Submitted to the Annals of Statistics

LEAVE-ONE-OUT SINGULAR SUBSPACE PERTURBATION ANALYSIS FOR SPECTRAL CLUSTERING

BY ANDERSON Y. ZHANG[1,*] AND HARRISON Y. ZHOU[2,†]

[1]*Department of Statistics and Data Science, University of Pennsylvania, ayz@wharton.upenn.edu*

[2]*Department of Statistics and Data Science, Yale University, huibin.zhou@yale.edu*

The singular subspaces perturbation theory is of fundamental importance in probability and statistics. It has various applications across different fields. We consider two arbitrary matrices where one is a leave-one-column-out submatrix of the other one and establish a novel perturbation upper bound for the distance between two corresponding singular subspaces. It is well-suited for mixture models and results in a sharper and finer statistical analysis than classical perturbation bounds such as Wedin's Theorem. Powered by this leave-one-out perturbation theory, we provide a deterministic entrywise analysis for the performance of the spectral clustering under mixture models. Our analysis leads to an explicit exponential error rate for the clustering of sub-Gaussian mixture models. For the mixture of isotropic Gaussians, the rate is optimal under a weaker signal-to-noise condition than that of Löffler et al. (2021).

1. Introduction. The matrix perturbation theory [36, 4] is a central topic in probability and statistics. It plays a fundamental role in spectral methods [10, 18], an umbrella term for algorithms involving eigendecomposition or singular value decomposition. It has a wide range of applications including principal component analysis [1, 7], covariance matrix estimation [14], clustering [38, 33, 34, 29], and matrix completion [27, 13], throughout different fields such as machine learning [5], network science [31, 2], and genomics [19].

Perturbation analysis for eigenspaces and singular subspaces dates back to seminal works of Davis and Kahan [11] and Wedin [40]. Davis-Kahan Theorem provides a clean bound for eigenspaces in terms of operator norm and Frobenius norm, and Wedin further extends it to singular subspaces. In recent years, there has been growing literature in developing fine-grained ℓ_∞ analysis for singular vectors [2, 14] and $\ell_{2,\infty}$ analysis for singular subspaces [24, 9, 6, 3], which often lead to sharp upper bounds. For clustering problems, they can be used to establish the exact recovery of spectral methods, but are usually not suitable for low signal-to-noise ratio regimes where only partial recovery is possible.

In this paper, we consider a special matrix perturbation case where one matrix differs from the other one by having one less column and investigate the difference between two corresponding left singular subspaces. Consider two matrices

$$Y = (y_1, \ldots, y_{n-1}) \in \mathbb{R}^{p\times(n-1)} \text{ and } \hat{Y} = (y_1, \ldots, y_{n-1}, y_n) \in \mathbb{R}^{p\times n}, \tag{1}$$

where Y is a leave-one-column-out submatrix of $\hat{Y}$ with the last column removed. Let U_r and $\hat{U}_r$ include the leading r left singular vectors of Y and $\hat{Y}$, respectively. The two corresponding left singular subspaces are $\text{span}(U_r)$ and $\text{span}(\hat{U}_r)$, where the former one can be interpreted as a leave-one-out counterpart of the latter one.

*Research supported in part by NSF grant DMS-2112988.

†Research supported in part by NSF grant DMS-2112918.

MSC2020 subject classifications: Primary 62H30.

Keywords and phrases: Mixture model, Spectral clustering, Singular subspace, Spectral perturbation, Leave-one-out analysis.

We establish a novel upper bound for the Frobenius norm of $\hat{U}_r\hat{U}_r^T - U_rU_r^T$ to quantify the distance between the two singular subspaces span(U_r) and span$(\hat{U}_r)$. A direct application of the generic Wedin's Theorem leads to a ratio of the magnitude of perturbation $(I - U_rU_r^T)y_n$ to the corresponding spectral gap $\sigma_r - \sigma_{r+1}$. We go beyond Wedin's Theorem and reveal that the interplay between $U_rU_r^Ty_n$ and $(I - U_rU_r^T)y_n$ plays a crucial role. Our new upper bound is a product of the aforementioned ratio and a factor determined $U_r^Ty_n$. That is, informally (see Theorem 2.1 for a precise statement),

$$\left\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\right\|_{\mathrm{F}} \lesssim \frac{\left\|(I - U_rU_r^T)y_n\right\|}{\sigma_r - \sigma_{r+1}} \times \text{ a factor from } U_r^Ty_n.$$

When this factor is smaller than some constant, it results in a sharper upper bound than Wedin's Theorem. The established upper bound is particularly suitable for mixture models where the contributions of $U_r^Ty_n$ are well-controlled, and consequently provides a key toolkit for the follow-up statistical analysis on spectral clustering.

Spectral clustering is one of the most popular approaches to group high-dimensional data. It first reduces the dimensionality of data by only using a few of its singular components, followed by a classical clustering method such as k-means to the data of reduced dimension. It is computationally appealing and often has remarkably good performance, and has been widely used in various problems. In recent years there has been growing interest in theoretical properties of spectral clustering, noticeably in community detection [2, 23, 17, 32, 33, 43, 15, 30, 22]. In spite of various polynomial-form upper bounds in terms of signal-to-noise ratios for the performance of spectral clustering, sharper exponential error rates are established in literature only for a few special scenarios, such as Stochastic Block Models with two equal-size communities [2]. Spectral clustering is also investigated in mixture models [29, 25, 1, 12, 39, 35]. For isotropic Gaussian mixture models, [25] shows spectral clustering achieves the optimal minimax rate. However, the proof technique used in [25] is very limited to the isotropic Gaussian noise and it is unclear whether it is possible to be extended to either sub-Gaussian distributed errors or unknown covariance matrices. Spectral clustering for sub-Gaussian mixture models is studied in [1] but only under special assumptions on the spectrum and geometry of the centers. It requires eigenvalues of the Gram matrix of centers to be all in the same order and sufficiently large, which rules out many interesting cases.

We study the theoretical performance of the spectral clustering under general mixture models where each observation X_i is equal to one of k centers plus some noise ϵ_i. The spectral clustering first projects X_i onto $\hat{U}_{1:r}^TX_i$ where $\hat{U}_{1:r}$ includes the leading r left singular vectors of the data matrix, and then performs k-means on this low-dimensional space. Powered by our leave-one-out perturbation theory, we provide a deterministic entrywise analysis for the spectral clustering and show that whether X_i is correctly clustered or not is determined by $\hat{U}_{-i,1:r}^T\epsilon_i$ where $\hat{U}_{-i,1:r}$ is the leave-one-out counterpart of $\hat{U}_{1:r}$ that uses all the observations except X_i. The independence between $\hat{U}_{-i,1:r}$ and ϵ_i enables us to derive explicit error risks when the noises are randomly generated from certain distributions. Specifically:

1. For sub-Gaussian mixture models, we establish an exponential error rate for the performance of the spectral clustering, assuming the centers are separated from each other and the smallest non-zero singular value is away from zero. Our conditions are more general than those needed in [1]. To remove the spectral gap condition, we further propose a variant of the spectral clustering where the number of singular vectors used is selected adaptively.
2. For Gaussian mixture models with isotropic covariance matrix, we fully recover the results of [25]. Empowered by the leave-one-out perturbation theory, our proof is completely different and is much shorter compared to that of [25]. In addition, the signal-to-noise ratio condition of [25] is improved.

3. For a two-cluster symmetric mixture model where coordinates of the noise ϵ_i are independently and identically distributed, we provide a matching upper and lower bound for the performance of the spectral clustering. This sharp analysis provides an answer to the optimality of the spectral clustering in this setting: it is in general sub-optimal and is optimal only if each coordinate of ϵ_i is normally distributed.

Organization. This paper is organized as follows. In Section 2, we first establish a general leave-one-out perturbation theory for singular subspaces, followed by its application in mixture models. In Section 3, we use our leave-one-out perturbation theory to provide theoretical guarantees for the spectral clustering under mixture models. The proofs of main results in Section 2 and Section 3 are given in Section 4 and in Section 5, respectively. The remaining proofs are included in the supplement [42].

Notation. For any positive integer r, let $[r]=\{1,2,\ldots,r\}$. For two scalars $a,b\in\mathbb{R}$, denote $a\wedge b=\min\{a,b\}$. For two matrices $A=(A_{i,j})$ and $B=(B_{i,j})$, we denote $\langle A,B\rangle=\sum_{i,j}A_{i,j}B_{i,j}$ to be the trace product, $\|A\|$ to be its operator norm, $\|A\|_{\mathrm{F}}$ to be its Frobenius norm, and $\mathrm{span}(A)$ to be the linear space spanned by columns of A. If both A,B are symmetric, we write $A\prec B$ if $B-A$ is positive semidefinite. For scalars $x_1,\ldots,x_d$, we denote $\mathrm{diag}(x_1,\ldots,x_d)$ to be a $d\times d$ diagonal matrix with diagonal entries being $x_1,\ldots,x_d$. For any integers $d,p\geq 0$, we denote $0_d\in\mathbb{R}^d$ to be a vector with all coordinates being 0, $\mathbb{1}_d\in\mathbb{R}^d$ to be a vector with all coordinates being 1, and $O_{d\times p}\in\mathbb{R}^{d\times p}$ to be a matrix with all entries being 0. We denote $I_{d\times d}$ and I_d to be the $d\times d$ identity matrix and we use I for short when the dimension of clear according to context. Let $\mathbb{O}^{d\times p}=\left\{V\in\mathbb{R}^{d\times p}:V^TV=I\right\}$ be the set of matrices in $\mathbb{R}^{d\times p}$ with orthonormal columns. We denote $\mathbb{I}\left\{\cdot\right\}$ to be the indicator function. For two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n\lesssim b_n$, $a_n=O(b_n)$, $b_n\gtrsim a_n$ all mean $a_n\leq Cb_n$ for some constant $C>0$ independent of n. We also write $a_n=o(b_n)$ when $\limsup_{n\to\infty}\frac{a_n}{b_n}=0$. For a random variable X, we say X is sub-Gaussian with variance proxy σ^2 (denoted as $X\sim\mathrm{SG}(\sigma^2)$) if $\mathbb{E}e^{tX}\leq\exp\left(\sigma^2t^2/2\right)$ for any $t\in\mathbb{R}$. For a random vector $X\in\mathbb{R}^d$, we say X is sub-Gaussian with variance proxy σ^2 (denoted as $X\sim\mathrm{SG}_d(\sigma^2)$) if $u^TX\sim\mathrm{SG}(\sigma^2)$ for any unit vector $u\in\mathbb{R}^d$.

2. Leave-one-out Singular Subspace Perturbation Analysis. In this section, we establish a general matrix perturbation theory for singular subspaces. In particular, we consider two arbitrary matrices with one having a less column than the other and study the difference between two corresponding left singular subspaces. We will first develop a general theory and then apply it to mixture models.

2.1. *General Results.* Consider two matrices as in (1) such that they are equal to each other except that $\hat{Y}$ has an extra last column. Let the Singular Value Decomposition (SVD) of these two matrices be

$$Y=\sum_{i\in[p\wedge(n-1)]}\sigma_iu_iv_i^T \text{ and } \hat{Y}=\sum_{i\in[p\wedge n]}\hat{\sigma}_i\hat{u}_i\hat{v}_i^T,$$

where $\sigma_1\geq\ldots\geq\sigma_{p\wedge(n-1)}$ and $\hat{\sigma}_1\geq\ldots\geq\hat{\sigma}_{p\wedge n}$. Consider any $r\in[p\wedge(n-1)]$. Define

$$U_r:=(u_1,\ldots,u_r)\in\mathbb{O}^{p\times r} \text{ and } \hat{U}_r:=(\hat{u}_1,\ldots,\hat{u}_r)\in\mathbb{O}^{p\times r}$$

to include the leading r left singular vectors of Y and $\hat{Y}$, respectively. Since Y can be viewed as a leave-one-out submatrix of $\hat{Y}$ that is without the last column y_n, U_r can be interpreted as a leave-one-out counterpart of $\hat{U}_r$.

The two matrices $U_r, \hat{U}_r$ correspond to two singular subspaces $\text{span}(U_r), \text{span}(\hat{U}_r)$, respectively. The difference between these two subspaces can be captured by $\sin\Theta$ distances, $\|\sin\Theta(\hat{U}_r, U_r)\|$ or $\|\sin\Theta(\hat{U}_r, U_r)\|_{\mathrm{F}}$, where

$$\Theta(\hat{U}_r, U_r) := \mathrm{diag}(\cos^{-1}(\alpha_1), \cos^{-1}(\alpha_2), \ldots, \cos^{-1}(\alpha_r))$$

with $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_r \geq 0$ being the r singular values of $\hat{U}_r^T U_r$. It is known (cf. Lemma 1 of [8]) that $\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\|_{\mathrm{F}} = \sqrt{2}\|\sin\Theta(\hat{U}_r, U_r)\|_{\mathrm{F}}$. Throughout this section, we will focus on establishing sharp upper bounds for $\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\|_{\mathrm{F}}$, i.e., the Frobenius norm of the difference between two corresponding projection matrices $U_rU_r^T$ and $\hat{U}_r\hat{U}_r^T$.

Since the augmented matrix $Y' := (Y, U_rU_r^T y_n) \in \mathbb{R}^{p\times n}$ concatenated by Y and $U_rU_r^T y_n$ has the same leading r left singular subspace and projection matrix as Y, a natural idea is to relate $\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\|_{\mathrm{F}}$ with the difference $\hat{Y} - Y'$. The classical spectral perturbation theory such as Wedin's Theorem [41, 8] leads to that if $\sigma_r - \sigma_{r+1} > 2\left\|(I - U_rU_r^T)y_n\right\|$, then

$$\left\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\right\|_{\mathrm{F}} \leq \frac{2\left\|(I - U_rU_r^T)y_n\right\|}{\sigma_r - \sigma_{r+1}}. \tag{2}$$

See Proposition D.1 in the supplement for its proof. The upper bound in (2) requires the spectral gap $\sigma_r - \sigma_{r+1}$ is away from zero. It also indicates the magnitude of the difference $\|\hat{Y} - Y'\| = \|(I - U_rU_r^T)y_n\|$ plays a crucial role. In spite of its simple form, (2) comes from generic spectral perturbation theories not specifically designed for the setting (1).

In the following Theroem 2.1, we provide a deeper and finer analysis for $\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\|_{\mathrm{F}}$, utilizing the fact that $\hat{Y}$ and Y differ by only one column and exploiting the interplay between $U_rU_r^T y_n$ and $(I - U_rU_r^T)y_n$.

THEOREM 2.1. *If*

$$\rho := \frac{\sigma_r - \sigma_{r+1}}{\|(I - U_rU_r^T)y_n\|} > 2, \tag{3}$$

we have

$$\left\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\right\|_{\mathrm{F}} \leq \frac{4\sqrt{2}}{\rho}\sqrt{\sum_{i=1}^{r}\left(\frac{u_i^T y_n}{\sigma_i}\right)^2}. \tag{4}$$

Theorem 2.1 gives an upper bound on $\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\|_{\mathrm{F}}$ that is essentially a product of ρ^{-1} and some quantity determined by $\{\sigma_i^{-1}u_i^T y_n\}_{i\in[r]}$. Since $(\sigma_i^{-1}u_i^T y_n)^2 \leq \sigma_r^{-2}(u_i^T y_n)^2$ for each $i \in [r]$, (4) leads to a simpler upper bound

$$\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\|_{\mathrm{F}} \leq \frac{4\sqrt{2}}{\rho}\frac{\|U_rU_r^T y_n\|}{\sigma_r}.$$

The condition (3) in Theorem 2.1 can be understood as a spectral gap assumption as it needs the gap $\sigma_r - \sigma_{r+1}$ to be larger than twice the magnitude of the perturbation $\|(I - U_rU_r^T)y_n\|$. This condition can be slightly weakened into $\sigma_r^2 - \sigma_{r+1}^2 - \|(I - U_rU_r^T)y_n\|^2 > 0$, though resulting in a more involved upper bound. See Theorem 4.1 in Section 4.1 for details.

We are ready to have a comparison of our result (4) and (2) that is from Wedin's Theorem. Under the assumption (3), the upper bound in (2) can be written equivalently as $2\rho^{-1}$. As a result, the comparison is about the magnitude of $(\sum_{i\in[r]}(\sigma_i^{-1}u_i^T y_n)^2)^{1/2}$. If it is smaller than $1/(2\sqrt{2})$, then (4) gives a sharper upper bound than (2). To further compare these two bounds, consider the following examples.

- *Example 1.* When $U_r^T y_n = 0$ and (3) is satisfied, (4) gives the correct upper bound 0. That is, $\hat{U}_r\hat{U}_r^T = U_r U_r^T$. On the contrary, (2) gives a non-zero bound $2/\rho^{-1}$. To be more concrete, let $Y = \sigma_1(p^{-1/2}\mathbb{1}_p)((n-1)^{-1/2}\mathbb{1}_{n-1})^T$ be a rank-one matrix and y_n be some vector that is orthogonal to $\mathbb{1}_p$. Then if $\sigma_1 > 2\|y_n\|$, we have $\hat{u}_1 = u_1 = p^{-1/2}\mathbb{1}_p$ up to sign. (4) gives the correct answer $\|\hat{u}_1\hat{u}_1^T - u_1u_1^T\|_{\mathrm{F}} = 0$ as $u_1^T y_n = 0$, while (2) leads to a loose upper bound $2\|y_n\|/\sigma_1$.
- *Example 2.* Let Y be a matrix with two unique columns such that y_j is equal to either θ or $-\theta$ for all $j \in [n-1]$ and for some vector $\theta \in \mathbb{R}^p$. Then Y is a rank-one matrix with $\sigma_1 = \|\theta\|\sqrt{n-1}$. Let $y_n = \theta + \epsilon$. As long as $\|\theta\|\sqrt{n-1} > 2\|\epsilon\|$, we have $\|\hat{u}_1\hat{u}_1^T - u_1u_1^T\|_{\mathrm{F}} \le 4\sqrt{2}\rho^{-1}(\|\theta\| + \|\epsilon\|)/\sigma_1$ from (4). If we further assume $\|\theta\| = 1$ and $\epsilon \sim \mathcal{N}(0, I_p)$ with $p \ll n$, we have $\|\hat{u}_1\hat{u}_1^T - u_1u_1^T\|_{\mathrm{F}} \lesssim \sqrt{p/n}\rho^{-1} = o(\rho^{-1})$ with high probability. In contrast, (2) only gives $2\rho^{-1}$.

In the next section, we consider mixture models where the magnitude of $(\sum_{i\in[r]}(\sigma_i^{-1}u_i^T y_n)^2)^{1/2}$ is well-controlled and (4) leads to a much sharper upper bound compared to (2).

2.2. *Singular Subspace Perturbation in Mixture Models.* The general perturbation theory presented in Theorem 2.1 is particularly suitable for analyzing singular subspaces of mixture models.

Mixture Models. We consider a mixture model with k centers $\theta_1^*, \theta_2^*, \ldots, \theta_k^* \in \mathbb{R}^p$ and a cluster assignment vector $z^* \in [k]^n$. The observations $X_1, X_2, \ldots, X_n \in \mathbb{R}^p$ are generated from

$$X_i = \theta^*_{z_i^*} + \epsilon_i, \tag{5}$$

where $\epsilon_1, \ldots, \epsilon_n \in \mathbb{R}^p$ are noises. The data matrix $X := (X_1, \ldots, X_n) \in \mathbb{R}^{p\times n}$ can be written equivalently in a matrix form

$$X = P + E, \tag{6}$$

where $P := (\theta^*_{z_1^*}, \theta^*_{z_2^*}, \ldots, \theta^*_{z_n^*})$ is the signal matrix and $E := (\epsilon_1, \ldots, \epsilon_n)$ is the noise matrix. Define $\beta := \frac{1}{n/k}\min_{a\in[k]}|\{i : z_i^* = a\}|$ such that $\beta n/k$ is the smallest cluster size.

We are interested in the left singular subspaces of X and its leave-one-out counterparts. For each $i \in [n]$, define X_{-i} to be a submatrix of X with its ith column removed. That is,

$$X_{-i} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \in \mathbb{R}^{p\times(n-1)}. \tag{7}$$

Let their SVDs be $X = \sum_{j\in[p\wedge n]}\hat{\lambda}_j\hat{u}_j\hat{v}_j^T$ and $X_{-i} = \sum_{j\in[p\wedge(n-1)]}\hat{\lambda}_{-i,j}\hat{u}_{-i,j}\hat{v}_{-i,j}^T$, where $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \ldots \ge \hat{\lambda}_{p\wedge n}$ and $\hat{\lambda}_{-i,1} \ge \hat{\lambda}_{-i,2} \ge \ldots \ge \hat{\lambda}_{-i,p\wedge(n-1)}$. Note that the signal matrix P is at most rank-k. Then for any $r \in [k]$, define

$$\hat{U}_{1:r} := (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_r) \in \mathbb{O}^{p\times r} \text{ and } \hat{U}_{-i,1:r} = (\hat{u}_{-i,1}, \ldots, \hat{u}_{-i,r}) \in \mathbb{O}^{p\times r}$$

to include the leading r left singular vectors of X and X_{-i}, respectively. We are interested in controlling the quantity $\|\hat{U}_{1:r}\hat{U}_{1:r}^T - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\|_{\mathrm{F}}$ for each $i \in [n]$.

In Theorem 2.2, we provide upper bounds for $\|\hat{U}_{1:\kappa}\hat{U}_{1:\kappa}^T - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\|_{\mathrm{F}}$ for all $i \in [n]$ where $\kappa \in [k]$ is the rank of the signal matrix P. In order to have such a uniform control across all $i \in [n]$, we consider the spectrum of the signal matrix P. Let $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_{p\wedge n}$ be the singular values of P and κ be the rank of P such that $\kappa \in [k]$, $\lambda_\kappa > 0$, and $\lambda_{\kappa+1} = 0$.

THEOREM 2.2. *Assume $\beta n/k^2 \geq 10$. Assume*

$$\rho_0 := \frac{\lambda_\kappa}{\|E\|} > 16. \tag{8}$$

For any $i \in [n]$, we have

$$\left\| \hat{U}_{1:\kappa}\hat{U}_{1:\kappa}^T - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T \right\|_{\mathrm{F}} \leq \frac{128}{\rho_0}\left(\sqrt{\frac{k\kappa}{\beta n}} + \frac{\left\| \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T \epsilon_i \right\|}{\lambda_\kappa} \right). \tag{9}$$

Theorem 2.2 exploits the mixture model structure (5) that the signal matrix P has only k unique columns with each appearing at least $\beta n/k$ times. The assumption $\beta n/k \geq 16$ helps ensure that spectrum and singular vectors of P are not much changed if any column of P is removed. We require the condition (8) so that $\hat{\lambda}_{-i,\kappa} - \hat{\lambda}_{-i,\kappa+1} > 2\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T X_i\|$ holds for each $i \in [n]$, and hence Theorem 2.1 can be applied uniformly for all $i \in [n]$. The upper bound (9) is a product of ρ_0^{-1} and a summation of two terms. The second term $\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T \epsilon_i\|/\lambda_\kappa$ can be trivially upper bounded by $\|E\|/\lambda_\kappa \leq \rho_0^{-1}$. The first term $\sqrt{k\kappa/(\beta n)} = o(1)$ if $\beta n/k^2 \gg 1$, for example, when β is a constant and $k \ll \sqrt{n}$. Then (9) leads to $\|\hat{U}_{1:\kappa}\hat{U}_{1:\kappa}^T - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\|_{\mathrm{F}} \lesssim o(1)\rho_0^{-1} + \rho_0^{-2}$, superior to the upper bound (2) obtained from the direct application of Wedin's Theorem that is in an order of ρ_0^{-1}.

Theorem 2.2 studies the perturbation for the leading κ singular subspaces where κ is the rank of P. In the following Theorem 2.3, we consider an extension to $\|\hat{U}_{1:r}\hat{U}_{1:r}^T - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\|_{\mathrm{F}}$ where r is not necessarily κ.

THEOREM 2.3. *Assume $\beta n/k^2 \geq 10$. Assume there exists some $r \in [k]$ such that*

$$\tilde{\rho}_0 := \frac{\lambda_r - \lambda_{r+1}}{\max\left\{ \|E\|, \sqrt{\frac{k^2}{\beta n}}\lambda_{r+1} \right\}} > 16. \tag{10}$$

For any $i \in [n]$, we have

$$\left\| \hat{U}_{1:r}\hat{U}_{1:r}^T - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T \right\|_{\mathrm{F}} \leq \frac{128}{\tilde{\rho}_0}\left(\frac{\sqrt{kr}}{\sqrt{\beta n}} + \frac{\left\| \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T \epsilon_i \right\|}{\lambda_r} \right). \tag{11}$$

In Theorem 2.3, $r \in [k]$ is any number such that (10) is satisfied. When r is chosen to be κ, (10) is reduced to (8), and (11) leads to the same upper bound as (9). When $r < \kappa$, λ_{r+1} is non-zero and in (10) it needs to be smaller than the spectral gap $\lambda_r - \lambda_{r+1}$ after some scaling factor.

To provide some intuition on the condition (10) when $r < \kappa$, let the SVD of the signal matrix P be $P = \sum_{j\in[p\wedge n]} \lambda_j u_j v_j^T$ and define $U_{1:r} := (u_1, u_2, \ldots, u_r) \in \mathbb{O}^{p\times r}$ and $U_{(r+1):\kappa} := (u_{r+1}, u_{r+2}, \ldots, u_\kappa) \in \mathbb{O}^{p\times(\kappa-r)}$. Then the data matrix (6) can be written equivalently as

$$X = P' + E', \text{ where } P' := U_{1:r}U_{1:r}^T P \text{ and } E' := E + U_{(r+1):\kappa}U_{(r+1):\kappa}^T P. \tag{12}$$

Since it is still a mixture model, Theorem 2.2 can be applied. Nevertheless, the condition (8) essentially requires $\lambda_r/(\|E\| + \lambda_{r+1}) > 16$ as $\|E'\| \leq \|E\| + \|U_{(r+1):\kappa}U_{(r+1):\kappa}^T P\| = \|E\| + \lambda_{r+1}$, which is stronger than the condition (10). In order to weaken the requirement on the spectral gap into (10), we study the contribution of $U_{(r+1):\kappa}U_{(r+1):\kappa}^T P$ towards to the leading r singular subspaces perturbation of E. It turns out that its contribution is roughly $\sqrt{k^2/(\beta n)}\lambda_{r+1}$ instead of λ_{r+1}, due to the fact that $U_{(r+1):\kappa}U_{(r+1):\kappa}^T P$ has at most k unique columns with each one appearing at least $\beta n/k$ times.

