Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

OPTIMALITY OF SPECTRAL CLUSTERING FOR GAUSSIAN MIXTURE MODEL

By Matthias Löffler, Anderson Y. Zhang, and Harrison H. Zhou

ETH Zürich, University of Pennsylvania, and Yale University

Spectral clustering is one of the most popular algorithms to group high dimensional data. It is easy to implement and computationally efficient. Despite its popularity and successful applications, its theoretical properties have not been fully understood. The spectral clustering algorithm is often used as a consistent initializer for more sophisticated clustering algorithms. However, in this paper, we show that spectral clustering is actually already optimal in the Gaussian Mixture Model, when the number of clusters of is fixed and consistent clustering is possible. Contrary to that spectral gap conditions are widely assumed in literature to analyze spectral clustering, these conditions are not needed in this paper to establish its optimality.

1. Introduction. Clustering is a central and fundamental problem in statistics and machine learning. Among all the clustering methods, spectral method [59, 64] has become particularly popular for clustering high dimensional data. It tracks back to [16, 24] and has enjoyed tremendous success. In computer science and machine learning, spectral clustering and its variants have been widely used to solve many different problems, including parallel computations [27, 58, 62], graph partition [7, 8, 13, 14, 15, 22, 23, 43, 52, 53, 66], and explanatory data mining and statistical data analysis [2, 6, 31, 46]. It also has many real data applications, including image segmentation [44, 57, 67], text mining [11, 12, 48], speech separation [4, 18], and many others. In recent years, spectral clustering has also been one of the most favored and studied methods for community detection [3, 5, 17, 29, 36, 38, 54, 56, 60].

Spectral clustering is easy to implement and has remarkably good performance. The idea behind spectral clustering is dimensionality reduction. First it performs a spectral decomposition on the dataset, or some related distance matrix, and only keeps the leading few spectral components. In this way, the dimension of data is then greatly reduced to the number of

MSC 2010 subject classifications: Primary 60G05
Keywords and phrases: Spectral Clustering, K-means, Gaussian Mixture Model, Random Matrix, Spectral Perturbation

spectral components used. Then a standard clustering method (for example, the k-means algorithm) is performed on the low dimensional denoised data to obtain an estimation of cluster assignment. Due to the dimensionality reduction, spectral clustering is computationally less demanding than many other classical clustering algorithms.

In spite of its popularity and successful applications, the theoretical properties of spectral clustering are not fully understood. Computer scientists associate spectral clustering with the graph Laplacian matrix. [6, 20, 25, 26, 65] provide various forms of asymptotic convergence for the graph Laplacian matrix and its related spectral properties. Recent years have seen more encouraging progress towards understanding the theoretical guarantee of spectral clustering, in the context of community detection. [29, 38, 53, 54, 69] show that spectral clustering applied to the adjacency matrix of network can consistently recover hidden community structure. However, their upper bounds on number of nodes incorrectly clustered are in some polynomial form of the signal-to-noise ratio. It differs from the optimal rate of community detection [68], which takes an exponential form of the signal-to-noise ratio. Therefore, in the literature spectral method is often used as a way to initialize iterative algorithms to eventually achieve the optimal clustering error rate.

In this paper, we investigate the theoretical performance of spectral clustering under the Gaussian Mixture Model [50, 61], which is arguably the most standard and used model for clustering analysis. Under the Gaussian Mixture Model, data points are generated from a mixture of Gaussian distributions, whose centers are separated from each other, resulting in a cluster structure. The goal is to recover the underlying true cluster assignment.

Maximum likelihood estimation for the labels in the Gaussian Mixture Model is equivalent to be the k-means algorithm, which is NP-hard [9, 42] and hence not practical when the dimension of data is large. As a result, various approximations have been used and studied. One direction is to relax the k-means clustering by semi-definite programming [21, 51, 55]. Another popular direction is to apply Lloyd's Algorithm [39], which is a greedy iterative method to approximate the k-means algorithm. Given a sufficiently good initializer, typically provided by spectral clustering [37], Lloyd's Algorithm achieves the optimal clustering rate [41, 45]. However, in fact, we show that spectral clustering itself is already optimal.

A closely related result about spectral clustering for the Gaussian Mixture Model is [63]. Under a strong separation condition which implies that exact cluster recovery is possible, spectral clustering is proved to achieve exact recovery with high probability. In this paper, we consider also situations where only partial recovery is possible. We measure the performance of the spectral clustering output $\hat{z}$ by the Hamming loss function $\ell(\cdot,\cdot)$, compared to the underlying true assignment z^*. The main result of this paper is summarized in Theorem 1.1 informally.

THEOREM 1.1 (Informal Statement of the Main Result). *For n data points generated from a Gaussian Mixture model, if*

- *the number of clusters is finite*
- *the sizes of clusters are in the same order*
- *the minimum distance among centers Δ goes to infinity*
- *the dimension p is at most in the same order of n*

then with high probability, spectral clustering achieves the optimal clustering rate, which is

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

To the best of our knowledge this provides arguably the first theoretical guarantee on the optimality of spectral clustering in a general setting. The separation parameter Δ covers a wide scale of values, ranging from consistent cluster estimation to the exact recovery. We refer readers to Theorem 2.1 and Corollary 2.1 for more rigorous statements and slightly stronger results, where we allow the number of clusters to grow with n, the cluster sizes are not necessarily in the same order, and the dimension p may go beyond n. Though in this paper we focus mainly on the Gaussian Mixture Model setting, we hope the technique can be extended to other clustering model, which may eventually leads to a general framework to understand the performance of spectral clustering.

It is worth mentioning that in Theorem 1.1, no spectral gap (i.e., singular value gap) condition is needed. This is contrary to the existing literature [1, 29, 38, 54], where various forms of eigenvalue gap or singular value gap are required. For technical reasons, they need them to be sufficiently large to apply matrix spectral perturbation theory. This does not match with the intuition, that what matters in a clustering problem should be the distances among centers, regardless of its spectral structure. In this paper, we completely drop any condition on the spectral gap by developing a novel technical analysis to show that the contribution of singular vectors from smaller singular values is essentially negligible.

A recent related paper by Abbe et al. [1] studies community detection under an idealized scenario, where the network has two equal-size communities and the connectivity probabilities are equal to $an^{-1}\log n$ or $bn^{-1}\log n$, where a and b are fixed constants. They show that, the performance of clustering on the second leading eigenvector matches with the minimax rate, by using a leave-one-out technique. The technical tools we use in this paper are different. We extend the spectral operator perturbation theory of [34, 35] and introduce new techniques to establish the optimality of spectral clustering and also to remove the spectral gap condition.

Organization. The paper is organized as follows. In Section 2, we first introduce the Gaussian Mixture Model, followed by the spectral clustering algorithm, and then state the main results. The proof of the main theorem is given in Section 3, which is started with a proof sketch. We include the proofs of all the lemmas in the supplement.

Notation. For any matrix M, we denote $\|M\|$ and $\|M\|_{\mathrm{F}}$ to be its operator norm and Frobenius norm. respectively. We use the notation $M_{i,\cdot}$ and $M_{\cdot,i}$ to indicate its ith row and columns, respectively. For any matrices M, N of the same dimension, their inner product is defined as $\langle M, N\rangle = \sum_{i,j} X_{i,j}Y_{i,j}$. For any d, we denote $\{e_a\}_{a=1}^d$ to be the standard Euclidean basis with $e_1 = (1,0,0,\ldots), e_2 = (0,1,0,\ldots,0),\ldots,e_d = (0,0,0,\ldots,1)$. We let 1_d be a vector of length d whose entries are all 1. We use $[d]$ to denote the set $\{1,2,\ldots,d\}$. We use $\mathbb{I}\{\cdot\}$ as a notation for the indicator function. For $y_1, y_2, \ldots, y_d \in \mathbb{R}$, $\mathrm{diag}\{y_1, y_2, \ldots, y_d\}$ is the $d \times d$ diagonal matrix, whose diagonal entries are $y_1, y_2, \ldots, y_d$ from the top-left to the bottom right, with off-diagonal entries being 0.

2. Main Results.

2.1. *Gaussian Mixture Model.*

Consider a Gaussian Mixture Model with $X = (X_1, \ldots, X_n) \in \mathbb{R}^{p\times n}$ with k centers $\theta_1^*, \ldots, \theta_k^* \in \mathbb{R}^p$. Let $z^* \in [k]^n$ be the underlying true cluster assignment vector. Under this model, the observations $\{X_i\}_{i\in[n]}$ are generated as follows:

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

where $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{N}(0, I_p)$.

The goal of clustering is to recover the cluster assignment z^*. For any $z \in [k]^n$, the performance of clustering can be captured by the Hamming distance: the number of coordinates that take different values in z^* and z. However, the cluster structure is invariant to the permutations of label symbols. We denote the loss $\ell(z, z^*)$ to be the number of data points misclustered, which is defined as follows, considering all the label permutation:

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

where $\Phi = \{\phi :$ bijection from $[k]$ to $[k]\}$.

Note that the difficulty of clustering is largely determined by the distances among centers $\{\theta_1^*, \ldots, \theta_k^*\}$. If two centers are exactly equal to each other, it is impossible to distinguish the corresponding two clusters. We define Δ to be the minimum distance among centers:

$$\Delta = \min_{j,l \in [k]: j \neq l} \left\| \theta_j^* - \theta_l^* \right\|. \tag{1}$$

Another quantity related to clustering is the sizes of clusters. When the size a cluster is small, recovering it might be more difficult. We quantify the size of the smallest cluster by β, defined as

$$\beta = \frac{\min_{j \in [k]} \left|\{i \in [n] : z_i^* = j\}\right|}{n/k}. \tag{2}$$

Note that β cannot be greater than 1. When β is a constant, all clusters sizes are of the same order. We allow the case $\beta = o(1)$ so that clusters sizes may differ in magnitude.

2.2. *Spectral Clustering.* Various forms of spectral clustering have been proposed and studied in the literature. Spectral clustering is an umbrella term for clustering after a dimension reduction through a spectral decomposition. The variants differ mostly in what matrix the spectral decomposition is applied on, and what spectral components are used for the subsequent clustering. The clustering method used mostly is the k-means method.

In the context of community detection, spectral clustering [29, 38, 53, 54, 69] is associated with the eigenvectors of the adjacency matrix. For general clustering settings, [6, 20, 25, 26, 64, 65] first obtain a similarity matrix from the original data points by certain kind of kernels. Then the Laplacian matrix is constructed, whose eigenvectors are used for clustering. In [32, 37], spectral clustering is performed directly on the original data matrix.

The spectral clustering algorithm considered in this paper is presented in Algorithm 1. It is simple, involves only one singular value decomposition (SVD) and one k-means clustering step. Despite its simplicity, it is powerful as it will be shown to achieve the optimal misclustering rate. The key step in the algorithm that leads to the optimal rate is to weight the empirical singular vectors by the corresponding empirical singular values.

Algorithm 1: Spectral Clustering

Input: Data matrix $X \in \mathbb{R}^{p\times n}$, number of clusters k
Output: Clustering label vector $\hat{z} \in [k]^n$

1 Perform SVD on X to have

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

where $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \ldots \geq \hat{\sigma}_{p\wedge n} \geq 0$ and $\{\hat{u}_i\}_{i=1}^{p\wedge n} \in \mathbb{R}^p, \{\hat{v}_i\}_{i=1}^{p\wedge n} \in \mathbb{R}^n$.

2 Consider the first k singular values and corresponding singular vectors. Define $\hat{\Sigma} \in \mathbb{R}^{k\times k}$ to be a diagonal matrix with $\hat{\Sigma}_{i,i} = \hat{\sigma}_i, \forall i \in [k]$, and $\hat{V} \in \mathbb{R}^{n\times k}$ to be $\hat{V} = (\hat{v}_1, \ldots, \hat{v}_k)$.

3 Define

$$\hat{Y} = \hat{\Sigma}\hat{V}^T \in \mathbb{R}^{k\times n}.$$

Let $\hat{Y}_{\cdot,i} \in \mathbb{R}^k$ be the ith column of $\hat{Y}$, for all $i \in [n]$. Perform k-means on $\{\hat{Y}_{\cdot,i}\}_{i=1}^n$. Let $\hat{z}$ be the clustering result it returns. That is,

$$\left(\hat{z}, \{\hat{c}_j\}_{j=1}^k\right) = \mathop{\arg\min}_{z\in[k]^n, \{c_j\}_{j=1}^k \in \mathbb{R}^k} \sum_{i\in[n]} \left\|\hat{Y}_{\cdot,i} - c_{z_i}\right\|^2. \tag{3}$$

As common in the clustering literature, we assume that k, the number of clusters, is known. The purpose of the SVD is to perform a dimensionality reduction while preserving the underlying data structure. After SVD, the dimension of data is reduced from p to k. This makes the follow-up k-means algorithm computationally feasible compared to applying it directly onto the columns of X.

The idea of weighting singular vectors by the corresponding singular values is natural. The importance of singular vectors is different: singular vectors with smaller singular values should carry relatively less useful information, and consequently deserve less attention. Clustering on $\hat{Y}$ instead of $\hat{V}$ is one of the reasons why we are able to remove the spectral gap condition. It is also worth pointing out, as we will later show in Lemma 3.1, that Algorithm 1 is equivalent to Algorithm 2, for which clustering is performed on the columns of the rank-k matrix approximation of X. Similar ideas of using low rank matrix approximations for clustering have also been proposed in [19, 37].

2.3. *Optimality.* In Theorem 2.1, we establish the theoretical optimality of spectral clustering in the Gaussian Mixture Model.

THEOREM 2.1. *Suppose that,*

$$\frac{\Delta}{k^{10.5}\beta^{-0.5}(1+\frac{p}{n})^{0.5}} \to \infty \tag{4}$$

then we have that

$$\ell(\hat{z}, z^*) \leq n \exp\left(-\left(1 - \left(\frac{\Delta}{k^{10.5}\beta^{-0.5}(1+\frac{p}{n})^{0.5}}\right)^{-0.1}\right)\frac{\Delta^2}{8}\right), \tag{5}$$

with probability at least $1 - \exp(-\Delta) - \exp(-0.08n)$.

In Theorem 2.1, we allow the number of clusters k to grow with n, the cluster sizes not to be in the same order (quantified by β), and the dimension p to go beyond n. This is slightly stronger than the informal statement we make in Theorem 1.1. Note that when k and β are both fixed constants and p is at most in the same order of n, Equation (4) is equivalent to assume $\Delta \to \infty$, and Theorem 2.1 is reduced to the case in Theorem 1.1.

COROLLARY 2.1. *In addition to the assumption stated in Theorem 2.1, if $p \leq Cn$ for some constant $C > 0$, then there exists another constant $C' > 0$, such that with probability at least $1 - \exp(-\Delta) - \exp(-0.08n)$, we have that*

$$\ell(\hat{z}, z^*) \leq n \exp\left(-\left(1 - C'\left(\frac{\Delta}{k^{10.5}\beta^{-0.5}}\right)^{-0.1}\right)\frac{\Delta^2}{8}\right), \tag{6}$$

which is minimax optimal.

Corollary 2.1 establishes the optimality of spectral clustering when p is at most in the same order of n. Under this scenario, Theorem 2.1 shows the spectral clustering provably attains the rate $n\exp\left(-(1-o(1))\Delta^2/8\right)$. This exactly matches the minimax lower bound established in [41]. Theorem 3.3 of [41] states that

$$\inf_{\hat{z}} \sup_{(\theta_1^*,\ldots,\theta_k^*),z} \mathbb{E}\ell(\hat{z}, z^*) \geq n \exp\left(-(1+o(1))\frac{\Delta^2}{8}\right), \tag{7}$$

when $\Delta/\log\left(k\beta^{-1}\right) \to \infty$. In Equation (7), the infimum is taken over all feasible estimators $\hat{z}$, and the supremum is taken over all possible parameters, where the true centers $(\theta_1^*, \ldots, \theta_k^*) \in \mathbb{R}^{p\times k}$ are separated by minimum distance Δ, and the true cluster assignment z^* has minimum cluster size

$\beta n/k$. The matching rates in Theorem 2.1 and Equation (6) indicate that spectral clustering is minimax optimal.

The separation parameter Δ is allowed to take a wide range of values. When k and β are fixed constants and p is at most in the same order of n, the condition needed for Δ in Theorem 2.1 and Corollary 2.1 is reduced to

$$\Delta \to \infty. \tag{8}$$

This turns out to be the necessary and sufficient condition to have a consistent cluster recover. Theorem 3.3 of [41] also shows that, if $\Delta = \mathcal{O}(1)$, the minimax rate is lower bounded by cn for some constant $c > 0$, indicating that no method is possible to achiever consistency. Equation (8) covers a wide range of settings from consistency to exact recovery. When $\liminf_{n\to\infty} \Delta^2/(8\log n) > 1$, we achieve exact recovery, where the output of spectral clustering $\hat{z}$ is exactly equal to the underlying true cluster assignment z^* with high probability.

There is no spectral gap (i.e., singular value gap) condition assumed in Theorem 2.1 and Corollary 2.1. It is entirely possible that the population matrix $\mathbb{E}X$ has a rank that is smaller than k. This is equivalent to allow for the smallest singular values of the population matrix $\mathbb{E}X$ to be 0. This is contrary to the existing literature [1, 29, 38, 54], where the spectral gap is assumed to be sufficiently large to apply matrix spectral perturbation theory. The spectral gap condition is not natural, as the minimax rate in Equation (7) only depends on Δ and is invariant to any spectral structure. In Theorem 2.1 and Corollary 2.1, we completely drop any spectral gap condition, and our results match with the intuition that the difficulty of cluster recovery is determined only by Δ, the minimum distance among the centers.

3. Proof of Main Results.

In this section, we give the proof of Theorem 2.1. In Section 3.1, we first introduce a few population quantities, which are population counterparts of the quantities appearing in Algorithm 1. After that, several key lemmas for the proof are presented in Section 3.2. Since the proof of Theorem 2.1 is long and involved, we first provide a proof sketch in Section 3.3, then followed by its complete and detailed proof in Section 3.4. Auxiliary lemmas are included in the supplement.

3.1. *Population Quantities.*

We define $P = \mathbb{E}X$ and $E = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^{p\times n}$ to be the noise matrix, so that we have a matrix representation $X = P + E$. We then define several quantities on P, in a similarly way as on X

as in Algorithm 1. Let the SVD of P be (note that P is at most rank-k)

$$P = \sum_{i=1}^{k} \sigma_i u_i v_i^T,$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k \geq 0$, $\{u_i\}_{i=1}^k \in \mathbb{R}^p$ and $\{v_i\}_{i=1}^k \in \mathbb{R}^n$. It can be written in matrix form $P = U\Sigma V^T$, where $U = (u_1, \ldots, u_k) \in \mathbb{R}^{n\times k}, V = (v_1, \ldots, v_k) \in \mathbb{R}^{n\times k}$, and $\Sigma \in \mathbb{R}^{k\times k}$ is a diagonal matrix with $\Sigma_{i,i} = \sigma_i, \forall i \in [k]$. Define

$$Y = \Sigma V^T \in \mathbb{R}^{k\times n}. \tag{9}$$

Let $P_{\cdot,i}, Y_{\cdot,i} \in \mathbb{R}^k$ be the ith column of P and Y respectively, for all $i \in [n]$. Note that $P_{\cdot,i} = \theta^*_{z_i^*}, \forall i \in [n]$. In Appendix A, we provide several propositions (Propositions A.1, A.2 and A.3) to characterize the structure of these population quantities especially related to V.

3.2. *Key Lemmas.* In this section, we present several key lemmas used in the proof of Theorem 2.1.

Algorithm 2: Clustering with Rank-k Approximation

Input: Data matrix $X \in \mathbb{R}^{p\times n}$, number of clusters k

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

1 Implement the same Step 1-2 as in Algorithm 1 to obtain $\hat{\Sigma} \in \mathbb{R}^{k\times k}$ and $\hat{V} \in \mathbb{R}^{n\times k}$. In addition, define $\hat{U} = (\hat{u}_1, \ldots, \hat{u}_k) \in \mathbb{R}^{p\times k}$.

2 Define

$$\hat{P} = \hat{U}\hat{\Sigma}\hat{V}^T \in \mathbb{R}^{p\times n}.$$

Let $\hat{P}_{\cdot,i} \in \mathbb{R}^p$ be the ith column of $\hat{P}$, for all $i \in [n]$. Perform k-means on $\{\hat{P}_{\cdot,i}\}_{i=1}^n$. Let $\hat{z}'$ be the clustering result it returns. That is,

$$\left(\hat{z}', \left\{\hat{\theta}_j\right\}_{j=1}^k\right) = \mathop{\arg\min}_{z\in[k]^n, \left\{\theta_j\right\}_{j=1}^k \in \mathbb{R}^k} \sum_{i\in[n]} \left\|\hat{P}_{\cdot,i} - \theta_{z_i}\right\|^2. \tag{10}$$

In Lemma 3.1, we show that Algorithm 1 has the same output as that of Algorithm 2, where clustering is performed on the columns of $\hat{U}\hat{Y}$ instead of $\hat{Y}$. Its proof is simple and straightforward, since $\hat{U}$ has orthonormal columns which enables $\hat{U}\hat{Y}$ to preserve all the structure in $\hat{Y}$. We defer its proof to the supplement.

