The Annals of Statistics
2024, Vol. 52, No. 5, 2004–2033
https://doi.org/10.1214/24-AOS2418

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

BY ANDERSON Y. ZHANG[1,a] AND HARRISON Y. ZHOU[2,b]

[1]*Department of Statistics and Data Science, University of Pennsylvania,* [a]*ayz@wharton.upenn.edu*
[2]*Department of Statistics and Data Science, Yale University,* [b]*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 the 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. Empowered by this leave-one-out perturbation theory, we provide a deterministic entrywise analysis for the performance of spectral clustering under mixture models. Our analysis leads to an explicit exponential error rate for spectral 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 [4, 37] is a central topic in probability and statistics. It plays a fundamental role in spectral methods [11, 19], an umbrella term for algorithms involving eigendecomposition or singular value decomposition. It has a wide range of applications including principal component analysis [1, 8], covariance matrix estimation [15], clustering [30, 34, 35, 41], and matrix completion [14, 28], throughout different fields such as machine learning [5], network science [2, 32], and genomics [20].

Perturbation analysis for eigenspaces and singular subspaces dates back to seminal works of Davis and Kahan [12] and Wedin [44]. 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, 15] and $\ell_{2,\infty}$ analysis for singular subspaces [3, 7, 10, 25], 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

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

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.

Received May 2022; revised January 2024.
MSC2020 subject classifications. 62H30.
Key words 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 $\mathrm{span}(U_r)$ and $\mathrm{span}(\hat{U}_r)$. 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 $U_rU_r^T y_n$ 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 $U_r^T y_n$, informally (see Theorem 2.1 for a precise statement),

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

When this factor is smaller than some constant, it results in a sharper upper bound than Wedin's theorem. The derived upper bound is particularly suitable for mixture models where the contributions of $U_r^T y_n$ 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 and then applies a classical clustering method, such as k-means, to the data of reduced dimension. It is computationally appealing and often delivers 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, 17, 18, 23, 24, 31, 33, 34, 47]. 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 [1, 6, 13, 26, 30, 36, 43]. For isotropic Gaussian mixture models, [26] shows spectral clustering achieves the optimal minimax rate. However, the proof technique used in [26] 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 of 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 $\hat{U}_{1:r}^T X_i$ 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. Building upon our leave-one-out perturbation theory, we provide a deterministic entrywise analysis for the spectral clustering. We demonstrate that the correctness of X_i's clustering 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 nonzero singular value is away from zero. Compared to [1], our assumptions on the spectrum and geometric distribution of the centers are weaker. In addition, we obtain an explicit constant $1/8$ in the exponent, which is sharp when the noises are further assumed to be isotropic Gaussian. To remove the spectral gap condition, we 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 [26]. Empowered by the leave-one-out perturbation theory, our proof adopts a completely different approach and is much shorter compared to that of [26]. In addition, the signal-to-noise ratio condition of [26] 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 suboptimal and is optimal only if each coordinate of ϵ_i is normally distributed.

Organization. The structure of this paper is 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. We discuss extensions and potential caveats of our analysis in Section 4. The proofs of main results in Section 2 and Section 3 are given in Section 5 and in Section 6, respectively. All other proofs can be found in the Supplementary Material [46].

Notation. For any positive integer r, let $[r] = \{1, 2, \dots, r\}$. 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 $x_1, \dots, x_d$, we denote $\mathrm{diag}(x_1, \dots, x_d)$ to be a $d \times d$ diagonal matrix with diagonal entries being $x_1, \dots, x_d$. 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 $\mathbb{O}^{d\times p} = \{V \in \mathbb{R}^{d\times p} : V^T V = I\}$ be the set of matrices in $\mathbb{R}^{d\times p}$ with orthonormal columns. We denote $\mathbb{I}\{\cdot\}$ 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 $a_n \leq C b_n$ 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 $\mathbb{E}e^{tX} \leq \exp(\sigma^2 t^2/2)$ 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 $u^T X \sim \mathrm{SG}(\sigma^2)$ for any unit vector $u \in \mathbb{R}^d$.

2. Leave-one-out singular subspace perturbation analysis. Classical singular subspace perturbation theory examines the relationship between the singular spaces of two matrices of the same dimension. However, prevailing upper bounds, such as those given by Wedin's theorem, often achieve tightness only in worst-case scenarios. They can be suboptimal, especially in situations like the one considered in this paper where one matrix is short of one column relative to the other.

In the domains of statistics and data science, it is common to work with data matrices wherein columns represent independent and identically distributed observations. Intuitively, when the number of observations is large, omitting a single observation should have minimal impact on the singular subspace. This intuition can guide entrywise perturbation analyses for spectral methods. As a case in point, the efficacy of spectral clustering under mixture models can largely be attributed to the perturbation of $\hat{U}_{1:r}^T X_i$, where X_i represents the ith observation and $\hat{U}_{1:r}$ encompasses the leading r left singular vectors of the data matrix. Directly analyzing $\hat{U}_{1:r}^T X_i$ is cumbersome due to the inherent dependence between $\hat{U}_{1:r}$ and

X_i. To disentangle this dependence, a logical strategy is to substitute $\hat{U}_{1:r}$ with its leave-one-out counterpart, $\hat{U}_{-i,1:r}$, which is formed using all observations except X_i. The resulting independence between $\hat{U}_{-i,1:r}$ and ϵ_i facilitates a more precise characterization of the tail probabilities of $\hat{U}_{-i,1:r}^T X_i$. This, in turn, yields a more defined bound on spectral clustering's performance. Such an analytical approach presumes that $\hat{U}_{1:r}$ and its leave-one-out version $\hat{U}_{-i,1:r}$ are sufficiently similar.

With this foundation laid, in this section, we focus on establishing a comprehensive leave-one-out perturbation theory for singular subspaces.

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

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

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

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

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

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

with $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r \geq 0$ being the r singular values of $\hat{U}_r^T U_r$. It is known (see Lemma 1 of [9]) that $\|\hat{U}_r\hat{U}_r^T - U_r U_r^T\|_{\mathrm{F}} = \sqrt{2}\|\sin\Theta(\hat{U}_r, U_r)\|_{\mathrm{F}}$. Throughout this section, we will focus on establishing sharp upper bounds for $\|\hat{U}_r\hat{U}_r^T - U_r U_r^T\|_{\mathrm{F}}$, the Frobenius norm of the difference between two corresponding projection matrices $U_r U_r^T$ and $\hat{U}_r\hat{U}_r^T$.

Since the augmented matrix $Y' := (Y, U_r U_r^T y_n) \in \mathbb{R}^{p\times n}$ concatenated by Y and $U_r U_r^T y_n$ has the same leading r left singular subspace and projection matrix as Y, a natural idea is to relate $\|\hat{U}_r\hat{U}_r^T - U_r U_r^T\|_{\mathrm{F}}$ with the difference $\hat{Y} - Y'$. The classical spectral perturbation theory such as Wedin's theorem [9, 45] leads to that if $\sigma_r - \sigma_{r+1} > 2\|(I - U_r U_r^T)y_n\|$, then

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

See Proposition E.1 in the Supplementary Material 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 $\|\hat{Y} - Y'\| = \|(I - U_r U_r^T)y_n\|$ 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 Theorem 2.1, we provide a deeper and finer analysis for $\|\hat{U}_r\hat{U}_r^T - U_r U_r^T\|_{\mathrm{F}}$, utilizing the fact that $\hat{Y}$ and Y differ by only one column and exploiting the interplay between $U_r U_r^T y_n$ and $(I - U_r U_r^T)y_n$.

THEOREM 2.1. *If*

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

we have

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

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

$$\|\hat{U}_r\hat{U}_r^T - U_rU_r^T\|_{\mathrm{F}} \le \frac{4\sqrt{2}}{\rho}\frac{\|U_rU_r^T y_n\|}{\sigma_r}. \tag{5}$$

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 5.1 in Section 5.1 for details.

We are ready to have a comparison of our result (4) and (2) from Wedin's theorem. Under the assumption (3), the upper bound in (2) can be written equivalently as $2\sqrt{2}\rho^{-1}$. 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 $1/2$, 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, $\hat{U}_r\hat{U}_r^T = U_rU_r^T$. On the contrary, (2) gives a nonzero bound $2\sqrt{2}/\rho^{-1}$. 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 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 $2\sqrt{2}\|y_n\|/\sigma_1$.

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 $2\sqrt{2}\rho^{-1}$.