3. Spectral Clustering for Mixture Models.

3.1. *Spectral Clustering and Polynomial Error Rate.* Recall the definition of the mixture model in (5) and also in (6). The goal of clustering is to estimate the cluster assignment vector z^* from the observations $X_1, X_2, \ldots, X_n$. Since the signal matrix P is of low rank, a natural idea is to project the observations $\{X_i\}_{i\in[n]}$ onto a low dimensional space before applying classical clustering methods such as variants of k-means. This leads to the spectral clustering presented in Algorithm 1.

Algorithm 1: Spectral Clustering

Input: Data matrix $X = (X_1, \ldots, X_n) \in \mathbb{R}^{p\times n}$, number of clusters k, number of singular vectors r

Output: Cluster assignment vector $\hat{z} \in [k]^n$

1 Perform SVD on X to have

$$X = \sum_{i=1}^{p\wedge n} \hat{\lambda}_i \hat{u}_i \hat{v}_i^T,$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_{p\wedge n} \geq 0$ and $\{\hat{u}_i\}_{i=1}^{p\wedge n} \in \mathbb{R}^p$, $\{\hat{v}_i\}_{i=1}^{p\wedge n} \in \mathbb{R}^n$. Let $\hat{U}_{1:r} := (\hat{u}_1, \ldots, \hat{u}_r) \in \mathbb{R}^{p\times r}$.

2 Perform k-means on the columns of $\hat{U}_{1:r}^T X$. That is,

$$\left(\hat{z}, \{\hat{c}_j\}_{j\in[k]}\right) = \underset{z\in[k]^n, \{c_j\}_{j\in[k]}\in\mathbb{R}^r}{\operatorname{argmin}} \sum_{i\in[n]} \left\| \hat{U}_{1:r}^T X_i - c_{z_i} \right\|^2. \tag{13}$$

In (13), the dimensionality of each data point $\hat{U}_{1:r}^T X_i$ is r, reduced from original dimensionality p. This is computationally appealing as r can be much smaller than p. The second step of Algorithm 1 is the k-means on the columns of $\hat{U}_{1:r}^T X$, which is equivalent to performing k-means onto the columns of $\hat{U}_{1:r}\hat{U}_{1:r}^T X \in \mathbb{R}^{p\times n}$. That is, define $\hat{\theta}_a = \hat{U}_{1:r}\hat{c}_a$ for each $a \in [k]$. It can be shown that (cf., Lemma 4.1 of [25])

$$\left(\hat{z}, \left\{\hat{\theta}_j\right\}_{j\in[k]}\right) = \underset{z\in[k]^n, \{\theta_j\}_{j\in[k]}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in[n]} \left\| \hat{U}_{1:r}\hat{U}_{1:r}^T X_i - \theta_{z_i} \right\|^2, \tag{14}$$

due to the fact that $\hat{U}_{1:r}$ has orthonormal columns. As a result, in the rest of the paper, we carry out our analysis on $\hat{z}$ using (14).

Before characterizing the theoretical performance of the spectral clustering $\hat{z}$, we give the definition of the misclustering error which quantifies the distance between an estimator and the ground truth z^*. For any $z \in [k]^n$, its misclustering error is defined as

$$\ell(z, z^*) = \min_{\phi\in\Phi} \frac{1}{n} \sum_{i\in[n]} \mathbb{I}\{z_i = \phi(z_i^*)\},$$

where $\Phi = \{\phi : \phi \text{ is a bijection from } [k] \text{ to } [k]\}$. The minimization of Φ is due to that the cluster assignment vector z^* is identifiable up to a permutation of the labels $[k]$. In addition to β that controls the smallest cluster size, another important quantity in this clustering task is the separation of the centers. Define Δ to be the minimum distance among centers, i.e.,

$$\Delta := \min_{a,b\in[k]:a\neq b} \|\theta_a^* - \theta_b^*\|.$$

As we will see later, Δ determines the difficulty of the clustering task and plays a crucial role.

In Proposition 3.1, we give a rough upper bound on the misclustering error $\ell(\hat{z}, z^*)$ that takes a polynomial expression (16) . It is worth mentioning that Proposition 3.1 is deterministic with no assumption on the distribution or the independence of the noises $\{\epsilon_i\}_{i\in[n]}$. In fact, the noise matrix E can be an arbitrary matrix as long as the data matrix has the decomposition (6) and the separation condition (15) is satisfied. In addition, it requires no spectral gap condition. Proposition 3.1 is essentially an extension of Lemma 4.2 in [25] which is only for the Gaussian mixture model and needs $r = k$. We include its proof in Appendix D for completeness. Recall κ is the rank of the signal matrix P.

PROPOSITION 3.1. *Consider the spectral clustering $\hat{z}$ of Algorithm 1 with $\kappa \le r \le k$. Assume*

$$\psi_0 := \frac{\Delta}{\beta^{-0.5} k n^{-0.5} \|E\|} \ge 16. \tag{15}$$

Then $\ell(\hat{z}, z^*) \le \beta/(2k)$. Furthermore, there exists one $\phi \in \Phi$ such that $\hat{z}$ satisfies

$$\ell(\hat{z}, z^*) = \frac{1}{n} |\{i \in [n] : \hat{z}_i \neq \phi(z_i^*)\}| \le \frac{C_0 k \|E\|^2}{n\Delta^2}, \tag{16}$$

and

$$\max_{a\in[k]} \left\| \hat{\theta}_{\phi(a)} - \theta_a^* \right\| \le C_0 \beta^{-0.5} k n^{-0.5} \|E\|, \tag{17}$$

where $C_0 = 128$.

Proposition 3.1 provides a starting point for our further theoretical analysis. In the following sections, we are going to provide a sharper analysis for the spectral clustering $\hat{z}$ that is beyond the polynomial rate stated in (16), with the help of singular subspaces perturbation established in Section 2.

3.2. *Entrywise Error Decompositions.* In this section, we are going to develop a fine-grained and entrywise analysis on the performance of $\hat{z}$. Proposition 3.1 points out that there exists a permutation $\phi \in \Phi$ such that $n\ell(\hat{z}, z^*) = |\{i \in [n] : \hat{z}_i \neq \phi(z_i^*)\}| \le n\beta/(2k)$. Since the smallest cluster size in z^* is at least $\beta n/k$, such permutation ϕ is unique. With ϕ identified, $\hat{z}_i \neq \phi(z_i^*)$ means that the ith data point X_i is incorrectly clustered in $\hat{z}$, for each $i \in [n]$. The following Lemma 3.2 studies the event $\hat{z}_i \neq \phi(z_i^*)$ and shows that it is determined by the magnitude of $\|\hat{U}_{1:r}\hat{U}_{1:r}^T \epsilon_i\|$.

LEMMA 3.1. *Consider the spectral clustering $\hat{z}$ of Algorithm 1 with $\kappa \le r \le k$. Assume (15) holds. Let $\phi \in \Phi$ be the permutation such that $\ell(\hat{z}, z^*) = \frac{1}{n}|\{i \in [n] : \hat{z}_i \neq \phi(z_i^*)\}|$. Then there exists a constant $C > 0$ such that for any $i \in [n]$,*

$$\mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\} \le \mathbb{I}\left\{ \left(1 - C\psi_0^{-1}\right)\Delta \le 2 \left\| \hat{U}_{1:r}\hat{U}_{1:r}^T \epsilon_i \right\| \right\}. \tag{18}$$

To understand Lemma 3.1, recall that in (14) $\hat{z}$ is obtained by k-means on $\{\hat{U}_{1:r}\hat{U}_{1:r}^T X_i\}_{i\in[n]}$. Since we have the decomposition $\hat{U}_{1:r}\hat{U}_{1:r}^T X_i = \hat{U}_{1:r}\hat{U}_{1:r}^T \theta^*_{z_i^*} + \hat{U}_{1:r}\hat{U}_{1:r}^T \epsilon_i$ for each $i \in [n]$, the data points $\{\hat{U}_{1:r}\hat{U}_{1:r}^T X_i\}_{i\in[n]}$ follow a mixture model with centers $\{\hat{U}_{1:r}\hat{U}_{1:r}^T \theta_a^*\}_{a\in[k]}$ and noises $\{\hat{U}_{1:r}\hat{U}_{1:r}^T \epsilon_i\}_{i\in[n]}$. In the proof of Lemma 3.1 we can show these k centers preserve the geometric structure of $\{\theta_a^*\}_{a\in[k]}$ with minimum distance around Δ. Intuitively, if $\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\|$ is smaller than half of the minimum distance, $\hat{U}_{1:r}\hat{U}_{1:r}^T X_i$ is closer to $\hat{U}_{1:r}\hat{U}_{1:r}^T\theta^*_{z_i^*}$ than any other centers, and thus z_i^* can be correctly recovered.

Lemma 3.1 itself is not sufficient to obtain explicit expressions for the performance of spectral clustering when the noises $\{\epsilon_i\}_{i\in[n]}$ are assumed to be random. The entrywise upper bound (18) shows that the event $\hat{z}_i \neq \phi(z_i^*)$ is determined by the $\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\|$, but the fact that $\hat{U}_{1:r}\hat{U}_{1:r}^T$ depends on ϵ_i makes any follow-up probability calculations challenging. The key to make use of Lemma 3.1 is our leave-one-out singular subspace perturbation theory, particularly, Theorem 2.2. To decouple the dependence between $\hat{U}_{1:r}\hat{U}_{1:r}^T$ and ϵ_i, we replace the former quantity by its leave-one-out counterpart $\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T$. Take r to be κ. Note that

$$\left\|\hat{U}_{1:\kappa}\hat{U}_{1:\kappa}^T\epsilon_i\right\| \leq \left\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\epsilon_i\right\| + \|U_{1:\kappa}\hat{U}_{1:\kappa}^T - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\|_{\mathrm{F}}\,\|\epsilon_i\|\,. \tag{19}$$

The perturbation $\|U_{1:\kappa}\hat{U}_{1:\kappa}^T - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\|_{\mathrm{F}}$ is well-controlled by Theorem 2.2, which shows the second term on the RHS of the above display is essentially $O(\rho_0^{-2})\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\epsilon_i\|$. This leads to the following Lemma 3.2 on the entrywise clustering errors.

LEMMA 3.2. *Consider the spectral clustering $\hat{z}$ of Algorithm 1 with $r=\kappa$. Assume $\beta n/k^2 \geq 10$, (8), and (15) hold. Let $\phi\in\Phi$ be the permutation such that $\ell(\hat{z},z^*) = \frac{1}{n}|\{i\in[n]:\hat{z}_i\neq\phi(z_i^*)\}|$. Then there exists a constant C such that for any $i\in[n]$,*

$$\mathbb{I}\{\hat{z}_i\neq\phi(z_i^*)\} \leq \mathbb{I}\left\{\left(1-C\left(\psi_0^{-1}+\rho_0^{-2}\right)\right)\Delta \leq 2\left\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\epsilon_i\right\|\right\}.$$

Consequently, if the noises $\{\epsilon_i\}_{i\in[n]}$ are random, we have the risk of $\hat{z}$ satisfy

$$\mathbb{E}\ell(\hat{z},z^*) \leq n^{-1}\sum_{i\in[n]}\mathbb{E}\mathbb{I}\left\{\left(1-C\left(\psi_0^{-1}+\rho_0^{-2}\right)\right)\Delta \leq 2\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|\right\}.$$

Lemma 3.2 needs three conditions. The first one $\beta n/k^2 \geq 10$ is on the smallest cluster sizes and can be easily satisfied if both β, k are constants. The second condition (8) is a spectral gap condition on the smallest non-zero singular value λ_κ. The third one is for the separation of the centers Δ. With all the three conditions satisfied, Lemma 3.2 shows that the entrywise clustering error for X_i boils down to $\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\epsilon_i\|$. When the noises $\{\epsilon_j\}_{j\in[n]}$ are assumed to be random and independent of each other, the projection matrix $\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T$ is independent of ϵ_i for each $i\in[n]$, a desired property crucial to our follow-up investigation on the risk $\mathbb{E}\ell(\hat{z},z^*)$. When $\{X_i\}_{i\in[n]}$ are generated randomly as in the following sections, Lemma 3.2 leads to explicit expressions for the performance of the spectral clustering.

The key towards establishing Lemma 3.2 is Theorem 2.2. Without Theorem 2.2, if the classical perturbation theory such as Wedin's theorem is used instead, then in order to obtain similar upper bounds in Lemma 3.2, the second term on the RHS of (19) needs to be much smaller than Δ. This essentially requires $\max_{i\in[p]}\|\epsilon_i\|^2 \lesssim \lambda_\kappa\Delta$, in addition to (8) and (15). As we will show in the next section, for sub-Gaussian noises, this additional condition requires $p\log n \lesssim \sqrt{n}$ in regimes where Lemma 3.2 only needs $p\lesssim n$.

, we provide an upper bound for $\left\|\hat{U}\hat{U}^T - \hat{U}_{-i}\hat{U}_{-i}^T\right\|$ showing that it is essentially determined by $\left\|U_{-i}U_{-i}^T\epsilon_i\right\|$ under an eigen-gap condition.

3.3. *Sub-Gaussian Mixture Models.* In this section, we investigate the performance of the spectral clustering $\hat{z}$ for mixture models with sub-Gaussian noises. Theorem 3.1 assumes that each noise ϵ_i is an independent sub-Gaussian random vector with zero mean and variance proxy σ^2 and establishes an exponential rate for the risk $\mathbb{E}\ell(\hat{z},z^*)$.

THEOREM 3.1. *Consider the spectral clustering $\hat{z}$ of Algorithm 1 with $r = \kappa$. Assume $\epsilon_i \sim SG_p(\sigma^2)$ independently with zero mean for each $i \in [n]$. Assume $\beta n/k^2 \geq 10$. There exist constants $C, C' > 0$ such that under the assumption that*

$$\psi_1 := \frac{\Delta}{\beta^{-0.5} k \left(1 + \sqrt{\frac{p}{n}}\right) \sigma} > C \tag{20}$$

and

$$\rho_1 := \frac{\lambda_\kappa}{\left(\sqrt{n} + \sqrt{p}\right) \sigma} > C, \tag{21}$$

we have

$$\mathbb{E}\ell(\hat{z}, z^*) \leq \exp\left(-\left(1 - C'\left(\psi_1^{-1} + \rho_1^{-2}\right)\right) \frac{\Delta^2}{8\sigma^2}\right) + \exp\left(-\frac{n}{2}\right).$$

Under this sub-Gaussian setting, standard concentration theory shows that the noise matrix E has its operator norm $\|E\| \lesssim \sigma(\sqrt{n} + \sqrt{p})$ with high probability (cf. Lemma D.1). Under this event, (20) and (21) are sufficient conditions for (8) and (15), respectively. The risk in Theorem 3.3 has two terms, where the first term takes an exponential form of $\Delta^2/(8\sigma^2)$ and the second term $\exp(-n/2)$ comes from the aforementioned event of $\|E\|$. The first term is the dominating one, as long as Δ^2/σ^2, which can be interpreted as the signal-to-noise ratio, is smaller than $n/2$. In fact, $\Delta^2/\sigma^2 \lesssim \log n$ is the most interesting regime as otherwise $\hat{z}$ already achieves the exact recovery (i.e., $\hat{z} = z^*$) with high probability, since $\mathbb{E}\{\ell(\hat{z}, z^*) = 0\} = o(1)$.

Theorem 3.1 makes a substantial improvement over Proposition 3.1. Using the aforementioned with-high-probability event on $\|E\|$, (16) only leads to $\mathbb{E}\ell(\hat{z}, z^*) \lesssim (1 + \sqrt{p/n})^2\sigma^2/\Delta^2 + \exp(-n/2)$ which takes a polynomial form of the Δ^2/σ^2. On the contrary, Theorem 3.1 provides a much sharper exponential rate.

Our leave-one-out singular subspace perturbation theory and its consequence Lemma 3.2 provide the key toolkit towards Theorem 3.1. Since $\hat{U}_{-i,1:\kappa}^T$ is independent of ϵ_i, we have $\hat{U}_{-i,1:\kappa}^T \epsilon_i \sim \mathrm{SG}_\kappa(\sigma^2)$ being another sub-Gaussian random vector. This makes it possible to control the tail probabilities of $\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\epsilon_i\|^2 = \|\hat{U}_{-i,1:\kappa}^T\epsilon_i\|^2$ which is a quadratic form of sub-Gaussian random vectors. Without using our perturbation theory, if the classical perturbation bounds such as Wedin's Theorem is used instead, the previous section shows that $\max_{i\in[p]} \|\epsilon_i\|^2 \lesssim \lambda_\kappa\Delta$ is additionally needed to obtain results similar to Lemma 3.2. This equivalently requires $\lambda_\kappa\Delta/(\sigma^2 p \log n) \gtrsim 1$. When $\Delta/\sigma, k, \beta$ are constants, this additional condition essentially requires $p \log n \lesssim \sqrt{n}$. In contrast, Theorem 3.1 only needs $p \lesssim n$.

Theorem 3.1 gives a finite-sample result for the performance of spectral clustering in sub-Gaussian mixture models. In the following Corollary 3.1, by slightly strengthening conditions (20) and (21), it immediately yields an asymptotic error bound with the exponent being $(1 - o(1))\Delta^2/(8\sigma^2)$.

COROLLARY 3.1. *Under the same setting as in Theorem 3.1, if $\psi_1, \rho_1 \to \infty$ is further assumed, we have*

$$\mathbb{E}\ell(\hat{z}, z^*) \leq \exp\left(-(1 - o(1))\frac{\Delta^2}{8\sigma^2}\right) + \exp\left(-\frac{n}{2}\right).$$

If $\Delta/\sigma \geq (1 + c)2\sqrt{2\log n}$ is further assumed where $c > 0$ is any constant, $\hat{z}$ achieves the exact recovery, i.e., $\mathbb{E}\mathbb{I}\{\ell(\hat{z}, z^*) \neq 0\} = o(1)$.

In the exponents of Theorem 3.1 and Corollary 3.1, we are able to obtain an explicit constant $1/8$. In addition, we obtain an explicit constant $2\sqrt{2}$ for the exact recovery in Corollary 3.1. These constants are sharp when the noises are further assumed to be isotropic Gaussian, as we will show in Section 3.5.

The recent related paper by [1] develops a ℓ_p perturbation theory and applies it to the spectral clustering for sub-Gaussian mixture models. It obtains exponential error rates but with unspecified constants in the exponents and under special assumptions on the spectrum and geometric distribution of the centers. It first assumes both β and k are constants. Let $G \in \mathbb{R}^{k\times k}$ be the Gram matrix of the centers such that $G_{i,j} = \theta_i^{*T}\theta_j^*$ for each $i, j \in [k]$. It requires $\bar{\lambda} I \prec G \prec c\bar{\lambda} I$ for some constant $c > 1$, i.e., all k eigenvalues of G are in the same order. It implies that the maximum and minimum distances among centers are comparable. This rules out many interesting cases such as all the centers are on one single line. In addition, [1] needs $\bar{\lambda}/\sigma \to \infty$. Equivalently it means that the leading k singular values $\lambda_1, \lambda_2, \dots, \lambda_k$ of the signal matrix P not only are all in the same order, but also $\lambda_k/(\sqrt{n}\sigma) \gg \max\{1, \sqrt{p/n}\}$. As a comparison, we allow collinearity of the centers such that the rank of G (and P) can be smaller than k. We allow the singular values $\lambda_1, \lambda_2, \dots, \lambda_\kappa$ not in the same order as long as the smallest one satisfies (21), which can be equivalently written as $\lambda_\kappa/(\sqrt{n}\sigma) \gtrsim \max\{1, \sqrt{p/n}\}$. The distances among the centers are also not necessarily in the same order as long as the smallest distance satisfies (20). Hence, our conditions are more general than those in [1].

The spectral gap condition (21) ensures that singular vectors corresponding to small nonzero singular values are well-behaved. It is not needed in Section 3.4 where we propose a variant of spectral clustering with adaptive dimension reduction. It can also be dropped in Section 3.5 when the noise is isotropic Gaussian. When the mixture model is symmetric with two components (for example, the model considered in Section 3.6), the signal matrix P is rank-one. Hence, (21) is also no longer needed as it can be directly implied from (20).

3.4. *Spectral Clustering with Adaptive Dimension Reduction.* The theoretical analysis for the spectral clustering $\hat{z}$ of Algorithm 1 that is carried out in Lemma 3.2 and Theorem 3.1 requires the use of all the κ singular vectors where κ is the rank of the signal matrix P. Nevertheless, not all singular components are equally useful towards the clustering task and the importance of an individual singular vector can be characterized by its corresponding singular value. This motivates us to propose the following algorithm where the number of singular vectors used is carefully picked.

Algorithm 2: Spectral Clustering with Adaptive Dimension Reduction

---

Input: Data matrix $X = (X_1, \dots, X_n) \in \mathbb{R}^{p\times n}$, number of clusters k, threshold T

Output: Clustering label vector $\tilde{z} \in [k]^n$

1 Perform SVD on X same as Step 1 of Algorithm 1.

2 Let $\hat{r}$ be the largest index in $[k]$ such that the difference between the corresponding two neighboring singular values is greater than T, i.e.,

$$\hat{r} = \max\{a \in [k] : \hat{\lambda}_a - \hat{\lambda}_{a+1} \geq T\}. \tag{22}$$

Let $\hat{U}_{1:\hat{r}} := (\hat{u}_1, \dots, \hat{u}_{\hat{r}}) \in \mathbb{R}^{p\times \hat{r}}$.

3 Perform k-means on the columns of $\hat{U}_{1:\hat{r}}^T X$. That is,

$$\left(\tilde{z}, \{\tilde{c}_j\}_{j=1}^k\right) = \mathop{\mathrm{argmin}}_{z\in[k]^n, \{c_j\}_{j=1}^k \in \mathbb{R}^{\hat{r}}} \sum_{i\in[n]} \left\| \hat{U}_{1:\hat{r}}^T X_i - c_{z_i} \right\|^2. \tag{23}$$

Algorithm 2 is a variant of Algorithm 1 with the number of singular vectors selected by (22), where $\hat{r}$ is the largest integer such that the empirical spectral gap $\hat{\lambda}_{\hat{r}} - \hat{\lambda}_{\hat{r}+1}$ is greater or equal to some threshold T. The choice of the threshold T matters. When T is small, $\hat{r}$ might be even bigger than the rank κ. When $T \gtrsim \|E\|$, it guarantees that the singular values of the signal matrix P satisfy $\lambda_{\hat{r}} - \lambda_{\hat{r}+1} \gtrsim T$ and $\lambda_{\hat{r}+1} \lesssim T$. When T is too large, the singular subspace $\hat{U}_{1:\hat{r}}$ misses singular vectors such as $\hat{u}_{\hat{r}+1}$ whose importance scales with $\lambda_{\hat{r}+1}$ that can not be ignored. This in turn deteriorates the clustering performance of $\tilde{z}$.

A rule of thumb for the threshold T is that $T/\|E\|$ is at least in a constant order. It is allowed to grow but not faster than $\tilde{\phi}_0$ defined in (24). The precise description of the choices of T needed is given below in Lemma 3.3, which provides an entrywise analysis of $\tilde{z}$ that is analogous to Lemma 3.2.

LEMMA 3.3. *Consider the estimator $\tilde{z}$ from Algorithm 2. Assume $\beta n/k^4 \geq 400$. Let $\phi \in \Phi$ be the permutation such that $\ell(\hat{z}, z^*) = \frac{1}{n}|\{i \in [n] : \hat{z}_i \neq \phi(z_i^*)\}|$. Define*

$$\tilde{\psi}_0 := \frac{\Delta}{\beta^{-0.5} k^2 n^{-0.5} \|E\|} \tag{24}$$

and $\tilde{\rho} := T/\|E\|$. Assume $256 < \tilde{\rho} < \tilde{\psi}_0/64$. There exist constants C, C' such that if $\tilde{\phi}_0 > C$, then

$$\mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\} \leq \mathbb{I}\left\{\left(1 - C'\left(\tilde{\rho}\tilde{\psi}_0^{-1} + \tilde{\rho}^{-1}\right)\right)\Delta \leq 2\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|\right\}.$$

Consequently, we have

$$\mathbb{E}\ell(\hat{z}, z^*) \leq n^{-1}\sum_{i\in[n]} \mathbb{E}\mathbb{I}\left\{\left(1 - C'\left(\tilde{\rho}\tilde{\psi}_0^{-1} + \tilde{\rho}^{-1}\right)\right)\Delta \leq 2\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|\right\}.$$

With a proper choice of the threshold T, Lemma 3.3 only poses requirements on the smallest cluster size $\beta n/k$ and minimum separation among the centers Δ. Compared to Lemma 3.2 and Theorem 3.1, it removes any condition on the smallest non-zero singular value such as (8) or (21). In addition, it requires no knowledge on the rank κ.

With Lemma 3.3, we have the following exponential error bound on the performance of $\tilde{z}$ on sub-Gaussian mixture models, analogous to Theorem 3.1 and Corollary 3.1 for $\hat{z}$.

THEOREM 3.2. *Consider the estimator $\tilde{z}$ from Algorithm 2. Assume $\epsilon_i \sim SG_p(\sigma^2)$ independently with zero mean for each $i \in [n]$. Assume $\beta n/k^4 \geq 400$. There exist constants $C, C', C_1, C_2 > 0$ such that under the assumption that*

$$\psi_2 := \frac{\Delta}{\beta^{-0.5} k^2 \left(1 + \sqrt{\frac{p}{n}}\right)\sigma} > C$$

and $\rho_2 := T/(\sigma(\sqrt{n} + \sqrt{p}))$ satisfies $C_1 \leq \rho_2 \leq \psi_2/C_2$, we have

$$\mathbb{E}\ell(\tilde{z}, z^*) \leq \exp\left(-\left(1 - C'\left(\rho_2\psi_2^{-1} + \rho_2^{-1}\right)\right)\frac{\Delta^2}{8\sigma^2}\right) + \exp\left(-\frac{n}{2}\right).$$

If $\psi_2, \rho_2 \to \infty$ and $\rho_2/\psi_2 = o(1)$ are further assumed, we have

$$\mathbb{E}\ell(\tilde{z}, z^*) \leq \exp\left(-(1 - o(1))\frac{\Delta^2}{8\sigma^2}\right) + \exp\left(-\frac{n}{2}\right).$$

3.5. *Isotropic Gaussian Mixture Models.* In this section, we consider the isotropic Gaussian mixture models where the noises are sampled from $\mathcal{N}(0,\sigma^2 I_p)$ independently. As a special case of the sub-Gaussian mixture models, Theorem 3.1 can be directly applied. Nevertheless, the isotropic Gaussian noises make it possible to remove the spectral gap condition (21). In addition, we study the performance of the spectral clustering $\hat{z}$ from Algorithm 1 with exactly the leading k singular vectors, regardless of κ the rank of matrix P. As a result, it requires no knowledge on κ and needs no adaptive dimension reduction such as Algorithm 2. We have the following theorem on its performance.