LEMMA 3.1. *Algorithm 1 and Algorithm 2 are equivalent. That is, let* $(\hat{z}, \{\hat{c}_j\}_{j=1}^k)$ *and* $(\hat{z}', \{\hat{\theta}_j\}_{j=1}^k)$ *be solutions to Equations (3) and (10) respectively. The two outputs* $\hat{z}, \hat{z}'$ *are identical up to a label permutation, i.e., there exists a* $\phi \in \Phi$ *such that*

$$\hat{z}_i' = \phi(\hat{z}_i), \forall i \in [n].$$

In addition,

$$\hat{\theta}_j = \hat{U}\hat{c}_{\phi(j)}, \forall j \in [k].$$

In Lemma 3.2, we give some preliminary results on the performance of Algorithm 2. By some straightforward random matrix analysis, we are able to show Algorithm 2 (equivalently, Algorithm 1) has at least polynomial convergence rate of order $k(n+p)/\Delta^2$ for clustering. This rate is far from being optimal. However, it indeed captures some useful information of estimated cluster assignments and centers. One important implication is that the estimated centers $\{\hat{\theta}_j\}_{j\in[n]}$ are close to their population ones $\{\theta_j^*\}_{j\in[n]}$, after some label permutation (i.e., Equation (13)). This is the start point for us to obtain sharper bounds for $\ell(\hat{z}, z^*)$.

Before stating Lemma 3.2, we first introduce an event:

$$\mathcal{F} = \left\{ \|E\| \leq \sqrt{2}(\sqrt{n} + \sqrt{p}) \right\}. \tag{11}$$

Standard random matrix theory shows that $\mathcal{F}$ occurs with high probability (Lemma B.1). The operator norm of E indicates the magnitude of difference between the empirical singular quantities and their population counterparts. The gap between the empirical and population quantities are well controlled under the event $\mathcal{F}$. In Lemma 3.2 and also in the proof of Theorem 2.1 (Section 3.4), we assume $\mathcal{F}$ holds. The proof of Lemma 3.2 is included in the supplement.

LEMMA 3.2. *Assume that the event* $\mathcal{F}$ *holds. We have that*

$$\left\| \hat{P} - P \right\|_F \leq 4\sqrt{k}(\sqrt{n} + \sqrt{p}).$$

In addition, if $\Delta/(\beta^{-0.5}k\,(1+p/n)^{0.5}) \geq C$ *for some constant* $C > 0$, *there exists another constant* C' *such that the solution to Equation (10)* $(\hat{z}', \{\hat{\theta}_j\}_{j=1}^k)$ *satisfies*

$$\ell(\hat{z}', z^*) \leq \frac{C'k\,(n+p)}{\Delta^2}, \tag{12}$$

$$\text{and} \quad \min_{\phi\in\Phi}\max_{j\in[k]} \left\| \hat{\theta}_j - \theta^*_{\phi(j)} \right\| \leq C'\beta^{-\frac{1}{2}}k\sqrt{1+\frac{p}{n}}. \tag{13}$$

Consequently, if the ratio $\Delta/(\beta^{-0.5}k\,(1+p/n)^{0.5})$ is sufficiently large, we have $\min_{j\in[k]}|\{i\in[n]:\hat{z}_i=j\}|\geq\frac{\beta n}{2k}$.

Lemma 3.3 studies the difference between empirical spectral projection matrix and its sample counterpart. It decomposes $\hat{V}_{a:b}\hat{V}_{a:b}^T - V_{a:b}V_{a:b}^T$, the difference between an empirical projection matrix and its population counterpart, into a linear part of the random noise matrix E and a remaining part, which can be shown to be negligible. The linear part has a very simple form, and is the main component that leads to the exponent $\Delta^2/8$ in Equation (5). The remaining non-linear part, though without an explicit expression, is well-behaved and concentrates strongly around its mean.

Lemma 3.3 holds for general matrices with underlying structure, not necessary from the clustering setup, as long as the noise is normally distributed. The result stated in Lemma 3.3 is a slight generalization of [34, 35], where $\sigma_a,\ldots,\sigma_b$ are assumed to be the same. That is, in [34, 35], $V_{a:b}V_{a:b}^T$ is the spectral projection matrix for one unique singular values, with $a-b$ being its multiplicity. Here we go beyond this assumption, by allowing the corresponding singular values to vary. The proof of Lemma 3.3 is quite involved but mainly follows [34, 35]. We include it in the supplement for completeness.

LEMMA 3.3. *Consider any rank-k matrix $M\in\mathbb{R}^{p\times n}$ with SVD $M=\sum_{j=1}^k\sigma_j u_j v_j^T$ where $\sigma_1\geq\sigma_2\ldots\geq\sigma_k>0$. Define $\sigma_0=+\infty$ and $\sigma_{k+1}=0$.*

Let E be a Gaussian noise matrix with all the entries $E_{i,j}\stackrel{iid}{\sim}\mathcal{N}(0,1),\forall i\in[p],j\in[n]$. Define $\hat{M}=M+E$. Let the SVD of $\hat{M}$ be $\sum_{j=1}^{p\wedge n}\hat{\sigma}_j\hat{u}_j\hat{v}_j^T$ where $\hat{\sigma}_1\geq\hat{\sigma}_2\geq\ldots\geq\hat{\sigma}_{p\wedge n}$.

For any two indexes a,b such that $1\leq a\leq b\leq k$, define $V_{a:b}=(v_a,\ldots,v_b)$ and $\hat{V}_{a:b}=(\hat{v}_a,\ldots,\hat{v}_b)$. Let $V=(v_1,\ldots,v_k)$. Define the singular value gap $g_{a:b}=\min\{\sigma_{a-1}-\sigma_a,\sigma_b-\sigma_{b+1}\}$. Let

$$S_{a:b}=\left(I-VV^T\right)\left(\hat{V}_{a:b}\hat{V}_{a:b}^T-V_{a:b}V_{a:b}^T\right)V_{a:b}-\sum_{a\leq j\leq b}\frac{1}{\sigma_j}\left(I-VV^T\right)E^Tu_jv_j^TV_{a:b}.$$

Assuming $\mathbb{E}\|E\|\leq\frac{g_{a:b}}{8}$, there exists some constant $C>0$ such that

$$\mathbb{P}\left(|\langle S_{a:b}-\mathbb{E}S_{a:b},W\rangle|\leq C\left(1+\frac{\sigma_a-\sigma_b}{g_{a:b}}\right)\frac{\sqrt{t}}{g_{a:b}}\left(\frac{\sqrt{n+p}+\sqrt{t}}{g_{a:b}}\right)\|W\|_*\right)\geq 1-2e^t,$$

for any $W\in\mathbb{R}^{n\times(b-a)}$, any $t\geq\log 4$ and where $\|\cdot\|_*$ denotes the nuclear norm.

Lemma 3.4 characterizes the distributions of empirical singular vectors. Similar to Lemma 3.3, Lemma 3.4 holds for matrices with any underlying structure, not necessary in the clustering setting, as long as the noise is normally distributed. The most important implication of Lemma 3.4 is that, for any empirical singular vector $\hat{v}_j$, its component that is orthogonal to the true signal V (i.e., $(I - VV^T)\hat{v}_j$) has a distribution that is invariant to the underlying signal structure, after a proper normalization. This observation appears and is utilized in [30, 49]. Lemma 3.4 is essentially the same as Theorem 6 of [49]. For completeness, we give the proof in the supplement.

LEMMA 3.4. *Consider any rank-k matrix $M \in \mathbb{R}^{p\times n}$ with SVD $M = \sum_{j=1}^{k} \sigma_j u_j v_j^T$ where $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_k > 0$.*

Let E be a Gaussian noise matrix with entries $E_{i,j} \overset{iid}{\sim} \mathcal{N}(0,1), \forall i \in [p], j \in [n]$. Define $\hat{M} = M + E$. Let the SVD of $\hat{M}$ be $\sum_{j=1}^{p\wedge n} \hat{\sigma}_j \hat{u}_j \hat{v}_j^T$ where $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \ldots \geq \hat{\sigma}_{p\wedge n}$.

Define $V = (v_1, \ldots, v_k)$. Then for any $j \in [k]$, the following holds:

(1) $\left(I - VV^T\right)\hat{v}_j / \left\|\left(I - VV^T\right)\hat{v}_j\right\|$ *is uniformly distributed on the unit sphere spanned by $\left(I - VV^T\right)$. That is, we have*

$$\frac{\left(I - VV^T\right)\hat{v}_j}{\left\|\left(I - VV^T\right)\hat{v}_j\right\|} \sim \frac{\left(I - VV^T\right)w}{\left\|\left(I - VV^T\right)w\right\|}, \text{ where } w \sim \mathcal{N}(0, I_n).$$

In particular, $\mathbb{E}\left(I - VV^T\right)\hat{v}_j / \left\|\left(I - VV^T\right)\hat{v}_j\right\| = 0$.

(2) $\left(I - VV^T\right)\hat{v}_j / \left\|\left(I - VV^T\right)\hat{v}_j\right\|$ *is independent of $VV^T\hat{v}_j$.*

(3) $\left(I - VV^T\right)\hat{v}_j / \left\|\left(I - VV^T\right)\hat{v}_j\right\|$ *is independent of $\left\|\left(I - VV^T\right)\hat{v}_j\right\|$.*

3.3. *Proof Sketch for Theorem 2.1.* In this section, we provide a sketch for proof of Theorem 2.1. The complete and detailed proof is given in section 3.4. Throughout the proof, we assume that the random event $\mathcal{F}$ (defined in Equation (11)) holds.

We use the equivalence between Algorithm 1 and Algorithm 2 (by Lemma 3.1), where clustering is performed on the columns of $\hat{P} = \hat{U}\hat{U}^T\hat{Y}$. It is sufficient to study the behavior of $(\hat{z}, \{\hat{\theta}_j\}_{j\in[n]})$, which is the solution to Equation (10). Note that Equation (10) implies

$$\hat{z}_i = \underset{j\in[k]}{\arg\min} \left\|\hat{P}_{\cdot,i} - \hat{\theta}_j\right\|^2, \forall i \in [n].$$

Then after some proper label permutation, $\ell(\hat{z}, z^*)$ can be bounded by

$$\ell(\hat{z}, z^*) \leq \sum_{i=1}^{n} \mathbb{I}\{\hat{z}_i \neq z_i^*\} \leq \sum_{i=1}^{n} \mathbb{I}\left\{\arg\min_{a\in[k]} \left\|\hat{P}_{\cdot,i} - \hat{\theta}_a\right\|^2 \neq z_i^*\right\}$$
$$\leq \sum_{i=1}^{n} \sum_{a\neq z_i^*} \mathbb{I}\left\{\left\|\hat{P}_{\cdot,i} - \hat{\theta}_a\right\|^2 \leq \left\|\hat{P}_{\cdot,i} - \hat{\theta}_{z_i^*}\right\|^2\right\}.$$

Hence, it is all about understanding the relative values of distances between the data points $\{\hat{P}_{\cdot,i}\}_{i\in[n]}$ and the centers $\{\hat{\theta}_j\}_{j\in[k]}$. We divide the remaining proof into four steps, corresponding to Sections 3.4.1 to 3.4.4 in the complete proof. We use c as a general notation for constants, and $o(1)$ as a general notation for sequences goes to 0 along with n going to infinity.

Step 1 (Sketch of Section 3.4.1). In this step, we are going to decompose $\ell(\hat{z}, z^*)$ into two parts: one is related to the leading large singular values, and the other one is related to the remaining ones. To achieve this, we split $\{\hat{P}_{\cdot,i}\}_{i\in[n]}$ and $\{\hat{\theta}_j\}_{j\in[k]}$ into two parts. To be more specific, letting $\sigma_{k+1} = 0$, we define $r \in [k]$ as follows

$$r = \max\left\{j \in [k] : \sigma_j - \sigma_{j+1} \geq \rho(\sqrt{n} + \sqrt{p})\right\},$$

where $\rho \to \infty$ is some quantity whose value will be given in the complete proof. There are two benefits in choosing such r: the remaining singular values are relative small; and the singular value $\sigma_r - \sigma_{r+1}$ is large enough for applying matrix spectral perturbation theory.

We split $\hat{U}$ into $(\hat{U}_{1:r}, \hat{U}_{(r+1):k})$. We have that $\hat{P}_{\cdot,i} = \hat{P}_{\cdot,i}^{(1)} + \hat{P}_{\cdot,i}^{(2)}, \forall i \in [n]$, where

$$\hat{P}_{\cdot,i}^{(1)} = \hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P}_{\cdot,i}, \quad \text{and } \hat{P}_{\cdot,i}^{(2)} = \hat{U}_{(r+1):k}\hat{U}_{(r+1):k}^T\hat{P}_{\cdot,i}.$$

Similarly we have $\hat{\theta}_j = \hat{\theta}_j^{(1)} + \hat{\theta}_j^{(2)}, \forall j \in [k]$. Then we have the decomposition

$$\ell(\hat{z}, z^*) = \sum_{i=1}^{n} \sum_{a\neq z_i^*} \mathbb{I}\left\{\left\|\hat{P}_{\cdot,i}^{(1)} - \hat{\theta}_a^{(1)}\right\|^2 - \left\|\hat{P}_{\cdot,i}^{(1)} - \hat{\theta}_{z_i^*}^{(1)}\right\|^2 \leq -\left\|\hat{P}_{\cdot,i}^{(2)} - \hat{\theta}_a^{(2)}\right\|^2 + \left\|\hat{P}_{\cdot,i}^{(2)} - \hat{\theta}_{z_i^*}^{(2)}\right\|^2\right\}$$
$$\leq \sum_{i=1}^{n} \sum_{a\neq z_i^*} \mathbb{I}\left\{\left\|\hat{P}_{\cdot,i}^{(1)} - \hat{\theta}_a^{(1)}\right\|^2 - \left\|\hat{P}_{\cdot,i}^{(1)} - \hat{\theta}_{z_i^*}^{(1)}\right\|^2 \leq -\gamma\Delta^2\right\}$$
$$+ \sum_{i=1}^{n} \sum_{a\neq z_i^*} \mathbb{I}\left\{\gamma\Delta^2 \leq -\left\|\hat{P}_{\cdot,i}^{(2)} - \hat{\theta}_a^{(2)}\right\|^2 + \left\|\hat{P}_{\cdot,i}^{(2)} - \hat{\theta}_{z_i^*}^{(2)}\right\|^2\right\},$$

for some quantity $\gamma = o(1)$ such that $\gamma\Delta/k \to \infty$. The value of γ will be given in the complete proof. Lemma 3.2 shows that $\{\hat{\theta}_j\}_{j\in[k]}$ are close to their true values $\{\theta^*_j\}_{j\in[k]}$:

$$\max_{j\in[k]} \left\|\hat{\theta}_j - \theta^*_j\right\| = o(\Delta).$$

Together with the fact that $\{\theta^*_j\}_{j\in[k]}$ differ at least Δ in distance from each other, and that $\max_{j\geq r+1}\hat{\sigma}_j$ are relative small by perturbation theory, we end up with

$$\ell(\hat{z},z^*) \leq \sum_{i=1}^n \sum_{a\neq z_i^*} \left(\mathbb{I}\left\{(1-o(1))\,\Delta \leq 2\left\|\hat{P}^{(1)}_{\cdot,i} - \hat{U}_{1:r}\hat{U}_{1:r}^T\theta^*_{z_i^*}\right\|\right\} + \mathbb{I}\left\{c\gamma\Delta^2 \leq \left|\left\langle \hat{P}^{(2)}_{\cdot,i}, \hat{\theta}^{(2)}_a - \hat{\theta}^{(2)}_{z_i^*}\right\rangle\right|\right\}\right).$$

Note that $\hat{P}^{(1)}_{\cdot,i} - \hat{U}_{1:r}\hat{U}_{1:r}^T\theta^*_{z_i^*} = \hat{U}_{1:r}\hat{U}_{1:r}^T(\hat{P}-P)e_i$. The projection of this quantity from the right on the space spanned by VV^T can be shown to be negligible, i.e. $\|\hat{U}_{1:r}\hat{U}_{1:r}^T(\hat{P}-P)VV^Te_i\| = o(\Delta)$. This leaves $\hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P}(I-VV^T)e_i$ to be analyzed. Note that $\|\hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P}(I-VV^T)e_i\| = \|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T(I-VV^T)e_i\|$, where $\hat{\Sigma}_{r\times r}$ and $\hat{V}_{1:r}$ are the leading part of $\hat{\Sigma}$ and $\hat{V}$, respectively. Moreover, we have that

$$\langle \hat{P}^{(2)}_{\cdot,i}, \hat{\theta}^{(2)}_a - \hat{\theta}^{(2)}_{z_i^*}\rangle = \sum_{l=r+1}^k \hat{\sigma}_l\hat{V}_{i,l}(\hat{u}_l^T\hat{\theta}_a - \hat{u}_l^T\hat{\theta}_{z_i^*}),$$

which, up to some constant scalar, can be upper bounded by $\sum_{l=r+1}^k \sqrt{n}|\hat{V}_{i,l}|$. Thus we obtain that

$$\begin{aligned}\ell(\hat{z},z^*) &\leq \sum_{i=1}^n \sum_{a\neq z_i^*}\left(\mathbb{I}\left\{(1-o(1))\,\Delta \leq 2\left\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T\left(I-VV^T\right)e_i\right\|\right\} + \mathbb{I}\left\{c\gamma\Delta^2 \leq \sum_{l=r+1}^k \sqrt{n}\left|e_i^T\hat{v}_l\right|\right\}\right)\\ &\leq k\sum_{i=1}^n \mathbb{I}\left\{(1-o(1))\,\Delta \leq 2\left\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T\left(I-VV^T\right)e_i\right\|\right\} + k\sum_{l=r+1}^k\sum_{i=1}^n \mathbb{I}\left\{c\gamma\Delta^2/k \leq \sqrt{n}\left|e_i^T\hat{v}_l\right|\right\}.\end{aligned}$$

We denote the two terms as A and B respectively. We are going to provide upper bounds on their expectations. Once we have upper bound on $\mathbb{E}\ell(\hat{z},z^*)\mathbb{I}\{\mathcal{F}\} \leq \mathbb{E}A\mathbb{I}\{\mathcal{F}\} + \mathbb{E}B\mathbb{I}\{\mathcal{F}\}$, Markov's inequality will lead to a with high probability result for $\ell(\hat{z},z^*)\mathbb{I}\{\mathcal{F}\}$.

Step 2 (Sketch of Section 3.4.2). In this step, we are going to establish an upper bound on $\mathbb{E}A$. First, we replace $\hat{\Sigma}_{r\times r}$ with $\Sigma_{r\times r}$. Indeed, we have that

$$\begin{aligned}\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T\left(I-VV^T\right)e_i\| &= \sup_{w\in\mathbb{R}^r:\|w\|=1} e_i^T\left(I-VV^T\right)\hat{V}_{1:r}\hat{\Sigma}_{r\times r}w\\ &= (1+o(1))\sup_{w\in\mathbb{R}^r:\|w\|=1} e_i^T\left(I-VV^T\right)\hat{V}_{1:r}\Sigma_{r\times r}w.\end{aligned}$$

We are going to connect $\hat{V}_{1:r}$ with $V_{1:r}$. However, since the singular values may vary in magnitude, a direct application of spectral perturbation theory on $\hat{V}_{1:r}$ and $V_{1:r}$ is not good enough. Instead, we split $[r]$ into disjoint sets $\cup_{1\leq m\leq s} J_m$, such that the condition number in each set, i.e., $\max_{j\in J_m}\sigma_j / \min_{j\in J_m}\sigma_j = 1+o(1)$, and also the the singular value gaps among $\{J_m\}_{m\in[s]}$ are sufficiently large. More detail of the split will be provided in the complete proof.

Let $\hat{\Sigma}_{J_m\times J_m}, \hat{V}_{J_m}, V_{J_m}, w_{J_m}$ be the corresponding part of the related quantities. We continue the approximation

$$\left\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T\left(I-VV^T\right)e_i\right\| = (1+o(1))\sup_{w\in\mathbb{R}^r:\|w\|=1}\sum_{m\in[s]} e_i^T\left(1-VV^T\right)\hat{V}_{J_m}\Sigma_{J_m\times J_m}w_{J_m}$$

$$= (1+o(1))\sup_{w\in\mathbb{R}^r:\|w\|=1}\sum_{m\in[s]} e_i^T\left(1-VV^T\right)\hat{V}_{J_m}\hat{V}_{J_m}^T V_{J_m}\Sigma_{J_m\times J_m}w_{J_m}$$

$$= (1+o(1))\sup_{w\in\mathbb{R}^r:\|w\|=1}\sum_{m\in[s]} e_i^T\left(1-VV^T\right)\left(\hat{V}_{J_m}\hat{V}_{J_m}^T - V_{J_m}V_{J_m}^T\right)V_{J_m}\Sigma_{J_m\times J_m}w_{J_m},$$

Here the second equation holds because $\hat{V}_{J_m}^T V_{J_m}$ is close to some orthonormal matrix, and the last equation is due to the fact $(I-VV^T)V_{J_m}=0$. Lemma 3.3 indicates the spectral projection matrix difference $(I-VV^T)(\hat{V}_{J_m}\hat{V}_{J_m}^T - V_{J_m}V_{J_m}^T)V_{J_m}$ can be decomposed into a linear term of E and a non-linear term. The linear terms of E will contribute to the optimal rate and dominate the non-linear terms. We have that

$$\left\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T\left(I-VV^T\right)e_i\right\|$$

$$= (1+o(1))\sup_{w\in\mathbb{R}^r:\|w\|=1}\sum_{m\in[s]} e_i^T\left(\sum_{l\in J_m}\frac{1}{\sigma_l}\left(I-VV^T\right)E^T u_l v_l^T V_{J_m} + S_m\right)\Sigma_{J_m\times J_m}w_{J_m}$$

$$= (1+o(1))\left(\sup_{w\in\mathbb{R}^r:\|w\|=1}\sum_{m\in[s]}\sum_{l\in J_m} w_l e_i^T\left(I-VV^T\right)E^T u_l + \sup_{w\in\mathbb{R}^r:\|w\|=1}\sum_{m\in[s]} e_i^T S_m\Sigma_{J_m\times J_m}w_{J_m}\right)$$

$$= (1+o(1))\left(\left\|U_{1:r}^T E\left(I-VV^T\right)e_i\right\| + o(\Delta)\right),$$

where the last inequality is due to Lemma 3.3 which enables us to control the non-linear term S_m.

