Covariance Projected Spectral Method: Exact Recovery and Hypothesis Test of Latent Class in High Dimension

1 Introduction

This paper focuses on the exact recovery and hypothesis testing of unobserved heterogeneous subgroups from high-dimensional categorical data. We consider the Latent Class Model (LCM, Goodman, 1974) that model multivariate dependent categorical responses which has wide applications in social, behavioral, and biomedical sciences. In LCMs, each subject belongs to an unobserved latent class. Subjects in the same latent class share homogeneity, whereas subjects from different latent classes have heterogeneous behaviors in the observed multivariate responses.

We focus on the general LCM in the modern large-scale and high-dimensional setting featuring a large number of n subjects responding to a large number of J items. This is the so-called *double-asymptotic regime*. For an integer M, denote $[M] = \{1, \ldots, M\}$. To be specific, consider that for each subject $i \in [n]$, there is a response vector $\mathbf{R}_i = (R_{i,1}, \ldots, R_{i,J}) \in \{0,1\}^J$ and a discrete latent class variable $z_i \in [K]$. The n response vectors $\mathbf{R}_1, \ldots, \mathbf{R}_n$ in the sample are independent and the generative process of each $\mathbf{R}_i$ is as follows. First, the latent class variable z_i follows a categorical distribution with $\mathbb{P}(z_i = k) = \pi_k^*$ where $\sum_{k=1}^K \pi_k^* = 1$. Then given the latent class variable $z_i = k$, the responses to J items are drawn from the conditional distribution distribution which is distinct between different classes. For example, in the case of binary responses, we can denote $\mathbb{P}(R_{i,j} = 1 \mid z_i = k) = \theta_{j,k}$. Collect all the conditional response probabilities $\theta_{j,k}$ in a $J \times K$ item paramter matrix $\mathbf{\Theta} \in \mathbb{R}^{J \times K}$.

And we define the indicator vector $\mathbf{Z}_i$ by letting the k th component be 1 and other components be zeroes when $z_i = k$. We also denote the number of subjects belonging to the kth latent class by n_k, its indices set by $C_k = \{i \in [n] : z_i = k\}$, and its sample proportion by $\pi_k = n_k/n$. Denote the $n \times K$ membership matrix by $\mathbf{Z} = (\mathbf{Z}_1^\top, \ldots, \mathbf{Z}_n^\top)^\top$, and the $n \times K$ response matrix by $\mathbf{R} = (\mathbf{R}_1^\top, \ldots, \mathbf{R}_n^\top)^\top$ where $R_{i,j} = 1$ We could write the expectation of $\mathbf{R}$ to be:

$$\mathbf{R}^* := \mathbb{E}[\mathbf{R}] = \mathbf{Z}\mathbf{\Theta}^\top \tag{1}$$

Next we give several different general examples of LCM that we will consider.

1. **LCM for Polytomous Responses.** In this setting, the response range of the j-th item is $[m_j] = \{1, 2, \ldots, m_j\}$ instead of $\{0, 1\}$, that is, we assume the response matrix is $\widetilde{\mathbf{R}} = \left(\widetilde{\mathbf{R}}_1, \ldots, \widetilde{\mathbf{R}}_n\right)^\top \in \mathbb{R}^{n \times J}$ where $\widetilde{R}_{i,j}$ takes value in $[m_j]$. specifically, if $m = m_i$ holds for all $j \in [J]$, the data matrix could be easily transformed into a tensor $\mathbf{R}^{\mathbf{tensor}} \in \mathbb{R}^{n \times J \times m}$. However, we do not want to go too far in the form of tensor since we introduce no order assumptions among responses with different levels. Conversely, we flatten the data matrix $\widetilde{\mathbf{R}}$ into $\mathbf{R}$ where $\mathbf{R}_{i, \sum_{k=1}^{j-1} m_k + c} = 1$ if and only if $\widetilde{R}_{i,j} = c$. In this way, the expectation of the flattened matrix $\mathbf{R}$ could be similarly expressed as

$$\mathbf{R}^* := \mathbb{E}[\mathbf{R}] = \mathbf{Z}\mathbf{\Theta}^\top \tag{2}$$

   where $\mathbf{\Theta} \in \mathbb{R}^{\sum_{i \in [J]} m_i \times K}$ and $\mathbf{\Theta}_{\sum_{k=1}^{j-1} m_k + c, k} = \mathbb{P}[R_{i,j} = c \mid z_i = k]$.

2. **Locally Dependent LCM.** Locally dependent LCM describes the LCM equipped with dependence structure in the local coordinates, such as Reboussin et al. (2008); Bowers and Culpepper (2022). And we can further assume that the dependence structure exists locally, such as

   (a) **Ising-type Graphical LCM.** Analogue to the FLaG-IRT model proposed in Chen et al. (2018), consider the Ising-type graphical LCM with quadratic interaction effect between different items. The conditional distribution is assumed to be

$$\mathbb{P}[\mathbf{R}_i = \boldsymbol{r} \mid z_i = k] \propto \exp\left((2\boldsymbol{r} - \mathbf{1})^\top \mathbf{A}_k (2\boldsymbol{r} - \mathbf{1}) + \boldsymbol{\beta}^\top \boldsymbol{r}\right) \tag{3}$$

   where $\mathbf{A}_k$ is a diagonal-free symmetric matrix.

   (b) **Discretized LCM from Underlying Multivariate Continuous Vector.** Another way to model the latent dependence structure is to assume that the binary responses are obtained from dichomatizing a underlying continuous random vector such as a multivariate Gaussian distributed vector (Zeng et al., 2022).

2 Contributions

In this paper, we consider the high dimensional latent class model with possibly locally dependent structure. In the challenging double-asymptotic regime where both the sample size n and the number of items J go to infinity, we make the following theoretical and methodological contributions.

1. **Clustering Error Risk Upper Bound in Weighted Spectral Clustering & Exact Recovery Guarantee of Latent Classes from High-dimensional Categorical Data**: We first derive the risk upper bound of misclustering in the weighted spectral clustering, by leveraging the singular subspace perturbation techniques in Zhang and Zhou (2022). Our risk upper bound also implies the exact recovery of the latent classes in high-dimensional categorical data.

2. **Asymptotic Normality of Singular Subspace Embeddings**: Secondly, we establish the row-wise asymptotic normality of the scaled left singular matrix, $\mathbf{UD}$, in the rank-K SVD under some regularity conditions. This result inspires us to propose the following estimation method and hypothesis testing procedure.

3. **Projected Covariance-Adjusted Weighted Spectral Clustering (PCW-Clust, or Covariance Projected Spectral Method)**: Taking advantage of the asymptotic normality of $\mathbf{UD}$ under high-dimensional regime, we propose the Projected Covariance Adjustment Method. This method significantly outperforms the traditional spectral clustering methods in the heteroskedastic noise setting.

4. **Post-Projection Hypothesis Testing and Uncertainty Quantification**: Given estimation results from our spectral method, we propose a hypothesis testing procedure to test whether a new data point belongs to a certain latent class.

5. **Tight $\ell_{2,\infty}$ Left Singular Space Perturbation Upper Bound for Polytomous LCMs:** Additionally, for the *Polytomous LCM*s, we establish a tight $\ell_{2,\infty}$ perturbation upper bound for the left singular subspace using the leave-two-out technique in Cai et al. (2021) and developed the theory of exact recovery guarantees.

3 Spectral Clustering Methods

3.1 Notations

Throughout the paper we denote by $[n]$ the set $\{1, \dots, n\}$. For any matrix $\mathbf{M}$, let $\mathbf{M}_k$ denote the $m-$th row of $\mathbf{M}$, $\mathbf{M}_{,k}$ denote the k-th column of $\mathbf{M}$, $\|M\|$ denote its spectral norm and $\|M\|_F$ denote its Frobenius norm. We also denote $d_{2,\infty}(\mathbf{U}, \widehat{\mathbf{U}})$ by the $\ell_{2,\infty}$ distance of the singular subspaces $\min_{H \in O^K} \left\|\mathbf{U}\mathbf{H} - \widehat{\mathbf{U}}\right\|_{2,\infty}$. Further, we denote by $\lambda_k(\mathbf{M})$ the k-th largest singular value of $\mathbf{M}$ and by λ_{min} the smallest one.

Furthermore, for any real valued function $f(n, J, K)$ and $g(n, J, K)$, we will write $f(n, J, K) \lesssim g(n, J, K)$ if $|f(n, J, K)| \leq C\,|g(n, J, K)|$ for some constant C. We also write $f(n, J, K) \ll g(n, J, K)$ when there exist some sufficiently small constat c such that $|f(n, J, K)| \leq c\,|g(n, J, K)|$ for all sufficiently large n and J. Finally, we write $f(n, J, K) = o(1)g(n, J, K)$ if $\frac{|f(n,J,K)|}{|g(n,J,K)|} \to 0$ as K is fixed and n, and J go to infinity simultaneously.

3.2 Spectral Clustering Methods

Spectral Clustering methods have been well studied in the literature of mixture model Kannan et al. (2009),Kumar and Kannan (2010), and the LCMs we consider is actaully a case of sub-gaussian mixture model. The elementary idea of recovering the membership is to make use of the information contained in the left singular matrix. Taking the binary outcome LCM for example, we suppose that the top-K SVD of $\mathbf{R}^*$ is $(\mathbf{U}, \mathbf{D}, \mathbf{V})$ where $\mathbf{U} \in \mathbb{R}^{n \times K}$, $\mathbf{V} \in \mathbb{R}^{J \times K}$, and $\mathbf{D} = diag(\sigma_1^*, \cdots, \sigma_L^*)$. Similarly, we suppose that the empirical top-K SVD of $\mathbf{R}$ is $\left(\widehat{\mathbf{U}}, \widehat{\mathbf{D}}, \widehat{\mathbf{V}}\right)$. Then there exists an invertible matrix $\mathbf{O}$ satisfying $\mathbf{U} = \mathbf{Z}\mathbf{O}$ in the population sense. Hence $\mathbf{U}_i^* = \mathbf{U}_j^*$ holds iff $z_i = z_j$. Therefore computing K-means clustering to the empirical singular matrix $\widehat{\mathbf{U}}$ could provide us the latent class information in a way. The ordinary spectral clustering based on the K-means is restated in Algorithm 1.

As pointed out in Löffler et al. (2019), the importance of singular vectors is different. To capture the information in the singular values, the idea in Zhang and Zhou (2022),Löffler et al. (2019) is to weight the empirical singular vectors with the corresponding singular values. The procedure of *Weighted Spectral Clustering* is stated in Algorithm 2.

Algorithm 1: Spectral Clustering

Input: Data matrix $\mathbf{R} = (\mathbf{R}_1, \ldots, \mathbf{R}_n) \in \mathbb{R}^{n \times J}$, number of latent class K

Output: Cluster assignment vector $\widehat{\mathbf{z}} \in [K]^n$

Perform the top-K SVD on $\mathbf{R}$ to have $\left(\widehat{\mathbf{U}}, \widehat{\mathbf{D}}, \widehat{\mathbf{V}}\right)$ where $\widehat{\mathbf{U}} \in \mathbb{R}^{n \times K}, \widehat{\mathbf{V}} \in \mathbb{R}^{J \times K}$, and $\widehat{\mathbf{D}} = \operatorname{diag}(\widehat{\sigma}_1, \ldots, \widehat{\sigma}_K)$;

Perform K-means on the rows of $\widehat{\mathbf{U}}$, that is

$$\left(\widehat{\mathbf{z}},\ \{\widehat{\mathbf{c}}_i\}_{i \in [K]}\right) = \underset{\mathbf{z} \in [K]^n, \{\mathbf{c}_i\}_{i \in [K]} \in \mathbb{R}^K}{\arg\min} \sum_{i \in [n]} \left\| \widehat{\mathbf{U}}_i - \mathbf{c}_{z_i} \right\|_2^2 \tag{4}$$

Algorithm 2: Weighted Spectral Clustering

Input: Data matrix $\mathbf{R} = \left(\mathbf{R}_1^\top, \ldots, \mathbf{R}_n^\top\right)^\top \in \mathbb{R}^{n \times J}$, number of latent class K

Output: Cluster assignment vector $\widehat{\mathbf{z}} \in [K]^n$

Perform the top-K SVD on $\mathbf{R}$ to have $\left(\widehat{\mathbf{U}}, \widehat{\mathbf{D}}, \widehat{\mathbf{V}}\right)$ where $\widehat{\mathbf{U}} \in \mathbb{R}^{n \times K}, \widehat{\mathbf{V}} \in \mathbb{R}^{J \times K}$, and $\widehat{\mathbf{D}} = \operatorname{diag}(\widehat{\sigma}_1, \ldots, \widehat{\sigma}_K)$;

Perform K-means on the rows of $\widehat{\mathbf{U}}$, that is

$$\left(\widehat{\mathbf{z}},\ \{\widehat{\mathbf{c}}_i\}_{i \in [K]}\right) = \underset{\mathbf{z} \in [K]^n, \{\mathbf{c}_1, \ldots, \mathbf{c}_K\} \in \mathbb{R}^K}{\arg\min} \sum_{i \in [n]} \left\| \widehat{\mathbf{U}}_i \widehat{\mathbf{D}} - \mathbf{c}_{z_i} \right\|_2^2 \tag{5}$$

4 Risk Upper Bound for Weighted Spectral Clustering and Exact Recovery Guarantees

Compared with controlling the error rate with a high probability, an upper bound for the error rate of clustering is more meaningful since it not only provides the partial recovery guarantees but also gives the exact recovery guarantees. With more delicate analysis, Zhang and Zhou (2022) recently gave a novel error risk upper bound for *weighted spectral clustering.*

Leveraging Theorem 3.1 in Zhang and Zhou (2022) and Lemma 11, we can easily give an upper bound for $\mathbb{E}[l(\widehat{z}, z)]$ where $l(\widehat{z}, z) = \min_{\phi \in perm(K)} \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\{\widehat{z}_i = \phi(z_i)\}$ when applying the weighted spectral clustering method.

Corollary 1. *Consider the weighted spectral clustering algorithm mentioned above in the polytomous LCM setting. Assume $\beta n / K^2 \geq 10$ where $\beta = \frac{\min_{i \in [K]} n_i}{\max_{j \in [K]} n_j}$. There exist constants $C, C' > 0$ such that under the assumption that:*

$$\psi_1 := \frac{\triangle}{\beta^{-0.5} K \left(1 + \sqrt{\frac{\sum_{i=1}^{J} m_i}{n}}\right) \bar{\sigma}} > C \tag{6}$$

where $\triangle := \min_{i \neq j \in [K]} \|\boldsymbol{\Theta}_{,i} - \boldsymbol{\Theta}_{,j}\|_2$ and

$$\begin{aligned}
&\rho_1 := \frac{\sigma_K^*}{(\sqrt{n} + \sqrt{\sum_{i=1}^{J} m_i})\bar{\sigma}} > C \\
&\bar{\sigma} = \min\left\{ \max_{k \in [K], j \in [J]} \frac{K\left(\max_{\sum_{l=1}^{j-1} m_l + 1 \leq h \leq \sum_{l=1}^{j} m_l} \Theta_{h,k}\right)^{1/2}}{1 - \max_{\sum_{l=1}^{j-1} m_l + 1 \leq h \leq \sum_{l=1}^{j} m_l} \Theta_{h,k}}, \max_{j \in [J], k \in [K]} \sqrt{m_j} \max_{c \in [m_j]} K(\Theta_{\sum_{i=1}^{j-1} m_i + c, k})^{1/2} \right\} \\
&K(p) = \begin{cases} 0 & p \in \{0, 1\} \\ 1/4 & p = \frac{1}{2} \\ \frac{p - q}{2(\log p - \log q)} & p \in (0, 1) \setminus \{\frac{1}{2}\} \end{cases}
\end{aligned} \tag{7}$$

we have:

$$\mathbb{E} l(\widehat{\mathbf{z}}, \mathbf{z}) \leq \exp\left(-\left(1 - C'\left(\psi_1^{-1} + \rho_1^{-2}\right)\right) \frac{\triangle^2}{8\bar{\sigma}^2}\right) + \exp\left(-\frac{n}{2}\right) \tag{8}$$

Remark. *To achieve exact recovery with probability at least $1 - n^{-\alpha}$, we require that $\mathbb{E} l(\widehat{z}, z) \leq n^{-(1+\alpha)}$. Assume that $m, K, \beta, R^*_{i,j} \asymp 1, J \lesssim n$, To satify $\mathbb{E} l(\widehat{z}, z) \leq n^{-(1+\alpha)}$, we require the* *following holds:*

$$\triangle \gtrsim \sqrt{\log n} \tag{9}$$

4.1 The relationship between singular values and $\triangle$, $\bar{\triangle}$

To better illustrate the conditions on $\triangle$ and σ_K^*, we also denote the noise matrix by $\mathbf{E} = \mathbf{R} - \mathbf{R}^* \in \mathbb{R}^{n \times \sum_{j=1}^J m_j}$, $\mathrm{diag}(\pi_1, \ldots, \pi_K)$ by $\mathbf{G}$, $\mathrm{diag}(\pi_1^*, \ldots, \pi_K^*)$ by $\mathbf{G}^*$, the SVD of $\mathbf{G}^{1/2}\mathbf{\Theta}^\top$ by $(\mathbf{L}, \mathbf{J}, \mathbf{R})$, and the smallest singular value of $\mathbf{G}^{1/2}\mathbf{\Theta}$ be σ_{min}. Note that

$$\mathbf{R}^* = \mathbf{Z}\mathbf{\Theta}^\top = \sqrt{n}\frac{1}{\sqrt{n}}\mathbf{Z}\mathbf{G}^{-1/2}\mathbf{G}^{1/2}\mathbf{\Theta}^\top = \sqrt{n}\left(\frac{1}{\sqrt{n}}\mathbf{Z}\mathbf{G}^{-1/2}\right)\mathbf{L}\mathbf{J}\mathbf{R}^\top \tag{10}$$

where columns of $\frac{1}{\sqrt{n}}\mathbf{Z}\mathbf{G}^{-1/2}$ is unit-normed and orthogonal with each others. Therefore, $\frac{1}{\sqrt{n}}\mathbf{Z}\mathbf{G}^{-1/2}\mathbf{L}$ is actually equal to the left singular matrix $\mathbf{U}$ of $\mathbf{R}^*$. And we also have

$$\mathbf{D}^* = \sqrt{n}\mathbf{J} \tag{11}$$

Then we have the following observation

Lemma 1. *For the LCMs we considered above, it holds that*

$$\kappa\sigma_{min}^* \max_{i,j:z_i \neq z_j} \sqrt{\frac{1}{\pi_i} + \frac{1}{\pi_j}} \geq \triangle \geq \sigma_{min}^* \min_{i,j:z_i \neq z_j} \sqrt{\frac{1}{\pi_i} + \frac{1}{\pi_j}} \tag{12}$$

where $\triangle := \min_{i \neq j \in [K]} \|\mathbf{\Theta}_{,i} - \mathbf{\Theta}_{,j}\|_2$, $\bar{\triangle} := \max_{i \neq j \in [K]} \|\mathbf{\Theta}_{,i} - \mathbf{\Theta}_{,j}\|_2$

When the proportion parameters π are fixed, we have

$$\sigma_{min}^* \lesssim \triangle \leq \bar{\triangle} \lesssim \kappa\sigma_{min}^* \tag{13}$$

5 Row-wise Asymptotic Normality of Singular Space

It's well known that the information distance of a sample away from its oracle center could be measured by its log-likelihood under Gaussian noise assumption Chen and Zhang (2021). However, under the existing difficulties in computing the log-likelihood of binary outcome model, we turn to ask how matrix $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{U}}\widehat{\mathbf{D}}$ behave under some regularity conditions when noises are restricted to some non-Gaussian type.