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

Regarding the sharpness of the bound in Theorem 2.1, it's worth noting that in Example 1 above, our theorem accurately derives an upper bound of 0, showcasing its optimality in that specific context. To further demonstrate the optimality of our theorem, consider a more intricate example.

EXAMPLE 3. Consider a rank-one matrix $Y = \mathbb{1}_p\mathbb{1}_{n-1}^T$ where $\sigma_1 = \sqrt{(n-1)p}$ and $u_1 = p^{-1/2}\mathbb{1}_p$. Now, define $y_n = \mathbb{1}_p + sw$, wherein s represents a scalar and w is a unit vector orthogonal to $\mathbb{1}_p$. This means that y_n matches each column of Y for $s = 0$ and introduces an orthogonal perturbation for $s \neq 0$. Given that $\rho = \sigma_1/s = \sqrt{(n-1)p}/s$ and $u_1^T y_n = \sqrt{p}$, it follows from Theorem 2.1 that $\|\hat{u}_1\hat{u}_1^T - u_1u_1^T\|_{\mathrm{F}} \le 4\sqrt{2}s/((n-1)\sqrt{p})$. Since $\hat{Y}$ is of rank-two, we can express $\hat{u}_1$ as $\hat{u}_1 = \sqrt{1-\alpha^2}u_1 + \alpha w$ where $|\alpha| \le 1$. Note that $\hat{u}_1^T\hat{Y} =$

$(\sqrt{(1-\alpha^2)p}\mathbb{1}_{n-1}^T, \sqrt{(1-\alpha^2)p}+\alpha s)$ and $\|\hat{u}_1^T\hat{Y}\|^2 = (1-\alpha^2)np + \alpha^2 s^2 + 2\sqrt{(1-\alpha^2)p}\alpha s$. For small s, we can approximate α (by maximizing $\|\hat{u}_1^T\hat{Y}\|^2$ over α) as $s/(n\sqrt{p})$. Since α is also small, we have $\|\hat{u}_1\hat{u}_1^T - u_1u_1^T\|_F \approx \alpha\sqrt{1-\alpha^2}\|u_1w^T + w^Tu_1\|_F = \sqrt{2}\alpha\sqrt{1-\alpha^2} \approx \sqrt{2}s/(n\sqrt{p})$. A comparison with the upper bound deduced from Theorem 2.1 underscores that the theorem captures the correct rate $s/(n\sqrt{p})$, albeit with a multiplicative constant.

However, the sharpness of Theorem 2.1 in diverse settings or under different conditions remains an area needing further investigation.

The leave-one-out singular subspace perturbation analysis established in this paper shares conceptual similarities with the leave-one-out technique grounded in random matrix theory and used in the ℓ_∞ or $\ell_{2,\infty}$ perturbation analysis [2, 11]. On a high level, for a matrix X with an eigenvector u, the goal of the ℓ_∞ analysis is to derive an upper bound for $\|u\|_\infty = \max_i |u_i|$, where $\{u_i\}$ represents the coordinates of u. To aid in this task, the leave-one-out technique introduces an auxiliary matrix, formed by excluding the ith column, X_i, of X, and the corresponding eigenvectors u_{-i}. It approximates u_i by a quantity involving both X_i and u_{-i}, leveraging the independence between them. Our approach aligns with this principle but subsequent analysis distinctly sets it apart. While both methods involve the difference between u and u_{-i}, the ℓ_∞ analysis predominantly uses it as a stepping stone towards $\|u\|_\infty$, dealing with it by a direct application of Wedin's theorem. In contrast, our methodology focuses on establishing a sharp bound for this difference. This distinction enables us to characterize the tail probabilities of u_i rather than just a general ℓ_∞ bound and paves the way for a more fine-grained investigation into the performance of spectral methods.

We conclude this section by mentioning that our current analytical framework might extend to scenarios wherein a matrix has multiple columns left out relative to another. Intuitively, as columns can be removed sequentially, Theorem 2.1 (or its more concise variant, (5)) can be invoked in a successive manner. This iterative application can provide an upper bound on the discrepancy between the two singular subspaces in question. A more intricate way to consider would be a direct extension of the proof of Theorem 2.1. Given that this theorem fundamentally revolves around the dynamics between $U_rU_r^Ty_n$ and $(I-U_rU_r^T)y_n$, its generalization is likely to encompass similar, yet more expansive, interactions.

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

$$X_i = \theta_{z_i^*}^* + \epsilon_i, \tag{6}$$

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

$$X = P + E, \tag{7}$$

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,

$$X_{-i} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \in \mathbb{R}^{p \times (n-1)}. \tag{8}$$

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 $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_{p \wedge n}$ and $\hat{\lambda}_{-i,1} \geq \hat{\lambda}_{-i,2} \geq \cdots \geq \hat{\lambda}_{-i,p \wedge (n-1)}$. Note that the signal matrix P is at most rank-k. Then for any $r \in [k]$, define

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

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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p \wedge n}$ 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*

$$\rho_0 := \frac{\lambda_\kappa}{\|E\|} > 16. \tag{9}$$

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

$$\|\hat{U}_{1:\kappa}\hat{U}_{1:\kappa}^T - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\|_{\mathrm{F}} \leq \frac{128}{\rho_0}\left(\sqrt{\frac{k\kappa}{\beta n}} + \frac{\|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T \epsilon_i\|}{\lambda_\kappa}\right). \tag{10}$$

Theorem 2.2 leverages the mixture model structure (6) that the signal matrix P has only k unique columns with each appearing at least $\beta n/k$ times. The assumption $\beta n/k^2 \geq 10$ helps ensure that spectrum and singular vectors of P do not change significantly if any column of P is removed. We require the condition (9) 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 (10) is a product of ρ_0^{-1} and a sum 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 (10) 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 of order 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 $\|\hat{U}_{1:r}\hat{U}_{1:r}^T - \hat{U}_{-i,1:r} \times \hat{U}_{-i,1:r}^T\|_{\mathrm{F}}$ 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*

$$\tilde{\rho}_0 := \frac{\lambda_r - \lambda_{r+1}}{\max\{\|E\|, \sqrt{\frac{k^2}{\beta n}}\lambda_{r+1}\}} > 16. \tag{11}$$

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

$$\|\hat{U}_{1:r}\hat{U}_{1:r}^T - \hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T\|_{\mathrm{F}} \leq \frac{128}{\tilde{\rho}_0}\left(\frac{\sqrt{kr}}{\sqrt{\beta n}} + \frac{\|\hat{U}_{-i,1:r}\hat{U}_{-i,1:r}^T \epsilon_i\|}{\lambda_r}\right). \tag{12}$$

In Theorem 2.3, $r \in [k]$ is any number such that (11) is satisfied. When r is chosen to be κ, (11) is reduced to (9), and (12) leads to the same upper bound as (10). When $r < \kappa$, λ_{r+1} is nonzero and in (11) 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 (11) 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 (7) can be written equivalently as

$$X = P' + E', \quad \text{where } P' := U_{1:r}U_{1:r}^T P \text{ and } E' := E + U_{(r+1):\kappa}U_{(r+1):\kappa}^T P. \tag{13}$$

Since it is still a mixture model, Theorem 2.2 can be applied. Nevertheless, the condition (9) essentially requires $\lambda_r/(\|E\| + \lambda_{r+1}) > 16$ as $\|E'\| \le \|E\| + \|U_{(r+1):\kappa}U_{(r+1):\kappa}^T P\| = \|E\| + \lambda_{r+1}$, which is stronger than the condition (11). In order to weaken the requirement on the spectral gap into (11), 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.

Theorem 2.2 and Theorem 2.3 require $\beta n/k^2$ be sufficiently large. Further in the paper, results such as Lemma 3.3 need an even stronger condition wherein $\beta n/k^4$ should be large. We acknowledge that these dependencies on k appear nonoptimal. The current formulations stem from challenges faced during our analysis, resulting in these inherent dependencies. We hope to explore more optimal dependency in future research.

3. Spectral clustering for mixture models.

3.1. *Spectral clustering and polynomial error rate.* Recall the definition of the mixture model in (6) and also in (7). 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 $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_{p\wedge n} \ge 0$ 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 $\hat{U}_{1:r}^T X.$,

$$(\hat{z}, \{\hat{c}_j\}_{j\in[k]}) = \operatorname*{argmin}_{z\in[k]^n, \{c_j\}_{j\in[k]}\in\mathbb{R}^r} \sum_{i\in[n]} \|\hat{U}_{1:r}^T X_i - c_{z_i}\|^2. \tag{14}$$

In (14), 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 $\hat{U}_{1:r}\hat{U}_{1:r}^T X \in \mathbb{R}^{p\times n}$, define $\hat{\theta}_a = \hat{U}_{1:r}\hat{c}_a$ for each $a \in [k]$. It can be shown that (see Lemma 4.1 of [26])

$$(\hat{z}, \{\hat{\theta}_j\}_{j\in[k]}) = \operatorname*{argmin}_{z\in[k]^n, \{\theta_j\}_{j\in[k]}\in\mathbb{R}^p} \sum_{i\in[n]} \|\hat{U}_{1:r}\hat{U}_{1:r}^T X_i - \theta_{z_i}\|^2, \tag{15}$$

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 (15).