The above approximations lead to a simplification of A:

$$A \leq k\sum_{i=1}^n \mathbb{I}\left\{(1-o(1))\Delta \leq 2(1+o(1))\left(\left\|U_{1:r}^T E\left(I-VV^T\right)e_i\right\| + o(1)\Delta\right)\right\}$$

$$\leq k\sum_{i=1}^n \mathbb{I}\left\{(1-o(1))\Delta \leq 2\left\|U_{1:r}^T E\left(I-VV^T\right)e_i\right\|\right\}.$$

Note that $\left\|U_{1:r}^T E\left(I - VV^T\right) e_i\right\|^2$ is stochastically dominated by χ_k^2. Thus, there exist $\{\xi_i\}_{i\in[n]} \sim \chi_k^2$, such that

$$A \leq k \sum_{i=1}^{n} \mathbb{I}\left\{(1-o(1))\,\Delta \leq 2\sqrt{\xi_i}\right\}.$$

The tail probability of the square root of a χ^2 distribution behaves similarly to $\mathcal{N}(0,1)$, when it is away from 0. This leads to the desired rate

$$\mathbb{E}A\mathbb{I}\left\{\mathcal{F}\right\} \leq k \sum_{i=1}^{n} \mathbb{E}\mathbb{I}\left\{(1-o(1))\,\Delta \leq 2\sqrt{\xi_i}\right\} \leq nk\exp\left(-\left(1-o(1)\right)\Delta^2/8\right),$$

as we assume the event $\mathcal{F}$ holds in the above analysis.

Step 3 (Sketch of Section 3.4.3). In this step, we provide an upper bound on $\mathbb{E}B$. Note that B is all about the singular vectors $\{\hat{v}_l\}_{r+1\leq l\leq k}$. For each one, the right projection $VV^T\hat{v}_l^T$ is easy to control, leaving $\left(I - VV^T\right)\hat{v}_l^T$ as the main term to be analyzed. Lemma 3.4 characterizes the distribution of $\left(I - VV^T\right)\hat{v}_l^T$, which has the same distribution as $\left(I - VV^T\right)\zeta_l$, where $\{\zeta_l\}_{l=r+1}^k \sim \mathcal{N}(0, I_n)$, after some normalization. Then $\sqrt{n}e_i^T\left(I - VV^T\right)\hat{v}_l^T$ is approximately Gaussian distributed. This yields

$$\mathbb{E}B\mathbb{I}\left\{\mathcal{F}\right\} \leq k \sum_{l=r+1}^{k}\sum_{i=1}^{n} \mathbb{E}\mathbb{I}\left\{c\gamma\Delta^2/k \leq \sqrt{n}\left|e_i^T\left(I - VV^T\right)\hat{v}_l\right|\right\} \leq k^2 n\exp\left(-c\left(\gamma\Delta k^{-1}\right)^2\Delta^2\right),$$

as we assume the event $\mathcal{F}$ holds in the above analysis.

Step 4 (Sketch of Section 3.4.4). Since $\mathbb{E}\ell(\hat{z},z^*)\mathbb{I}\left\{\mathcal{F}\right\} \leq \mathbb{E}A\mathbb{I}\left\{\mathcal{F}\right\} + \mathbb{E}B\mathbb{I}\left\{\mathcal{F}\right\}$, so far we have obtained

$$\begin{aligned}\mathbb{E}\ell(\hat{z},z^*)\mathbb{I}\left\{\mathcal{F}\right\} &\leq nk\exp\left(-\left(1-o(1)\right)\frac{\Delta^2}{8}\right) + k^2 n\exp\left(-c\left(\gamma\Delta k^{-1}\right)^2\Delta^2\right)\\ &= n\exp\left(-\left(1-o(1)\right)\frac{\Delta^2}{8}\right).\end{aligned}$$

By Markov's inequality, with very high probability, we achieve

$$\ell(\hat{z},z^*)\mathbb{I}\left\{\mathcal{F}\right\} \leq n\exp\left(-\left(1-o(1)\right)\Delta^2/8\right).$$

A simple union bound with $\mathbb{P}\left(\mathcal{F}\right)$ leads to the desired rate for $\ell(\hat{z},z^*)$.

3.4. *Proof of Theorem 2.1.* In this section, we are going to give a complete and detailed proof of Theorem 2.1. We divide this section into four parts, following the same structure as in the proof sketch (i.e, Section 3.3). In Section 3.4.1, we establish the decomposition $\ell(\hat{z}, z^*) \leq A + B$. Then in Section 3.4.2 and Section 3.4.3, we provide upper bounds on $\mathbb{E}A$ and $\mathbb{E}B$, respectively. Finally in Section 3.4.4, we wrap everything up to achieve the desired rate. Again, throughout the whole proof, we assume the random event $\mathcal{F}$ (defined in Equation (11)) holds.

By Lemma 3.1, studying $\ell(\hat{z}, z^*)$ is equivalent to studying $\ell(\hat{z}', z^*)$ where $\hat{z}'$ is the output of Algorithm 2. Indeed, Lemma 3.1 indicates there exists a label permutation $\phi_0 \in \Phi$ such that $\hat{z}_i = \phi_0(\hat{z}'_i)$ for all $i \in [n]$. Without loss of generality, we can assume ϕ_0 is the identity mapping. Then we have that

$$\hat{z} = \hat{z}', \qquad \text{and } \hat{\theta}_j = \hat{U}\hat{c}_j, \forall j \in [k].$$

Together with Equation (10), we have that

$$\left(\hat{z}, \left\{\hat{\theta}_j\right\}_{j=1}^k\right) = \underset{z \in [k]^n, \{\theta_j\}_{j=1}^k \in \mathbb{R}^k}{\arg\min} \sum_{i \in [n]} \left\|\hat{P}_{\cdot,i} - \theta_{z_i}\right\|^2. \tag{14}$$

This implies that

(15)
$$\hat{z}_i = \underset{j \in [k]}{\arg\min} \left\|\hat{P}_{\cdot,i} - \hat{\theta}_j\right\|^2, \forall i \in [n], \text{ and } \hat{\theta}_j = \frac{\sum_{i\in[n]} \hat{P}_{\cdot,i}\mathbb{I}\left\{\hat{z}_i = j\right\}}{\left|\left\{i \in [n] : \hat{z}_i = j\right\}\right|}, \forall j \in [k].$$

By Lemma 3.2, we have that $\max_{j\in[k]} \|\hat{\theta}_j - \theta^*_{\phi'(j)}\| \leq 8\sqrt{2}\sqrt{\beta^{-1}k^2\left(1 + p/n\right)}$ for some label permutation mapping $\phi' \in \Phi$. Without loss of generality, we can again assume ϕ' is the identity mapping. Then we have that

$$\max_{j \in [k]} \left\|\hat{\theta}_j - \theta^*_j\right\| \leq 8\sqrt{2}\sqrt{\beta^{-1}k^2\left(1 + \frac{p}{n}\right)}. \tag{16}$$

From Equation (15), we have

$$\begin{aligned}
\ell(\hat{z}, z^*) &= \min_{\phi \in \Phi} \sum_{i=1}^n \mathbb{I}\left\{\underset{a \in [k]}{\arg\min} \left\|\hat{P}_{\cdot,i} - \hat{\theta}_a\right\|^2 \neq \phi(z^*_i)\right\} \leq \sum_{i=1}^n \mathbb{I}\left\{\underset{a \in [k]}{\arg\min} \left\|\hat{P}_{\cdot,i} - \hat{\theta}_a\right\|^2 \neq z^*_i\right\} \\
&\leq \sum_{i=1}^n \sum_{a \neq z^*_i} \mathbb{I}\left\{\left\|\hat{P}_{\cdot,i} - \hat{\theta}_a\right\|^2 \leq \left\|\hat{P}_{\cdot,i} - \hat{\theta}_{z^*_i}\right\|^2\right\} \triangleq \sum_{i=1}^n \sum_{a \neq z^*_i} T_{i,a},
\end{aligned}$$

where we denote $T_{i,a} = \mathbb{I}\{\|\hat{P}_{\cdot,i} - \hat{\theta}_a\|^2 \leq \|\hat{P}_{\cdot,i} - \hat{\theta}_{z^*_i}\|^2\}$ for all $i \in [n]$ and $a \in [k], a \neq z^*_i$. In the following, we focus on simplifying $T_{i,a}$.

3.4.1. *Establishing the decomposition $\ell(\hat{z}, z^*) \le A + B$.* A key step is to decompose each of $\{\hat{P}_{\cdot,i}\}_{i\in[n]}, \{\hat{\theta}_j\}_{j\in[k]}$ into a summation of two parts: one is consisted of coordinates corresponding to the larger singular values, and the other is consisted of the remaining coordinates related to the relative smaller singular values. Let $\sigma_{k+1} = 0$. We define $r \in [k]$ to be

$$r = \max\left\{j \in [k] : \sigma_j - \sigma_{j+1} \ge \rho\sqrt{n+p}\right\}, \tag{17}$$

for some quantity $\rho \to \infty$ to be determined later. We note that if $\Delta/(k^{\frac{3}{2}}\rho\beta^{\frac{1}{2}}(1+p/n)^{\frac{1}{2}}) \to \infty$, the set $\{j \in [k] : \sigma_j - \sigma_{j+1} \ge \rho\sqrt{n+p}\}$ is not empty. Otherwise, this would imply $\sigma_1 \le k\rho\sqrt{n+p}$ which would contradict Proposition A.1.

Thus, r is the largest index in $[k]$ such that the corresponding singular value gap is greater than or equal to g. An immediate implication is

$$\max_{r+1\le j\le k} \sigma_j \le kg = k\rho\sqrt{n+p}. \tag{18}$$

We split $\hat{U}$ into $(\hat{U}_{1:r}, \hat{U}_{(r+1):k})$ where $\hat{U}_{1:r} = (\hat{u}_1, \ldots, \hat{u}_r)$ and $\hat{U}_{(r+1):k} = (\hat{u}_1, \ldots, \hat{u}_r)$. Recall that $\hat{P}_{\cdot,i} = \hat{U}\hat{Y}_{\cdot,i}, \forall i \in [n]$ and $\hat{\theta}_j = \hat{U}\hat{c}_j, \forall j \in [k]$. For each $i \in [n]$, we decompose $\hat{P}_{\cdot,i} = \hat{P}^{(1)}_{\cdot,i} + \hat{P}^{(2)}_{\cdot,i}$ where

$$\hat{P}^{(1)}_{\cdot,i} = \hat{U}_{1:r}\hat{U}^T_{1:r}\hat{P}_{\cdot,i}, \quad \text{and } \hat{P}^{(2)}_{\cdot,i} = \hat{U}_{(r+1):k}\hat{U}^T_{(r+1):k}\hat{P}_{\cdot,i}.$$

Similarly, for each $j \in [k]$, we have $\hat{\theta}_j = \hat{\theta}^{(1)}_j + \hat{\theta}^{(2)}_j$ where

$$\hat{\theta}^{(1)}_j = \hat{U}_{1:r}\hat{U}^T_{1:r}\hat{\theta}_j, \quad \text{and } \hat{\theta}^{(2)}_j = \hat{U}_{(r+1):k}\hat{U}^T_{(r+1):k}\hat{\theta}_j,$$

Due to the orthogonality of $\{\hat{u}_l\}_{l\in[k]}$, we have

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

Let $\rho'' = o(1)$ be a sequence whose value will be determined later. The above indicator function can be split into two: $\mathbb{I}\{\|\hat{\theta}^{(1)}_{z_i^*} - \hat{\theta}^{(1)}_a\|^2 - \rho''\Delta^2 - \|\hat{\theta}^{(2)}_{z_i^*}\|^2 \le -2\langle \hat{P}^{(1)}_{\cdot,i} - \hat{\theta}^{(1)}_{z_i^*}, \hat{\theta}^{(1)}_{z_i^*} - \hat{\theta}^{(1)}_a\rangle\}$ and $\mathbb{I}\{\rho''\Delta^2 \le 2\langle \hat{P}^{(2)}_{\cdot,i}, \hat{\theta}^{(2)}_a - \hat{\theta}^{(2)}_{z_i^*}\rangle\}$. Hence, we have that

$$T_{i,a} \le \mathbb{I}\left\{\left\|\hat{\theta}^{(1)}_{z_i^*} - \hat{\theta}^{(1)}_a\right\| - \frac{\rho''\Delta^2 + \left\|\hat{\theta}^{(2)}_{z_i^*}\right\|^2}{\left\|\hat{\theta}^{(1)}_{z_i^*} - \hat{\theta}^{(1)}_a\right\|} \le 2\left\|\hat{P}^{(1)}_{\cdot,i} - \hat{\theta}^{(1)}_{z_i^*}\right\|\right\} + \mathbb{I}\left\{\rho''\Delta^2 \le 2\left\langle \hat{P}^{(2)}_{\cdot,i}, \hat{\theta}^{(2)}_a - \hat{\theta}^{(2)}_{z_i^*}\right\rangle\right\},$$

where we use the Cauchy-Schwarz inequality. We are going to simplify the terms inside the above indicator functions.

- Recall $\hat{\theta}_j$ is determined by $\hat{z}$ as in Equation (15). Define $\hat{Z} \in \{0,1\}^{n\times k}$ to be the estimated label matrix. That is

$$\hat{Z}_{i,j} = \mathbb{I}\left\{\hat{z}_i = j\right\}, \forall i \in [n], j \in [k].$$

Then Equation (15) is equivalent to be stated in matrix form:

$$\hat{\theta}_j = \frac{\hat{P}\hat{Z}_{\cdot,j}}{|\{i\in[n]:\hat{z}_i=j\}|} = \frac{\hat{U}\hat{\Sigma}\hat{V}^T\hat{Z}_{\cdot,j}}{|\{i\in[n]:\hat{z}_i=j\}|} = \frac{\sum_{l\in[k]}\hat{\sigma}_l\hat{u}_l\hat{v}_l^T\hat{Z}_{\cdot,j}}{|\{i\in[n]:\hat{z}_i=j\}|}, \quad \forall j \in [k].$$

As a result, we obtain that

$$\left|\left\langle \hat{u}_l, \hat{\theta}_j\right\rangle\right| = \frac{\left|\hat{\sigma}_l\hat{v}_l^T\hat{Z}_{\cdot,j}\right|}{|\{i\in[n]:\hat{z}_i=j\}|} \leq \frac{\hat{\sigma}_l\left\|\hat{v}_l\right\|\left\|\hat{Z}_{\cdot,j}\right\|}{|\{i\in[n]:\hat{z}_i=j\}|} = \frac{\hat{\sigma}_l}{\sqrt{|\{i\in[n]:\hat{z}_i=j\}|}}.$$

By Equation (18) and Lemma B.2, we have

$$\max_{r+1\leq l\leq k}\hat{\sigma}_j \leq \sqrt{2}(\sqrt{n}+\sqrt{p}) + \max_{r+1\leq l\leq k}\sigma_j \leq (k\rho+4)\sqrt{n+p}. \tag{19}$$

Lemma 3.2 shows that $\min_{j\in[k]}|\{i\in[n]:\hat{z}_i=j\}| \geq \frac{\beta n}{2k}$. Hence,

$$\max_{j\in[k]}\max_{r+1\leq l\leq k}\left|\left\langle \hat{u}_l, \hat{\theta}_j\right\rangle\right| \leq (k\rho+4)\sqrt{\frac{2k}{\beta}\left(1+\frac{p}{n}\right)}. \tag{20}$$

Consequently,

$$\max_{j\in[k]}\left\|\hat{\theta}_j^{(2)}\right\|^2 = \max_{j\in[k]}\left\|\hat{U}_{(r+1):k}\hat{U}_{(r+1):k}^T\hat{\theta}_j\right\|^2 = \max_{j\in[k]}\sum_{r+1\leq l\leq k}\left\langle \hat{u}_l, \hat{\theta}_j\right\rangle^2$$
$$\leq \frac{2k^2}{\beta}\left(1+\frac{p}{n}\right)(k\rho+4)^2. \tag{21}$$

- From Lemma 3.2, for any $a,b\in[k]$ such that $a\neq b$, we have $\|\hat{\theta}_b-\hat{\theta}_a\| \geq \|\theta_b^*-\theta_a^*\| - \|\hat{\theta}_b-\theta_b^*\| - \|\theta_a^*-\hat{\theta}_a\| \geq \Delta - 16\sqrt{2}\sqrt{\beta^{-1}k^2(1+p/n)}$. Using Equation (21), we have that

$$\min_{a,b\in[k]:a\neq b}\left\|\hat{\theta}_b^{(1)}-\hat{\theta}_a^{(1)}\right\| \geq \min_{a,b\in[k]:a\neq b}\left(\left\|\hat{\theta}_b-\hat{\theta}_a\right\| - \left\|\hat{\theta}_a^{(2)}\right\| - \left\|\hat{\theta}_b^{(2)}\right\|\right)$$
$$\geq \Delta - \left(16\sqrt{2}+2\sqrt{2}(k\rho+4)\right)\sqrt{\beta^{-1}k^2\left(1+\frac{p}{n}\right)}. \tag{22}$$

- Recall that $\hat{P}^{(2)}_{\cdot,i} = (\hat{U}_{(r+1):k}\hat{U}^T_{(r+1):k})\hat{P}_{\cdot,i} = \sum_{l=r+1}^k \hat{u}_l \hat{Y}_{l,i} = \sum_{l=r+1}^k \hat{u}_l\hat{\sigma}_l\hat{V}_{i,l}$ and $\hat{\theta}^{(2)}_a - \hat{\theta}^{(2)}_{z_i^*} = (\hat{U}_{(r+1):k}\hat{U}^T_{(r+1):k})(\hat{\theta}_a - \hat{\theta}_{z_i^*})$. We have that

$$\left\langle \hat{P}^{(2)}_{\cdot,i}, \hat{\theta}^{(2)}_a - \hat{\theta}^{(2)}_{z_i^*}\right\rangle = \left\langle \sum_{l=r+1}^k \hat{u}_l\hat{\sigma}_l\hat{V}_{i,l}, \hat{\theta}_a - \hat{\theta}_{z_i^*}\right\rangle = \sum_{l=r+1}^k \hat{\sigma}_l\hat{V}_{i,l}\left(\hat{u}_l^T\hat{\theta}_a - \hat{u}_l^T\hat{\theta}_{z_i^*}\right).$$

Note that $|\hat{u}_l^T\hat{\theta}_a - \hat{u}_l^T\hat{\theta}_{z_i^*}| \le 2\max_{j\in[k]}\max_{r+1\le l\le k}|\langle \hat{u}_l, \hat{\theta}_j\rangle|$. Using Equations (19) and (20) , we have

$$\left|\left\langle \hat{P}^{(2)}_{\cdot,i}, \hat{\theta}^{(2)}_a - \hat{\theta}^{(2)}_{z_i^*}\right\rangle\right| \le 2\left(k\rho+4\right)^2\sqrt{\frac{2nk}{\beta}\left(1+\frac{p}{n}\right)^2}\sum_{l=r+1}^k\left|\hat{V}_{i,l}\right|. \tag{23}$$

As a result, using Equations (21), (22) and (23), we have that

$$\begin{aligned}T_{i,a} \le& \mathbb{I}\Bigg\{\left(\Delta - \left(16\sqrt{2}+2\sqrt{2}\left(k\rho+6\right)\right)\sqrt{\beta^{-1}k^2\left(1+\frac{p}{n}\right)}\right)\\ &- \frac{\rho''\Delta^2 + \frac{2k^2}{\beta}\left(1+\frac{p}{n}\right)\left(k\rho+4\right)^2}{\Delta - \left(16\sqrt{2}+2\sqrt{2}\left(k\rho+4\right)\right)\sqrt{\beta^{-1}k^2\left(1+\frac{p}{n}\right)}} \le 2\left\|\hat{P}^{(1)}_{\cdot,i} - \hat{\theta}^{(1)}_{z_i^*}\right\|\Bigg\}\\ &+ \mathbb{I}\left\{\rho''\Delta^2 \le 4\left(k\rho+4\right)^2\sqrt{\frac{2nk}{\beta}\left(1+\frac{p}{n}\right)^2}\sum_{l=r+1}^k\left|\hat{V}_{i,l}\right|\right\}.\end{aligned}$$

For simplicity, define

$$\eta = \sqrt{1+p/n}.$$

Under the assumption, $\rho\to\infty$ and $\Delta/(k^2\rho\beta^{-1/2}\eta)\to\infty$, there exists some constant $c_1 > 0$, such that the above formula can be simplified into

$$\begin{aligned}T_{i,a} &\le \mathbb{I}\left\{\left(1 - c_1\rho'' - \frac{c_1k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\right)\Delta \le 2\left\|\hat{P}^{(1)}_{\cdot,i} - \hat{\theta}^{(1)}_{z_i^*}\right\|\right\} + \mathbb{I}\left\{\rho''\Delta^2 \le c_1k^{\frac{5}{2}}\rho^2\beta^{-\frac{1}{2}}\eta\sum_{l=r+1}^k\sqrt{2n}\left|\hat{V}_{i,l}\right|\right\}\\ &\le \mathbb{I}\left\{\left(1 - c_1\rho'' - \frac{c_1k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\right)\Delta \le 2\left\|\hat{P}^{(1)}_{\cdot,i} - \hat{\theta}^{(1)}_{z_i^*}\right\|\right\} + \sum_{l=r+1}^k\mathbb{I}\left\{k^{-1}\rho''\Delta^2 \le c_1k^{\frac{5}{2}}\rho^2\beta^{-\frac{1}{2}}\eta\sqrt{2n}\left|\hat{V}_{i,l}\right|\right\}.\end{aligned}$$

