The Annals of Statistics
2021, Vol. 49, No. 5, 2506–2530
https://doi.org/10.1214/20-AOS2044

OPTIMALITY OF SPECTRAL CLUSTERING IN THE GAUSSIAN MIXTURE MODEL

BY MATTHIAS LÖFFLER[1], ANDERSON Y. ZHANG[2], AND HARRISON H. ZHOU[3]

[1]*Seminar for Statistics, ETH Zürich, matthias.loeffler@stat.math.ethz.ch*
[2]*Department of Statistics, University of Pennsylvania, ayz@wharton.upenn.edu*
[3]*Department of Statistics and Data Science, Yale University, huibin.zhou@yale.edu*

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. In this paper, we show that spectral clustering is minimax optimal in the Gaussian mixture model with isotropic covariance matrix, when the number of clusters is fixed and the signal-to-noise ratio is large enough. Spectral gap conditions are widely assumed in the literature to analyze spectral clustering. On the contrary, these conditions are not needed to establish optimality of spectral clustering in this paper.

1. Introduction. Clustering is a central and fundamental problem in statistics and machine learning. One popular approach to clustering of high-dimensional data is to use a spectral method [60, 65]. It tracks back to [17, 25] 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 computation [29, 59, 63], graph partitioning [8, 10, 13, 15, 24, 45, 54, 67] and explanatory data mining and statistical data analysis [3, 7, 34, 49]. It also has many real data applications, including image segmentation [46, 58, 69], text mining [11, 12, 51], speech separation [5, 19], and many others. In recent years, spectral clustering has also been one of the most favored and studied methods for community detection [4, 6, 18, 31, 40, 55, 57].

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. This way the dimensionality of the data is greatly reduced. Then a standard clustering method (e.g., the k-means algorithm) is performed on the low-dimensional denoised data to obtain an estimate of the cluster assignments. Due to the dimensionality reduction, spectral clustering is computationally less demanding than many other classical clustering algorithms.

In spite of its popularity, the theoretical properties of spectral clustering are not fully understood. One line of theoretical investigation of spectral clustering is to consider the performance under general conditions when applied to eigenvectors of the graph Laplacian. For instance, [7, 21, 27, 28, 66] provide various forms of asymptotic convergence guarantees for the graph Laplacian, related spectral properties and spectral clustering. Another approach is to consider the performance of spectral clustering in a specific statistical model. Particularly, spectral clustering for community detection in the stochastic blockmodel has been investigated frequently. Papers including [31, 40, 54, 55, 71] show that spectral clustering applied to the adjacency matrix of the network can consistently recover hidden community structure. However, their upper bounds on the number of nodes incorrectly clustered are polynomial in

Received November 2019; revised December 2020.
MSC2020 subject classifications. 62H30.
Key words and phrases. Spectral clustering, K-means, Gaussian mixture model, Spectral perturbation.

the signal-to-noise ratio, whereas the optimal rate of community detection is exponential in the signal-to-noise ratio [70]. Therefore, in the literature spectral clustering is often used as a way to initialize (i.e., "warm start") iterative algorithms which eventually achieve the optimal misclustering error rate.

In this paper, we investigate the theoretical performance of spectral clustering in the isotropic Gaussian mixture model. In this model, data points are generated from a mixture of Gaussian distributions with identity covariance each, 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 cluster assignment labels in the isotropic Gaussian mixture model is equivalent to the k-means algorithm. Finding an exact solution to the k-means objective has an exponential dependence on the dimension of the data points [30, 44], and hence is not feasible, even in moderate dimensions. As a result, various approximations have been used and studied. One direction is to relax the k-means objective by semidefinite programming (SDP) [16, 23, 53, 56]. These relaxations are more robust to outliers than spectral methods [61], but have a slower running time. Another possibility is to apply Lloyd's algorithm [41, 43], which is a greedy iterative method to approximately find a solution to the k-means objective. Given a sufficiently good initializer, typically provided by spectral clustering [37], Lloyd's algorithm achieves the optimal misclustering rate [43, 48]. However, we show that spectral clustering itself is already optimal when the error variance is isotropic and the dimensionality of the data does not grow faster than the number of samples.

A closely related result about spectral clustering for the Gaussian mixture model is [64]. Under a strong separation condition, spectral clustering is proved to achieve exact recovery of the underlying cluster structure 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 normalized Hamming loss function $\ell(\cdot, \cdot)$. We summarize our main result informally in Theorem 1.1.

THEOREM 1.1 (Informal statement of the main result). *For n data points generated from a Gaussian mixture model with an isotropic covariance matrix, we assume that*:

- *the number of clusters is finite*
- *the size of the clusters are of the same order*
- *the minimum distance among the centers, Δ, goes to infinity*
- *the dimension p of each data point is at most of the same order as n.*

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

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

This provides 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 exact recovery. We refer readers to Theorem 2.1 for a rigorous statement and a slightly stronger result, where we allow the number of clusters to grow with n, the cluster sizes to be not necessarily of the same order and the dimension p to grow slightly faster than n.

In particular, in Theorem 1.1, no spectral gap (i.e., singular value gap) condition is needed. This is contrary to the existing literature [2, 31, 40, 55], where various forms of eigenvalue gap or singular value gap conditions are required to apply matrix perturbation theory. This does not match the intuition that the difficulty of clustering should be determined by the

distances between the cluster centers, regardless of the spectral structure. In this paper, we completely drop any condition on the spectral gap. We achieve this by showing that the contribution of singular vectors from smaller singular values is negligible.

A recent related paper by Abbe et al. [2] studies community detection in 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 spectral operator perturbation theory of [35, 36] and introduce new techniques to establish optimality of spectral clustering and 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. We discuss extensions and potential caveats of our analysis in Section 3. The proof of the main theorem is given in Section 4, which is started with a proof sketch. We include the proofs of all the lemmas in the Supplementary Material [42].

Notation. For any matrix M, we denote by $\|M\|$ and $\|M\|_{\mathrm{F}}$ its operator norm and Frobenius norm, respectively. $M_{i,\cdot}$ denotes the ith row of M and $M_{\cdot,i}$ its ith column. For matrices M, N of the same dimension, their inner product is defined as $\langle M, N\rangle = \sum_{i,j} M_{ij}N_{ij}$. For any d, we denote by $\{e_a\}_{a=1}^d$ 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 $\{1, 2, \ldots, d\}$ and $\mathbb{I}\{\cdot\}$ to denote the indicator function. For $y_1, y_2, \ldots, y_d \in \mathbb{R}$, $\mathrm{diag}(y_1, y_2, \ldots, y_d)$ denotes the $d \times d$ diagonal matrix with diagonal entries $y_1, y_2, \ldots, y_d$.