Recently, Fan et al. (2022) gave a delicate description to the asymptotic properties of eigenvectors for the generalized Wigner matrix equipped with low-rank structure. Taking convenience of their result, we could investigate the asymptotic behavior of $\widehat{\mathbf{U}}\widehat{\mathbf{D}}$ under independent noise assumption.

Firstly we introduce some regularity condition consistent with those in Fan et al. (2022):

Condition 1. *Assume that*

$$\alpha_n^2 = \max_{k\in[K]} \sum_{j\in[J]} \theta_{j,k}(1-\theta_{j,k}) \to \infty$$

as $n + J \to \infty$

Condition 2. *There exists a positive constant* $c_0 < 1$ *such that* $\min\{\sigma_i/\sigma_j : 1 \le i < j \le K+1\} \ge 1 + c_0$. *Additionally, either of the following conditions holds:*

1. $\frac{|\sigma_K|}{(n+J)^\epsilon \alpha_n} \to \infty$ *with some small positive constant* ϵ.
2. $\max_{k\in[K],j\in[J]} \theta_{j,k}(1-\theta_{j,k}) \le (c_1^2\alpha_n^2)/(n+J)$ *and* $|\sigma_K| > c\alpha_n \log(n+J)$ *with some constants* $c_1 \ge 1$ *and* $c > 4c_1\left(1+\frac{c_0}{2}\right)$.

Theorem 1. *Assume the noise matrix* $\mathbf{E}$ *is entrywise independent and* Condition 1,2 *hold, considering the singular vector* $\widehat{\mathbf{u}}_k$ *such that* $\widehat{\mathbf{u}}_k^T\mathbf{u}_k \ge 0$ *for* $k \in [K]$ *we have the following properties:*

1. *For arbitrary* $k \in [K]$, *if* $\boldsymbol{var}(\mathbf{u}_k^\top \mathbf{E}\mathbf{v}_k) \gg \alpha_n^2\sigma_k^{-2}$, $\boldsymbol{var}(\mathbf{u}_k^\top \mathbf{E}\mathbf{v}_k) \ge c_0$, *and* $|\mathbf{v}|_\infty \ll 1$, *it has:*
$$\frac{\widehat{\sigma}_k - t_k}{[\boldsymbol{var}(\mathbf{u}_k^\top \mathbf{E}\mathbf{v}_k)]^{1/2}} \overset{d}{\Longrightarrow} N(0,1) \tag{14}$$
where t_k *is defined below and* c_0 *is a positive constant.*

2. *Fixing an arbitrary* $i \in [n]$, *for the vector* $\mathbf{o}_i$ *whose i-th coordinate takes* 1 *and others take 0 and the vector* $\bar{\mathbf{o}}_{z_i} = \frac{\sum_{j\in[n]} \mathbf{1}\{z_j = z_i\} o_j}{\sum_{j\in[n]} \mathbf{1}\{z_j = z_i\}}$, *it holds for arbitrary* $k \in [K]$ *that:*
$$\widehat{\sigma}_k \left(\mathbf{o}_i - \bar{\mathbf{o}}_{z_i}\right)^\top \widehat{\mathbf{u}}_k = \left(\mathbf{o}_i - \bar{\mathbf{o}}_{z_i}\right)^\top \mathbf{E}\mathbf{v}_k + O_p(\alpha_n/t_k) \tag{15}$$

where t_k is defined as the solution to equation:

$$f_k(z) = 1 + \sigma_k \left\{ \mathcal{R}(\mathbf{t}_k, \mathbf{t}_k, z) - \mathcal{R}(\mathbf{t}_k, \mathbf{T}_{-k}, z) \left[(\mathbf{D}_{-k})^{-1} + \mathcal{R}(\mathbf{T}_{-k}, \mathbf{T}_{-k}, z) \right]^{-1} \mathcal{R}(\mathbf{T}_{-k}, \mathbf{t}_k, z) \right\} = 0 \tag{16}$$

$$\mathbf{t}_i = \frac{1}{\sqrt{2}} \left(\mathbf{u}_i^\top, \mathbf{v}_i^\top \right)^\top \text{ for } 1 \le i \le K, \ \mathbf{t}_i = \frac{1}{\sqrt{2}} \left(-\mathbf{u}_i^\top, \mathbf{v}_i^\top \right)^\top \text{ for } K+1 \le i \le 2K \tag{17}$$

$$\mathbf{T}_{-k} = (\mathbf{t}_1, \dots, \mathbf{t}_{k-1}, \mathbf{t}_{k+1}, \dots, \mathbf{t}_{2K}) \tag{18}$$

$$\widetilde{\mathbf{E}} = \begin{pmatrix} & \mathbf{E} \\ \mathbf{E}^\top & \end{pmatrix} \tag{19}$$

$$\mathcal{R}(\mathbf{M}_1, \mathbf{M}_2, t) = - \sum_{l=0, l \neq 1, K}^{L} t^{-(l+1)} \mathbf{M}_1^\top \mathbb{E} \widetilde{\mathbf{E}}^l \mathbf{M}_2 \tag{20}$$

where L in (20) is some sufficiently large integer speciefied in (66) of Fan et al. (2022).

3. *Furthermore, if we additionally assume that $\max_{k\in[K]} \left\| \left(\mathbf{V}^\top \mathbf{\Sigma}_k \mathbf{V} \right)^{-1/2} \right\| \lesssim 1$ and $\min \sum_{i \in [n]} \mathbf{1}\{z_i = k\} \to \infty$, then we have:*

$$\begin{aligned} &\left(\mathbf{R}_i \widehat{\mathbf{V}} - \frac{\sum_{j\in[n]} 1\{z_j = z_i\} \mathbf{R}_j \widehat{\mathbf{V}}}{\sum_{j\in[n]} 1\{z_j = z_i\}} \right) \left(\mathbf{V}^\top \mathbf{\Sigma}_{z_i} \mathbf{V} \right)^{-1/2} \\ &= \left(\widehat{\mathbf{U}}_i \widehat{\mathbf{D}} - \widehat{\theta}(z)_{z_i}^\top \widehat{\mathbf{V}} \right) \left(\mathbf{V}^\top \mathbf{\Sigma}_{z_i} \mathbf{V} \right)^{-1/2} \overset{d}{\Rightarrow} N\left(0, \mathbf{I}_k\right) \end{aligned} \tag{21}$$

as J goes to infinity with n is fixed or increasing either.

4. *With the assumptions in 3, if $\frac{\delta_e}{\min_{k\in[L]} \left\| \mathbf{V}^\top \mathbf{\Sigma}_k \mathbf{V} \right\|^{1/2}} = o_p(1)$, it has:*

$$\left(\mathbf{R}_i \widehat{\mathbf{V}} - \widehat{\theta}^\top \widehat{\mathbf{V}} \right) \left(\mathbf{V}^\top \mathbf{\Sigma}_{z_i} \mathbf{V} \right)^{-1/2} = \left(\widehat{\mathbf{U}}_i \widehat{\mathbf{D}} - \widehat{\theta}^\top \widehat{\mathbf{V}} \right) \left(\mathbf{V}^\top \mathbf{\Sigma}_{z_i} \mathbf{V} \right)^{-1/2} \overset{d}{\Rightarrow} N\left(0, \mathbf{I}_K\right) \tag{22}$$

as J goes to infinity with n is fixed or increasing either.

5. *With the assumptions in 4, if $\frac{\delta_s}{\min_{k\in[L]} \left\| \mathbf{V}^\top \mathbf{\Sigma}_k \mathbf{V} \right\|} = o_p(1)$ holds additionally, it has:*

$$\left(\mathbf{R}_i \widehat{\mathbf{V}} - \widehat{\theta}^\top \widehat{\mathbf{V}} \right) \left(\widehat{\mathbf{S}}_{z_i} \right)^{-1/2} = \left(\widehat{\mathbf{U}}_i \widehat{\mathbf{D}} - \widehat{\theta}^\top \widehat{\mathbf{V}} \right) \left(\widehat{\mathbf{S}}_{z_i} \right)^{-1/2} \overset{d}{\Rightarrow} N\left(0, \mathbf{I}_K\right) \tag{23}$$

The proof is in Appendix A.

6 Projected Covariance-adjusted Weighted Clustering Method

Heteroskedastic noises in different subgroups is a common setting in the real data application, that is, the covariance matrices of multivariate distributions $\mathcal{E}$ differ significantly for different latent classes. In the literature of Gaussian mixture models, some theory has been established recently Chen and Zhang (2021). However, beyond the perspective of Gaussian mixture model, we point out that the sub-gaussian model especially the binary outcome case is more complicated than gaussian one. Firstly, the information of the noise distribution can not be completely captured by the covariance in the sub-gaussian case. Secondly, the full-rank assumption to the covariance matrix is usually unrealistic, particularly in the polytomous LCMs. Thirdly, without introducing sparsity assumption, it is hard to accurately estimate the covariance matrix in the high-dimensional regime.

In this section we proposed a Projected Covariance-adjusted Weighted Clustering (PCW-Clust) method to deal with the challenging cases where noise suffers from severe heteroscedasticity or dependence. Taking the advantages of the results in the above section, we are able to turn our focus to the linear subspace spanned by the centers of latent classes. Instead of estimating the complete covariance matrix, we are interested in capturing the covariance information in the K-dimensional subspace and adjust our distance measurement.

We assume that

$$R_{i,} = \Theta_{,z_i} + e_i, \quad e_i \sim \mathcal{E}_{z_i}, \quad \mathrm{Cov}(\mathcal{E}_{z_i}) = \Sigma_{z_i} \tag{24}$$

6.1 Algorithm

We proposed a one-step spectral clustering algorithm 3 named *Projected Covariance-adjusted Weighted Clustering (PCW-Clust).*

Algorithm 3: Projected Covariance-adjusted Weighted Clustering (**PCW-Clust**)

Input: Data matrix $\mathbf{R} = \left(\mathbf{R}_1^\top, \ldots, \mathbf{R}_n^\top\right)^\top \in \mathbb{R}^{n \times J}$, number of latent class K, an initial cluster estimate $\widehat{\mathbf{z}}^{(0)}$

Output: Cluster assignment vector $\widehat{\mathbf{z}} \in [K]^n$

1 Given an initial estimate $\{\widehat{z}_i^{(0)}\}$, estimate the centers $\{\widehat{\theta}_k^{(0)}\}$ by

$$\widehat{\theta}_k^{(0)} = \frac{\sum_{i \in [n]} \mathbf{1}\left\{\widehat{z}_i^{(0)} = k\right\} \mathbf{R}_i}{\sum_{i \in [n]} \mathbf{1}\left\{\widehat{z}_i^{(0)} = k\right\}}, \ k \in [K] \tag{25}$$

For each $k \in [K]$, compute the statistic:

$$\widehat{\mathbf{S}}_k := \frac{\sum_{i \in c_k} \widehat{\mathbf{V}}^\top (\mathbf{R}_i - \widehat{\theta}_k^{(0)})^\top (\mathbf{R}_i - \widehat{\theta}_k^{(0)}) \widehat{\mathbf{V}}}{\sum_{i \in [n]} \mathbf{1}\left\{\widehat{z}_i^{(0)} = k\right\}} = \widehat{\mathbf{V}}^\top \widehat{\mathbf{\Sigma}}_k \widehat{\mathbf{V}} \tag{26}$$

Estimate the cluster

$$\widehat{z}_i^{(1)} = \min_{k \in [K]} (\mathbf{R}_i - \widehat{\theta}_k^{(0)})^\top \widehat{\mathbf{V}} \widehat{\mathbf{S}}_k^{-1} \widehat{\mathbf{V}}^\top (\mathbf{R}_i - \widehat{\theta}_k^{(0)}) \tag{27}$$

which is equivalent with

$$\widehat{z}_i^{(1)} = \min_{k \in [K]} \left((\mathbf{R}_i - \widehat{\theta}_k^{(0)})^\top \widehat{\mathbf{V}} \right) \left(\widehat{\mathbf{V}}^\top \widehat{\mathbf{\Sigma}}_k \widehat{\mathbf{V}} \right)^{-1} \left(\widehat{\mathbf{V}}^\top (\mathbf{R}_i - \widehat{\theta}_k^{(0)}) \right) \tag{28}$$

Theorem 2. *Consider the PCW-Clust algorithm mentioned above, and assume that* $\mathbf{e}_i$ *independently with zero mean, covariance* $\boldsymbol{\Sigma}_{z_i}$*, and sub-gaussian norm* σ_{z_i}*. For arbitrary constant* $\alpha > 1$*, assume that the mis-specified number of the initial cluster estimate* $nl(\widehat{z}^{(0)}, z) \leq \zeta$ *with probability at least* $1 - C_0 n^{-\alpha}$*. Additionally, we require that* $\bar{\sigma}\sqrt{\frac{K(K+\alpha\log n)}{\beta n}} < 1$ *where* $\bar{\sigma} = \max_{k\in[K]} \sigma_k$*,* $K < \min\{n, J\}$*,* $\beta > 0$*,* $\frac{\sqrt{n}+\bar{\sigma}^2\sqrt{J}}{\lambda_{min}(Z\Theta^\top)} < \frac{1}{2}$*,* $\frac{\zeta K}{n\beta} \leq \frac{1}{e}$ *and* $\sigma_{max}(\mathbf{S}_k) \leq \bar{\lambda}$*,* $\lambda_{min}(\mathbf{S}_k) \geq \underline{\lambda}$*,* $\|\boldsymbol{\Sigma}_k\| \leq \Gamma$ *holds for all* $k \in [K]$ *where* $\bar{\lambda}, \underline{\lambda}, \Gamma$ *are positive constants, and further assume that*

$$\psi_1 := C_1 \frac{\left(\bar{\sigma}\sqrt{K + \log n} + \kappa\triangle\right)\sqrt{\delta_s} + \frac{K\zeta\kappa\triangle}{n\beta} + \frac{\bar{\sigma}K}{\beta}\sqrt{\frac{n+J}{n}}}{\boldsymbol{SNR}} \tag{29}$$

$$\psi_2 := C_2 \frac{\left(\sqrt{\frac{K^2}{n\beta}} + \frac{\bar{\sigma}\sqrt{K+\alpha\log n}}{\lambda_{min}(\mathbf{Z\Theta}^\top)}\right)\sqrt{J + \log n}\,\delta_b}{\boldsymbol{SNR}} \tag{30}$$

$$\psi_0 := C_3(\psi_1 + \psi_2) \leq 1 \tag{31}$$

where C_1, C_2, C_3 *are some constants determined by* $\bar{\lambda}, \underline{\lambda}, \Gamma, C_0$ *and* δ_b, δ_s *are defined as*

$$\delta_b = \frac{\sqrt{n} + \bar{\sigma}^2\sqrt{J}}{\lambda_{min}(\mathbf{Z\Theta}^\top)} < 1 \tag{32}$$

$$\begin{aligned}\delta_s = \sigma_k^2\sqrt{\frac{\alpha\log nK}{n\beta}} + \frac{\bar{\sigma}^2K^2}{n\beta} + \left(\frac{\bar{\sigma}\alpha K}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{K\kappa\triangle^2\zeta}{n\beta}\right) \\ + \delta_b^2\left(\sigma_k^2\sqrt{\frac{(J+\alpha\log n)K}{n\beta}} + \frac{\bar{\sigma}^2KJ}{n\beta}\right) < 1\end{aligned} \tag{33}$$

which represent the singular space perturbation and the projected covariance estimation error respectively.

Then it holds that

$$\mathbb{E}l(\widehat{z}, z) \leq \exp\left(-\min_{k\in[K]} \frac{\lambda_{min}(\mathbf{S}_k)(1-\psi_0)}{2\sigma_k^2}\boldsymbol{SNR}^2\right) + Cn^{-\alpha} \tag{34}$$

where the constant C *is determined by* $\bar{\lambda}, \underline{\lambda}, \Gamma, C_0$ *and* $\boldsymbol{SNR}$ *is defined as*

$$\boldsymbol{SNR} := \min_{i,j} \frac{\lambda_{min}\left(\mathbf{S}_i^{-1/2}\mathbf{S}_j^{1/2}\right)\left\|\mathbf{S}_j^{-1/2}\mathbf{V}^\top(\theta_i - \theta_j)\right\|}{1 + \lambda_{min}\left(\mathbf{S}_i^{-1/2}\mathbf{S}_j^{1/2}\right)} \tag{35}$$

The proof is in Appendix B.