In the following, we are going to establish an upper bound for $\left\|\hat{P}^{(1)}_{\cdot,i} - \hat{\theta}^{(1)}_{z_i^*}\right\|$ to simplify $T_{i,a}$. We have

$$\begin{aligned}&\left\|\hat{P}^{(1)}_{\cdot,i} - \hat{\theta}^{(1)}_{z_i^*}\right\| \le \left\|\hat{P}^{(1)}_{\cdot,i} - \hat{U}_{1:r}\hat{U}^T_{1:r}\theta^*_{z_i^*}\right\| + \left\|\hat{\theta}^{(1)}_{z_i^*} - \hat{U}_{1:r}\hat{U}^T_{1:r}\theta^*_{z_i^*}\right\|\\ &\le \left\|\hat{P}^{(1)}_{\cdot,i} - \hat{U}_{1:r}\hat{U}^T_{1:r}\theta^*_{z_i^*}\right\| + \left\|\hat{\theta}_{z_i^*} - \theta^*_{z_i^*}\right\| \le \left\|\hat{P}^{(1)}_{\cdot,i} - \hat{U}_{1:r}\hat{U}^T_{1:r}\theta^*_{z_i^*}\right\| + 8\sqrt{2}\sqrt{\beta^{-1}k^2\left(1+\frac{p}{n}\right)},\end{aligned}$$

where the last inequality is due to Lemma 3.2. Since $\theta^*_{z_i^*} = P_{\cdot,i}$, we have that $\hat{P}^{(1)}_{\cdot,i} - \hat{U}_{1:r}\hat{U}_{1:r}^T\theta^*_{z_i^*} = (\hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P} - \hat{U}_{1:r}\hat{U}_{1:r}^T P)e_i$. Then, we have that

$$\begin{aligned}\hat{P}^{(1)}_{\cdot,i} - \hat{U}_{1:r}\hat{U}_{1:r}^T\theta^*_{z_i^*} &= \left(\hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P} - \hat{U}_{1:r}\hat{U}_{1:r}^T P^*\right)VV^Te_i + \left(\hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P} - \hat{U}_{1:r}\hat{U}_{1:r}^T P\right)\left(I - VV^T\right)e_i \\ &= \hat{U}_{1:r}\hat{U}_{1:r}^T\left(\hat{P} - P\right)VV^Te_i + \hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P}\left(I - VV^T\right)e_i.\end{aligned}$$

We first bound the euclidean norm of $\hat{U}_{1:r}\hat{U}_{1:r}^T(\hat{P} - P)VV^Te_i$. Proposition A.1 implies that the projection $(\hat{P} - P)VV^T$ has k unique columns, and two columns are the same when the two corresponding indexes have the same value in z^*. This leads to

$$\begin{aligned}&\sqrt{\left|i' \in [n] : z^*_{i'} = z^*_i\right|}\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\left(\hat{P} - P\right)VV^Te_i\right\| \\ &\leq \sqrt{\left|i' \in [n] : z^*_{i'} = z^*_i\right|}\left\|\left(\hat{P} - P\right)VV^Te_i\right\| \leq \left\|\left(\hat{P} - P\right)VV^T\right\|_{\mathrm{F}} \leq \left\|\hat{P} - P\right\|_{\mathrm{F}}.\end{aligned}$$

By Lemma 3.2, we have that

$$\left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\left(\hat{P} - P\right)VV^Te_i\right\| \leq \frac{4\sqrt{k}(\sqrt{n} + \sqrt{p})}{\sqrt{\beta n/k}} = 4\sqrt{\beta^{-1}k^2}\left(1 + \sqrt{\frac{p}{n}}\right).$$

As a result, we obtain that

$$\begin{aligned}\left\|\hat{P}^{(1)}_{\cdot,i} - \hat{\theta}^{(1)}_{z_i^*}\right\| &\leq \left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\left(\hat{P} - P^*\right)VV^Te_i\right\| + \left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P}\left(I - VV^T\right)e_i\right\| + 8\sqrt{2}\sqrt{\beta^{-1}k^2\left(1 + \frac{p}{n}\right)} \\ &\leq \left\|\hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P}\left(I - VV^T\right)e_i\right\| + 20\sqrt{\beta^{-1}k^2}\left(1 + \sqrt{\frac{p}{n}}\right).\end{aligned}$$

As we can see, the above expression is all about the leading r singular values and vectors. We are going to use $\hat{\Sigma}_{r\times r}, \hat{V}_{1:r}, \Sigma_{r\times r}, V_{1:r}$ to denote the corresponding matrices and their population counterparts. Define $\hat{\Sigma}_{r\times r} = \mathrm{diag}\{\hat{\sigma}_1, \ldots, \hat{\sigma}_r\}, \Sigma_{r\times r} = \mathrm{diag}\{\sigma_1, \ldots, \sigma_r\}$ to be two $r \times r$ diagonal matrices. Also define $\hat{V}_{1:r} = (\hat{v}_1, \ldots, \hat{v}_r)$ and $V_{1:r} = (v_1, \ldots, v_r)$. Then, we have that

$$\left\|\hat{P}^{(1)}_{\cdot,i} - \hat{\theta}^{(1)}_{z_i^*}\right\| \leq \left\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T\left(I - VV^T\right)e_i\right\| + 20\sqrt{\beta^{-1}k^2\left(1 + \frac{p}{n}\right)}.$$

Consequently, we obtain that

$$
\begin{aligned}
T_{i,a} &\leq \mathbb{I}\left\{\left(1-c_1\rho''-\frac{c_1k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\right)\Delta-20\sqrt{\beta^{-1}k^2\left(1+\frac{p}{n}\right)}\leq 2\left\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T\left(I-VV^T\right)e_i\right\|\right\}\\
&\quad+\sum_{l=r+1}^{k}\mathbb{I}\left\{\rho''\Delta^2\leq c_1k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta\sqrt{2n}\left|\hat{V}_{i,l}\right|\right\}\\
&\leq \mathbb{I}\left\{\left(1-c_1\rho''-\frac{c_2k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\right)\Delta\leq 2\left\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T\left(I-VV^T\right)e_i\right\|\right\}\\
&\quad+\sum_{l=r+1}^{k}\mathbb{I}\left\{\rho''\Delta^2\leq c_1k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta\sqrt{2n}\left|\hat{V}_{i,l}\right|\right\},
\end{aligned}
$$

for some constant $c_2 > 0$. Since $\ell(\hat{z}, z^*) = \sum_{i\in[n]}\sum_{a\neq z_i^*} T_{i,a}$ and the upper bound we establish above on $T_{i,a}$ does not depend on a, we have that

$$
\begin{aligned}
\ell(\hat{z},z^*)\leq k\sum_{i\in[n]}\mathbb{I}&\left\{\left(1-c_1\rho''-\frac{c_2k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\right)\Delta\leq 2\left\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T\left(I-VV^T\right)e_i\right\|\right\}\\
&+k\sum_{i\in[n]}\sum_{l=r+1}^{k}\mathbb{I}\left\{\rho''\Delta^2\leq c_1k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta\sqrt{2n}\left|\hat{V}_{i,l}\right|\right\}\triangleq A+B.
\end{aligned}
$$

Hence, we complete the key decomposition $\ell(\hat{z}, z^*) \leq A + B$.

We are going to provide upper bounds of $\mathbb{E}A$ and $\mathbb{E}B$ respectively in Section 3.4.2 and Section 3.4.3. To be more precise, we let $\mathbb{I}\{\mathcal{F}'\}$ be the indicator function for the event $\mathcal{F}'$. Here $\mathcal{F}' = \mathcal{F} \cup \mathcal{H}_1 \cup \mathcal{H}_2$, where $\mathcal{H}_1, \mathcal{H}_2$ are two more random events we will define later in the proof that also hold with high probability. Once we have an upper bound on $\mathbb{E}\ell(\hat{z}, z^*)\mathbb{I}\{\mathcal{F}'\} \leq \mathbb{E}A\mathbb{I}\{\mathcal{F}'\} + \mathbb{E}B\mathbb{I}\{\mathcal{F}'\}$, we apply Markov's inequality to have a with-high-probability result for $\ell(\hat{z}, z^*)\mathbb{I}\{\mathcal{F}'\}$. Then a union bound with $\mathbb{P}(\mathcal{F}')$ leads to the desired result.

Before starting Section 3.4.2 and Section 3.4.3, we introduce a few more notation. We define

$$
A_i=\mathbb{I}\left\{\left(1-c_1\rho''-\frac{c_2k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\right)\Delta\leq 2\left\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T\left(I-VV^T\right)e_i\right\|\right\},\forall i\in[n],
$$

and

$$
B_{i,l}=\mathbb{I}\left\{\rho''\Delta^2\leq c_1k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta\sqrt{2n}\left|\hat{V}_{i,l}\right|\right\},\forall i\in[n],\forall r+1\leq l\leq k.
$$

With these definitions, we have $A = k\sum_i A_i$ and $B = \sum_{l=r+1}^{k} B_l$ where $B_l = k\sum_i B_{i,l}$.

3.4.2. *Upper Bounds on* $\mathbb{E}A$. In this section, we focus on studying one A_i. Once its behavior is well understood, we can easily generalize it to obtain an upper bound on $\mathbb{E}A = \sum_i \mathbb{E}A_i$. For any $i \in [n]$, A_i can be written as

$$A_i = \mathbb{I}\left\{\left(1 - c_1\rho'' - \frac{c_2 k^2 \rho \beta^{-\frac{1}{2}}\eta}{\Delta}\right)\Delta \le 2 \sup_{w\in\mathbb{R}^r:\|w\|\le 1} e_i^T\left(I - VV^T\right)\hat{V}_{1:r}\hat{\Sigma}_{r\times r} w\right\},$$

where in the last equation we used the symmetry of $I - VV^T$ and $\hat{\Sigma}_{r\times r}$. For any unit vector $w \in \mathbb{R}^r$, define $w' = \Sigma_{r\times r}^{-1}\hat{\Sigma}_{r\times r}w$. In this way, we have the identity $\Sigma_{r\times r}w' = \hat{\Sigma}_{r\times r}w$ and also $\|w'\| \le 1 + 6\rho^{-1}$. The latter one is due to Lemma B.2. For any coordinate of w, w', Lemma B.2 shows that

$$\max_{j\in[r]}\left|\frac{w_j'}{w_j}\right| = \max_{j\in[r]}\frac{\hat{\sigma}_j}{\sigma_j} \le \max_{j\in[r]}\frac{\sigma_j + 4\sqrt{n+p}}{\sigma_j} \le 1 + 6\rho^{-1}.$$

Thus,

$$A_i \le \mathbb{I}\left\{\left(1 - c_1\rho'' - \frac{c_2 k^2 \rho \beta^{-\frac{1}{2}}\eta}{\Delta}\right)\Delta \le 2 \sup_{w\in\mathbb{R}^r:\|w\|\le 1+6\rho^{-1}} e_i^T\left(I - VV^T\right)\hat{V}_{1:r}\Sigma_{r\times r} w\right\},$$

and consequently,

$$\begin{aligned}\sum_{i\in[n]} A_i &\le \sum_{i\in[n]}\mathbb{I}\left\{\left(1 - c_1\rho'' - \frac{c_2 k^2 \rho \beta^{-\frac{1}{2}}\eta}{\Delta}\right)\Delta \le 2 \sup_{w\in\mathbb{R}^r:\|w\|\le 1+6\rho^{-1}} e_i^T\left(I - VV^T\right)\hat{V}_{1:r}\Sigma_{r\times r} w\right\}\\ &= \sum_{i\in[n]}\mathbb{I}\left\{\frac{1 - c_1\rho'' - \frac{c_2 k^2 \rho \beta^{-\frac{1}{2}}\eta}{\Delta}}{1+6\rho^{-1}}\Delta \le 2 \sup_{w\in\mathbb{R}^r:\|w\|\le 1} e_i^T\left(I - VV^T\right)\hat{V}_{1:r}\Sigma_{r\times r} w\right\}.\end{aligned}$$

In the following part, we will focus on investigating $e_i^T\left(I - VV^T\right)\hat{V}_{1:r}\Sigma_{r\times r}w$. First we are going to give a partition of the leading $[r]$ singular values. Let s be the cardinarlity of the set

$$\left\{l \in [r] : \frac{\sigma_l - \sigma_{l+1}}{\sigma_{l+1}} \ge \frac{1}{\rho' k}\right\}, \tag{24}$$

for some $\rho' \to \infty$ whose value will be specified later. Denote all the entries it contains as $j_1' < j_2' < \ldots < j_s'$ from the smallest one to the largest one. It is easy to check that $j_s' = r$, by Equation (18). Define $j_0' = 0$ and

$$j_m = j_{m-1}' + 1, \forall m \in [s].$$

In this way, we split $[r]$ into disjoint sets $\{J_m\}_{m=1}^s$ where $J_m = \{j_m, j_m + 1, \ldots, j_m'\}$. Such partition enjoys the following properties:

- Define the singular value gaps among $J_1, \dots, J_s$ and $(\cup_{m=1}^s J_m)^C$ to be
$$g_m = \min\left\{\sigma_{j'_{m-1}} - \sigma_{j_m}, \sigma_{j'_m} - \sigma_{j_{m+1}}\right\}, \forall m \in [s],$$
with $j_{s+1} = r+1$ and $\sigma_0 = +\infty$. By Equation (24), for any $m \in [s-1]$, we have $\sigma_{j'_m} - \sigma_{j_{m+1}} = \sigma_{j'_m} - \sigma_{j'_m+1} \geq \frac{\sigma_{j'_m+1}}{\rho' k} \geq \frac{\sigma_r}{\rho' k} \geq \frac{\rho\sqrt{n+p}}{\rho' k}$. Then,
$$\min_{m\in[s]} g_m \geq \frac{\rho\sqrt{n+p}}{\rho' k}. \tag{25}$$
- The set defined in Equation (24) is equivalent to $\{l \in [r] : \sigma_l/\sigma_{l+1} > 1+1/(\rho' k)\}$. As a result, for any $l \in J_m$, we have $\sigma_l/\sigma_{l+1} < 1+1/(\rho' k)$. Under the assumption $\rho' \to \infty$, we have that
$$\max_{m\in[s]} \frac{\sigma_{j_m}}{\sigma_{j'_m}} \leq \left(1 + \frac{1}{\rho' k}\right)^{|J_m|} \leq \left(1 + \frac{1}{\rho' k}\right)^{k} \leq 1 + \frac{2}{\rho'}. \tag{26}$$
- We have that $\max_{m\in[s]} \sigma_{j'_m}/(\sigma_{j'_m} - \sigma_{j'_m+1}) \leq \max_{m\in[s]}(1 + \sigma_{j'_m+1}/(\sigma_{j'_m} - \sigma_{j'_m+1})) \leq 1 + \rho' k$ due to Equation (24). Together with Equation (26), under the assumption $\rho' \to \infty$, we have
$$\max_{m\in[s]} \frac{\sigma_{j_m} - \sigma_{j'_m}}{g_m} \leq \frac{2}{\rho'} \max_{m\in[s]} \frac{\sigma_{j'_m}}{\sigma_{j'_m} - \sigma_{j'_m+1}} \leq \frac{2}{\rho'}\left(1 + \rho' k\right) \leq 3k, \tag{27}$$
and
$$\max_{m\in[s]} \frac{\sigma_{j_m}}{g_m} \leq \left(1 + \frac{2}{\rho'}\right) \max_{m\in[s]} \frac{\sigma_{j'_m}}{\sigma_{j'_m} - \sigma_{j'_m+1}} \leq 1 + 2\rho' k. \tag{28}$$

Now, consider any fixed $w \in \mathbb{R}^r$. For $m \in [s]$, we define $\hat{V}_{J_m} = (\hat{v}_{j_m}, \dots, \hat{v}_{j'_m}) \in \mathbb{R}^{n\times|J_m|}$, ${}_{J_m} = (v_{j_m}, \dots, v_{j'_m}) \in \mathbb{R}^{n\times|J_m|}$, $\Sigma_{J_m\times J_m} = \mathrm{diag}\{\sigma_{j_m}, \dots, \sigma_{j'_m}\} \in \mathbb{R}^{|J_m|\times|J_m|}$, and $w_{J_m} = (w_{j_m}, \dots, w_{j'_m}) \in \mathbb{R}^{|J_m|}$. Then, we have that
$$e_i^T\left(I - VV^T\right)\hat{V}_{1:r}\Sigma_{r\times r} w = \sum_{m\in[s]} e_i^T\left(I - VV^T\right)\hat{V}_{J_m}\Sigma_{J_m\times J_m} w. \tag{29}$$

For any $m \in [s]$, by the Davis-Kahan-Wedin $sin(\Theta)$ Theorem (see Lemma B.3), there exists an orthonormal matrix $O_m \in \mathbb{R}^{|J_m|\times|J_m|}$ such that
$$\begin{aligned}\left\|\hat{V}_{J_m} - V_{J_m}O_m\right\| &\leq \sqrt{2}\left\|\hat{V}_{J_m}\hat{V}_{J_m}^T - V_{J_m}V_{J_m}^T\right\| \\ &\leq \frac{4\sqrt{2}\,\|E\|}{g_m} \leq \frac{16\sqrt{2}\sqrt{n+p}}{\rho'^{-1}k^{-1}\rho\sqrt{n+p}} = \frac{16\sqrt{2}\rho' k}{\rho},\end{aligned} \tag{30}$$

by Equation (25). Since $\|\hat{V}_{J_m}^T V_{J_m} O_m - I\| = \|(\hat{V}_{J_m} - V_{J_m} O_m)^T V_{J_m} O_m\|$, we have that

$$\left\|\hat{V}_{J_m}^T V_{J_m} O_m - I\right\| \leq \left\|\hat{V}_{J_m} - V_{J_m} O_m\right\| \leq \frac{16\sqrt{2}\rho' k}{\rho}. \tag{31}$$

Assuming that $\rho/(\rho' k) \to \infty$, $O_m^T V_{J_m}^T \hat{V}_{J_m}$ is invertible. Now, we define $w' = \left(w'_{J_1}, \ldots, w'_{J_s}\right) \in \mathbb{R}$ such that

$$w'_{J_m} = \Sigma_{J_m \times J_m}^{-1} O_m \left(\hat{V}_{J_m}^T V_{J_m} O_m\right)^{-1} \Sigma_{J_m \times J_m} w_{J_m}, \forall m \in [s], \tag{32}$$

which implies $w_{J_m} = \Sigma_{J_m \times J_m}^{-1} (\hat{V}_{J_m}^T V_{J_m} O_m) O_m^T \Sigma_{J_m \times J_m} w'_{J_m} = \Sigma_{J_m \times J_m}^{-1} \hat{V}_{J_m}^T V_{J_m} \Sigma_{J_m \times J_m} w'_{J_m}$. Plugging the above it into Equation (29), we have that

$$e_i^T \left(I - VV^T\right) \hat{V}_{1:r} \Sigma_{r \times r} w = \sum_{m \in [s]} e_i^T \left(I - VV^T\right) \hat{V}_{J_m} \hat{V}_{J_m}^T V_{J_m} \Sigma_{J_m \times J_m} w'_{J_m}.$$

Due to Equations (26) and (31), we have that

$$\begin{aligned} \max_{m \in [s]} \frac{\left\|w'_{J_m}\right\|}{\|w_{J_m}\|} &\leq \max_{m \in [s]} \left\|\Sigma_{J_m \times J_m}^{-1} O_m \left(\hat{V}_{J_m}^T V_{J_m} O_m\right)^{-1} \Sigma_{J_m \times J_m}\right\| \\ &\leq \max_{m \in [s]} \left\|\Sigma_{J_m \times J_m}^{-1}\right\| \|\Sigma_{J_m \times J_m}\| \|O_m\| \left\|\left(I + \left(\hat{V}_{J_m}^T V_{J_m} O_m - I\right)\right)^{-1}\right\| \\ &\leq \left(1 + 2\rho'^{-1}\right) \max_{m \in [s]} \left\|\left(I + \left(\hat{V}_{J_m}^T V_{J_m} O_m - I\right)\right)^{-1}\right\|, \end{aligned}$$

where in the last inequality we used $\max_{m \in [s]} \left\|\Sigma_{J_m \times J_m}^{-1}\right\| \|\Sigma_{J_m \times J_m}\| = \max_{m \in [s]} \sigma_{j_m} / \sigma_{j'_m} \leq 1 + 2\rho'^{-1}$ by Equation (26). Thus, using also Equation (31) we obtain that

$$\max_{m \in [s]} \frac{\left\|w'_{J_m}\right\|}{\|w_{J_m}\|} \leq \max_{m \in [s]} \frac{1 + 2\rho'^{-1}}{1 - \left\|\hat{V}_{J_m}^T V_{J_m} O_m - I\right\|} \leq \frac{1 + 2\rho'^{-1}}{1 - 24\sqrt{2}\rho' k \rho^{-1}}.$$