THEOREM 3.3. *Consider the spectral clustering $\hat{z}$ of Algorithm 1 with $r = k$. Assume $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2 I_p)$ for each $i \in [n]$. Assume $\beta n/k^4 \geq 100$ and*

$$\frac{\Delta}{k^{3.5}\beta^{-0.5}\left(1+\frac{p}{n}\right)\sigma} \to \infty. \tag{25}$$

We have

$$\mathbb{E}\ell(\hat{z},z^*) \leq \exp\left(-\left(1 - C\left(\frac{\Delta}{k^{3.5}\beta^{-0.5}\left(1+\frac{p}{n}\right)\sigma}\right)^{-0.25}\right)\frac{\Delta^2}{8\sigma^2}\right) + 2e^{-0.08n}, \tag{26}$$

where $C > 0$ is some constant.

Theorem 3.3 shows that asymptotically $\mathbb{E}\ell(\hat{z},z^*) \leq \exp(-(1-o(1))\Delta^2/(8\sigma^2)) + 2\exp(-0.08n)$ where the first term dominates when $\Delta^2/\sigma^2 = o(n)$. The minmax lower bound for recovering z^* under the given model is established in [26]: $\inf_{\hat{z}} \sup_{(\theta_1^*,\ldots,\theta_k^*),z^*} \mathbb{E}\ell(\hat{z},z^*) \geq \exp(-(1+o(1))\Delta^2/(8\sigma^2))$ as long as $\Delta^2/\sigma^2 \gg \log(k\beta^{-1})$. This immediately implies that the considered estimator is minimax optimal. Theorem 3.3 also implies $\hat{z}$ achieves the exact recovery $\mathbb{E}\{\ell(\hat{z},z^*) \neq 0\} = o(1)$ when $\Delta/\sigma \geq (1+c)2\sqrt{2\log n}$ for any small constant $c > 0$. When $\Delta/\sigma \leq (1-c)2\sqrt{2\log n}$, no algorithm is able to recover z^* exactly with high probability according to the minimax lower bound.

It is worth mentioning that Theorem 3.3 requires no spectral gap condition such as (8) or (21). The purpose of such conditions is to ensure that singular vectors of X are well controlled, especially those corresponding to small non-zero singular values of the signal matrix P. When the noises are isotropic Gaussian, the distribution of each right singular vector $\hat{v}_j$ is well-behaved for any $j \in [p \wedge n]$. Lemma 4.4 of [25] shows that each $(I - V_{1:\kappa}V_{1:\kappa}^T)\hat{v}_j$ is Haar distributed on the sphere spanned by $(I - V_{1:\kappa}V_{1:\kappa}^T)$, where $V_{1:\kappa} := (v_1, v_2, \ldots, v_\kappa) \in \mathbb{O}^{n\times\kappa}$ is the right singular subspace of the signal matrix P. Theorem 3.3 is about the singular subspace $\hat{U}_{1:k}$. In its proof, we decompose it into $\hat{U}_{1:r}$ and $\hat{U}_{(r+1):k}$, for some index $r \in [\kappa]$ with sufficient large spectral gap $\lambda_r - \lambda_{r+1}$ so that the contribution of $\hat{U}_{1:r}$ can be precisely quantified following similar arguments used to establish Lemma 3.3 and Theorem 3.1. The contribution of each $\hat{u}_j$ where $j \in \{r+1,\ldots,k\}$ is eventually connected with properties of the corresponding right singular vector $\hat{v}_j$, particularly, the distribution of $(I - V_{1:\kappa}V_{1:\kappa}^T)\hat{v}_j$. These two sources of errors together lead to the upper bound (26).

The performance of Algorithm 1 with $r = k$ under the same isotropic Guassian mixture model is the main topic of [25] which derives a similar upper bound for $\mathbb{E}\ell(\hat{z},z^*)$ assuming $\Delta/(\beta^{-0.5}k^{10.5}(1+p/n)) \to \infty$. The key technical tool used in [25] is spectral operator perturbation theory of [20, 21] on the difference between empirical singular subspaces and population ones, which works for the Gaussian noise case and it is not clear whether it is possible to be extended to other distributions including sub-Gaussian distributions. In this paper, the proof of Theorem 3.3 is completely different, using Theorem 2.3 on the difference between empirical singular subspaces and their leave-one-out counterparts. We not only recover the main result of [25] with a much shorter proof, but also improve the dependence of k. Despite that Theorem 3.3 needs an extra condition $\beta n/k^4 \geq 100$, it only requires $k^{3.5}$ to satisfy (25), while [25] needs $k^{10.5}$ instead which is a stronger condition.

3.6. *Lower Bounds and Sub-optimality of Spectral Clustering.* In the above sections, we focus on quantifying the performance of spectral clustering under mixture models. An interesting question is whether the spectral clustering is optimal or not. When the noise is the isotropic Gaussian, Theorem 3.3 matches with the minimax rate assuming (25) holds, showing that the spectral clustering is indeed optimal in this case. It remains unclear whether the spectral clustering is optimal or not when the noise is beyond the isotropic Gaussian model.

To answer this question, in this section we consider a two-cluster symmetric mixture model whether the centers are proportional to $\mathbb{1}_p$ and the noises have i.i.d. entries. This setup makes it possible to apply the central limit theorem to characterize the performance of the spectral clustering with sharp upper and lower bounds, as $\mathbb{1}_p^T\epsilon_i$ is asymptotically normal for each $i \in [n]$ when p is large.

A Two-cluster Symmetric Mixture Model. Consider a mixture model (5) with two clusters such that

$$\theta_1^* = -\theta_2^* = \delta\mathbb{1}_p, \text{ and } \{\epsilon_{i,j}\}_{i\in[n],j\in[p]} \overset{iid}{\sim} F, \tag{27}$$

for some $\delta \in \mathbb{R}$ and some distribution F, where $\{\epsilon_{i,j}\}_{j\in[p]}$ are entries of ϵ_i for each $i \in [n]$.

Under the above model (27), we have $k = 2$, $\Delta = 2\sqrt{p}\delta$ and the largest singular value $\lambda_1 = \delta\sqrt{np}$. Since the signal matrix matrix P is rank-one (i.e., $\kappa = 1$) with $u_1 = (1/\sqrt{p})\mathbb{1}_p$, a natural idea is to cluster using the first singular vector only. Define

$$\left(\check{z}, \{\check{c}_j\}_{j=1}^2\right) = \operatorname*{argmin}_{z\in[2]^n, \{c_j\}_{j=1}^2\in\mathbb{R}} \sum_{i\in[n]} \left(\hat{u}_1^T X_i - c_{z_i}\right)^2. \tag{28}$$

The performance of the spectral estimator $\check{z}$ will be the focus in this section. Note that $\hat{u}_1^T X = \hat{\lambda}_1 \hat{v}_1^T$ where $\hat{v}_1$ is the leading right singular vector of X, so $\check{z}$ equivalently performs clustering on $\{\hat{v}_{1,i}\}_{i\in[n]}$, the entries of $\hat{v}_1$. This is closely related to the sign estimator $\{\text{sign}(\hat{v}_{1,i})\}_{i\in[n]}$, which estimates the cluster assignment by the signs of $\{\hat{v}_{1,i}\}_{i\in[n]}$.

Since $\check{z}$ is exactly the spectral clustering $\hat{z}$ of Algorithm 1 with $r = 1$, Theorem 3.1 can be directly applied when noises are sub-Gaussian and yields the following result. Under the model (27), assume that F is a $\text{SG}(\sigma^2)$ distribution with zero mean and $\beta n > 40$. There exist constants $C, C' > 0$ such that under the assumption that

$$\psi_3 := \frac{\Delta}{\beta^{-0.5}\left(1 + \sqrt{\frac{p}{n}}\right)\sigma} > C,$$

we have $\mathbb{E}\ell(\check{z}, z^*) \leq \exp(-(1 - C'\psi_3^{-1})\Delta^2/(8\sigma^2)) + \exp(-n/2)$.

The special structure of (27) makes it possible to derive a sharper upper bound than the above one and a matching lower bound on the performance of $\check{z}$ with some additional assumption on the distribution F. Instead of directly using Lemma 3.2 (which leads to Theorem 3.1 and then the above upper bound), we can further connect the clustering error with $u_1^T\epsilon_i$ where $u_1^T\epsilon_i = p^{-1/2}\sum_{j=1}^p \epsilon_{i,j}$ is approximately normally distributed when p is large. On the other hand, the structure of (27) enables us to have a lower bound for $\mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\}$ that is in an opposite direction of Lemma 3.2. See Lemma 5.1 for details. The key technical tool used is Theorem 2.2 on the perturbation $|\hat{u}_1\hat{u}_1^T - \hat{u}_{-i,1}\hat{u}_{-i,1}^T|$ for all $i \in [n]$. These together give a sharp and matching lower bound for $\mathbb{E}\ell(\check{z}, z^*)$ where the clustering error is essentially determined by Δ and the variance $\bar{\sigma}^2$.

THEOREM 3.4. *Consider the model (27). For any $\xi \sim F$, assume $\mathbb{E}\xi = 0, Var(\xi) = \bar{\sigma}^2$, and $\xi \sim SG(\sigma^2)$ where $\sigma \leq C\bar{\sigma}$ for some constant $C > 0$. Assume $\beta n > 40$. Then there exist constants $C', C'', C''' > 0$ such that if $\psi_3 \geq C'$, we have*

$$\mathbb{E}\ell(\check{z}, z^*) \leq \exp\left(-\frac{\left(1 - C''\psi_3^{-1}\right)^2 \Delta^2}{8\bar{\sigma}^2}\right) + \exp\left(-C''\sqrt{p}\right) + \exp\left(-\frac{n}{2}\right),$$

$$\text{and} \quad \mathbb{E}\ell(\check{z}, z^*) \geq \exp\left(-\frac{\left(1 + C'''\psi_3^{-1}\right)^2 \Delta^2}{8\bar{\sigma}^2}\right) - \exp\left(-C'''\sqrt{p}\right) - \exp\left(-\frac{n}{2}\right).$$

In Theorem 3.4, the term $\exp(-C''\sqrt{p})$ is due to the normal approximation of $u_1^T\epsilon_i$ and decays when the dimensionality p increases. The term $\exp(-n/2)$ is due to a with-high-probability event on $\|E\|$. If additionally $\Delta/\bar{\sigma} \ll \max\{p^{1/4}, n^{1/2}\}$ is assumed, Theorem 3.4 concludes asymptotically

$$\mathbb{E}\ell(\check{z}, z^*) = \exp\left(-\frac{(1+c)\Delta^2}{8\bar{\sigma}^2}\right), \tag{29}$$

for some small constant c.

The upper and lower bounds in Theorem 3.4 give a sharp characterization on the performance of $\check{z}$. To answer the question of whether it is optimal or not, we need to establish the minimax rate for the clustering task under the model (27). Since the model (27) is essentially about a testing between two parametric distributions, the optimal procedure is the likelihood ratio test. According to the classical asymptotics theory [37], the likelihood ratio behaves like a normal random variable as $p \to \infty$ under some regularity condition. This leads to an error rate determined by Δ and the Fisher information.

LEMMA 3.4. *Consider the model (27). Assume the distribution F has a positive, continuously differentiable density f with mean zero and finite Fisher information $\mathcal{I} := \int (f'/f)^2 f \mathrm{d}x$. Assume Δ is a constant. We have*

$$C_1 \exp\left(-\frac{\Delta^2}{8\mathcal{I}^{-1}}\right) \leq \lim_{p\to\infty} \inf_{z} \sup_{z^* \in [2]^n} \mathbb{E}\ell(z, z^*) \leq C_2 \exp\left(-\frac{\Delta^2}{8\mathcal{I}^{-1}}\right), \tag{30}$$

for some constants $C_1, C_2 > 0$.

With Lemma 3.4, the question of whether $\check{z}$ is optimal or not boils down to a comparison of the variance $\bar{\sigma}^2$ and the inverse of the Fisher information $\mathcal{I}^{-1}$. Due to the fact that $\mathcal{I}^{-1} \leq \bar{\sigma}^2$ and the equation holds if and only if F is a normal distribution, we have the following conclusion.

THEOREM 3.5. *Consider the model (27). Assume all the assumptions needed in Theorem 3.4 and Lemma 3.4 hold. Then the spectral clustering $\check{z}$ is in general suboptimal, i.e., it fails to achieve the minimax rate (30). It is optimal if and only if the noise distribution F is $N(0, \bar{\sigma}^2)$.*

Theorem 3.5 establishes the sub-optimality of the spectral clustering $\check{z}$ under the model (27). Though $\check{z}$ achieves an exponential error rate, it has a fundamentally sub-optimal exponent involving $\bar{\sigma}^2$ instead of $\mathcal{I}^{-1}$. This is due to the fact $\check{z}$ clusters data points based on Euclidean distances while the optimal procedure is the likelihood ratio test. Only when the noise is normally distributed, the likelihood ratio test is equivalent to a comparison of two Euclidean distances, leading to the optimality of $\check{z}$ in the Gaussian case. Despite that Theorem 3.5 is only limited to the model (27), the above reasoning suggests the spectral clustering is generally sub-optimal under mixture models beyond (27) unless the noise is Gaussian.

4. Proof of Main Results in Section 2. In this section, we give the proofs of Theorem 2.1 and Theorem 2.2. The proof of Theorem 2.3 is included in the supplement [42] due to page limit.

4.1. *Proof of Theorem 2.1.* Before giving the proof of Theorem 2.1, we first present and prove a slightly more general perturbation result, Theorem 4.1, which only requires $\sigma_r^2 - \sigma_{r+1}^2 - \|(I-U_rU_r^T)y_n\|^2 > 0$ instead of assuming $\rho > 2$. We defer the proof of Theorem 2.1 to the end of this section, which is an immediate consequence of Theorem 4.1.

THEOREM 4.1. *If $\sigma_r^2 - \sigma_{r+1}^2 - \|(I-U_rU_r^T)y_n\|^2 > 0$, we have*

$$\left\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\right\|_{\mathrm{F}} \leq \frac{2\sqrt{2}\sigma_r\left\|(I-U_rU_r^T)y_n\right\|}{\sigma_r^2 - \sigma_{r+1}^2 - \|(I-U_rU_r^T)y_n\|^2}\sqrt{\sum_{i=1}^r\left(\frac{u_i^Ty_n}{\sigma_i}\right)^2}.$$

PROOF. Decompose y_n into $y_n = \theta + \epsilon$ with $\theta := U_rU_r^Ty_n$ and $\epsilon := (I-U_rU_r^T)y_n$. Then we have $u_i^T\theta = u_i^Ty_n$ for each $i\in[r]$.

Throughout the proof, we denote

$$\alpha^2 = \left\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\right\|_{\mathrm{F}}^2.$$

Denote $d = p\wedge(n-1)$. If $p \leq n-1$, we have $d = p$ and denote $U := (u_1,\ldots,u_p)\in\mathbb{R}^{p\times p}$ which is an orthogonal matrix. If $p > n-1$, we let $U\in\mathbb{R}^{p\times p}$ be an orthogonal matrix with the first $p\wedge(n-1)$ columns being $u_1,\ldots,u_{p\wedge(n-1)}$. In both cases, we have U being an orthogonal matrix. Then $\hat{U}_r$ can be written as $\hat{U}_r = U\hat{B}$ for some $\hat{B} = (\hat{B}_{i,j})\in\mathbb{R}^{p\times r}$. Let $\hat{B}_{i,\cdot}$ be the ith row of $\hat{B}$ for each $i\in[p]$. Define $b_i^2 = 1 - \|\hat{B}_{i,\cdot}\|^2$ for each $i\in[r]$ and $b_i^2 = \|\hat{B}_{i,\cdot}\|^2$ for each $i > r$. Then we have

$$\begin{aligned}\alpha^2 &= \left\|\hat{U}_r\hat{U}_r^T\right\|_{\mathrm{F}}^2 + \left\|U_rU_r^T\right\|_{\mathrm{F}}^2 - 2\left\langle\hat{U}_r\hat{U}_r^T, U_rU_r^T\right\rangle\\ &= 2k - 2\left\|U_r^T\hat{U}_r\right\|_{\mathrm{F}}^2 = 2k - 2\sum_{i\in[r]}\sum_{j\in[r]}\hat{B}_{i,j}^2\\ &= 2\sum_{i\in[r]}b_i^2 = 2\sum_{i=r+1}^p b_i^2,\end{aligned}\tag{31}$$

where in the last equation we use the fact that $\|\hat{B}\|_{\mathrm{F}}^2 = r$.

Note that $\hat{U}_r\hat{U}_r^T\hat{Y}$ is the best rank-r approximation of $\hat{Y}$. We have

$$\left\|\left(I-\hat{U}_r\hat{U}_r^T\right)\hat{Y}\right\|_{\mathrm{F}}^2 \leq \left\|\left(I-U_rU_r^T\right)\hat{Y}\right\|_{\mathrm{F}}^2.$$

Due to the fact $\hat{Y} = (Y, y_n)$, we have

$$\left\|\left(I - \hat{U}_r\hat{U}_r^T\right)Y\right\|_{\mathrm{F}}^2 + \left\|\left(I - \hat{U}_r\hat{U}_r^T\right)y_n\right\|^2 \leq \left\|\left(I - U_rU_r^T\right)Y\right\|_{\mathrm{F}}^2 + \left\|\left(I - U_rU_r^T\right)y_n\right\|^2,$$

which implies

(32)

$$\left\|\left(I - \hat{U}_r\hat{U}_r^T\right)Y\right\|_{\mathrm{F}}^2 - \left\|\left(I - U_rU_r^T\right)Y\right\|_{\mathrm{F}}^2 \leq \left\|\left(I - U_rU_r^T\right)y_n\right\|^2 - \left\|\left(I - \hat{U}_r\hat{U}_r^T\right)y_n\right\|^2.$$

We are going to simplify terms in (32).

(Simplification of the LHS of (32)). Recall the decomposition $Y = \sum_{i\in[d]}\sigma_i u_i v_i^T$. Since $\left(I - U_rU_r^T\right)Y = \sum_{i>r}^d \sigma_i u_i v_i^T$, we have $\left\|\left(I - U_rU_r^T\right)Y\right\|_{\mathrm{F}}^2 = \sum_{i>r}^d \sigma_i^2$. Since

$$U^TY = U^T\left(\sum_{i\in[d]}\sigma_i u_i v_i^T\right) = \begin{pmatrix}\sigma_1 v_1^T\\ \cdots\\ \sigma_d v_d^T\\ 0_{p-d}\end{pmatrix} = \mathrm{diag}(\sigma_1,\ldots,\sigma_d,0_{p-d})\begin{pmatrix}v_1^T\\ \cdots\\ v_d^T\\ O_{(p-d)\times n}\end{pmatrix},$$

we have

$$\begin{aligned}\left\|\left(I - \hat{U}_r\hat{U}_r^T\right)Y\right\|_{\mathrm{F}}^2 &= \left\|U\left(I - U^T\hat{U}_r\hat{U}_r^TU\right)U^TY\right\|_{\mathrm{F}}^2\\ &= \left\|\left(I - \hat{B}\hat{B}^T\right)\mathrm{diag}(\sigma_1,\ldots,\sigma_d,0_{p-d})\begin{pmatrix}v_1^T\\ \cdots\\ v_d^T\\ O_{(p-d)\times n}\end{pmatrix}\right\|_{\mathrm{F}}^2\\ &= \mathrm{tr}\left(\mathrm{diag}(\sigma_1,\ldots,\sigma_d,0_{p-d})\left(I - \hat{B}\hat{B}^T\right)\mathrm{diag}(\sigma_1,\ldots,\sigma_d,0_{p-d})\begin{pmatrix}I_{d\times d} & \\ & O_{(p-d)\times(p-d)}\end{pmatrix}\right),\end{aligned}$$

where in the last equation we use the following facts: (1) for any two square matrices of the same size A, D, we have $\|AD\|_{\mathrm{F}}^2 = \mathrm{tr}(D^TA^TAD) = \mathrm{tr}(A^TADD^T)$; (2) $\hat{B}$ has orthogonal columns such that $(I - \hat{B}\hat{B}^T)^2 = I - \hat{B}\hat{B}^T$; and (3) $\{v_1,\ldots,v_d\} \in \mathbb{R}^{n-1}$ are orthogonal vectors. Since the diagonal entries of $\hat{B}\hat{B}^T$ are $\{\|\hat{B}_{i,\cdot}\|^2\}_{i\in[p]}$, we have

$$\begin{aligned}\left\|\left(I - \hat{U}_r\hat{U}_r^T\right)Y\right\|_{\mathrm{F}}^2 &= \mathrm{tr}\left(\mathrm{diag}(\sigma_1,\ldots,\sigma_d,0_{p-d})\left(I - \hat{B}\hat{B}^T\right)\mathrm{diag}(\sigma_1,\ldots,\sigma_d,0_{p-d})\right)\\ &= \sum_{i=1}^d \sigma_i^2\left(1 - \left\|\hat{B}_{i,\cdot}\right\|_{\mathrm{F}}^2\right).\end{aligned}$$

Then we have

$$\text{LHS of (32)} = \sum_{i=1}^r \sigma_i^2\left(1 - \left\|\hat{B}_{i,\cdot}\right\|_{\mathrm{F}}^2\right) - \sum_{i>r}^d \sigma_i^2\left\|\hat{B}_{i,\cdot}\right\|_{\mathrm{F}}^2 = \sum_{i=1}^r \sigma_i^2 b_i^2 - \sum_{i>r}^d \sigma_i^2 b_i^2 \geq \sum_{i=1}^r \sigma_i^2 b_i^2 - \sigma_{r+1}^2\frac{\alpha^2}{2},$$

where we use $\sum_{i>r}^d b_i^2 \leq \sum_{i>r}^p b_i^2 = \alpha^2/2$ from (31) in the last inequality .

(Simplification of the RHS of (32)). Recall that $\hat{U}_r = U\hat{B}$. We decompose it into $\hat{B} = (\hat{B}_1^T, \hat{B}_2^T)^T$ where $\hat{B}_1 \in \mathbb{R}^{r\times r}$ are the first r rows and $\hat{B}_2 \in \mathbb{R}^{(p-r)\times r}$. We have

$$\text{RHS of (32)} = y_n^T\left(I - U_rU_r^T\right)y_n - y_n^T\left(I - \hat{U}_r\hat{U}_r^T\right)y_n$$

$$= y_n^T \left(\hat{U}_r \hat{U}_r^T - U_r U_r^T\right) y_n$$

$$= y_n^T U \begin{pmatrix} \hat{B}_1 \hat{B}_1^T - I_{r\times r} & \hat{B}_1 \hat{B}_2^T \\ \hat{B}_2 \hat{B}_1^T & \hat{B}_2 \hat{B}_2^T \end{pmatrix} U^T y_n.$$

Define $\hat{B}^\perp \in \mathbb{R}^{p\times(p-r)}$ to be the matrix such that $(\hat{B}, \hat{B}^\perp) \in \mathbb{R}^{p\times p}$ is an orthonormal matrix. We can further decompose it into $\hat{B}^\perp = (\hat{B}_1^{\perp T}, \hat{B}_2^{\perp Y})^T$ where $\hat{B}_1^\perp \in \mathbb{R}^{r\times(p-r)}$ including the first r rows and $\hat{B}_2^\perp \in \mathbb{R}^{(p-r)\times(p-r)}$. Since $(\hat{B}, \hat{B}^\perp)$ has orthogonal columns, we have

$$(\hat{B}_1, \hat{B}_1^\perp)(\hat{B}_1, \hat{B}_1^\perp)^T = \hat{B}_1 \hat{B}_1^T + \hat{B}_1^\perp \hat{B}_1^{\perp T} = I_{r\times r},$$

and $(\hat{B}_1, \hat{B}_1^\perp)(\hat{B}_2, \hat{B}_2^\perp)^T = O_{r\times(p-r)}$, which implies

$$\hat{B}_1 \hat{B}_2^T = -\hat{B}_1^\perp \hat{B}_2^{\perp T}.$$

We also decompose the matrix $U =: (U_r, U_\perp)$. Then

$$\begin{aligned} \text{RHS of (32)} &= y_n^T (U_r, U_\perp) \begin{pmatrix} -\hat{B}_1^\perp \hat{B}_1^{\perp T} & -\hat{B}_1^\perp \hat{B}_2^{\perp T} \\ -\hat{B}_2^\perp \hat{B}_1^{\perp T} & \hat{B}_2 \hat{B}_2^T \end{pmatrix} (U_r, U_\perp)^T y_n \\ &= -y_n^T U_r \hat{B}_1^\perp \hat{B}_1^{\perp T} U_r^T y_n - 2 y_n^T U_r \hat{B}_1^\perp \hat{B}_2^{\perp T} U_\perp^T y_n + y_n^T U_\perp \hat{B}_2 \hat{B}_2^T U_\perp^T y_n \\ &\le -\left\|\hat{B}_1^{\perp T} U_r^T y_n\right\|^2 + 2\left\|\hat{B}_1^{\perp T} U_r^T y_n\right\| \left\|\hat{B}_2^{\perp T}\right\| \left\|U_\perp^T y_n\right\| + \left\|\hat{B}_2^T\right\|^2 \left\|U_\perp^T y_n\right\|^2. \end{aligned}$$

Note that $\|\hat{B}_2^{\perp T}\| \le 1$ and $\|\hat{B}_2^T\|^2 \le \|\hat{B}_2^T\|_{\mathrm{F}}^2 = \sum_{i>r}^p \|\hat{B}_{i,\cdot}\|^2 = \alpha^2/2$ which is by (31). We also have

$$\left\|U_\perp^T y_n\right\| = \|\epsilon\|.$$

Since $\|\hat{B}_1^\perp\|_{\mathrm{F}}^2 = \sum_{i=1}^r \left(1 - \|\hat{B}_{i,\cdot}\|^2\right) = \alpha^2/2$ according to (31), we have $\|\hat{B}_1^\perp\| \le \alpha/\sqrt{2}$. Thus, using $U_r^T \epsilon = 0$, we have

$$\left\|\hat{B}_1^{\perp T} U_r^T y_n\right\| = \left\|\hat{B}_1^{\perp T} U_r^T \theta\right\|.$$

Then,

$$\text{RHS of (32)} \le 2\left\|\hat{B}_1^{\perp T} U_r^T \theta\right\| \|\epsilon\| + \frac{\alpha^2}{2} \|\epsilon\|^2.$$

To simplify $\|\hat{B}_1^{\perp T} U_r^T \theta\|$, denote $w_i = u_i^T \theta$ and $s_i = |w_i|/\sigma_i$ for each $i \in [r]$. Recall that $u_i^T \theta = u_i^T y_n$ for each $i \in [r]$. We have

$$s_i = \left|\frac{u_i^T y_n}{\sigma_i}\right|, \forall i \in [r].$$

We then have

$$\left\|\hat{B}_1^{\perp T} U_r^T \theta\right\| = \left\|\sum_{i=1}^r w_i \hat{B}_{i,\cdot}^\perp\right\| \le \sum_{i=1}^r |w_i| \left\|\hat{B}_{i,\cdot}^\perp\right\| = \sum_{i=1}^r s_i \sigma_i |b_i| \le \|s\| \sqrt{\sum_{i=1}^r \sigma_i^2 b_i^2},$$

where we denote the ith row of $\hat{B}_1^\perp$ as $\hat{B}_{i,\cdot}^\perp$ and we use the fact that $\|\hat{B}_{i,\cdot}^\perp\|^2 = 1 - \|\hat{B}_{i,\cdot}\|^2 = b_i^2$ for each $i \in [r]$. As a result,

$$\text{RHS of (32)} \le 2\|s\| \sqrt{\sum_{i=1}^r \sigma_i^2 b_i^2}\, \|\epsilon\| + \frac{\alpha^2}{2} \|\epsilon\|^2.$$

(Combining the above simplifications for (32)). From the above simplifications on the LHS and RHS of (32), we have

$$\sum_{i=1}^{r}\sigma_i^2 b_i^2-\sigma_{r+1}^2\frac{\alpha^2}{2}\leq 2\left\|s\right\|\sqrt{\sum_{i=1}^{r}\sigma_i^2 b_i^2}\left\|\epsilon\right\|+\frac{\alpha^2}{2}\left\|\epsilon\right\|^2.$$

Define $t=\sqrt{\sum_{i=1}^{r}\sigma_i^2 b_i^2}$. Then after arrangement, the above display becomes

$$t^2-2\left\|s\right\|\left\|\epsilon\right\|t\leq\sigma_{r+1}^2\frac{\alpha^2}{2}+\frac{\alpha^2}{2}\left\|\epsilon\right\|^2.$$

Note that the function $t^2-2\left\|s\right\|\left\|\epsilon\right\|t$ is increasing as long as $t\geq t_0$ where we define $t_0:=\left\|s\right\|\left\|\epsilon\right\|$. On the other hand, from (31), we have the domain $t\geq\alpha\sigma_r/\sqrt{2}$. We consider the following two scenarios.