2. Main results.

2.1. *Gaussian mixture model.* We consider an isotropic Gaussian mixture model with k centers $\theta_1^*, \ldots, \theta_k^* \in \mathbb{R}^p$ and a cluster assignment vector $z^* \in [k]^n$. In this model, independent observations $\{X_i\}_{i\in[n]}$ are generated as follows:

$$X_i = \theta^*_{z_i^*} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, I_p). \tag{1}$$

The goal of clustering is to recover the cluster assignment z^*. We measure the quality of a clustering algorithm by the average number of misclustered labels. Since the cluster structure is invariant to permutation of the label symbols, we define the misclustering error as

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

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

The difficulty of clustering is mainly determined by the distances between the 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:

$$\Delta = \min_{j,l\in[k]:j\neq l} \|\theta_j^* - \theta_l^*\|. \tag{2}$$

Another quantity that determines the possibility of consistent clustering is the size of the clusters. When the size of a cluster is small, recovery might be more difficult. We quantify the size of the smallest cluster by β, defined as

$$\beta = \frac{\min_{j\in[k]} |\{i \in [n] : z_i^* = j\}|}{n/k}. \tag{3}$$

Note that β cannot be greater than 1. We allow the case $\beta = o(1)$, such that cluster 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 for the matrix on which the spectral decomposition is applied, and which spectral components are used for the subsequent clustering. The clustering method used most commonly is the k-means algorithm.

In the context of community detection, spectral clustering [31, 40, 54, 55, 71] is usually performed on the eigenvectors of the adjacency matrix. For general clustering settings, [7, 21, 27, 28, 65, 66] first obtain a similarity matrix from the original data points by applying a kernel function. Then the graph Laplacian is constructed, whose eigenvectors are used for clustering. In [33, 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 the simplicity of this approach, it is powerful, as it achieves 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.

As common in the clustering literature, we assume that k, the number of clusters, is known. The purpose of the SVD is to reduce the dimensionality of the data while preserving underlying structure. After SVD, the dimensionality of the data vectors is reduced from p to k.[1] This makes the follow-up k-means algorithm computationally feasible compared to applying it directly onto the columns of X. Finding an exact solution for the k-means objective of the projected data (i.e., (4)) has computational complexity $O(n^{k^2+1})$ [30], which is polynomial in n if k is constant. In Section 2.5, we show how to modify Algorithm 1, using a $(1+\varepsilon)$-solution for the k-means algorithm to achieve linear (in n) complexity.

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 val-

Algorithm 1: Spectral clustering

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

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

1 Perform SVD on X to decompose

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

where $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_{p\wedge n} \geq 0$ 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 $\hat{\Sigma} := \operatorname{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_k)$, $\hat{V} := (\hat{v}_1, \ldots, \hat{v}_k)$, $\hat{U} := (\hat{u}_1, \ldots, \hat{u}_k)$ and

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

3 Perform k-means on the columns of $\hat{Y}$ and return an estimator $\hat{z}$ for the clustering assignment vector, that is,

$$(\hat{z}, \{\hat{c}_j\}_{j=1}^k) = \operatorname*{arg\,min}_{z\in[k]^n, \{c_j\}_{j=1}^k \in \mathbb{R}^k} \sum_{i\in[n]} \|\hat{Y}_{\cdot,i} - c_{z_i}\|^2. \tag{4}$$

[1] Here, we assume $p \geq k$. If $p < k$, then the dimensionality reduction is not needed and Algorithm 1 reduces to the k-means algorithm. To accommodate both $p \geq k$ and $p < k$, Step 2 of Algorithm 1 can be slightly changed by using the leading $\min\{k, p\}$ singular vectors instead.

ues should carry relatively less useful information, and consequently deserve less attention. Clustering on $\hat{Y}$ instead of $\hat{V}$ is also the main reason why we are able to remove the spectral gap condition. In particular, we will show in Lemma 4.1 that Algorithm 1 is equivalent to Algorithm 3, which performs clustering 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 [20, 37].

2.3. *Consistency.* We first present a preliminary result that proves consistency of the estimator $\hat{z}$ obtained in Algorithm 1.

PROPOSITION 2.1. *Assume that* $\Delta/(\beta^{-0.5}k(1+p/n)^{0.5}) \geq C$ *for some large enough constant* $C > 0$. *Then the output of Algorithm* 1, $\hat{z}$, *satisfies for another constant* $C' > 0$,

$$\ell(\hat{z}, z^*) \leq \frac{C'k(1+\frac{p}{n})}{\Delta^2} \tag{5}$$

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

Proposition 2.1 is an immediate consequence of Lemma 4.1 and Lemma 4.2, which are stated in Section 4. It is worth mentioning that there is no spectral gap condition assumed. In addition, Proposition 2.1 can be extended to mixture models where the errors $\{\epsilon_i\}$ are not necessarily $\mathcal{N}(0, I_p)$ distributed. We include this extension in Appendix D as Proposition D.1.

2.4. *Optimality.* In the next theorem we establish that Algorithm 1 achieves in fact an exponential convergence rate in the Gaussian mixture model when the covariance matrix of the Gaussian noise variables is isotropic.

THEOREM 2.1. *Suppose that*

$$\frac{\Delta}{k^{10.5}\beta^{-0.5}(1+\frac{p}{n})(\frac{n-k}{n})^{-0.5}} \to \infty. \tag{6}$$

Then the output of Algorithm 1, $\hat{z}$, *satisfies*

$$\ell(\hat{z}, z^*) \leq \exp\left(-\left(1 - \left(\frac{\Delta}{k^{10.5}\beta^{-0.5}(1+\frac{p}{n})(\frac{n-k}{n})^{-0.5}}\right)^{-0.1}\right)\frac{\Delta^2}{8}\right) \tag{7}$$

with probability at least $1 - \exp(-\Delta) - 3nk\exp(-0.08(n-k))$.