This immediately leads to $\|w'\| = \frac{\|w'\|}{\|w\|} \|w\| \leq \frac{1 + 2\rho'^{-1}}{1 - 24\sqrt{2}\rho' k \rho^{-1}}$. Hence,

$$\begin{aligned} &\sup_{w \in \mathbb{R}^r : \|w\| \leq 1} w^T \Sigma_{r \times r} \hat{V}_{1:r}^T \left(I - VV^T\right) e_i \\ &\leq \frac{1 + 2\rho'^{-1}}{1 - 24\sqrt{2}\rho' k \rho^{-1}} \sup_{w \in \mathbb{R}^r : \|w\| \leq 1} \sum_{m \in [s]} e_i^T \left(I - VV^T\right) \left(\hat{V}_{J_m} \hat{V}_{J_m}^T - V_{J_m} V_{J_m}^T\right) V_{J_m} \Sigma_{J_m \times J_m} w_{J_m}. \end{aligned}$$

where we also used that $\left(I - VV^T\right) V_{J_m} = 0$. As a consequence, we obtain that

$$\sum_{i\in[n]} A_i = \sum_{i\in[n]} \mathbb{I}\Bigg\{ \frac{1-24\sqrt{2}\rho' k\rho^{-1}}{\left(1+6\rho^{-1}\right)\left(1+2\rho'^{-1}\right)} \left(1 - c_1\rho'' - \frac{c_2 k^2 \rho \beta^{-\frac{1}{2}}\eta}{\Delta}\right)\Delta$$

$$\leq 2 \sup_{w\in\mathbb{R}^r:\|w\|\leq 1} \sum_{m\in[s]} e_i^T \left(I - VV^T\right)\left(\hat{V}_{J_m}\hat{V}_{J_m}^T - V_{J_m}V_{J_m}^T\right) V_{J_m}\Sigma_{J_m\times J_m} w_{J_m} \Bigg\}.$$

We continue to simplify the term on the right hand side of what is inside the indicator function. From Lemma 3.3, we have the following decomposition such that for all $m \in [s]$,

$$\left(I - VV^T\right)\left(\hat{V}_{J_m}\hat{V}_{J_m}^T - V_{J_m}V_{J_m}^T\right) V_{J_m} = \sum_{l\in J_m} \frac{1}{\sigma_l}\left(I - VV^T\right) E^T u_l v_l^T V_{J_m} + S_m. \tag{33}$$

Hence, we obtain that

$$\begin{aligned} &e_i^T \left(I - VV^T\right)\left(\hat{V}_{J_m}\hat{V}_{J_m}^T - V_{J_m}V_{J_m}^T\right) V_{J_m}\Sigma_{J_m\times J_m} w_{J_m} \\ =& \sum_{l\in J_m} w_l e_i^T \left(I - VV^T\right) E^T u_l + e_i^T \mathbb{E} S_m \Sigma_{J_m\times J_m} w_{J_m} + e_i^T \left(S_m - \mathbb{E}S_m\right)\Sigma_{J_m\times J_m} w_{J_m}, \end{aligned}$$

where we use $v_l^T V_{J_m}\Sigma_{J_m\times J_m} w_{J_m} = \sigma_l w_l$ in the last equation. We are going to show the middle term above is 0, as $\mathbb{E}S_m = 0$. Taking the expectation on both sides of Equation (33), we have that

$$\mathbb{E}S_m = (I - VV^T)(\mathbb{E}(\hat{V}_{J_m}\hat{V}_{J_m}^T) - V_{J_m}V_{J_m}^T)V_{J_m} = (I - VV^T)\mathbb{E}(\hat{V}_{J_m}\hat{V}_{J_m}^T)V_{J_m}.$$

Further investigation of the last term yields that

$$\begin{aligned} \mathbb{E}S_m &= \left(I - VV^T\right)\mathbb{E}\left(\sum_{j\in J_m} \hat{v}_j\hat{v}_j^T\right) V_{J_m} = \sum_{j\in J_m} \mathbb{E}\left(\left(I - VV^T\right)\hat{v}_j\right)\left(V_{J_m}^T\hat{v}_j\right)^T \\ &= \sum_{j\in J_m} \mathbb{E}\left(\frac{\left(I - VV^T\right)\hat{v}_j}{\left\|\left(I - VV^T\right)\hat{v}_j\right\|}\right)\left(\left\|\left(I - VV^T\right)\hat{v}_j\right\| V_{J_m}^T VV^T\hat{v}_j\right)^T, \end{aligned}$$

where the last inequality is due to $VV^TV_{J_m} = V_{J_m}$. By Lemma 3.4, we have that $\left(I - VV^T\right)\hat{v}_j/\left\|\left(I - VV^T\right)\hat{v}_j\right\|$ and $\left\|\left(I - VV^T\right)\hat{v}_j\right\| V_{J_m}^T VV^T\hat{v}_j$

are independent for all $j \in [r]$. Hence, using Lemma 3.4 again to obtain that $\mathbb{E}\frac{\left(I-VV^T\right)\hat{v}_j}{\|(I-VV^T)\hat{v}_j\|} = 0$, we have that

$$\mathbb{E}S_m = \sum_{j\in J_m} \mathbb{E}\left(\frac{\left(I-VV^T\right)\hat{v}_j}{\left\|\left(I-VV^T\right)\hat{v}_j\right\|}\right)\mathbb{E}\left(\left\|\left(I-VV^T\right)\hat{v}_j\right\| V_{J_m}^T VV^T\hat{v}_j\right)^T = 0.$$

Hence, we obtain that

$$\begin{aligned}&\sup_{w\in\mathbb{R}^r:\|w\|\le 1}\sum_{m\in[s]} e_i^T\left(1-VV^T\right)\left(\hat{V}_{J_m}\hat{V}_{J_m}^T - V_{J_m}V_{J_m}^T\right)V_{J_m}\Sigma_{J_m\times J_m}w_{J_m}\\ &\le \sup_{w\in\mathbb{R}^r:\|w\|\le 1} e_i^T\left(I-VV^T\right)E^TU_{1:r}w + \sup_{w\in\mathbb{R}^r:\|w\|\le 1}\sum_{m\in[s]} e_i^T\left(S_m-\mathbb{E}S_m\right)\Sigma_{J_m\times J_m}w_{J_m}.\end{aligned}$$

Since $\sup_{w\in\mathbb{R}^r:\|w\|\le 1} e_i^T\left(I-VV^T\right)E^TU_{1:r}w = \left\|U_{1:r}^TE\left(I-VV^T\right)e_i\right\|$, we have that

$$\begin{aligned}\sum_{i\in[n]}A_i = \sum_{i\in[n]}\mathbb{I}\Bigg\{&\frac{1-24\sqrt{2}\rho' k\rho^{-1}}{\left(1+6\rho^{-1}\right)\left(1+2\rho'^{-1}\right)}\left(1-c_1\rho''-\frac{c_2k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\right)\Delta\\ &\le 2\left\|U_{1:r}^TE\left(I-VV^T\right)e_i\right\| + \sup_{w\in\mathbb{R}^r:\|w\|\le 1}\sum_{m\in[s]} e_i^T\left(S_m-\mathbb{E}S_m\right)\Sigma_{J_m\times J_m}w_{J_m}\Bigg\}.\end{aligned}$$

We are now going to further investigate the two terms above.

First term. We have that $U_{1:r}^TE\left(I-VV^T\right)e_i \sim \mathcal{N}\left(0, \left\|\left(I-VV^T\right)e_i\right\|^2 I_{r\times r}\right)$. Moreover, since $\left\|\left(I-VV^T\right)e_i\right\| \le 1$, we have that $\left\|U_{1:r}^TE\left(I-VV^T\right)e_i\right\|^2$ is stochastically dominated by a χ_k^2 random variable. Hence, for some $\xi_i \sim \chi_k^2$, we have that

$$\left\|U_{1:r}^TE\left(I-VV^T\right)e_i\right\| \le \sqrt{\xi_i}.$$

Second term. By Lemma B.1, we have that $g_m \ge 8\mathbb{E}\|E\|$, and hence we can apply Lemma 3.3. Note that $\left\|\Sigma_{J_m\times J_m}w_{J_m}e_i^T\right\|_* = \left\|\Sigma_{J_m\times J_m}w_{J_m}\right\|\left\|e_i^T\right\| \le \sigma_{j_m}\left\|w_{J_m}\right\|$. Together with Equations (25), (27) and (28), for some constant $c_0 > 0$, we have with probability at least $1-2e^{-t^2}$ that

$$\begin{aligned}&\left|e_i^T\left(S_m-\mathbb{E}S_m\right)\Sigma_{J_m\times J_m}w_{J_m}\right| \le c_0\left(1+\frac{\sigma_{j_m}-\sigma_{j'_m}}{g_m}\right)\frac{t}{g}\left(\frac{\sqrt{n+p}+t}{g}\right)\sigma_{j_m}\left\|w_{J_m}\right\|\\ &\le c_0\rho^{-1}\left(1+2k\right)\left(1+2\rho' k\right)\rho' k\left(1+\frac{t}{\sqrt{n+p}}\right)t\left\|w_{J_m}\right\| \le 8c_0\rho^{-1}k^3\rho'^2\left(1+\frac{t}{\sqrt{n+p}}\right)t\left\|w_{J_m}\right\|.\end{aligned}$$

Taking $t = \Delta \wedge \sqrt{n+p}$, we obtain that $\left|e_i^T (S_m - \mathbb{E}S_m) \Sigma_{J_m \times J_m} w_{J_m}\right| \leq 16c_0\rho^{-1}k^3\rho'^2\Delta \|w_{J_m}\|$ holds with probability at least $1-2\exp\left(-\Delta^2\right)\mathbb{I}\{\Delta \leq \sqrt{n+p}\} - 2\exp(-n)\mathbb{I}\{\Delta > \sqrt{n+p}\}$. This implies that

$$\sum_{m\in[s]} e_i^T (S_m - \mathbb{E}S_m) \Sigma_{J_m \times J_m} w_{J_m} \leq 16c_0\rho^{-1}k^{\frac{7}{2}}\rho'^2\Delta,$$

with probability at least $1-2k\exp\left(-\Delta^2\right)\mathbb{I}\{\Delta \leq \sqrt{n+p}\} - 2k\exp(-n)\mathbb{I}\{\Delta > \sqrt{n+p}\}$, by the Cauchy–Schwarz inequality. By applying a standard ϵ-net argument with a union bound, we have that

$$\sup_{w\in\mathbb{R}^r:\|w\|\leq 1} \sum_{m\in[s]} e_i^T (S_m - \mathbb{E}S_m) \Sigma_{J_m \times J_m} w_{J_m} \leq 32c_0\rho^{-1}k^{\frac{7}{2}}\rho'^2\Delta$$

holds with probability at least

$$1 - 2ke^k \exp\left(-\Delta^2\right)\mathbb{I}\left\{\Delta \leq \sqrt{n+p}\right\} - 2ke^k\exp(-n)\mathbb{I}\left\{\Delta > \sqrt{n+p}\right\}.$$

Note that if $\Delta > \sqrt{n+p}$, we can use a union bound, such that the above inequality holds for all $i \in [n]$ with probability at least $1 - 2nke^k\exp(-n)$, which is greater than $1 - \exp(-n/2)$. Defining the event $\mathcal{H}_1$ as

$$\left\{\left\{\Delta > \sqrt{n+p}\right\} \cup \sum_{i\in[n]} \mathbb{I}\left\{32c_0\rho^{-1}k^{\frac{7}{2}}\rho'^2\Delta \leq \sup_{w\in\mathbb{R}^r:\|w\|\leq 1} \sum_{m\in[s]} e_i^T (S_m - \mathbb{E}S_m) \Sigma_{J_m \times J_m} w_{J_m}\right\} = 0\right\},$$

we have that $\mathbb{P}(\mathcal{H}_1) \geq 1 - 1 - 2nke^k\exp(-n)$. Assuming that $\mathcal{H}_1$ holds, we only need to deal with the case where $\Delta \leq \sqrt{n+p}$. For all $i \in [n]$, we have that

$$\mathbb{E}\mathbb{I}\left\{\mathcal{H}_1, \sup_{w\in\mathbb{R}^r:\|w\|\leq 1} \sum_{m\in[s]} e_i^T (S_m - \mathbb{E}S_m) \Sigma_{J_m \times J_m} w_{J_m} \geq 32c_0\rho^{-1}k^{\frac{7}{2}}\rho'^2\Delta\right\} \leq 2ke^k\exp\left(-\Delta^2\right).$$

Combining the two terms. As a consequence, we have that

$$\sum_{i\in[n]} A_i \leq \sum_{i\in[n]} \mathbb{I}\left\{\frac{\left(1-24\sqrt{2}\rho' k\rho^{-1}\right)\left(1 - c_1\rho'' - \frac{c_2k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\right)}{(1+6\rho^{-1})(1+2\rho'^{-1})}\Delta - 32c_0\rho^{-1}k^{\frac{7}{2}}\rho'^2\Delta \leq 2\|\xi_i\|\right\}$$

$$+ \sum_{i\in[n]} \mathbb{I}\left\{32c_0\rho^{-1}k^{\frac{7}{2}}\rho'^2\Delta \leq \sup_{w\in\mathbb{R}^r:\|w\|\leq 1} \sum_{m\in[s]} e_i^T (S_m - \mathbb{E}S_m) \Sigma_{J_m \times J_m} w_{J_m}\right\}.$$

Under the assumption that $\rho' \to \infty$, $\rho/(k^{7/2}\rho'^2) \to \infty$ and $\Delta/(k^2\rho\beta^{-1/2}\eta) \to \infty$, there exists a constant $c_3 > 0$, such that

$$\sum_{i\in[n]} A_i \le \sum_{i\in[n]} \mathbb{I}\left\{\left(1 - c_3\rho'' - \frac{c_3 k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta} - \frac{c_3 k^{\frac{7}{2}}\rho'^2}{\rho}\right)\Delta \le 2\left\|\xi_i\right\|\right\}$$
$$+ \sum_{i\in[n]} \mathbb{I}\left\{32c\rho^{-1}k^{\frac{7}{2}}\rho'^2\Delta \le \sup_{w\in\mathbb{R}^r:\|w\|\le 1} \sum_{m\in[s]} e_i^T\left(S_m - \mathbb{E}S_m\right)\Sigma_{J_m\times J_m} w_{J_m}\right\}.$$

Finally, using the tail probability of χ^2 distribution, we obtain that

$$\sum_{i\in[n]} \mathbb{E}A_i\mathbb{I}\left\{\mathcal{F}\cup\mathcal{H}_1\right\} \le n\exp\left(-\frac{1}{8}\left(1 - c_3\rho'' - \frac{c_3k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta} - \frac{c_3k^{\frac{7}{2}}\rho'^2}{\rho}\right)^2\Delta^2\right) + 2nke^k\exp\left(-\Delta^2\right),$$

as we assume that the events $\mathcal{F}$ and $\mathcal{H}_1$ hold in the above analysis.

3.4.3. *Upper Bounds on* $\mathbb{E}B$. In this section, we study $B_{i,l}$. Consider any $i \in [n], r+1 \le l \le k$. Recall that $B_{i,l} = \mathbb{I}\{\rho''\Delta^2 \le c_1 k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta\sqrt{2n}|\hat{V}_{i,l}|\}$. Note that

$$\left|\hat{V}_{i,l}\right| = \left|e_i^T\left(VV^T\hat{v}_l\right)\right| + \left|e_i^T\left(I - VV^T\right)\hat{v}_l\right| \le \sum_{j=1}^k \left|e_i^T v_j\right|\left|v_j^T\hat{v}_l\right| + \left|e_i^T\left(I - VV^T\right)\hat{v}_l\right|$$
$$\le \sum_{j=1}^k \left|e_i^T v_j\right| + \left|e_i^T\left(I - VV^T\right)\hat{v}_l\right| \le \sqrt{\beta^{-1}k^3/n} + \left|e_i^T\left(I - VV^T\right)\hat{v}_l\right|,$$

where the last inequality is due to Proposition A.2. Moreover, we have that

$$B_{i,l} \le \mathbb{I}\left\{\frac{\rho''\Delta^2}{c_1k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta} - \sqrt{2\beta^{-1}k^3} \le \sqrt{2n}\frac{\left|e_i^T\left(I - VV^T\right)\hat{v}_l\right|}{\left\|\left(I - VV^T\right)\hat{v}_l\right\|}\right\}$$
$$\overset{d}{=} \mathbb{I}\left\{\frac{\rho''\Delta^2}{c_1k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta} - \sqrt{2\beta^{-1}k^3} \le \sqrt{2n}\frac{\left|e_i^T\left(I - VV^T\right)\zeta_{i,l}\right|}{\left\|\left(I - VV^T\right)\zeta_{i,l}\right\|}\right\},$$

where $\zeta_{i,l} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_n)$ and where we applied Lemma 3.4. We further estimate

$$B_{i,l} \le \mathbb{I}\left\{\frac{\rho''\Delta^2}{c_1k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta} - \sqrt{2\beta^{-1}k^3} \le \sqrt{2n}\frac{\left|e_i^T\left(I - VV^T\right)\zeta_{i,l}\right|}{2\sqrt{n}}\right\} + \mathbb{I}\left\{\left\|\left(I - VV^T\right)\zeta_{i,l}\right\| > 2\sqrt{n}\right\}.$$

Since $\|e_i^T(I-VV^T)\zeta_{i,l}\|^2$ is stochastically dominated by χ_n^2, we have with probability at least $1-\exp(-n)$, that $\left\|e_i^T\left(I-VV^T\right)\zeta_{i,l}\right\| \leq 2\sqrt{n}$. Now, we define the event $\mathcal{H}_2$ as

$$\mathcal{H}_2 = \left\{ \sum_{i\in[n]} \sum_{l=r+1}^{n} \mathbb{I}\left\{\left\|\left(I-VV^T\right)\zeta_{i,l}\right\| > 2\sqrt{n}\right\} = 0 \right\}.$$

A union bound yields that $\mathbb{P}(\mathcal{H}_2) \geq 1 - 2nk\sqrt{n}$. Then, assuming that $\mathcal{H}_2$ holds, we have that

$$B_{i,l} \leq \mathbb{I}\left\{ \frac{\rho''\Delta^2}{c_1 k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta} - \sqrt{2\beta^{-1}k^3} \leq \sqrt{2n}\frac{\left|e_i^T\left(I-VV^T\right)\zeta_{i,l}\right|}{2\sqrt{n}} \right\}.$$

Note that $\left|e_i^T\left(I-VV^T\right)\zeta_{i,l}\right|$ is stochastically dominated by the absolute value of an univariate standard Gaussian random variable. Hence, using the tail probability of the standard normal distribution, under the assumption that $\rho''\Delta/\left(k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta\right) \to \infty$, there exists a constant $c_4 > 0$, such that

$$\mathbb{E}B_{i,l}\mathbb{I}\left\{\mathcal{F}\cup\mathcal{H}_2\right\} \leq \exp\left(-c_4 \left(\frac{\rho''\Delta}{k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta} \right)^2 \Delta^2 \right).$$

Hence, we obtain that

$$\sum_{i=1}^{n} \sum_{l=r+1}^{k} \mathbb{E}B_{i,l}\mathbb{I}\left\{\mathcal{F}\cup\mathcal{H}_2\right\} \leq nk\exp\left(-c_4 \left(\frac{\rho''\Delta}{k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta} \right)^2 \Delta^2 \right).$$

3.4.4. *Obtaining the final Result.* Combining the above upper bounds together, we have that

$$\begin{aligned}
&\mathbb{E}\ell(\hat{z},z^*)\mathbb{I}\left\{\mathcal{F}\cup\mathcal{H}_1\cup\mathcal{H}_2\right\} \leq \sum_{i\in[n]} \mathbb{E}A_i\mathbb{I}\left\{\mathcal{F}\cup\mathcal{H}_1\right\} + \sum_{i=1}^{n}\sum_{l=r+1}^{k} \mathbb{E}B_{i,l}\mathbb{I}\left\{\mathcal{F}\cup\mathcal{H}_2\right\} \\
&\leq n\exp\left(-\frac{1}{8}\left(1 - c_3\rho'' - \frac{c_3k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta} - \frac{c_3k^{\frac{7}{2}}\rho'^2}{\rho} \right)^2 \Delta^2 \right) + nke^k\exp\left(-\Delta^2\right) \\
&\quad + nk\exp\left(-c_4 \left(\frac{\rho''\Delta}{k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta} \right)^2 \Delta^2 \right).
\end{aligned}$$

Under the assumption

$$\frac{\Delta}{k^{10.5}\beta^{-0.5}\eta} \to \infty,$$

we can take

$$\rho = \frac{k^{\frac{7}{2}}}{8c_3}\left(\frac{\Delta}{k^{10.5}\beta^{-0.5}\eta}\right)^{0.3}, \quad \rho' = \frac{1}{8c_3}\left(\frac{\Delta}{k^{10.5}\beta^{-0.5}\eta}\right)^{0.1}, \text{ and } \rho'' = \frac{1}{8c_3}\left(\frac{\Delta}{k^{10.5}\beta^{-0.5}\eta}\right)^{-0.1},$$

such that

$$\mathbb{E}\ell(\hat{z}, z^*)\mathbb{I}\left\{\mathcal{F}\cup\mathcal{H}_1\cup\mathcal{H}_2\right\} \leq n\exp\left(-\left(1-\frac{1}{2}\left(\frac{\Delta}{k^{10.5}\beta^{-0.5}\eta}\right)^{-0.1}\right)\frac{\Delta^2}{8}\right).$$

Applying Markov's inequality, we obtain that

$$\ell(\hat{z}, z^*)\mathbb{I}\left\{\mathcal{F}\cup\mathcal{H}_1\cup\mathcal{H}_2\right\} \leq n\exp\left(-\left(1-\left(\frac{\Delta}{k^{10.5}\beta^{-0.5}\eta}\right)^{-0.1}\right)\frac{\Delta^2}{8}\right),$$

with probability at least $1-\exp(-\Delta)-\exp(-n/2)$. Finally, the proof is completed by using a union bound accounting for the events $\mathcal{F}, \mathcal{H}_1$ and $\mathcal{H}_2$.

Acknowledgments. We would like to thank Zhou Fan from Yale University for pointing out the references [30, 49], which leads to the establishment of Lemma 3.4.