Remark. *This algorithm only requires the matrices* $\mathbf{V}^\top\boldsymbol{\Sigma}_k\mathbf{V}$ *(instead of* Σ_k*) to be full-rank for each* $k \in [K]$*, which means it could adapts to more general settings compared with Chen and Zhang (2021)'s method.*

Remark. *Since the assumption* $\bar{\sigma}\sqrt{\frac{K(J+\log n)}{n}} = o(1)$ *is necessary in estimating the covariance matrix for each latent class, the upper bound given by Chen and Zhang (2021) could not adapt to the high-dimensional case*($J \gtrsim n$). *However, when the signal-to-noise ratio* ***SNR*** *is strong enough, our method could extend the setting to* $J \lesssim n^{\alpha}$ *where* $\alpha \in (0, 2)$. *Even when the signal is relatively weak, that is the order of* ***SNR*** *is* $\sqrt{\log n}$, *our method still permit* $J \lesssim n(\log n)^{\alpha}$ *where* $\alpha \in (0, 1/2)$.

7 Post-Projection Hypothesis Testing and Uncertainty Quantification

One important advantage of the row-wise asymptotic normality of $\widehat{\mathbf{U}}\widehat{\mathbf{D}}$ is that after given a data matrix and a cluster estimate, we could test whether a certain individual in the data set or an individual out of the dataset belongs to a certain latent class.

More precisely, given a cluster estimate $\widehat{\mathbf{z}}$, which could be the result of weighted spectral clustering or PCW-Clust, we are interested in testing an individual null hypothesis $H_{0,i}$: $z_i = k$ versus the alternative $H_{A,i} : z_i \neq k$. We construct a p-value P_i for the test $H_{0,i}$ as follows

$$P_i = 1 - F\left(\left(\mathbf{R}_i\mathbf{V} - \widehat{\theta}_k\right)^{\top}\left(\widehat{\mathbf{V}}^T\widehat{\mathbf{\Sigma}}_k\widehat{\mathbf{V}}_k\right)^{\top}\left(\mathbf{R}_i\mathbf{V} - \widehat{\theta}_k\right)\right) \tag{36}$$

where F is the c.d.f of the chi-squared distribution with K degrees of freedom and $\widehat{\theta}_k$ is the mean center estimation according to the given cluster estimate.

8 Simulations

In this section, we perform simulation studies to assess the performance of the proposed method **PCW-Clust**. Specifically, we examine the error rate of the **PCW-Clust** method compared with **Spectral Clustering** and **Weighted Spectral Clustering** in the heteroskedastic noises setting.

8.1 Performace of PCW-Clust method under heteroskedastic noises in LCMs

We consider the setting $\{n = 500, J = 50, L = 3\}$ and $\{n = 500, J = 100, L = 3\}$ first. For the first class, we let the mean vector be $\theta_1 = (0.5, \ldots, 0.5)^\top$ which represent a class has large variance. For the other two classes, we generate the true θ's elements $\theta_{j,a}$ independently and uniformly from $\{0.1, 0.2, 0.8, 0.9\}$. Repeating for 200 iterations, we have the result in figure 8.1.

And we further investigate the rate of exact recovery as the number of J increases. We present the figure 8.1.

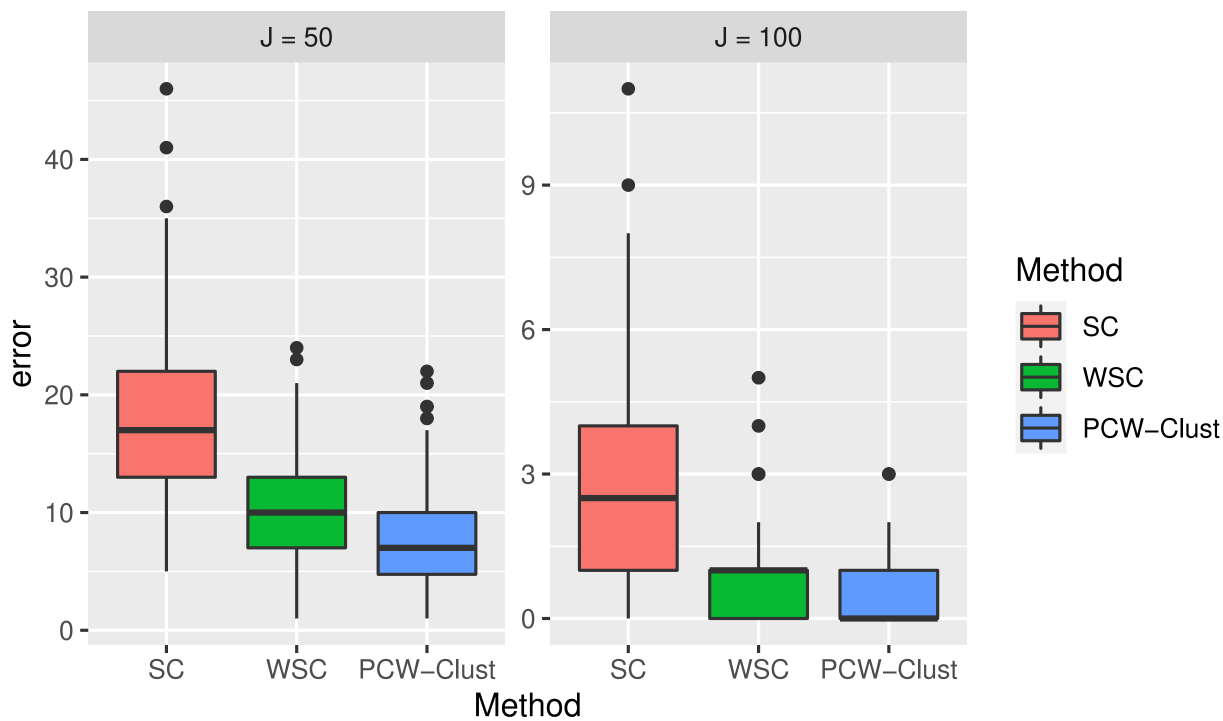

Figure 1: Heteroskedastic LCM, $n = 500, J = \{50, 100\}, L = 3$

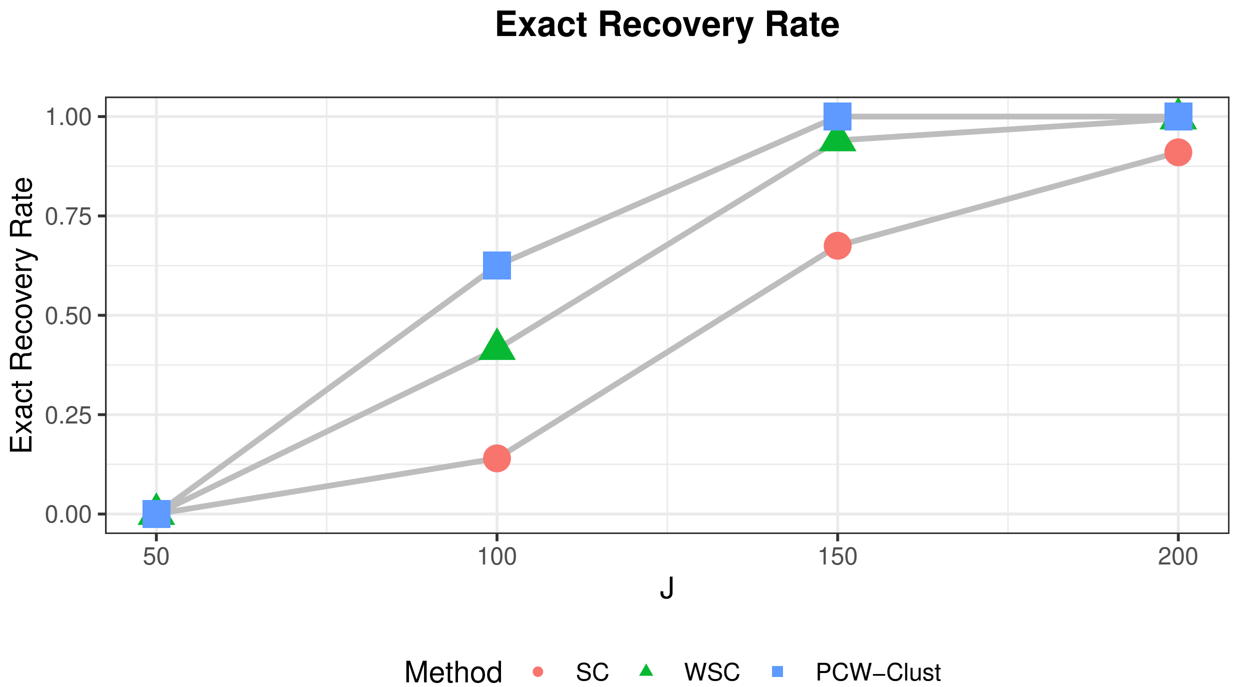

Figure 2: Heteroskedastic LCM, $n = 500, J = \{50, 100, 150, 200\}, L = 3$

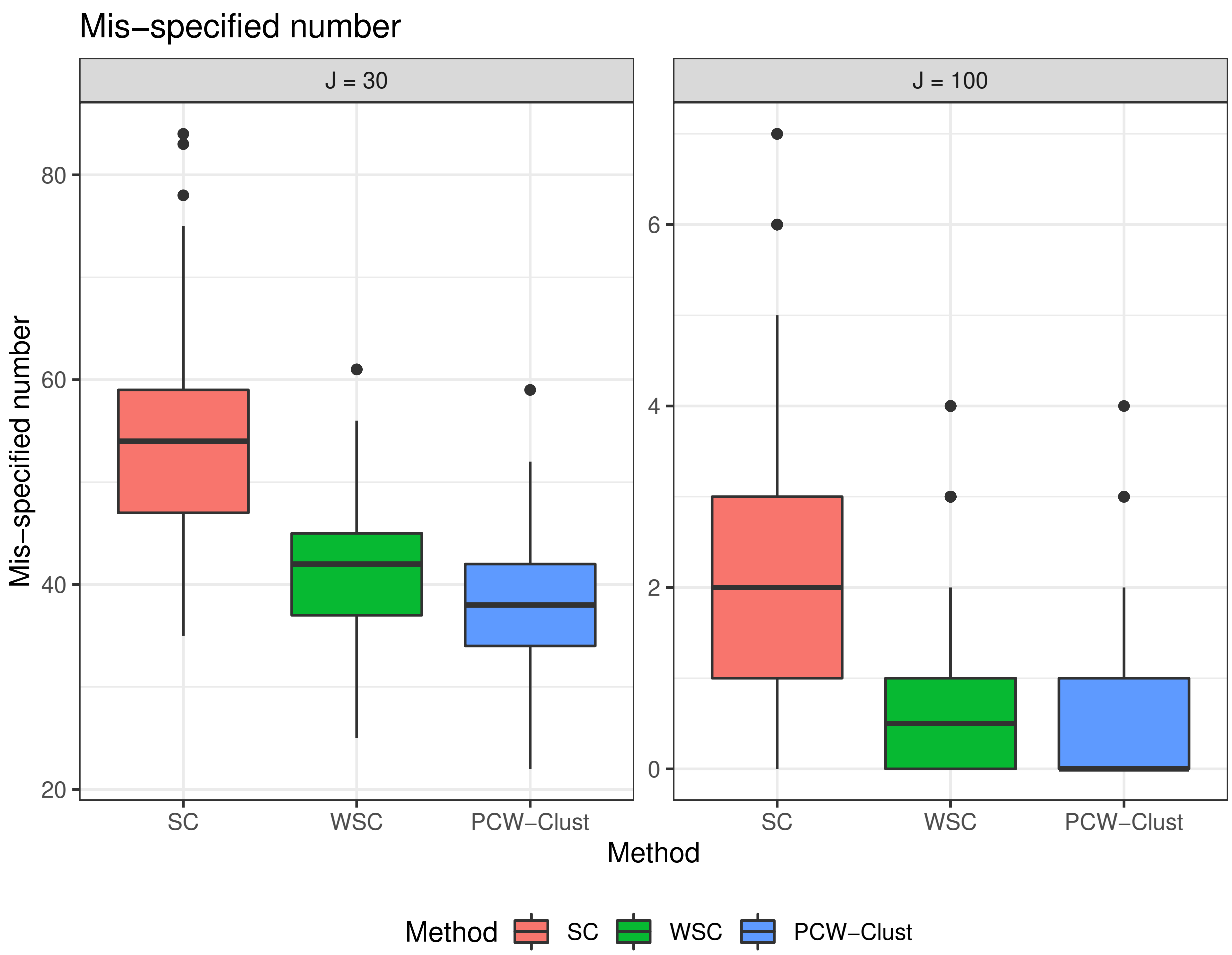

Figure 3: Heteroskedastic Polytomous LCM, $n = 600, J = \{30, 100\}, L = 3, m = 3$

8.2 Performace of PCW-Clust method under heteroskedastic noises in Polytomous LCMs

Here we consider the setting $\{n = 600, J = 30, 100, L = 3, m_i = 3\}$. For the first latent class, we set the mean vector to be $\boldsymbol{\Theta}_1 = (\frac{1}{3}, \ldots, \frac{1}{3})$. For the other two class, we generate the true $\boldsymbol{\Theta}$ for each item independently. For the j-th item, we sample the values of $\boldsymbol{\Theta}_{3(j-1)+1:3j,k}$ from the set $\{0.1, 0.3, 0.6\}$ without replacement. Then we sample the j-th item for the i-th individual from the multinomial distribution with parameter $(\Theta_{3(j-1)+1,z_i}, \Theta_{3(j-1)+2,z_i}, \Theta_{3j,z_i}$. The results is displayed in 8.1.

9 Real Data Analysis

9.1 1000 Genomes Data

In this subsection, we illustrate the application of *PCW-Clust* algorithm to the genetic variation data from the 1000 Genome Project Consortium et al. (2015). The 1000 Genomes Project is ”a deep catalog of human Genetic Variation” as it describes, collecting the genetic data from more than 100 regions of the genomes and different populations. Our goal is to estimate the population cluster from the variant data collected from people from diverse populations. Specifically, we use the genomic variant data on chromosome 17 from 2504 people from the 1000 Genomes Project. Due to the high dimensionality of variant information (1,812,841 variant positions), we select the first 30,000 variants from three super populations EUR,AFR,EAS to compare the performance of spectral clustering, weighted spectral clustering, and PCW-Clust. Compared with the factual super population distribution, the misspecified numbers of individuals with various clustering methods are in 9.1 and the cluster distribution is plotted in 9.1.

	Spectral Clustering	Weighted Spectral Clustering	PCW-Clust
Error	97	97	80

Table 1: misspecified number among 1668 individuals

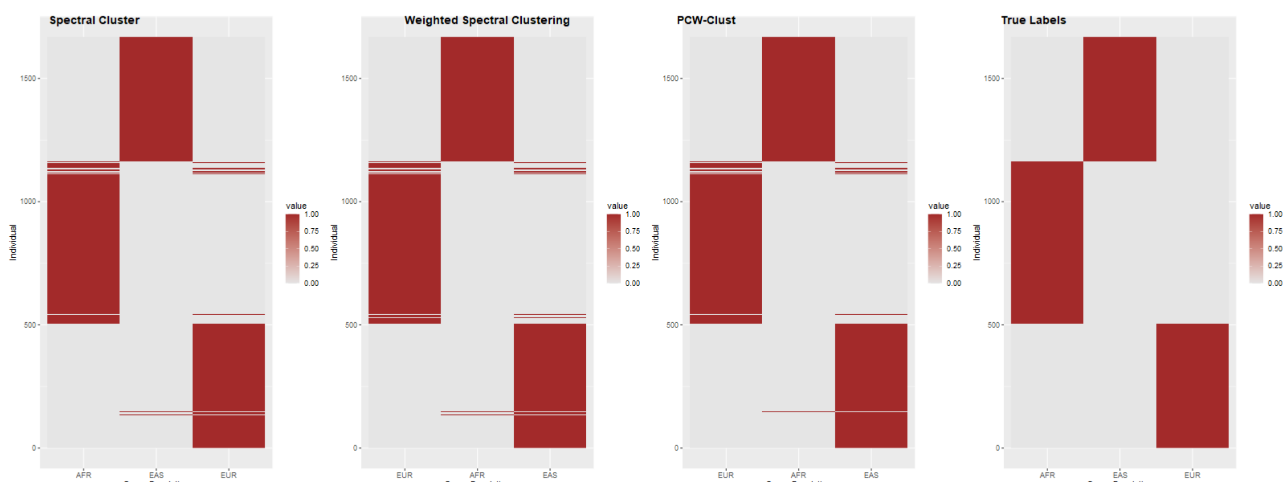

Figure 4: Cluster Distribution Using Various Methods

10 Exact Recovery Guarantees for Polytomous LCMs

In this section, we mainly discuss the performance of the ordinary *Spectral Clustering*. Here we give a sufficient condition for exact recovery in the spirit of Lemma 5.1 in Lei (2019).

Lemma 2. *We use the above notations and assume that $\mathbf{U}_i = \mathbf{c}_s, \forall i \in C_s$. Then the K-means algorithm exactly recovers $C_1, \cdots, C_k$ if*

$$d_{2,\infty}(\widehat{\mathbf{U}}, \mathbf{U}) \leq \frac{\min \pi_r}{6} \min_{s,s' \in [L]} \|\mathbf{c}_s - \mathbf{c}_{s'}\|_2 \tag{37}$$

The proof is in the appendix D.

10.1 $\ell_{2,\infty}$ Perturbation Upper Bound and Exact Recovery Guarantees for the Flattened Matrix R

Here we consider a specific polytomous LCM where $m_1 = \ldots = m_J = m$. Following the idea of Cai et al. (2021), we could give a tighter bound for the top-K left singular matrix of the flattened matrix $\mathbf{R} \in \{0,1\}^{n \times mJ}$ from the tensor $\mathbf{R}^{\mathbf{tensor}} \in \{0,1\}^{n \times J \times m}$, which could provide us a exact recovery guarantee result for the *Spectral Clustering* method by lemma 2. For simplicity, We denote the expectation of matrices and tensors by superscript $*$ in this section.