In Theorem 2.1, we allow the number of clusters k to grow with n, the cluster sizes not to be of the same order (quantified by β), and the dimension p to be of larger order than n. This is slightly stronger than the informal statement we make in Theorem 1.1. In addition, the proof of Theorem 2.1 yields a version of (7) that holds in expectation:

$$\mathbb{E}\ell(\hat{z}, z^*) \leq \exp\left(-(1-o(1))\frac{\Delta^2}{8}\right) + \exp(-(1-o(1))0.08n), \tag{8}$$

where the first term dominates as long as $\Delta^2 = o(n)$.

The following minimax lower bound for recovering z^* in the Gaussian mixture model is established in [43]:

$$\inf_{\hat{z}} \sup_{(\theta_1^*,\ldots,\theta_k^*),z^*} \mathbb{E}\ell(\hat{z}, z^*) \geq \exp\left(-(1-o(1))\frac{\Delta^2}{8}\right) \quad \text{if } \frac{\Delta}{\log(k\beta^{-1})} \to \infty. \tag{9}$$

Here, the infimum is taken over all feasible estimators $\hat{z}$, and the supremum is taken over all possible parameters, where the true centers $(\theta_1^*, \dots, \theta_k^*) \in \mathbb{R}^{p\times k}$ are separated by at least Δ, and the true cluster assignment z^* has a minimum cluster size of $\beta n/k$.

When $\Delta \to \infty$, $p = o(n\Delta)$ and k and β are constants, the convergence rate in (7) matches the minimax lower bound (9) up to a $(1 + o(1))$ factor in the exponent. Moreover, when additionally $\liminf_{n\to\infty} \Delta^2/(8\log n) > 1$, $\hat{z}$ equals z^* with high probability and we achieve exact recovery. This sharply matches the exact recovery threshold [9, 43].

Whereas $\Delta \to \infty$ is a necessary condition for consistent recovery [43, 48], the condition in (6) is not optimal. The assumption that $p = o(n\Delta)$ is an artifact of our proof technique. It can be improved to $p = o(n\Delta^2)$ under additional assumptions on the singular values of the population matrix $\mathbb{E}X$. When $n\Delta^2 = o(p)$, Algorithm 1 may only achieve suboptimal convergence rates and we discuss the intuition behind this in Section 3.3. The dependence on k is suboptimal as well, mainly due to higher order perturbation terms in our proof. In contrast, [16, 23] only need to assume $kp = o(n\Delta^2)$ for their SDP relaxation of k-means to achieve exponential rates (but with suboptimal constant in the exponent).

We emphasize that, as in Proposition 2.1, there is no spectral gap (i.e., singular value gap) condition assumed in Theorem 2.1. It is possible that the population matrix $\mathbb{E}X$ has a rank that is smaller than k, such that the smallest singular values of the population matrix $\mathbb{E}X$ are 0 or near 0. For instance, this occurs when some of the centers are (nearly) collinear. This is contrary to the existing literature [2, 31, 40, 55], where the spectral gap is assumed to be sufficiently large to apply spectral perturbation theory. The spectral gap condition is not natural, as the minimax rate in (9) only depends on Δ and is invariant to any spectral structure. In Theorem 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.

2.5. $(1+\varepsilon)$-*Solutions to k-means*. Computing the k-means objective in Algorithm 1 has complexity $O(n^{k^2+1})$ [30] and quickly becomes impractical, even for moderate values of k. A potential alternative is to use an $(1+\varepsilon)$-solution. An $(1+\varepsilon)$-solution is a pair $(\tilde{z}, \{\tilde{c}_j\}_{j=1}^k)$, such that its k-means objective value is within a factor of $(1+\varepsilon)$ of the global minimum of the k-means objective. For instance, [38] proposed an $(1+\varepsilon)$-approximation algorithm with complexity $O(2^{(k/\varepsilon)^{O(1)}}n)$, which is linear in n when k is constant and polynomial in n as long as k grows sublogarithmically in n. Proposition 2.1 is still valid when an $(1+\varepsilon)$-solution is used. However, $(1+\varepsilon)$-solutions do not necessarily enjoy a *local* optimality guarantee for the estimated labels, that is, $\|\hat{Y}_i - \tilde{c}_{\tilde{z}_i}\| \le \|\hat{Y}_i - \tilde{c}_j\|, \forall i \in [n], j \ne \tilde{z}_i$, which is required in the proof of Theorem 2.1. To overcome this problem, we propose to run an extra one step Lloyd's algorithm [41] as described in Algorithm 2. Consequently, the statement of Theorem 2.1 still holds for Algorithm 2, which we present below in Theorem 2.2.

THEOREM 2.2. *Assume that*

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

holds. Then the output of Algorithm 2, $\tilde{z}$, satisfies

$$\ell(\tilde{z}, z^*) \le \exp\left(-\left(1 - \left(\frac{\Delta}{k^{10.5}\beta^{-0.5}(1+\frac{p}{n})(\frac{n-k}{n})^{-0.5}(1+\varepsilon)^{0.5}}\right)^{-0.1}\right)\frac{\Delta^2}{8}\right) \tag{10}$$

with probability at least $1 - \exp(-\Delta) - 3nk\exp(-0.08(n-k))$.

The proof of Theorem 2.2 is almost identical to that of Theorem 2.1 and we sketch the necessary modifications in Appendix E.

Algorithm 2: Spectral clustering with $(1+\varepsilon)$-solution

Input: Data matrix $X \in \mathbb{R}^{p\times n}$, number of clusters k, approximation level ε
Output: Clustering assignment vector $\tilde{z} \in [k]^n$

1 Implement Steps 1-2 of Algorithm 1 to obtain $\hat{Y} \in \mathbb{R}^{k\times n}$. Compute a $(1+\varepsilon)$-solution (e.g., [38]) for the k-means algorithm on the columns of $\hat{Y}$ and return $(\check{z}, \{\check{c}_j\}_{j=1}^k)$, the cluster assignment vector and centers, such that

$$\sum_{i\in[n]} \|\hat{Y}_{\cdot,i} - \check{c}_{\check{z}_i}\|^2 \le (1+\varepsilon) \inf_{\{c_j\}_{j=1}^k \in \mathbb{R}^k} \sum_{i\in[n]} \min_{j\in[k]} \|\hat{Y}_{\cdot,i} - c_j\|^2$$

3 Update the centers

$$\tilde{c}_j = \frac{\sum_{i\in[n]} \hat{Y}_{\cdot,i} \mathbb{I}\{\check{z}_i = j\}}{\sum_{i\in[n]} \mathbb{I}\{\check{z}_i = j\}}, \quad j = 1,\dots,k.$$

4 Update the labels

$$\tilde{z}_i = \arg\min_{j\in[k]} \|\hat{Y}_{\cdot,i} - \tilde{c}_j\|, \quad i = 1,\dots,n.$$

3. Discussion.

3.1. *Unknown covariance matrix and sub-Gaussian errors.* The consistency guarantee established in Proposition 2.1 can be extended to more general settings where the noise variables $\{\epsilon_i\}_{i=1}^n$ have covariance matrix Σ or are sub-Gaussian. We include this extension in Appendix D as Proposition D.1.