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 only 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 among the centers. Define Δ to be the minimum distance among centers,

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

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

In Proposition 3.1, a rough upper bound is provided on the misclustering error $\ell(\hat{z}, z^*)$ that takes a polynomial expression (17). Notably, 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 (7) and the separation condition (16) is satisfied. In addition, it requires no spectral gap condition. Proposition 3.1 is essentially an extension of Lemma 4.2 in [26] which is only for the Gaussian mixture model and needs $r = k$. We include its proof in Appendix E for completeness. Recall κ denotes 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*

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

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

$$\ell(\hat{z}, z^*) = \frac{1}{n}\big|\{i \in [n] : \hat{z}_i \neq \phi(z_i^*)\}\big| \le \frac{C_0 k \|E\|^2}{n\Delta^2}, \tag{17}$$

and

$$\max_{a\in[k]} \|\hat{\theta}_{\phi(a)} - \theta_a^*\| \le C_0 \beta^{-0.5} k n^{-0.5} \|E\|, \tag{18}$$

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 $\hat{z}$ beyond the polynomial rate stated in (17), 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 $n\ell(\hat{z}, z^*) = |\{i \in [n] : \hat{z}_i \neq \phi(z_i^*)\}| \leq n\beta/(2k)$. 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 3.1 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 $\kappa \leq r \leq k$. Assume (16) 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 $i \in [n]$,*

$$\mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\} \leq \mathbb{I}\{(1 - C\psi_0^{-1})\Delta \leq 2\|\hat{U}_{1:r}\hat{U}_{1:r}^T\epsilon_i\|\}. \tag{19}$$

To understand Lemma 3.1, recall that in (15) $\hat{z}$ 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.

While Lemma 3.1 lays foundational understanding, it alone is not sufficient for deriving 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 (19) 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

$$\|\hat{U}_{1:\kappa}\hat{U}_{1:\kappa}^T\epsilon_i\| \leq \|\hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\epsilon_i\| + \|\hat{U}_{1:\kappa}\hat{U}_{1:\kappa}^T - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\|_{\mathrm{F}}\|\epsilon_i\|. \tag{20}$$

The perturbation $\|\hat{U}_{1:\kappa}\hat{U}_{1:\kappa}^T - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T\|_{\mathrm{F}}$ is well controlled by Theorem 2.2, which shows the second term on the RHS of the above display is essentially $O(\rho_0^{-2})\|\hat{U}_{-i,1:\kappa} \times \hat{U}_{-i,1:\kappa}^T\epsilon_i\|$. 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 $\beta n/k^2 \geq 10$, (9), and (16) 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 $i \in [n]$,*

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

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

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

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 (9) is a spectral gap condition on the smallest nonzero 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 discussed in subsequent 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 (20) needs to be much smaller than Δ. This essentially requires $\max_{i\in[n]}\|\epsilon_i\|^2 \lesssim \lambda_\kappa\Delta$, in addition to (9) and (16). 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 $p \lesssim n$.

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*

$$\psi_1 := \frac{\Delta}{\beta^{-0.5}k(1+\sqrt{\frac{p}{n}})\sigma} > C \tag{21}$$

and

$$\rho_1 := \frac{\lambda_\kappa}{(\sqrt{n}+\sqrt{p})\sigma} > C, \tag{22}$$

we have

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

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 (see Lemma E.1). Under this event, (21) and (22) are sufficient conditions for (9) and (16), 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 high-probability event on $\|E\|$, (17) 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 $\max_{i\in[n]}\|\epsilon_i\|^2 \lesssim \lambda_\kappa\Delta$ 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 Δ/σ, k, β 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 (21) and (22), we immediately obtain 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*

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

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

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 further requires $\bar{\lambda}I\prec G\prec c\bar{\lambda}I$ for some constant $c>1$, all k eigenvalues of G are of 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 of 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 of the same order as long as the smallest one satisfies (22), 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 of the same order as long as the smallest distance satisfies (21). Hence, our conditions are more general than those in [1].

The spectral gap condition (22) 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 (e.g., the model considered in Section 3.6), the signal matrix P is rank-one. Hence, (22) is also no longer needed as it can be directly implied from (21).

3.4. *Spectral clustering with adaptive dimension reduction.* The theoretical analysis for the spectral clustering $\hat{z}$ of Algorithm 1 presented 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

Algorithm 2: Spectral Clustering with Adaptive Dimension Reduction

Input: Data matrix $X = (X_1, \ldots, X_n) \in \mathbb{R}^{p\times n}$, 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 two neighboring singular values is greater than T, ,

$$\hat{r} = \max\{a \in [k] : \hat{\lambda}_a - \hat{\lambda}_{a+1} \geq T\}. \tag{23}$$

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

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

$$(\tilde{z}, \{\tilde{c}_j\}_{j=1}^k) = \underset{z\in[k]^n, \{c_j\}_{j=1}^k \in \mathbb{R}^{\hat{r}}}{\operatorname{argmin}} \sum_{i\in[n]} \|\hat{U}_{1:\hat{r}}^T X_i - c_{z_i}\|^2. \tag{24}$$

motivates us to propose the following algorithm where the number of singular vectors used is carefully picked.

Algorithm 2 is a variant of Algorithm 1 with the number of singular vectors selected by (23), 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 criterion in (23) for choosing $\hat{r}$ has two purposes. Firstly, it ensures the presence of a desirable spectral gap. More crucially, it is intended to encompass important singular vectors while disregarding those that are noisy or of lesser relevance. This is illuminated by an implication from (23) that $\hat{\lambda}_{\hat{r}+1} \leq \hat{\lambda}_{k+1} + kT$ and that the significance of a singular vector can be characterized by the magnitude of its associated singular value. To illustrate this further, let us compare our approach with an alternative selection mechanism that simply choose an arbitrary index from $\{a \in [k] : \hat{\lambda}_a - \hat{\lambda}_{a+1} \geq T\}$ instead of the largest one. While such a criterion would indeed ensure a spectral gap, it is possible that $\hat{\lambda}_{\hat{r}+1}$ and subsequent singular values remain large, suggesting that the corresponding singular vectors are of importance. Omission of these pivotal vectors from the clustering algorithm would result in a decline in its performance.

The choice of the threshold T is crucial. 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 of constant order. It is allowed to grow but not faster than $\tilde{\phi}_0$ defined in (25). The precise description of the choices of T needed is given below in Lemma 3.3, which provides an entrywise analysis of $\tilde{z}$ 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*

$$\tilde{\psi}_0 := \frac{\Delta}{\beta^{-0.5}k^2 n^{-0.5}\|E\|} \tag{25}$$

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

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

Consequently, we have

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

With a proper choice of the threshold T, Lemma 3.3 only poses requirements on the smallest cluster size $\beta n/k$ and the minimum separation among the centers Δ. Compared to Lemma 3.2 and Theorem 3.1, it removes any condition on the smallest nonzero singular value such as (9) or (22). In addition, it requires no knowledge on the rank κ. Note that under the conditions of Lemma 3.3, $\hat{r}$ defined in (23) always exists (see Lemma B.1).

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 $\beta n/k^4 \ge 400$. There exist constants $C, C', C_1, C_2 > 0$ such that under the assumption that*

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

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

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

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

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

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 (22). 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 $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_p)$ for each $i \in [n]$. Assume $\beta n/k^4 \ge 100$ and*

$$\frac{\Delta}{k^{3.5}\beta^{-0.5}(1 + \frac{p}{n})\sigma} \to \infty. \tag{26}$$

We have

$$\mathbb{E}\ell(\hat{z}, z^*) \le \exp\left(-\left(1 - C\left(\frac{\Delta}{k^{3.5}\beta^{-0.5}(1 + \frac{p}{n})\sigma}\right)^{-0.25}\right)\frac{\Delta^2}{8\sigma^2}\right) + 2e^{-0.08n}, \tag{27}$$

where $C > 0$ is some constant.