Theorem 3. *Assume that the following conditions hold*

$$\begin{aligned}
&B\sqrt{\log(n \vee J)} \leq c_0 \min\{\sqrt{n}, \sqrt{J}\}\sigma \\
&\frac{\sigma}{\sigma_L^*} \leq c_1 \min\{\frac{1}{\kappa(nJ)^{1/4}\sqrt{\log(n \vee J)}}, \frac{1}{\kappa^3\sqrt{n\log(n \vee J)}}, \frac{1}{\kappa^3\sqrt{J\log(n \vee J)}}\} \\
&n \wedge J \geq c_4 \mu L \log(n \vee J) \\
&L \leq \frac{c_2 n}{\mu_1 \kappa^4} \\
&\sum_i \left\|\mathbf{R}^{\boldsymbol{tensor}*}{}_{,,i}\right\|^2 \leq c_3 \kappa^2 {\sigma_K^*}^2 \\
&\sum_i \left\|\mathbf{R}^{\boldsymbol{tensor}*}{}_{,,i}\right\| \asymp \|\mathbf{R}^*\|
\end{aligned} \tag{38}$$

where $c_0, c_1, c_2, c_3, c_4 > 0$ are some sufficiently small constants, $(\sigma_1^*, \cdots, \sigma_K^*)$ is the singular value of $\mathbf{R}^*$ and $\kappa = \frac{\sigma_1^*}{\sigma_K^*}$. Then with probability at least $1 - O((n \vee J)^{-10})$, the matrices $\mathbf{U}$

satisfy

$$
\begin{aligned}
d_{2,\infty}(\widehat{\mathbf{U}}, \mathbf{U}) &\lesssim \kappa^2 \sqrt{\frac{\mu r}{n}} \cdot \frac{\zeta_{op} + \sigma_{row}^2}{{\sigma_K^*}^2} \\
&\lesssim \kappa^2 \sqrt{\frac{\mu r}{n}} \frac{\sigma \sigma_1^* \left(\sqrt{n \log(n \vee J)} + \sqrt{J \log(n \vee J)}\right) + \sigma_{row}^2}{{\sigma_K^*}^2}
\end{aligned}
\tag{39}
$$

and

$$
\zeta_{op} = (\sigma_{row}+\sigma_{col})(\sigma_{col}+\sum_i \left\| \mathbf{R}_{,,i}^{\boldsymbol{tensor}*\top} \right\|_{2,\infty}) \log(n\vee J)+(\sigma_{col}+\sigma_{row})\sqrt{\log(n \vee J)} \sum_i \left\| \mathbf{R}_{,,i}^{\boldsymbol{tensor}*} \right\|
\tag{40}
$$

where

$$
\begin{aligned}
&\mu := \max\{\mu_0, \mu_1, \mu_2\}, \mu_0 := \frac{nJ \max |\mathbf{R}_{i,j}^*|^2}{\|\mathbf{R}^*\|_F^2}, \mu_1 := \frac{n}{K} \max_i \left\|\mathbf{U}_{i,:}^*\right\|_2^2, \mu_2 := \frac{J}{K} \max_i \left\|V_{i,:}^*\right\|_2^2 \\
&\max_{i \in [n]} \sqrt{\sum_{j \in [J]} \mathbb{E}[\mathbf{E}_{i,j}^2]} \le \sigma \sqrt{mJ} =: \sigma_{row} \\
&\max_{j \in [J]} \sqrt{\sum_{i \in [n]} \mathbb{E}[\mathbf{E}_{i,j}^2]} \le \sigma \sqrt{n} =: \sigma_{col}
\end{aligned}
\tag{41}
$$

Remark. *To achieve exact recovery, by lemma 2 we require that* $\|\mathbf{UH} - \mathbf{U}^*\|_{2,\infty} \lesssim \frac{1}{\sqrt{n}}$.

Assume that $m, K = O(1)$,*to satisfy the conditions, by Theorem 3 the requirement to* σ_{min} *is:*

$$
\begin{aligned}
\sigma_{min}^* &\ge c_1 \sigma \kappa (\frac{J}{n})^{1/4} \sqrt{\log(n \vee J)} \\
\sigma_{min}^* &\ge c_2 \sigma \sqrt{\mu} \kappa^3 \sqrt{\log(n \vee J)} \\
\sigma_{min}^* &\ge c_3 \sigma \sqrt{\frac{\mu J}{n}} \kappa^3 \sqrt{\log(n \vee J)} \\
\sigma_{min}^* &\gg \sigma \kappa \mu^{1/4} \sqrt{\frac{J}{n}}
\end{aligned}
\tag{42}
$$

where c_1, c_2, c_3, c_4 *are sufficiently small constant.*

Remark. *We compare spectral clustering and weighted spectral clustering methods under the binary LCM with independent noise setting. Assume that* $J \lesssim n$,$\mu \asymp 1$*(defined in 3) and* $R_{i,j}^* \asymp 1$, *to achieve exact recovery, by 3 we know that spectral clustering method requires that*

$$
\sigma_{min}^* \ge c_1 \kappa^3 \sqrt{\log n}
\tag{43}
$$

By (1)*, we know that weighted spectral clustering method requires that:*

$$\triangle \geq c_2\sqrt{\log n} \tag{44}$$

By (13) *we could that the condition in the weighted spectral clustering method is more mild especially when* $\kappa \to \infty$*.*

Appendices

A Proof of Theorem 1

To adapt the asymmetric matrix in our setting to the symmetric case in Fan et al. (2022), we consider the following symmetric matrices

$$\mathbf{X} = \begin{pmatrix} \mathbf{O} & \mathbf{R} \\ \mathbf{R}^\top & \mathbf{O} \end{pmatrix} \tag{45}$$

$$\mathbf{H} = \begin{pmatrix} \mathbf{O} & \mathbf{R}^* \\ \mathbf{R}^{*\top} & \mathbf{O} \end{pmatrix} \tag{46}$$

The top-$2K$ eigen decomposition of $\mathbf{X}$ can be write as

$$\begin{pmatrix} \frac{\widehat{\mathbf{U}}}{\sqrt{2}} & -\frac{\widehat{\mathbf{U}}}{\sqrt{2}} \\ \frac{\widehat{\mathbf{V}}}{\sqrt{2}} & \frac{\widehat{\mathbf{V}}}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{D}} & \\ & -\widehat{\mathbf{D}} \end{pmatrix} \begin{pmatrix} \frac{\widehat{\mathbf{U}}}{\sqrt{2}} & -\frac{\widehat{\mathbf{U}}}{\sqrt{2}} \\ \frac{\widehat{\mathbf{V}}}{\sqrt{2}} & \frac{\widehat{\mathbf{V}}}{\sqrt{2}} \end{pmatrix}^\top \tag{47}$$

Correspondingly, the eigen decomposition of $\mathbf{H}$ is

$$\mathbf{M}^* = \begin{pmatrix} \frac{\mathbf{U}}{\sqrt{2}} & -\frac{\mathbf{U}}{\sqrt{2}} \\ \frac{\mathbf{V}}{\sqrt{2}} & \frac{\mathbf{V}}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \mathbf{D} & \\ & -\mathbf{D} \end{pmatrix} \begin{pmatrix} \frac{\mathbf{U}}{\sqrt{2}} & -\frac{\mathbf{U}}{\sqrt{2}} \\ \frac{\mathbf{V}}{\sqrt{2}} & \frac{\mathbf{V}}{\sqrt{2}} \end{pmatrix}^\top \tag{48}$$

For the first part, by Theorem 1 in Fan et al. (2022), we just need to verify

$$\frac{\mathbf{u}_k^\top \mathbf{E}\mathbf{v}_k}{\mathbf{var}(\mathbf{u}_k^\top \mathbf{E}\mathbf{v}_k)^{1/2}} \overset{d}{\Rightarrow} N(0,1) \tag{49}$$

We know that:

$$\mathbf{u}_k^\top \mathbf{E}\mathbf{v}_k = \sum_{i\in[n], j\in[J]} u_{i,k} E_{i,j} v_{j,k} \tag{50}$$

Then it follows from the assumption $|\mathbf{u}_k\mathbf{v}_k^\top|_\infty \ll 1$ and the fact $|E_{i,j}| \le 1$ and $\mathbf{var}(u_k^\top E v_k) \le 1$:

$$\begin{aligned}
&\frac{1}{\mathbf{var}(\mathbf{u}_k^\top \mathbf{E}\mathbf{v}_k)^{3/2}} \sum_{i\in[n],j\in[J]} \mathbb{E}|E_{i,j}|^3 |u_{i,k}v_{j,k}|^3 \\
\ll &\frac{1}{\mathbf{var}(\mathbf{u}_k^\top \mathbf{E}\mathbf{v}_k)^{3/2}} \sum_{i\in[n],j\in[J]} \mathbb{E}|E_{i,j}|^2 |u_{i,k}v_{j,k}|^2 \\
\le &\frac{1}{\mathbf{var}(\mathbf{u}_k^\top \mathbf{E}\mathbf{v}_k)^{1/2}}
\end{aligned} \tag{51}$$

Therefore by Lyapunov condition, we conclude that (49) holds.

For the second part in 1, by the assumptions we make and (50) in Fan et al. (2022), for arbitrary j satisfying $z_j = z_i$ it holds that:

$$t_k \widetilde{\mathbf{o}}_j \widehat{\mathbf{t}}_k + t_k A_{\widetilde{\mathbf{o}}_j,k,t_k} \widetilde{\mathcal{P}}_{k,t_k}^{1/2} = t_k \mathbf{o}_j \widehat{\mathbf{u}}_k + t_k A_{\widetilde{\mathbf{o}}_j,k,t_k} \widetilde{\mathcal{P}}_{k,t_k}^{1/2} = \left(\mathbf{b}_{\widetilde{\mathbf{o}}_j,k,t_k}^\top - \widetilde{\mathbf{o}}_j^\top \mathbf{t}_k \mathbf{t}_k^\top\right) \widetilde{\mathbf{E}}\mathbf{t}_k + O_p(\alpha_n/t_k) \tag{52}$$

where $\mathbf{b}_{\mathbf{u},k,t}$, $A_{\mathbf{u},\mathbf{v},t}$ and $\widetilde{\mathcal{P}}_{k,t}$ is defined in (7),(8),(9) in Fan et al. (2022).

Since for distinct j, l such that $z_j = z_l = z_i$, it has:

$$\begin{aligned}
&A_{\widetilde{\mathbf{o}}_j,k,t_k} \widetilde{\mathcal{P}}_{k,t_k}^{1/2} = A_{\widetilde{\mathbf{o}}_l,k,t_k} \widetilde{\mathcal{P}}_{k,t_k}^{1/2} \\
&\mathbf{T}_{-j}\left[(\mathbf{D}_{-j})^{-1} + \mathcal{R}(\mathbf{T}_{-j},\mathbf{T}_{-j},t_k)\right]^{-1} \mathcal{R}(o_j,\mathbf{V}_{-j},t_k)^\top = \mathbf{T}_{-l}\left[(\mathbf{D}_{-l})^{-1} + \mathcal{R}(\mathbf{T}_{-l},\mathbf{T}_{-l},t_k)\right]^{-1} \mathcal{R}(o_l,\mathbf{V}_{-l},t_k)^\top
\end{aligned} \tag{53}$$

Therefore it follows that:

$$t_k\left(\mathbf{o}_i^\top \widehat{\mathbf{u}}_k - \bar{\mathbf{o}}_{z_i}^\top \widehat{\mathbf{u}}_k\right) = \left(\mathbf{o}_i - \bar{\mathbf{o}}_{z_i}\right)^\top \mathbf{E}\mathbf{v}_k + O_p(\alpha_n/t_k) \tag{54}$$

By the first part we know that:

$$\frac{\widehat{\sigma}_k - t_k}{[\mathbf{var}(\mathbf{u}_k^\top \mathbf{E}\mathbf{v}_k)]^{1/2}} \overset{d}{\Rightarrow} N(0,1) \tag{55}$$

and by Condition 2 and Lemma 3 in Fan et al. (2022), we also know that $t_k \to \infty$.

Combining the above results and condition 2, it follows that:

$$\widehat{\sigma}_k\left(\mathbf{o}_i^\top \widehat{\mathbf{u}}_k - \bar{\mathbf{o}}_{z_i}^\top \widehat{\mathbf{u}}_k\right) = \left(\mathbf{o}_i - \bar{\mathbf{o}}_{z_i}\right)^\top \mathbf{E}\mathbf{v}_k + O_p(\alpha_n/t_k) \tag{56}$$

For the third part, it holds that:

$$\begin{aligned}
&\mathbf{R}_i\widehat{\mathbf{V}} - \frac{\sum_{j\in[n]} 1\{z_j = z_i\}\mathbf{R}_j\widehat{\mathbf{V}}}{\sum_{j\in[n]} 1\{z_j = z_i\}} \\
=&\widehat{\mathbf{U}}_i\widehat{\mathbf{D}} - \widehat{\theta}(z)^\top\widehat{\mathbf{V}} \\
=&\left(\mathbf{o}_i - \bar{\mathbf{o}}_{z_i}\right)^\top \left(\mathbf{E}\mathbf{v}_1, \ldots, \mathbf{E}\mathbf{v}_k\right) + O_p(\alpha_n/t_k) \\
=&\mathbf{o}_i^\top \left(\mathbf{E}\mathbf{v}_1, \ldots, \mathbf{E}\mathbf{v}_k\right) - \frac{1}{\sqrt{\sum_{j\in[n]} \mathbf{1}\{z_j = z_i\}}}\sqrt{\sum_{j\in[n]} \mathbf{1}\{z_j = z_i\}}\bar{\mathbf{o}}_{z_i} \left(\mathbf{E}\mathbf{v}_1, \ldots, \mathbf{E}\mathbf{v}_k\right) + O_p(\alpha_n/t_k)
\end{aligned} \tag{57}$$

Since we assume that $\left\|\left(\mathbf{V}^\top\Sigma_{z_i}\mathbf{V}\right)^{-1/2}\right\| \lesssim 1 \ll \frac{t_k}{\alpha_n}$, it follows that:

$$\begin{aligned}
&\left(\mathbf{R}_i\widehat{\mathbf{V}} - \frac{\sum_{j\in[n]} 1\{z_j = z_i\}\mathbf{R}_j\widehat{\mathbf{V}}}{\sum_{j\in[n]} 1\{z_j = z_i\}}\right)\left(\mathbf{V}^\top\Sigma_{z_i}\mathbf{V}\right)^{-1/2} \\
=&\mathbf{o}_i^\top \left(\mathbf{E}\mathbf{v}_1, \ldots, \mathbf{E}\mathbf{v}_k\right)\left(\mathbf{V}^\top\Sigma_{z_i}\mathbf{V}\right)^{-1/2} \\
&-\frac{1}{\sqrt{\sum_{j\in[n]} \mathbf{1}\{z_j = z_i\}}}\sqrt{\sum_{j\in[n]} \mathbf{1}\{z_j = z_i\}}\bar{\mathbf{o}}_{z_i} \left(\mathbf{E}\mathbf{v}_1, \ldots, \mathbf{E}\mathbf{v}_k\right)\left(\mathbf{V}^\top\Sigma_{z_i}\mathbf{V}\right)^{-1/2} + o_p(1)
\end{aligned} \tag{58}$$

Consider the term $\mathbf{o}_i^\top \left(\mathbf{E}\mathbf{v}_1, \ldots, \mathbf{E}\mathbf{v}_k\right)\left(\mathbf{V}^\top\Sigma_{z_i}\mathbf{V}\right)^{-1/2}$ first. Denote the term $\left(\mathbf{V}^\top\Sigma_{z_i}\mathbf{V}\right)^{-1/2}\mathbf{c}$ by $\widetilde{\mathbf{c}}$. For arbitrary unit vector $\mathbf{c} \in \mathbb{R}^K$ by the conditions $\|\mathbf{v}\|_\infty \ll 1$ and $\left(\left\|\mathbf{V}^\top\Sigma_{z_i}\mathbf{V}\right\|\right)^{-1/2} \lesssim 1$ it has

$$\sum_{j\in[J]} |E_{i,j}|^3| \sum_{k\in[K]} v_{j,k}\widetilde{v}_k|^3 \ll \sum_{j\in[J]} |E_{i,j}|^2| \sum_{k\in[K]} v_{j,k}\widetilde{v}_k|^2 = \mathbf{var}\left(\mathbf{o}_i^\top \left(\mathbf{E}\mathbf{v}_1, \ldots, \mathbf{E}\mathbf{v}_k\right)\left(\mathbf{V}^\top\Sigma_{z_i}\mathbf{V}\right)^{-1/2}\mathbf{c}\right) = 1 \tag{59}$$

By the Lyapunov condition, it holds that:

$$\mathbf{o}_i^\top \left(\mathbf{E}\mathbf{v}_1, \ldots, \mathbf{E}\mathbf{v}_k\right)\left(\mathbf{V}^\top\Sigma_{z_i}\mathbf{V}\right)^{-1/2} \overset{d}{\Rightarrow} N(\mathbf{0}, \mathbf{I}_K) \tag{60}$$

Similarly, for the term $\sqrt{\sum_{j\in[n]} \mathbf{1}\{z_j = z_i\}}\bar{\mathbf{o}}_{z_i} \left(\mathbf{E}\mathbf{v}_1, \ldots, \mathbf{E}\mathbf{v}_k\right)\left(\mathbf{V}^\top\Sigma_{z_i}\mathbf{V}\right)^{-1/2}$, we can also prove that:

$$\sqrt{\sum_{j\in[n]} \mathbf{1}\{z_j = z_i\}}\bar{\mathbf{o}}_{z_i}\bar{\mathbf{o}}_{z_i} \left(\mathbf{E}\mathbf{v}_1, \ldots, \mathbf{E}\mathbf{v}_k\right)\left(\mathbf{V}^\top\Sigma_{z_i}\mathbf{V}\right)^{-1/2} \overset{d}{\Rightarrow} N(\mathbf{0}, \mathbf{I}_K) \tag{61}$$

By the condition $\min_{k\in[K]} \sum_{j\in[n]} \mathbf{1}\{z_j = k\} \to \infty$, it follows that:

$$\left(\mathbf{R}_i\widehat{\mathbf{V}} - \frac{\sum_{j\in[n]} 1\{z_j = z_i\}\mathbf{R}_j\widehat{\mathbf{V}}}{\sum_{j\in[n]} 1\{z_j = z_i\}}\right)\left(\mathbf{V}^\top\boldsymbol{\Sigma}_{z_i}\mathbf{V}\right)^{-1/2} \overset{d}{\Rightarrow} N(\mathbf{0}, \mathbf{I}_K) \tag{62}$$

For the last part, it is straightforward by the assumption we make.

