 user's guide*. Statistical Innovations Inc.
Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* (Vol. *47*). Cambridge University Press.
Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*, 395–416.
Vu, V. (2018). A simple SVD algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, *27*(1), 124–140.
Wang, M., & Hanges, P. J. (2011). Latent class procedures: Applications to organizational research. *Organizational Research Methods*, *14*(1), 24–31.
Zeng, Z., Gu, Y., & Xu, G. (2023). A tensor-EM method for large-scale latent class analysis with binary responses. *Psychometrika*, *88*(2), 580–612.
Zhang, A. Y. (2024). Fundamental limits of spectral clustering in stochastic block models. *IEEE Transactions on Information Theory*, *70*(10), 7320–7348.
Zhang, A. Y., & Zhou, H. Y. (2024). Leave-one-out singular subspace perturbation analysis for spectral clustering. *The Annals of Statistics*, *52*(5), 2004–2033.
Zhang, H., Chen, Y., & Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, *85*(2), 358–372.

A. Appendix of the proofs

We present a general version of Proposition 3. To this end, we introduce some additional notation. For $\forall j \in [J]$ and $k_1 \neq k_2 \in [K]$, define $\Delta^2_{j,k_1,k_2} := \left(\theta_{j,k_1} - \theta_{j,k_2}\right)^2 / \left[8(\sigma^2_{\theta_{j,k_1}} \vee \sigma^2_{\theta_{j,k_2}})\right]$. We have the following result, whose proof can be found in Section B.5.

Proposition A.1. *For any $k_1 \neq k_2 \in [K]$, define*

$$\tau_{k_1,k_2} := \frac{\sum_{j\in\Omega_0(k_1,k_2)} \Delta^2_{j,k_1,k_2} \cdot \tau\left(\theta_{j,k_1}\right)}{\sum_{j\in\Omega_0(k_1,k_2)} \Delta^2_{j,k_1,k_2}} (1+\eta), \quad \tau(x) = \frac{1-2x}{2x(1-x)\log\left((1-x)/x\right)},$$

for some $\eta = o(1)$. Assume that (a) $\max_{k_1\neq k_2\in[K]} |\Omega_0^c(k_1,k_2)| = O(1)$, where $\Omega_0(k_1,k_2) := \left\{j \in [J] : \left|\theta_{j,k_1} - \theta_{j,k_2}\right| = o(1)\right\}$, then we have

$$\exp\left(-\frac{I^\star}{2}\right) \lesssim \exp\left(-\frac{\Delta^2}{8\tilde{\sigma}^2}(1+\underline{\tau})\right),$$

where $\underline{\tau} := \min_{k_1\neq k_2\in[K]} \tau_{k_1,k_2}$. Furthermore, if (b) $\min_{k\in[K]} \left|\left\{j \in [J] : \left|\theta_{j,k} - 1/2\right| > c_0\right\}\right| \gtrsim J$ for some universal constant $c_0 > 0$, then we have $\underline{\tau} > c_\tau$ for some universal constant $c_\tau > 0$.

B. Proofs of main results

B.1. Proof of Theorem 1

The proof can be adapted from Theorem S.4 in Lyu et al. (2025) by fixing the individual degree parameters therein to be all 1's. We only sketch essential modifications here. Let

$$\mathcal{P}_\Theta := \left\{\widetilde{\mathbf{\Theta}} : \widetilde{\theta_{j,k}} \in [c_\theta, C_\theta], j \in [J], k \in [K], \max_{j\in[J]} \max_{k\neq k'} \left|\widetilde{\theta}_{j,k} - \widetilde{\theta}_{j,k'}\right| = o(1)\right\}.$$

Following the construction of $\mathcal{V}^*$ and the proof of Theorem S.4 in Lyu et al. (2025), it suffices to consider the following space:

$$\tilde{\mathcal{P}}_K(\boldsymbol{s}, \mathbf{\Theta}) := \left\{\tilde{\mathbf{R}} : \tilde{R}_{i,j} = \theta_{j,s_i}, \boldsymbol{s} \in \mathcal{V}^*, \mathbf{\Theta} \in \mathcal{P}_\Theta\right\}.$$

Then we can follow the proof of Theorem S.4 in Lyu et al. (2025). Notice that in the proof of Lemma S.2, the argument therein still holds by allowing $p_{1,j} \asymp p_{2,j} \asymp 1$ and $\left|p_{1,j} - p_{2,j}\right| = o\left(p_{1,j}\right)$. We can then get the desired result by directly applying the modified Lemma S.2 without further lower bounding it.

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

B.2. Proof of Theorem 2

Initialization: We start with investigating the performance of $\widehat{\theta}_{j,k}^{(m)}$'s. To this end, for any $\rho = o(1)$, we define the event

$$\mathcal{E}_{\rho,m} := \left\{ \min_{\pi \in \mathfrak{S}_K} \max_{j \in [J], k \in [K]} \left| \widehat{\theta}_{j,k}^{(m)} - \theta_{j,\pi(k)} \right| \le \rho I^\star / J \right\}. \tag{B.1}$$

The following lemma indicates that (B.1) holds with high probability, whose proof is deferred to Section C.1.

Lemma B.1. *There exists some absolute constant* $c_0 \in (0,1)$ *such that* $\mathbb{P}(\mathcal{E}_{\rho,m}) \ge 1 - \exp(-(1-c_0)\Delta^2/(8\tilde{\sigma}^2))$ *for* $m \in [2]$ *with* $\rho = o(1/K)$.

One-step refinement: In the following, we will derive the clustering error for $i \in \mathcal{S}_1$. Fix any $k \in [K] \setminus \{s_i^\star\}$. Without loss of generality, we assume that $\pi = \mathrm{Id}$ (identity map with $\pi(k) = k$ for all k) in (B.1) as otherwise we can replace $\theta_{j,k}$ with $\theta_{j,\pi(k)}$ and $\theta_{j,s_i^\star}$ with $\theta_{j,\pi(s_i^\star)}$ in the following derivation. By definition, we have

$$\begin{aligned}
\mathbb{I}(\widehat{s}_i = k) &\le \mathbb{I}\left(\sum_{j\in[J]} R_{i,j} \log \frac{\widehat{\theta}_{j,k}^{(2)} \left(1 - \widehat{\theta}_{j,s_i^\star}^{(2)}\right)}{\widehat{\theta}_{j,s_i^\star}^{(2)} \left(1 - \widehat{\theta}_{j,k}\right)} \ge \sum_{j\in[J]} \log \frac{1 - \widehat{\theta}_{j,s_i^\star}^{(2)}}{1 - \widehat{\theta}_{j,k}^{(2)}} \right) \\
&\le \mathbb{I}\left(\sum_{j\in[J]} R_{i,j} \log \frac{\theta_{j,k}\left(1 - \theta_{j,s_i^\star}\right)}{\theta_{j,s_i^\star}\left(1 - \theta_{j,k}\right)} \ge \sum_{j\in[J]} \log \frac{1 - \theta_{j,s_i^\star}}{1 - \theta_{j,k}} - \delta I^\star \right) \\
&\quad + \mathbb{I}\left(\sum_{j\in[J]} R_{i,j} \left(\log \frac{\widehat{\theta}_{j,k}^{(2)} \left(1 - \widehat{\theta}_{j,s_i^\star}^{(2)}\right)}{\widehat{\theta}_{j,s_i^\star}^{(2)} \left(1 - \widehat{\theta}_{j,k}^{(2)}\right)} - \log \frac{\theta_{j,k}\left(1 - \theta_{j,s_i^\star}\right)}{\theta_{j,s_i^\star}\left(1 - \theta_{j,k}\right)} \right) \right) \\
&\ge \sum_{j\in[J]} \left(\log \frac{1 - \widehat{\theta}_{j,s_i^\star}^{(2)}}{1 - \widehat{\theta}_{j,k}^{(2)}} - \log \frac{1 - \theta_{j,s_i^\star}}{1 - \theta_{j,k}} \right) + \delta I^\star \Bigg),
\end{aligned} \tag{B.2}$$