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to "Optimality of Spectral Clustering for Gaussian Mixture Model"
(url to be specified). In the supplement [40], we first present some propositions that characterize the population quantities in Appendix A. Then in Appendix B, we give several auxiliary lemmas related to the noise matrix E. In Appendix C, we include proofs of Lemma 3.1, Lemma 3.2 and Lemma 3.4. The proof of Lemma 3.3 is given in Appendix D.

References.

[1] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

[2] So-Zen Yao Charles J Alpert. Spectral partitioning: the more eigenvectors, the better. In *32nd Design Automation Conference*, pages 195–200. IEEE, 1995.

[3] Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research*, 15(1):2239–2312, 2014.
[4] Francis R Bach and Michael I Jordan. Learning spectral clustering, with application to speech separation. *Journal of Machine Learning Research*, 7(Oct):1963–2001, 2006.
[5] Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems*, pages 954–962, 2011.
[6] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.
[7] Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory*, pages 35–1, 2012.
[8] Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.
[9] S. Dasgupta. The hardness of k-means clustering. *Department of Computer Science and Engineering, University of California, San Diego*, 2008.
[10] K.R. Davison and S.J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces*, volume 1, pages 317–366. North-Holland, Amsterdam, 2001.
[11] Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.
[12] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 606–610. SIAM, 2005.
[13] Chris HQ Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst D Simon. A min-max cut algorithm for graph partitioning and data clustering. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 107–114. IEEE, 2001.
[14] William E Donath and Alan J Hoffman. Lower bounds for the partitioning of graphs. In *Selected Papers Of Alan J Hoffman: With Commentary*, pages 437–442. World Scientific, 2003.
[15] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.
[16] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.
[17] Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013.
[18] Saduoki Furui. Unsupervised speaker adaptation based on hierarchical spectral clustering. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(12):1923–1930, 1989.
[19] Chao Gao, Zongming Ma, Anderson Y Zhang, Harrison H Zhou, et al. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185, 2018.
[20] Evarist Giné, Vladimir Koltchinskii, et al. Empirical graph laplacian approximation of laplace–beltrami operators: Large sample results. In *High dimensional probability*, pages 238–259. Institute of Mathematical Statistics, 2006.
[21] C. Giraud and N. Verzelen. Partial recovery bounds for clustering with the relaxed

k-means. *Mathematical Statistics and Learning*, 1(3/4):317–374, 2019.

[22] Stephen Guattery and Gary L Miller. On the quality of spectral separators. *SIAM Journal on Matrix Analysis and Applications*, 19(3):701–719, 1998.

[23] Lars Hagen and Andrew B Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE transactions on computer-aided design of integrated circuits and systems*, 11(9):1074–1085, 1992.

[24] Kenneth M Hall. An r-dimensional quadratic placement algorithm. *Management science*, 17(3):219–229, 1970.

[25] Matthias Hein. Uniform convergence of adaptive graph-based regularization. In *International Conference on Computational Learning Theory*, pages 50–64. Springer, 2006.

[26] Matthias Hein, Jean-Yves Audibert, and Ulrike Von Luxburg. From graphs to manifolds–weak and strong pointwise consistency of graph laplacians. In *International Conference on Computational Learning Theory*, pages 470–485. Springer, 2005.

[27] Bruce Hendrickson and Robert Leland. An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM Journal on Scientific Computing*, 16(2):452–469, 1995.

[28] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[29] Jiashun Jin et al. Fast community detection by score. *The Annals of Statistics*, 43 (1):57–89, 2015.

[30] Iain M Johnstone and Debashis Paul. Pca in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292, 2018.

[31] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515, 2004.

[32] Ravindran Kannan, Santosh Vempala, et al. Spectral algorithms. *Foundations and Trends® in Theoretical Computer Science*, 4(3–4):157–288, 2009.

[33] Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

[34] Vladimir Koltchinskii and Dong Xia. Perturbation of linear forms of singular vectors under gaussian noise. In *High Dimensional Probability VII*, pages 397–423. Springer, 2016.

[35] Vladimir Koltchinskii, Karim Lounici, et al. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 1976–2013. Institut Henri Poincaré, 2016.

[36] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[37] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE, 2010.

[38] Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[39] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, 1982.

[40] Matthias Löffler, Anderson Y Zhang, and Harrison H Zhou. Supplement to "optimality of spectral clustering for gaussian mixture model". 2019.

[41] Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

[42] M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is np-hard. *International Workshop on Algorithms and Computation*, pages 274–285, 2009.
[43] Frank McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE, 2001.
[44] Marina Meila and Jianbo Shi. Learning segmentation by random walks. In *Advances in neural information processing systems*, pages 873–879, 2001.
[45] M. Ndaoud. Sharp optimal recovery in the two component gaussian mixture model. *arXiv preprint*, 2019.
[46] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.
[47] Sean O'Rourke, Van Vu, and Ke Wang. Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications*, 540:26–59, 2018.
[48] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760. ACM, 2010.
[49] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.
[50] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.
[51] J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM J. on Optimization*, 18(1):186–205, 2007.
[52] Alex Pothen, Horst D Simon, and Kang-Pu Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM journal on matrix analysis and applications*, 11(3): 430–452, 1990.
[53] Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.
[54] Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.
[55] M. Royer. Adaptive clustering through semidefinite programming. *Advances in Neural Information Processing Systems*, pages 1795–1803, 2017.
[56] Purnamrita Sarkar, Peter J Bickel, et al. Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics*, 43(3):962–990, 2015.
[57] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Departmental Papers (CIS)*, page 107, 2000.
[58] Horst D Simon. Partitioning of unstructured problems for parallel processing. *Computing systems in engineering*, 2(2-3):135–148, 1991.
[59] Daniel A Spielman and Shang-Hua Teng. Spectral partitioning works: Planar graphs and finite element meshes. In *Proceedings of 37th Conference on Foundations of Computer Science*, pages 96–105. IEEE, 1996.
[60] Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.
[61] D Michael Titterington, Adrian FM Smith, and Udi E Makov. *Statistical analysis of finite mixture distributions*. Wiley,, 1985.
[62] Rafael Van Driessche and Dirk Roose. An improved spectral bisection algorithm and its application to dynamic load balancing. *Parallel computing*, 21(1):29–48, 1995.
[63] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *J.*

Comput. Syst. Sci., 68(4):841–860, 2004.

[64] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17 (4):395–416, 2007.

[65] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

[66] Van Vu. A simple svd algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018.

[67] Stella Yu and Jianbo Shi. Multiclass spectral clustering. In *null*, page 313. IEEE, 2003.

[68] A.Y. Zhang and H.H. Zhou. Minimax rates of community detection in stochastic block models. *Ann. Statist.*, 44(5):2252–2280, 2016.

[69] Zhixin Zhou and Arash A Amini. Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *Journal of Machine Learning Research*, 20(47):1–47, 2019.

SUPPLEMENT TO "OPTIMALITY OF SPECTRAL CLUSTERING FOR GAUSSIAN MIXTURE MODEL"

BY Matthias Löffler, Anderson Y. Zhang and Harrison H. Zhou

ETH Zürich, University of Pennsylvania and Yale University

APPENDIX A: CHARACTERISTICS OF THE POPULATION QUANTITIES

In this section, we include several propositions that characterize the population quantities defined in Section 3.1. We first define two matrices related to z^*. Let $D \in \mathbb{R}^{k\times k}$ be a diagonal matrix with

$$D_{j,j} = \left|\{i \in [n] : z_i^* = j\}\right|, \forall j \in [k],$$

and $Z^* \in \{0,1\}^{n\times k}$ be a matrix such that

$$Z^*_{i,j} = \mathbb{I}\{z_i^* = j\}, \forall i \in [n], j \in [k]. \tag{34}$$

That is, Z^* can be viewed as a label matrix which serves a similar role as z^*. For any $j \in [k]$, $Z^*_{\cdot,j} \in \mathbb{R}^n$ indicates all the indexes belonging to the jth cluster.

PROPOSITION A.1. *There exists an orthonomal matrix $W \in \mathbb{R}^{k\times k}$ such that*

$$V = Z^* D^{-\frac{1}{2}} W.$$

Consequently, $V_{i,\cdot} = V_{j,\cdot}$ for all $i, j \in [n]$ such that $z_i^* = z_j^*$. In addition,

$$\sigma_1 \geq \sqrt{\frac{\beta n}{k}}\frac{\Delta}{2}.$$

PROOF. First note that

$$P = (\theta_1^*, \ldots, \theta_k^*)\, Z^{*T} = (\theta_1^*, \ldots, \theta_k^*)\, D^{\frac{1}{2}} D^{-\frac{1}{2}} Z^{*T} = (\theta_1^*, \ldots, \theta_k^*)\, D^{\frac{1}{2}} \left(Z^* D^{-\frac{1}{2}}\right)^T,$$

and observe that $Z^* D^{-\frac{1}{2}}$ has orthonormal columns. Now, we decompose $(\theta_1^*, \ldots, \theta_k^*)\, D^{\frac{1}{2}} = U\Lambda W^T$ into its SVD. Here W is some orthonomal matrix $W \in \mathbb{R}^{k\times k}$. Then we have that

$$P = U\Lambda \left(Z^* D^{-\frac{1}{2}} W\right)^T,$$

with $Z^*D^{-\frac{1}{2}}W$ having orthonormal columns. Then $\Sigma = \Lambda$ and $V = Z^*D^{-\frac{1}{2}}W$. The structure of Z^* immediately leads to the second statement presented in the proposition.

Due to Equation (1), the largest singular value of $(\theta_1^*, \ldots, \theta_k^*)$ must be greater than $\Delta/2$. Since $(\theta_1^*, \ldots, \theta_k^*)\,D^{\frac{1}{2}} = U\Sigma W^T$, we have

$$\sigma_1 \geq \sqrt{\frac{\beta n}{k}}\frac{\Delta}{2}.$$

□

PROPOSITION A.2. *All the coordinates of V satisfy*

$$\max_{i\in[n],j\in[k]} |V_{i,j}| \leq \sqrt{\frac{k}{\beta n}}.$$

PROOF. By Proposition A.1, due to the structure of D and Z^* we have that

$$V_{i,j} = \sum_{l=1}^{k} \frac{1}{D_{l,l}^{\frac{1}{2}}} Z_{i,l}^* W_{l,j} = \frac{W_{z_i^*,j}}{D_{z_i^*,z_i^*}^{\frac{1}{2}}},$$

Hence, we obtain that

$$\max_{i\in[n],j\in[k]} |V_{i,j}| \leq \frac{1}{\min_{i\in[n]} D_{z_i^*,z_i^*}^{\frac{1}{2}}} \leq \frac{1}{\sqrt{\beta n/k}}.$$

□

PROPOSITION A.3. *We have*

$$\left|\langle u_l, \theta_j^*\rangle\right| \leq \sigma_l \sqrt{\frac{k}{\beta n}}, \forall j, l \in [k].$$

PROOF. Since $P = U\Sigma V^T$ and $P_{\cdot,i} = \theta_{z_i^*}^*, \forall i \in [n]$, we have for any $u, l \in [k]$ that

$$\langle u_l, \theta_j^*\rangle = \sigma_l V_{i,l}, \text{ where } i \in [n] \text{ is any index such that } z_i^* = j.$$

The proof is then completed by applying Proposition A.2. □

APPENDIX B: AUXILIARY LEMMAS RELATED TO THE NOISE MATRIX E

In this section, we include some auxiliary lemmas related to singular values and perturbation of singular vectors used in the proof of Theorem 2.1.

Lemma B.1. *For a random matrix $E \in \mathbb{R}^{p\times n}$ with $\{E_{i,j}\} \overset{iid}{\sim} \mathcal{N}(0,1)$, define the event $\mathcal{F} = \{\|E\| \le \sqrt{2}(\sqrt{n}+\sqrt{p})\}$. We have that*

$$\mathbb{P}\left(\|E\| \ge \sqrt{n}+\sqrt{p}+t\right) \le e^{-t^2/2}$$

and particularly,

$$\mathbb{P}(\mathcal{F}) \ge 1 - e^{-0.08n}$$

Proof. By Theorem 2.13 in [10] we have that $\mathbb{E}\|E\| \le \sqrt{n}+\sqrt{p}$. Moreover, as $\|E\| = \sup_{\|u\|=\|v\|=1}\langle u, Ev\rangle$, we have by Borell's inequality that $\mathbb{P}(\|E\| \ge \mathbb{E}\|E\| + t) \le e^{-t^2/2}$. The result follows. □

Weyl's inequality (Theorem 4.3.1 of [28]) and the fact that $X = P + E$, Lemma B.1 implies the following lemma.

Lemma B.2. *Assume that the random event $\mathcal{F}$ holds. We have that*

$$\hat{\sigma}_j \le \sigma_j + \sqrt{2}(\sqrt{n}+\sqrt{p}), \quad \forall j \in [k].$$

The last lemma included in this section is the Davis-Kahan-Wedin $sin(\Theta)$ Theorem, which characterizes the distance between empirical and population singular vector spaces. We refer readers to Theorem 21 of [47] for its proof.

Lemma B.3 (Davis-Kahan-Wedin $sin(\Theta)$ Theorem). *Consider any rank-s matrices $W, \hat{W}$. Let $W = \sum_{i=1}^s \sigma_i u_i v_i^T$ be its SVD with $\sigma_1 \ge \ldots \ge \sigma_s$. Similarly, let $\hat{W} = \sum_{i=1}^s \hat{\sigma}_i \hat{u}_i \hat{v}_i^T$ be its SVD with $\hat{\sigma}_1 \ge \ldots \ge \hat{\sigma}_s$. For any $1 \le j \le l \le s$, define $V = (v_j, \ldots, v_l)$ and $\hat{V} = (\hat{v}_j, \ldots, \hat{v}_l)$. Then, we have that*

$$\inf_{O:\ orthonormal\ matrix} \left\|\hat{V} - VO\right\| \le \sqrt{2}\left\|\hat{V}\hat{V}^T - VV^T\right\| \le \frac{4\sqrt{2}\left\|\hat{W} - W\right\|}{\min\{\sigma_{j-1}-\sigma_j, \sigma_l - \sigma_{l+1}\}},$$

where we define $\sigma_0 = +\infty$ and $\sigma_{s+1} = 0$.

APPENDIX C: PROOFS OF KEY LEMMAS

In this section, we provide proofs of the key lemmas stated in Section 3, except the one of Lemma 3.3, which is deferred to Appendix D. Throughout this section, for any matrix W, we denote span(W) to be the space spanned by the columns of W.

Proof of Lemma 3.1. Since all the $\hat{P}_{\cdot,i} = \hat{U}\hat{Y}_{\cdot,i} = \left(\hat{U}\hat{U}^T\right)\hat{U}\hat{Y}_{\cdot,i}$ lie in the column space span$(\hat{U})$, any $\{\theta_j\}_{j=1}^k$ that achieves the minimum of Equation 10 must also lie in span$(\hat{U})$. That is

$$\begin{aligned}\min_{z\in[k]^n,\{\theta_j\}_{j=1}^k\in\mathbb{R}^k}\sum_{i\in[n]}\left\|\hat{P}_{\cdot,i}-\theta_{z_i}\right\|^2 &= \min_{z\in[k]^n,\{c_j\}_{j=1}^k\in\mathbb{R}^k}\sum_{i\in[n]}\left\|\hat{U}\hat{Y}_{\cdot,i}-\hat{U}c_{z_j}\right\|^2\\ &= \min_{z\in[k]^n,\{c_j\}_{j=1}^k\in\mathbb{R}^k}\sum_{i\in[n]}\left\|\hat{Y}_{\cdot,i}-c_{z_j}\right\|^2,\end{aligned}$$

where the last equation is due to the fact that all columns of $\hat{U}$ are orthonormal to each other. Thus they yield the same result after a proper label permutation. □

Proof of Lemma 3.2. Due to the fact $\hat{P}$ is the best rank-k approximation of X and P is also rank-k, we have that

$$\left\|\hat{P}-X\right\|_{\mathrm{F}}^2 \leq \|P-X\|_{\mathrm{F}}^2.$$

This, Hölder's inequality and the fact that we work on the event $\mathcal{F}$ imply that,

$$\begin{aligned}\left\|\hat{P}-P\right\|_{\mathrm{F}} &\leq 2\left\langle \frac{\hat{P}-P}{\left\|\hat{P}-P\right\|_{\mathrm{F}}}, X-P\right\rangle = 2\left\langle \frac{\hat{P}-P}{\left\|\hat{P}-P\right\|_{\mathrm{F}}}, E\right\rangle\\ &\leq \sup_{M:\|M\|_{\mathrm{F}}=1,\mathrm{rank}(M)\leq 2k} 2\langle M,E\rangle \leq 2\sqrt{2k}\|E\|\\ &\leq 4\sqrt{k}(\sqrt{n}+\sqrt{p})\end{aligned}$$

Now, denote by $\hat{\Theta}$ the centre matrix after solving Equation (10). That is, the ith column of $\hat{\Theta}$ is $\hat{\theta}_{\hat{z}_i'}$. Since $\hat{\Theta}$ is the solutions to the k-means objective, we have that

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

Hence, by the triangle inequality, we obtain that

$$\left\|\hat{\Theta} - P\right\|_{\mathrm{F}} \leq 2\left\|\hat{P} - P\right\|_{\mathrm{F}} \leq 8\sqrt{k}(\sqrt{n} + \sqrt{p}).$$

Now, define the set S as

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

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

$$|S| \leq \frac{\left\|\hat{\Theta} - P\right\|_{\mathrm{F}}^2}{(\Delta/2)^2} \leq \frac{256k\,(n+p)}{\Delta^2}.$$

Assuming that

$$\frac{\beta\Delta^2}{k^2\left(1 + \frac{p}{n}\right)} \geq 512,$$

we have that

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

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

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

We have the following arguments:

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

  $$\left\|\theta^*_j - \theta^*_l\right\| = \left\|\theta^*_{z^*_i} - \theta^*_{z^*_{i'}}\right\| \leq \left\|\theta^*_{z^*_i} - \hat{\theta}_{\hat{z}'_i}\right\| + \left\|\hat{\theta}_{\hat{z}'_i} - \hat{\theta}_{\hat{z}'_{i'}}\right\| + \left\|\hat{\theta}_{\hat{z}'_{i'}} - \theta^*_{z^*_{i'}}\right\| < \Delta,$$

  contradicting Equation (1).

Since $\hat{z}'_i$ can only take values in $[k]$, we conclude that $\{\hat{z}'_i : i \in C_j\}$ contains only one and different element for all $j \in [k]$. That is, there exists a permutation $\phi \in \Phi$, such that

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

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

$$\ell(\hat{z},z^*)\leq|S|\leq\frac{256k\,(n+p)}{\Delta^2}.$$

When the ratio $\Delta^2/\left(k^2\,(n+p)\right)$ is large enough, an immediate implication is that $\min_{j\in[k]}|\{i\in[n]:\hat{z}_i=j\}|\geq\frac{\beta n}{k}-|S|\geq\frac{\beta n}{2k}$. Moreover, in this case we obtain that

$$\max_j\left\|\hat{\theta}_j-\theta^*_{\phi(j)}\right\|^2\leq\frac{\left\|\hat{\Theta}-P\right\|_{\mathrm{F}}^2}{\frac{\beta n}{k}-|S|}\leq\frac{128k^2\,(n+p)}{\beta n}\leq\frac{\Delta^2}{4}.$$

□

Proof of Lemma 3.4. In this proof, we use the notation "$\overset{d}{=}$" to indicate two quantities having the same distribution. Recall that M has SVD $M=U\Sigma V^T$ where $U=(u_1,\ldots,u_k)$, $V=(v_1,\ldots,v_k)$ and $\Sigma=\mathrm{diag}\{\sigma_1,\ldots,\sigma_k\}\in\mathbb{R}^{k\times k}$ with $\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_k\geq 0$. We denote $\mathbb{S}=\left\{x\in\mathrm{span}\left(I-VV^T\right):\|x\|=1\right\}$ to be the unit sphere in $\mathrm{span}\left(I-VV^T\right)$. We also denote $\mathcal{O}$ to be the set of all orthonormal matrices in $\mathbb{R}^{n\times n}$ and furthermore

$$\mathcal{O}'=\{O\in\mathcal{O}:OV=V\}.$$

Let $V_\perp$ be an orthonormal extension of V such that $(V,V_\perp)\in\mathcal{O}$. Then for any $O\in\mathcal{O}'$, due to the fact that $O(V,V_\perp)\in\mathcal{O}$ and $O(V,V_\perp)=(V,OV_\perp)$, we have that $OV_\perp$ is another orthonormal extension of V. This implies that

$$Ox\in\mathrm{span}\left(I-VV^T\right),\ \forall x\in\mathrm{span}\left(I-VV^T\right).\tag{35}$$

Hence $\mathcal{O}'$ includes all rotation matrices in $\mathrm{span}\left(I-VV^T\right)$. In the following, we prove the three assertions of Lemma 3.4 one by one.

Assertion (1). Recall that $\hat{M}=M+E=U\Sigma V^T+E$ and $\hat{M}=\sum_{j=1}^{p\wedge n}\hat{\sigma}_j\hat{u}_j\hat{v}_j^T$. For any $O\in\mathcal{O}'$, since $EO^T\overset{d}{=}E$, we have that $\hat{M}O^T=\left(U\Sigma V^T+E\right)O^T=U\Sigma V^T+EO^T\overset{d}{=}\hat{M}$. On the other hand, $\hat{M}O^T$ has SVD

$$\hat{M}O^T=\sum_{j=1}^{p\wedge n}\hat{\sigma}_j\hat{u}_j\left(O\hat{v}_j\right)^T.$$

Hence, for any $j\in[k]$, we have that $\hat{v}_j\overset{d}{=}O\hat{v}_j$.