B Proof of Theorem 2

WLOG, we assume that $\arg\min_{\phi \in \mathbf{perm}(L)} l(z, \widehat{z}) = \mathbf{id}$. We introduce some notations at the begining:

$$
\begin{aligned}
&\zeta = nl(\widehat{z}^{(0)}, z) \\
&\bar{\sigma} = \max_{k \in [L]} \sigma_k \\
&\delta_b := \frac{\sqrt{n} + \bar{\sigma}^2 \sqrt{J}}{\lambda_{min}(\mathbf{Z}\boldsymbol{\Theta})}
\end{aligned}
\tag{63}
$$

Aiming at controlling the error risk we decompose the risk for each sample

$$
\mathbb{E}[l(z, \widehat{z})] = \sum_{i \in [n]} \mathbb{E}\left[\mathbf{1}\left\{\bigcup_{k \neq z_i} \left\{\left\|\widehat{S}_{z_i}^{-1/2} \widehat{V}^\top \left(R_{i,} - \widehat{\theta}_{z_i}^{(0)}\right)\right\| \geq \left\|\widehat{S}_k^{-1/2} \widehat{V}^\top \left(R_{i,} - \widehat{\theta}_k^{(0)}\right)\right\|\right\}\right\}\right] \tag{64}
$$

Inspired by the procedure in Zhang and Zhou (2022), we analysis the risk term by term

$$
\begin{aligned}
&\mathbf{1}\left\{\bigcup_{k \neq z_i} \left\{\left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top \left(\mathbf{R}_i - \widehat{\theta}_{z_i}^{(0)}\right)\right\|_2 \geq \left\|\widehat{\mathbf{S}}_k^{-1/2} \widehat{V}^\top \left(\mathbf{R}_{i,} - \widehat{\theta}_k^{(0)}\right)\right\|_2\right\}\right\} && (65) \\
\leq &\mathbf{1}\left\{\bigcup_{k \neq z_i} \left\{\left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top \left(\mathbf{R}_i - \widehat{\theta}_{z_i}^{(0)}\right)\right\|_2 \geq \widehat{\eta}_{k,z_i} \left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top \left(\mathbf{R}_i - \widehat{\theta}_k^{(0)}\right)\right\|_2\right\}\right\} && (66) \\
\leq &\mathbf{1}\left\{\left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top \left(\mathbf{R}_i - \widehat{\theta}_{z_i}^{(0)}\right)\right\|_2 \geq \min_k \frac{\widehat{\eta}_{k,z_i}}{1 + \widehat{\eta}_{k,z_i}} \left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{V}^\top \left(\widehat{\theta}_{z_i}^{(0)} - \widehat{\theta}_k^{(0)}\right)\right\|_2\right\} && (67)
\end{aligned}
$$

Replacing the empirical quantities with the population ones, it has

$$
(67) \leq \mathbf{1}\left\{\left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top \mathbf{e}_i\right\| + \left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top \left(\widehat{\theta}_{z_i} - \theta_{z_i}\right)\right\|_2 \geq \min_k \frac{\widehat{\eta}_{k,z_i}}{1 + \widehat{\eta}_{k,z_i}} \left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top \left(\widehat{\theta}_{z_i}^{(0)} - \widehat{\theta}_k^{(0)}\right)\right\|_2\right\} \tag{68}
$$

$$
\begin{aligned}
\leq \mathbf{1}\Bigg\{&\left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top \mathbf{e}_i\right\|_2 + \left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top \left(\widehat{\theta}_{z_i}^{(0)} - \theta_{z_i}\right)\right\|_2 \geq \min_k \frac{\widehat{\eta}_{k,z_i}}{1 + \widehat{\eta}_{k,z_i}} \left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top (\theta_{z_i} - \theta_k)\right\|_2 \\
&- \max_k \frac{2\widehat{\eta}_{k,z_i}}{1 + \widehat{\eta}_{k,z_i}} \max_{j \in [L]} \left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \left(\widehat{\theta}_j^{(0)} - \theta_j\right)\right\|_2\Bigg\}
\end{aligned}
\tag{69}
$$

$$
\leq \mathbf{1}\left\{\left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top \mathbf{e}_i\right\| \geq \min_k \frac{\widehat{\eta}_{k,z_i}}{1 + \widehat{\eta}_{k,z_i}} \left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \widehat{\mathbf{V}}^\top (\theta_{z_i} - \theta_k)\right\| - \max_k \frac{3\widehat{\eta}_{k,z_i} + 1}{1 + \widehat{\eta}_{k,z_i}} \max_{j \in [L]} \left\|\widehat{\mathbf{S}}_{z_i}^{-1/2} \left(\widehat{\theta}_j^{(0)} - \theta_j\right)\right\|\right\} \tag{70}
$$

where $\widehat{\eta}_{k,z_i}$ is the smallest singular value of $\widehat{\mathbf{S}}_k^{-1/2} \widehat{\mathbf{S}}_{z_i}^{1/2}$ and η_{k,z_i} is the smallest singular value of $\mathbf{S}_k^{-1/2} \mathbf{S}_{z_i}^{1/2}$.

Observing that

$$
\begin{aligned}
&\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1/2}\widehat{\mathbf{V}}^\top\mathbf{e}_i\right\| \\
=&\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1/2}\widehat{\mathbf{V}}^\top\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\mathbf{e}_i\right\| \\
\leq&\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1/2}\widehat{\mathbf{V}}^\top\widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^\top}\mathbf{e}_i\right\| + \left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1/2}\widehat{\mathbf{V}}^\top\left(\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^\top}\right)\mathbf{e}_i\right\| \\
\leq&\left\|\mathbf{V}\mathbf{S}_{z_i}^{-1/2}\mathbf{V}^\top\widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^\top}\mathbf{e}_i\right\| + \left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top\right\|^{1/2}\left\|\widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^\top}\mathbf{e}_i\right\| \\
+&\left\|\mathbf{V}\mathbf{S}_{z_i}^{-1/2}\mathbf{V}^\top\right\|\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^\top}\right\|\|\mathbf{e}_i\| + \left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top\right\|^{1/2}\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^\top}\right\|\|\mathbf{e}_i\|
\end{aligned} \tag{71}
$$

(70) can be write as

$$
\begin{aligned}
(70) \leq \mathbf{1}\Bigg\{&\left\|\mathbf{S}_{z_i}^{-1/2}\mathbf{V}^\top\widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^\top}\mathbf{e}_i\right\| \geq \min_k \frac{\eta_{k,z_i}}{1+\eta_{k,z_i}}\left\|\mathbf{S}_{z_i}^{-1/2}\mathbf{V}^\top(\theta_{z_i}-\theta_k)\right\| \\
&-\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top\right\|^{1/2}\max_{k\in[K]}\|(\theta_{z_i}-\theta_k)\| - \max_{k\in[L]}|\widehat{\eta}_{k,z_i}-\eta_{k,z_i}|\left\|\mathbf{S}_{z_i}^{-1/2}\mathbf{V}^\top(\theta_{z_i}-\theta_k)\right\| \\
&-\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top\right\|^{1/2}\max_{k\in[L]}|\widehat{\eta}_{k,z_i}-\eta_{k,z_i}|\,\|(\theta_{z_i}-\theta_k)\| - 3\left\|\mathbf{S}_{z_i}^{-1/2}\right\|\max_{j\in[L]}\left\|\left(\widehat{\theta}_j^{(0)}-\theta_j\right)\right\| \\
&-3\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top\right\|^{1/2}\max_{j\in[L]}\left\|\left(\widehat{\theta}_j^{(0)}-\theta_j\right)\right\| - \left\|\mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top\right\|^{1/2}\left\|\widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^T}\mathbf{e}_i\right\| \\
&-\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top\right\|^{1/2}\left\|\widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^\top}\mathbf{e}_i\right\| - \left\|\mathbf{V}\mathbf{S}_{z_i}^{-1/2}\mathbf{V}^\top\right\|\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^\top}\right\|\|\mathbf{e}_i\| \\
&-\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top\right\|^{1/2}\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^\top}\right\|\|\mathbf{e}_i\|\Bigg\}
\end{aligned} \tag{72}
$$

where the event $\mathcal{F}$ are defined as

$$
\begin{aligned}
\mathcal{F} = \Bigg\{&\max_{j\in[K]}\left\|\widehat{\theta}_j - \theta_j\right\| \leq c_1\xi_g \\
&\max_{i,j}|\widehat{\eta}_{i,j}-\eta_{i,j}| \leq \widetilde{c}_2\sqrt{\frac{\delta_s\bar{\lambda}}{\underline{\lambda}^2}} = c_2\sqrt{\delta_s} \\
&\max_{i\in[K]}\left\|\left(\mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top - \widehat{\mathbf{V}}\widehat{\mathbf{S}}_{z_i}^{-1}\widehat{\mathbf{V}}^\top\right)\right\|^{1/2} \leq \widetilde{c}_3\sqrt{\frac{\delta_s}{\underline{\lambda}^2}} = c_3\sqrt{\delta_s} && (73) \\
&\max_{i\in[n]}\|\mathbf{e}_i\| \leq c_4\bar{\sigma}\sqrt{J+\alpha\log n} \\
&\max_{i\in[n]}\left\|\mathbf{V}^{-i}\mathbf{V}^{-i^\top}\mathbf{e}_i\right\|_2 \leq c_5\bar{\sigma}\sqrt{K+\alpha\log n} \\
&\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^T}\right\| \leq c_6\delta_h\Bigg\}
\end{aligned} \tag{74}
$$

and the quantity $\xi_g, \delta_h, \delta_s$ are defined as

$$\xi_g = \frac{\bar{\sigma} K \sqrt{n^{\alpha+1}\tau}}{n\beta} + \frac{n^{\alpha} K \kappa \triangle \tau}{\beta} + \sqrt{\frac{K(\alpha \log n + K)}{n\beta}} \sigma_k \tag{75}$$

$$\begin{aligned} \delta_s = &\sigma_k^2 \sqrt{\frac{\alpha \log n K}{n\beta}} + \frac{\bar{\sigma}\alpha K}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K \bar{\triangle}^2 \zeta}{n\beta} \\ &+ \delta_b^2 \sigma_k^2 \sqrt{\frac{(J + \alpha \log n) K}{n\beta}} + \delta_b^2 \frac{\bar{\sigma}^2 K J}{n\beta} \end{aligned} \tag{76}$$

$$\delta_h := \delta_b \left(\sqrt{\frac{K^2}{\beta n}} + \frac{\bar{\sigma}\sqrt{K + \log n}}{\lambda_{min}(\mathbf{Z}\boldsymbol{\Theta}^\top)} \right) \tag{77}$$

By Lemma 3, 4, 5, 6, 7, 8, 9 and Theorem 2.2 in Zhang and Zhou (2022), we know that

$$\mathbb{P}(\mathcal{F}^{\mathcal{C}}) = O(n^{-\alpha}) \tag{78}$$

Then it follows that

$$\begin{aligned} (72) \leq\ &\mathbf{1}\left\{\mathcal{F}^C\right\} + \mathbf{1}\left\{\mathcal{F}\right\} \bigcap \Bigg\{ \left\| \mathbf{S}_{z_i}^{-1/2} \mathbf{V}^\top \widehat{\mathbf{V}}^{-i} \widehat{\mathbf{V}}^{-i^\top} \mathbf{e}_i \right\| \geq \min_k \frac{\eta_{k,z_i}}{1+\eta_{k,z_i}} \left\| \mathbf{S}_{z_i}^{-1/2} \mathbf{V}^\top (\theta_{z_i} - \theta_k) \right\| \\ &- \left\| \widehat{\mathbf{V}} \widehat{\mathbf{S}}_{z_i}^{-1} \widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top \right\|^{1/2} \max_{k \in [K]} \left\| (\theta_{z_i} - \theta_k) \right\| - \max_{k\in[L]} \left| \widehat{\eta}_{k,z_i} - \eta_{k,z_i} \right| \left\| \mathbf{S}_{z_i}^{-1/2} \mathbf{V}^\top (\theta_{z_i} - \theta_k) \right\| \\ &- \left\| \widehat{\mathbf{V}} \widehat{\mathbf{S}}_{z_i}^{-1} \widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top \right\|^{1/2} \max_{k\in[L]} \left| \widehat{\eta}_{k,z_i} - \eta_{k,z_i} \right| \left\| (\theta_{z_i} - \theta_k) \right\| - 3 \left\| \mathbf{S}_{z_i}^{-1/2} \right\| \max_{j \in [L]} \left\| \left(\widehat{\theta}_j^{(0)} - \theta_j \right) \right\| \\ &-3 \left\| \widehat{\mathbf{V}} \widehat{\mathbf{S}}_{z_i}^{-1} \widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top \right\|^{1/2} \max_{j\in[L]} \left\| \left(\widehat{\theta}_j^{(0)} - \theta_j \right) \right\| - \left\| \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top \right\|^{1/2} \left\| \widehat{\mathbf{V}}^{-i} \widehat{\mathbf{V}}^{-i^T} \mathbf{e}_i \right\| \\ &- \left\| \widehat{\mathbf{V}} \widehat{\mathbf{S}}_{z_i}^{-1} \widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top \right\|^{1/2} \left\| \widehat{\mathbf{V}}^{-i} \widehat{\mathbf{V}}^{-i^\top} \mathbf{e}_i \right\| - \left\| \mathbf{V}\mathbf{S}_{z_i}^{-1/2}\mathbf{V}^\top \right\| \left\| \widehat{\mathbf{V}} \widehat{\mathbf{V}}^\top - \widehat{\mathbf{V}}^{-i} \widehat{\mathbf{V}}^{-i^\top} \right\| \left\| \mathbf{e}_i \right\| \\ &- \left\| \widehat{\mathbf{V}} \widehat{\mathbf{S}}_{z_i}^{-1} \widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_{z_i}^{-1}\mathbf{V}^\top \right\|^{1/2} \left\| \widehat{\mathbf{V}} \widehat{\mathbf{V}}^\top - \widehat{\mathbf{V}}^{-i} \widehat{\mathbf{V}}^{-i^\top} \right\| \left\| \mathbf{e}_i \right\| \Bigg\} \end{aligned} \tag{79}$$

$$\begin{aligned} \leq\ &\mathbf{1}\left\{\mathcal{F}^C\right\} + \mathbf{1}\left\{\mathcal{F}\right\} \bigcap \Bigg\{ \left\| \mathbf{S}_{z_i}^{-1/2} \mathbf{V}^\top \widehat{\mathbf{V}}^{-i} \widehat{\mathbf{V}}^{-i^\top} \mathbf{e}_i \right\| \geq \min_k \frac{\eta_{k,z_i}}{1+\eta_{k,z_i}} \left\| \mathbf{S}_{z_i}^{-1/2} \mathbf{V}^\top (\theta_{z_i} - \theta_k) \right\| \\ &-c_3\sqrt{\delta_s} - c_3\sqrt{\delta_s}\frac{\bar{\triangle}}{\underline{\lambda}} - \delta_s \bar{\triangle} - 3c_1 \frac{\xi_g}{\sqrt{\underline{\lambda}}} - 3c_1\sqrt{\delta_s}\xi_g - c_5 \bar{\sigma} \sqrt{\frac{L + \alpha \log n}{\underline{\lambda}}} \\ &-c_3 c_5 \sqrt{\delta_s} \bar{\sigma} \sqrt{L + \alpha \log n} - c_6 \frac{\bar{\sigma}\delta_h \sqrt{J + \alpha \log n}}{\sqrt{\underline{\lambda}}} - c_3 c_6 \sqrt{\delta_s}\delta_h \bar{\sigma} \sqrt{J + \alpha \log n} \Bigg\} \end{aligned} \tag{80}$$

$$
\begin{aligned}
&\leq \mathbf{1}\left\{\mathcal{F}^C\right\} + \mathbf{1}\left\{\mathcal{F}\right\} \bigcap \left\{ \left\| \mathbf{S}_{z_i}^{-1/2} \mathbf{V}^\top \widehat{\mathbf{V}}^{-i} \widehat{\mathbf{V}}^{-i\top} \mathbf{e}_i \right\| \geq \min_{k \in [K]} \frac{\eta_{k,z_i}}{1+\eta_{k,z_i}} \left\| \mathbf{S}_{z_i}^{-1/2} \mathbf{V}^\top \left(\theta_{z_i} - \theta_k\right) \right\| \right. \\
&\quad \left. -\overline{C}\left(\bar{\sigma}\sqrt{K+\log n} + \bar{\triangle} + \xi_g\right)\sqrt{\delta_s} - \widetilde{C}\xi_g - \widehat{C}\delta_h\sqrt{J+\log n} \right\}
\end{aligned}
\tag{81}
$$

$$
\leq \mathbf{1}\left\{\mathcal{F}^C\right\} + \mathbf{1}\left\{\mathcal{F}\right\} \bigcap \left\{ \left\| \mathbf{S}_{z_i}^{-1/2} \mathbf{V}^\top \widehat{\mathbf{V}}^{-i} \widehat{\mathbf{V}}^{-i\top} \mathbf{e}_i \right\| \geq \left(1-\psi_1-\psi_2\right) \mathbf{SNR} \right\} \tag{82}
$$

where constants $\overline{C}, \widetilde{C}, \widehat{C}$ are determined by $\underline{\lambda}, \bar{\lambda}, \Gamma$. and ψ_1 and ψ_2 are defined as

$$
\psi_1 := C_1 \frac{\left(\bar{\sigma}\sqrt{K+\log n} + \bar{\triangle}\right)\sqrt{\delta_s} + \frac{L\zeta\kappa\triangle}{n\beta} + \frac{\bar{\sigma}L}{\beta}\sqrt{\frac{n+J}{n}}}{\mathbf{SNR}} \tag{83}
$$

$$
\psi_2 := C_2 \frac{\left(\sqrt{\frac{K^2}{n\beta}} + \frac{\bar{\sigma}\sqrt{K+\alpha\log n}}{\lambda_{min}(\mathbf{Z}\boldsymbol{\Theta}^\top)}\right)\sqrt{J+\log n}\,\delta_b}{\mathbf{SNR}} \tag{84}
$$

where C_1, C_2 are constants determined by $\underline{\lambda}, \bar{\lambda}, \Gamma$, (84) is straightforward derived from the definition of δ_h and (83) holds since

$$
\begin{aligned}
&\overline{C}\left(\bar{\sigma}\sqrt{K+\log n} + \bar{\triangle} + \xi_g\right)\sqrt{\delta_s} + \widetilde{C}\xi_g \\
&\lesssim \left(\bar{\sigma}\sqrt{K+\log n} + \bar{\triangle}\right)\sqrt{\delta_s} + \left(\frac{\bar{\sigma}K\sqrt{\zeta(J+n)}}{n\beta} + \frac{K\zeta\bar{\triangle}}{n\beta} + \sqrt{\frac{K\left(\alpha\log n + J\right)}{n\beta}}\sigma_k\right) \\
&\overset{(a)}{\lesssim} \left(\bar{\sigma}\sqrt{K+\log n} + \bar{\triangle}\right)\sqrt{\delta_s} + \frac{K\zeta\kappa\triangle}{n\beta} + \frac{\bar{\sigma}K}{\beta}\sqrt{\frac{n+J}{n}}
\end{aligned}
\tag{85}
$$

where (a) holds since $\frac{K}{\beta} > \sqrt{\frac{K}{\beta}}$ and $\frac{\zeta}{n} < 1$.