In contrast, it is not possible to extend Theorem 2.1 and Theorem 2.2 to either sub-Gaussian distributed errors or unknown covariance matrices with our current proof techniques. This is due to the fact that the proof is highly reliant on both the isoperimetric inequality (cf. (52)) and rotation invariance of the singular vectors of the noise matrix $(\epsilon_1,\dots,\epsilon_n)$ (as in Lemma 4.4). An isoperimetric inequality would also be fulfilled by strongly log-concave distributed errors [50]. On the other hand, the rotation invariance of the singular vectors of $(\epsilon_1,\dots,\epsilon_n)$ is equivalent to ϵ_i being spherically Gaussian distributed.

3.2. *Unknown k.* Algorithm 1 and Theorem 2.1 require that the number of clusters, k, is known. In practice, k might be unknown and might need to be estimated. For this purpose, several approaches have been developed, including cross-validation [68], the gap-statistic [62], eigenvalue based heuristics [65] and resampling strategies [47]. However, while these methods often work well empirically, their theoretical performances are not fully understood, especially in high-dimensional regimes with growing p and n. One may estimate k by the aforementioned methods and use the resulting estimate in Algorithm 1, but further investigation is beyond the scope of this paper.

3.3. *Parameter regime $n\Delta^2 = O(p)$.* Proposition 2.1 and Theorems 2.1 and 2.2 are limited to the parameter regimes $p = o(n\Delta^2)$ and $p = o(n\Delta)$, respectively, beyond which the performance of Algorithm 1 remains unclear. When $n\Delta^2 = o(p)$, Theorem 2.2 in [14] indicates that, in general, the leading empirical singular values of X are all equal to $(1+o(1))(\sqrt{n}+\sqrt{p})$. As a result, running k-means on $\hat{Y}$ in Algorithm 1 is the same as on $\hat{V}$, and its performance may depend on the structure of the population singular values [26]. On

the other hand, [1, 48] consider a simplified model where $X_i = z_i^* \theta^* + \epsilon_i$ with $z^* \in \{-1, 1\}^n$ and study a different variant of spectral clustering which utilizes the leading eigenvector of the hollowed Gram matrix ($X^T X$ with its diagonal entries replaced by zero). This approach leads to a bias reduction and yields optimal misclustering rates in the high-dimensional setting when $n\Delta^2 = o(p)$. Likewise, a debiasing step is also employed for SDP relaxations of k-means [23, 56] to show exponential misclustering rates in this regime. It would be highly interesting to investigate whether similar debiasing ideas can be used to modify Algorithm 1 and prove optimal convergence rates in the general k-cluster case when $n\Delta^2 = o(p)$.

3.4. *Adaptive dimension reduction.* The population matrix $(\theta^*_{z_1^*}, \ldots, \theta^*_{z_n^*})$ might have smaller rank than k. For instance, when the centers are collinear, the rank of the population matrix equals one. Hence, in such cases it is conceivable to use a smaller number of singular vectors in Algorithm 1, as this further reduces the computational burden of computing the k-means objective. One way to achieve this, while still retaining the theoretical guarantees of Theorem 2.1, is to use the leading $\hat{r}$ singular vectors for the projection Step 2 in Algorithm 1, where $\hat{r}$ is an empirical version of r defined in (15). This preselection step keeps all the informative singular vectors without involving the noisy part of the projected data corresponding to small population singular values and allows to shorten the proof of Theorem 2.1. On the other hand, estimating r requires the noise level to be known or to be estimated, which adds additional computational complexity and introduces an additional tuning parameter.

4. Proof of main results. In Section 4.1, we first introduce the population counterparts of the quantities appearing in Algorithm 1. After that, several key lemmas for the proof are presented in Section 4.2. Since the proof of Theorem 2.1 is long and involved, we provide a proof sketch in Section 4.3, followed by its complete and detailed proof in Section 4.4. Auxiliary lemmas are included in the Supplementary Material [42].

4.1. *Population quantities.* We define $P = \mathbb{E}X$ and $E = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^{p\times n}$, such that we have the matrix representation $X = P + E$. We define several quantities related to P, the population version of X. We denote the SVD of P (note that P is at most rank of $k \wedge p$)

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

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$, $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$, $U = (u_1, \ldots, u_k) \in \mathbb{R}^{p\times k}$, $V = (v_1, \ldots, v_k) \in \mathbb{R}^{n\times k}$. Moreover, we define

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

In Appendix A, we provide several propositions (Propositions A.1, A.2 and A.3) to characterize the structure of these population quantities.

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

In Lemma 4.1, we show that Algorithm 1 has the same output as Algorithm 3, where clustering is performed on the columns of $\hat{U}\hat{Y}$ instead of $\hat{Y}$. We defer its proof to the Supplementary Material [42].

LEMMA 4.1. *Denote by* $(\hat{z}, \{\hat{c}_j\}_{j=1}^k)$ *and* $(\hat{z}', \{\hat{\theta}_j\}_{j=1}^k)$ *the outputs of Algorithm* 1 *and Algorithm* 3, *respectively. Then, after a label permutation,* $\hat{z}$ *equals* $\hat{z}'$, *that is, there exists a* $\phi \in \Phi$ *such that*

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

Algorithm 3: Clustering with rank-k approximation

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

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

1 Implement Steps 1–2 of Algorithm 1 to obtain $\hat{\Sigma} \in \mathbb{R}^{k\times k}$, $\hat{V} \in \mathbb{R}^{n\times k}$ and $\hat{U} \in \mathbb{R}^{p\times k}$. In addition, define

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

2 Perform k-means on the columns of $\hat{P}$ and return the estimated clustering assignment vector $\hat{z}'$ and estimated centers $\{\hat{\theta}_j\}_{j=1}^k$, that is,

$$(\hat{z}', \{\hat{\theta}_j\}_{j=1}^k) = \operatorname*{arg\,min}_{z\in[k]^n, \{\theta_j\}_{j=1}^k\in\mathbb{R}^k} \sum_{i\in[n]} \|\hat{P}_{\cdot,i} - \theta_{z_i}\|^2. \tag{11}$$

In addition, we have that

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

In Lemma 4.2, we show consistency of Algorithm 3 on the following event:

$$\mathcal{F} = \{\|E\| \le \sqrt{2}(\sqrt{n} + \sqrt{p})\}. \tag{12}$$

which occurs with high probability (as proven in Lemma B.1).