Theorem 3.3 shows that asymptotically $\mathbb{E}\ell(\hat{z}, z^*) \le \exp(-(1 - o(1))\Delta^2/(8\sigma^2)) + 2\exp(-0.08n)$ 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 [27]: $\inf_{\hat{z}} \sup_{(\theta_1^*,\ldots,\theta_k^*),z^*} \mathbb{E}\ell(\hat{z}, z^*) \geq \exp(-(1+o(1))\Delta^2/(8\sigma^2))$ 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 (9) or (22). The purpose of such conditions is to ensure that singular vectors of X are well controlled, especially those corresponding to small nonzero 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 [26] 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 sufficiently 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 (27).

The performance of Algorithm 1 with $r = k$ under the same isotropic Gaussian mixture model is the main topic of [26] 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 [26] is spectral operator perturbation theory of [21, 22] 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 [26] 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 (26), while [26] needs $k^{10.5}$ instead which is a stronger condition.

3.6. *Lower bounds and suboptimality 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. When the noise is the isotropic Gaussian, Theorem 3.3 matches with the minimax rate assuming (26) 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 where 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 (6) with two clusters such that

$$\theta_1^* = -\theta_2^* = \delta\mathbb{1}_p, \quad \text{and} \quad \{\epsilon_{i,j}\}_{i\in[n], j\in[p]} \overset{\text{iid}}{\sim} F, \tag{28}$$

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 (28), we have $k = 2$, $\Delta = 2\sqrt{p}\delta$ and the largest singular value $\lambda_1 = \delta\sqrt{np}$. Since the signal 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

$$(\check{z}, \{\check{c}_j\}_{j=1}^2) = \operatorname*{argmin}_{z\in[2]^n, \{c_j\}_{j=1}^2\in\mathbb{R}} \sum_{i\in[n]} (\hat{u}_1^T X_i - c_{z_i})^2. \tag{29}$$

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 $\{\operatorname{sign}(\hat{v}_{1,i})\}_{i\in[n]}$, 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 (28), assume that F is a $\mathrm{SG}(\sigma^2)$ distribution with zero mean and $\beta n > 40$. There exist constants $C, C' > 0$ such that under the assumption that

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

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

The special structure of (28) makes it possible to derive a sharper upper bound than the one above 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 (28) enables us to have a lower bound for $\mathbb{I}\{\hat{z}_i \ne \phi(z_i^*)\}$ that is in an opposite direction of Lemma 3.2. See Lemma D.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* (28). *For any* $\xi \sim F$, *assume* $\mathbb{E}\xi = 0$, $Var(\xi) = \bar{\sigma}^2$, *and* $\xi \sim SG(\sigma^2)$ *where* $\sigma \le C\bar{\sigma}$ *for some constant* $C > 0$. *Assume* $\beta n > 40$. *Then there exist constants* $C', C'', C''' > 0$ *such that if* $\psi_3 \ge C'$, *we have*

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

and

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

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 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

$$\mathbb{E}\ell(\check{z}, z^*) = \exp\left(-\frac{(1+c)\Delta^2}{8\bar{\sigma}^2}\right), \tag{30}$$

for some small constant c.

The upper and lower bounds in Theorem 3.4 give a sharp characterization of 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 (28). Since the model (28) is essentially about a testing between two parametric distributions, the optimal procedure is the likelihood

ratio test. According to the classical asymptotics theory [39], the likelihood ratio behaves like a normal random variable as $p \to \infty$ under some regularity conditions. This leads to an error rate determined by Δ and the Fisher information.

LEMMA 3.4. *Consider the model* (28). *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*

$$C_1 \exp\left(-\frac{\Delta^2}{8\mathcal{I}^{-1}}\right) \le \lim_{p\to\infty} \inf_{z} \sup_{z^* \in [2]^n} \mathbb{E}\ell(z, z^*) \le C_2 \exp\left(-\frac{\Delta^2}{8\mathcal{I}^{-1}}\right), \tag{31}$$

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 $\mathcal{I}^{-1} \le \bar{\sigma}^2$ and the equation is true if and only if F is a normal distribution, we have the following conclusion.

THEOREM 3.5. *Consider the model* (28). *Assume all the assumptions needed in Theorem 3.4 and Lemma 3.4 hold. Then the spectral clustering $\check{z}$ is in general suboptimal, it fails to achieve the minimax rate* (31). *It is optimal if and only if the noise distribution F is $N(0, \bar{\sigma}^2)$.*

Theorem 3.5 establishes the suboptimality of the spectral clustering $\check{z}$ under the model (28). Though $\check{z}$ achieves an exponential error rate, it has a fundamentally suboptimal 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, whereas the optimal procedure uses 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. Even though that Theorem 3.5 is only limited to the model (28), the above reasoning suggests the spectral clustering is generally suboptimal under mixture models beyond (28) unless the noise follows a Gaussian distribution.

4. Discussion.

4.1. *Potential applications of leave-one-out singular subspace perturbation analysis.* In this paper, we have primarily applied the developed leave-one-out singular subspace perturbation toolkit to study the performance of spectral clustering in the context of mixture models. However, it is important to highlight that this toolkit holds promise for various other applications that exhibit low-rank structures and require entrywise analysis. Examples of such applications include low-rank matrix denoising, matrix completion, factor analysis, biclustering, and more.

To illustrate the versatility of our approach, consider a simple scenario where the data matrix W is approximately rank-one and can be expressed as $W = \lambda u v^T + E$. Here, λ is a scalar, and u and v are unit vectors. Let $\hat{\lambda}$, $\hat{u}$, $\hat{v}$ be the leading singular value, left singular vector, and right singular vector of W. Specifically, $\hat{v}_i$, the ith coordinate of $\hat{v}$, can be expressed as $\hat{v}_i = \hat{u}^T W_i/\hat{\lambda} = (\lambda \hat{u}^T u/\hat{\lambda}) v_i + \hat{u}^T \epsilon_i/\hat{\lambda}$, where X_i and ϵ_i represent the ith column of X and E, respectively. Under suitable regularity conditions, we can observe that the first term, $(\lambda \hat{u}^T u/\hat{\lambda}) v_i$, is well controlled, leaving the perturbation of $\hat{v}_i$ to be predominantly determined by the second term, $\hat{u}^T \epsilon_i/\hat{\lambda}$, which can be approximated as $\hat{u}^T \epsilon_i/\lambda$. Since $|\hat{u}^T \epsilon_i| = \|\hat{u}\hat{u}^T \epsilon_i\|$, we can leverage Theorem 2.1 to establish a connection between $|\hat{u}^T \epsilon_i|$

and $\|\hat{u}_{-i}\hat{u}_{-i}^T\epsilon_i\| = |\hat{u}_{-i}^T\epsilon_i|$, where $\hat{u}_{-i}$ represents the leading left singular vector of the data matrix with the ith column removed. Importantly, the independence between $\hat{u}_{-i}$ and ϵ_i can be exploited to analyze the magnitude of $|\hat{u}_{-i}^T\epsilon_i|$, facilitating an entrywise perturbation analysis for $\hat{v}_i$. This demonstrates the potential broader applicability of our leave-one-out singular subspace perturbation analysis beyond spectral clustering and mixture models.

4.2. *Extension to eigenspace perturbation.* In this paper, we primarily focus on the analysis of singular subspace perturbations. However, it is worth considering the potential extension of our findings to eigenspace perturbation scenarios. Let us consider two symmetric matrices, $Y \in \mathbb{R}^{(n-1)\times(n-1)}$ and $\hat{Y} \in \mathbb{R}^{n\times n}$. Here, $\hat{Y}$ is obtained from Y by removing the last row and column of $\hat{Y}$. For simplicity, we assume that $\hat{Y}_{n,n} = 0$. We introduce a vector $y_n \in \mathbb{R}^{n-1}$ such that the last row and column of $\hat{Y}$ can be represented as $(y_n^T, 0)$ and $(y_n^T, 0)^T$, respectively. Let the leading eigenspaces of Y and $\hat{Y}$ be denoted as $U_r \in \mathbb{R}^{(n-1)\times r}$ and $\hat{U}_r \in \mathbb{R}^{n\times r}$, respectively. In contrast to the singular subspace analysis, we note that Y and $\hat{Y}$ have different dimensionalities. To address this, we consider an augmented matrix $\tilde{U}_r = (U_r^T, 0)^T \in \mathbb{R}^{n\times r}$. Analyzing $\|\tilde{U}_r\tilde{U}_r^T - U_rU_r^T\|_{\mathrm{F}}$ leads us to follow a similar proof strategy as employed in Theorem 2.1. However, extending the proof from Theorem 2.1 to cover $\|\tilde{U}_r\tilde{U}_r^T - U_rU_r^T\|_{\mathrm{F}}$ appears to be nontrivial and potentially challenging.