where $\delta = o(1)$ is some sequence to be determined later. We start with bounding the first term in (B.2). By the Chernoff bound, we obtain that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j\in[J]} R_{i,j} \log \frac{\theta_{j,k}\left(1 - \theta_{j,s_i^\star}\right)}{\theta_{j,s_i^\star}\left(1 - \theta_{j,k}\right)} \ge \sum_{j\in[J]} \log \frac{1 - \theta_{j,s_i^\star}}{1 - \theta_{j,k}} - \delta I^\star \right) \\
&\le \exp\left(- \sum_{j\in[J]} \log \sqrt{\frac{1 - \theta_{j,s_i^\star}}{1 - \theta_{j,k}}} + \frac{\delta I^\star}{2} \right) \prod_{j=1}^{J} \left(\sqrt{\theta_{j,s_i^\star}\theta_{j,k}} \sqrt{\frac{1 - \theta_{j,s_i^\star}}{1 - \theta_{j,k}}} + 1 - \theta_{j,s_i^\star} \right) \\
&\le \exp\left(\frac{\delta I^\star}{2} \right) \exp\left(-\frac{I^\star}{2} \right) = \exp\left(-\frac{I^\star}{2}(1 - \delta) \right).
\end{aligned} \tag{B.3}$$

Using this, we can bound the second term in (B.2) as

$$\begin{aligned}
&\mathbb{I}\left(\sum_{j\in[J]} R_{i,j} \left(\log \frac{\widehat{\theta}_{j,k}^{(2)} \left(1 - \widehat{\theta}_{j,s_i^\star}^{(2)}\right)}{\widehat{\theta}_{j,s_i^\star}^{(2)} \left(1 - \widehat{\theta}_{j,k}^{(2)}\right)} - \log \frac{\theta_{j,k}\left(1 - \theta_{j,s_i^\star}\right)}{\theta_{j,s_i^\star}\left(1 - \theta_{j,k}\right)} \right) \ge \sum_{j\in[J]} \left(\log \frac{1 - \widehat{\theta}_{j,s_i^\star}^{(2)}}{1 - \widehat{\theta}_{j,k}^{(2)}} - \log \frac{1 - \theta_{j,s_i^\star}}{1 - \theta_{j,k}} \right) + \delta I^\star \right) \\
&\le \mathbb{I}\left(\sum_{j\in[J]} R_{i,j} \log \frac{\widehat{\theta}_{j,k}^{(2)}}{\theta_{j,k}} > \frac{\delta}{4} I^\star \right) + \mathbb{I}\left(\sum_{j\in[J]} R_{i,j} \log \frac{\theta_{j,s_i^\star}}{\widehat{\theta}_{j,s_i^\star}} > \frac{\delta}{4} I^\star \right) \\
&\quad + \mathbb{I}\left(\sum_{j\in[J]} \left(R_{i,j} - 1\right) \log \frac{1 - \widehat{\theta}_{j,s_i^\star}^{(2)}}{1 - \theta_{j,s_i^\star}} > \frac{\delta}{4} I^\star \right) + \mathbb{I}\left(\sum_{j\in[J]} \left(R_{i,j} - 1\right) \log \frac{1 - \theta_{j,k}}{1 - \widehat{\theta}_{j,k}^{(2)}} > \frac{\delta}{4} I^\star \right).
\end{aligned} \tag{B.4}$$

For the first term in (B.4), we proceed by noticing that

$$\begin{aligned}
&\mathbb{I}\left(\sum_{j\in[J]} R_{i,j} \log \frac{\widehat{\theta}_{j,k}^{(2)}}{\theta_{j,k}} > \frac{\delta}{4} I^\star \right) \\
&\le \mathbb{I}\left(\sum_{j\in[J]} R_{i,j} \log \frac{\widehat{\theta}_{j,k}^{(2)}}{\theta_{j,k}} \mathbb{I}\left(\mathcal{E}_{\rho,2}\right) > \frac{\delta}{8} I^\star \right) + \mathbb{I}\left(\sum_{j\in[J]} R_{i,j} \log \frac{\widehat{\theta}_{j,k}^{(2)}}{\theta_{j,k}} \mathbb{I}\left(\mathcal{E}_{\rho,2}^c\right) > \frac{\delta}{8} I^\star \right).
\end{aligned} \tag{B.5}$$

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

Let $\delta = \rho^{\varepsilon}$ for $\varepsilon \in (0,1)$ and $\lambda = c\rho^{-\varepsilon}$ for some constant $c > 0$ depending only on c_θ and C_θ, then by Chernoff bound on the first term of (B.5), we obtain that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j\in[J]} R_{i,j}\log\frac{\widehat{\theta}^{(2)}_{j,k}}{\theta_{j,k}}\mathbb{I}\left(\mathcal{E}_{\rho,2}\right) > \frac{\delta}{8}I^\star\right) \overset{(a)}{\leq} \mathbb{P}\left(\sum_{j\in[J]} R_{i,j}\frac{\rho I^\star}{J} > \frac{c_\theta\delta}{8}I^\star\right)\\
&\leq \exp\left(-\lambda\frac{c_\theta\delta}{8}I^\star\right)\prod_{j=1}^{J}\mathbb{E}\exp\left(\lambda R_{i,j}\frac{\rho I^\star}{J}\right) \overset{(b)}{\leq} \exp\left(-\lambda\frac{c_\theta\delta}{8}I^\star\right)\prod_{j=1}^{J}\left(1+2\lambda\theta_{i,j}\frac{\rho I^\star}{J}\right)\\
&\leq \exp\left(-\lambda\frac{c_\theta\delta}{8}I^\star\right)\exp\left(\rho\cdot 2C_\theta I^\star\right) \overset{(c)}{\leq} \exp\left(-\frac{I^\star}{2}\right),
\end{aligned}$$

where (a) holds due to (B.1) and $\log(1+x) \leq x$ for $x > -1$, (b) holds due to $e^x - 1 \leq 2x$ for $0 < x < 1$ and the fact that $\lambda\rho I^\star/J \lesssim \rho^{1-\varepsilon}\Delta^2/J = o(1)$, and (c) holds due to the choice of $\lambda = c\rho^{-\varepsilon}$. For the second term of (B.5), by the independence between $\widehat{\theta}^{(2)}_{j,k}$ and $\left\{R_{i,j}, j\in[J]\right\}$, we get

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j\in[J]} R_{i,j}\log\frac{\widehat{\theta}^{(2)}_{j,k}}{\theta_{j,k}}\mathbb{I}\left(\mathcal{E}^c_{\rho,2}\right) > \frac{\delta}{8}I^\star\right) \leq \exp\left(-\lambda\frac{\delta}{8}I^\star\right)\prod_{j=1}^{J}\mathbb{E}\exp\left(\lambda R_{i,j}\log\frac{\widehat{\theta}^{(2)}_{j,k}}{\theta_{j,k}}\mathbb{I}\left(\mathcal{E}^c_{\rho,2}\right)\right)\\
&\leq \exp\left(-\lambda\frac{\delta}{8}I^\star\right)\prod_{j=1}^{J}\left[1+\theta_{j,k}\left(\mathbb{E}\exp\left(\lambda\log\frac{\widehat{\theta}^{(2)}_{j,k}}{\theta_{j,k}}\mathbb{I}\left(\mathcal{E}^c_{\rho,2}\right)\right)-1\right)\right].
\end{aligned}$$