Define the event $\mathcal{G}$ and the quantity ψ_4 by

$$
\mathcal{G} = \left\{ \left\| \widehat{\mathbf{V}}^{-i} \widehat{\mathbf{V}}^{-i\top} - \mathbf{V}\mathbf{V}^\top \right\| \leq \dot{C}\delta_b \right\} \tag{86}
$$

we know that

$$
\mathbb{P}[\mathcal{G}^C] \leq O(n^{-\alpha}) \tag{87}
$$

(82) becomes

$$
(82) \leq \mathbf{1}\left\{\mathcal{F}^C\right\} + \mathbf{1}\left\{\mathcal{G}^C\right\} + \mathbf{1}\left\{ \mathcal{F} \bigcap \mathcal{G} \bigcap \left\| \mathbf{S}_{z_i}^{-1/2} \mathbf{V}^\top \widehat{\mathbf{V}}^{-i} \widehat{\mathbf{V}}^{-i\top} \mathbf{e}_i \right\| \geq \left(1-\psi_1-\psi_2-\psi_3\right) \mathbf{SNR} \right\} \tag{88}
$$

Denote the matrix $\mathbf{S}_{z_i}^{-1/2} \mathbf{V}^\top \widehat{\mathbf{V}}^{-i} \widehat{\mathbf{V}}^{-i\top}$ by $\mathbf{A}$, by the tail inequality estabilished by Hsu et al. (2012), conditioned on the other $n-1$ samples it has

$$
\mathbb{P}\left[\|\mathbf{A}\mathbf{e}_i\|_2^2 \geq \sigma_{z_i}^2 \left(\mathbf{tr}\left(\mathbf{A}^\top\mathbf{A}\right) + 2\sqrt{\mathbf{tr}\left(\mathbf{A}^\top\mathbf{A}\mathbf{A}^\top\mathbf{A}\right)t} + \left\|\mathbf{A}^\top\mathbf{A}\right\| t \right) \right] \leq \exp(-t) \tag{89}
$$

under the event $\mathcal{G}$ it becomes

$$\begin{aligned}
&\exp(-t) \geq \mathbb{P}\left[\left\{\|\mathbf{A}\mathbf{e}_i\|_2^2 \geq \sigma_{z_i}^2\left(\mathbf{tr}\left(\mathbf{A}^\top\mathbf{A}\right) + 2\sqrt{\mathbf{tr}\left(\mathbf{A}^\top\mathbf{A}\mathbf{A}^\top\mathbf{A}\right)t} + \left\|\mathbf{A}^\top\mathbf{A}\right\| t\right)\right\} \bigcup \mathcal{G}\right] \\
\geq&\mathbb{P}\left[\|\mathbf{A}\mathbf{e}_i\|_2^2 \geq \sigma_{z_i}^2(1+2\dot{C}\delta_b)\left(\mathbf{tr}\left(\mathbf{S}_{z_i}^{-1}\right) + 2\sqrt{\mathbf{tr}\left(\mathbf{S}_{z_i}^{-2}\right)t} + \left\|\mathbf{S}_{z_i}^{-1}\right\| t\right)\right] \\
\geq&\mathbb{P}\left[\|\mathbf{A}\mathbf{e}_i\|_2^2 \geq \sigma_{z_i}^2(1+2\dot{C}\delta_b)\left(\frac{K}{\underline{\lambda}} + 2\sqrt{\frac{K}{\underline{\lambda}^2}t} + \frac{2}{\sigma_{min}(\mathbf{S}_{z_i})}t\right)\right] \\
\geq&\mathbb{P}\left[\|\mathbf{A}\mathbf{e}_i\|_2^2 \geq 2\sigma_{z_i}^2(1+2\dot{C}\delta_b)\left(\sqrt{\frac{K\sigma_{min}(\mathbf{S}_{z_i})}{2\underline{\lambda}^2}} + \sqrt{\frac{t}{\sigma_{min}(\mathbf{S}_{z_i})}}\right)^2\right]
\end{aligned} \tag{90}$$

Define $\psi_0 = 2\psi_1 + 2\psi_2 + 6\psi_s + 4\dot{C}\delta_b$, $\psi_s = \sqrt{\frac{2K\sigma_{min}(\mathbf{S}_{z_i})\sigma_{z_i}}{\underline{\lambda}^2\mathbf{SNR}^2}} \leq \sqrt{\frac{K\sigma_{min}(\mathbf{S}_{z_i})^2}{2\bar{\lambda}^2 t}}$ and $t = \frac{(1-\psi_0)\mathbf{SNR}^2\sigma_{min}(\mathbf{S}_{z_i})}{2\sigma_{z_i}^2}$. Then we have

$$\begin{aligned}
&2\sigma_{z_i}^2(1+2\dot{C}\delta_b)\left(\sqrt{\frac{K\sigma_{min}(\mathbf{S}_{z_i})}{2\bar{\lambda}^2}} + \sqrt{\frac{t}{\sigma_{min}(\mathbf{S}_{z_i})}}\right)^2 \\
=&2\sigma_{z_i}^2(1+2\dot{C}\delta_b)\left(1 + \sqrt{\frac{K\sigma_{min}(\mathbf{S}_{z_i})^2}{2\bar{\lambda}^2 t}}\right)^2 \frac{t}{\sigma_{min}(\mathbf{S}_{z_i})} \\
\leq&(1+2\dot{C}\delta_b)(1+3\psi_s)(1-\psi_0)\mathbf{SNR}^2 \\
\leq&(1-2\psi_1-2\psi_2)\mathbf{SNR}^2 \\
\leq&\left((1-\psi_1-\psi_2)\,\mathbf{SNR}\right)^2
\end{aligned} \tag{91}$$

where we use the facts that $(1+x)^2 \leq 1+3x$, $1-2x \leq (1-x)^2$ when $x \leq 1$ and $(1+\dot{C}\delta_b)(1+3\psi_s) \leq 1+2\dot{C}\delta_b+6\psi_s$ when $\psi_0 \leq \frac{1}{2}$.

Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left[\left\|\mathbf{S}_{z_i}^{-1/2}\mathbf{V}^\top\widehat{\mathbf{V}}^{-i}\widehat{\mathbf{V}}^{-i^\top}\mathbf{e}_i\right\| \geq (1-\psi_1-\psi_2)\,\mathbf{SNR}\right] \\
&\leq \exp(-t) = \exp\left\{-\frac{(1-\psi_0)\mathbf{SNR}^2\sigma_{min}(\mathbf{S}_{z_i})}{2\sigma_{z_i}^2}\right\}
\end{aligned} \tag{92}$$

Combining (88), (84) and (87), we complete the proof.

C Lemmas

Lemma 3. *Assume that* $e_i \sim \boldsymbol{SG}\left(\sigma_{z_i}^2\right)$*. With probability at least* $1-O(n^{-\alpha-1})$ *it holds for all* $k \in [K]$*:*

$$\left\|\mathbf{V}^\top\left(\frac{1}{\sum_{i\in[n]}1\{z_i=k\}}\sum_{i\in[n]}1\{z_i=k\}\mathbf{e}_i\mathbf{e}_i^\top-\Sigma_k\right)\mathbf{V}\right\| \lesssim \sigma_k^2\sqrt{\frac{(K+\alpha\log n)\,K}{n\beta}} \tag{93}$$

Proof: By (4.22) in Vershynin (2018) it is proved.

Lemma 4. *Assume that* $\mathbf{e}_i \sim \boldsymbol{SG}(\sigma_{z_i}^2)$*,* $\bar{\sigma}\sqrt{\frac{K(K+\alpha\log n)}{\beta n}} < 1$*,*$K < n$*,* $\delta_b < \frac{1}{2}$ *and with probability* $1-O(n^{-\alpha})$ *it holds that* $l(\widehat{z}^{(0)}, z) \le \frac{\beta}{eK}$*. Then with probability at least* $1-O(n^{-\alpha})$ *where* $\alpha > 1$ *it holds for all* $k\in[K]$ *and :*

$$\left\|\mathbf{V}^\top\left(\widehat{\theta}_k^{(0)}-\widehat{\theta}_k(z)\right)\right\|_2 \lesssim \frac{\bar{\sigma}K\sqrt{\zeta\left(K+\alpha\zeta\log\frac{n}{\zeta}\right)}}{n\beta}+\frac{K\zeta\bar{\triangle}}{n\beta} =: \delta_e \tag{94}$$

Proof. Similar to the proof of (118) in Gao and Zhang (2019), for arbitrary $k\in[K]$ we decompose $\left\|\mathbf{V}^\top\left(\widehat{\theta}_k^{(0)}-\widehat{\theta}_k(z)\right)\right\|_2$ as follows:

$$\begin{aligned}\left\|\mathbf{V}^\top\left(\widehat{\theta}_k^{(0)}-\widehat{\theta}_k(z)\right)\right\|_2 \le{}& \left\|\mathbf{V}^\top\left(\frac{\sum_{i\in[n]}1\{\widehat{z}_i=k\}\mathbf{e}_i}{\sum_{i\in[n]}1\{\widehat{z}_i=k\}}-\frac{\sigma\sum_{i\in[n]}1\{z_i=k\}\mathbf{e}_i}{\sum_{i\in[n]}1\{z_i=k\}}\right)\right\| \\ &+\left\|\mathbf{V}^\top\left(\frac{\sum_{i\in[n]}1\{\widehat{z}_i=k\}\theta_{z_i}}{\sum_{i\in[n]}1\{\widehat{z}_i=k\}}-\theta_k\right)\right\|\end{aligned} \tag{95}$$

For the term $\left\|\mathbf{V}^\top\left(\frac{\sum_{i\in[n]}1\{\widehat{z}_i=k\}\mathbf{e}_i}{\sum_{i\in[n]}1\{\widehat{z}_i=k\}}-\frac{\sum_{i\in[n]}1\{z_i=k\}\mathbf{e}_i}{\sum_{i\in[n]}1\{z_i=k\}}\right)\right\|$, with probability at least $1-O(n^{-\alpha-1})$ it has:

$$\begin{aligned}&\left\|\mathbf{V}^\top\left(\frac{\sum_{i\in[n]}1\{\widehat{z}_i=k\}\mathbf{e}_i}{\sum_{i\in[n]}1\{\widehat{z}_i=k\}}-\frac{\sum_{i\in[n]}1\{z_i=k\}\mathbf{e}_i}{\sum_{i\in[n]}1\{z_i=k\}}\right)\right\| \\ \le&\left\|\mathbf{V}^\top\left(\frac{\sum_{i\in[n]}1\{\widehat{z}_i=k\}\mathbf{e}_i}{\sum_{i\in[n]}1\{\widehat{z}_i=k\}}-\frac{\sum_{i\in[n]}1\{z_i=k\}\mathbf{e}_i}{\sum_{i\in[n]}1\{\widehat{z}=k\}}\right)\right\| \\ +&\left\|\mathbf{V}^\top\left(\frac{\sum_{i\in[n]}1\{z_i=k\}\mathbf{e}_i}{\sum_{i\in[n]}1\{\widehat{z}_i=k\}}-\frac{\sum_{i\in[n]}1\{z_i=k\}\mathbf{e}_i}{\sum_{i\in[n]}1\{z_i=k\}}\right)\right\| \\ \lesssim&\frac{\bar{\sigma}K\sqrt{\zeta\left(K+\alpha\zeta\log\frac{n}{\zeta}\right)}}{n\beta}+\frac{\sigma_kK^{3/2}\zeta\sqrt{K+\alpha\log n}}{n^{3/2}\beta^{3/2}} \\ \lesssim&\frac{\bar{\sigma}K\sqrt{\zeta\left(K+\alpha\zeta\log\frac{n}{\zeta}\right)}}{n\beta}\end{aligned} \tag{96}$$

where the last two lines hold by arguments similar to lemma 6 under the assumption $l(\widehat{z}^{(0)}, z) \leq \frac{\beta}{eK}$ and the fact that $\zeta \log \frac{n}{\zeta} \geq \log n$ when $\zeta < \frac{n}{e}$.

For the second term $\left\| \mathbf{V}^\top \left(\frac{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}\theta_{z_i}}{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}} - \theta_k \right) \right\|_2$, it has:

$$\begin{aligned} & \left\| \mathbf{V}^\top \left(\frac{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}\theta_{z_i}}{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}} - \theta_k \right) \right\|_2 \\ \leq & \left\| \mathbf{V}^\top \left(\frac{\sum_{i\in[n]} 1\{z_i = k, \widehat{z}_i \neq k\} \left(\theta_{\widehat{z}_i} - \theta_k\right)}{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}} \right) \right\|_2 \\ \lesssim & \frac{K\zeta\bar{\triangle}}{n\beta} \end{aligned} \tag{97}$$

Combining the above bound, with probability at least $1 - O(n^{-\alpha-1})$ it has

$$\left\| \mathbf{V}^\top \left(\widehat{\theta}_k^{(0)} - \widehat{\theta}_k(z) \right) \right\|_2 \lesssim \frac{\bar{\sigma} K \sqrt{\zeta \left(K + \alpha\zeta \log \frac{n}{\zeta} \right)}}{n\beta} + \frac{K\zeta\bar{\triangle}}{n\beta} \tag{98}$$

which lead to the conclusion. □

Lemma 5. *Considering the setting in lemma 3, for arbitrary $k \in [K]$ with probability at least $1 - O(n^{-\alpha})$ it has:*

$$\left\| \widehat{\theta}_k - \theta_k \right\| \lesssim \left(\frac{\bar{\sigma} K \sqrt{\zeta(J+n)}}{n\beta} + \frac{K\zeta\bar{\triangle}}{n\beta} + \sqrt{\frac{K\left(\alpha \log n + J\right)}{n\beta}}\sigma_k \right) =: \xi_g \tag{99}$$

Proof. See the arguments of (118) in Gao and Zhang (2019) and (120) . □

The following lemma is analogous to Lemma 7.4 in Chen and Zhang (2021).

Lemma 6. *Assume $\mathbf{e}_i \sim \boldsymbol{SG}(\bar{\sigma}^2)$ and $s = c_1 n$ where c_1 is a sufficiently small constant. Then for any constant C, with probability at least $1 - n^{-C'}$ where C' is a constant determined by C, it has:*

$$\max_{T\subset[n]:|T|\leq s} \frac{1}{\bar{\sigma}^2 \left(|T| \log \frac{n}{|T|} + K \right)} \left\| \sum_{i\in[T]} \mathbf{V}^\top \mathbf{e}_i \mathbf{e}_i^\top \mathbf{V} \right\| \leq 3C' + 2C_0 \tag{100}$$

Proof. Consider a fixed $T \subset [n]$ where $|T| \leq s$. We can take the ϵ covering of S^{K-1} with $\epsilon < \frac{1}{4}$ and $|N_\epsilon| \leq 9^K$. Then we have:

$$\left\| \sum_{i\in T} \mathbf{V}^\top \mathbf{e}_i \mathbf{e}_i^\top \mathbf{V} \right\| = \sup_{\|\mathbf{w}\|=1} \sum_{j\in[T]} \left(\mathbf{w}^\top \mathbf{V}^\top \mathbf{e}_j \right) \leq 2 \max_{\mathbf{w}\in N_\epsilon} \sum_{i\in T} \left(\mathbf{w}^\top \mathbf{V}^\top \mathbf{e}_i \right)^2 \tag{101}$$

By Hsu et al. (2012), it has :

$$\mathbb{P}\left(\sum_{i\in T}\left(\mathbf{w}^\top\mathbf{V}^\top\mathbf{e}_i\right)^2 \geq \bar{\sigma}^2\left(|T|+2\sqrt{|T|t}+2t\right)\right) \leq \exp\left(-t\right) \tag{102}$$

There exists constant C_0 such that $|T| \leq C_0|T|\log\frac{n}{|T|}$. We can take $t = \widetilde{C}\left(a\log\frac{n}{a}+K\right)$ where $\widetilde{C} = \frac{C}{3} - \frac{2C_0}{3}$. Thus we have:

$$\mathbb{P}\left(\sum_{i\in T}\left(\mathbf{w}^\top\mathbf{V}^\top\mathbf{e}_i\right)^2 \geq C\bar{\sigma}^2\left(|T|\log\frac{n}{|T|}+K\right)\right) \leq \exp\left(-\widetilde{C}\left(a\log\frac{n}{a}+d\right)\right) \tag{103}$$

Hence,

$$\mathbb{P}\left(\left\|\sum_{i\in T}\mathbf{V}^\top\mathbf{e}_i\mathbf{e}_i^\top\mathbf{V}\right\| \geq 2C\bar{\sigma}^2\left(|T|\log\frac{n}{|T|}+K\right)\right) \leq 9^K\exp\left(-\widetilde{C}\left(a\log\frac{n}{a}+d\right)\right) \tag{104}$$

And we have:

$$\begin{aligned}\mathbb{P}\left(\max_{T\subset[n],1\leq|T|\leq s}\frac{\left\|\sum_{i\in T}\mathbf{V}^\top\mathbf{e}_i\mathbf{e}_i^\top\mathbf{V}\right\|}{2\bar{\sigma}^2\left(|T|\log\frac{n}{|T|}+K\right)} \geq C\right) &\leq \sum_{a\in[s]}\binom{n}{a}9^K\exp\left(-\widetilde{C}\left(a\log\frac{n}{a}+K\right)\right)\\ &\leq n^{-C'}\end{aligned} \tag{105}$$

where $C' = \widetilde{C} - 3$. □

Lemma 7. *Combining the assumptions in lemma 4, we further assume that* $\mathbf{e}_i \sim \boldsymbol{SG}\left(\sigma_{z_i}^2\right)$, $\bar{\sigma}\sqrt{\frac{K(\alpha\log n+K)}{\beta n}} \leq 1$, $\frac{K\zeta\bar{\triangle}}{n\beta} \leq 1$, *and* $\sigma_{max}(\mathbf{S}_k) \leq \bar{\lambda}$, $\sigma_{min}(\mathbf{S}_k) \geq \underline{\lambda}$, $\|\mathbf{\Sigma}_k\| \leq \Gamma$ *holds for all* $k \in [K]$ *where* $\bar{\lambda}, \underline{\lambda}$ *are positive constants. Then with probability at least* $1 - O(n^{-\alpha})$ *it holds simultaneously for all* k *in* $[K]$*:*

$$\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k\mathbf{V}^\top\right\| \lesssim \delta_s \tag{106}$$

where δ_s *is defined as*

$$\begin{aligned}\delta_s := \sigma_k^2\sqrt{\frac{\alpha\log nK}{n\beta}} + \frac{\bar{\sigma}\alpha K}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2K^2}{n\beta} + \frac{K\bar{\triangle}^2\zeta}{n\beta}\\ + \delta_b^2\sigma_k^2\sqrt{\frac{\left(J+\alpha\log n\right)K}{n\beta}} + \delta_b^2\frac{\bar{\sigma}^2KJ}{n\beta}\end{aligned} \tag{107}$$

Proof. For notation simplicity, we denote the covariance estimation according to a certain cluster $\widetilde{z}$ by $\widehat{\Sigma}(\widetilde{z})$. By triangle inequality it has:

$$\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k\mathbf{V}^\top\right\| \lesssim \left\|\widetilde{\mathbf{S}}_k - \mathbf{S}_k\right\| + \left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\right\|^2 \left\|\widehat{\mathbf{\Sigma}}_k\right\| \tag{108}$$

where $\widetilde{\mathbf{S}}_k = \mathbf{V}^\top \widehat{\mathbf{\Sigma}}_k \mathbf{V}$.