The reason for this challenge lies in the perturbation between $\tilde{U}_r\tilde{U}_r^T$ and $U_rU_r^T$ that not only involves the last column but also the last row of $\hat{Y}$. In particular, the contribution of the last row $(y_n^T, 0)$ to the upper bound of $\|\tilde{U}_r\tilde{U}_r^T - U_rU_r^T\|_{\mathrm{F}}$ remains unclear, as it is not accounted for in the current analysis presented in Theorem 2.1. Hence, we defer the analysis of eigenspace perturbations to future research endeavors, recognizing the need for a more comprehensive and specialized treatment of this aspect.

4.3. *Approximated solution to k-means.* Solving the k-means problem exactly, as detailed in (14), can be computationally challenging, particularly for large datasets. To enhance practicality, one might opt for an approximate solution to k-means, where the solution's objective value remains within a factor of $(1+\varepsilon)$ of the global minimum. It's worth noting, however, that such an approximate solution may lack a property intrinsic to the global minimizer in (14): $\hat{z}_i = \mathrm{argmin}_{a\in[k]} \|\hat{U}_{1:r}^T X_i - \hat{c}_a\|^2$ for every $i \in [n]$, which is critical to our theoretical analysis. To circumvent this issue, we can use a strategy delineated in Section 2.5 of [26]. This approach, devised for addressing a similar problem for spectral clustering under Gaussian mixture models, executes an additional step of Lloyd's algorithm after obtaining the $(1+\varepsilon)$ solution. As evidenced by Theorem 2.2 in [26], the theoretical analysis for this augmented method closely mirrors that of the original. The cost of having the approximate solution is the need for a slightly more stronger signal-to-noise condition. In our context, this means Theorem 3.1 would remain valid, albeit with ψ_1 carrying an extra $\sqrt{1+\varepsilon}$ factor in its denominator.

4.4. *High-dimensional regime $p \gg n$.* In the context where k, β, σ are constants, Corollary 3.1 and Theorem 3.3 demand the conditions $\Delta/(1+\sqrt{p/n}) \to \infty$ and $\Delta/(1+p/n) \to \infty$ respectively. In the low-dimensional scenario, where $p \lesssim n$, these conditions can be equivalently expressed as $\Delta \to \infty$ recognized as optimal. Nevertheless, in the high-dimensional case $p \gg n$, these conditions are deemed suboptimal. For a two-component symmetric isotropic Gaussian mixture model, [9] demonstrates that spectral clustering remains consistent as long as $\Delta/(p/n)^{1/4} \to \infty$. More recently, for sub-Gaussian mixture models, under this condition, exponential misclustering errors are obtained in [16] through semi-definite programming (SDP) and in [1, 30] through a variant of spectral clustering that employs the

leading eigenvectors of a hollowed gram matrix $\mathcal{H}(X^TX) \in \mathbb{R}^{n\times n}$, where $\mathcal{H}(\cdot)$ is the hollowing operator that zeros out all diagonal entries of a square matrix. In addition, it is suggested in [1] that hollowing is crucial for spectral clustering in high-dimensional and heteroscedastic scenarios. It provides counterexamples showing that the leading eigenvectors of X^TX can be asymptotically orthogonal to their population counterparts. In contrast, those of the hollowed matrix $\mathcal{H}(X^TX)$ remain consistent. Our more stringent conditions, as compared to $\Delta/(p/n)^{1/4} \to \infty$, stem from challenges inherent in our analysis, possibly related to our use of the gram matrix, as opposed to $\mathcal{H}(X^TX)$.

4.5. *Explicit error rate of spectral clustering under other mixture models.* As our analysis in this paper establishes an explicit error rate under sub-Gaussian mixture models, a natural question is whether our analysis framework can be extended to other mixture models. A key observation is that the clustering error bound in Lemma 3.2 imposes no specific assumptions on the noise distribution $\{\epsilon_i\}$, allowing for potential applicability to a wide range of mixture models. However, this flexibility comes with challenges. Lemma 3.2 highlights that the clustering error is intimately tied to the tail probabilities of $\|\hat{U}^T_{-i,1:\kappa}\epsilon_i\|$. While the independence between $\hat{U}_{-i,1:\kappa}$ and ϵ_i is advantageous, the lack of explicit expressions for $\hat{U}_{-i,1:\kappa}$ poses difficulties when dealing with other noise distributions.

When ϵ_i follows a sub-Gaussian distribution, existing concentration inequalities can be applied to analyze the norm of $\hat{U}^T_{-i,1:\kappa}\epsilon_i$, providing a sharp upper bound as in Theorem 3.1. However, in scenarios where ϵ_i is assumed to follow a specific distribution, such as a centered Bernoulli random vector with success probability q decreasing as n grows (as encountered in community detection tasks), issues arise. Despite modeling ϵ_i as $\mathrm{SG}_p(1)$, the correct variance is q, leading to a loose upper bound for spectral clustering performance. Directly analyzing $\|\hat{U}^T_{-i,1:\kappa}\epsilon_i\|$ becomes challenging in such cases due to the lack of explicit expressions for $\hat{U}_{-i,1:\kappa}$ and uncertainties about the behavior of its entries. It is important to acknowledge that our current analysis framework has limitations when confronted with these complexities. Future research in this direction may involve exploring novel techniques or adapting existing methodologies to handle non-sub-Gaussian noise distributions more effectively, thereby establishing sharp analysis for spectral clustering under diverse mixture models.

4.6. *Unknown k or σ.* In this paper, we assume k, the number of clusters, is known. If k is unknown, one can employ existing methodologies, as found in the literature [29, 38, 40, 42], to estimate its value prior to applying our spectral clustering method. Our theoretical results maintain their validity, given that k is accurately estimated, albeit with an added term accounting for the estimation error of k. However, while such methods have empirically demonstrated decent performance, their theoretical performances are not fully understood, especially in contexts where both p, n are large. Regarding σ, the noise level in sub-Gaussian mixture models, both Algorithm 1 and Algorithm 2 require no prior knowledge of σ. However, in Theorem 3.2, the threshold T is needed to satisfy a condition involving σ. More generally, in Lemma 3.3, $T/\|E\|$ needs to be bounded away from 0. To endow the algorithm with enhanced adaptability, one possible approach is to consider $\hat{\lambda}_{k+1}$, the $(k+1)$th largest singular value of the data matrix, as a surrogate of $\|E\|$. The intuition is that when entries of the noise matrix E are independent and identically distributed, asymptotic behavior of its singular values can be characterized using random matrix theory, building a connection between $\|E\|$ and its leading singular values. Further investigation is beyond the scope of this paper.

5. 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 Supplementary Material [46] due to page limit.

5.1. *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 5.1, which only requires $\sigma_r^2 - \sigma_{r+1}^2 - \|(I - U_r U_r^T) y_n\|^2 > 0$ 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 5.1.

THEOREM 5.1. *If* $\sigma_r^2 - \sigma_{r+1}^2 - \|(I - U_r U_r^T) y_n\|^2 > 0$, *we have*

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

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

Throughout the proof, we denote

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

Denote $d = p \wedge (n-1)$. If $p \le n-1$, 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

$$\begin{aligned} \alpha^2 &= \|\hat{U}_r \hat{U}_r^T\|_{\mathrm{F}}^2 + \|U_r U_r^T\|_{\mathrm{F}}^2 - 2\langle \hat{U}_r \hat{U}_r^T, U_r U_r^T \rangle \\ &= 2k - 2\|U_r^T \hat{U}_r\|_{\mathrm{F}}^2 = 2k - 2 \sum_{i \in [r]} \sum_{j \in [r]} \hat{B}_{i,j}^2 \\ &= 2 \sum_{i \in [r]} b_i^2 = 2 \sum_{i=r+1}^{p} b_i^2, \end{aligned} \tag{32}$$

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

$$\|(I - \hat{U}_r \hat{U}_r^T) \hat{Y}\|_{\mathrm{F}}^2 \le \|(I - U_r U_r^T) \hat{Y}\|_{\mathrm{F}}^2.$$

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

$$\|(I - \hat{U}_r \hat{U}_r^T) Y\|_{\mathrm{F}}^2 + \|(I - \hat{U}_r \hat{U}_r^T) y_n\|^2 \le \|(I - U_r U_r^T) Y\|_{\mathrm{F}}^2 + \|(I - U_r U_r^T) y_n\|^2,$$

which implies

$$\|(I - \hat{U}_r \hat{U}_r^T) Y\|_{\mathrm{F}}^2 - \|(I - U_r U_r^T) Y\|_{\mathrm{F}}^2 \le \|(I - U_r U_r^T) y_n\|^2 - \|(I - \hat{U}_r \hat{U}_r^T) y_n\|^2. \tag{33}$$

We are going to simplify terms in (33).