To bound the above term, notice that

$$\mathbb{E}\left(\exp\left(\lambda\log\frac{\widehat{\theta}^{(2)}_{j,k}}{\theta_{j,k}}\mathbb{I}\left(\mathcal{E}^c_{\rho,2}\right)\right)-1\right) \leq \left(\exp\left(\lambda\log\frac{1}{c_\theta}\right)-1\right)\mathbb{P}\left(\mathcal{E}^c_{\rho,2}\right) \leq \exp\left(\lambda\log\frac{1}{c_\theta}-\frac{\Delta^2}{16\tilde{\sigma}^2}\right).$$

Since we can always choose $\rho \to 0$ sufficiently slow such that

$$\lambda\log\frac{1}{c_\theta} \lesssim \frac{1}{\rho^{\varepsilon}} \lesssim \frac{\Delta^2}{\tilde{\sigma}^2} \asymp I^\star,$$

we can conclude that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j\in[J]} R_{i,j}\log\frac{\widehat{\theta}^{(2)}_{j,k}}{\theta_{j,k}}\mathbb{I}\left(\mathcal{E}^c_{\rho,2}\right) > \frac{\delta}{8}I^\star\right) \leq \exp\left(-\lambda\frac{\delta}{8}I^\star\right)\prod_{j=1}^{J}\left[1+\theta_{j,k}\exp\left(-cI^\star\right)\right]\\
&\leq \exp\left(-\lambda\frac{\delta}{8}I^\star\right)\left(1+\frac{C_\theta}{\exp\left(cI^\star\right)}\right)^{J} \leq \exp\left(-\frac{I^\star}{2}\right).
\end{aligned}$$

where the last inequality due to the fact that $\left(1+C_\theta\exp(-cI^\star)\right)^J = O(1)$ provided that $\exp(cI^\star) \geq J$. We thereby arrive at

$$\mathbb{P}\left(\sum_{j\in[J]} R_{i,j}\log\frac{\widehat{\theta}^{(2)}_{j,k}}{\theta_{j,k}} > \frac{\delta}{4}I^\star\right) \leq 2\exp\left(-\frac{I^\star}{2}\right). \tag{B.6}$$

The second term in (B.4) can be bounded in the same fashion. For the third term in (B.4), we have

$$\begin{aligned}
&\mathbb{I}\left(\sum_{j\in[J]}\left(R_{i,j}-1\right)\log\frac{1-\widehat{\theta}^{(2)}_{j,s_i^\star}}{1-\theta_{j,s_i^\star}} > \frac{\delta}{4}I^\star\right)\\
&\leq \mathbb{I}\left(\sum_{j\in[J]}\left(R_{i,j}-1\right)\log\frac{1-\widehat{\theta}^{(2)}_{j,s_i^\star}}{1-\theta_{j,s_i^\star}}\mathbb{I}\left(\mathcal{E}_{\rho,2}\right) > \frac{\delta}{8}I^\star\right) + \mathbb{I}\left(\sum_{j\in[J]}\left(R_{i,j}-1\right)\log\frac{1-\widehat{\theta}^{(2)}_{j,s_i^\star}}{1-\theta_{j,s_i^\star}}\mathbb{I}\left(\mathcal{E}^c_{\rho,2}\right) > \frac{\delta}{8}I^\star\right).
\end{aligned}$$

Similarly, we can bound the expectation of the first term in the above equation as

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j\in[J]}\left(R_{i,j}-1\right)\log\frac{1-\widehat{\theta}^{(2)}_{j,s_i^\star}}{1-\theta_{j,s_i^\star}}\mathbb{I}\left(\mathcal{E}_{\rho,2}\right) > \frac{\delta}{8}I^\star\right)\\
&\leq \exp\left(-\lambda\frac{(1-C_\theta)\,\delta}{8}I^\star\right)\prod_{j=1}^{J}\mathbb{E}\exp\left(\lambda\left(R_{i,j}-1\right)\frac{\rho I^*}{J}\right)
\end{aligned}$$

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

$$\overset{(a)}{\leq} \exp\left(-\lambda\frac{(1-C_\theta)\,\delta}{8}I^\star\right)\prod_{j=1}^{J}\left(1+\frac{C_\theta\rho I^\star}{2J}\right)$$
$$\leq \exp\left(-\lambda\frac{(1-C_\theta)\,\delta}{8}I^\star\right)\exp\left(\rho\cdot\frac{C_\theta I^\star}{2}\right)\leq \exp\left(-\frac{I^\star}{2}\right),$$

where (a) holds due to $e^{-x}\leq 1-x/2$ for $x\in[0,1]$. Similarly, we have that

$$\mathbb{P}\left(\sum_{j\in[J]}\left(R_{i,j}-1\right)\log\frac{1-\widehat{\theta}^{(2)}_{j,s_i^\star}}{1-\theta_{j,s_i^\star}}\mathbb{I}\left(\mathcal{E}^c_{\rho,2}\right)>\frac{\delta}{8}I^\star\right)$$
$$\leq \exp\left(-\lambda\frac{\delta}{8}I^\star\right)\prod_{j=1}^{J}\left[\theta_{j,s_i^\star}+\left(1-\theta_{j,s_i^\star}\right)\mathbb{E}\exp\left(\lambda\log\frac{1-\widehat{\theta}^{(2)}_{j,s_i^\star}}{1-\theta_{j,s_i^\star}}\mathbb{I}\left(\mathcal{E}^c_{\rho,2}\right)\right)\right]$$
$$\leq \exp\left(-\lambda\frac{\delta}{8}I^\star\right)\prod_{j=1}^{J}\left[\theta_{j,s_i^\star}+\left(1-\theta_{j,s_i^\star}\right)\exp\left(\lambda\log\frac{1}{1-C_\theta}\right)\mathbb{P}\left(\mathcal{E}^c_{\rho,2}\right)\right]$$
$$\leq \exp\left(-\lambda\frac{\delta}{8}I^\star\right)\prod_{j=1}^{J}\left[\theta_{j,s_i^\star}+\left(1-\theta_{j,s_i^\star}\right)\exp\left(-cI^\star\right)\right]$$
$$\leq \exp\left(-\lambda\frac{\delta}{8}I^\star\right)C_\theta^J\left(1+\frac{1}{c_\theta\exp\left(cI^\star\right)}\right)^J\leq \exp\left(-\frac{I^\star}{2}\right),$$

provided that $\exp(cI^\star)\geq J$. We thus get that

$$\mathbb{P}\left(\sum_{j\in[J]}\left(R_{i,j}-1\right)\log\frac{1-\widehat{\theta}^{(2)}_{j,s_i^\star}}{1-\theta_{j,s_i^\star}}>\frac{\delta}{4}I^\star\right)\leq 2\exp\left(-\frac{I^\star}{2}\right). \tag{B.7}$$

The last term in (B.4) can be bounded in the same fashion. By (B.2), (B.3), (B.4), (B.6), and (B.7), we thereby arrive at

$$\mathbb{P}\left(\widehat{s}_i=k\right)\leq \exp\left(-\frac{I^\star}{2}\left(1-\rho^\varepsilon\right)\right)+8\exp\left(-\frac{I^\star}{2}\right)\leq \exp\left(-\frac{I^\star}{2}\left(1-o(1)\right)\right).$$

We thus conclude that

$$\frac{1}{|\mathcal{S}_1|}\mathbb{E}\sum_{i\in\mathcal{S}_1}\mathbb{I}\left(\widehat{s}_i\neq s_i^\star\right)\leq \exp\left(-\frac{I^\star}{2}\left(1-o(1)\right)\right).$$

The proof for the case $i\in\mathcal{S}_2$ can be conducted in the same fashion. We can conclude that

$$\frac{1}{|\mathcal{S}_2|}\mathbb{E}\sum_{i\in\mathcal{S}_2}\mathbb{I}\left(\widehat{s}_i\neq \pi_0\left(s_i^\star\right)\right)\leq \exp\left(-\frac{I^\star}{2}\left(1-o(1)\right)\right)$$

for some $\pi_0\in\mathfrak{G}_K$.