For any $x \in \mathbb{R}^n$, we define the mapping $f : \mathbb{R}^n \to \mathbb{S}$ as $f(x) = (I - VV^T)x/\|(I - VV^T)x\|$. Applying f on both $\hat{v}_j$ and $O\hat{v}_j$, we obtain that

$$\frac{(I - VV^T)O\hat{v}_j}{\|(I - VV^T)O\hat{v}_j\|} \stackrel{d}{=} \frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|}.$$

Since $\hat{v}_j = VV^T\hat{v}_j + (I - VV^T)\hat{v}_j$, we have $O\hat{v}_j = VV^T\hat{v}_j + O(I - VV^T)\hat{v}_j$. By Equation (35), we have that $O(I - VV^T)\hat{v}_j \in \text{span}\left(I - VV^T\right)$. Hence, we obtain that

$$VV^TO\hat{v}_j = VV^T\hat{v}_j \tag{36}$$

$$(I - VV^T)O\hat{v}_j = (I - VV^T)O(I - VV^T)\hat{v}_j = O(I - VV^T)\hat{v}_j. \tag{37}$$

As a consequence of Equation (37), we obtain

$$O\frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|} = \frac{O(I - VV^T)\hat{v}_j}{\|O(I - VV^T)\hat{v}_j\|} \stackrel{d}{=} \frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|}, \forall\, O \in \mathcal{O}'. \tag{38}$$

In particular, $(I - VV^T)\hat{v}_j/\|(I - VV^T)\hat{v}_j\|$ is contained in $\mathbb{S}$ and is rotation-invariant. Hence, we obtain that $(I - VV^T)\hat{v}_j/\|(I - VV^T)\hat{v}_j\|$ is uniformly distributed on $\mathbb{S}$.

Assertion (2). For any $x \in \mathbb{R}^n$, we define another mapping $g : \mathbb{R}^n \to \mathbb{R}^n$ as $g(x) = ((VV^Tx)^T, ((I - VV^T)x)^T/\|(I - VV^T)x\|)^T$. Recall that $\hat{v}_j \stackrel{d}{=} O\hat{v}_j$, $\forall O \in \mathcal{O}'$. Applying g on both $\hat{v}_j$ and $O\hat{v}_j$ and using Equations (36), (37) and (38), we obtain that

$$\begin{pmatrix} VV^T\hat{v}_j \\ \frac{(I-VV^T)\hat{v}_j}{\|(I-VV^T)\hat{v}_j\|} \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} VV^T\hat{v}_j \\ O\frac{(I-VV^T)\hat{v}_j}{\|(I-VV^T)\hat{v}_j\|} \end{pmatrix}. \tag{39}$$

Let $\mathcal{A}$ be a Borel subset of $\text{span}(VV^T)$ and $\mathcal{B}$ a Borel subset of $\mathbb{S}$. By (39) we have for any $O \in \mathcal{O}'$ that

$$\mathbb{P}\left(\frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|} \in \mathcal{B}\middle| VV^T\hat{v}_j \in \mathcal{A}\right) = \mathbb{P}\left(O\frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|} \in \mathcal{B}\middle| VV^T\hat{v}_j \in \mathcal{A}\right).$$

Hence, we obtain that $\frac{(I-VV^T)\hat{v}_j}{\|(I-VV^T)\hat{v}_j\|}\Big|VV^T\hat{v}_j$ is also uniformly distributed on $\mathbb{S}$, invariant to the value of $VV^T\hat{v}_j$. This implies that $\frac{(I-VV^T)\hat{v}_j}{\|(I-VV^T)\hat{v}_j\|}$ is independent of $VV^T\hat{v}_j$.

Assertion (3). Since $\|(I - VV^T)\hat{v}_j\| = \sqrt{1 - \|VV^T\hat{v}_j\|^2}$ is a function of only $VV^T\hat{v}_j$, it is an immediate consequence of the second assertion.

□

APPENDIX D: SPECTRAL PROJECTION MATRIX PERTURBATION THEORY

In this section, we give the proof of Lemma 3.3. Before that, we first introduce two lemmas used in the proof of Lemma 3.3.

The following lemma gives an upper bound on the operator norm of $\|S_{a:b}\|$. The setting considered here is slightly more general than that in Lemma 3.3, as E is not necessarily a Gaussian noise matrix. The proof of Lemma D.1 mainly follows that of Lemma 2 in [35]. It is included in the later part of this section for completeness.

LEMMA D.1. *Consider any rank-k matrix $M \in \mathbb{R}^{p\times n}$ with SVD $M = \sum_{j=1}^{k} \sigma_j u_j v_j^T$ where $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_k > 0$. Define $\sigma_0 = \sigma_{k+1} = 0$.*

Consider any matrix $E \in \mathbb{R}^{p\times n}$. Define $\hat{M} = M + E$. Let the SVD of $\hat{M}$ be $\sum_{j=1}^{p\wedge n} \hat\sigma_j \hat u_j \hat v_j^T$ where $\hat\sigma_1 \geq \hat\sigma_2 \geq \ldots \geq \hat\sigma_{p\wedge n}$.

For any two indexes a, b such that $1 \leq a \leq b \leq k$, define $V_{a:b} = (v_a, \ldots, v_b)$ and $\hat V_{a:b} = (\hat v_a, \ldots, \hat v_b)$. Let $V = (v_1, \ldots, v_k)$. Define the singular value gap $g_{a:b} = \min\{\sigma_{a-1} - \sigma_a, \sigma_b - \sigma_{b+1}\}$. Define

$$S_{a:b} = \left(I - VV^T\right)\left(\hat V_{a:b}\hat V_{a:b}^T - V_{a:b}V_{a:b}^T\right)V_{a:b} - \sum_{a\leq j\leq b}\frac{1}{\sigma_j}\left(I - VV^T\right)E^T u_j v_j^T V_{a:b}. \tag{40}$$

We have that

$$\|S_{a:b}\| \leq \left(\frac{32(\sigma_a - \sigma_b)}{\pi g_{a:b}} + 16\right)\frac{\|E\|^2}{g_{a:b}^2}.$$

The $S_{a:b}$ in Lemma 3.3 and Lemma D.1 depends on E. It can be written as $S_{a:b}(E)$ with $S_{a:b}(\cdot)$ treated as a function of the noise matrix. Lemma D.2 studies the Lipschitz continuity of $S_{a:b}(\cdot)$. It slightly generalizes Lemma 2.4 in [34] and follows along the same arguments. Its proof will be given in the later part of this section for completeness.

LEMMA D.2. *Consider the same setting as in Lemma D.1. Define $S_{a:b}(E)$ as in Equation (40). Consider another matrix $E' \in \mathbb{R}^{p\times n}$. Let $\hat M' = M + E'$. Define $S_{a:b}(E')$ analogously. Under the assumption that $\max\{\|E\|, \|E'\|\} \leq g_{a:b}/4$, we have that*

$$\|S_{a:b}(E) - S_{a:b}(E')\| \leq 1024\left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right)\frac{\max\{\|E\|, \|E'\|\}}{g_{a:b}^2}\|E - E'\|. \tag{41}$$

With Lemma D.1 and Lemma D.2, we are able to prove Lemma 3.3. It generalizes Theorem 1.1 in [34], and its proof follows the same argument.

Proof of Lemma 3.3. Define ϕ as follows

$$\phi(s) = \begin{cases} 1, & s \leq 1 \\ 3-2s, & s < 1 < 3/2 \\ 0, & s \geq 3/2 \end{cases}$$

and note that ϕ is Lipschitz with Lipschitz constant 2. As we mention earlier in this section, we can write $S_{a:b}(E)$ instead of $S_{a:b}$, and treat $S_{a:b}(\cdot)$ as a function.

Step 1. Define a function

$$h_\delta(E) = \langle S_{a:b}(E), W\rangle \phi\left(\frac{6\|E\|}{\delta}\right).$$

We are going to show that h_δ is also Lipschitz for any $\delta \leq g_{a:b}/4$. We use the notation $\|\cdot\|_*$ for the nuclear norm of a matrix.

- First suppose that $\max\{\|E\|, \|E'\|\} \leq \delta$. Then, by Lemma D.1, Lemma D.2 and the fact that ϕ is Lipschitz, we obtain that

$$\begin{aligned}
&|h_\delta(E) - h_\delta(E')| \\
&\leq \langle S_{a:b}(E) - S_{a:b}(E'), W\rangle \phi\left(\frac{6\|E\|}{\delta}\right) + \left\langle S_{a:b}(E'), W\right\rangle \left(\phi\left(\frac{6\|E\|}{\delta}\right) - \phi\left(\frac{6\|E'\|}{\delta}\right)\right) \\
&\leq \|S_{a:b}(E) - S_{a:b}(E')\| \|W\|_* \phi\left(\frac{6\|E\|}{\delta}\right) + \|S_{a:b}(E')\| \|W\|_* \left|\phi\left(\frac{6\|E\|}{\delta}\right) - \phi\left(\frac{6\|E'\|}{\delta}\right)\right| \\
&\leq 1024\left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right) \frac{\max\{\|E\|, \|E'\|\}}{g_{a:b}^2} \|E - E'\| \|W\|_* \\
&\quad + 16\left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right) \frac{\|E'\|^2}{g_{a:b}^2} \|W\|_* \frac{12\,|\|E\| - \|E'\||}{\delta} \\
&\leq C_1\left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right) \frac{\delta}{g_{a:b}^2} \|E - E'\| \|W\|_*,
\end{aligned}$$

  for some constant $C_1 > 0$ that is independent of E, E'.
- If $\min\{\|E\|, \|E'\|\} \geq \delta$ then $h(E) = h(E') = 0$ and the above inequality trivially holds.

- Finally, if $\|E\| < \delta \le \|E'\|$, by a similar argument as above, we obtain that

$$\begin{aligned} |h_\delta(E) - h_\delta(E')| = |h_\delta(E)| &= \left| \langle S_{a:b}(E), W\rangle \phi\left(\frac{6\|E\|}{\delta}\right) - \phi\left(\frac{6\|E'\|}{\delta}\right)\right| \\ &\le \|S_{a:b}(E)\|\|W\|_* \frac{\|E - E'\|}{\delta} \left|\phi\left(\frac{6\|E\|}{\delta}\right) - \phi\left(\frac{6\|E'\|}{\delta}\right)\right| \\ &\le C_1\left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right)\frac{\delta}{g_{a:b}^2}\|E - E'\|\|W\|_*, \end{aligned}$$

and the same bound holds if we switch the places of E and E' in the last case.

Combining the above cases together, we have shown that for any δ such that $\delta \le g_{a:b}/4$, h_δ is a Lipschitz function with Lipschitz constant bounded by

$$C_1\left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right)\frac{\delta}{g_{a:b}^2}\|W\|_*.$$

Step 2. In the following, for any two sequences $\{x_n\}, \{y_n\}$, we adopt the notation $x_n \lesssim y_n$ meaning there exists some constant $c > 0$ independent of n, such that $x_n \le c y_n$.

By lemma B.1, we have that for all $t > 0$,

$$\mathbb{P}\left(|\|E\| - \mathbb{E}\|E\|| \ge \sqrt{2t}\right) \le \exp(-t).$$

Set $\delta = \delta(t) = \mathbb{E}\|E\| + \sqrt{2t}$. We consider the following two scenarios depending on the values of t.

- We first consider the case where $\sqrt{2t} \le g_{a:b}/24$, which implies $\delta(t) \le g_{a:b}/6$. By the definition of $h_\delta(\cdot)$, we have that $h_\delta(E) = \langle S_{a:b}(E), W\rangle$ in this case. Denoting by m the median of $\langle S_{a:b}(E), W\rangle$ we have that

$$\begin{aligned} \mathbb{P}(h_\delta(E) \ge m) \ge \mathbb{P}(h_\delta(E) \ge m, \|E\| \le \delta(t)) &= \mathbb{P}(\langle S_{a:b}(E), W\rangle \ge m, \|E\| \le \delta(t)) \\ &\ge \mathbb{P}(\langle S_{a:b}, W\rangle \ge m) - \mathbb{P}(\|E\| > \delta(t)) \ge \frac{1}{2} - \frac{1}{2}e^{-t} \ge \frac{1}{4}, \end{aligned}$$

and likewise $\mathbb{P}(h_\delta(E) \le m) \ge 1/4$. Hence, since h_δ is Lipschitz, we can apply Lemma 2.6 in [34], which is a corollary to the the Gaussian isoperimetric inequality, to show that with probability at least $1 - e^{-t}$ that

$$|h_\delta(E) - m| \lesssim \sqrt{t}\left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right)\frac{\delta(t)}{g_{a:b}^2}\|W\|_*. \tag{42}$$

By Lemma B.1, we have that $\mathbb{E}\,\|E\| \lesssim \sqrt{n+p}$. Thus, we obtain that

$$|h_\delta(E) - m| \lesssim \left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right) \frac{\sqrt{t}}{g_{a:b}} \left(\frac{\sqrt{n+p} + \sqrt{t}}{g_{a:b}}\right) \|W\|_*. \tag{43}$$

Moreover, the event where $\|E\| \le \delta(t)$ occurs with probability at least $1 - e^{-t}$ and on this event h_δ coincides with $\langle S_{a:b}(E), W\rangle$. Hence, with probability at least $1 - 2e^{-t}$

$$|\langle S_{a:b}(E), W\rangle - m| \lesssim \left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right) \frac{\sqrt{t}}{g_{a:b}} \left(\frac{\sqrt{n+p} + \sqrt{t}}{g_{a:b}}\right) \|W\|_*. \tag{44}$$

- We need to prove a similar inequality in the case $\sqrt{2t} > g_{a:b}/24$. In this case we have that $\mathbb{E}\|E\| \lesssim \sqrt{t}$ as by assumption $\mathbb{E}\|E\| \le g_{a:b}/8$. Hence, applying lemma D.1, we have that

$$|\langle S_{a:b}(E), W\rangle| \le \|S_{a:b}(E)\|\,\|W\|_* \lesssim \left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right) \frac{t}{g_{a:b}^2} \|W\|_*. \tag{45}$$

  Hence, since $t \ge \log(4)$ and $e^{-t} \le 1/4$, we conclude that we can bound

$$|m| \lesssim \left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right) \frac{t}{g_{a:b}^2} \|W\|_*. \tag{46}$$

  Equation (45) and Equation (46) together immediately imply that the inequality in (44) also holds for $\sqrt{2t} > g_{a:b}/24$.

So far we have proved that Equation (44) holds for all $t > \log 4$. Integrating out the tails in the inequality in (44) we obtain that

$$|\mathbb{E}\langle S_{a:b}(E), W\rangle - m| \le \mathbb{E}|\langle S_{a:b}(E), W\rangle - m| \lesssim \left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right) \frac{\sqrt{n+p}}{g_{a:b}^2} \|W\|_*,$$

and hence we can substitute the median by the mean in the concentration inequality (44).

□

The last two things left are the proofs of Lemma D.1 and Lemma D.2.

Proof of Lemma D.1. As in the proof of Lemma 3.4, we use self-adjoint dilation. As before, we define for any matrix W

$$D\,(W) = \begin{pmatrix} 0 & W \\ W^T & 0 \end{pmatrix}.$$

As a result, we have that

$$D(M) = \begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix}, \quad D(E) = \begin{pmatrix} 0 & E \\ E^T & 0 \end{pmatrix}$$

and $D(\hat{M}) = D(M) + D(E)$. Note that all three matrices are symmetric and consequently have eigendecompositions. Particularly,

$$D(M) = \sum_{1 \le |i| \le k} \sigma_i P_i, \tag{47}$$

where for $i \in [k]$,

$$\sigma_{-i} = -\sigma_i,\ P_i = \frac{1}{2}\begin{pmatrix} u_i u_i^T & u_i v_i^T \\ v_i u_i^T & v_i v_i^T \end{pmatrix},\ P_{-i} = \frac{1}{2}\begin{pmatrix} u_i u_i^T & -u_i v_i^T \\ -v_i u_i^T & v_i v_i^T \end{pmatrix} \tag{48}$$

Similarly, we have that

$$D(\hat{M}) = \sum_{1 \le |i| \le p \wedge n} \hat{\sigma}_i \hat{P}_i., \tag{49}$$

where for each $i \in [k]$, $\hat{\sigma}_{-i}, \hat{P}_i$ and $\hat{P}_{-i}$ are defined analogously. Denote

$$P = \sum_{|i| \in \{a,\dots,b\}} P_i, \quad \text{and } \hat{P} = \sum_{|i| \in \{a,\dots,b\}} \hat{P}_i. \tag{50}$$

By doing so, we have

(51)

$$\left(I - VV^T\right)\left(\hat{V}_{a:b}\hat{V}_{a:b}^T - V_{a:b}V_{a:b}^T\right)V_{a:b} = \left(O_{n\times p} \quad (I - VV^T)\right)\left(\hat{P} - P\right)\begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix}.$$

The following proof can be divided into three steps.

Step 1. In this step, we decompose $\left(I - VV^T\right)\left(\hat{V}_{a:b}\hat{V}_{a:b}^T - V_{a:b}V_{a:b}^T\right)V_{a:b}$. Despite a little abuse of notation, denote $[\sigma_a, \sigma_b]$ to be the corresponding interval on the real axis of the complex plane $\mathbb{C}$. Define γ^+ to be the contour on $\mathbb{C}$ that circle around the intervals $[\sigma_a, \sigma_b]$ by a distance equal to $g_{a:b}/2$. That is,

$$\gamma^+ = \left\{\eta \in \mathbb{C} : \mathrm{dist}(\eta, [\sigma_a, \sigma_b]) = \frac{g_{a:b}}{2}\right\}, \tag{52}$$

where for any point $\eta \in \mathbb{C}$ and interval $B \in \mathbb{C}$, $\text{dist}(\eta, B) = \min_{\eta' \in \mathbb{B}} \|\eta - \eta'\|$. Likewise we define γ^- as

$$\gamma^- = \left\{ \eta \in \mathbb{C} : \text{dist}(\eta, [\sigma_{-b}, \sigma_{-a}]) = \frac{g_{a:b}}{2} \right\}. \tag{53}$$

In this way, among all the singular values of $D(M)$, only those with indexes in $\{a, \ldots, b\}$ and $\{-b, \ldots, -a\}$ are included in γ^+ and γ^- respectively, and the rest ones lie outside of the contours.

By the Riesz representation Theorem for spectral projectors (c.f. page 39 of [33]), we have that

$$\hat{P} = -\frac{1}{2\pi i} \oint_{\gamma^+} (D(\hat{M}) - \eta I)^{-1} d\eta - \frac{1}{2\pi i} \oint_{\gamma^-} (D(\hat{M}) - \eta I)^{-1} d\eta. \tag{54}$$

For any matrix W and any $\eta \in \mathbb{C}$, define the resolvent operator

$$R_W(\eta) = (D(W) - \eta I)^{-1}.$$

Then Equation (54) can be written as

$$\hat{P} = -\frac{1}{2\pi i} \oint_{\gamma^+} R_{\hat{M}}(\eta) d\eta - \frac{1}{2\pi i} \oint_{\gamma^-} R_{\hat{M}}(\eta) d\eta$$

Recall that $D(\hat{M}) = D(M) + D(E)$. Note that $R_M(\eta) = (D(M) - \eta I)^{-1}$. We are going to expand $R_{\hat{M}}(\eta)$ into its Neumann series:

$$\begin{aligned} R_{\hat{M}}(\eta) &= (D(M) - \eta I + D(E))^{-1} = ((D(M) - \eta I)(I + R_M(\eta) D(E)))^{-1} \\ &= (I + R_M(\eta) D(E))^{-1} R_M(\eta) = \sum_{j=0}^{\infty} (-1)^j [R_M(\eta) D(E)]^j R_M(\eta) \\ &= R_M(\eta) - R_M(\eta) D(E) R_M(\eta) + \sum_{j=2}^{\infty} (-1)^j [R_M(\eta) D(E)]^j R_M(\eta). \end{aligned} \tag{55}$$

Applying the Riesz representation Theorem on P, we have that

$$\begin{aligned} P &= -\frac{1}{2\pi i} \oint_{\gamma^+} (D(M) - \eta I)^{-1} d\eta - \frac{1}{2\pi i} \oint_{\gamma^-} (D(M) - \eta I)^{-1} d\eta \\ &= -\frac{1}{2\pi i} \oint_{\gamma^+} R_M(\eta) d\eta - \frac{1}{2\pi i} \oint_{\gamma^-} R_M(\eta) d\eta. \end{aligned}$$

As a result, we have the decomposition

$$\hat{P} - P = L(E) + S(E) \tag{56}$$

where $L(E)$ and $S(E)$ are operators on E, defined as

(57)
$$L(E)=\frac{1}{2\pi i}\oint_{\gamma^+}R_M(\eta)D(E)R_M(\eta)d\eta+\frac{1}{2\pi i}\oint_{\gamma^-}R_M(\eta)D(E)R_M(\eta)d\eta.$$

and

$$S(E)=-\frac{1}{2\pi i}\oint_{\gamma^+}\sum_{j=2}^{\infty}(-1)^j[R_M(\eta)D(E)]^jR_M(\eta)d\eta$$
$$-\frac{1}{2\pi i}\oint_{\gamma^-}\sum_{j=2}^{\infty}(-1)^j[R_M(\eta)D(E)]^jR_M(\eta)d\eta. \tag{58}$$

By Equation (51), we have that

$$\left(I-VV^T\right)\left(\hat{V}_{a:b}\hat{V}_{a:b}^T-V_{a:b}V_{a:b}^T\right)V_{a:b}=\begin{pmatrix}O_{n\times p} & (I-VV^T)\end{pmatrix}L(E)\begin{pmatrix}O_{n\times p}\\ V_{a:b}\end{pmatrix}$$
$$+\begin{pmatrix}O_{n\times p} & (I-VV^T)\end{pmatrix}S(E)\begin{pmatrix}O_{n\times p}\\ V_{a:b}\end{pmatrix}.$$

Step 2. In the following, we are going to show the first term on the right hand side of the above formula is exactly $\sum_j\frac{1}{\sigma_j}\left(I-VV^T\right)E^Tu_je_j^T$, which will imply that the second term is equivalent to $S_{a:b}$.