(Simplification of the LHS of (33)). Recall the decomposition $Y = \sum_{i \in [d]} \sigma_i u_i v_i^T$. Since $(I - U_r U_r^T) Y = \sum_{i>r}^{d} \sigma_i u_i v_i^T$, we have $\|(I - U_r U_r^T) Y\|_{\mathrm{F}}^2 = \sum_{i>r}^{d} \sigma_i^2$. Since

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

we have

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

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

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

Then we have

$$\begin{aligned}\text{LHS of (33)} &= \sum_{i=1}^r\sigma_i^2(1-\|\hat{B}_{i,\cdot}\|_{\mathrm{F}}^2)-\sum_{i>r}^d\sigma_i^2\|\hat{B}_{i,\cdot}\|_{\mathrm{F}}^2\\ &= \sum_{i=1}^r\sigma_i^2b_i^2-\sum_{i>r}^d\sigma_i^2b_i^2\ge\sum_{i=1}^r\sigma_i^2b_i^2-\sigma_{r+1}^2\frac{\alpha^2}{2},\end{aligned}$$

where we use $\sum_{i>r}^d b_i^2\le\sum_{i>r}^p b_i^2=\alpha^2/2$ from (32) in the last inequality.

(Simplification of the RHS of (33)). 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

$$\begin{aligned}\text{RHS of (33)} &= y_n^T(I-U_rU_r^T)y_n-y_n^T(I-\hat{U}_r\hat{U}_r^T)y_n\\ &= y_n^T(\hat{U}_r\hat{U}_r^T-U_rU_r^T)y_n\\ &= y_n^TU\begin{pmatrix}\hat{B}_1\hat{B}_1^T-I_{r\times r} & \hat{B}_1\hat{B}_2^T\\ \hat{B}_2\hat{B}_1^T & \hat{B}_2\hat{B}_2^T\end{pmatrix}U^Ty_n.\end{aligned}$$

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

$$\begin{aligned}\text{RHS of (33)} &= y_n^T (U_r, U_\perp) \begin{pmatrix} -\hat{B}_1^\perp \hat{B}_1^{\perp T} & -\hat{B}_1^\perp \hat{B}_2^{\perp T} \\ -\hat{B}_2^\perp \hat{B}_1^{\perp T} & \hat{B}_2 \hat{B}_2^T \end{pmatrix} (U_r, U_\perp)^T y_n \\ &= -y_n^T U_r \hat{B}_1^\perp \hat{B}_1^{\perp T} U_r^T y_n - 2 y_n^T U_r \hat{B}_1^\perp \hat{B}_2^{\perp T} U_\perp^T y_n + y_n^T U_\perp \hat{B}_2 \hat{B}_2^T U_\perp^T y_n \\ &\le -\|\hat{B}_1^{\perp T} U_r^T y_n\|^2 + 2\|\hat{B}_1^{\perp T} U_r^T y_n\| \|\hat{B}_2^{\perp T}\| \|U_\perp^T y_n\| + \|\hat{B}_2^T\|^2 \|U_\perp^T y_n\|^2.\end{aligned}$$

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 (32). We also have

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

Since $\|\hat{B}_1^\perp\|_{\mathrm{F}}^2 = \sum_{i=1}^r (1 - \|\hat{B}_{i,\cdot}\|^2) = \alpha^2/2$ according to (32), we have $\|\hat{B}_1^\perp\| \le \alpha/\sqrt{2}$. Thus, using $U_r^T \epsilon = 0$, we have

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

Then,

$$\text{RHS of (33)} \le 2\|\hat{B}_1^{\perp T} U_r^T \theta\| \|\epsilon\| + \frac{\alpha^2}{2}\|\epsilon\|^2.$$

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

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

We then have

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

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,

$$\text{RHS of (33)} \le 2\|s\| \sqrt{\sum_{i=1}^r \sigma_i^2 b_i^2} \|\epsilon\| + \frac{\alpha^2}{2}\|\epsilon\|^2.$$

(*Combining the above simplifications for* (33).) From the above simplifications on the LHS and RHS of (33), we have

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

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

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

Note that the function $t^2 - 2\|s\|\|\epsilon\| t$ is increasing as long as $t \ge t_0$ where we define $t_0 := \|s\|\|\epsilon\|$. On the other hand, from (32), we have the domain $t \ge \alpha\sigma_r/\sqrt{2}$. We consider the following two scenarios.

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

$$\alpha \le \frac{\sqrt{2}t_0}{\sigma_r} = \frac{\sqrt{2}\|s\|\|\epsilon\|}{\sigma_r}. \tag{34}$$

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

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

Hence, we have an inequality of α:

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

which can be arranged into

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

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

$$\alpha \le \frac{2\sqrt{2}\sigma_r\|s\|\|\epsilon\|}{\sigma_r^2 - \sigma_{r+1}^2 - \|\epsilon\|^2}. \tag{35}$$

Since $2\sigma_r^2 > \sigma_r^2 - \sigma_{r+1}^2 - \|\epsilon\|^2$, the upper bound in (34) is strictly below that in (35). Hence, (35) holds for both scenarios. The proof is complete. □

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

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

Together with Theorem 5.1, we obtain the desired bound. □

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, define

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

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

$$P_{-i} := (\theta_{z_1^*}^*, \ldots, \theta_{z_{i-1}^*}^*, \theta_{z_{i+1}^*}^*, \ldots, \theta_{z_n^*}^*) \in \mathbb{R}^{p\times(n-1)}. \tag{36}$$

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

$$\begin{aligned}\lambda_{-i,\kappa}^2 &= \min_{w\in\Theta:\|w\|=1} \|w^T P_{-i}\|^2 = \min_{w\in\Theta:\|w\|=1} \sum_{j\in[n]:j\neq i} (w^T\theta_{z_j^*}^*)^2 \\ &\geq \frac{\frac{\beta n}{k}-1}{\frac{\beta n}{k}} \min_{w\in\Theta:\|w\|=1} \sum_{j\in[n]} (w^T\theta_{z_j^*}^*)^2 = \left(1-\frac{k}{\beta n}\right) \min_{w\in\Theta:\|w\|=1} \|w^T P\|^2 \\ &\geq \left(1-\frac{k}{\beta n}\right)\lambda_\kappa^2.\end{aligned} \tag{37}$$

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

$$E_{-i} := (\epsilon_1, \ldots, \epsilon_{i-1}, \epsilon_{i+1}, \ldots, \epsilon_n) \in \mathbb{R}^{p\times(n-1)}, \tag{38}$$

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

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

and

$$\hat{\lambda}_{-i,\kappa} - \hat{\lambda}_{-i,\kappa+1} \geq \lambda_{-i,\kappa} - \lambda_{-i,\kappa+1} - 2\|E\| \geq \sqrt{1-\frac{k}{\beta n}}\lambda_\kappa - 2\|E\|. \tag{40}$$

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

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

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

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

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

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

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

$$\|(I - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T)\theta_{z_i^*}^*\| \leq \frac{\|(I - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T)P_{-i}\|_{\mathrm{F}}}{\sqrt{\frac{\beta n}{k}-1}} \leq \frac{3\sqrt{\kappa}\|E\|}{\sqrt{\frac{\beta n}{k}-1}}, \tag{41}$$

and consequently,

$$\begin{aligned}\|(I - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T)X_i\| &\leq \|(I - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T)\theta_{z_i^*}^*\| + \|(I - \hat{U}_{-i,1:\kappa}\hat{U}_{-i,1:\kappa}^T)\epsilon_i\| \\ &\leq \frac{3\sqrt{\kappa}\|E\|}{\sqrt{\frac{\beta n}{k}-1}} + \|E\|.\end{aligned} \tag{42}$$

From (40) and (42), we have

$$\rho_{-i} \geq \frac{\sqrt{1-\frac{k}{\beta n}}\lambda_\kappa - 2\|E\|}{\|E\| + \frac{3\sqrt{\kappa}\|E\|}{\sqrt{\frac{\beta n}{k}-1}}} \geq \frac{\rho_0}{8} > 2, \tag{43}$$

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

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

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

$$|\hat{u}_{-i,a}^T \theta^*_{z_i^*}| \leq \hat{\lambda}_{-i,a}\frac{\sqrt{|S|}}{|S|} + \frac{\sqrt{|S|}\|E_{-i}\|}{|S|} \leq \frac{\hat{\lambda}_{-i,a}}{\sqrt{\frac{\beta n}{k}-1}} + \frac{\|E\|}{\sqrt{\frac{\beta n}{k}-1}}. \tag{44}$$

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

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

Consequently,

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

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

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

Since $\beta n/k^2 \geq 10$ and $\rho_0 > 16$ are assumed, we have $\hat{\lambda}_{-i,\kappa} \geq \lambda_\kappa/2$ by (39). Then together with (43), the above display can be simplified into

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

This concludes the proof of Theorem 2.2. □

6. Proof of main results in Section 3. In this section, we include proofs of Lemma 3.1, Lemma 3.2, and Theorem 3.1. The proofs of all other results of Section 3 are included in the Supplementary Material [46] due to page limit.