Label alignment: We then proceed with the proof for label alignment. Recall that without loss of generality, we have assumed that $\ell(\widehat{\boldsymbol{s}}^{(1)},\boldsymbol{s}^\star_{\mathcal{S}_1})=\ell_0(\widehat{\boldsymbol{s}}^{(1)},\boldsymbol{s}^\star_{\mathcal{S}_1})$ and $\ell(\widehat{\boldsymbol{s}}^{(2)},\boldsymbol{s}^\star_{\mathcal{S}_2})=\ell_0(\widehat{\boldsymbol{s}}^{(2)},\pi_0(\boldsymbol{s}^\star_{\mathcal{S}_2}))$ for some $\pi_0\in\mathfrak{G}_K$. Notice that

$$\ell_0\left(\widehat{\boldsymbol{s}}_{\mathcal{S}_2},\boldsymbol{s}^\star_{\mathcal{S}_2}\right)=\ell_0\left(\widehat{\pi}(\widehat{\boldsymbol{s}}^{(2)}),\boldsymbol{s}^\star_{\mathcal{S}_2}\right)=\ell_0\left(\widehat{\boldsymbol{s}}^{(2)},\widehat{\pi}^{-1}(\boldsymbol{s}^\star_{\mathcal{S}_2})\right).$$

It suffices to show that $\widehat{\pi}=\pi_0^{-1}$, then we have

$$\mathbb{E}\ell_0\left(\widehat{\boldsymbol{s}}^{(2)},\widehat{\pi}^{-1}(\boldsymbol{s}^\star_{\mathcal{S}_2})\right)=\mathbb{E}\ell_0\left(\widehat{\boldsymbol{s}}^{(2)},\pi_0(\boldsymbol{s}^\star_{\mathcal{S}_2})\right)\leq \exp\left(-\frac{I^\star}{2}\left(1-o(1)\right)\right),$$

as desired. By definition, we have

$$\left\|\widehat{\boldsymbol{\Theta}}^{(1)}-\widehat{\boldsymbol{\Theta}}^{(2)}\mathbf{G}_{\pi_0^{-1}}\right\|_{\mathrm{F}}^2=\sum_{j\in[J]}\sum_{k\in[K]}\left|\widehat{\theta}^{(1)}_{j,k}-\widehat{\theta}^{(2)}_{j,\pi_0^{-1}(k)}\right|^2$$
$$\leq \sum_{j\in[J]}\sum_{k\in[K]}\left|\widehat{\theta}^{(1)}_{j,k}-\theta_{j,k}\right|^2+\sum_{j\in[J]}\sum_{k\in[K]}\left|\widehat{\theta}^{(2)}_{j,\pi_0^{-1}(k)}-\theta_{j,k}\right|^2\leq 2\rho KI^\star.$$

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

On the other hand, for any $\pi \in \mathfrak{S}_K$ and $\pi \neq \pi_0^{-1}$, we have

$$\left\|\widehat{\boldsymbol{\Theta}}^{(1)} - \widehat{\boldsymbol{\Theta}}^{(2)}\mathbf{G}_\pi\right\|_{\mathrm{F}}^2 = \sum_{j\in[J]}\sum_{k\in[K]}\left|\widehat{\theta}_{j,k}^{(1)} - \widehat{\theta}_{j,\pi(k)}^{(2)}\right|^2 \geq \sum_{j\in[J]}\sum_{k\in[K]}\left|\widehat{\theta}_{j,\pi(k)}^{(2)} - \theta_{j,k}\right|^2 - \rho KI^\star.$$

It suffices to show that $\sum_{j\in[J]}\sum_{k\in[K]}\left|\widehat{\theta}_{j,\pi(k)}^{(2)} - \theta_{j,k}\right|^2 > 4\rho KI^\star$. Observe that

$$\left|\widehat{\theta}_{j,\pi(k)}^{(2)} - \theta_{j,k}\right| = \left|\widehat{\theta}_{j,\pi(k)}^{(2)} - \theta_{j,\pi\circ\pi_0(k)} + \theta_{j,\pi\circ\pi_0(k)} - \theta_{j,k}\right| \geq \left|\theta_{j,\pi\circ\pi_0(k)} - \theta_{j,k}\right| - \left|\widehat{\theta}_{j,\pi(k)}^{(2)} - \theta_{j,\pi\circ\pi_0(k)}\right|.$$

Notice that $\left|\widehat{\theta}_{j,\pi(k)}^{(2)} - \theta_{j,\pi\circ\pi_0(k)}\right| = \left|\widehat{\theta}_{j,k}^{(2)} - \theta_{j,\pi_0(k)}\right| \leq \rho I^\star/J$, we have

$$\begin{aligned}
&\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\widehat{\theta}_{j,\pi(k)}^{(2)} - \theta_{j,k}\right)^2 \\
&\geq \sum_{j=1}^{J}\sum_{k=1}^{K}\left(\theta_{j,\pi\circ\pi_0(k)} - \theta_{j,k}\right)^2 - 2\sum_{j=1}^{J}\sum_{k=1}^{K}\left|\theta_{j,\pi\circ\pi_0(k)} - \theta_{j,k}\right|\left|\widehat{\theta}_{j,\pi(k)}^{(2)} - \theta_{j,\pi\circ\pi_0(k)}\right| \\
&\geq \sum_{j=1}^{J}\sum_{k=1}^{K}\left(\theta_{j,\pi\circ\pi_0(k)} - \theta_{j,k}\right)^2 - \frac{2\rho I^\star}{J}\sum_{j=1}^{J}\sum_{k=1}^{K}\left|\theta_{j,\pi\circ\pi_0(k)} - \theta_{j,k}\right| \\
&\geq \left(\sum_{k=1}^{K}\mathbb{I}\left(\pi\circ\pi_0(k)\neq k\right)\right)\left(\Delta^2 - 4C_\theta\rho I^\star\right) \geq \Delta^2,
\end{aligned}$$

where the last inequality holds due to the fact that $\pi \neq \pi_0^{-1}$, $\Delta^2 \asymp I^\star \to \infty$, and $\rho = o(1)$. The proof is thus completed by using $\rho = o(1/K)$.

B.3. Proof of Proposition 2

Notice that $\lambda_K \geq \lambda_K(\mathbf{Z})\lambda_K(\boldsymbol{\Theta})$. To get a lower bound for λ_K, it suffices to obtain lower bounds for $\lambda_K(\mathbf{Z})$ and $\lambda_K(\boldsymbol{\Theta})$. First note that

$$\mathbf{Z}^\top\mathbf{Z} = \mathrm{diag}\left(\left|\left\{i : s_i^\star = 1\right\}\right|, \ldots, \left|\left\{i : s_i^\star = K\right\}\right|\right).$$

We thus obtain $\lambda_K(\mathbf{Z}) \geq \sqrt{\alpha N/K}$. On the other hand, by Proposition 2 in Lyu et al. (2025), we obtain that for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\lambda_K\left(\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\right) \geq (1-\delta)BJK\right) \leq K\left(e^{-\delta^2B^2J/16} + e^{-3\delta BJ/8}\right).$$

Denote the above event as $\mathcal{E}_\theta$, under which we arrive at $\lambda_K(\boldsymbol{\Theta}) \geq (1-\delta/2)\sqrt{BJK}$.

To obtain a lower bound for Δ, it suffices to use Lemma S.1 in Lyu et al. (2025) to get $\Delta \geq \sqrt{2}\lambda_K(\boldsymbol{\Theta})$. We get the desired result by setting $\delta = 1/2$.