LEMMA 4.2. *Assume that the event $\mathcal{F}$ holds and that $\Delta/(\beta^{-0.5}k(1 + p/n)^{0.5}) \ge C$ for some constant $C > 0$. Then there exists another constant C' such that the output of Algorithm 3 $(\hat{z}', \{\hat{\theta}_j\}_{j=1}^k)$ satisfies*

$$\ell(\hat{z}', z^*) \le \frac{C'k(1 + \frac{p}{n})}{\Delta^2} \quad \textit{and} \tag{13}$$

$$\min_{\phi\in\Phi} \max_{j\in[k]} \|\hat{\theta}_j - \theta^*_{\phi(j)}\| \le C'\beta^{-\frac{1}{2}}k\sqrt{1 + \frac{p}{n}}. \tag{14}$$

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

The proof of Lemma 4.2 is included in the Supplementary Material [42]. The results of Lemma B.1, Lemma 4.2 and Lemma 4.1 immediately imply Proposition 2.1.

Lemma 4.3 studies the difference between the empirical spectral projection matrix and its sample counterpart. It decomposes $\hat{V}_{a:b}\hat{V}_{a:b}^T - V_{a:b}V_{a:b}^T$ 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 simple form, and is the main component that leads to the exponent $\Delta^2/8$ in (7). The remaining nonlinear part, though without an explicit expression, is well behaved and concentrates strongly around 0. Lemma 4.3 is a slight generalization of results due to [35, 36], where $\sigma_a, \ldots, \sigma_b$ are assumed to be the same. Here, we relax this assumption, by allowing the corresponding singular values to vary. The proof of Lemma 4.3 is involved but mainly follows the line of arguments in [35, 36]. We include the proof in the Supplementary Material [42] for completeness.

LEMMA 4.3. *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* $\sigma_{k+1} = 0$.

Suppose that E *is a matrix with i.i.d. Gaussian entries,* $E_{i,j}$. *Define* $\hat{M} = M + E$ *and suppose that* $\hat{M}$ *has SVD* $\sum_{j=1}^{p\wedge n} \hat{\sigma}_j \hat{u}_j \hat{v}_j^T$ *where* $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_{p\wedge n}$. *For any two indices* a, b *such that* $1 \leq a \leq b \leq k$, *define* $V_{a:b} = (v_a, \ldots, v_b)$, $\hat{V}_{a:b} = (\hat{v}_a, \ldots, \hat{v}_b)$ *and* $V = (v_1, \ldots, v_k)$. *Moreover, define the singular value gap* $g_{a:b} = \min\{\sigma_{a-1} - \sigma_a, \sigma_b - \sigma_{b+1}\}$ *and denote*

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

Suppose that $\mathbb{E}\|E\| \leq \frac{g_{a:b}}{8}$. *Then there exists some constant* $C > 0$ *such that with probability at least* $1 - 2e^{-t}$

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

for any $W \in \mathbb{R}^{n\times(b-a)}$, *any* $t \geq \log 4$ *and where* $\|\cdot\|_*$ *denotes the nuclear* (*Schatten*-1) *norm*.

The next lemma, Lemma 4.4, characterizes the distribution of empirical singular vectors. Similar to Lemma 4.3, Lemma 4.4 holds for matrices with any underlying structure, not necessarily in the clustering setting, as long as the noise is Gaussian distributed. The most important implication of Lemma 4.4 is that, for any empirical singular vector $\hat{v}_j$, its component that is orthogonal to the true signal V (i.e., $(I - VV^T)\hat{v}_j$) is after normalization Haar distributed on the sphere spanned by $(I - VV^T)$. This observation appears and is utilized in [32, 52]. Lemma 4.4 is essentially the same as Theorem 6 of [52]. For completeness, we give the proof in the Supplementary Material [42].

LEMMA 4.4. *Consider a 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* $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_k > 0$. *Suppose that* E *is a matrix with i.i.d. Gaussian entries,* $E_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. *Define* $\hat{M} = M + E$ *and suppose that* $\hat{M}$ *has SVD* $\sum_{j=1}^{p\wedge n} \hat{\sigma}_j \hat{u}_j \hat{v}_j^T$ *where* $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_{p\wedge n}$. *Define* $V = (v_1, \ldots, v_k)$. *Then for any* $j \in [k]$, *the following holds*:

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

$$\frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|} \overset{d}{=} \frac{(I - VV^T)w}{\|(I - VV^T)w\|} \quad \text{where } w \sim \mathcal{N}(0, I_n)$$

and where $\overset{d}{=}$ *denotes equality in distribution. In particular, we have that*

$$\mathbb{E}\frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|} = 0.$$

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

4.3. *Proof sketch for Theorem* 2.1. In this section, we provide a sketch for the proof of Theorem 2.1. The complete and detailed proof is given in Section 4.4. Throughout the proof, we assume that the random event $\mathcal{F}$ (defined in (12)) holds.

We use the equivalence between Algorithm 1 and Algorithm 3 (by Lemma 4.1), where clustering is performed on the columns of $\hat{P} = \hat{U}\hat{Y}$. Hence, it is sufficient to study the behavior of $(\hat{z}, \{\hat{\theta}_j\}_{j\in[n]})$. Particularly, (11) implies a *local* optimality result of the estimated labels, that is,

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

Then after label permutation, which without loss of generality we assume to be $\phi = \mathrm{Id}$, $n\ell(\hat{z}, z^*)$ can be bounded by

$$\begin{aligned} n\ell(\hat{z}, z^*) &= \sum_{i=1}^{n} \mathbb{I}\{\operatorname*{arg\,min}_{a\in[k]} \|\hat{P}_{\cdot,i} - \hat{\theta}_a\|^2 \neq z_i^*\} \\ &\le \sum_{i=1}^{n} \sum_{a\neq z_i^*} \mathbb{I}\{\|\hat{P}_{\cdot,i} - \hat{\theta}_a\|^2 \le \|\hat{P}_{\cdot,i} - \hat{\theta}_{z_i^*}\|^2\}. \end{aligned}$$

We divide the remaining proof into four steps, corresponding to Sections 4.4.1 to 4.4.4 in the complete proof.