6.1. *Proofs 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 (15), we know $\hat{z}_i$ must satisfy

$$\hat{z}_i = \operatorname*{argmin}_{a\in[k]} \|\hat{U}\hat{U}^T X_i - \hat{\theta}_a\|,$$

where $\{\hat{\theta}_a\}_{a\in[k]}$ satisfies (18) according to Proposition 3.1. Hence, we have

$$\mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\} = \mathbb{I}\Big\{\min_{a\in[k]:a\neq\phi(z_i^*)} \|\hat{U}\hat{U}^T X_i - \hat{\theta}_a\| \le \|\hat{U}\hat{U}^T X_i - \hat{\theta}_{\phi(z_i^*)}\|\Big\}.$$

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 $\|x-y\| \le \|x-w\|$, then we must have $\|y-w\|/2 \le \|x-w\|$. Hence, we have

$$\begin{aligned}
&\mathbb{I}\{\|\hat{U}\hat{U}^T X_i - \hat{\theta}_a\| \le \|\hat{U}\hat{U}^T X_i - \hat{\theta}_{\phi(z_i^*)}\|\}\\
&\quad= \mathbb{I}\Big\{\frac{1}{2}\|\hat{\theta}_{\phi(z_i^*)} - \hat{\theta}_a\| \le \|\hat{U}\hat{U}^T X_i - \hat{\theta}_{\phi(z_i^*)}\|\Big\}\\
&\quad\le \mathbb{I}\Big\{\frac{1}{2}\|\hat{\theta}_{\phi(z_i^*)} - \hat{\theta}_a\| \le \|\hat{U}\hat{U}^T \epsilon_i - \hat{\theta}_{\phi(z_i^*)}\| + \|\hat{U}\hat{U}^T\theta^*_{z_i^*} - \hat{\theta}_{\phi(z_i^*)}\|\Big\}\\
&\quad\le \mathbb{I}\{\|\hat{\theta}_{\phi(z_i^*)} - \hat{\theta}_a\| - 2\|\theta^*_{z_i^*} - \hat{\theta}_{\phi(z_i^*)}\| \le 2\|\hat{U}\hat{U}^T\epsilon_i - \hat{\theta}_{\phi(z_i^*)}\|\},
\end{aligned}$$

where we use the fact that $X_i = \theta^*_{z_i^*} + \epsilon_i$ and $\|\hat{U}\hat{U}^T\theta^*_{z_i^*} - \hat{\theta}_{\phi(z_i^*)}\| \le \|\theta^*_{z_i^*} - \hat{\theta}_{\phi(z_i^*)}\|$. 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

$$\begin{aligned}
&\mathbb{I}\{\|\hat{U}\hat{U}^T X_i - \hat{\theta}_a\| \le \|\hat{U}\hat{U}^T X_i - \hat{\theta}_{\phi(z_i^*)}\|\}\\
&\quad\le \mathbb{I}\{\|\theta^*_{z_i^*} - \theta^*_{\phi^{-1}(a)}\| - \|\hat{\theta}_{\phi(z_i^*)} - \theta^*_{z_i^*}\| - \|\theta^*_{\phi^{-1}(a)} - \hat{\theta}_a\|\\
&\quad\quad - 2\|\theta^*_{z_i^*} - \hat{\theta}_{\phi(z_i^*)}\| \le 2\|\hat{U}\hat{U}^T\epsilon_i\|\} \qquad (45)\\
&\quad\le \mathbb{I}\Big\{\|\theta^*_{z_i^*} - \theta^*_{\phi^{-1}(a)}\| - 4\max_{b\in[k]}\|\theta^*_b - \hat{\theta}_{\phi(b)}\| \le 2\|\hat{U}\hat{U}^T\epsilon_i\|\Big\}\\
&\quad\le \mathbb{I}\Big\{\Big(1 - \frac{4C_0\beta^{-0.5}kn^{-0.5}\|E\|}{\Delta}\Big)\Delta \le 2\|\hat{U}\hat{U}^T\epsilon_i\|\Big\},
\end{aligned}$$

where in the last inequality, we use the fact that $\max_{b\in[k]}\|\theta^*_b - \hat{\theta}_{\phi(b)}\| \le C_0\beta^{-0.5}kn^{-0.5}\|E\|$ 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 $a \in [k]$ not $\phi(z_i^*)$, we have

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

where in the last inequality we use the definition of ψ_0 in (16). □

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 $\hat{U}\hat{U}^T\epsilon_i$:

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

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

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

In Lemma 3.1 we establish (19). From there, we have

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

for some constant $C' > 0$, where in the last inequality we use the assumption $\rho_0 > 16$ from (9). The upper bound on $\mathbb{E}\ell(\hat{z}, z^*)$ is an immediate consequence as $\mathbb{E}\ell(\hat{z}, z^*) = n^{-1}\sum_{i\in[n]}\mathbb{E}\mathbb{I}\{\hat{z}_i \ne \phi(z_i^*)\}$. □

6.2. *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

$$\mathbb{P}\bigl(\|E\| \le 8\sigma(\sqrt{n} + \sqrt{p})\bigr) \ge 1 - e^{-n/2}, \tag{46}$$

by Lemma E.1. Denote $\mathcal{F}$ to be this event. Under $\mathcal{F}$, as long as $\psi_1, \rho_1 \ge 128$, we have both (16) and (9) hold. Let $\phi \in \Phi$ satisfy $\ell(\hat{z}, z^*) = n^{-1}\sum_{i\in[n]}\mathbb{I}\{\hat{z}_i \ne \phi(z_i^*)\}$. Consider a fixed $i \in [n]$. Then from Lemma 3.2, we have

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

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

$$\begin{aligned}\mathbb{E}\ell\bigl(\hat{z}, z^*\bigr) &\le \mathbb{E}\mathbb{I}\bigl\{\mathcal{F}^{\complement}\bigr\} + \mathbb{E}\ell\bigl(\hat{z}, z^*\bigr)\mathbb{I}\{\mathcal{F}\} \\ &\le e^{-n/2} + n^{-1}\sum_{i\in[n]}\mathbb{E}\mathbb{I}\bigl\{(1 - C_1\psi)\Delta \le 2\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\|\bigr\}.\end{aligned} \tag{47}$$

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 E.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 $k/t \le 16k^2\sigma^2/\Delta^2 \le 16\psi_1^2$ where we use the fact that $\frac{\Delta}{k\sigma} > \psi_1^{-1}$ from (21) as $\beta \le 1$. Then we have

$$\begin{aligned}\sigma^2(\kappa + 2\sqrt{\kappa t} + 2t) &= 2\sigma^2 t\biggl(\frac{1}{2}\frac{\kappa}{t} + \sqrt{\frac{\kappa}{t}} + 1\biggr) \le 2\sigma^2 t\bigl(8\psi_1^2 + 4\psi_1 + 1\bigr) \le 2\sigma^2 t(1 + 8\psi_1) \\ &\le (1 - C_2\psi)\Delta^2/\bigl(8\sigma^2\bigr)(1 + 8\psi) \le (1 - C_1\psi)\Delta^2/\bigl(8\sigma^2\bigr),\end{aligned}$$

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

$$\mathbb{E}\mathbb{I}\bigl\{(1 - C_1\psi)\Delta \le 2\|\hat{U}_{-i}\hat{U}_{-i}^T\epsilon_i\|\bigr\} \le \exp(-t) = \exp\biggl(-(1 - C_2\psi)\frac{\Delta^2}{8\sigma^2}\biggr). \qquad \square$$

Acknowledgments. The authors are grateful to an anonymous Associate Editor and anonymous referees for careful reading of the manuscript and their valuable remarks and suggestions.