B.4. Proof of Corollary 1

The proof follows the same argument as that of Theorem 2. We only sketch the modifications here. Without loss of generality, we only consider sample data points in $\mathcal{S}_1$. First note that by Proposition 3.1 in Zhang and Zhou (2024), we have $\ell(\widetilde{\boldsymbol{s}}^{(2)}, \boldsymbol{s}^\star_{\mathcal{S}_2}) = o(1)$ with probability at least $1 - O(e^{-N\vee J})$. Hence, we get that

$$\log\left(\frac{\widehat{p}_k}{\widehat{\pi}_{s_i^\star}}\right) = \log\left(\frac{\widehat{p}_k}{p_k}\right) + \log\left(\frac{\pi_{s_i^\star}}{\widehat{\pi}_{s_i^\star}}\right) + \log\left(\frac{p_k}{\pi_{s_i^\star}}\right) \lesssim \log\left(\frac{1}{\alpha}\right) = o(I)$$

with probability at least $1 - O(e^{-N\vee J})$, where we've used Assumption 2. This indicates that the impact of this additional term is ignorable, that is,

$$\mathbb{P}\left(\log\left(\frac{\widehat{p}_k}{\widehat{\pi}_{s_i^\star}}\right) > c\delta I^\star\right) \leq O\left(e^{-N\vee J}\right)$$

for some universal constant $c > 0$. The remaining proof follows from that of Theorem 2 line by line.

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

B.5. Proof of Proposition 4

Specifically, we fix $j \in \Omega_0(k_1,k_2)$ and $k_1 \neq k_2 \in [K]$ in the following analysis. Note that

$$\begin{aligned}
&-2\log\left(\sqrt{\theta_{j,k_1}\theta_{j,k_2}}+\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}\right)\\
&=-\log\left(1-\left(\sqrt{\theta_{j,k_1}}-\sqrt{\theta_{j,k_2}}\right)^2-2\sqrt{\theta_{j,k_1}\theta_{j,k_2}}\left(1-\sqrt{\theta_{j,k_1}\theta_{j,k_2}}-\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}\right)\right)\\
&\geq\left(\sqrt{\theta_{j,k_1}}-\sqrt{\theta_{j,k_2}}\right)^2+2\sqrt{\theta_{j,k_1}\theta_{j,k_2}}\left(1-\sqrt{\theta_{j,k_1}\theta_{j,k_2}}-\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}\right).
\end{aligned}$$

When $\theta_{j,k_1}+\theta_{j,k_2}\leq 1$, we have

$$\begin{aligned}
1-\sqrt{\theta_{j,k_1}\theta_{j,k_2}}-\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}&=\frac{\left(1-\sqrt{\theta_{j,k_1}\theta_{j,k_2}}\right)^2-\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}{1-\sqrt{\theta_{j,k_1}\theta_{j,k_2}}+\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}}\\
&=\frac{\left(\sqrt{\theta_{j,k_1}}-\sqrt{\theta_{j,k_2}}\right)^2}{1-\sqrt{\theta_{j,k_1}\theta_{j,k_2}}+\sqrt{\theta_{j,k_1}\theta_{j,k_2}+1-\theta_{j,k_1}-\theta_{j,k_2}}}\\
&=C_{\theta_{j,k_1},\theta_{j,k_2}}\left(\sqrt{\theta_{j,k_1}}-\sqrt{\theta_{j,k_2}}\right)^2.
\end{aligned}$$

When $\theta_{j,k_2}+\theta_{j,k_1}>1$, we have that

$$\begin{aligned}
1-\sqrt{\theta_{j,k_1}\theta_{j,k_2}}-\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}&=\frac{\left(1-\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}\right)^2-\theta_{j,k_1}\theta_{j,k_2}}{1-\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}+\sqrt{\theta_{j,k_1}\theta_{j,k_2}}}\\
&=\frac{\left(\sqrt{1-\theta_{j,k_1}}-\sqrt{1-\theta_{j,k_2}}\right)^2}{1+\sqrt{\theta_{j,k_1}\theta_{j,k_2}}-\sqrt{\theta_{j,k_1}\theta_{j,k_2}-\left(\theta_{j,k_1}+\theta_{j,k_2}-1\right)}}\\
&=C_{\theta_{j,k_1},\theta_{j,k_2}}\left(\sqrt{1-\theta_{j,k_1}}-\sqrt{1-\theta_{j,k_2}}\right)^2.
\end{aligned}$$

Here, for any $\theta_1,\theta_2\in[0,1]$, we define

$$C_{\theta_1,\theta_2}:=\begin{cases}\left(1-\sqrt{\theta_1\theta_2}+\sqrt{\theta_1\theta_2+1-\theta_1-\theta_2}\right)^{-1}, & \theta_1+\theta_2\leq 1\\ \left(1+\sqrt{\theta_1\theta_2}-\sqrt{\theta_1\theta_2-(\theta_1+\theta_2-1)}\right)^{-1}, & \theta_1+\theta_2>1.\end{cases}$$

In addition, let $\bar{C}_{\theta_1,\theta_2}:=2C_{\theta_1,\theta_2}\sqrt{\theta_1\theta_2}$. Hence, we conclude that

$$-2\log\left(\sqrt{\theta_{j,k_1}\theta_{j,k_2}}+\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}\right)\geq\left(1+\bar{C}_{\theta_{j,k_1},\theta_{j,k_2}}\right)\mathsf{D}_2\left(\theta_{j,k_1},\theta_{j,k_2}\right),$$

where for any $\theta_1,\theta_2\in[0,1]$, we define

$$\mathsf{D}_2(\theta_1,\theta_2):=\begin{cases}\left(\sqrt{\theta_1}-\sqrt{\theta_2}\right)^2, & \theta_1+\theta_2\leq 1\\ \left(\sqrt{1-\theta_1}-\sqrt{1-\theta_2}\right)^2, & \theta_1+\theta_2>1.\end{cases}$$

By the definition of $\Omega_0(k_1,k_2)$, we have either (i) $\theta_{j,k_1}\vee\theta_{j,k_2}<1/2$ and $\theta_{j,k_1}+\theta_{j,k_2}\leq 1$ or (ii) $\theta_{j,k_1}\wedge\theta_{j,k_2}>1/2$ and $\theta_{j,k_1}+\theta_{j,k_2}>1$.

We start with case (i). Without loss of generality, we assume $\frac{1}{2}>\theta_{j,k_1}>\theta_{j,k_2}$ and denote $\theta:=\theta_2$, $\delta:=\theta_1-\theta_2=o(1)$. By definition, we have

$$\begin{aligned}
-\log\left(\sqrt{\theta_{j,k_1}\theta_{j,k_2}}+\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}\right)&=\frac{\left(1+\bar{C}_{\theta+\delta,\theta}\right)}{2}\left(\sqrt{\theta+\delta}-\sqrt{\theta}\right)^2(1+o(1))\\
&=\delta^2\frac{\left(1+\bar{C}_{\theta+\delta,\theta}\right)}{8\theta}(1+o(1))\\
&=\frac{\delta^2}{8\theta(1-\theta)}(1+o(1)).
\end{aligned}$$

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

On the other hand, we have

$$\frac{\left(\theta_{j,k_1}-\theta_{j,k_2}\right)^2}{8\left(\sigma^2_{\theta_{j,k_1}}\vee\sigma^2_{\theta_{j,k_1}}\right)}=\delta^2\frac{\log\left((1-\theta)/\theta\right)}{4(1-2\theta)}.$$

This implies that

$$\frac{-\log\left(\sqrt{\theta_{j,k_1}\theta_{j,k_2}}+\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}\right)}{\left(\theta_{j,k_1}-\theta_{j,k_2}\right)^2/\left(8\left(\sigma^2_{\theta_{j,k_1}}\vee\sigma^2_{\theta_{j,k_1}}\right)\right)}=\frac{1-2\theta}{2\theta(1-\theta)\log\left((1-\theta)/\theta\right)}(1+o(1)).$$