For arbitrary $k \in [K]$, it has:

$$\left\|\widetilde{\mathbf{S}}_k - \mathbf{S}_k\right\| \leq \left\|\widetilde{\mathbf{S}}_k(z) - \mathbf{S}_k\right\| + \left\|\widetilde{\mathbf{S}}_k - \widetilde{\mathbf{S}}_k(z)\right\| \tag{109}$$

where $\widetilde{\mathbf{S}}(z) = \mathbf{V}\widehat{\mathbf{\Sigma}}(z)\mathbf{V}^\top$.

For the first part $\left\|\widetilde{\mathbf{S}}_k(z) - \mathbf{S}_k\right\|$, by lemma 3 with probability at least $1 - O(n^{-\alpha})$ it holds simultaneously for all $k \in [K]$ that

$$\left\|\widetilde{\mathbf{S}}_k(z) - \mathbf{S}_k\right\| \lesssim \sigma_k^2 \sqrt{\frac{(K + \alpha \log n)\, K}{n\beta}} \tag{110}$$

For the second part $\left\|\widetilde{\mathbf{S}}_k - \widetilde{\mathbf{S}}_k(z)\right\|$, it has:

$$\begin{aligned} &\left\|\widetilde{\mathbf{S}}_k - \widetilde{\mathbf{S}}_k(z)\right\| \\ &\leq K_1 + K_2 \end{aligned} \tag{111}$$

where

$$\begin{aligned} K_1 := &\left\| \frac{1}{\sum 1\{\widehat{z}_i = k\}} \sum_{i\in[n]} \left(1\{\widehat{z}_i^{(0)} = k\} \mathbf{V}^\top \left(\mathbf{R}_i - \widehat{\theta}^{(0)}_{\widehat{z}_i^{(0)}}\right)\left(\mathbf{R}_i - \widehat{\theta}^{(0)}_{\widehat{z}_i^{(0)}}\right)^\top \mathbf{V} \right.\right. \\ &\left.\left. - 1\{z_i = k\}\mathbf{V}^\top \left(\mathbf{R}_i - \widehat{\theta}_{z_i}(z)\right)\left(\mathbf{R}_i - \widehat{\theta}_{z_i}(z)\right)^\top \mathbf{V} \right) \right\| \end{aligned} \tag{112}$$

$$K_2 := \left\| \left(\frac{1}{\sum 1\{\widehat{z}_i^{(0)} = k\}} - \frac{1}{\sum 1\{z_i = k\}} \right) \sum_{i\in[n]} 1\{z_i = k\}\mathbf{V}^\top \left(\mathbf{R}_i - \widehat{\theta}_{z_i}(z)\right)\left(\mathbf{R}_i - \widehat{\theta}_{z_i}(z)\right)^\top \mathbf{V} \right\| \tag{113}$$

For the term K_1, it could be decomposed as:

$$K_1 \leq L_1 + L_2 + L_3 \tag{114}$$

where

$$L_1$$

$$:= \left\| \frac{1}{\sum 1\{\widehat{z}_i = k\}} \sum_{i\in[n]} 1\{\widehat{z}_i = z_i = k\} \mathbf{V}^\top \left(\left(\mathbf{R}_i - \widehat{\theta}^{(0)}_{\widehat{z}_i}\right) \left(\mathbf{R}_i - \widehat{\theta}^{(0)}_{\widehat{z}_i}\right)^\top - \left(\mathbf{R}_i - \widehat{\theta}_{z_i}(z)\right) \left(\mathbf{R}_i - \widehat{\theta}_{z_i}(z)\right)^\top \right) \mathbf{V} \right\| \tag{115}$$

$$L_2 := \left\| \frac{1}{\sum 1\{\widehat{z}_i = k\}} \sum_{i\in[n]} 1\{\widehat{z}_i = k, z_i \neq k\} \mathbf{V}^\top \left(\mathbf{R}_i - \widehat{\theta}^{(0)}_{\widehat{z}_i}\right) \left(\mathbf{R}_i - \widehat{\theta}^{(0)}_{\widehat{z}_i}\right)^\top \mathbf{V} \right\| \tag{116}$$

$$L_3 := \left\| \frac{1}{\sum 1\{\widehat{z}_i = k\}} \sum_{i\in[n]} 1\{\widehat{z}_i \neq k, z_i = k\} \mathbf{V}^\top \left(\mathbf{R}_i - \widehat{\theta}_{z_i}(z)\right) \left(\mathbf{R}_i - \widehat{\theta}_{z_i}(z)\right)^\top \mathbf{V} \right\| \tag{117}$$

For L_1, it has:

$$\begin{aligned}
L_1 \leq & \left\| \frac{1}{\sum 1\{\widehat{z}_i = k\}} \sum_{i\in[n]} 1_{\widehat{z}_i = z_i = k} \mathbf{V}^\top \left(\widehat{\theta}_{\widehat{z}_i}(z) - \widehat{\theta}^{(0)}_{\widehat{z}_i}\right) \left(\widehat{\theta}_{\widehat{z}_i}(z) - \widehat{\theta}^{(0)}_{\widehat{z}_i}\right)^\top \mathbf{V} \right\| \\
& + 2 \left\| \frac{1}{\sum 1\{\widehat{z}_i = k\}} \sum_{i\in[n]} 1_{\widehat{z}_i = z_i = k} \mathbf{V}^\top \left(\mathbf{R}_i - \widehat{\theta}_{\widehat{z}_i}(z)\right) \left(\widehat{\theta}_{\widehat{z}_i}(z) - \widehat{\theta}^{(0)}_{\widehat{z}_i}\right)^\top \mathbf{V} \right\| \\
\lesssim & \left\| \mathbf{V}^\top \left(\widehat{\theta}^{(0)}_k - \widehat{\theta}_k(z)\right) \right\|_2^2 \frac{\sum_{i\in[n]} 1\{z_i = k\}}{\sum_{i\in[n]} 1\{\widehat{z}^{(0)}_i = k\}} \\
& + \left\| \mathbf{V}^\top \left(\theta_k - \widehat{\theta}_k(z)\right) \right\|_2 \left\| \mathbf{V}^\top \left(\widehat{\theta}^{(0)}_k - \widehat{\theta}_k(z)\right) \right\|_2 \frac{\sum_{i\in[n]} 1\{z_i = k\}}{\sum_{i\in[n]} 1\{\widehat{z}^{(0)}_i = k\}} \\
& + \left\| \mathbf{V}^\top \left(\widehat{\theta}^{(0)}_k - \widehat{\theta}_k(z)\right) \right\|_2 \left\| \frac{1}{\sum_{i\in[n]} 1\{\widehat{z}^{(0)}_i = k\}} \sum_{i\in[n]} 1\{z_i = \widehat{z}^{(0)}_i = k\} \mathbf{e}_i \mathbf{V} \right\|
\end{aligned} \tag{118}$$

By lemma 4 in (118) with probability at least $1 - O(n^{-\alpha})$ it has:

$$\left\| \mathbf{V}^\top \left(\widehat{\theta}^{(0)}_k - \widehat{\theta}_k(z)\right) \right\|_2 \lesssim \frac{\bar{\sigma} K \sqrt{\zeta \left(K + \alpha \zeta \log \frac{n}{\zeta}\right)}}{n\beta} + \frac{K \zeta \bar{\triangle}}{n\beta} =: \delta_e \tag{119}$$

By the sub-gaussian property of random vector, it is straightforward to derive that

$$\left\| \mathbf{V} \left(\widehat{\theta}_k(z) - \theta_k\right) \right\|_2 \lesssim \sqrt{\frac{K\left(\alpha \log n + K\right)}{n\beta}} \sigma_k \tag{120}$$

with probability at least $1 - O(n^{-\alpha})$ for all $k \in [K]$. Hence, with probability at least

$1 - O(n^{-\alpha})$ it holds for all $k \in [K]$ that

$$\left\| \mathbf{V} \left(\widehat{\theta}_k^{(0)} - \theta_k \right) \right\|_2 \lesssim \sigma_k \sqrt{\frac{K\left(\alpha \log n + K\right)}{n\beta}} + \frac{\bar{\sigma} K \sqrt{\zeta \left(K + \alpha \zeta \log \frac{n}{\zeta} \right)}}{n\beta} + \frac{K \zeta \bar{\triangle}}{n\beta} =: \xi_e \quad (121)$$

Additionally, by the assumption $l(\widehat{z}^{(0)}, z) \leq \frac{\beta}{eK}$, we have

$$\frac{\sum_{i\in[n]} 1\{z_i = k\}}{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}} \leq 2 \quad (122)$$

And by (122) and procedures similar to those in 6, it holds with probability at least $1 - O(n^{-\alpha})$ for all $k \in [K]$ uniformly

$$\left\| \frac{1}{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}} \mathbf{V}^\top \sum_{i\in[n]} 1\{z_i = \widehat{z}_i = k\} \mathbf{e}_i \right\|_2 \lesssim \frac{\sigma_k \sqrt{K + \zeta \log \frac{n}{\zeta}}}{\sqrt{\frac{n\beta}{K}}} \lesssim 1 \quad (123)$$

By (118) and the assumption $\sigma_k \sqrt{\frac{K(K+\alpha \log n)}{n\beta}} < 1$ we derive that:

$$L_1 \lesssim \left\| V \left(\widehat{\theta}_k^{(0)} - \widehat{\theta}_k(z) \right) \right\|_2 = \delta_e \quad (124)$$

Now we turn to the term L_2. Decompose it similarly:

$$L_2 \lesssim M_1 + M_2 \quad (125)$$

where

$$M_1 := \left\| \frac{1}{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}} \sum_{i\in[n]} 1\{\widehat{z}_i = k, z_i \neq k\} V^\top \left(R_i - \theta_k\right) \left(R_i - \theta_k\right)^\top V \right\| \quad (126)$$

and

$$M_2 = \left\| \frac{1}{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}} \sum_{i\in[n]} 1\{\widehat{z}_i = k, z_i \neq k\} V^\top \left(\widehat{\theta}_k^{(0)} - \theta_k \right) \left(\widehat{\theta}_k^{(0)} - \theta_k \right)^\top V \right\| \quad (127)$$

For the term M_1, by lemma 6, with probability at least $1 - O(n^{\alpha})$ it uniformly holds for all $k \in [K]$

$$
\begin{aligned}
M_1 &\lesssim \left\| \frac{1}{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}} \sum_{i\in[n]} 1\{\widehat{z}_i = k, z_i \neq k\} V^\top e_i e_i^\top V \right\| \\
&+ \left\| \frac{1}{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}} \sum_{i\in[n]} 1\{\widehat{z}_i = k, z_i \neq k\} \left(\theta_{z_i} - \theta_k\right)\left(\theta_{z_i} - \theta_k\right)^\top \right\| \\
&\lesssim \frac{\bar{\sigma}^2 \left(\zeta \log\left(\frac{n}{\zeta}\right) + K\right)}{\beta n / K} + \frac{K\zeta \bar{\triangle}^2}{n\beta} \qquad (128) \\
&\overset{(1)}{\lesssim} \frac{\bar{\sigma}^2}{\beta} K \sqrt{\frac{\zeta}{n}} \sqrt{\frac{\zeta}{n}\left(\log\left(\frac{n}{\zeta}\right)\right)^2} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\zeta\bar{\triangle}^2}{n\beta} \\
&\lesssim \frac{\bar{\sigma}^2}{\beta} K \sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\zeta\bar{\triangle}^2}{n\beta}
\end{aligned}
$$

where (1) follows since

$$
x \log \frac{n}{x} \le \sqrt{x} \sqrt{x(\log \frac{n}{x})^2} \le \frac{2}{e}\sqrt{nx} \qquad (129)
$$

For the term M_2, with probabiliity at least $1 - O(n^{-\alpha})$ it holds simultaneously for all $k \in [K]$

$$
M_2 \lesssim \xi_e^2 \frac{K\zeta}{n\beta} \lesssim \frac{K\zeta}{n\beta} \lesssim \frac{K\zeta\bar{\triangle}^2}{n\beta} \qquad (130)
$$

by (121) where ξ_e is also defined there.

Hence, with probability at least $1 - O(n^{-\alpha})$, uniformlly for all $k \in [K]$ it has:

$$
L_2 \lesssim \frac{\bar{\sigma}^2}{\beta} K \sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\zeta\bar{\triangle}^2}{n\beta} \qquad (131)
$$

Analogue to the procedures above, for the term L_3, we have:

$$
\begin{aligned}
L_3 &\lesssim \left\| \frac{1}{\sum_{i\in[n]} 1\{\widehat{z}_i = k\}} \sum_{i\in[n]} 1\{\widehat{z}_i \neq k, z_i = k\} \mathbf{V}^\top \mathbf{e}_i \mathbf{e}_i^\top \mathbf{V} \right\| + \frac{K\zeta}{n\beta} \left\| \theta_k - \widehat{\theta}_k(z) \right\|_2^2 \\
&\lesssim \frac{\bar{\sigma}^2 K}{\beta} \sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\zeta}{n\beta}\xi_e^2 \qquad (132) \\
&\lesssim \frac{\bar{\sigma}^2 K}{\beta} \sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta}
\end{aligned}
$$

where the last inequality follows since $\frac{\bar{\sigma}^2 K}{\beta}\sqrt{\frac{\zeta}{n}} \ge \frac{\bar{\sigma}^2 K\zeta}{n\beta} \ge \frac{K\zeta}{n\beta}\xi_e^2$.

Therefore, with probability at least $1-O(n^{-\alpha})$, it holds for all $k \in [K]$ that

$$
\begin{aligned}
K_1 \leq & L_1 + L_2 + L_3 \\
\lesssim & \delta_e + \frac{\bar{\sigma}^2}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\bar{\triangle}^2\zeta}{n\beta} \\
= & \frac{\bar{\sigma} K \sqrt{\zeta\left(K + \alpha\zeta \log\frac{n}{\zeta}\right)}}{n\beta} + \frac{\bar{\sigma}^2}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\bar{\triangle}^2\zeta}{n\beta} \\
\overset{(a)}{\lesssim} & \frac{\bar{\sigma} K\sqrt{\zeta K}}{n\beta} + \frac{\alpha\bar{\sigma} K \zeta^{3/4}}{n^{3/4}\beta} + \frac{\bar{\sigma}^2}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\bar{\triangle}^2\zeta}{n\beta} \\
\lesssim & \frac{\bar{\sigma}\alpha K}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\bar{\triangle}^2\zeta}{n\beta}
\end{aligned}
\tag{133}
$$

where (a) follows by (129) and the last line follows since $\bar{\sigma} < 1, \alpha > 1, \sqrt{\zeta} < \zeta, \frac{\zeta}{n} < 1$.

For the second part K_2, it has:

$$
K_2 = \frac{\left|\sum_{i\in[n]} 1\{z_i = k\} - \sum_{i\in[n]} 1\{\widehat{z}_i = k\}\right|}{\sum_{i\in[n]} 1\{z_i = k\}} \left\|\widetilde{\mathbf{S}}_k(z)\right\| \lesssim \frac{K\zeta}{n\beta}\left\|\mathbf{S}_k\right\| + \frac{K\zeta}{n\beta}\left\|\mathbf{S}_k - \widetilde{\mathbf{S}}_k(z)\right\| \tag{134}
$$

Combining the above bounds for all $k \in [K]$ with probability at least $1-O(n^{-\alpha})$ it has

$$
\left\|\widetilde{\mathbf{S}}_k - \widetilde{\mathbf{S}}_k(z)\right\| \lesssim \left(\frac{\bar{\sigma}\alpha K}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\bar{\triangle}^2\zeta}{n\beta}\right) + \frac{K\zeta}{n\beta} + \frac{K\zeta}{n\beta}\left\|\mathbf{S}_k - \widetilde{\mathbf{S}}_k(z)\right\| \tag{135}
$$

by the assumption that $\|\mathbf{S}_k\| \leq \bar{\lambda}$ for all $k \in [K]$.