Funding. The first author was supported in part by NSF Grant DMS-2112988.
The second author was supported in part by NSF Grant DMS-2112918.

SUPPLEMENTARY MATERIAL

Supplement to "Leave-one-out singular subspace perturbation analysis for spectral clustering" (DOI: 10.1214/24-AOS2418SUPP; .pdf). In the supplement [46], we first provide the proof of Theorem 2.3 in Appendix A, followed by the proofs of results of Section 3.4 in Appendix B. The proof of Theorem 3.3 is given in Appendix C. The proofs of results of Section 3.6 are given in Appendix D. Auxiliary lemmas and propositions and their proofs are included in Appendix E.

REFERENCES

[1] ABBE, E., FAN, J. and WANG, K. (2022). An ℓ_p theory of PCA and spectral clustering. *Ann. Statist.* **50** 2359–2385. MR4474494 https://doi.org/10.1214/22-aos2196

[2] ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* **48** 1452–1474. MR4124330 https://doi.org/10.1214/19-AOS1854

[3] AGTERBERG, J., LUBBERTS, Z. and PRIEBE, C. E. (2022). Entrywise estimation of singular vectors of low-rank matrices with heteroskedasticity and dependence. *IEEE Trans. Inf. Theory* **68** 4618–4650. MR4449064 https://doi.org/10.1109/tit.2022.3159085

[4] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2567175 https://doi.org/10.1007/978-1-4419-0661-8

[5] BELABBAS, M.-A. and WOLFE, P. J. (2009). Spectral methods in machine learning and new strategies for very large datasets. *Proc. Natl. Acad. Sci. USA* **106** 369–374.

[6] BLUM, A., COJA-OGHLAN, A., FRIEZE, A. and ZHOU, S. (2007). Separating populations with wide data: A spectral analysis. In *Algorithms and Computation. Lecture Notes in Computer Science* **4835** 439–451. Springer, Berlin. MR2472630 https://doi.org/10.1007/978-3-540-77120-3_39

[7] 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. *Ann. Statist.* **49** 944–967. MR4255114 https://doi.org/10.1214/20-aos1986

[8] CAI, T., LI, H. and MA, R. (2021). Optimal structured principal subspace estimation: Metric entropy and minimax rates. *J. Mach. Learn. Res.* **22** Paper No. 46, 45. MR4253739

[9] CAI, T. T. and ZHANG, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.* **46** 60–89. MR3766946 https://doi.org/10.1214/17-AOS1541

[10] CAPE, J., TANG, M. and PRIEBE, C. E. (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *Ann. Statist.* **47** 2405–2439. MR3988761 https://doi.org/10.1214/18-AOS1752

[11] CHEN, Y., CHI, Y., FAN, J. and MA, C. (2021). Spectral methods for data science: A statistical perspective. *Found. Trends Mach. Learn.* **14** 566–806.

[12] DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450 https://doi.org/10.1137/0707001

[13] DAVIS, D., DIAZ, M. and WANG, K. (2021). Clustering a mixture of Gaussians with unknown covariance. Preprint. Available at arXiv:2110.01602.

[14] DING, L. and CHEN, Y. (2020). Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Trans. Inf. Theory* **66** 7274–7301. MR4173640 https://doi.org/10.1109/TIT.2020.2992769

[15] FAN, J., WANG, W. and ZHONG, Y. (2017). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **18** Paper No. 207, 42. MR3827095

[16] GIRAUD, C. and VERZELEN, N. (2018). Partial recovery bounds for clustering with the relaxed K-means. *Math. Stat. Learn.* **1** 317–374. MR4059724

[17] HAN, R., LUO, Y., WANG, M. and ZHANG, A. R. (2022). Exact clustering in tensor block model: Statistical optimality and computational limit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1666–1698. MR4515554

[18] JIN, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89. MR3285600 https://doi.org/10.1214/14-AOS1265

[19] KANNAN, R. and VEMPALA, S. (2008). Spectral algorithms. *Found. Trends Theor. Comput. Sci.* **4** 157–288. MR2558901 https://doi.org/10.1561/0400000025

[20] KISELEV, V. Y., ANDREWS, T. S. and HEMBERG, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20** 273–282. https://doi.org/10.1038/s41576-018-0088-9

[21] KOLTCHINSKII, V. and LOUNICI, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Probab. Stat.* **52** 1976–2013. MR3573302 https://doi.org/10.1214/15-AIHP705

[22] KOLTCHINSKII, V. and XIA, D. (2016). Perturbation of linear forms of singular vectors under Gaussian noise. In *High Dimensional Probability VII. Progress in Probability* **71** 397–423. Springer, Cham. MR3565274 https://doi.org/10.1007/978-3-319-40519-3_18

[23] LEI, J. and LIN, K. Z. (2023). Bias-adjusted spectral clustering in multi-layer stochastic block models. *J. Amer. Statist. Assoc.* **118** 2433–2445. MR4681594 https://doi.org/10.1080/01621459.2022.2054817

[24] LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

[25] LEI, L. (2019). Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. Preprint. Available at arXiv:1909.04798.

[26] LÖFFLER, M., ZHANG, A. Y. and ZHOU, H. H. (2021). Optimality of spectral clustering in the Gaussian mixture model. *Ann. Statist.* **49** 2506–2530. MR4338373 https://doi.org/10.1214/20-aos2044

[27] LU, Y. and ZHOU, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. Preprint. Available at arXiv:1612.02099.

[28] MA, C., WANG, K., CHI, Y. and CHEN, Y. (2018). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. In *International Conference on Machine Learning* 3345–3354. PMLR.

[29] MONTI, S., TAMAYO, P., MESIROV, J. and GOLUB, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52** 91–118.

[30] NDAOUD, M. (2022). Sharp optimal recovery in the two component Gaussian mixture model. *Ann. Statist.* **50** 2096–2126. MR4474484 https://doi.org/10.1214/22-aos2178

[31] NDAOUD, M., SIGALLA, S. and TSYBAKOV, A. B. (2022). Improved clustering algorithms for the bipartite stochastic block model. *IEEE Trans. Inf. Theory* **68** 1960–1975. MR4395508 https://doi.org/10.1109/tit.2021.3130683

[32] NEWMAN, M. E. (2013). Spectral methods for community detection and graph partitioning. *Phys. Rev. B* **88** 042822.

[33] QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. *Adv. Neural Inf. Process. Syst.* 3120–3128.

[34] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887

[35] SCHIEBINGER, G., WAINWRIGHT, M. J. and YU, B. (2015). The geometry of kernelized spectral clustering. *Ann. Statist.* **43** 819–846. MR3325711 https://doi.org/10.1214/14-AOS1283

[36] SRIVASTAVA, P. R., SARKAR, P. and HANASUSANTO, G. A. (2023). A robust spectral clustering algorithm for sub-Gaussian mixture models with outliers. *Oper. Res.* **71** 224–244. MR4560196 https://doi.org/10.1287/opre.2022.2317

[37] STEWART, G. W. (1990). Matrix perturbation theory.

[38] TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 411–423. MR1841503 https://doi.org/10.1111/1467-9868.00293

[39] VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[40] VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. MR2409803 https://doi.org/10.1007/s11222-007-9033-z

[41] VON LUXBURG, U., BELKIN, M. and BOUSQUET, O. (2008). Consistency of spectral clustering. *Ann. Statist.* **36** 555–586. MR2396807 https://doi.org/10.1214/009053607000000640

[42] WANG, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika* **97** 893–904. MR2746159 https://doi.org/10.1093/biomet/asq061

[43] WANG, K., YAN, Y. and DIAZ, M. (2020). Efficient clustering for stretched mixtures: Landscape and optimality. *Adv. Neural Inf. Process. Syst.* **33** 21309–21320.

[44] WEDIN, P. (1972). Perturbation bounds in connection with singular value decomposition. *BIT* **12** 99–111. MR0309968 https://doi.org/10.1007/bf01932678

[45] YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika* **102** 315–323. MR3371006 https://doi.org/10.1093/biomet/asv008

[46] ZHANG, A. Y. and ZHOU, H. Y. (2024). Supplement to "Leave-one-out singular subspace perturbation analysis for spectral clustering." https://doi.org/10.1214/24-AOS2418SUPP

[47] ZHOU, Z. and AMINI, A. A. (2019). Analysis of spectral clustering algorithms for community detection: The general bipartite setting. *J. Mach. Learn. Res.* **20** Paper No. 47, 47. MR3948087