Step 1 (Sketch of Section 4.4.1). We decompose $\ell(\hat{z}, z^*)$ into two parts: the first part corresponds 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. We define $r \in [k]$ as follows (with $\sigma_{k+1} := 0$):

$$r := \max\{j \in [k] : \sigma_j - \sigma_{j+1} \ge \rho(\sqrt{n} + \sqrt{p})\}, \tag{15}$$

where $\rho \to \infty$ is some quantity whose value will be given in the complete proof. There are two benefits in choosing r this way: singular values with index larger than r are relatively small and the singular value gap $\sigma_r - \sigma_{r+1}$ is large enough to apply matrix spectral perturbation theory. We split $\hat{U}$ into $(\hat{U}_{1:r}, \hat{U}_{(r+1):k})$, and hence we obtain that $\hat{P}_{\cdot,i} = \hat{P}^{(1)}_{\cdot,i} + \hat{P}^{(2)}_{\cdot,i}$, where

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

Likewise, we decompose $\hat{\theta}_j = \hat{\theta}_j^{(1)} + \hat{\theta}_j^{(2)}$. Then we estimate

$$\begin{aligned} n\ell(\hat{z}, z^*) &\le \sum_{i=1}^{n} \sum_{a\neq z_i^*} \mathbb{I}\{\|\hat{P}^{(1)}_{\cdot,i} - \hat{\theta}^{(1)}_a\|^2 - \|\hat{P}^{(1)}_{\cdot,i} - \hat{\theta}^{(1)}_{z_i^*}\|^2 \le \gamma\Delta^2\} \\ &\quad + \sum_{i=1}^{n} \sum_{a\neq z_i^*} \mathbb{I}\{\gamma\Delta^2 \le -\|\hat{P}^{(2)}_{\cdot,i} - \hat{\theta}^{(2)}_a\|^2 + \|\hat{P}^{(2)}_{\cdot,i} - \hat{\theta}^{(2)}_{z_i^*}\|^2\} \\ &=: \sum_{i=1}^{n} \sum_{a\neq z_i^*} A_{i,a} + \sum_{i=1}^{n} \sum_{a\neq z_i^*} B_{i,a}. \end{aligned} \tag{16}$$

for some $\gamma = o(1)$ such that $\gamma\Delta/k \to \infty$. The value of γ will be given in the complete proof. We now investigate the two double-sums above separately.

Step 2 (Sketch of Section 4.4.2). Here, we consider the terms $A_{i,a}$ in the first double-sum above. Lemma 4.2 shows that $\{\hat{\theta}_j\}_{j\in[k]}$ are close to their true values $\{\theta_j^*\}_{j\in[k]}$:

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

Together with the fact that the centers $\{\theta_j^*\}_{j=1}^k$ are separated by Δ and that $\max_{j\geq r+1}\hat{\sigma}_j$ is relatively small, we bound

$$\begin{aligned}A_{i,a} &= \mathbb{I}\{\|\hat{P}_{\cdot,i}^{(1)} - \hat{\theta}_a^{(1)}\|^2 - \|\hat{P}_{\cdot,i}^{(1)} - \hat{\theta}_{z_i^*}^{(1)}\|^2 \leq \gamma\Delta^2\}\\ &\leq \mathbb{I}\{(1-o(1))\Delta \leq 2\|\hat{P}_{\cdot,i}^{(1)} - \hat{U}_{1:r}\hat{U}_{1:r}^T\theta_{z_i^*}^*\|\}.\end{aligned}$$

Next, we observe that $\hat{P}_{\cdot,i}^{(1)} - \hat{U}_{1:r}\hat{U}_{1:r}^T\theta_{z_i^*}^* = \hat{U}_{1:r}\hat{U}_{1:r}^T(\hat{P}-P)e_i$ and show that $\|\hat{U}_{1:r}\hat{U}_{1:r}^T(\hat{P}-P)VV^Te_i\| = o(\Delta)$ by using that $|V_{i,j}| \leq \sqrt{k/(n\beta)}$. Hence, we obtain that

$$\begin{aligned}A_{i,a} &\leq \mathbb{I}\{(1-o(1))\Delta \leq 2\|\hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P}(I-VV^T)e_i\|\}\\ &= \mathbb{I}\{(1-o(1))\Delta \leq 2\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T(I-VV^T)e_i\|\}.\end{aligned}$$

Since the singular values may vary in magnitude, a direct application of spectral perturbation theory on $\hat{V}_{1:r}$ is not sufficient. Instead, we split $[r]$ into disjoint sets $\bigcup_{1\leq m\leq s} J_m$, such that the condition number in each set equals approximately 1, that is, $\max_{j\in J_m}\sigma_j / \min_{j\in J_m}\sigma_j = 1+o(1)$, and such that the the singular value gaps among $\{J_m\}_{m\in[s]}$ are sufficiently large. We carefully explain how to construct these sets in the complete proof. We define $\hat{\Sigma}_{J_m\times J_m}$, $\hat{V}_{J_m}$, V_{J_m}, w_{J_m} as the corresponding parts of the related quantities. We first replace $\hat{\Sigma}_{r\times r}$ above with $\Sigma_{\times r}$. Indeed, using the variational characterization of the Euclidean norm we have for some $w = (w_{J_1},\dots,w_{J_s})$, $\|w\|=1$, that

$$\begin{aligned}\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T(I-VV^T)e_i\| &= \sum_{m\in[s]} e_i^T(I-VV^T)\hat{V}_{J_m}\Sigma_{J_m\times J_m}w_{J_m}\\ &= \sum_{m\in[s]} e_i^T(I-VV^T)\hat{V}_{J_m}\hat{V}_{J_m}^T V_{J_m}\Sigma_{J_m\times J_m}w'_{J_m},\end{aligned}$$

for some $w'\in\mathbb{R}^r$. Since in each set J_m the condition number is bounded by $1+o(1)$ and since $\|(\hat{V}_{1:r}^T V_{1:r})^{-1}\| = 1+o(1)$, we can estimate $\|w'\| \leq 1+o(1)$. Thus, we obtain that

$$\begin{aligned}&\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T(I-VV^T)e_i\|\\ &\quad\leq (1+o(1))\sup_{w\in\mathbb{R}^r:\|w\|=1}\sum_{m\in[s]} e_i^T(I-VV^T)(\hat{V}_{J_m}\hat{V}_{J_m}^T - V_{J_m}V_{J_m}^T)V_{J_m}\Sigma_{J_m\times J_m}w_{J_m}.\end{aligned}$$