Hence with probability at least $1-O(n^{-\alpha})$ for all $k \in [K]$ it has

$$
\left\|\widetilde{\mathbf{S}}_k - \widetilde{\mathbf{S}}_k(z)\right\| \lesssim \frac{\bar{\sigma}\alpha K}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\bar{\triangle}^2\zeta}{n\beta} \tag{136}
$$

since $\frac{K\zeta}{n\beta} \leq \frac{1}{2}$.

and by lemma 3 it has

$$
\begin{aligned}
\left\|\widetilde{\mathbf{S}}_k - \mathbf{S}_k\right\| &\lesssim \sigma_k^2\sqrt{\frac{(K+\alpha\log n)K}{n\beta}} + \frac{\bar{\sigma}\alpha K}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\bar{\triangle}^2\zeta}{n\beta} \\
&\lesssim \sigma_k^2\sqrt{\frac{\alpha\log nK}{n\beta}} + \frac{\bar{\sigma}\alpha K}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2 K^2}{n\beta} + \frac{K\bar{\triangle}^2\zeta}{n\beta}
\end{aligned}
\tag{137}
$$

For the second part $\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\right\|^2 \left\|\widehat{\mathbf{\Sigma}}_k\right\|$, by some arguments similar to above, with probability at least $1-O(n^{-\alpha})$ it holds for all $k \in [K]$ simultaneously that

$$
\left\|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k\right\| \lesssim \sigma_k^2\sqrt{\frac{(J+\alpha\log n)K}{n\beta}} + \frac{\bar{\sigma}^2 KJ}{n\beta} + \frac{K\zeta\bar{\triangle}^2}{n\beta} + \frac{K\zeta}{n\beta}\left\|\mathbf{\Sigma}_k\right\| \tag{138}
$$

Then it follows that:

$$\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\right\|^2 \left\|\widehat{\boldsymbol{\Sigma}}_k\right\| \lesssim \delta_b^2 \left(\sigma_k^2\sqrt{\frac{(J+\alpha\log n)\,K}{n\beta}} + \frac{\bar{\sigma}^2 KJ}{n\beta} + \frac{K\zeta\bar{\triangle}^2}{n\beta} + \frac{K\zeta}{n\beta}\|\boldsymbol{\Sigma}_k\|\right) \tag{139}$$

To conclude, with probability at least $1-O(n^{-\alpha})$ it has:

$$\begin{aligned}
\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k\mathbf{V}^\top\right\| &\lesssim \sigma_k^2\sqrt{\frac{\alpha\log nK}{n\beta}} + \frac{\bar{\sigma}\alpha K}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2K^2}{n\beta} + \frac{K\bar{\triangle}^2\zeta}{n\beta} \\
&\quad + \delta_b^2\left(\sigma_k^2\sqrt{\frac{(J+\alpha\log n)\,K}{n\beta}} + \frac{\bar{\sigma}^2 KJ}{n\beta} + \frac{K\zeta\bar{\triangle}^2}{n\beta} + \frac{K\zeta}{n\beta}\|\boldsymbol{\Sigma}_k\|\right) \\
&\lesssim \sigma_k^2\sqrt{\frac{\alpha\log nK}{n\beta}} + \frac{\bar{\sigma}\alpha K}{\beta}\sqrt{\frac{\zeta}{n}} + \frac{\bar{\sigma}^2K^2}{n\beta} + \frac{K\bar{\triangle}^2\zeta}{n\beta} \\
&\quad + \delta_b^2\sigma_k^2\sqrt{\frac{(J+\alpha\log n)\,K}{n\beta}} + \delta_b^2\frac{\bar{\sigma}^2 KJ}{n\beta} \\
&= \delta_s
\end{aligned} \tag{140}$$

by the assumption that $\delta_b < 1$ and $\|\boldsymbol{\Sigma}_k\| \leq \Gamma$ for all $k \in [K]$.

□

Lemma 8. *With the assumptions in lemma 7,we additionally assume that $\frac{\delta_s}{\underline{\lambda}} < \frac{1}{2}$. Then, with probability at least $1-O(n^{-\alpha})$ it has for all $k \in [K]$*

$$\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\right\| \lesssim \frac{\delta_s}{\underline{\lambda}^2} \tag{141}$$

Proof. It is straightforward to derive that

$$\begin{aligned}
&\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\right\| \\
\leq &\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k^{-1}\widehat{\mathbf{V}}^\top\right\| \left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k^{1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k^{1}\mathbf{V}^\top\right\| \|\mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\| \\
\leq &\left(\|\mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\| + \left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\right\|\right) \left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k\mathbf{V}^\top\right\| \|\mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\| \\
\leq &\|\mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\| \left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k\mathbf{V}^\top\right\| \|\mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\| + \left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\right\| \frac{\delta_s}{\underline{\lambda}}
\end{aligned} \tag{142}$$

which yeilds that

$$\left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k^{-1}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\right\| \lesssim \|\mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\| \left\|\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k\mathbf{V}^\top\right\| \|\mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top\| \lesssim \frac{\delta_s}{\underline{\lambda}^2} \tag{143}$$

□

Lemma 9. *With the assumptions in Lemma 8, for all $k \in [K]$ with probability at least $1 - O(n^{-\alpha})$ there exists an orthogonal matrix $\mathbf{H}_k$ such that*

$$\left\| \mathbf{V}\mathbf{S}_k^{-1/2}\mathbf{V}^\top \widehat{\mathbf{V}}\widehat{\mathbf{S}}_k^{1/2}\widehat{\mathbf{V}}^\top - \mathbf{H}_k \right\| \lesssim \sqrt{\frac{\delta_s}{\underline{\lambda}}} \tag{144}$$

Proof. Firstly, we could easily see that

$$\begin{aligned} &\left\| \mathbf{V}\mathbf{S}_k^{-1/2}\mathbf{V}^\top \widehat{\mathbf{V}}\widehat{\mathbf{S}}_k\widehat{\mathbf{V}}^\top \mathbf{V}\mathbf{S}_k^{-1/2}\mathbf{V}^\top - \mathbf{I} \right\| = \left\| \mathbf{V}\mathbf{S}_k^{-1/2}\mathbf{V}^\top \left(\widehat{\mathbf{V}}\widehat{\mathbf{S}}_k\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k\mathbf{V}^\top \right) \mathbf{V}\mathbf{S}_k^{-1/2}\mathbf{V}^\top \right\| \\ &\leq \left\| \mathbf{V}\mathbf{S}_k^{-1}\mathbf{V}^\top \right\| \left\| \widehat{\mathbf{V}}\widehat{\mathbf{S}}_k\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{S}_k\mathbf{V}^\top \right\| \lesssim \frac{\delta_s}{\underline{\lambda}} \end{aligned} \tag{145}$$

Hence there exists an orthogonal matrix $\mathbf{H}_k$ such that

$$\left\| \mathbf{V}\mathbf{S}_k^{-1/2}\mathbf{V}^\top \widehat{\mathbf{V}}\widehat{\mathbf{S}}_k^{1/2}\widehat{\mathbf{V}}^\top - \mathbf{H}_k \right\| \lesssim \sqrt{\frac{\delta_s}{\underline{\lambda}}} \tag{146}$$

□

D Proofs for Section 10

Proof of Lemma 2 We have:

$$\begin{aligned} \sum_i \left\| \widehat{\mathbf{U}}_i - \widehat{\mathbf{c}}_{\widehat{z}_i} \right\|_2 &\leq \sqrt{n \sum_i \left\| \widehat{\mathbf{U}}_i - \widehat{\mathbf{c}}_{\widehat{z}_i} \right\|_2^2} \\ &\leq \sqrt{n \sum_i \left\| \mathbf{U}_i - \widehat{\mathbf{U}}_i \right\|_2^2} \end{aligned} \tag{147}$$

$$\sum_i \left\| \widehat{\mathbf{U}}_i - \mathbf{U}_i \right\|_2 \leq \sqrt{n \sum_i \left\| \mathbf{U}_i - \widehat{\mathbf{U}}_i \right\|_2^2} \tag{148}$$

which yields that

$$\sum_i \left\| \mathbf{U}_i - \widehat{\mathbf{c}}_{\widehat{z}_i} \right\|_2 \leq 2\sqrt{n \sum_i \left\| \mathbf{U}_i - \widehat{\mathbf{U}}_i \right\|_2^2} \leq 2n \left\| \widehat{\mathbf{U}} - \mathbf{U} \right\|_{2,\infty} \tag{149}$$

On the other hand, letting $r_k = \min_{i \in [n]: z_i = k, j \in [K]} \left\| \mathbf{U}_{z_i} - \widehat{\mathbf{c}}_j \right\|$ for all $k \in [K]$, we have:

$$\sum_i \min_{r \in [L]} \left\| \mathbf{U}_i - \widehat{\mathbf{c}}_r \right\|_2 = n \sum_{s=1}^{L} \pi_s \left\| \mathbf{c}_s^* - \widehat{\mathbf{c}}_{r_s} \right\|_2 \leq \sum_i \left\| \widehat{\mathbf{U}}_i - \widehat{\mathbf{c}}_{z_i} \right\|_2 \tag{150}$$

which implies that for $\forall s \in [K]$, it holds that

$$\min_s \pi_s \left\|\mathbf{c}_s^* - \widehat{\mathbf{c}}_{r_s}\right\|_2 \le 2 \left\|U - U^*\right\|_{2,\infty} \tag{151}$$

For $s_1, s_2 \in [K], s_1 \neq s_2$, it holds that

$$\begin{aligned}
\left\|\widehat{v}_{r_{s_1}} - \widehat{v}_{r_{s_2}}\right\|_2 &\ge \left\|v_{s_1}^* - v_{s_2}^*\right\|_2 - \left\|v_{s_1}^* - \widehat{v}_{r_{s_1}}\right\|_2 - \left\|v_{s_2}^* - \widehat{v}_{r_{s_2}}\right\|_2 \\
&\ge \min_{s,s'} \left\|v_s^* - v_{s'}^*\right\| - \frac{4}{\min_s \pi_s} \left\|U - U^*\right\|_{2,\infty} > 0
\end{aligned} \tag{152}$$

This implies that $\{r_s\}$ must be a permutation of $[L]$. WLOG, we assume that $r_s = s, \forall s \in [L]$.

It is left to prove that for any $i \in C_s$,

$$\arg\min_{r \in [K]} \left\|U_i - \widehat{v}_r\right\|_2 = s \tag{153}$$

By triangle inequality, we have:

$$\begin{aligned}
\left\|U_i - \widehat{v}_s\right\|_2 &\le \left\|U_i - U_i^*\right\|_2 + \left\|U_i^* - \widehat{v}_s\right\|_2 \\
&\le \left\|U - U^*\right\|_{2,\infty} + \frac{2}{\min_s \pi_s} \left\|U - U^*\right\|_{2,\infty} \\
&< \frac{3}{\min_s \pi_s} \left\|U - U^*\right\|_{2,\infty}
\end{aligned} \tag{154}$$

And for any $s' \neq s$, we have:

$$\left\|U_i - \widehat{v}_{s'}\right\| \ge \left\|v_s^* - v_{s'}^*\right\|_2 - \left\|v_{s'} - \widehat{v}_{s'}\right\|_2 - \left\|U_i - v_s^*\right\|_2 \ge \frac{3}{\min_s \pi_s} \left\|U - U^*\right\|_{2,\infty} \tag{155}$$

This completes the proof.

E Auxiliary Lemmas

Proof of Lemma 1 For the response matrix R, by definition we have:

$$\begin{aligned}
\triangle &= \min_{i \neq j \in [K]} \left\|\mathbf{\Theta}_{,i} - \mathbf{\Theta}_{,j}\right\|_2 \\
&= \min_{i \neq j \in [n]} \left\|\mathbf{R}_i^* - \mathbf{R}_{j,}^*\right\|_2 \\
&= \min_{i \neq j \in [n]} \left\|\mathbf{U}_i \mathbf{D} - \mathbf{U}_j \mathbf{D}\right\|_2
\end{aligned} \tag{156}$$

where $(\mathbf{U}, \mathbf{D}, \mathbf{V})$ is the top-K SVD of R^*. Then we have:

$$\max_{i \neq j \in [n]} \left\|\mathbf{U}_i - \mathbf{U}_j\right\|_2 \sigma_1^* \ge \triangle \ge \min_{i \neq j \in [n]} \left\|\mathbf{U}_i - \mathbf{U}_j\right\|_2 \sigma_K^* \tag{157}$$

that is:

$$\max_{i,j:z_i\neq z_j} \sqrt{n}\left\|\mathbf{U}_i-\mathbf{U}_j\right\|_2 \kappa\sigma^*_{min} \geq \triangle \geq \min_{i,j:z_i\neq z_j} \sqrt{n}\left\|\mathbf{U}_i-\mathbf{U}_j\right\|_2 \sigma^*_{min} \tag{158}$$

where σ^*_{min} is the smallest singular value of $\mathbf{G}^{1/2}\mathbf{\Theta}^\top$. (Recall that $G = diag\left(\sqrt{\pi_1},\cdots,\sqrt{\pi_K}\right)$).

Since $\left\|\mathbf{U}_i-\mathbf{U}_j\right\|_2 = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{\pi_i}+\frac{1}{\pi_j}}$ when $z_i \neq z_j$ (we know that the rows of $\mathbf{U}$ are orthogonal with each other by 10), we have:

$$\kappa\sigma^*_{min}\max_{i,j:z_i\neq z_j}\sqrt{\frac{1}{\pi_i}+\frac{1}{\pi_j}} \geq \triangle \geq \sigma^*_{min}\min_{i,j:z_i\neq z_j}\sqrt{\frac{1}{\pi_i}+\frac{1}{\pi_j}} \tag{159}$$

Lemma 10 (Theorem 2.1 in Buldygin and Moskvichova (2013))**.** *Let X be a Bernoulli random variable with $\mathbb{P}[X=1]=p$, $\mathbb{P}[X=0]=p$. Then it has*

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \exp\left\{\frac{K(p)\lambda^2}{2}\right\} \tag{160}$$

where

$$K(p) = \begin{cases} 0 & p \in \{0,1\} \\ 1/4 & p = \frac{1}{2} \\ \frac{p-q}{2(\log p - \log q)} & p \in (0,1)\setminus\{\frac{1}{2}\} \end{cases} \tag{161}$$

Lemma 11. *Suppose $\mathbf{X}\in\mathbb{R}^m$ is a random vector such that $\mathbb{P}\left[\mathbf{X} = \left(0,\ldots,\underbrace{1}_{i\ th},\ldots,0\right)^T\right] = p_i$ and $\sum_{i=1}^m p_i = 1$. Then the subgaussian norm of $\mathbf{x}$ is upper bounded by $\frac{K(\max_{i\in[m]} p_i)^{1/2}}{(1-\max_{i\in[m]} p_i)}$ where $K(p)$ is defined in the above lemma.*

Proof. By the definition of subgaussian norm of a random vector and Lemma 10, for arbitrary $\alpha \in \mathbb{S}^{m-1}$ and $\lambda\in\mathbb{R}$ it has:

$$\begin{aligned}
&\mathbb{E}[e^{\alpha^T X - \mathbb{E}[\alpha^T X]}] \\
\leq &\sum_{i=1}^m p_i e^{\lambda(\alpha_i - \sum_{j=1}^m \alpha_j p_j)} \\
\leq &\left(\sum_{j=1}^m p_i e^{\lambda a_i}\right) e^{\lambda\sum_{j=1}^m p_j^2} \\
\leq &(\max_i p_i) e^{\lambda(1+\sum_{j=1}^m p_j^2)} \\
\leq &\max_{i\in[m]} p_i e^{\lambda(1+\sum_{j=1}^m p_j^2)} + (1-\max_{i\in[m]} p_i) e^{-\frac{\max_{i\in[m]} p_i}{1-\max_{i\in[m]} p_i}\lambda(1+\sum_{j=1}^m p_j^2)} \\
\leq &\exp\left\{\frac{K(\max_{i\in[n]} p_i)\lambda}{2(1-\max_{i\in[m]} p_i)^2}\right\}
\end{aligned} \tag{162}$$

□

References

Bowers, J. and Culpepper, S. (2022). Dependent latent class models. *arXiv preprint arXiv: Arxiv-2205.08677*.

Buldygin, V. and Moskvichova, K. (2013). The sub-gaussian norm of a binary random variable. *Theory of probability and mathematical statistics*, 86:33–49.

Cai, C., Li, G., Chi, Y., Poor, H. V., and Chen, Y. (2021). Subspace estimation from unbalanced and incomplete data matrices: statistical guarantees. *The Annals of Statistics*, 49(2):944–967.

Chen, X. and Zhang, A. Y. (2021). Optimal clustering in anisotropic gaussian mixture models. *arXiv preprint arXiv:2101.05402*.

Chen, Y., Li, X., Liu, J., and Ying, Z. (2018). Robust measurement via a fused latent and graphical item response theory model. *Psychometrika*, 83(3):538–562.

Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

Fan, J., Fan, Y., Han, X., and Lv, J. (2022). Asymptotic theory of eigenvectors for random matrices with diverging spikes. *Journal of the American Statistical Association*, 117(538):996–1009.

Gao, C. and Zhang, A. Y. (2019). Iterative algorithm for discrete structure recovery. *arXiv preprint arXiv: Arxiv-1911.01018*.

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

Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6.

Kannan, R., Vempala, S., et al. (2009). Spectral algorithms. *Foundations and Trends® in Theoretical Computer Science*, 4(3–4):157–288.

Kumar, A. and Kannan, R. (2010). Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE.

Lei, L. (2019). Unified eigenspace perturbation theory for symmetric random matrices. *arXiv preprint arXiv:1909.04798*.

Löffler, M., Zhang, A. Y., and Zhou, H. H. (2019). Optimality of spectral clustering in the gaussian mixture model. *arXiv preprint arXiv: Arxiv-1911.00538*.

Reboussin, B. A., Ip, E. H., and Wolfson, M. (2008). Locally dependent latent class models with covariates: an application to under-age drinking in the USA. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(4):877–897.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Zeng, Z., Gu, Y., and Xu, G. (2022). A tensor-em method for large-scale latent class analysis with binary responses. *Psychometrika*, pages 1–33.

Zhang, A. Y. and Zhou, H. H. (2022). Leave-one-out singular subspace perturbation analysis for spectral clustering. *arXiv preprint arXiv: Arxiv-2205.14855*.