If $\alpha\sigma_r/\sqrt{2}\leq t_0$, we have

$$\alpha\leq\frac{\sqrt{2}t_0}{\sigma_r}=\frac{\sqrt{2}\left\|s\right\|\left\|\epsilon\right\|}{\sigma_r}. \tag{33}$$

If $\alpha\sigma_r/\sqrt{2}>t_0$, we have

$$t^2-2\left\|s\right\|t\geq\frac{\alpha^2\sigma_r^2}{2}-\sqrt{2}\left\|s\right\|\left\|\epsilon\right\|\alpha\sigma_r.$$

Hence, we have an inequality of α:

$$\frac{\alpha^2\sigma_r^2}{2}-\sqrt{2}\left\|s\right\|\left\|\epsilon\right\|\alpha\sigma_r\leq\sigma_{r+1}^2\frac{\alpha^2}{2}+\frac{\alpha^2}{2}\left\|\epsilon\right\|^2,$$

which can be arranged into

$$\frac{\alpha}{2}\left(\sigma_r^2-\sigma_{r+1}^2-\left\|\epsilon\right\|^2\right)\leq\sqrt{2}\left\|s\right\|\sigma_r\left\|\epsilon\right\|.$$

Hence, under the assumption $\sigma_r^2-\sigma_{r+1}^2-\left\|\epsilon\right\|^2>0$, we have

$$\alpha\leq\frac{2\sqrt{2}\sigma_r\left\|s\right\|\left\|\epsilon\right\|}{\sigma_r^2-\sigma_{r+1}^2-\left\|\epsilon\right\|^2}. \tag{34}$$

Since $2\sigma_r^2>\sigma_r^2-\sigma_{r+1}^2-\left\|\epsilon\right\|^2$, the upper bound in (33) is strictly below that in (34). Hence, (34) holds for both scenarios. The proof is complete. □

PROOF OF THEOREM 2.1. Since we assume $\rho>2$, we have

$$\begin{aligned}\sigma_r^2-\sigma_{r+1}^2-\left\|(I-U_rU_r^T)\epsilon_i\right\|^2&\geq\sigma_r(\sigma_r-\sigma_{r+1})-(\sigma_r-\sigma_{r+1})^2/4\\&\geq\sigma_r(\sigma_r-\sigma_{r+1})/2=\rho\sigma_r\left\|(I-U_rU_r^T)\epsilon_i\right\|/2.\end{aligned}$$

Together with Theorem 4.1, we obtain the desired bound. □

4.2. *Proof of Theorem 2.2.*

PROOF OF THEOREM 2.2. Consider any $i\in[n]$. In order to apply Theorem 2.1, we need to verify that the spectral gap assumption (3) is satisfied. That is, define

$$\rho_{-i}:=\frac{\hat{\lambda}_{-i,\kappa}-\hat{\lambda}_{-i,\kappa+1}}{\left\|\left(I-\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\right)X_i\right\|}.$$

We need to show $\rho_{-i} > 2$. In the following, we provide a lower bound for the numerator $\hat{\lambda}_{-i,\kappa} - \hat{\lambda}_{-i,\kappa+1}$.

Define $\lambda_{-i,1} \geq \lambda_{-i,2} \geq \ldots \geq \lambda_{-i,p\wedge(n-1)}$ to be singular values of P_{-i}, the leave-one-out counterpart of the signal matrix P where

$$P_{-i} := (\theta^*_{z^*_1}, \ldots, \theta^*_{z^*_{i-1}}, \theta^*_{z^*_{i+1}}, \ldots, \theta^*_{z^*_n}) \in \mathbb{R}^{p\times(n-1)}. \tag{35}$$

We are interested in the value of $\lambda_{-i,\kappa}$. Recall that λ_κ is the κth largest singular value of P which is rank-κ. Since P has k unique columns $\{\theta^*_a\}_{a\in[k]}$, its left singular vectors $u_j \in \Theta$ for each $j \in [k]$ where $\Theta := \text{span}(\{\theta^*_a\}_{a\in[k]})$. Note that each θ^*_a appears at least $\beta n/k$ times in the columns of P. Then P_{-i} also has these k unique columns with each appearing at least $\beta n/k - 1$ times. This concludes that P_{-i} has the same leading left singular vector space as P. We then have

$$\begin{aligned}
\lambda^2_{-i,\kappa} &= \min_{w\in\Theta:\|w\|=1} \left\|w^T P_{-i}\right\|^2 = \min_{w\in\Theta:\|w\|=1} \sum_{j\in[n]:j\neq i} (w^T\theta^*_{z^*_j})^2 \\
&\geq \frac{\frac{\beta n}{k} - 1}{\frac{\beta n}{k}} \min_{w\in\Theta:\|w\|=1} \sum_{j\in[n]} (w^T\theta^*_{z^*_j})^2 = \left(1 - \frac{k}{\beta n}\right) \min_{w\in\Theta:\|w\|=1} \left\|w^T P\right\|^2 \\
&\geq \left(1 - \frac{k}{\beta n}\right) \lambda^2_\kappa. \qquad (36)
\end{aligned}$$

We also have $\lambda_{-i,\kappa+1} = 0$ as P_{-i} is rank-κ.

Next, we are going to analyze $\hat{\lambda}_{-i,\kappa}$ and $\hat{\lambda}_{-i,\kappa+1}$, the κth and $(\kappa+1)$th largest singular values of X_{-i}. Recall the SVD of X_{-i} in Section 2.2. Define

$$E_{-i} := (\epsilon_1, \ldots, \epsilon_{i-1}, \epsilon_{i+1}, \ldots, \epsilon_n) \in \mathbb{R}^{p\times(n-1)}, \tag{37}$$

so that $X_{-i} = P_{-i} + E_{-i}$. By Weyl's inequality, we have $|\lambda_{-i,\kappa} - \hat{\lambda}_{-i,\kappa}|, |\lambda_{-i,\kappa+1} - \hat{\lambda}_{-i,\kappa+1}| \leq \|E_{-i}\| \leq \|E\|$. Then we have

$$\hat{\lambda}_{-i,\kappa} \geq \lambda_{-i,\kappa} - \|E\| \geq \sqrt{1 - \frac{k}{\beta n}}\lambda_\kappa - \|E\| \tag{38}$$

and

$$\hat{\lambda}_{-i,\kappa} - \hat{\lambda}_{-i,\kappa+1} \geq \lambda_{-i,\kappa} - \lambda_{-i,\kappa+1} - 2\|E\| \geq \sqrt{1 - \frac{k}{\beta n}}\lambda_\kappa - 2\|E\|. \tag{39}$$

Next, we study $\|(I - \hat{U}_{-i,1:\kappa}\hat{U}^T_{-i,1:\kappa})X_i\|$. Since $\hat{U}_{-i,1:\kappa}\hat{U}^T_{-i,1:\kappa}X_{-i}$ is the best rank-κ approximation of X_{-i}, we have

$$\left\|\hat{U}_{-i,1:\kappa}\hat{U}^T_{-i,1:\kappa}X_{-i} - X_{-i}\right\| \leq \|P_{-i} - X_{-i}\| = \|E_{-i}\|,$$

where we use the fact that P_{-i} is rank-κ. Then by the triangle inequality, we have

$$\begin{aligned}
&\left\|\left(I - \hat{U}_{-i,1:\kappa}\hat{U}^T_{-i,1:\kappa}\right)P_{-i}\right\| \\
&= \left\|\hat{U}_{-i,1:\kappa}\hat{U}^T_{-i,1:\kappa}P_{-i} - P_{-i}\right\| \\
&\leq \left\|\hat{U}_{-i,1:\kappa}\hat{U}^T_{-i,1:\kappa}(P_{-i} - X_{-i})\right\| + \left\|\hat{U}_{-i,1:\kappa}\hat{U}^T_{-i,1:\kappa}X_{-i} - X_{-i}\right\| + \|X_{-i} - P_{-i}\| \\
&\leq 3\|E_{-i}\|.
\end{aligned}$$

Using the fact P_{-i} is rank-κ again, we have

$$\left\|\left(I-\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\right)P_{-i}\right\|_{\mathrm{F}} \leq \sqrt{\kappa}\left\|\left(I-\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\right)P_{-i}\right\| \leq 3\sqrt{\kappa}\left\|E_{-i}\right\| \leq 3\sqrt{\kappa}\left\|E\right\|.$$

Since P_{-i} has at least $\beta n/k-1$ columns being exactly $\theta^*_{z_i^*}$, we have

$$\left\|\left(I-\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\right)\theta^*_{z_i^*}\right\| \leq \frac{\left\|\left(I-\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\right)P_{-i}\right\|_{\mathrm{F}}}{\sqrt{\frac{\beta n}{k}-1}} \leq \frac{3\sqrt{\kappa}\left\|E\right\|}{\sqrt{\frac{\beta n}{k}-1}}, \tag{40}$$

and consequently,

$$\begin{aligned}\left\|\left(I-\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\right)X_i\right\| &\leq \left\|\left(I-\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\right)\theta^*_{z_i^*}\right\| + \left\|\left(I-\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\right)\epsilon_i\right\| \\ &\leq \frac{3\sqrt{\kappa}\left\|E\right\|}{\sqrt{\frac{\beta n}{k}-1}} + \left\|E\right\|. \end{aligned}\tag{41}$$

From (39) and (41), we have

$$\rho_{-i} \geq \frac{\sqrt{1-\frac{k}{\beta n}}\lambda_\kappa - 2\left\|E\right\|}{\left\|E\right\| + \frac{3\sqrt{\kappa}\|E\|}{\sqrt{\frac{\beta n}{k}-1}}} \geq \frac{\rho_0}{8} > 2, \tag{42}$$

where the last inequality is due to the assumption $\rho_0 > 16$ and $\beta n/k^2 \geq 10$.

The next thing to do is to study $\{\hat{u}_{-i,a}^T X_i\}_{a\in[\kappa]}$. Denote the columns of P_{-i} and E_{-i} as $\{(P_{-i})_{\cdot,j}\}_{j\in[n-1]}$ and $\{(E_{-i})_{\cdot,j}\}_{j\in[n-1]}$, respectively. Define $S := \{j\in[n-1] : (P_{-i})_{\cdot,j} = \theta^*_{z_i^*}\}$. Then for any $a\in[\kappa]$, by the SVD of X_{-i}, we have

$$\begin{aligned}\hat{u}_{-i,a}^T\theta^*_{z_i^*} &= \frac{1}{|S|}\sum_{j\in S}\hat{u}_{-i,a}^T(P_{-i})_{\cdot,j} = \frac{1}{|S|}\sum_{j\in S}\hat{u}_{-i,a}^T(X_{-i})_{\cdot,j} + \frac{1}{|S|}\sum_{j\in S}\hat{u}_{-i,a}^T(E_{-i})_{\cdot,j} \\ &= \frac{1}{|S|}\sum_{j\in S}\hat{\lambda}_{-i,a}(v_{-i,a})_j + \frac{1}{|S|}\hat{u}_{-i,a}^T\left(\sum_{j\in S}(E_{-i})_{\cdot,j}\right).\end{aligned}$$

Hence, by Cauchy-Schwarz inequality and the fact that $\|v_{-i,a}\|=1$, we have

$$\left|\hat{u}_{-i,a}^T\theta^*_{z_i^*}\right| \leq \hat{\lambda}_{-i,a}\frac{\sqrt{|S|}}{|S|} + \frac{\sqrt{|S|}\left\|E_{-i}\right\|}{|S|} \leq \frac{\hat{\lambda}_{-i,a}}{\sqrt{\frac{\beta n}{k}-1}} + \frac{\|E\|}{\sqrt{\frac{\beta n}{k}-1}}. \tag{43}$$

Since $|\hat{u}_{-i,a}^T X_i| \leq |\hat{u}_{-i,a}^T\theta^*_{z_i^*}| + |\hat{u}_{-i,a}^T\epsilon_i|$, we have

$$\begin{aligned}\frac{|\hat{u}_{-i,a}^T X_i|}{\hat{\lambda}_{-i,a}} &\leq \frac{1}{\sqrt{\frac{\beta n}{k}-1}} + \frac{1}{\hat{\lambda}_{-i,a}}\left(\frac{\|E\|}{\sqrt{\frac{\beta n}{k}-1}} + |\hat{u}_{-i,a}^T\epsilon_i|\right) \\ &\leq \frac{1}{\sqrt{\frac{\beta n}{k}-1}} + \frac{1}{\hat{\lambda}_{-i,\kappa}}\frac{\|E\|}{\sqrt{\frac{\beta n}{k}-1}} + \frac{1}{\hat{\lambda}_{-i,\kappa}}|\hat{u}_{-i,a}^T\epsilon_i|.\end{aligned}$$

Consequently,

$$\sqrt{\sum_{a\in\kappa}\left(\frac{\hat{u}_{-i,a}^T X_i}{\hat{\lambda}_{-i,a}}\right)^2} \leq \frac{\sqrt{\kappa}}{\sqrt{\frac{\beta n}{k}-1}} + \frac{1}{\hat{\lambda}_{-i,\kappa}}\frac{\|E\|\sqrt{\kappa}}{\sqrt{\frac{\beta n}{k}-1}} + \frac{1}{\hat{\lambda}_{-i,\kappa}}\left\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\epsilon_i\right\|,$$

where we use the fact $\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\epsilon_i\| = \|\hat{U}_{-i,1:\kappa}^T\epsilon_i\| = (\sum_{i\in[\kappa]}(\hat{u}_{-i,a}^T\epsilon_i)^2)^{1/2}$.

Lastly, by Theorem 2.1, we have

$$\left\|\hat{U}_{1:\kappa}\hat{U}_{1:\kappa}^T - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\right\|_{\mathrm{F}} \leq \frac{4\sqrt{2}}{\rho_{-i}}\left(\frac{\sqrt{\kappa}}{\sqrt{\beta n/k-1}} + \frac{1}{\hat{\lambda}_{-i,\kappa}}\left(\frac{\sqrt{\kappa}\,\|E\|}{\sqrt{\beta n/k-1}} + \left\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\epsilon_i\right\|\right)\right).$$

Since $\beta n/k^2 \geq 10$ and $\rho_0 > 16$ are assumed, we have $\hat{\lambda}_{-i,\kappa} \geq \lambda_\kappa/2$ by (38). Then together with (42), the above display can be simplified into

$$\begin{aligned}\left\|\hat{U}_{1:\kappa}\hat{U}_{1:\kappa}^T - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\right\|_{\mathrm{F}} &\leq \frac{32\sqrt{2}}{\rho_0}\left(\frac{2\sqrt{k\kappa}}{\sqrt{\beta n}} + \frac{2\left\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\epsilon_i\right\|}{\lambda_\kappa}\right)\\ &\leq \frac{128}{\rho_0}\left(\frac{\sqrt{k\kappa}}{\sqrt{\beta n}} + \frac{\left\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\epsilon_i\right\|}{\lambda_\kappa}\right).\end{aligned}$$

This concludes the proof of Theorem 2.2. □

5. Proof of Main Results in Section 3.

In this section, we include proofs of main results in Section 3 except Lemma 3.3, Theorem 3.2, and Theorem 3.3. Their proofs are included in the supplement [42] due to page limit.

5.1. *Proof of Lemma 3.1 and Lemma 3.2.*

PROOF OF LEMMA 3.1. For simplicity, we denote $\hat{U}$ to be short for $\hat{U}_{1:r}$ throughout the proof. From (14), we know $\hat{z}_i$ must satisfy

$$\hat{z}_i = \underset{a\in[k]}{\mathrm{argmin}}\left\|\hat{U}\hat{U}^TX_i - \hat{\theta}_a\right\|,$$

where $\{\hat{\theta}_a\}_{a\in[k]}$ satisfies (17) according to Proposition 3.1. Hence, we have

$$\mathbb{I}\left\{\hat{z}_i \neq \phi(z_i^*)\right\} = \mathbb{I}\left\{\min_{a\in[k]:a\neq\phi(z_i^*)}\left\|\hat{U}\hat{U}^TX_i - \hat{\theta}_a\right\| \leq \left\|\hat{U}\hat{U}^TX_i - \hat{\theta}_{\phi(z_i^*)}\right\|\right\}.$$

Consider a fixed $a \in [k]$ such that $a \neq \phi(z_i^*)$. Note that for any vectors x, y, w of same dimension, if $\|x-y\| \leq \|x-w\|$, then we must have $\|y-w\|/2 \leq \|x-w\|$. Hence, we have

$$\begin{aligned}&\mathbb{I}\left\{\left\|\hat{U}\hat{U}^TX_i - \hat{\theta}_a\right\| \leq \left\|\hat{U}\hat{U}^TX_i - \hat{\theta}_{\phi(z_i^*)}\right\|\right\}\\ &= \mathbb{I}\left\{\frac{1}{2}\left\|\hat{\theta}_{\phi(z_i^*)} - \hat{\theta}_a\right\| \leq \left\|\hat{U}\hat{U}^TX_i - \hat{\theta}_{\phi(z_i^*)}\right\|\right\}\\ &\leq \mathbb{I}\left\{\frac{1}{2}\left\|\hat{\theta}_{\phi(z_i^*)} - \hat{\theta}_a\right\| \leq \left\|\hat{U}\hat{U}^T\epsilon_i - \hat{\theta}_{\phi(z_i^*)}\right\| + \left\|\hat{U}\hat{U}^T\theta_{z_i^*}^* - \hat{\theta}_{\phi(z_i^*)}\right\|\right\}\\ &\leq \mathbb{I}\left\{\left\|\hat{\theta}_{\phi(z_i^*)} - \hat{\theta}_a\right\| - 2\left\|\theta_{z_i^*}^* - \hat{\theta}_{\phi(z_i^*)}\right\| \leq 2\left\|\hat{U}\hat{U}^T\epsilon_i - \hat{\theta}_{\phi(z_i^*)}\right\|\right\},\end{aligned}$$

where we use the fact that $X_i = \theta_{z_i^*}^* + \epsilon_i$ and $\|\hat{U}\hat{U}^T\theta_{z_i^*}^* - \hat{\theta}_{\phi(z_i^*)}\| \leq \|\theta_{z_i^*}^* - \hat{\theta}_{\phi(z_i^*)}\|$. Since $\hat{\theta}_{\phi(z_i^*)} - \hat{\theta}_a = \hat{\theta}_{\phi(z_i^*)} - \theta_{z_i^*}^* + \theta_{z_i^*}^* - \theta_{\phi^{-1}(a)}^* + \theta_{\phi^{-1}(a)}^* - \hat{\theta}_a$, we have

$$\mathbb{I}\left\{\left\|\hat{U}\hat{U}^TX_i - \hat{\theta}_a\right\| \leq \left\|\hat{U}\hat{U}^TX_i - \hat{\theta}_{\phi(z_i^*)}\right\|\right\}$$

$$\begin{aligned}
&\leq \mathbb{I}\Big\{ \left\|\theta^*_{z_i^*} - \theta^*_{\phi^{-1}(a)}\right\| - \left\|\hat{\theta}_{\phi(z_i^*)} - \theta^*_{z_i^*}\right\| - \left\|\theta^*_{\phi^{-1}(a)} - \hat{\theta}_a\right\| \\
&\qquad - 2\left\|\theta^*_{z_i^*} - \hat{\theta}_{\phi(z_i^*)}\right\| \leq 2\left\|\hat{U}\hat{U}^T\epsilon_i\right\| \Big\} \\
&\leq \mathbb{I}\left\{ \left\|\theta^*_{z_i^*} - \theta^*_{\phi^{-1}(a)}\right\| - 4\max_{b\in[k]}\left\|\theta^*_b - \hat{\theta}_{\phi(b)}\right\| \leq 2\left\|\hat{U}\hat{U}^T\epsilon_i\right\| \right\} \\
&\leq \mathbb{I}\left\{ \left(1 - \frac{4C_0\beta^{-0.5}kn^{-0.5}\|E\|}{\Delta}\right)\Delta \leq 2\left\|\hat{U}\hat{U}^T\epsilon_i\right\| \right\}, \qquad (44)
\end{aligned}$$

where in the last inequality, we use the fact that $\max_{b\in[k]}\|\theta^*_b - \hat{\theta}_{\phi(b)}\| \leq C_0\beta^{-0.5}kn^{-0.5}\|E\|$ from Proposition 3.1 and $\min_{b,b'\in[k]:b\neq b'}\|\theta^*_b - \theta^*_{b'}\| = \Delta$. Since the above display holds for each $a \in [k]$ that is not $\phi(z_i^*)$, we have

$$\begin{aligned}
\mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\} &\leq \mathbb{I}\left\{ \left(1 - \frac{4C_0\beta^{-0.5}kn^{-0.5}\|E\|}{\Delta}\right)\Delta \leq 2\left\|\hat{U}\hat{U}^T\epsilon_i\right\| \right\} \\
&= \mathbb{I}\left\{ \left(1 - 4C_0\psi_0^{-1}\right)\Delta \leq 2\left\|\hat{U}\hat{U}^T\epsilon_i\right\| \right\},
\end{aligned}$$

where in the last inequality we use the definition of ψ_0 in (15). □

PROOF OF LEMMA 3.2. For simplicity, throughout the proof we denote $\hat{U}$ and $\hat{U}_{-i}$ to be short for $\hat{U}_{1:\kappa}$ and $\hat{U}_{-i,1:\kappa}$, respectively. We have the following decomposition for $\hat{U}\hat{U}^T\epsilon_i$,

$$\left\|\hat{U}\hat{U}^T\epsilon_i\right\| \leq \left\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\right\| + \left\|\hat{U}\hat{U}^T - \hat{U}_{-i}\hat{U}_{-i}^T\right\|_{\mathrm{F}}\|\epsilon_i\|.$$

Using the fact that $\|\epsilon_i\| \leq \|E\|$ and Theorem 2.2, after rearrangement, we have

$$\begin{aligned}
\left\|\hat{U}\hat{U}^T\epsilon_i\right\| &\leq \frac{128k\|E\|}{\sqrt{n\beta}\rho_0} + \left(1 + \frac{128\|E\|}{\rho_0\lambda_k}\right)\left\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\right\| \\
&= 128\psi_0^{-1}\rho_0^{-1}\Delta + \left(1 + \frac{128}{\rho_0^2}\right)\left\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\right\|.
\end{aligned}$$

In Lemma 3.1 we establish (18). From there we have

$$\begin{aligned}
\mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\} &\leq \mathbb{I}\left\{ \left(1 - C\psi_0^{-1}\right)\Delta \leq 256\psi_0^{-1}\rho_0^{-1}\Delta + 2\left(1 + \frac{128}{\rho_0^2}\right)\left\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\right\| \right\} \\
&\leq \mathbb{I}\left\{ \left(1 - C'\left(\psi_0^{-1} + \rho_0^{-2}\right)\right)\Delta \leq 2\left\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\right\| \right\},
\end{aligned}$$

for some constant $C' > 0$, where in the last inequality we use the assumption $\rho_0 > 16$ from (8). The upper bound on $\mathbb{E}\ell(\hat{z}, z^*)$ is an immediate consequence as $\mathbb{E}\ell(\hat{z}, z^*) = n^{-1}\sum_{i\in[n]}\mathbb{E}\mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\}$. □

5.2. *Proofs of Theorem 3.1.*

PROOF OF THEOREM 3.1. For simplicity, we denote $\hat{U}_{-i}$ to be short for $\hat{U}_{-i,1:\kappa}$ throughout the proof. Define $\psi := \psi_1^{-1} + \rho_1^{-2}$. Then $\psi < 2/C$.