The rest of the proof in this section consists of using spectral perturbation theory to show that $(I-VV^T)\hat{V}_{1:r}^T$ equals (up to a small order error term) a linear function of the noise matrix E. Applying Lemma 4.3, we show that the above sum is linear in E (up to a $o(\Delta)$ error term) and obtain

$$\begin{aligned}&\|\Sigma_{r\times r}\hat{V}_{1:r}^T(I-VV^T)e_i\|\\ &\quad= \sup_{w\in\mathbb{R}^r:\|w\|=1}\sum_{m\in[s]} e_i^T\Bigg(\sum_{l\in J_m}\frac{1}{\sigma_l}(I-VV^T)E^Tu_lv_l^TV_{J_m} + S_m\Bigg)\Sigma_{J_m\times J_m}w_{J_m}\\ &\quad= \|U_{1:r}^TE(I-VV^T)e_i\| + o(\Delta).\end{aligned}$$

Hence, summarizing, on the event $\mathcal{F}\cap\mathcal{H}_G$ we bound

$$\sum_{i=1}^n\sum_{a\neq z_i^*} A_{i,a} \leq k\sum_{i=1}^n \mathbb{I}\{(1-o(1))\Delta \leq 2\|U_{1:r}^TE(I-VV^T)e_i\|\}.$$

The tail probability and expectation of $\|U_{1:r}^T E(I - VV^T)e_i\|^2$ are bounded by the tail probability and expectation of a chi-square distributed random variable with k degrees of freedom, χ_k^2. Thus, there exist $\{\xi_i\}_{i\in[n]} \overset{\text{i.i.d.}}{\sim} \chi_k^2$, such that on the event $\mathcal{F} \cap \mathcal{F}'$,

$$\sum_{i=1}^{n} \sum_{a \neq z_i^*} A_{i,a} \leq k \sum_{i=1}^{n} \mathbb{I}\{(1 - o(1))\Delta \leq 2\sqrt{\xi_i}\}.$$

The tail probability of the square root of a χ^2 distribution can be bounded by using Borell's inequality, and hence we obtain that

$$\mathbb{E} \sum_{i=1}^{n} \sum_{a \neq z_i^*} A_{i,a} \mathbb{I}\{\mathcal{F} \cap \mathcal{F}'\} \leq nk \exp\bigl(-(1 - o(1))\Delta^2/8\bigr).$$

Step 3 (Sketch of Section 4.4.3). We next provide an upper bound on the $B_{i,a}$-terms in (16), corresponding to small singular values. We have that

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

which, up to some constant scalar, can be upper bounded by $\sum_{l=r+1}^{k} \sqrt{n} |\hat{V}_{i,l}|$ by construction of r and Weyl's inequality. Hence, on the event $\mathcal{F}$ we obtain that

$$\begin{aligned} B_{i,a} &:= \mathbb{I}\{\gamma \Delta^2 \leq -\|\hat{P}_{\cdot,i}^{(2)} - \hat{\theta}_a^{(2)}\|^2 + \|\hat{P}_{\cdot,i}^{(2)} - \hat{\theta}_{z_i^*}^{(2)}\|^2\} \\ &\leq \sum_{l=r+1}^{k} \mathbb{I}\{c\gamma \Delta^2/k \leq \sqrt{n} |e_i^T \hat{v}_l|\}. \end{aligned}$$

We decompose $e_i^T \hat{v}_l = e_i^T VV^T \hat{v}_l + e_i^T (I - VV^T)\hat{v}_l$. Since, by Lemma A.2 $|V_{ij}| \leq \sqrt{k/(n\beta)}$, the first term in this decomposition is negligible, leaving $(I - VV^T)\hat{v}_l^T$ as the main term to be analyzed.

We apply Lemma 4.4 to show that, after normalization, $(I - VV^T)\hat{v}_l^T$ is Haar distributed on the unit sphere spanned by $I - VV^T$. Hence, on an event $\mathcal{T}$, $e_i^T (I - VV^T)\hat{v}_l^T$ has a Gaussian tail and variance at most $3/(n - k)$. This yields

$$\begin{aligned} \mathbb{E} B_{i,a} \mathbb{I}\{\mathcal{F} \cap \mathcal{T}\} &\leq \sum_{l=r+1}^{k} \mathbb{E}\mathbb{I}\{c'\gamma \Delta^2/k \leq \sqrt{n} |e_i^T (I - VV^T)\hat{v}_l|\} \mathbb{I}\{\mathcal{T}\} \\ &\leq k \exp\bigl(-c''(\gamma \Delta^2 k^{-1})^2\bigr). \end{aligned}$$

Step 4 (Sketch of Section 4.4.4). Summarizing the previous two sections, we obtain that

$$\begin{aligned} \mathbb{E} n \ell(\hat{z}, z) \mathbb{I}\{\mathcal{F} \cap \mathcal{H}_G \cap \mathcal{T}\} &\leq \sum_{i=1}^{n} \sum_{a \neq z_i^*} \mathbb{E}(A_{i,a} + B_{i,a}) \mathbb{I}\{\mathcal{F} \cap \mathcal{H}_G \cap \mathcal{T}\} \\ &\leq nk \exp\biggl(-(1 - o(1))\frac{\Delta^2}{8}\biggr) + k^2 n \exp\bigl(-c(\gamma \Delta k^{-1})^2 \Delta^2\bigr) \\ &= n \exp\biggl(-(1 - o(1))\frac{\Delta^2}{8}\biggr). \end{aligned}$$

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

$$\ell(\hat{z}, z^*) \mathbb{I}\{\mathcal{F} \cap \mathcal{H}_G \cap \mathcal{T}\} \leq \exp\bigl(-(1 - o(1))\Delta^2/8\bigr).$$

Finally, a union bound with $\mathbb{P}(\mathcal{F} \cap \mathcal{H}_G \cap \mathcal{T})$ leads to the desired rate for $\ell(\hat{z}, z^*)$.