Define the function

$$\tau(x)=\frac{1-2x}{2x(1-x)\log\left((1-x)/x\right)},\quad g(x)=\left(-2x^2+2x-1\right)\log\left(\frac{1}{x}-1\right)-2x+1$$

for $0<x<1/2$. Some algebra gives that

$$\tau'(x)=\frac{g(x)}{2x^2(1-x)^2\log^2\left(\frac{1}{x}-1\right)}.$$

Note that

$$g'(x)=\frac{(1-2x)^2}{x(1-x)}+2(1-2x)\log\left(\frac{1}{x}-1\right)>0,$$

for $0<x<1/2$, which implies that $g(x)<\lim_{x\to\frac{1}{2}}g(x)=0$. Hence, we have $\tau'(x)<0$ and $\tau(x)>\lim_{x\to\frac{1}{2}}\tau(x)=1$. Thus, we have

$$\frac{-\log\left(\sqrt{\theta_{j,k_1}\theta_{j,k_2}}+\sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}\right)}{\left(\theta_{j,k_1}-\theta_{j,k_2}\right)^2/\left(8\left(\sigma^2_{\theta_{j,k_1}}\vee\sigma^2_{\theta_{j,k_1}}\right)\right)}=\tau(\theta)(1+o(1))>1,$$

where $\tau(\theta)$ is monotone decreasing as θ increases to $1/2$ and $\tau(\theta)\to\infty$ as $\theta\to 0$.

The analysis for case (ii) is essentially the same and hence omitted. Combining (i) and (ii), we can arrive at

$$I_{k_1,k_2}=\sum_{j\in\Omega_0(k_1,k_2)}I\left(\theta_{j,k_1},\theta_{j,k_2}\right)+\sum_{j\in\Omega_0^c(k_1,k_2)}I\left(\theta_{j,k_1},\theta_{j,k_2}\right)$$

and

$$\begin{aligned}\exp\left(-\frac{I_{k_1,k_2}}{2}\right)&=\exp\left(-\frac{1}{2}\sum_{j\in\Omega_0(k_1,k_2)}I\left(\theta_{j,k_1},\theta_{j,k_2}\right)\right)\exp\left(-\frac{1}{2}\sum_{j\in\Omega_0^c(k_1,k_2)}I\left(\theta_{j,k_1},\theta_{j,k_2}\right)\right)\\&\gtrsim\exp\left(-\sum_{j\in\Omega_0(k_1,k_2)}\Delta^2_{j,k_1,k_2}\cdot\frac{I\left(\theta_{j,k_1},\theta_{j,k_2}\right)/2}{\Delta^2_{j,k_1,k_2}}\right)\\&\gtrsim\exp\left(-\frac{\sum_{j\in[J]}\left(\theta_{j,k_1}-\theta_{j,k_2}\right)^2}{8\tilde{\sigma}^2}(1+\tau_{k_1,k_2})\right),\end{aligned}$$

where the inequality holds due to the assumption $|\Omega_0^c|=O(1)$, and

$$\tau_{k_1,k_2}:=\frac{\sum_{j\in\Omega_0(k_1,k_2)}\Delta^2_{j,k_1,k_2}\cdot\tau\left(\theta_{j,k_1}\right)}{\sum_{j\in\Omega_0(k_1,k_2)}\Delta^2_{j,k_1,k_2}}(1+o(1)).$$

B.6. Proof of Lemma 1

In light of Corollary 3.12, Remark 3.13 in Bandeira and Van Handel (2016), and Theorem 6.10 in Boucheron et al. (2013), for $t\geq 0$, we have

$$\mathbb{P}\left(\|\mathbf{E}\|\geq\frac{(1+\varepsilon)\sqrt{2}}{2}\left(\sqrt{N}+\sqrt{J}\right)+C_\varepsilon\sqrt{\log(N\wedge J)}+t\right)\leq(N\wedge J)e^{-t^2/2}.$$

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

By choosing $t = 2\sqrt{\log(N \wedge J)}$ and $\varepsilon = \sqrt{2} - 1$, for N, J large enough, we obtain that

$$\mathbb{P}\left(\|\mathbf{E}\| \geq 2\left(\sqrt{N} + \sqrt{J}\right)\right) \leq (N \wedge J)^{-1}.$$

This implies that $\lambda_{K+1}(\mathbf{R}) \leq \|\mathbf{E}\| \leq 2\left(\sqrt{J} + \sqrt{N}\right)$ with probability $1 - o(1)$. On the other hand, by assumption in Proposition 1 on $\sigma_K(\mathbb{E}\mathbf{R})$, we obtain that

$$\lambda_K(\mathbf{R}) \geq \lambda_K(\mathbb{E}\mathbf{R}) - \|\mathbf{E}\| = \omega\left(\sqrt{N} + \sqrt{J}\right)$$

with probability $1 - o(1)$. This completes the proof of Lemma 1.

B.7. Special case when there exists $\theta_{j,k}$ close to $1/2$

For $\theta < 1/2$, we have

$$\sigma_\theta^2 = \frac{1-2\theta}{2\log((1-\theta)/\theta)} = \left(\frac{1}{2} - \theta\right)\frac{1}{\log\left(1 + \frac{1}{\theta} - 2\right)} \geq \frac{\theta}{2}.$$

For $\theta > 1/2$, we have

$$\sigma_\theta^2 = \frac{2\theta-1}{2\log(\theta/(1-\theta))} = \left(\theta - \frac{1}{2}\right)\frac{1}{\log\left(1 + \frac{2\theta-1}{1-\theta}\right)} \geq \frac{1-\theta}{2}.$$

Denote $\tilde{\theta} := \arg\min_{\theta \in \{\theta_{j,k}, j \in [J], k \in [K]\}} \left|\frac{1}{2} - \theta\right|$ and $\tilde{\delta} := \left|\frac{1}{2} - \tilde{\theta}\right|$, then we simply have

$$\tilde{\sigma}^2 = \sigma_{\tilde{\theta}}^2 \geq \frac{\tilde{\theta} \wedge (1-\tilde{\theta})}{2} = \frac{1}{4}(1 - 2\tilde{\delta}).$$

We thereby have

$$\frac{1}{8\tilde{\sigma}^2}\left(\theta_{j,k_1} - \theta_{j,k_2}\right)^2 \leq \frac{1}{2\left(1-2\tilde{\delta}\right)}\left(\theta_{j,k_1} - \theta_{j,k_2}\right)^2.$$

We have $\tilde{\delta} = o(1)$ by assumption, and without loss of generality assume $\frac{1}{2} > \theta_{j,k_1} > \theta_{j,k_2}$ and denote $\theta := \theta_2$, $\delta := \theta_1 - \theta_2 = o(1)$. Then we have

$$\frac{\left(\theta_{j,k_1} - \theta_{j,k_2}\right)^2}{8\tilde{\sigma}^2} = \frac{\delta^2}{2}(1 + o(1)).$$

On the other hand, we have

$$\begin{aligned}
-\log\left(\sqrt{\theta_{j,k_1}\theta_{j,k_2}} + \sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}\right) &= \frac{\left(1 + \tilde{C}_{\theta+\delta,\theta}\right)}{2}\left(\sqrt{\theta+\delta} - \sqrt{\theta}\right)^2(1 + o(1)) \\
&= \delta^2\frac{\left(1 + \tilde{C}_{\theta+\delta,\theta}\right)}{8\theta}(1 + o(1)) \\
&= \frac{\delta^2}{8\theta(1-\theta)}(1 + o(1)),
\end{aligned}$$

where we've used the fact that $\tilde{C}_{\theta+\delta,\theta} = \frac{\theta}{1-\theta}(1 + o(1))$. This implies that

$$\frac{-\log\left(\sqrt{\theta_{j,k_1}\theta_{j,k_2}} + \sqrt{\left(1-\theta_{j,k_1}\right)\left(1-\theta_{j,k_2}\right)}\right)}{\left(\theta_{j,k_1} - \theta_{j,k_2}\right)^2/(8\tilde{\sigma}^2)} = \frac{1}{4\theta(1-\theta)}(1 + o(1)) > 1,$$

where the last inequality holds due to $\theta < 1/2$.