Since E is a random matrix with independent sub-Gaussian columns, we have

$$\mathbb{P}\left(\|E\| \leq 8\sigma(\sqrt{n} + \sqrt{p})\right) \geq 1 - e^{-n/2}, \qquad (45)$$

by Lemma D.1. Denote $\mathcal{F}$ to be this event. Under $\mathcal{F}$, as long as $\psi_1, \rho_1 \geq 128$, we have both (15) and (8) hold. Let $\phi \in \Phi$ satisfy $\ell(\hat{z}, z^*) = n^{-1}\sum_{i\in[n]} \mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\}$. Consider a fixed $i \in [n]$. Then from Lemma 3.2, we have

$$\begin{aligned}\mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\}\mathbb{I}\{\mathcal{F}\} &\leq \mathbb{I}\left\{(1-C_1\psi)\,\Delta \leq 2\left\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\right\|\right\}\mathbb{I}\{\mathcal{F}\} \\ &\leq \mathbb{I}\left\{(1-C_1\psi)\,\Delta \leq 2\left\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\right\|\right\},\end{aligned}$$

where $C_1 > 0$ is some constant that does not depend on C. Then,

$$\begin{aligned}\mathbb{E}\ell(\hat{z}, z^*) &\leq \mathbb{E}\mathbb{I}\left\{\mathcal{F}^{\complement}\right\} + \mathbb{E}\ell(\hat{z}, z^*)\mathbb{I}\{\mathcal{F}\} \\ &\leq e^{-n/2} + n^{-1}\sum_{i\in[n]}\mathbb{E}\mathbb{I}\left\{(1-C_1\psi)\,\Delta \leq 2\left\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\right\|\right\}. \end{aligned} \tag{46}$$

Since $\epsilon_i \sim \mathrm{SG}_p(\sigma^2)$ and it is independent of $\hat{U}_{-i}\hat{U}_{-i}^T$, we can apply concentration inequalities for $\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\|$ from Lemma D.2. Define $t = (1-C_2\psi)\Delta^2/(8\sigma^2)$ where $C_2 = C_1 + 16$. Since C_2 does not depend on C, we can let $C > \max\{4C_2, 128\}$ such that $1 - C_2\psi > 1/2$. Then we have $k/t \leq 16k^2\sigma^2/\Delta^2 \leq 16\psi_1^2$ where we use the fact that $\frac{\Delta}{k\sigma} > \psi_1^{-1}$ from (20) as $\beta \leq 1$. Then we have

$$\begin{aligned}\sigma^2(\kappa + 2\sqrt{\kappa t} + 2t) = 2\sigma^2 t\left(\frac{1}{2}\frac{\kappa}{t} + \sqrt{\frac{\kappa}{t}} + 1\right) &\leq 2\sigma^2 t\left(8\psi_1^2 + 4\psi_1 + 1\right) \leq 2\sigma^2 t\,(1 + 8\psi_1) \\ &\leq (1-C_2\psi)\Delta^2/(8\sigma^2)\,(1+8\psi) \leq (1-C_1\psi)\Delta^2/(8\sigma^2),\end{aligned}$$

where we use that $\psi_1 < 1/128$ and $\psi < 1/64$ as we let $C > 128$. Then from Lemma D.2, we have

$$\mathbb{E}\mathbb{I}\left\{(1-C_1\psi)\,\Delta \leq 2\left\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\right\|\right\} \leq \exp(-t) = \exp\left(-(1-C_2\psi)\frac{\Delta^2}{8\sigma^2}\right).$$

□

5.3. *Proof of Theorem 3.4.* The proof of Theorem 3.4 relies on the following entrywise decomposition that is analogous to Lemma 3.2 but in an opposite direction. Note the the singular vectors $\hat{u}_1$, and $\{\hat{u}_{1,-i}\}_{i\in[n]}$ are all identifiable up to sign. Without loss of generality, we assume $\langle \hat{u}_1, u_1\rangle \geq 0$ and $\langle \hat{u}_{1,-i}, u_1\rangle \geq 0$ for all $i \in [n]$.

LEMMA 5.1. *Consider the model (27). Let $\phi \in \Phi$ be the permutation such that $\ell(\check{z}, z^*) = \frac{1}{n}|\{i \in [n] : \check{z}_i \neq \phi(z_i^*)\}|$. Then there exists a constants $C, C_1 > 0$ such that if*

$$\frac{\Delta}{\beta^{-0.5}n^{-0.5}\|E\|} \geq C, \tag{47}$$

then for any $i \in [n]$,

$$\mathbb{I}\{\check{z}_i \neq \phi(z_i^*)\} \geq \mathbb{I}\left\{\left(1 + \frac{C_1\beta^{-0.5}n^{-0.5}\|E\|}{\Delta}\right)\Delta \leq -2(\hat{u}_{1,-i}^T\epsilon_i)\mathit{sign}(u_1^T\theta_{\phi(z_i^*)})\right\}. \tag{48}$$

PROOF. The proof mainly follows the proofs of Lemma 3.1 and Lemma 3.2 with some modifications such as adding a negative term instead of a positive term in order to obtain a lower bound.

We first write $\check{z}$ equivalently as

$$\left(\check{z}, \{\check{\theta}_j\}_{j=1}^2\right) = \operatorname*{argmin}_{z\in[2]^n, \{\theta_j\}_{j=1}^2 \in \mathbb{R}^p} \sum_{i\in[n]} \left\|\hat{u}_1\hat{u}_1^T X_i - \theta_{z_i}\right\|^2,$$

where $\check{\theta}_a = \hat{u}_1 \check{c}_a$ for each $a\in[2]$. Note that $k=2$. From Proposition 3.1, we have

$$\frac{1}{n}\left|\{i\in[n] : \check{z}_i \neq \phi(z_i^*)\}\right| \leq \frac{C_0 k \|E\|^2}{n\Delta^2},$$

and

$$\max_{a\in[2]} \left\|\check{\theta}_{\phi(a)} - \theta_a^*\right\| \leq C_0 \beta^{-0.5} k n^{-0.5} \|E\|, \tag{49}$$

for some permutation $\phi : [2]\to[2]$ and some constant $C_0 > 0$. Without loss of generality, assume $\phi = \mathrm{Id}$.

Recall that $\theta_1^* = -\theta_2^* = \delta \mathbb{1}_p$, $u_1 = 1/\sqrt{p}\mathbb{1}_p$, $\lambda_1 = \delta\sqrt{np} = \frac{\Delta\sqrt{n}}{2}$, and $|u_1^T(\theta^*_{z_i^*} - (-\theta^*_{z_i^*}))| = 2\delta\sqrt{p} = \Delta$. By Davis-Kahan Theorem, we have

$$\min_{s\in\pm 1} \|\hat{u}_1 - s u_1\| \leq \frac{\|E\|}{\lambda_1} = \frac{2\|E\|}{\sqrt{n}\Delta} \leq 1/16,$$

where the last inequality is due to the assumption (15). Since we assume $\langle \hat{u}_1, u_1\rangle \geq 0$, we have $\|\hat{u}_1 - s u_1\| = \min_{s\in\pm1}\|\hat{u}_1 - s u_1\|$.

Consider any $i\in[n]$ and any $a\in[2]$ such that $a\neq z_i^*$. Note that for any scalars x, y, w, if $|x-y| \leq |x-w|$, we have equivalently $\mathrm{sign}(w-y)(y+w)/2 \geq \mathrm{sign}(w-y)x$. Since $(y+w)/2 = (y-w)/2 + w$, a sufficient condition is $|w-y|/2 + |w| \leq (-\mathrm{sign}(w-y))x$. Hence, we have

$$\begin{aligned}
&\mathbb{I}\left\{\left\|\hat{u}_1\hat{u}_1^T X_i - \check{\theta}_a\right\| \leq \left\|\hat{u}_1\hat{u}_1^T X_i - \check{\theta}_{z_i^*}\right\|\right\} \\
&= \mathbb{I}\left\{\left|\hat{u}_1^T X_i - \hat{u}_1^T\check{\theta}_a\right| \leq \left|\hat{u}_1^T X_i - \hat{u}_1^T\check{\theta}_{z_i^*}\right|\right\} \\
&= \mathbb{I}\left\{\left|\hat{u}_1^T\epsilon_i - \hat{u}_1^T\left(\check{\theta}_a - \theta^*_{z_i^*}\right)\right| \leq \left|\hat{u}_1^T\epsilon_i - \hat{u}_1^T\left(\check{\theta}_{z_i^*} - \theta^*_{z_i^*}\right)\right|\right\} \\
&\geq \mathbb{I}\left\{\frac{1}{2}\left|\hat{u}_1^T(\check{\theta}_{z_i^*} - \check{\theta}_a)\right| + \left|\hat{u}_1^T\left(\check{\theta}_{z_i^*} - \theta^*_{z_i^*}\right)\right| \leq -(\hat{u}_1^T\epsilon_i)\mathrm{sign}(\hat{u}_1^T(\check{\theta}_{z_i^*} - \check{\theta}_a))\right\} \\
&\geq \mathbb{I}\left\{\left\|\check{\theta}_{z_i^*} - \check{\theta}_a\right\| + 2\left\|\check{\theta}_{z_i^*} - \theta^*_{z_i^*}\right\| \leq -2(\hat{u}_1^T\epsilon_i)\mathrm{sign}(\hat{u}_1^T(\check{\theta}_{z_i^*} - \check{\theta}_a))\right\}.
\end{aligned}$$

We are going to show $\mathrm{sign}(\hat{u}_1^T(\check{\theta}_{z_i^*} - \check{\theta}_a)) = \mathrm{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a))$. By (49), we have

$$\begin{aligned}
\left\langle \check{\theta}_{z_i^*} - \check{\theta}_a, \theta^*_{z_i^*} - \theta^*_a\right\rangle &= \left\|\theta^*_{z_i^*} - \theta^*_a\right\|^2 + \left\langle \check{\theta}_{z_i^*} - \theta^*_{z_i^*}, \theta^*_{z_i^*} - \theta^*_a\right\rangle + \left\langle \check{\theta}_a - \theta^*_a, \theta^*_{z_i^*} - \theta^*_a\right\rangle \\
&\geq \Delta^2\left(1 - \frac{2C_0 k\beta^{-0.5} n^{-0.5}\|E\|}{\Delta}\right) \\
&> 0,
\end{aligned}$$

where the last inequality holds as long as $\Delta > 2C_0\beta^{-0.5} k n^{-0.5}\|E\|$. Due to the fact $\theta^*_{z_i^*} - \theta^*_a \in \mathrm{span}(u_1)$, $\check{\theta}_{z_i^*} - \check{\theta}^*_a \in \mathrm{span}(\hat{u}_1)$, and $\langle \hat{u}_1, u_1\rangle \geq 0$, if $u_1, \theta^*_{z_i^*} - \theta^*_a$ are in the same direction, then $\hat{u}_1, \check{\theta}_{z_i^*} - \check{\theta}^*_a$ must also be in the same direction, and vice versa. Hence, we have $\mathrm{sign}(\hat{u}_1^T(\check{\theta}_{z_i^*} - \check{\theta}_a)) = \mathrm{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a))$. Thus,

$$\begin{aligned}
&\mathbb{I}\left\{\left\|\hat{u}_1\hat{u}_1^T X_i - \check{\theta}_a\right\| \leq \left\|\hat{u}_1\hat{u}_1^T X_i - \check{\theta}_{z_i^*}\right\|\right\} \\
&\geq \mathbb{I}\left\{\left\|\check{\theta}_{z_i^*} - \check{\theta}_a\right\| + 2\left\|\check{\theta}_{z_i^*} - \theta^*_{z_i^*}\right\| \leq -2(\hat{u}_1^T\epsilon_i)\mathrm{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a))\right\}.
\end{aligned}$$

Following the same analysis as in the proof of Lemma 3.1, we can get the following result that is analogous to (44):

$$\mathbb{I}\left\{\left\|\hat{u}_1\hat{u}_1^T X_i - \check{\theta}_a\right\| \leq \left\|\hat{u}_1\hat{u}_1^T X_i - \check{\theta}_{z_i^*}\right\|\right\}$$
$$\geq \mathbb{I}\left\{\left(1 + \frac{4C_0\beta^{-0.5}kn^{-0.5}\left\|E\right\|}{\Delta}\right)\Delta \leq -2(\hat{u}_1^T\epsilon_i)\text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a))\right\}.$$

Next, we are going to decompose $\hat{u}_1^T\epsilon_i$ following the proof of Lemma 3.2. Denote $\hat{u}_{1,-i}$ be the leave-one-out counterpart of $\hat{u}_1$, i.e., $\hat{u}_{1,-i}$ is the leading left singular vector of X_{-i}. Since we assume $\langle \hat{u}_{1,-i}, u_1\rangle \geq 0$, we have $\|\hat{u}_{1,-i} - u_1\| \leq 2\left\|E\right\|/(\sqrt{n-1}\Delta)$. As a result, we have $\|\hat{u}_{1,-i} - \hat{u}_1\| \leq 4\left\|E\right\|/(\sqrt{n-1}\Delta)$ which leads to

$$\langle \hat{u}_{1,-i}, \hat{u}_1\rangle \geq 1 - 4\left\|E\right\|/(\sqrt{n-1}\Delta) > 0. \tag{50}$$

We have the following decomposition:

$$\begin{aligned}
&(\hat{u}_1^T\epsilon_i)\text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a)) \\
&= \left\langle \hat{u}_1, \hat{u}_1\hat{u}_1^T\epsilon_i\right\rangle \text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a)) \\
&= \left\langle \hat{u}_1, (\hat{u}_{1,-i}\hat{u}_{1,-i}^T)\epsilon_i\right\rangle \text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a)) + \left\langle \hat{u}_1, (\hat{u}_1\hat{u}_1^T - \hat{u}_{1,-i}\hat{u}_{1,-i}^T)\epsilon_i\right\rangle \text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a)) \\
&= \langle \hat{u}_1, \hat{u}_{1,-i}\rangle\,(\hat{u}_{1,-i}^T\epsilon_i)\text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a)) + \left\langle \hat{u}_1, (\hat{u}_1\hat{u}_1^T - \hat{u}_{1,-i}\hat{u}_{1,-i}^T)\epsilon_i\right\rangle \text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a)) \\
&\leq \langle \hat{u}_1, \hat{u}_{1,-i}\rangle\,(\hat{u}_{1,-i}^T\epsilon_i)\text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a)) + \left\|\hat{u}_1\hat{u}_1^T - \hat{u}_{1,-i}\hat{u}_{1,-i}^T\right\|\left\|\epsilon_i\right\|.
\end{aligned}$$

Note that $\lambda_1/\left\|E\right\| = \Delta\sqrt{n}/(2\left\|E\right\|)$ is greater than 16 under the assumption (47) holds for a large constant C. From Theorem 2.2 we have

$$\left\|\hat{u}_1\hat{u}_1^T - \hat{u}_{1,-i}\hat{u}_{1,-i}^T\right\| \leq \frac{128}{\lambda_1/\left\|E\right\|}\left(\frac{k}{\sqrt{\beta n}} + \frac{\left\|\hat{u}_{1,-i}\hat{u}_{1,-i}^T\epsilon_i\right\|}{\lambda_1}\right).$$

Then,

$$\begin{aligned}
&(\hat{u}_1^T\epsilon_i)\text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a)) \\
&\leq \langle \hat{u}_1, \hat{u}_{1,-i}\rangle\,(\hat{u}_{1,-i}^T\epsilon_i)\text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a)) + \left(\frac{128k}{\sqrt{n\beta}(\lambda_1/\left\|E\right\|)} + \frac{128\left\|\hat{u}_{1,-i}\hat{u}_{1,-i}^T\epsilon_i\right\|}{\lambda_1^2/\left\|E\right\|}\right)\left\|E\right\| \\
&= \langle \hat{u}_1, \hat{u}_{1,-i}\rangle\,(\hat{u}_{1,-i}^T\epsilon_i)\text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a)) + \frac{256n^{-0.5}k\beta^{-0.5}\left\|E\right\|^2}{\Delta} + \frac{512\left|\hat{u}_{1,-i}^T\epsilon_i\right|n^{-1}\left\|E\right\|^2}{\Delta^2}.
\end{aligned}$$

So far we have obtained

$$\begin{aligned}
&\mathbb{I}\left\{\left\|\hat{u}_1\hat{u}_1^T X_i - \check{\theta}_a\right\| \leq \left\|\hat{u}_1\hat{u}_1^T X_i - \check{\theta}_{z_i^*}\right\|\right\} \\
&\geq \mathbb{I}\Bigg\{\left(1 + \frac{4C_0\beta^{-0.5}kn^{-0.5}\left\|E\right\|}{\Delta}\right)\Delta \leq -2\langle \hat{u}_1, \hat{u}_{1,-i}\rangle\,(\hat{u}_{1,-i}^T\epsilon_i)\text{sign}(u_1^T(\theta^*_{z_i^*} - \theta^*_a)) \\
&\qquad - \frac{256n^{-0.5}k\beta^{-0.5}\left\|E\right\|^2}{\Delta} - \frac{512\left|\hat{u}_{1,-i}^T\epsilon_i\right|n^{-1}\left\|E\right\|^2}{\Delta^2}\Bigg\} \\
&= \mathbb{I}\Bigg\{\left(1 + \frac{4C_0\beta^{-0.5}kn^{-0.5}\left\|E\right\|}{\Delta} + \frac{256n^{-0.5}k\beta^{-0.5}\left\|E\right\|^2}{\Delta^2}\right)\Delta
\end{aligned}$$

$$\leq -2\left\langle \hat{u}_1, \hat{u}_{1,-i}\right\rangle (\hat{u}_{1,-i}^T\epsilon_i)\text{sign}(u_1^T(\theta^*_{z_i^*}-\theta^*_a)) - \frac{512\left|\hat{u}_{1,-i}^T\epsilon_i\right| n^{-1}\left\|E\right\|^2}{\Delta^2}\Bigg\}.$$

From (50) we have

$$\begin{aligned}\left\langle \hat{u}_{1,-i}, \hat{u}_1\right\rangle - \frac{512n^{-1}\left\|E\right\|^2}{\Delta^2} &\geq 1 - 4\frac{\left\|E\right\|(n-1)^{-0.5}}{\Delta} - \frac{512n^{-1}\left\|E\right\|^2}{\Delta^2}\\ &\geq 1-\frac{16n^{-0.5}\left\|E\right\|}{\Delta} \geq \frac{1}{2},\end{aligned}$$

assuming $\frac{\Delta}{n^{-0.5}\|E\|} \geq 64$. For any $x,y,z,w\in\mathbb{R}$ such that $x\geq 0$, $1\geq z\geq 0$, and $z\left|y\right| > w \geq 0$, we have $\mathbb{I}\left\{x\leq zy-w\right\} \geq \mathbb{I}\left\{x \leq (z-w/|y|)\,y\right\}$. We then have,

$$\begin{aligned}&\mathbb{I}\left\{\left\|\hat{u}_1\hat{u}_1^TX_i-\check{\theta}_a\right\| \leq \left\|\hat{u}_1\hat{u}_1^TX_i-\check{\theta}_{z_i^*}\right\|\right\}\\ &\geq \mathbb{I}\Bigg(\left(1+\frac{4C_0\beta^{-0.5}kn^{-0.5}\left\|E\right\|}{\Delta}+\frac{256n^{-0.5}k\beta^{-0.5}\left\|E\right\|^2}{\Delta^2}\right)\Delta\\ &\qquad \leq -2\left(1-\frac{16n^{-0.5}\left\|E\right\|}{\Delta}\right)(\hat{u}_{1,-i}^T\epsilon_i)\text{sign}(u_1^T(\theta^*_{z_i^*}-\theta^*_a))\Bigg)\\ &\geq \mathbb{I}\left\{\left(1+\frac{C_1\beta^{-0.5}n^{-0.5}\left\|E\right\|}{\Delta}\right)\Delta \leq -2(\hat{u}_{1,-i}^T\epsilon_i)\text{sign}(u_1^T(\theta^*_{z_i^*}-\theta^*_a))\right\}.\end{aligned}$$

Since $\theta^*_a = -\theta^*_{z_i^*}$, we have $\text{sign}(u_1^T(\theta^*_{z_i^*}-\theta^*_a)) = \text{sign}(u_1^T\theta^*_{z_i^*})$. The proof is complete. $\square$

PROOF OF THEOREM 3.4. Recall that $\lambda_1 = \Delta\sqrt{n}/2$. Same as the proof of Theorem 3.1, we work on the with-high-probability event (45).

For the upper bound, from Lemma 3.2, there exists some $\phi\in\Phi$ such that for any $i\in[n]$,

$$\mathbb{I}\left\{\hat{z}_i\neq\phi(z_i^*)\right\} \leq \mathbb{I}\left\{\left(1-C_1\psi_3^{-1}\right)\Delta \leq 2\left\|\hat{u}_{1,-i}\hat{u}_{-i}^T\epsilon_i\right\|\right\} = \mathbb{I}\left\{\left(1-C_1\psi_3^{-1}\right)\Delta\leq 2\left|\hat{u}_{1,-i}^T\epsilon_i\right|\right\},$$

for some $C_1>0$, where the last inequality is due to that ψ_3 is large. By Davis-Kahan Theorem, we know there exists some $s_i\in\{-1,1\}$ such that $\left\|\hat{u}_{1,-i}-s_iu_1\right\| \leq 2\left\|E\right\|/(\sqrt{n-1}\Delta) \leq 4\psi_3^{-1}$. Since $\left\langle \hat{u}_{1,-i}, u_1\right\rangle \geq 0$ is assumed, we have $s_i=1$ for all $i\in[n]$. Then

$$\begin{aligned}\mathbb{I}\left\{\hat{z}_i\neq\phi(z_i^*)\right\} &\leq \mathbb{I}\left\{\left(1-C_1\psi_3^{-1}\right)\Delta\leq 2\left|u_1^T\epsilon_i\right| + 2\left|\left(\hat{u}_{1,-i}-s_iu_1\right)^T\epsilon_i\right|\right\}\\ &\leq \mathbb{I}\left\{\left(1-(C_1+C_2)\psi_3^{-1}\right)\Delta\leq 2\left|u_1^T\epsilon_i\right|\right\} + \mathbb{I}\left\{C_2\psi_3^{-1}\Delta\leq 2\left|\left(\hat{u}_{1,-i}-s_iu_1\right)^T\epsilon_i\right|\right\},\end{aligned}$$

where $C_2>0$ is a constant whose value will be determined later. Due to the independence of $\hat{u}_{1,-i}-s_iu_1$ and ϵ_i, we have $\left(\hat{u}_{1,-i}-s_iu_1\right)^T\epsilon_i \sim \text{SG}(16\psi_3^{-2}\sigma^2)$ and then

$$\mathbb{E}\mathbb{I}\left\{C_2\Delta\leq 2\left|\left(\hat{u}_{1,-i}-s_iu_1\right)^T\epsilon_i\right|\right\} \leq 2\exp\left(-\frac{C_2^2\Delta^2}{128\sigma^2}\right).$$

On the other hand, $u_1^T\epsilon_i = p^{-\frac{1}{2}}\sum_{j=1}^p\epsilon_{i,j}$ where $\{\epsilon_{i,j}\}_{j\in[p]}$ are i.i.d. with variance $\bar{\sigma}^2$, which can be approximated by a normal distribution. Since the distribution F is sub-Gaussian, its moment generating function exists. Then we can use the following KMT quantile inequality (cf., Proposition [KMT] of [28]). Let $Y \stackrel{d}{=} \bar{\sigma}^{-1}p^{-\frac{1}{2}}\sum_{j=1}^p\epsilon_{i,j}$. There exist some constants $D,\eta>0$ and $Z\sim\mathcal{N}(0,1)$, such that whenever $|Y|\leq\eta\sqrt{p}$, we have

$$|Y-Z|\leq\frac{DY^2}{\sqrt{p}}+\frac{D}{\sqrt{p}}.$$

Then,

$$\mathbb{E}\mathbb{I}\left\{\left(1-(C_1+C_2)\psi_3^{-1}\right)\Delta \leq 2\left|u_1^T\epsilon_i\right|\right\}$$
$$=\mathbb{E}\mathbb{I}\left\{\left(1-(C_1+C_2)\psi_3^{-1}\right)\frac{\Delta}{\bar\sigma}\leq 2\left|Y\right|\right\}$$
$$\leq \mathbb{E}\mathbb{I}\left\{\left(1-(C_1+C_2)\psi_3^{-1}\right)\frac{\Delta}{\bar\sigma}\leq 2\left|Z\right|+\frac{2DY^2}{\sqrt{p}}+\frac{2D}{\sqrt{p}}\right\}+\mathbb{E}\mathbb{I}\left\{|Y|>\eta\sqrt{p}\right\}$$
$$\leq \mathbb{E}\mathbb{I}\left\{\left(1-(C_1+C_2+C_3+2D)\psi_3^{-1}\right)\frac{\Delta}{\bar\sigma}\leq 2\left|Z\right|\right\}+\mathbb{E}\mathbb{I}\left\{\frac{2DY^2}{\sqrt{p}}\geq C_3\right\}+\mathbb{E}\mathbb{I}\left\{|Y|>\eta\sqrt{p}\right\},$$

where $C_3>0$ is a constant. Using the fact that $Y\sim \mathrm{SG}(1)$ with zero mean, we have

$$\mathbb{E}\mathbb{I}\left\{\left(1-(C_1+C_2)\psi_3^{-1}\right)\Delta \leq 2\left|u_1^T\epsilon_i\right|\right\}$$
$$\leq 2\exp\left(-\frac{\left(1-(C_1+C_2+C_3+2D)\psi_3^{-1}\right)^2\Delta^2}{8\bar\sigma^2}\right)+2\exp\left(-\frac{C_3\sqrt{p}}{4D}\right)+2\exp\left(-\frac{\eta^2 p}{2}\right).$$

Then we have

$$\mathbb{E}\ell(\check z,z^*)$$
$$\leq \frac{1}{n}\sum_{i=1}^n\mathbb{E}\mathbb{I}\left\{\left(1-(C_1+C_2)\psi_3^{-1}\right)\Delta \leq 2\left|u_1^T\epsilon_i\right|\right\}+\frac{1}{n}\sum_{i=1}^n\mathbb{E}\mathbb{I}\left\{C_2\Delta\leq 2\left|\left(\hat u_{1,-i}-s_iu_1\right)^T\epsilon_i\right|\right\}+e^{-0.5n}$$
$$\leq 2\exp\left(-\frac{\left(1-(C_1+C_2+C_3+2D)\psi_3^{-1}\right)^2\Delta^2}{8\bar\sigma^2}\right)$$
$$+2\exp\left(-\frac{C_2^2\Delta^2}{128\sigma^2}\right)+2\exp\left(-\frac{C_3\sqrt{p}}{4D}\right)+2\exp\left(-\frac{\eta^2p}{2}\right)+e^{-0.5n},$$

where $e^{-0.5n}$ is the probability that (45) does not hold. Since $\sigma\leq C\bar\sigma$, when C_2 is chosen to satisfy $C_2^2/(128C^2)\geq 16$, we have

$$\mathbb{E}\ell(\check z,z^*)\leq 2\exp\left(-\frac{\left(1-C''\psi_3^{-1}\right)^2\Delta^2}{8\bar\sigma^2}\right)+\exp\left(-C''\sqrt{p}\right)+e^{-0.5n},$$

for some constant $C''>0$.