Define

$$L_{a:b}=\begin{pmatrix}O_{n\times p} & (I-VV^T)\end{pmatrix}L(E)\begin{pmatrix}O_{n\times p}\\ V_{a:b}\end{pmatrix}. \tag{59}$$

To simplify it, we first simplify $L(E)$. Recalling Equation (48), for any i such that $|i|\leq k$, we have that $P_i=\theta_i\theta_i^T$, where $\theta_i=\frac{1}{\sqrt{2}}(u_i^T,v_i^T)^T,\theta_{-i}=\frac{1}{\sqrt{2}}(u_i^T,-v_i^T)^T,\forall i\in[k]$. We can expand it so that $\{\theta_i,\theta_{-i}\}_{i\in[k]}\cup\{\theta_j\}_{k+1\leq j\leq p+n-k}$ gives an orthonormal basis for $\mathbb{R}^{p+n}$. There are two immediate implications.

- For $k+1\leq j\leq p+n-k$, we define $P_j=\theta_j\theta_j^T$. Then we have a decomposition for the identity matrix

$$I=\sum_{i\in\{1,\ldots,p+n-k\}\cup\{-k,\ldots,-1\}}P_i$$

  In the rest of the proof, by default we treat $\{1,\ldots,p+n-k\}\cup\{-k,\ldots,-1\}$ to be the whole set for the index i. We drop it when there is no ambiguity. For instance, the above equation can be simply written as $I=\sum_i P_i$.

- We also define

$$\sigma_j = 0, \forall k+1 \leq j \leq p+n-k. \tag{60}$$

Then Equation (47) is equivalent to be written as

$$D(M) = \sum_i \sigma_i P_i.$$

- For $k+1 \leq j \leq p+n-k$, θ_j is orthogonal to $\theta_i - \theta_{-i}, \forall i \in [k]$. This implies that the second part of θ_j (i.e., from the $(p+1)$th coordinate to the $(p+n)$th coordinate) is 0, or orthogonal to $\text{span}(v_1, \dots, v_k)$. Thus,

$$\begin{pmatrix} O_{n\times p} & (I - VV^T) \end{pmatrix} P_i = O, \forall i \text{ s.t. } |i| \leq k, \tag{61}$$

$$P_i \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix} = O, \forall i \text{ s.t. } |i| \notin \{a, \dots, b\}. \tag{62}$$

and

$$\begin{pmatrix} O_{n\times p} & (I - VV^T) \end{pmatrix} \sum_{i>k} P_i \begin{pmatrix} O_{p\times n} \\ O_{I_{n\times n}} \end{pmatrix} = I - VV^T. \tag{63}$$

By Equation (47), we have that

$$\begin{aligned} R_M(\eta) = (D(M) - \eta I)^{-1} = \left(\sum_i \sigma_i P_i - \eta I\right)^{-1} = \left(\sum_i (\sigma_i - \eta) P_i\right)^{-1} \\ = \sum_i \frac{1}{\sigma_i - \eta} P_i = \sum_{i\in\{a,\dots,b\}} \frac{1}{\sigma_i - \eta} P_i + \sum_{i\notin\{a,\dots,b\}} \frac{1}{\sigma_i - \eta} P_i, \end{aligned} \tag{64}$$

defined as $R_1^+(\eta)$ and $R_2^+(\eta)$ respectively. With this, for the first term of $L(E)$ in Equation (57), we have that

$$\frac{1}{2\pi i} \oint_{\gamma^+} R_M(\eta) D(E) R_M(\eta) d\eta = \frac{1}{2\pi i} \oint_{\gamma^+} \left(R_1^+(\eta) + R_2^+(\eta)\right) D(E) \left(R_1^+(\eta) + R_2^+(\eta)\right) d\eta.$$

Observe that by the Cauchy-Goursat Theorem,

$$\begin{aligned} &\oint_{\gamma^+} R_1^+(\eta) D(E) R_1^+(\eta) d\eta \\ =& \sum_{i\in\{a,\dots,b\}} P_i D(E) P_i \oint_{\gamma^+} \frac{1}{(\sigma_i - \eta)^2} d\eta + \sum_{i\neq j,\ i,j\in\{a,\dots,b\}} P_i D(E) P_j \oint_{\gamma^+} \frac{1}{(\sigma_i - \eta)(\sigma_j - \eta)} d\eta \\ =& 0, \end{aligned}$$

since there is no singularity inside γ^+. The identical result holds for $\oint_{\gamma^+} R_2^+(\eta)D(E)R_2^+(\eta)d\eta$. By the Cauchy integral formula

$$\frac{1}{2\pi i}\oint_{\gamma^+} R_1^+(\eta)D(E)R_2^+(\eta)d\eta = \sum_{i\in\{a,\ldots,b\}}\sum_{j\notin\{a,\ldots,b\}}\frac{1}{2\pi i}\oint_{\gamma^+}\frac{d\eta}{(\sigma_i-\eta)(\sigma_j-\eta)}P_iEP_j$$

$$= \sum_{i\in\{a,\ldots,b\}}\sum_{j\notin\{a,\ldots,b\}}\frac{P_iEP_j}{\sigma_i-\sigma_j}.$$

And similar result holds for $\frac{1}{2\pi i}\oint_{\gamma^+} R_2^+(\eta)D(E)R_1^+(\eta)d\eta$. Hence, we obain that

$$\frac{1}{2\pi i}\oint_{\gamma^+} R_M(\eta)D(E)R_M(\eta)d\eta = \sum_{i\in\{a,\ldots,b\}}\sum_{j\notin\{a,\ldots,b\}}\frac{P_iD(E)P_j+P_jD(E)P_i}{\sigma_i-\sigma_j}.$$

In the same manner, splitting

$$R_M(\eta) = R_1^-(\eta)+R_2^-(\eta) \triangleq \sum_{i\in\{-b,\ldots,-a\}}\frac{P_i}{\sigma_i-\eta}+\sum_{i\notin\{-b,\ldots,-a\}}\frac{P_i}{\sigma_i-\eta}, \tag{65}$$

we also obtain that

$$\frac{1}{2\pi i}\oint_{\gamma^-} R_M(\eta)D(E)R_M(\eta)d\eta = \sum_{i\in\{-b,\ldots,-a\}}\sum_{j\notin\{-b,\ldots,-a\}}\frac{P_iD(E)P_j+P_jD(E)P_i}{\sigma_i-\sigma_j}. \tag{66}$$

Hence, we have that

$$L(E) = \left(\sum_{i\in\{a,\ldots,b\}}\sum_{j\notin\{a,\ldots,b\}}+\sum_{i\in\{-b,\ldots,-a\}}\sum_{j\notin\{-b,\ldots,-a\}}\right)\frac{P_iD(E)P_j+P_jD(E)P_i}{\sigma_i-\sigma_j}. \tag{67}$$

Note that for any i such $|i|\in\{a,\ldots,b\}$ and any $|j|\notin\{a,\ldots,b\}$ Equations (61) and (62) imply

$$\begin{pmatrix}O_{n\times p} & (I-VV^T)\end{pmatrix}P_iD(E)P_j\begin{pmatrix}O_{n\times p}\\ V_{a:b}\end{pmatrix}=0.$$

Together with Equation (59), we have that

$$\begin{aligned}L_{a:b} &= \left(\sum_{i\in\{a,\ldots,b\}}\sum_{j\notin\{a,\ldots,b\}}+\sum_{i\in\{-b,\ldots,-a\}}\sum_{j\notin\{-b,\ldots,-a\}}\right)\begin{pmatrix}O_{n\times p} & (I-VV^T)\end{pmatrix}\frac{P_jD(E)P_i}{\sigma_i-\sigma_j}\begin{pmatrix}O_{n\times p}\\ V_{a:b}\end{pmatrix}\\ &= \left(\sum_{i\in\{a,\ldots,b\}}+\sum_{i\in\{-b,\ldots,-a\}}\right)\sum_{j>k}\begin{pmatrix}O_{n\times p} & (I-VV^T)\end{pmatrix}\frac{P_jD(E)P_i}{\sigma_i}\begin{pmatrix}O_{n\times p}\\ V_{a:b}\end{pmatrix},\end{aligned}$$

where in the last equation, we used Equation (60). Recall that for all $i \leq k$, $\sigma_{-i} = -\sigma_i$. This yields

$$
\begin{aligned}
L_{a:b} &= \sum_{i \in \{a,\ldots,b\}} \frac{1}{\sigma_i} \begin{pmatrix} O_{n\times p} & (I - VV^T) \end{pmatrix} \left(\sum_{j>k} P_j \right) D(E) \left(P_i - P_{-i}\right) \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix} \\
&= \sum_{i \in \{a,\ldots,b\}} \frac{1}{\sigma_i} \begin{pmatrix} O_{n\times p} & (I - VV^T) \end{pmatrix} \left(\sum_{j>k} P_j \right) \begin{pmatrix} O & E \\ E^T & O \end{pmatrix} \begin{pmatrix} O & u_i v_i^T \\ v_i^T u_i & O \end{pmatrix} \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix} \\
&= \sum_{i \in \{a,\ldots,b\}} \frac{1}{\sigma_i} \begin{pmatrix} O_{n\times p} & (I - VV^T) \end{pmatrix} \left(\sum_{j>k} P_j \right) \begin{pmatrix} O_{p\times n} \\ I_{n\times n} \end{pmatrix} E^T u_i v_i^T V_{a:b} \\
&= \sum_{i \in \{a,\ldots,b\}} \frac{1}{\sigma_i} \left(I - VV^T\right) E^T u_i v_i^T V_{a:b}.
\end{aligned}
$$

where the last equation is due to Equation (63). This implies

$$
\begin{aligned}
S_{a:b} &= \left(I - VV^T\right) \left(\hat{V}_{a:b} \hat{V}_{a:b}^T - V_{a:b} V_{a:b}^T \right) V_{a:b} - L_{a:b} \\
&= \begin{pmatrix} O_{n\times p} & (I - VV^T) \end{pmatrix} S(E) \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix}. \qquad (68)
\end{aligned}
$$

Step 3. In the final step, we are going to upper bound $\|S_{a:b}\|$ via the above formula. By Equation (64), for any $\eta \in \gamma^+$ or $\eta \in \gamma^-$, we have that

$$\|R_M(\eta)\| \leq \frac{2}{g_{a:b}}. \qquad (69)$$

Moreover, we have that

$$|\gamma^+| = |\gamma^-| \leq 2(\sigma_a - \sigma_b) + \pi g_{a:b}.$$

Recall the definition of $S(E)$ in Equation (58). Note that $\|D(E)\| = \|E\|$.

- Under the assumption that $\|E\| \leq g_{a:b}/4$, we have that

$$
\begin{aligned}
\|S_{a:b}\| \leq \|S(E)\| &\leq \frac{|\gamma^+| + |\gamma^-|}{2\pi} \sum_{j=2}^{\infty} \|R_M(\eta)\|^{j+1} \|D(E)\|^j \\
&\leq \frac{2(\sigma_a - \sigma_b) + \pi g_{a:b}}{\pi} \|E\|^2 \left(\frac{2}{g_{a:b}}\right)^3 \sum_{j=0}^{\infty} \|E\|^j \left(\frac{2}{g_{a:b}}\right)^j \\
&\leq \left(\frac{16(\sigma_a - \sigma_b)}{\pi g_{a:b}} + 8 \right) \frac{\|E\|^2}{g_{a:b}^2} \sum_{j=0}^{\infty} \|E\|^j \left(\frac{2}{g_{a:b}}\right)^j \\
&\leq \left(\frac{32(\sigma_a - \sigma_b)}{\pi g_{a:b}} + 16 \right) \frac{\|E\|^2}{g_{a:b}^2}. \qquad (70)
\end{aligned}
$$

- If $\|E\| > g_{a:b}/4$, by Equation (68) we have that

$$\|S_{a:b}\| \leq \left\|\hat{V}_{a:b}\hat{V}_{a:b}^T - V_{a:b}V_{a:b}^T\right\| + \|L_{a:b}\|.$$

The first term can be trivially upper bounded by 2. By the definition of $\|L_{a:b}\|$, the second term

$$\|L_{a:b}\| = \left\|\left(I - VV^T\right)E^T\left(\sum_{i\in\{a,\dots,b\}}\frac{1}{\sigma_i}u_iv_i^T\right)V_{a:b}\right\| \leq \frac{\|E\|}{\min_{i\in\{a,\dots,b\}}\sigma_i} \leq \frac{\|E\|}{g_{a:b}}.$$

Hence, we finally obtain that

$$\|S_{a:b}\| \leq 2 + \frac{\|E\|}{g_{a:b}} \leq 16\frac{\|E\|^2}{g_{a:b}^2}.$$

□

The very last thing is to prove Lemma D.2.

Proof of Lemma D.2. We follow the same decomposition and notation as in the proof of D.1. Recall the definition of $\hat{P}$ and P in Equation (50). In Equation (56), we have

$$\hat{P} - P = L(E) + S(E),$$

where $L(E)$ and $S(E)$ are defined in Equation (57) and Equation (58), respectively. Define $\hat{P}', L(E'), S(E')$ in the same manner for M'. Then we have

$$S(E') - S(E) = \hat{P}' - \hat{P} - \left(L(E') - L(E)\right).$$

As a consequence, by Equation (68)

$$\begin{aligned}
S_{a:b}(E') - S_{a:b}(E) &= \begin{pmatrix} O_{n\times p} & (I - VV^T)\end{pmatrix}\left(S(E') - S(E)\right)\begin{pmatrix} O_{n\times p} \\ V_{a:b}\end{pmatrix} \\
&= \begin{pmatrix} O_{n\times p} & (I - VV^T)\end{pmatrix}\left(\hat{P}' - \hat{P}\right)\begin{pmatrix} O_{n\times p} \\ V_{a:b}\end{pmatrix} \\
&\quad - \begin{pmatrix} O_{n\times p} & (I - VV^T)\end{pmatrix}\left(L(E') - L(E)\right)\begin{pmatrix} O_{n\times p} \\ V_{a:b}\end{pmatrix}.
\end{aligned}$$

In the proof of Lemma D.1, we study the perturbation between $\hat{P}$ and P. By the exactly the same argument, we can study the perturbation between $\hat{P}'$ and $\hat{P}$. Analogous to Equation (56), we have that

$$\hat{P}' - \hat{P} = \hat{L}(E' - E) + \hat{S}(E' - E),$$

where

$$\hat{L}(E'-E) = \frac{1}{2\pi i}\oint_{\gamma^+} R_{\hat{M}}(\eta)D(E'-E)R_{\hat{M}}(\eta)d\eta + \frac{1}{2\pi i}\oint_{\gamma^-} R_{\hat{M}}(\eta)D(E'-E)R_{\hat{M}}(\eta)d\eta. \tag{71}$$

and

$$\hat{S}(E'-E) = -\frac{1}{2\pi i}\oint_{\gamma^+}\sum_{j=2}^{\infty}(-1)^j[R_{\hat{M}}(\eta)D(E'-E)]^j R_{\hat{M}}(\eta)d\eta - \frac{1}{2\pi i}\oint_{\gamma^-}\sum_{j=2}^{\infty}(-1)^j[R_{\hat{M}}(\eta)D(E'-E)]^j R_{\hat{M}}(\eta)d\eta,$$

with γ^+,γ^- defied in Equation (52) and Equation (53) As a result,

$$S_{a:b}(E') - S_{a:b}(E) = \begin{pmatrix} O_{n\times p} & (I-VV^T)\end{pmatrix}\hat{S}(E'-E)\begin{pmatrix} O_{n\times p} \\ V_{a:b}\end{pmatrix} + \begin{pmatrix} O_{n\times p} & (I-VV^T)\end{pmatrix}\left(\hat{L}(E'-E) - \big(L(E')-L(E)\big)\right)\begin{pmatrix} O_{n\times p} \\ V_{a:b}\end{pmatrix},$$

which implies

$$\left\|S_{a:b}(E') - S_{a:b}(E)\right\| \le \left\|\hat{S}(E'-E)\right\| + \left\|\hat{L}(E'-E) - \big(L(E')-L(E)\big)\right\|. \tag{72}$$

We are going to establish upper bounds on the two terms individually.

Step 1. For the term related to $\hat{L}, L$, note that by Equation (57), Equation (71) and the fact $D(E'-E) = D(E') - D(E)$, we have

$$\begin{aligned}&\hat{L}(E'-E) - \big(L(E')-L(E)\big)\\ &= \frac{1}{2\pi i}\oint_{\gamma^+}\big(R_{\hat{M}}(\eta)D(E'-E)R_{\hat{M}}(\eta)d\eta - R_M(\eta)D(E'-E)R_M(\eta)\big)\,d\eta\\ &\quad + \frac{1}{2\pi i}\oint_{\gamma^-}\big(R_{\hat{M}}(\eta)D(E'-E)R_{\hat{M}}(\eta)d\eta - R_M(\eta)D(E'-E)R_M(\eta)\big)\,d\eta.\end{aligned}$$

By Weyl's inequality (Theorem 4.3.1 of [28]), we have $|\hat{\sigma}_i - \sigma_i| \le \|E\|, \forall i \in [p\wedge n]$. Under the assumption that $\|E\| \le g_{a:b}/4$, the minimum distance between γ^+,γ^- to the points $\{(\hat{\sigma}_i, 0)\}$ is at least $g_{a:b}/2 - \|E\| \ge g_{a:b}/4$, for all $i \in [p\wedge n]$. Similar to Equation (69), we obtain that

$$\left\|R_{\hat{M}}(\eta)\right\| \le \frac{4}{g_{a:b}}, \forall \eta \in \gamma^+,\gamma^-.$$

Then, together with the fact that $\|D(E'-E)\| = \|E'-E\|$, we have that

$$\begin{aligned}
&\left\| \oint_{\gamma^+} \left(R_{\hat{M}}(\eta) D(E'-E) R_{\hat{M}}(\eta) d\eta - R_M(\eta) D(E'-E) R_M(\eta) \right) d\eta \right\| \\
&\leq \left\| \oint_{\gamma^+} R_{\hat{M}}(\eta) D(E'-E)(R_{\hat{M}}(\eta) - R_M(\eta)) d\eta \right\| + \left\| \oint_{\gamma^+} (R_{\hat{M}}(\eta) - R_M(\eta)) D(E'-E) R_M(\eta) d\eta \right\| \\
&\leq \frac{8|\gamma^+|}{g_{a_b}} \|E'-E\| \sup_{\eta \in \gamma^+} \|R_{\hat{M}}(\eta) - R_M(\eta)\|.
\end{aligned}$$

Moreover, by the expansion of the resolvent into a Neumann series in (55), we have that

$$\begin{aligned}
\|R_{\hat{M}}(\eta) - R_M(\eta)\| &\leq \sum_{j=1}^{\infty} (\|R_M(\eta)\| \|E\|)^j \|R_M(\eta)\| \leq \|R_M(\eta)\|^2 \|E\| \sum_{j=0}^{\infty} (\|R_M(\eta)\| \|E\|)^j \\
&\leq \frac{8\|E\|}{g_{a:b}^2}, \forall \eta \in \gamma^+,
\end{aligned}$$

where the last inequality is due to Equation (69). Hence, as $|\gamma^+| \leq \pi g_{a:b} + 2(\sigma_a - \sigma_b)$, we have

$$\begin{aligned}
&\left\| \oint_{\gamma^+} \left(R_{\hat{M}}(\eta) D(E'-E) R_{\hat{M}}(\eta) d\eta - R_M(\eta) D(E'-E) R_M(\eta) \right) d\eta \right\| \\
&\leq 64 \left(\pi + \frac{2(\sigma_a - \sigma_b)}{g_{a:b}} \right) \frac{\|E\| \|E'-E\|}{g_{a:b}^2}.
\end{aligned}$$

The same result holds for the other integral over γ^-. Hence, we obtain that

$$\left\| \hat{L}(E'-E) - \left(L(E') - L(E) \right) \right\| \leq 64 \left(1 + \frac{2(\sigma_a - \sigma_b)}{\pi g_{a:b}} \right) \frac{\|E\| \|E'-E\|}{g_{a:b}^2}.$$

Step 2. For the term related to $\hat{S}$, we bound it analogously as in the proof of Lemma D.1. Following Equation (70), we have that

$$\left\| \hat{S}(E'-E) \right\| \leq 64 \left(\frac{32(\sigma_a - \sigma_b)}{\pi g_{a:b}} + 16 \right) \frac{\|E'-E\|^2}{g_{a:b}^2}.$$

Combining the above result together and by Equation (72), we have that

$$\left\| S_{a:b}(E') - S_{a:b}(E) \right\| \leq 1024 \left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}} \right) \frac{\max\{\|E\|, \|E'\|\}}{g_{a:b}^2} \|E - E'\|.$$

□

Seminar for Statistics
ETH Zürich
Rämistrasse 101
8092 Zürich Switzerland
E-mail: matthias.loeffler@stat.ethz.ch
URL: https://people.math.ethz.ch/~mloeffler/

Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104
E-mail: ayz@wharton.upenn.edu

Department of Statistics
Yale University
New Haven, CT 06511
E-mail: huibin.zhou@yale.edu
URL: http://www.stat.yale.edu/~hz68/