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

C. Proofs of lemmas

C.1. Proof of Lemma B.1

To keep notation simple in the analysis, for any $\boldsymbol{s} \in [K]^N$, we define

$$\theta_{j,k}(\boldsymbol{s}) := \frac{\sum_{i\in[N]} R_{i,j}\mathbb{I}(s_i = k)}{\sum_{i\in[N]} \mathbb{I}(s_i = k)}, \quad j \in [J], k \in [K].$$

The following result is needed, whose proof is deferred to Section C.2.

Lemma C.1. *Suppose Assumptions 1 and 2 hold. For any $\gamma_0 > 0$, there exist some absolute constants $C_1, C_2 > 0$ depending only on γ_0 such that with probability at least $1 - 2JKN^{-(2\gamma_0+1)}$,*

$$\max_{s:\ell_0(\boldsymbol{s},\boldsymbol{s}^\star)\le\alpha/(2K)} \max_{j\in[J],k\in[K]} \frac{\left|\theta_{j,k}(\boldsymbol{s}) - \theta_{j,k}(\boldsymbol{s}^\star)\right|}{\ell_0(\boldsymbol{s},\boldsymbol{s}^\star)} \le \frac{C_1 K}{\alpha}\left(1 + \tilde{\sigma}\sqrt{\frac{K\log N}{\alpha N}}\right), \tag{C.1}$$

and

$$\max_{j\in[J],k\in[K]} \left|\theta_{j,k}(\boldsymbol{s}^\star) - \theta_{j,k}\right| > C_2\tilde{\sigma}\sqrt{\frac{K\log N}{\alpha N}}, \tag{C.2}$$

where $\ell_0(\boldsymbol{s},\boldsymbol{s}') := \frac{1}{N}\sum_{i\in[N]}\mathbb{I}(s_i \neq s_i')$. As a result, for any (possibly random) $\boldsymbol{s} \in [K]^N$ such that $\ell_0(\boldsymbol{s},\boldsymbol{s}^\star) \le \alpha/(2K)$, we have

$$\max_{j\in[J],k\in[K]} \left|\theta_{j,k}(\boldsymbol{s}) - \theta_{j,k}\right| \le \ell_0(\boldsymbol{s},\boldsymbol{s}^\star)\cdot\frac{C_1 K}{\alpha}\left(1 + \tilde{\sigma}\sqrt{\frac{K\log N}{\alpha N}}\right) + C_2\tilde{\sigma}\sqrt{\frac{K\log N}{\alpha N}}$$

with probability at least $1 - 2JKN^{-(2\gamma_0+1)}$.

By Proposition 1 and Markov's inequality, we obtain that if

$$\min\left\{\frac{\Delta/\tilde{\sigma}}{K\left(1+\sqrt{\frac{J}{N}}\right)}, \frac{\lambda_K/\tilde{\sigma}}{\left(\sqrt{N}+\sqrt{J}\right)}\right\} \to \infty,$$

then for $m \in [2]$, with probability exceeding at least $1 - \exp(-(1-c_0)\Delta^2/(8\tilde{\sigma}^2))$, we have

$$\ell\left(\widehat{\boldsymbol{s}}^{(m)}, \boldsymbol{s}^\star_{\mathcal{S}_m}\right) \le \exp\left(-(1-c_0)\frac{\Delta^2}{8\tilde{\sigma}^2}\right),$$

where $\boldsymbol{s}^\star_{\mathcal{S}_m}$ is sub-vector of $\boldsymbol{s}^\star$ with the index set $\mathcal{S}_m$, and we've used that $\exp(-N/2) < \exp(-(1-c_0)\Delta^2/(8\tilde{\sigma}^2))$. Without loss of generality, we assume $\ell\left(\widehat{\boldsymbol{s}}^{(m)}, \boldsymbol{s}^\star_{\mathcal{S}_m}\right) = \ell_0\left(\widehat{\boldsymbol{s}}^{(m)}, \boldsymbol{s}^\star_{\mathcal{S}_m}\right)$. Combined with Lemma C.1, we then have

$$\max_{j\in[J],k\in[K]} \left|\widehat{\theta}^{(m)}_{j,k} - \theta_{j,k}\right| \le C_1\exp\left(-(1-c_0)\frac{\Delta^2}{8\tilde{\sigma}^2}\right) + C_2\tilde{\sigma}\sqrt{\frac{K\log N}{N}}$$

with probability at least $1 - \exp\left(-(1-c_0)\Delta^2/\left(8\tilde{\sigma}^2\right)\right) - \exp\left(-C\sqrt{N/K}\right) - N^{-(1+\gamma_0)}$, provided that there exists some universal constant $\gamma_0 > 0$ such that

$$JK \lesssim N^{\gamma_0}, \quad N \gtrsim K\log N.$$

The proof is completed by noticing that $I^\star \asymp \Delta^2/\tilde{\sigma}^2$, $JK \le \exp\left((1-c_1)\Delta^2/\left(8\tilde{\sigma}^2\right)\right)$ for some $c_1 \in (c_0, 1)$, $\tilde{\sigma} \asymp 1$, and

$$\frac{\Delta}{J^{1/2}K^{3/4}(\log N)^{1/4}/N^{1/4}} \to \infty.$$

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

C.2. Proof of Lemma C.1

We introduce the following notations:

$$\bar{\theta}_{j,k}(\boldsymbol{s}) := \frac{\sum_{i\in[N]} \theta_{i,s_i^\star} \mathbb{I}(s_i = k)}{\sum_{i\in[N]} \mathbb{I}(s_i = k)}, \quad \mathfrak{E}_{j,k}(s) := \frac{\sum_{i\in[N]} E_{i,j} \mathbb{I}(s_i = k)}{\sum_{i\in[N]} \mathbb{I}(s_i = k)}.$$

By definition, we have $\bar{\theta}_{j,k}(\boldsymbol{s}^\star) = \theta_{j,k}$ and $\mathfrak{E}_{j,k}(s) = \theta_{j,k}(\boldsymbol{s}) - \bar{\theta}_{j,k}(\boldsymbol{s})$, and hence

$$\begin{aligned}
&\left|\theta_{j,k}(\boldsymbol{s}) - \theta_{j,k}\right| \\
&= \left|\theta_{j,k}(\boldsymbol{s}) - \bar{\theta}_{j,k}(\boldsymbol{s}) + \bar{\theta}_{j,k}(\boldsymbol{s}) - \bar{\theta}_{j,k}(\boldsymbol{s}^\star) + \bar{\theta}_{j,k}(\boldsymbol{s}^\star) - \theta_{j,k}(\boldsymbol{s}^\star) + \theta_{j,k}(\boldsymbol{s}^\star) - \theta_{j,k}\right| \\
&\le \left|\mathfrak{E}_{j,k}(s) - \mathfrak{E}_{j,k}(\boldsymbol{s}^\star)\right| + \left|\bar{\theta}_{j,k}(\boldsymbol{s}) - \bar{\theta}_{j,k}(\boldsymbol{s}^\star)\right| + \left|\theta_{j,k}(\boldsymbol{s}^\star) - \theta_{j,k}\right|.
\end{aligned} \tag{C.3}$$