For the lower bound, from (48) we know

$$\mathbb{I}\left\{\check z_i\neq\phi(z_i^*)\right\}\geq\mathbb{I}\left\{\left(1+C_4\psi_3^{-1}\right)\Delta\leq -2(\hat u_{1,-i}^T\epsilon_i)\mathrm{sign}(u_1^T(\theta_{\phi(z_i^*)}-\theta_{3-\phi(z_i^*)}))\right\},$$

for some constant $C_4>0$ assuming ψ_3 is large. Using the same argument as in the upper bound, we are going to decompose $\hat u_{1,-i}^T\epsilon_i$ into $u_1^T\epsilon_i$ and $(\hat u_{1,-i}-y_1)^T\epsilon_i$. Hence,

$$\begin{aligned}\mathbb{I}\left\{\check z_i\neq\phi(z_i^*)\right\}&\geq\mathbb{I}\left\{\left(1+C_4\psi_3^{-1}\right)\Delta\leq -2(u_1^T\epsilon_i)\mathrm{sign}(u_1^T(\theta_{\phi(z_i^*)}-\theta_{3-\phi(z_i^*)}))-2\left|(\hat u_{1,-i}-s_iu_1)^T\epsilon_i\right|\right\}\\&\geq\mathbb{I}\left\{\left(1+\left(C_4+C_5\right)\psi_3^{-1}\right)\Delta\leq -2(u_1^T\epsilon_i)\mathrm{sign}(u_1^T(\theta_{\phi(z_i^*)}-\theta_{3-\phi(z_i^*)}))\right\}\\&\quad-\mathbb{I}\left\{C_5\psi_3^{-1}\Delta\leq 2\left|(\hat u_{1,-i}-s_iu_1)^T\epsilon_i\right|\right\},\end{aligned}$$

for some constant $C_5>0$ whose value to be chosen. Let

$$Y'\overset{d}{=}\bar\sigma^{-1}(u_1^T\epsilon_i)\mathrm{sign}(u_1^T(\theta_{\phi(z_i^*)}-\theta_{3-\phi(z_i^*)}))=\mathrm{sign}(u_1^T(\theta_{\phi(z_i^*)}-\theta_{3-\phi(z_i^*)}))\bar\sigma^{-1}p^{-\frac{1}{2}}\sum_{j=1}^p\epsilon_{i,j}.$$

Then using the same argument above, there exists some $Z' \sim \mathcal{N}(0,1)$ such that whenever $Y' \leq \eta'\sqrt{p}$, we have $|Y'-Z'| \leq \frac{D'Y'^2}{\sqrt{p}} + \frac{D'}{\sqrt{p}}$ where $D', \eta' > 0$ are constants. Then

$$
\begin{aligned}
&\mathbb{E}\mathbb{I}\left\{\left(1+(C_4+C_5)\psi_3^{-1}\right)\Delta \leq -2(u_1^T\epsilon_i)\text{sign}(u_1^T(\theta_{\phi(z_i^*)}-\theta_{3-\phi(z_i^*)}))\right\} \\
&= \mathbb{E}\mathbb{I}\left\{\left(1+(C_4+C_5)\psi_3^{-1}\right)\frac{\Delta}{\bar{\sigma}} \leq -2Y'\right\} \\
&\geq \mathbb{E}\mathbb{I}\left\{\left(1+(C_4+C_5)\psi_3^{-1}\right)\frac{\Delta}{\bar{\sigma}} \leq -2Z' - \frac{2DY'^2}{\sqrt{p}} - \frac{2d}{\sqrt{p}}\right\}\mathbb{I}\left\{Y' \leq \eta'\sqrt{p}\right\} \\
&\geq \mathbb{E}\mathbb{I}\left\{\left(1+(C_4+C_5+2D+C_6)\psi_3^{-1}\right)\frac{\Delta}{\bar{\sigma}} \leq -2Z'\right\} - \mathbb{E}\mathbb{I}\left\{\frac{2DY'^2}{\sqrt{p}} \geq C_6\right\} - \mathbb{E}\mathbb{I}\left\{Y' > \eta'\sqrt{p}\right\},
\end{aligned}
$$

where $C_6 > 0$ is a constant. Then following the proof of the upper bound, and by a proper choice of C_5, we have

$$
\mathbb{E}\ell(\check{z}, z^*) \geq 2\exp\left(-\frac{\left(1+C'''\psi_3^{-1}\right)^2\Delta^2}{8\bar{\sigma}^2}\right) - \exp\left(-C'''\sqrt{p}\right) - e^{-0.5n},
$$

for some constant $C''' > 0$. ☐

5.4. *Proofs of Lemma 3.4 and Theorem 3.5.*

PROOF OF LEMMA 3.4. For the upper bound, we consider the following likelihood ratio test. For any $x \in \mathbb{R}^p$, define the two log-likelihood functions as

$$
l_1(x) = \sum_{j=1}^p \log f(x_j - \delta), \text{ and } l_2(x) = \sum_{j=1}^p \log f(x_j + \delta).
$$

Then for each $i \in [n]$, define the likelihood ratio test as

$$
\hat{z}_i^{\text{LRT}} = \begin{cases} 1, \text{ if } l_1(X_i) \geq l_2(X_i), \\ 2, \text{ otherwise.} \end{cases}
$$

Then for any $i \in [n]$ such that $z_i^* = 1$, we have

$$
\mathbb{E}\mathbb{I}\left\{\hat{z}_i^{\text{LRT}} = 2\right\} = \mathbb{P}\left(l_2(X_i) > l_1(X_i)\right) = \mathbb{P}\left(\sum_{j=1}^p \log\frac{f(2\delta+\epsilon_{i,j})}{f(\epsilon_{i,j})} > 0\right) = \mathbb{P}\left(\sum_{j=1}^p \log\frac{f_{\frac{\Delta}{\sqrt{p}}}(\epsilon_{i,j})}{f_0(\epsilon_{i,j})} > 0\right),
$$

where we use the fact $2\delta = \frac{\Delta}{\sqrt{p}}$. Since Δ is a constant, by local asymptotic normality (c.f., Chapter 7, [37]), we have

$$
\sum_{j=1}^p \log\frac{f_{\frac{\Delta}{\sqrt{p}}}(\epsilon_{i,j})}{f_0(\epsilon_{i,j})} \xrightarrow{d} \mathcal{N}\left(-\frac{\mathcal{I}\Delta^2}{2}, \mathcal{I}\Delta^2\right).
$$

Then, $\lim_{p\to\infty}\mathbb{E}\mathbb{I}\left\{\hat{z}_i^{\text{LRT}} = 2\right\} \leq C_1\exp\left(-\mathcal{I}\Delta^2/8\right)$ for some constant $C_1 > 0$. We have the same upper bound if $z_i^* = 2$ instead. Hence,

$$
\lim_{p\to\infty}\inf_z\sup_{z^*\in[2]^n}\mathbb{E}\ell(z,z^*) \leq \lim_{p\to\infty}\sup_{z^*\in[2]^n}\mathbb{E}\ell(\hat{z}^{\text{LRT}}, z^*) \leq \exp\left(-\frac{\mathcal{I}\Delta^2}{8}\right).
$$

For the lower bound, instead of allowing $z^* \in [2]^n$, we consider a slightly smaller parameter space. Define $\mathcal{Z} = \{z \in [2]^n : z_i = 1, \forall 1 \leq i \leq n/3, z_i = 2, \forall n/3+1 \leq i \leq 2n/3\}$.

Then for any $z, z' \in \mathcal{Z}$ we have $\ell(z, z') = n^{-1}\sum_{i=1}^n \mathbb{I}\{z_i \neq z_i'\} \leq 1/3$ due to the fact $n^{-1}\sum_{i=1}^n \mathbb{I}\{\phi(z_i) \neq z_i'\} \geq 1/3$ if $\phi \neq \mathrm{Id}$. Hence,

$$\inf_z \sup_{z^* \in [2]^n} \mathbb{E}\ell(z, z^*) \geq \inf_z \sup_{z^* \in \mathcal{Z}} \mathbb{E}\ell(z, z^*) \geq n^{-1} \inf_z \sup_{z^* \in \mathcal{Z}} \mathbb{E} \sum_{i \in [n]} \mathbb{I}\{z_i \neq z_i^*\}$$

$$\geq n^{-1} \sum_{i > 2n/3} \inf_{z_i} \sup_{z_i^* \in [2]} \mathbb{E}\mathbb{I}\{z_i \neq z_i^*\} = \frac{1}{3} \inf_{z_n} \sup_{z_n^* \in [2]} \mathbb{E}\mathbb{I}\{z_n \neq z_n^*\},$$

where it is reduced into a testing problem on whether X_n has mean θ_1^* or θ_2^*. According to the Neyman-Pearson Lemma, the optimal procedure is the likelihood ratio test $\hat{z}_n^{\mathrm{LRT}}$ defined above. By the same argument, we have

$$\liminf_{p \to} \inf_z \sup_{z^* \in [2]^n} \mathbb{E}\ell(z, z^*) \geq \frac{1}{3} \liminf_{p \to} \inf_{z_n} \sup_{z_n^* \in [2]} \mathbb{E}\mathbb{I}\{z_n \neq z_n^*\} \geq C_2 \exp\left(-\frac{\mathcal{I}\Delta^2}{8}\right),$$

for some constant $C_2 > 0$. $\square$

PROOF OF THEOREM 3.5. First, we have the following connection between the Fisher information $\mathcal{I}$ and the variance $\bar{\sigma}^2$:

$$\mathcal{I}\bar{\sigma}^2 = \left(\int \left(\frac{f'}{f}\right)^2 f \mathrm{d}x\right)\left(\int x^2 f \mathrm{d}x\right) \geq \left(\int \frac{f'}{f} x f \mathrm{d}x\right)^2 = \left(\int x f' \mathrm{d}x\right)^2 = 1,$$

where we use Cauchy-Schwarz inequality and the integral by part $\int x f' \mathrm{d}x = \int x f \mathrm{d}x - \int f \mathrm{d}x = 0 - 1 = -1$. The equation holds if and only if $f'/f \propto x$, which is equivalent to F being normally distributed. $\square$

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to "Leave-one-out Singular Subspace Perturbation Analysis for Spectral Clustering"
(url to be specified). In the supplement [42], we first provide the proof of Theorem 2.3 in Appendix A, followed by the proofs of Lemma 3.3 and Theorem 3.2 in Appendix B. The proof of Theorem 3.3 is given in Appendix C. Auxiliary lemmas and propositions and their proofs are included in Appendix D.

REFERENCES

[1] ABBE, E., FAN, J. and WANG, K. (2020). An ℓ_p theory of PCA and spectral clustering. *arXiv preprint arXiv:2006.14062*.

[2] ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics* **48** 1452.

[3] AGTERBERG, J., LUBBERTS, Z. and PRIEBE, C. (2021). Entrywise estimation of singular vectors of low-rank matrices with heteroskedasticity and dependence. *arXiv preprint arXiv:2105.13346*.

[4] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral analysis of large dimensional random matrices* **20**. Springer.

[5] BELABBAS, M.-A. and WOLFE, P. J. (2009). Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences* **106** 369–374.

[6] CAI, C., LI, G., CHI, Y., POOR, H. V. and CHEN, Y. (2021). Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics* **49** 944–967.

[7] CAI, T. T., LI, H. and MA, R. (2021). Optimal Structured Principal Subspace Estimation: Metric Entropy and Minimax Rates. *J. Mach. Learn. Res.* **22** 46–1.

[8] CAI, T. T. and ZHANG, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics* **46** 60–89.

[9] CAPE, J., TANG, M. and PRIEBE, C. E. (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics* **47** 2405–2439.
[10] CHEN, Y., CHI, Y., FAN, J. and MA, C. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning* **14** 566–806.
[11] DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis* **7** 1–46.
[12] DAVIS, D., DIAZ, M. and WANG, K. (2021). Clustering a mixture of gaussians with unknown covariance. *arXiv preprint arXiv:2110.01602*.
[13] DING, L. and CHEN, Y. (2020). Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory* **66** 7274–7301.
[14] FAN, J., WANG, W. and ZHONG, Y. (2018). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research* **18** 1–42.
[15] HAN, R., LUO, Y., WANG, M. and ZHANG, A. R. (2020). Exact clustering in tensor block model: Statistical optimality and computational limit. *arXiv preprint arXiv:2012.09996*.
[16] 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.
[17] JIN, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57-89.
[18] KANNAN, R. and VEMPALA, S. (2009). Spectral algorithms. *Foundations and Trends® in Theoretical Computer Science* **4** 157–288.
[19] KISELEV, V. Y., ANDREWS, T. S. and HEMBERG, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20** 273–282.
[20] KOLTCHINSKII, V. and LOUNICI, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **52** 1976–2013. Institut Henri Poincaré.
[21] KOLTCHINSKII, V. and XIA, D. (2016). Perturbation of linear forms of singular vectors under Gaussian noise. In *High Dimensional Probability VII* 397–423. Springer.
[22] LEI, J. and LIN, K. Z. (2020). Bias-adjusted spectral clustering in multi-layer stochastic block models. *arXiv preprint arXiv:2003.08222*.
[23] LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237.
[24] LEI, L. (2019). Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. *arXiv preprint arXiv:1909.04798*.
[25] LÖFFLER, M., ZHANG, A. Y. and ZHOU, H. H. (2021). Optimality of Spectral Clustering in the Gaussian Mixture Model. *Annals of Statistics* **49** 2506-2530.
[26] LU, Y. and ZHOU, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*.
[27] MA, C., WANG, K., CHI, Y. and CHEN, Y. (2018). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning* 3345–3354. PMLR.
[28] MASON, D. M. and ZHOU, H. H. (2012). Quantile coupling inequalities and their applications. *Probability Surveys* **9** 439–479.
[29] NDAOUD, M. (2018). Sharp optimal recovery in the two-component Gaussian mixture model. *arXiv preprint arXiv:1812.08078*.
[30] NDAOUD, M., SIGALLA, S. and TSYBAKOV, A. B. (2021). Improved clustering algorithms for the bipartite stochastic block model. *IEEE Transactions on Information Theory*.
[31] NEWMAN, M. E. (2013). Spectral methods for community detection and graph partitioning. *Physical Review E* **88** 042822.
[32] QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems* 3120–3128.
[33] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics* **39** 1878–1915.
[34] SCHIEBINGER, G., WAINWRIGHT, M. J. and YU, B. (2015). The geometry of kernelized spectral clustering. *The Annals of Statistics* **43** 819–846.
[35] SRIVASTAVA, P. R., SARKAR, P. and HANASUSANTO, G. A. (2019). A robust spectral clustering algorithm for sub-Gaussian mixture models with outliers. *arXiv preprint arXiv:1912.07546*.
[36] STEWART, G. W. (1990). Matrix perturbation theory.
[37] VAN DER VAART, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press.
[38] VON LUXBURG, U., BELKIN, M. and BOUSQUET, O. (2008). Consistency of spectral clustering. *The Annals of Statistics* 555–586.

[39] WANG, K., YAN, Y. and DIAZ, M. (2020). Efficient clustering for stretched mixtures: Landscape and optimality. *Advances in Neural Information Processing Systems* **33** 21309–21320.
[40] WEDIN, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics* **12** 99–111.
[41] YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323.
[42] ZHANG, A. Y. and ZHOU, H. H. (2022). Supplement to "Leave-one-out Singular Subspace Perturbation Analysis for Spectral Clustering".
[43] ZHOU, Z. and AMINI, A. A. (2019). Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *J. Mach. Learn. Res.* **20** 1–47.

SUPPLEMENT TO "LEAVE-ONE-OUT SINGULAR SUBSPACE PERTURBATION ANALYSIS FOR SPECTRAL CLUSTERING"

BY Anderson Y. Zhang and Harrison H. Zhou

University of Pennsylvania and Yale University

APPENDIX A: PROOF OF THEOREM 2.3

The proof idea is similar to that of Theorem 2.2 but with more involved calculation as r is not necessarily κ. Consider any $i \in [n]$. Define

$$\tilde{\rho}_{-i} := \frac{\hat{\lambda}_{-i,r} - \hat{\lambda}_{-i,r+1}}{\left\| \left(I - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T \right) X_i \right\|}.$$

We need to verify $\tilde{\rho}_{-i} > 2$ first in order to apply Theorem 2.1. Recall the definition of P_{-i} in (35) and E_{-i} in (37). Let the SVD of P_{-i} be

$$P_{-i} = \sum_{j=1}^{p\wedge(n-1)} \lambda_{-i,j} u_{-i,j} v_{-i,j}^T,$$

where $\lambda_{-i,1} \geq \lambda_{-i,2} \geq \ldots \geq \lambda_{-i,p\wedge(n-1)}$. Denote $U_{-i,1:r} = (u_{-i,1}, u_{-i,2}, \ldots, u_{-i,r}) \in \mathbb{O}^{p\times r}$. Then by Weyl's inequality, we have

$$|\hat{\lambda}_{-i,r} - \lambda_{-i,r}|, |\hat{\lambda}_{-i,r+1} - \lambda_{-i,r+1}| \leq \|E_{-i}\| \leq \|E\|. \tag{51}$$

Then the numerator

$$\hat{\lambda}_{-i,r} - \hat{\lambda}_{-i,r+1} \geq \lambda_{-i,r} - \lambda_{-i,r+1} - 2\|E\|. \tag{52}$$

In the following, we are going to connect $\lambda_{-i,r} - \lambda_{-i,r+1}$ with $\lambda_r - \lambda_{r+1}$.

To bridge the gap between $\lambda_{-i,r}, \lambda_{-i,r+1}$ and λ_r, λ_{r+1}, define

$$\tilde{P}_{-i} := (\theta^*_{z_1^*}, \ldots, \theta^*_{z_{i-1}^*}, U_{-i,1:r}U_{-i,1:r}^T\theta^*_{z_i^*}, \theta^*_{z_{i+1}^*}, \ldots, \theta^*_{z_n^*}) \in \mathbb{R}^{p\times n}.$$

Let $\tilde{\lambda}_{-i,1} \geq \tilde{\lambda}_{-i,2} \geq \ldots \geq \tilde{\lambda}_{-i,p\wedge n}$ be its singular values. Note that $U_{-i,1:r}U_{-i,1:r}^T\tilde{P}_{-i}$ is the best rank-r approximation of $\tilde{P}_{-i}$. This is because for any rank-r projection matrix $M \in \mathbb{R}^{p\times p}$ such that $M^2 = M$, we have

$$\begin{aligned}
\left\| \tilde{P}_{-i} - MM^T\tilde{P}_{-i} \right\|_{\mathrm{F}}^2 &= \left\|(I - MM^T)P_{-i}\right\|_{\mathrm{F}}^2 + \left\|(I - MM^T)U_{-i,1:r}U_{-i,1:r}^T\theta^*_{z_i^*}\right\|_{\mathrm{F}}^2 \\
&\geq \left\|(I - U_{-i,1:r}U_{-i,1:r}^T)P_{-i}\right\|_{\mathrm{F}}^2 + 0 \\
&= \left\|\tilde{P}_{-i} - U_{-i,1:r}U_{-i,1:r}^T\tilde{P}_{-i}\right\|_{\mathrm{F}}^2,
\end{aligned}$$

where we use the fact $U_{-i,1:r}U_{-i,1:r}^T P_{-i}$ is the best rank-r approximation of P_{-i}. Hence, $\mathrm{span}(U_{-i,1:r})$ is exactly the leading r left singular space of $\tilde{P}_{-i}$. It immediately implies:

- $\tilde{\lambda}_{-i,j} = \lambda_{-i,j}$ for any $j \geq r+1$, including

$$\tilde{\lambda}_{-i,r+1} = \lambda_{-i,r+1}. \tag{53}$$

- Since $U_{-i,1:r}U_{-i,1:r}^T\tilde{P}_{-i}$ and $U_{-i,1:r}U_{-i,1:r}^T P_{-i}$ only differ by one column where the latter one can be seen as the leave-one-out counterpart of the former one, using the same argument as in (36), we have

$$\lambda_{-i,r}^2 \geq \left(1-\frac{k}{\beta n}\right)\tilde{\lambda}_{-i,r}^2. \tag{54}$$

Then from (52), we have

$$\hat{\lambda}_{-i,r} - \hat{\lambda}_{-i,r+1} \geq \sqrt{1-\frac{k}{\beta n}}\tilde{\lambda}_{-i,r} - \tilde{\lambda}_{-i,r+1} - 2\|E\|. \tag{55}$$

For the difference between $\tilde{\lambda}_{-i,r}, \tilde{\lambda}_{-i,r+1}$ and λ_r, λ_{r+1}, we use the Weyl's inequality again:

$$\max_{j\in[k]}\left|\tilde{\lambda}_{-i,j} - \lambda_j\right| \leq \left\|P - \tilde{P}_{-i}\right\| = \left\|\theta_{z_i^*}^* - U_{-i,1:r}U_{-i,1:r}^T\theta_{z_i^*}^*\right\|.$$

In the proof of Theorem 2.2, we show $u_{-i,j} \in \text{span}(\{\theta_a^*\}_{a\in[k]})$ for each $j\in[\kappa]$. Then

$$\begin{aligned}\left\|\theta_{z_i^*}^* - U_{-i,1:r}U_{-i,1:r}^T\theta_{z_i^*}^*\right\| &= \left\|\left(u_{-i,r+1},\dots,u_{-i,\kappa}\right)\left(u_{-i,r+1},\dots,u_{-i,\kappa}\right)^T\theta_{z_i^*}^*\right\| \\ &= \sqrt{\sum_{a\in[\kappa]:a\geq r+1}\left(u_{-i,a}^T\theta_{z_i^*}^*\right)^2}.\end{aligned}$$

For any $a\in[\kappa]$ such $a\geq r+1$, we have

$$\begin{aligned}\left(u_{-i,a}^T\theta_{z_i^*}^*\right)^2 &\leq \frac{1}{\left|\left\{j\in[n]: z_j^* = z_i^*\right\}\right| - 1}\sum_{j\in[n]:j\neq i, z_j^*=z_i^*}\left(u_{-i,a}^T\theta_{z_j^*}^*\right)^2 \leq \frac{1}{\frac{\beta n}{k}-1}(u_{-i,a}^T P_{-i})^2 \\ &\leq \frac{\lambda_{-i,a}^2}{\frac{\beta n}{k}-1} \leq \frac{\lambda_{-i,r+1}^2}{\frac{\beta n}{k}-1}.\end{aligned}$$

Hence, we obtain $\|\theta_{z_i^*}^* - U_{-i,1:r}U_{-i,1:r}^T\theta_{z_i^*}^*\| \leq \sqrt{\kappa}\lambda_{-i,a}/\sqrt{\beta n/k-1}$ and consequently,

$$\max_{j\in[k]}\left|\tilde{\lambda}_{-i,j} - \lambda_j\right| \leq \frac{\sqrt{\kappa}\lambda_{-i,r+1}}{\sqrt{\frac{\beta n}{k}-1}}. \tag{56}$$

Then together with (53), we have $|\lambda_{-i,r+1} - \lambda_{r+1}| \leq \sqrt{\kappa}\lambda_{-i,r+1}/\sqrt{\beta n/k-1}$ and hence

$$\lambda_{-i,r+1} \leq \frac{\lambda_{r+1}}{1-\frac{\sqrt{\kappa}}{\sqrt{\frac{\beta n}{k}-1}}}. \tag{57}$$

Denote $d := \beta n/k$. With (55), we have

$$\begin{aligned}\hat{\lambda}_{-i,r} - \hat{\lambda}_{-i,r+1} &\geq \sqrt{\frac{d-1}{d}}\left(\lambda_r - \frac{\lambda_{-i,r+1}}{\sqrt{d-1}}\right) - \left(\lambda_{r+1} + \frac{\lambda_{-i,r+1}}{\sqrt{d-1}}\right) - 2\|E\| \\ &\geq \sqrt{\frac{d-1}{d}}\lambda_r - \lambda_{r+1}\left(1 + \left(\frac{1}{\sqrt{d}} + \frac{1}{\sqrt{d-1}}\right)\frac{1}{1-\frac{\sqrt{\kappa}}{\sqrt{d-1}}}\right) - 2\|E\| \\ &\geq \sqrt{\frac{d-1}{d}}\left(\lambda_r - \lambda_{r+1} - \frac{4}{\sqrt{d}}\lambda_{r+1}\right) - 2\|E\| \\ &\geq \frac{3}{4}\left(\lambda_r - \lambda_{r+1} - \frac{4}{\sqrt{d}}\lambda_{r+1}\right) - 2\|E\|, \qquad (58)\end{aligned}$$

where in the last two inequalities we use the assumption that $d/k \geq 10$. As a consequence, we have

$$\tilde{\rho}_{-i} \geq \frac{\hat{\lambda}_{-i,r} - \hat{\lambda}_{-i,r+1}}{\left\| \left(I - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T \right) X_i \right\|} \geq \frac{\frac{3}{4}\left(\lambda_r - \lambda_{r+1} - \frac{4}{\sqrt{d}}\lambda_{r+1} \right) - 2\left\| E \right\|}{\left\| \left(I - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T \right) X_i \right\|}.$$

Next, we are going to simplify the denominator of the above display. Using the orthogonality of the singular vectors, we have

$$\begin{aligned}
&\left\| \left(I - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T \right) \theta^*_{z_i^*} \right\| \\
&\leq \left\| \left(I - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T \right) \theta^*_{z_i^*} \right\| + \left\| \left(\hat{u}_{-i,r+1}, \ldots, \hat{u}_{-i,\kappa} \right) \left(\hat{u}_{-i,r+1}, \ldots, \hat{u}_{-i,\kappa} \right)^T \theta^*_{z_i^*} \right\| \\
&= \left\| \left(I - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T \right) \theta^*_{z_i^*} \right\| + \sqrt{\sum_{j=r+1}^{\kappa} \left(\hat{u}_{-i,j}^T \theta^*_{z_i^*} \right)^2} \\
&\leq \frac{3\sqrt{\kappa}\left\| E \right\|}{\sqrt{\frac{\beta n}{k} - 1}} + \sqrt{\sum_{j=r+1}^{\kappa} \left(\frac{\hat{\lambda}_{-i,j}}{\sqrt{\frac{\beta n}{k} - 1}} + \frac{\left\| E \right\|}{\sqrt{\frac{\beta n}{k} - 1}} \right)^2} \\
&\leq \frac{3\sqrt{\kappa}\left\| E \right\|}{\sqrt{\frac{\beta n}{k} - 1}} + \sqrt{\kappa}\left(\frac{\hat{\lambda}_{-i,r+1}}{\sqrt{\frac{\beta n}{k} - 1}} + \frac{\left\| E \right\|}{\sqrt{\frac{\beta n}{k} - 1}} \right),
\end{aligned}$$

where the second to the inequality is due to (40) and (43). By (57) and the Weyl's inequality, we have

$$\hat{\lambda}_{-i,r+1} \leq \lambda_{-i,r+1} + \left\| E \right\| \leq \frac{1}{1 - \frac{\sqrt{\kappa}}{\sqrt{\frac{\beta n}{k} - 1}}} \lambda_{r+1} + \left\| E \right\|.$$

Then, with the assumption $\beta n / k^2 \geq 10$, we have

$$\begin{aligned}
\left\| \left(I - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T \right) \theta^*_{z_i^*} \right\| &\leq \frac{3\sqrt{\kappa}\left\| E \right\|}{\sqrt{\frac{\beta n}{k} - 1}} + \sqrt{\kappa}\left(\frac{\lambda_{r+1}}{\sqrt{\frac{\beta n}{k} - 1} - \sqrt{\kappa}} + \frac{2\left\| E \right\|}{\sqrt{\frac{\beta n}{k} - 1}} \right) \\
&\leq \frac{\sqrt{k\kappa}}{\sqrt{\beta n}}(6\left\| E \right\| + 2\lambda_{r+1}).
\end{aligned}$$

Hence,

$$\begin{aligned}
\left\| \left(I - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T \right) X_i \right\| &\leq \left\| \left(I - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T \right) \theta^*_{z_i^*} \right\| + \left\| \left(I - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T \right) \epsilon_i \right\| \\
&\leq \frac{\sqrt{k\kappa}}{\sqrt{\beta n}}(6\left\| E \right\| + 2\lambda_{r+1}) + \left\| E \right\|.
\end{aligned}$$

As a result,

$$\tilde{\rho}_{-i} \geq \frac{\frac{3}{4}\left(\lambda_r - \lambda_{r+1} - \frac{4}{\sqrt{\beta n / k}}\lambda_{r+1} \right) - 2\left\| E \right\|}{\frac{\sqrt{k\kappa}}{\sqrt{\beta n}}(6\left\| E \right\| + 2\lambda_{r+1}) + \left\| E \right\|} \geq \frac{\tilde{\rho}_0}{8} > 2,$$

under the assumption that $\beta n / (k^2) \geq 10$ and (10).

The remaining part of the proof is to study $\{\hat{u}_{-i,a}^T X_i\}_{a\in[r]}$ and then apply Theorem 2.1. Following the exact argument as in the proof of Theorem 2.2, we have

$$\sqrt{\sum_{a\in r}\left(\frac{\hat{u}_{-i,a}^T X_i}{\hat{\lambda}_{-i,a}}\right)^2} \leq \frac{\sqrt{r}}{\sqrt{\frac{\beta n}{k}-1}} + \frac{1}{\hat{\lambda}_{-i,r}}\frac{\|E\|\sqrt{r}}{\sqrt{\frac{\beta n}{k}-1}} + \frac{1}{\hat{\lambda}_{-i,r}}\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|.$$

Under the assumption that $\beta n/(k^2) \geq 10$ and (10), (58) is lower bounded by $\lambda_r/2$. This also implies $\hat{\lambda}_{-i,r} \geq \lambda_r/2$. Then a direct application of Theorem 2.1 leads to

$$\begin{aligned}\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\right\|_{\mathrm{F}} &\leq \frac{4\sqrt{2}}{\tilde{\rho}_{-i}}\left(\frac{\sqrt{r}}{\sqrt{\beta n/k-1}} + \frac{1}{\hat{\lambda}_{-i,r}}\left(\frac{\sqrt{r}\,\|E\|}{\sqrt{\beta n/k-1}} + \left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|\right)\right)\\ &\leq \frac{128}{\tilde{\rho}_0}\left(\frac{\sqrt{kr}}{\sqrt{\beta n}} + \frac{\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|}{\lambda_r}\right).\end{aligned}$$

APPENDIX B: PROOFS OF LEMMA 3.3 AND THEOREM 3.2

PROOF OF LEMMA 3.3. Note that $\hat{r} \in [k]$ is a random variable. We are going to associate it with some deterministic index in $[k]$. Recall $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{p\wedge n}$ are singular values of the signal matrix P and κ is the its rank. Let its SVD be $P = \sum_{i\in[p\wedge n]}\lambda_i u_i v_i^T$ with $\{u_j\}_{j\in[p\wedge n]} \in \mathbb{R}^p$ being its left singular vectors.

By the definition of $\hat{r}$ in (22) and the definition of $\tilde{\rho}$, we know $\hat{\lambda}_{\hat{r}} - \hat{\lambda}_{\hat{r}+1} \geq \tilde{\rho}\,\|E\|$ and $\hat{\lambda}_{\hat{r}+1} \leq k\tilde{\rho}\,\|E\|$. By Weyl's inequality, we have $|\hat{\lambda}_a - \lambda_a| \leq \|E\|$ for all singular values of X and P. Then we have $\lambda_{\hat{r}} - \lambda_{\hat{r}+1} \geq (\tilde{\rho}-2)\,\|E\|$ and $\lambda_{\hat{r}+1} \leq (k\tilde{\rho}+1)\,\|E\|$. Note that $(\tilde{\rho}-2)\,\|E\| > 0$ is as long as $\tilde{\rho} > 2$. Define

$$\mathcal{R} := \{a \in [k] : \lambda_a - \lambda_{a+1} \geq (\tilde{\rho}-2)\,\|E\| \text{ and } \lambda_{a+1} \leq (k\tilde{\rho}+1)\,\|E\|\}, \tag{59}$$

which is a deterministic subset of $[\kappa]$. Then $\hat{r} \in \mathcal{R}$.

Consider an arbitrary $r \in \mathcal{R}$ and define $\hat{U}_{1:r} := (\hat{u}_1,\ldots,\hat{u}_r) \in \mathbb{R}^{p\times r}$. Perform k-means on the columns of $\hat{U}_{1:r}\hat{U}_{1:r}^T X$ and let the output be

$$\left(\check{z}(r), \{\check{\theta}_j(r)\}_{j=1}^k\right) = \underset{z\in[k]^n,\{\theta_j\}_{j=1}^k\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in[n]}\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T X - \theta_{z_i}\right\|^2.$$

In the following, we are going to establish statistical properties for $\check{z}(r)$ and eventually obtain a desired upper bound for $\ell(\check{z}(r), z^*)$. Since performing k-means on the columns of $\hat{U}_{1:r}^T X$ is equivalent to k-means on the columns of $\hat{U}_{1:r}\hat{U}_{1:r}^T X$, and since $\hat{r} \in \mathcal{R}$, we have $\tilde{z} = \check{z}(\hat{r})$ and thus the desired upper bound also holds for $\ell(\tilde{z}, z^*)$.

In the rest of the proof we are going to analyze $\check{z}(r)$ for any $r \in \mathcal{R}$. For simplicity, we use the notation $\check{z}, \{\check{\theta}_j\}_{j\in[n]}$ instead of $\check{z}(r), \{\check{\theta}_j(r)\}_{j\in[n]}$. The remaining proof can be decomposed into several parts.

(Preliminary Results for $\check{z}, \{\check{\theta}_j\}_{j\in[n]}$). We are going to use Proposition 3.1 to have some preliminary results. Define $U_{1:r} := (u_1,\ldots,u_r)$ and $U_{(r+1):k} := (u_{r+1},\ldots,u_k)$. Instead of the decomposition (5), we can write

$$X_i = U_{1:r}U_{1:r}^T\theta_{z_i^*}^* + U_{(r+1):k}U_{(r+1):k}^T\theta_{z_i^*}^* + \epsilon_i = U_{1:r}U_{1:r}^T\theta_{z_i^*}^* + \check{\epsilon}_i,$$

where $\check{\epsilon}_i := U_{(r+1):k}U_{(r+1):k}^T\theta^*_{z_i^*} + \epsilon_i$. In this way, we have a new mixture model with the centers being $\{U_{1:r}U_{1:r}^T\theta^*_a\}_{a\in[k]}$ and the additive noises being $\{\check{\epsilon}_i\}$. Define $\check{E} := (\check{\epsilon}_1, \ldots, \check{\epsilon}_n)$. Then

$$\begin{aligned}\left\|\check{E}\right\| &\leq \|E\| + \left\|\left(U_{(r+1):k}U_{(r+1):k}^T\theta^*_{z_1^*}, \ldots, U_{(r+1):k}U_{(r+1):k}^T\theta^*_{z_n^*}\right)\right\| \\ &= \|E\| + \left\|U_{(r+1):k}U_{(r+1):k}^T P\right\| = \|E\| + \lambda_{r+1} \\ &\leq (k\tilde{\rho} + 2)\|E\|. \end{aligned} \tag{60}$$

The separation among the new centers is no longer Δ. Define

$$\check{\Delta} := \min_{a,b\in[k]:a\neq b} \left\|U_{1:r}U_{1:r}^T\theta^*_a - U_{1:r}U_{1:r}^T\theta^*_b\right\|.$$

For any $a, b \in [k]$, $U_{1:r}U_{1:r}^T\theta^*_a - U_{1:r}U_{1:r}^T\theta^*_b = (\theta^*_a - \theta^*_b) - U_{(r+1):k}U_{(r+1):k}^T\theta^*_a + U_{(r+1):k}U_{(r+1):k}^T\theta^*_b$. Also,

$$\begin{aligned}\max_{a\in[k]} \left\|U_{(r+1):k}U_{(r+1):k}^T\theta^*_a\right\| &= \max_{a\in[k]} \sqrt{\frac{\sum_{i\in[n]:z_i^*=a}\left\|U_{(r+1):k}U_{(r+1):k}^T\theta^*_a\right\|^2}{|\{i\in[n]:z_i^*=a\}|}} \leq \frac{\left\|U_{(r+1):k}U_{(r+1):k}^T P\right\|_{\mathrm{F}}}{\sqrt{\beta n/k}} \\ &\leq \frac{2\sqrt{k}\lambda_{r+1}}{\sqrt{\beta n/k}} \leq \frac{\sqrt{k}(k\tilde{\rho}+1)\|E\|}{\sqrt{\beta n/k}}. \end{aligned} \tag{61}$$

Hence, we have

$$\check{\Delta} \geq \min_{a,b\in[k]:a\neq b}\|\theta^*_a - \theta^*_b\| - 2\max_{a\in[k]}\left\|U_{(r+1):k}U_{(r+1):k}^T\theta^*_a\right\| \geq \Delta - \frac{2\sqrt{k}(k\tilde{\rho}+1)\|E\|}{\sqrt{\beta n/k}}. \tag{62}$$

Then from Proposition 3.1, as long as (which will be verified later)

$$\check{\psi}_0 := \frac{\check{\Delta}}{\beta^{-0.5}kn^{-0.5}\left\|\check{E}\right\|} \geq 16, \tag{63}$$

we have

$$\ell(\check{z}, z^*) = \frac{1}{n}|i\in[n] : \check{z}_i \neq \phi(z_i^*)| \leq \frac{C_0 k\left\|\check{E}\right\|^2}{n\check{\Delta}^2},$$

and

$$\max_{a\in[k]}\left\|\check{\theta}_{\phi(z)} - U_{1:r}U_{1:r}^T\theta^*_a\right\| \leq C_0\beta^{-0.5}kn^{-0.5}\left\|\check{E}\right\|.$$

where $C_0 = 128$.

(Entrywise Decomposition for $\check{z}$). Next, we are going to have an entrywise decomposition for $\mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\}$ that is analogous to that of Lemma 3.2. When (63) is satisfied, from Lemma 3.1, we have

$$\mathbb{I}\{\check{z}_i \neq \phi(z_i^*)\} \leq \mathbb{I}\left\{\left(1 - C_0\check{\psi}_0^{-1}\right)\check{\Delta} \leq 2\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\check{\epsilon}_i\right\|\right\}.$$

By the definition of $\check{\epsilon}_i$ and (61), we have

$$\begin{aligned}\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\check{\epsilon}_i\right\| &\leq \left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\right\| + \left\|\hat{U}_{1:r}\hat{U}_{1:r}^T U_{(r+1):k}U_{(r+1):k}^T\theta^*_{z_i^*}\right\| \\ &\leq \left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\right\| + \left\|U_{(r+1):k}U_{(r+1):k}^T\theta^*_{z_i^*}\right\| \\ &\leq \left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\right\| + \frac{\sqrt{k}(k\tilde{\rho}+1)\|E\|}{\sqrt{\beta n/k}}. \end{aligned}$$

Then, we have

$$\mathbb{I}\{\check{z}_i \neq \phi(z_i^*)\} \leq \mathbb{I}\left\{\left(1 - C_0\check{\psi}_0^{-1}\right)\check{\Delta} \leq 2\left(\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\right\| + \frac{\sqrt{k}(k\tilde{\rho}+1)\|E\|}{\sqrt{\beta n/k}}\right)\right\}$$
$$= \mathbb{I}\left\{\left(1 - C_0\check{\psi}_0^{-1} - \frac{2\sqrt{k}(k\tilde{\rho}+1)\|E\|}{\sqrt{\beta n/k}\check{\Delta}}\right)\check{\Delta} \leq 2\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\right\|\right\}.$$

From (59), under the assumption that $\tilde{\rho} > 4$ and $\beta n/k^4 > 400$, we have $\tilde{\rho}_0$ defined as in (10) to satisfy

$$\tilde{\rho}_0 \geq \frac{(\tilde{\rho}-1)\|E\|}{\max\left\{\|E\|, \sqrt{\frac{k^2}{\beta n}}(k\tilde{\rho}+1)\|E\|\right\}} \geq 2.$$

Then Theorem 2.3 can be applied, with which we have

$$\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\right\|_{\mathrm{F}} \leq \frac{256\sqrt{rk}}{\sqrt{n\beta}} + \frac{256\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|}{\lambda_r}.$$

Then following the proof of Lemma 3.2, we have

$$\mathbb{I}\{\check{z}_i \neq \phi(z_i^*)\}$$
$$\leq \mathbb{I}\left\{\left(1 - C_0\check{\psi}_0^{-1} - \frac{2\sqrt{k}(k\tilde{\rho}+1)\|E\|}{\sqrt{\beta n/k}\check{\Delta}}\right)\check{\Delta} \leq 2\left(\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\| + \left\|\hat{U}_{1:r}\hat{U}_{1:r}^T - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\right\|_{\mathrm{F}}\|E\|\right)\right\}$$
$$\leq \mathbb{I}\left\{\left(1 - C_0\check{\psi}_0^{-1} - \frac{2\sqrt{k}(k\tilde{\rho}+1)\|E\|}{\sqrt{\beta n/k}\check{\Delta}}\right)\check{\Delta} \leq 2\left(\frac{256\sqrt{rk}\|E\|}{\sqrt{n\beta}} + \left(1 + \frac{256\|E\|}{\lambda_r}\right)\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|\right)\right\}$$
$$\leq \mathbb{I}\left\{\left(1 - C_0\check{\psi}_0^{-1} - \frac{2\sqrt{k}(k\tilde{\rho}+257)\|E\|}{\sqrt{\beta n/k}\check{\Delta}}\right)\check{\Delta} \leq 2\left(1 + \frac{256\|E\|}{\lambda_r}\right)\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|\right\}$$
$$\leq \mathbb{I}\left\{\left(1 - C_0\check{\psi}_0^{-1} - \frac{2\sqrt{k}(k\tilde{\rho}+257)\|E\|}{\sqrt{\beta n/k}\check{\Delta}}\right)\check{\Delta} \leq 2\left(1 + \frac{256}{\tilde{\rho}-2}\right)\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|\right\},$$

where in the last inequality we use $\lambda_r \geq (\tilde{\rho}-2)\|E\| > 0$ (as long as $\tilde{\rho} > 2$) from (59).

The last step of the proof is to simplify the above display using Δ instead of $\check{\Delta}$. Then, under the assumption that $\tilde{\rho} > 256$, we have $(1 + 256/(\tilde{\rho}-2))^{-1} \leq (1 - 512/\tilde{\rho})$. Recall the definition of $\tilde{\psi}_0$ in (24). Under the assumption that $\tilde{\rho} \leq \tilde{\psi}_0/64$, we have

$$\check{\Delta} \geq \Delta\left(1 - \frac{4\beta^{-0.5}k^2n^{-0.5}\tilde{\rho}\|E\|}{\Delta}\right) = \Delta\left(1 - \frac{4\tilde{\rho}}{\tilde{\psi}_0}\right) \geq \frac{\Delta}{2}, \tag{64}$$

according to (62). Then together with (60), we can verify (63) holds due to

$$\check{\psi}_0 \geq \frac{\Delta/2}{\beta^{-0.5}kn^{-0.5}(k\tilde{\rho}+2)\|E\|} \geq \frac{\Delta}{4\beta^{-0.5}k^2n^{-0.5}\tilde{\rho}\|E\|} = \frac{\tilde{\psi}_0}{4\tilde{\rho}} \geq 16.$$

Rearranging all the terms with the help of (64), we can simplify $\mathbb{I}\{\check{z}_i \neq \phi(z_i^*)\}$ into

$$\mathbb{I}\{\check{z}_i \neq \phi(z_i^*)\}$$
$$\leq \mathbb{I}\left\{\left(1 - 4C_0\tilde{\rho}\tilde{\psi}_0 - \frac{4\beta^{-0.5}k^2n^{-0.5}\tilde{\rho}\|E\|}{\Delta/2}\right)\left(1 - \frac{256}{\tilde{\rho}}\right)\left(1 - \frac{4\tilde{\rho}}{\tilde{\psi}_0}\right)\Delta \leq 2\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|\right\}$$
$$\leq \mathbb{I}\left\{\left(1 - 5C_0\tilde{\rho}\tilde{\psi}_0^{-1} - 256\tilde{\rho}^{-1}\right)\Delta \leq 2\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|\right\}.$$

□

PROOF OF THEOREM 3.2. Recall the definition of $\mathcal{F}$ in (45). Then if $\mathcal{F}$ holds, by appropriate choices of C_1, C_2, we can verify the assumptions needed in Lemma 3.3 hold, which lead to

$$\mathbb{I}\left\{\tilde{z}_i \neq \phi(z_i^*)\right\}\mathbb{I}\left\{\mathcal{F}\right\} \leq \mathbb{I}\left\{\left(1 - C''(\rho_2\psi_2^{-1} + \rho_2^{-1})\right)\Delta \leq 2\left\|\hat{U}_{-i,1:\hat{r}}\hat{U}_{-i,1:\hat{r}}^T\epsilon_i\right\|\right\}\mathbb{I}\left\{\mathcal{F}\right\},$$

for some constant $C'' > 0$. Though $\hat{r}$ is random, the proof of Lemma 3.3 shows that $\hat{r} \in \mathcal{R} \subset [k]$ where $\mathcal{R}$ is defined in (59). Note that for any $r \in [k]$, we can follow the proof of Theorem 3.1 to show

$$\mathbb{E}\mathbb{I}\left\{\left(1 - C''(\rho_2\psi_2^{-1} + \rho_2^{-1})\right)\Delta \leq 2\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|\right\} \leq \exp\left(-(1 - C'''(\rho_2\psi_2^{-1} + \rho_2^{-1}))\frac{\Delta^2}{8\sigma^2}\right),$$

for some constant $C''' > 0$. Hence, the same upper bound holds for $\mathbb{E}\mathbb{I}\{(1 - C''(\rho_2\psi_2^{-1} + \rho_2^{-1}))\Delta \leq 2\|\hat{U}_{-i,1:\hat{r}}\hat{U}_{-i,1:\hat{r}}^T\epsilon_i\|\}$. The rest of the proof follows that of Theorem 3.1 and is omitted here. □

APPENDIX C: PROOF OF THEOREM 3.3

Define $\mathcal{F} = \left\{\|E\| \leq \sqrt{2}(\sqrt{n} + \sqrt{p})\sigma\right\}$. Then by Lemma B.1 of [25], we have $\mathbb{P}(\mathcal{F}) \geq 1 - e^{-0.08n}$. Then under the event $\mathcal{F}$, the assumption (25) implies (15) holds, and hence (16) and (17) hold. For simplicity, and without loss of generality, we can let ϕ in (16)-(17) to be the identity, and we get

$$\ell(\hat{z}, z^*) = \frac{1}{n}|\{i \in [n] : \hat{z}_i \neq z_i^*\}| \leq \frac{C_0 k \left(1 + \sqrt{\frac{p}{n}}\right)^2 \sigma^2}{\Delta^2},$$

and

$$\max_{a \in [k]}\left\|\hat{\theta}_a - \theta_a^*\right\| \leq C_0\beta^{-0.5}k\left(1 + \sqrt{\frac{p}{n}}\right)\sigma,$$

where $C_0 > 0$ is some constant.

Denote $\hat{P} = \hat{U}_{1:k}\hat{U}_{1:k}^T X$ and let $\hat{P}_{\cdot,i}$ be its ith column so that $\hat{P}_{\cdot,i} = \hat{U}_{1:k}\hat{U}_{1:k}^T X_i$. We define $r \in [k]$ as (with $\lambda_{k+1} := 0$)

$$r = \max\left\{j \in [k] : \lambda_j - \lambda_{j+1} \geq \tau\sqrt{n+p}\sigma\right\}, \tag{65}$$

for a sequence $\tau \to \infty$ to be determined later. We note that if $\Delta/(k^{\frac{3}{2}}\tau\beta^{\frac{1}{2}}(1+p/n)^{\frac{1}{2}}\sigma) \to \infty$, the set $\{j \in [k] : \lambda_j - \lambda_{j+1} \geq \tau\sqrt{n+p}\sigma\}$ is not empty. Otherwise, this would imply $\lambda_1 \leq k\tau\sqrt{n+p}\sigma$ which would contradict with the fact $\lambda_1 \geq \sqrt{\beta n/k}\Delta/(2\sigma)$ (cf. Proposition A.1 of [25]). By the definition of r in (65), we immediately have

$$\lambda_r - \lambda_{r+1} \geq \tau\sqrt{n+p}\sigma, \tag{66}$$

$$\text{and } \lambda_{r+1} \leq k\tau\sqrt{n+p}\sigma. \tag{67}$$

We split $\hat{U}_{1:k}$ into $(\hat{U}_{1:r}, \hat{U}_{(r+1):k})$ where $\hat{U}_{1:r} := (\hat{u}_1, \ldots, \hat{u}_r)$ and $\hat{U}_{(r+1):k} := (\hat{u}_{r+1}, \ldots, \hat{u}_k)$. We decompose $\hat{P}_{\cdot,i} = \hat{P}_{\cdot,i}^{(1)} + \hat{P}_{\cdot,i}^{(2)}$, where $\hat{P}_{\cdot,i}^{(1)} := \hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P}_{\cdot,i}$ and $\hat{P}_{\cdot,i}^{(2)} := \hat{U}_{(r+1):k}\hat{U}_{(r+1):k}^T\hat{P}_{\cdot,i}$. Similarly, for each $a \in [k]$, we decompose $\hat{\theta}_a = \hat{\theta}_a^{(1)} + \hat{\theta}_a^{(2)}$, where $\hat{\theta}_a^{(1)} := \hat{U}_{1:r}\hat{U}_{1:r}^T\hat{\theta}_a$ and

$\hat{\theta}_a^{(2)} := \hat{U}_{(r+1):k}\hat{U}_{(r+1):k}^T\hat{\theta}_a$. Due to the orthogonality of $\{\hat{u}_l\}_{l\in[k]}$, we obtain that for any $i\in[n]$ and any $a\in[k]$ such that $a\neq z_i^*$,

$$\begin{aligned}\mathbb{I}\{\hat{z}_i=a\}&\leq \mathbb{I}\left\{\left\|\hat{P}_{\cdot,i}^{(1)}+\hat{P}_{\cdot,i}^{(2)}-\hat{\theta}_a^{(1)}-\hat{\theta}_a^{(2)}\right\|^2\leq\left\|\hat{P}_{\cdot,i}^{(1)}+\hat{P}_{\cdot,i}^{(2)}-\hat{\theta}_{z_i^*}^{(1)}-\hat{\theta}_{z_i^*}^{(2)}\right\|^2\right\}\\&=\mathbb{I}\left\{2\left\langle \hat{P}_{\cdot,i}^{(1)}-\hat{\theta}_{z_i^*}^{(1)},\hat{\theta}_{z_i^*}^{(1)}-\hat{\theta}_a^{(1)}\right\rangle+\left\|\hat{\theta}_{z_i^*}^{(1)}-\hat{\theta}_a^{(1)}\right\|^2\leq 2\left\langle \hat{P}_{\cdot,i}^{(2)},\hat{\theta}_a^{(2)}-\hat{\theta}_{z_i^*}^{(2)}\right\rangle-\left\|\hat{\theta}_a^{(2)}\right\|^2+\left\|\hat{\theta}_{z_i^*}^{(2)}\right\|^2\right\}\end{aligned}$$

We denote $\tau''=o(1)$ to be another sequence which we will specify later. Then the above display can be decomposed and upper bounded by

$$\begin{aligned}\mathbb{I}\{\hat{z}_i=a\}\leq&\mathbb{I}\left\{\left\|\hat{\theta}_{z_i^*}^{(1)}-\hat{\theta}_a^{(1)}\right\|-\frac{\tau''\Delta^2+\left\|\hat{\theta}_{z_i^*}^{(2)}\right\|^2}{\left\|\hat{\theta}_{z_i^*}^{(1)}-\hat{\theta}_a^{(1)}\right\|}\leq 2\left\|\hat{P}_{\cdot,i}^{(1)}-\hat{\theta}_{z_i^*}^{(1)}\right\|\right\}\\&+\mathbb{I}\left\{\tau''\Delta^2\leq 2\left\langle\hat{P}_{\cdot,i}^{(2)},\hat{\theta}_a^{(2)}-\hat{\theta}_{z_i^*}^{(2)}\right\rangle\right\}=:A_{i,a}+B_{i,a}.\end{aligned}$$

Then

$$\begin{aligned}\mathbb{E}\ell(\hat{z},z^*)&\leq\frac{1}{n}\sum_{i\in[n]}\sum_{a\in[k]:a\neq z_i^*}\mathbb{E}\mathbb{I}\{\hat{z}_i=a\}\\&\leq\mathbb{P}\left(\mathcal{F}^{\complement}\right)+\frac{1}{n}\sum_{i\in[n]}\sum_{a\in[k]:a\neq z_i^*}\mathbb{E}A_{i,a}\mathbb{I}\{\mathcal{F}\}+\frac{1}{n}\sum_{i\in[n]}\sum_{a\in[k]:a\neq z_i^*}\mathbb{E}B_{i,a}\mathbb{I}\{\mathcal{F}\}.\end{aligned}\tag{68}$$

We are going to establish upper bounds first for $n^{-1}\sum_{i\in[n]}\sum_{a\in[k]:a\neq z_i^*}\mathbb{E}B_{i,a}\mathbb{I}\{\mathcal{F}\}$ and then for $n^{-1}\sum_{i\in[n]}\sum_{a\in[k]:a\neq z_i^*}\mathbb{E}A_{i,a}\mathbb{I}\{\mathcal{F}\}$.