We first bound the term $\left|\mathfrak{E}_{j,k}(s) - \mathfrak{E}_{j,k}(\boldsymbol{s}^\star)\right|$. Notice that

$$\begin{aligned}
&\left|\mathfrak{E}_{j,k}(s) - \mathfrak{E}_{j,k}(\boldsymbol{s}^\star)\right| \\
&\le \left|\frac{\sum_{i\in[N]} E_{i,j}\mathbb{I}(s_i = k)}{\sum_{i\in[N]} \mathbb{I}(s_i = k)} - \frac{\sum_{i\in[N]} E_{i,j}\mathbb{I}(s_i^\star = k)}{\sum_{i\in[N]} \mathbb{I}(s_i = k)}\right| + \left|\frac{\sum_{i\in[N]} E_{i,j}\mathbb{I}(s_i^\star = k)}{\sum_{i\in[N]} \mathbb{I}(s_i = k)} - \frac{\sum_{i\in[N]} E_{i,j}\mathbb{I}(s_i^\star = k)}{\sum_{i\in[N]} \mathbb{I}(s_i^\star = k)}\right| \\
&\le \frac{1}{\sum_{i\in[N]} \mathbb{I}(s_i = k)} \left(\left|\sum_{i\in[N]} E_{i,j}\mathbb{I}(s_i = k, s_i^\star \ne k)\right| + \left|\sum_{i\in[N]} E_{i,j}\mathbb{I}(s_i \ne k, s_i^\star = k)\right|\right) \\
&+ \left|\frac{\sum_{i\in[N]} \left[\mathbb{I}(s_i^\star = k) - \mathbb{I}(s_i = k)\right]}{\sum_{i\in[N]} \mathbb{I}(s_i = k)}\right| \left|\frac{\sum_{i\in[N]} E_{i,j}\mathbb{I}(s_i^\star = k)}{\sum_{i\in[N]} \mathbb{I}(s_i^\star = k)}\right|.
\end{aligned} \tag{C.4}$$

Note that the following fact holds for any $\boldsymbol{s} \in [K]^N$ such that $\ell(\boldsymbol{s}, \boldsymbol{s}^\star) \le \alpha/(2K)$:

$$\sum_{i\in[N]} \mathbb{I}(s_i = k) \ge \sum_{i\in[N]} \mathbb{I}(s_i^\star = k) - \sum_{i\in[N]} \mathbb{I}(s_i^\star \ne s_i) \ge \frac{\alpha N}{K} - \frac{\alpha N}{2K} \ge \frac{\alpha N}{2K}.$$

Combined with the fact that $|E_{i,j}| \le 1$, we thus obtain the following bound for the first term in (C.4):

$$\frac{1}{\sum_{i\in[N]} \mathbb{I}(s_i = k)} \left(\left|\sum_{i\in[N]} E_{i,j}\mathbb{I}(s_i = k, s_i^\star \ne k)\right| + \left|\sum_{i\in[N]} E_{i,j}\mathbb{I}(s_i \ne k, s_i^\star = k)\right|\right) \le \frac{4K}{\alpha}\ell(\boldsymbol{s}, \boldsymbol{s}^\star).$$

Next, we bound the second term in (C.4). By General Hoeffding's inequality, we obtain that

$$\mathbb{P}\left(\left|\frac{\sum_{i\in[N]} E_{i,j}\mathbb{I}(s_i^\star = k)}{\sum_{i\in[N]} \mathbb{I}(s_i^\star = k)}\right| \ge t\right) \le 2\exp\left(-\frac{ct^2}{\bar{\sigma}^2 N_k}\right).$$

By choosing $t = \bar{\sigma}\sqrt{CN_k \log N}$ for some sufficiently large $C > 0$, we have that

$$\left|\frac{\sum_{i\in[N]} E_{i,j}\mathbb{I}(s_i^\star = k)}{\sum_{i\in[N]} \mathbb{I}(s_i^\star = k)}\right| \le C\bar{\sigma}\sqrt{\frac{K\log N}{\alpha N}}$$

with probability exceeding $1 - N^{-2\gamma_0 - 1}$ for some large constant $\gamma_0 > 0$. In addition, we have

$$\begin{aligned}
\left|\frac{\sum_{i\in[N]} \left[\mathbb{I}(s_i^\star = k) - \mathbb{I}(s_i = k)\right]}{\sum_{i\in[N]} \mathbb{I}(s_i = k)}\right| &\le \frac{2K}{\alpha N}\left(\left|\sum_{i\in[N]} \mathbb{I}(s_i^\star = k, s_i \ne k)\right| + \left|\sum_{i\in[N]} \mathbb{I}(s_i^\star \ne k, s_i = k)\right|\right) \\
&\lesssim \frac{K}{\alpha}\ell(\boldsymbol{s}, \boldsymbol{s}^\star).
\end{aligned}$$

From (C.4) and a union bound over $j \in [J]$ and $k \in [K]$, we arrive at

$$\max_{j\in[J], k\in[K]} \left|\mathfrak{E}_{j,k}(s) - \mathfrak{E}_{j,k}(\boldsymbol{s}^\star)\right| \lesssim \alpha^{-1} K \ell(\boldsymbol{s}, \boldsymbol{s}^\star)\left(1 + \bar{\sigma}\sqrt{\frac{K\log N}{\alpha N}}\right)$$

with probability at least $1 - JKN^{-2\gamma_0 - 1}$.

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press

Next, we bound the term $\left|\tilde{\theta}_{j,k}(\boldsymbol{s}) - \tilde{\theta}_{j,k}(\boldsymbol{s}^\star)\right|$. Notice that by definition $\tilde{\theta}_{j,k}(\boldsymbol{s}^\star) = \theta_{j,k}$, hence

$$\left|\tilde{\theta}_{j,k}(\boldsymbol{s}) - \tilde{\theta}_{j,k}(\boldsymbol{s}^\star)\right| = \left|\frac{\sum_{i\in[N]}\sum_{k'\in[K]}(\theta_{i,k}-\theta_{i,k'})\mathbb{I}(s_i = k, s_i^\star = k')}{\sum_{i\in[N]}\mathbb{I}(s_i = k)}\right| \le \frac{2K}{\alpha}\ell(\boldsymbol{s},\boldsymbol{s}^\star).$$

It suffices to bound $\left|\theta_{j,k}(\boldsymbol{s}^\star) - \theta_{j,k}\right|$ by

$$\max_{j\in[J],k\in[K]}\left|\theta_{j,k}(\boldsymbol{s}^\star) - \theta_{j,k}\right| = \max_{j\in[J],k\in[K]}\left|\frac{\sum_{i\in[N]}E_{i,j}\mathbb{I}(s_i^\star = k)}{\sum_{i\in[N]}\mathbb{I}(s_i^\star = k)}\right| \le C\tilde{\sigma}\sqrt{\frac{K\log N}{\alpha N}}$$

with probability exceeding $1 - JKN^{-2\gamma_0-1}$.

Collecting (C.3) and the bounds for $\left|\mathfrak{E}_{j,k}(s) - \mathfrak{E}_{j,k}(\boldsymbol{s}^\star)\right|$, $\left|\tilde{\theta}_{j,k}(\boldsymbol{s}) - \tilde{\theta}_{j,k}(\boldsymbol{s}^\star)\right|$, and $\left|\theta_{j,k}(\boldsymbol{s}^\star) - \theta_{j,k}\right|$, we can complete the proof.

Cite this article: Lyu, Z. and Gu, Y. (2026). Spectral Clustering with Likelihood Refinement for High-Dimensional Latent Class Recovery. *Psychometrika*, 695–723. https://doi.org/10.1017/psy.2026.10095

https://doi.org/10.1017/psy.2026.10095 Published online by Cambridge University Press