(Analysis on $n^{-1}\sum_{i\in[n]}\sum_{a\neq z_i^*}\mathbb{E}B_{i,a}\mathbb{I}\{\mathcal{F}\}$). For $\sum_{i\in[n]}\sum_{a\neq z_i^*}\mathbb{E}B_{i,a}\mathbb{I}\{\mathcal{F}\}$, we can directly use upper bounds established in Section 4.4.3 of [25][1]. It proves that for any $i\in[n]$,

$$\sum_{a\in[k]:a\neq z_i^*}B_{i,a}\mathbb{I}\{\mathcal{F}\cap\mathcal{T}\}\leq 2\exp\left(-\frac{1}{2}\left(c_4\frac{\tau''\Delta}{k^{\frac{7}{2}}\tau^2\beta^{-\frac{1}{2}}(1+\frac{p}{n})\sigma}\sqrt{\frac{n-k}{3n}}\right)^2\frac{\Delta^2}{\sigma^2}\right),$$

where $c_4>0$ is some constant, and $\mathcal{T}$ is some with-high-probability event in the sense that

$$\mathbb{P}(\mathcal{T})\geq 1-nk\exp\left(-\frac{(n-k)}{9}\right).$$

Hence,

$$\begin{aligned}\frac{1}{n}\sum_{i\in[n]}\sum_{a\in[k]:a\neq z_i^*}\mathbb{E}B_{i,a}\mathbb{I}\{\mathcal{F}\}&\leq\frac{1}{n}\sum_{i\in[n]}\sum_{a\in[k]:a\neq z_i^*}\mathbb{E}B_{i,a}\mathbb{I}\{\mathcal{F}\cap\mathcal{T}\}+\mathbb{P}\left(\mathcal{T}^{\complement}\right)\\&\leq 2\exp\left(-\frac{1}{2}\left(c_4\frac{\tau''\Delta}{k^{\frac{7}{2}}\tau^2\beta^{-\frac{1}{2}}(1+\frac{p}{n})\sigma}\sqrt{\frac{n-k}{3n}}\right)^2\frac{\Delta^2}{\sigma^2}\right)+nk\exp\left(-\frac{(n-k)}{9}\right).\end{aligned}$$

[1]The model in [25] assumes $\{\epsilon_j\}\overset{iid}{\sim}\mathcal{N}(0,I)$ while in this paper we assume $\{\epsilon_j\}\overset{iid}{\sim}\mathcal{N}(0,\sigma^2 I)$. To directly use results from [25], we can re-scale our data to have $X_j'=X_j/\sigma$ for all $j\in[n]$. Then $\{X_j'\}$ has $\mathcal{N}(0,I)$ noise and the separation between their centers becomes Δ/σ. Then all the results from [25] can be used here with Δ replaced by Δ/σ.

(Analysis on $n^{-1}\sum_{i\in[n]}\sum_{a\neq z_i^*}\mathbb{E}A_{i,a}\mathbb{I}\{\mathcal{F}\}$). We first follow some algebra as in Section 4.4.2 of [25] to simplify $A_{i,a}\mathbb{I}\{\mathcal{F}\}$. For any $i\in[n]$ and $a\neq z_i^*$, it proves

$$A_{i,a}\mathbb{I}\{\mathcal{F}\}\leq\mathbb{I}\left\{\left(1-c_1\tau''-\frac{c_1k^2\tau\beta^{-\frac{1}{2}}\sqrt{1+\frac{p}{n}}\sigma}{\Delta}\right)\Delta\leq 2\left\|\hat{P}^{(1)}_{\cdot,i}-\hat{\theta}^{(1)}_{z_i^*}\right\|\right\}\mathbb{I}\{\mathcal{F}\}, \tag{69}$$

for some constant $c_1>0$. Still working on the event $\mathcal{F}$, it also proves

$$\left\|\hat{P}^{(1)}_{\cdot,i}-\hat{\theta}^{(1)}_{z_i^*}\right\|\leq\left\|\hat{P}^{(1)}_{\cdot,i}-\hat{U}_{1:r}\hat{U}_{1:r}^T\theta^*_{z_i^*}\right\|+8\sqrt{2}\sqrt{\beta^{-1}k^2\left(1+\frac{p}{n}\right)}\sigma. \tag{70}$$

Our following analysis on $A_{i,a}\mathbb{I}\{\mathcal{F}\}$ is different from the rest proof in Section 4.4.2 of [25]. Note that $\hat{P}^{(1)}_{\cdot,i}-\hat{U}_{1:r}\hat{U}_{1:r}^T\theta^*_{z_i^*}=\hat{U}_{1:r}\hat{U}_{1:r}^TX_i-\hat{U}_{1:r}\hat{U}_{1:r}^T\theta^*_{z_i^*}=\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i$. Then (69) and (70) give

$$A_{i,a}\mathbb{I}\{\mathcal{F}\}\leq\mathbb{I}\left\{\left(1-c_2\tau''-\frac{c_2k^2\tau\beta^{-\frac{1}{2}}\left(1+\sqrt{\frac{p}{n}}\right)\sigma}{\Delta}\right)\Delta\leq 2\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\right\|\right\}\mathbb{I}\{\mathcal{F}\}, \tag{71}$$

where we use $\tau\to\infty$ and the fact that $1+\sqrt{p/n}$, $\sqrt{1+p/n}$ are in the same order.

Recall the definition of X_{-i} in (7) and $\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T$ is the leave-one-out counterpart of $\hat{U}_{1:r}\hat{U}_{1:r}^T$. For (71), we can decompose $\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\|$ into

$$\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\right\|\leq\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|+\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T-\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\right\|_{\mathrm{F}}\left\|\epsilon_i\right\|.$$

To upper bound $\|\hat{U}_{1:r}\hat{U}_{1:r}^T-\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\|_{\mathrm{F}}$, we are going to use Theorem 2.3. Since (66)-(67) hold, under the assumption $\beta n/k^4\geq 100$, we have

$$\frac{\lambda_r-\lambda_{r+1}}{\max\left\{\|E\|,\sqrt{\frac{k^2}{n\beta}}\lambda_{r+1}\right\}}\geq\frac{\tau}{2}.$$

Applying Theorem 2.3, we have

$$\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T-\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\right\|_{\mathrm{F}}\leq\frac{256\sqrt{rk}}{\sqrt{n\beta}}+\frac{256\left\|\hat{U}_{-i,1:r}\hat{U}_{-i.1:r}^T\epsilon_i\right\|}{\lambda_r}.$$

Hence,

$$\begin{aligned}\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\right\|&\leq\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|+\left(\frac{256\sqrt{rk}}{\sqrt{n\beta}}+\frac{256\left\|\hat{U}_{-i,1:r}\hat{U}_{-i.1:r}^T\epsilon_i\right\|}{\lambda_r}\right)\|E\|\\&=\frac{256k\|E\|}{\sqrt{n\beta}}+\left(1+\frac{256\|E\|}{\lambda_r}\right)\left\|\hat{U}_{-i,1:r}\hat{U}_{-i.1:r}^T\epsilon_i\right\|\\&\leq\frac{256\sqrt{2}k(\sqrt{n}+\sqrt{p})\sigma}{\sqrt{n\beta}}+\left(1+\frac{256\sqrt{2}(\sqrt{n}+\sqrt{p})\sigma}{\tau\sqrt{n+p}\sigma}\right)\left\|\hat{U}_{-i,1:r}\hat{U}_{-i.1:r}^T\epsilon_i\right\|\\&\leq 512k\beta^{-0.5}\left(1+\sqrt{\frac{p}{n}}\right)\sigma+\left(1+512\tau^{-1}\right)\left\|\hat{U}_{-i,1:r}\hat{U}_{-i.1:r}^T\epsilon_i\right\|,\end{aligned}$$

where in the second to the last inequality, we use (66) for λ_r and the event $\mathcal{F}$ for $\|E\|$. Then (71) leads to

$$A_{i,a}\mathbb{I}\{\mathcal{F}\} \leq \mathbb{I}\left\{\left(1 - c_3\tau'' - \frac{c_3 k^2 \tau \beta^{-\frac{1}{2}}\left(1+\sqrt{\frac{p}{n}}\right)\sigma}{\Delta}\right)\Delta \leq 2\left(1+512\tau^{-1}\right)\left\|\hat{U}_{-i,1:r}\hat{U}_{-i.1:r}^T\epsilon_i\right\|\right\}\mathbb{I}\{\mathcal{F}\}$$
$$\leq \mathbb{I}\left\{\left(1 - c_4\left(\frac{k^2 \tau \beta^{-\frac{1}{2}}\left(1+\sqrt{\frac{p}{n}}\right)\sigma}{\Delta} + \tau^{-1}\right)\right)\Delta \leq 2\left\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\right\|\right\},$$

where $c_3, c_4 > 0$ are some constants. As long as $1 - c_4(k^2\tau\beta^{-0.5}(1+\sqrt{p/n})\sigma/\Delta + \tau^{-1}) > 1/2$, we can use Lemma D.2 to calculate the tail probability of $\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\epsilon_i\|$. Following the proof of Theorem 3.1, we have

$$\mathbb{E}A_{i,a}\mathbb{I}\{\mathcal{F}\} \leq \exp\left(-\left(1 - c_5\left(\frac{k^2 \tau \beta^{-\frac{1}{2}}\left(1+\sqrt{\frac{p}{n}}\right)\sigma}{\Delta} + \tau^{-1}\right)\right)\frac{\Delta^2}{8\sigma^2}\right),$$

for some constant $c_5 > 0$. Then we have,

$$n^{-1}\sum_{i\in[n]}\sum_{a\in[k]:a\neq z_i^*}\mathbb{E}A_{i,a}\mathbb{I}\{\mathcal{F}\} \leq k\exp\left(-\left(1 - c_5\left(\frac{k^2 \tau \beta^{-\frac{1}{2}}\left(1+\sqrt{\frac{p}{n}}\right)\sigma}{\Delta} + \tau^{-1}\right)\right)\frac{\Delta^2}{8\sigma^2}\right).$$

(Obtaining the Final Result.) From (68) and the above upper bounds on $n^{-1}\sum_{i\in[n]}\sum_{a\in[k]:a\neq z_i^*}\mathbb{E}B_{i,a}\mathbb{I}\{\mathcal{F}\}$ and $n^{-1}\sum_{i\in[n]}\sum_{a\in[k]:a\neq z_i^*}\mathbb{E}A_{i,a}\mathbb{I}\{\mathcal{F}\}$, we have

$$\mathbb{E}\ell(\hat{z}, z^*) \leq e^{-0.08n} + 2\exp\left(-\frac{1}{2}\left(c_4\frac{\tau''\Delta}{k^{\frac{7}{2}}\tau^2\beta^{-\frac{1}{2}}(1+\frac{p}{n})\sigma}\sqrt{\frac{n-k}{3n}}\right)^2\frac{\Delta^2}{\sigma^2}\right) + nk\exp\left(-\frac{(n-k)}{9}\right)$$
$$+ k\exp\left(-\left(1 - c_5\left(\frac{k^2 \tau \beta^{-\frac{1}{2}}\left(1+\sqrt{\frac{p}{n}}\right)\sigma}{\Delta} + \tau^{-1}\right)\right)\frac{\Delta^2}{8\sigma^2}\right).$$

Since we assume $\beta n/k^4 \geq 100$, we have $(n-k)/n > 0.99$. Hence, under the assumption that $\Delta/(k^{3.5}\beta^{-0.5}(1+\frac{p}{n})\sigma) \to \infty$, we can take τ, τ'' to be

$$\tau = \tau''^{-1} := \left(\frac{\Delta}{k^{3.5}\beta^{-0.5}\left(1+\frac{p}{n}\right)\sigma}\right)^{0.25}$$

such that $\tau \to \infty$ and $\tau'' = o(1)$. Then for some constant $c_6 > 0$, we have

$$\mathbb{E}\ell(\hat{z}, z^*) \leq e^{-0.08n} + 2\exp\left(-\frac{c_4^2}{12}\left(\frac{\Delta}{k^{3.5}\beta^{-0.5}\left(1+\frac{p}{n}\right)\sigma}\right)^{0.5}\frac{\Delta^2}{\sigma^2}\right) + nke^{-0.1n}$$
$$+ k\exp\left(-\left(1 - 2c_5\left(\frac{\Delta}{k^{3.5}\beta^{-0.5}\left(1+\frac{p}{n}\right)\sigma}\right)^{-0.25}\right)\frac{\Delta^2}{8\sigma^2}\right)$$
$$\leq \exp\left(-\left(1 - c_6\left(\frac{\Delta}{k^{3.5}\beta^{-0.5}\left(1+\frac{p}{n}\right)\sigma}\right)^{-0.25}\right)\frac{\Delta^2}{8\sigma^2}\right) + 2e^{-0.08n}.$$

APPENDIX D: AUXILIARY LEMMAS AND PROPOSITIONS AND THEIR PROOFS

PROPOSITION D.1. *For Y and $\hat{Y}$ defined in (1), we have (2) holds assuming $\sigma_r - \sigma_{r+1} > 2\left\|(I - U_rU_r^T)y_n\right\|$.*

PROOF. Recall the definition of the augmented matrix Y'. Note that $U_rU_r^TY$ is the best rank-r approximation of Y. Since

$$\left\|\left(I - U_rU_r^T\right)Y'\right\|_{\mathrm{F}} = \left\|\left(\left(I - U_rU_r^T\right)Y, 0\right)\right\|_{\mathrm{F}} = \left\|\left(I - U_rU_r^T\right)Y\right\|_{\mathrm{F}},$$

we have $U_rU_r^TY'$ also being the best rank-r approximation of Y'. This proves that span(U_r) and $U_rU_r^T$ are also the leading r left singular subspace and projection matrix of Y'. Then $\hat{U}_r\hat{U}_r^T - U_rU_r^T$ is about the perturbation between $\hat{Y}$ and Y'.

Let σ_r', σ_{r+1}' be the rth and $(r+1)$th largest singular values of Y', respectively. By Wedin's Thereom (cf. Section 2.3 of [8]), if $\sigma_r' - \hat{\sigma}_{r+1} > 0$, then we have

$$\left\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\right\|_{\mathrm{F}} \le \frac{\left\|\hat{Y} - Y'\right\|_{\mathrm{F}}}{\sigma_r' - \hat{\sigma}_{r+1}} = \frac{\left\|(I - U_rU_r^T)y_n\right\|}{\sigma_r' - \hat{\sigma}_{r+1}}.$$

Regarding the values of σ_r' and σ_{r+1}', first we have $\sigma_r' \ge \sigma_r$. This is because

$$\sigma_r' = \inf_{x \in \mathrm{span}(U_r)} \left\|x^TY'\right\| = \inf_{x \in \mathrm{span}(U_r)} \left\|\left(x^TY, x^Ty_n\right)\right\| \ge \inf_{x \in \mathrm{span}(U_r)} \left\|x^TY\right\| \ge \sigma_r.$$

In addition, we have $\sigma_{r+1}' = \sigma_{r+1}$, due to the fact that $(I - U_rU_r^T)Y' = ((I - U_rU_r^T)Y, 0)$. By Weyl's inequality, we have

$$|\hat{\sigma}_{r+1} - \sigma_{r+1}'| \le \left\|Y - Y'\right\| = \left\|(I - U_rU_r^T)y_n\right\|.$$

Hence, if $\sigma_r - \sigma_{r+1} > 2\left\|(I - U_rU_r^T)y_n\right\|$ is further assumed, we have

$$\sigma_r' - \hat{\sigma}_{r+1} \ge \sigma_r - \sigma_{r+1} - \left\|(I - U_rU_r^T)y_n\right\| \ge \frac{1}{2}\left(\sigma_r - \sigma_{r+1}\right).$$

The proof is complete. ☐

LEMMA D.1. *Let $E = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^{p\times n}$ be a random matrix with each column $\epsilon_i \sim SG_p(\sigma^2), \forall i \in [n]$ independently. Then*

$$\mathbb{P}\left(\|E\| \ge 4t\sigma(\sqrt{n} + \sqrt{p})\right) \le \exp\left(-\frac{(t^2-3)n}{2}\right),$$

for any $t \ge 2$.

PROOF. We follow a standard ϵ-net argument. Let $\mathcal{U}$ and $\mathcal{V}$ be a 1/4 covering set of the unit sphere in $\mathbb{R}^p$ and in $\mathbb{R}^n$, respectively. That is, for any $u \in \mathbb{R}^p$ such that $\|u\| = 1$, there exists a $u' \in \mathcal{U}$ such that $\|u'\| = 1$ and $\|u - u'\| \le 1/4$. Similarly, for any $v \in \mathbb{R}^n$ such that $\|v\| = 1$, there exists a $v' \in \mathcal{V}$ such that $\|v'\| = 1$ and $\|v - v'\| \le 1/4$. Then

$$\begin{aligned}\left|u^TEv\right| &= \left|u'^TEv' + u'^TE(v - v') + (u - u')^TEv' + (u - u')^TE(v - v')\right| \\ &\le \left|u'^TEv'\right| + \left|u'^TE(v - v')\right| + \left|(u - u')^TEv'\right| + \left|(u - u')^TE(v - v')\right|.\end{aligned}$$

Maximizing over u, v on both sides, we have

$$\|E\| = \max_{u\in\mathbb{R}^p, v\in\mathbb{R}^n: \|u\|=\|v\|=1} \left|u^TEv\right| \le \max_{u'\in\mathcal{U}, v'\in\mathcal{V}} \left|u'^TEv'\right| + \frac{1}{4}\|E\| + \frac{1}{4}\|E\| + \frac{1}{16}\|E\|.$$

Hence,

$$\|E\| \le 4 \max_{u' \in \mathcal{U}, v' \in \mathcal{V}} \left| u'^T E v' \right|.$$

For any $u' \in \mathcal{U}, v' \in \mathcal{V}$, we have each $u'^T \epsilon_i$ being an independent SG(σ^2) and then $u'^T E v' \sim$ SG(σ^2). Note $|U| \le 9^p \le e^{3p}$ and similarly $|V| \le e^{3n}$. Then by the tail probability of sub-Gaussian random variable and by the union bound, we have

$$\begin{aligned}
\mathbb{P}\left(\|E\| \le 4t\sigma(\sqrt{n}+\sqrt{p})\right) &\le \mathbb{P}\left(\max_{u' \in \mathcal{U}, v' \in \mathcal{V}} \left|u'^T E v'\right| \le t\sigma(\sqrt{n}+\sqrt{p})\right) \\
&\le |U|\,|V| \exp\left(-\frac{t^2\left(\sqrt{n}+\sqrt{p}\right)^2}{2}\right) \\
&\le \exp\left(-\frac{(t^2-3)n}{2}\right),
\end{aligned}$$

for any $t \ge 2$. $\square$

LEMMA D.2. *Let $X \sim SG_d(\sigma^2)$. Consider any $k \le d$. For any matrix $U = (u_1, \dots, u_k) \in \mathbb{R}^{d \times k}$ that is independent of X and is with orthogonal columns $\{u_i\}_{i \in [k]}$. We have*

$$\mathbb{P}\left(\left\|UU^T X\right\|^2 \ge \sigma^2(k + 2\sqrt{kt} + 2t)\right) \le e^{-t}.$$

PROOF. Note that $\mathrm{tr}(UU^T) = \mathrm{tr}((UU^T)^2) = k$ and $\left\|UU^T\right\| = 1$. This is a direct consequence of Theorem 1 in [16] for concentration of quadratic forms of sub-Gaussian random vectors. $\square$

PROOF OF PROPOSITION 3.1. Define $\hat{P} = \sum_{i \in [r]} \hat{\lambda}_i \hat{u}_i \hat{v}_i^T$. Due to the fact that $\hat{P}$ is the best rank-r approximation of X in spectral norm and P is rank-κ, under the assumption that $\kappa \le r$, we have that

$$\left\|\hat{P} - X\right\| \le \|P - X\| = \|E\|.$$

Since $r \le k$ is assumed, the rank of $\hat{P} - P$ his at most $2k$, and we have

$$\left\|\hat{P} - P\right\|_{\mathrm{F}} \le \sqrt{2k}\left\|\hat{P} - P\right\| \le \sqrt{2k}\left(\left\|\hat{P} - X\right\| + \|P - X\|\right) \le 2\sqrt{2k}\,\|E\| \tag{72}$$

Now, denote $\hat{\Theta} := (\hat{\theta}_{\hat{z}_1}, \hat{\theta}_{\hat{z}_2}, \dots, \hat{\theta}_{\hat{z}_n})$. Since $\hat{\Theta}$ is the solution to the k-means objective (14), we have that

$$\left\|\hat{\Theta} - \hat{P}\right\|_{\mathrm{F}} \le \left\|P - \hat{P}\right\|_{\mathrm{F}}.$$

Hence, by the triangle inequality, we obtain that

$$\left\|\hat{\Theta} - P\right\|_{\mathrm{F}} \le 2\left\|\hat{P} - P\right\|_{\mathrm{F}} \le 4\sqrt{2k}\,\|E\|.$$

Now, define the set S as

$$S = \left\{ i \in [n] : \left\|\hat{\theta}_{\hat{z}_i} - \theta^*_{z_i^*}\right\| > \frac{\Delta}{2} \right\}.$$

Since $\left\{\hat{\theta}_{\hat{z}_i} - \theta^*_{z_i^*}\right\}_{i\in[n]}$ are exactly the columns of $\hat{\Theta} - P$, we have that

$$|S| \leq \frac{\left\|\hat{\Theta} - P\right\|_{\mathrm{F}}^2}{(\Delta/2)^2} \leq \frac{128k\|E\|^2}{\Delta^2}.$$

Under the assumption (15) we have

$$\frac{\beta\Delta^2 n}{k^2\|E\|^2} \geq 256,$$

which implies

$$|S| \leq \frac{\beta n}{2k}.$$

We now show that all the data points in S^C are correctly clustered. We define

$$C_j = \left\{i \in [n] : z_i^* = j, i \in S^C\right\},\ j \in [k].$$

The following holds:

- For each $j \in [k]$, C_j cannot be empty, as $|C_j| \geq |\{i : z_i^* = j\}| - |S| > 0$.
- For each pair $j, l \in [k], j \neq l$, there cannot exist some $i \in C_j, i' \in C_l$ such that $\hat{z}_i = \hat{z}_{i'}$. Otherwise $\hat{\theta}_{\hat{z}_i} = \hat{\theta}_{\hat{z}_{i'}}$ which would imply

$$\begin{aligned}\left\|\theta_j^* - \theta_l^*\right\| &= \left\|\theta^*_{z_i^*} - \theta^*_{z_{i'}^*}\right\| \\ &\leq \left\|\theta^*_{z_i^*} - \hat{\theta}_{\hat{z}_i}\right\| + \left\|\hat{\theta}_{\hat{z}_i} - \hat{\theta}_{\hat{z}_{i'}}\right\| + \left\|\hat{\theta}_{\hat{z}_{i'}} - \theta^*_{z_{i'}^*}\right\| < \Delta,\end{aligned}$$

  contradicting with the definition of Δ.

Since $\hat{z}_i$ can only take values in $[k]$, we conclude that the sets $\{\hat{z}_i : i \in C_j\}$ are disjoint for all $j \in [k]$. That is, there exists a permutation $\phi \in \Phi$, such that

$$\hat{z}_i = \phi(j),\ i \in C_j,\ j \in [k].$$

This implies that $\sum_{i\in S^C} \mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\} = 0$. Hence, we obtain that

$$|\{i \in [n] : \hat{z}_i \neq \phi(z_i^*)\}| \leq |S| \leq \frac{128k\|E\|^2}{\Delta^2}.$$

Since $|S| \leq \frac{\beta n}{2k}$ (which means $\ell(\hat{z}, z^*) \leq \frac{\beta n}{2k}$ from the above display), for any $\psi \in \Phi$ such that $\psi \neq \phi$, we have $|\{i \in [n] : \hat{z}_i \neq \psi(z_i^*)\}| \geq 2\beta n/k - |S| \geq \beta n/k$. As a result, we have

$$\ell(\hat{z}, z^*) = \frac{1}{n}|\{i \in [n] : \hat{z}_i \neq \phi(z_i^*)\}| \leq \frac{128k\|E\|^2}{n\Delta^2}.$$

Moreover, for each $a \in [k]$, we have

$$\left\|\hat{\theta}_{\phi(a)} - \theta_a^*\right\|^2 \leq \frac{\left\|\hat{\Theta} - P\right\|_{\mathrm{F}}^2}{|\{i \in [n] : \hat{z}_i = \phi(a), z_i^* = a\}|} \leq \frac{\left\|\hat{\Theta} - P\right\|_{\mathrm{F}}^2}{\frac{\beta n}{k} - |S|} \leq \frac{64k^2\|E\|^2}{\beta n}$$

□