4.4. *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 4.3). In Section 4.4.1, we establish the decomposition $\ell(\hat{z}, z^*) \le A + B$. Then in Section 4.4.2 and Section 4.4.3, we provide upper bounds on $\mathbb{E}A$ and $\mathbb{E}B$, respectively. Finally in Section 4.4.4, we wrap everything up to achieve the desired rate. Again, throughout the whole proof, we assume the random event $\mathcal{F}$ (defined in (12)) holds.

Applying Lemma 4.1, we obtain that it suffices to bound $\ell(\hat{z}', z^*)$ where $\hat{z}'$ is the output of Algorithm 3. Indeed, Lemma 4.1 proves that 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 assume that ϕ_0 is the identity mapping. By definition of the k-means objective in (11), we have that

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

In particular, $\hat{z}$ fulfills the *local* optimality condition

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

Hence, assuming without loss of generality that $\phi = \mathrm{Id}$, we obtain that

$$n\ell(\hat{z}, z^*) = \sum_{i=1}^n \mathbb{I}\{\operatorname*{arg\,min}_{a\in[k]} \|\hat{P}_{\cdot,i} - \hat{\theta}_a\|^2 \ne z_i^*\} \tag{17}$$

$$\le \sum_{i=1}^n \sum_{a\ne z_i^*} \mathbb{I}\{\|\hat{P}_{\cdot,i} - \hat{\theta}_a\|^2 \le \|\hat{P}_{\cdot,i} - \hat{\theta}_{z_i^*}\|^2\} \triangleq \sum_{i=1}^n \sum_{a\ne z_i^*} T_{i,a}. \tag{18}$$

4.4.1. *Decomposing $\ell(\hat{z}, z^*)$.* We decompose $\{\hat{P}_{\cdot,i}\}_{i\in[n]}, \{\hat{\theta}_j\}_{j\in[k]}$ into two parts: the first part corresponds to singular values that are above the detection threshold and where $\hat{P}_{\cdot,i}$ contains signal and the second part consists of the remainder noise term. We define $r \in [k]$ as (with $\sigma_{k+1} := 0$)

$$r := \max\{j \in [k] : \sigma_j - \sigma_{j+1} \ge \rho\sqrt{n+p}\}, \tag{19}$$

for a sequence $\rho \to \infty$ 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 $\rho\sqrt{n+p}$. An immediate implication is

$$\max_{r+1\le j\le k} \sigma_j \le k\rho\sqrt{n+p}. \tag{20}$$

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

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

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

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

With this notation and due to the orthogonality of $\{\hat{u}_l\}_{l\in[k]}$, we obtain that

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

We denote by $\rho''=o(1)$ another sequence which we will specify later. We split the indicator function above according to our decomposition and obtain that

$$\begin{aligned}T_{i,a} \le \mathbb{I}\biggl\{\|\hat{\theta}^{(1)}_{z_i^*}-\hat{\theta}^{(1)}_a\|-\frac{\rho''\Delta^2+\|\hat{\theta}^{(2)}_{z_i^*}\|^2}{\|\hat{\theta}^{(1)}_{z_i^*}-\hat{\theta}^{(1)}_a\|}\le 2\|\hat{P}^{(1)}_{\cdot,i}-\hat{\theta}^{(1)}_{z_i^*}\|\biggr\}\\ +\mathbb{I}\{\rho''\Delta^2\le 2\langle \hat{P}^{(2)}_{\cdot,i},\hat{\theta}^{(2)}_a-\hat{\theta}^{(2)}_{z_i^*}\rangle\}=:A_{i,a}+B_{i,a},\end{aligned}$$

where we also used the Cauchy–Schwarz inequality. We now consider $A_{i,a}$ and $B_{i,a}$ separately.

4.4.2. *Upper bounds on* $\mathbb{E}A_{i,a}$. By Lemma 4.2, we have on the event $\mathcal{F}$ that $\max_{j\in[k]}\|\hat{\theta}_j-\theta^*_{\phi'(j)}\|\le 8\sqrt{2}\sqrt{\beta^{-1}k^2(1+p/n)}$ for some label permutation mapping $\phi'\in\Phi$. Without loss of generality, we assume again that $\phi'=\mathrm{Id}$. Define $\hat{Z}\in\{0,1\}^{n\times k}$ to be the estimated label matrix, that is, $\hat{Z}_{i,j}=\mathbb{I}\{\hat{z}_i=j\}$. With this notation and by definition of the k-means objective, we obtain that

$$\hat{\theta}_j=\frac{\sum_{\hat{z}_i=j}\hat{P}_{\cdot,i}}{\sum_{\hat{z}_i=j}1}=\frac{\hat{P}\hat{Z}_{\cdot,j}}{|\{i\in[n]:\hat{z}_i=j\}|}=\frac{\sum_{l\in[k]}\hat{\sigma}_l\hat{u}_l\hat{v}_l^T\hat{Z}_{\cdot,j}}{|\{i\in[n]:\hat{z}_i=j\}|}. \tag{21}$$

Hence, using the above, we obtain that

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

By (20) and Lemma B.2, we have on the event $\mathcal{F}$ that

$$\max_{r+1\le j\le k}\hat{\sigma}_j\le\sqrt{2}(\sqrt{n}+\sqrt{p})+\max_{r+1\le j\le k}\sigma_j\le(k\rho+4)\sqrt{n+p}. \tag{22}$$

By Lemma 4.2, we have that $|\{i\in[n]:\hat{z}_i=j\}|\ge\frac{\beta n}{2k}$, and thus we obtain

$$\max_{j\in[k]}\max_{r+1\le l\le k}|\langle\hat{u}_l,\hat{\theta}_j\rangle|\mathbb{I}\{\mathcal{F}\}\le(k\rho+4)\sqrt{\frac{2k}{\beta}\biggl(1+\frac{p}{n}\biggr)}. \tag{23}$$

Consequently, we bound, working on the event $\mathcal{F}$,

$$\begin{aligned}\max_{j\in[k]}\|\hat{\theta}^{(2)}_j\|^2&=\max_{j\in[k]}\sum_{r+1\le l\le k}\langle\hat{u}_l,\hat{\theta}_j\rangle^2\\ &\le\frac{2k^2}{\beta}\biggl(1+\frac{p}{n}\biggr)(k\rho+4)^2.\end{aligned} \tag{24}$$

Applying Lemma 4.2, we have on the event $\mathcal{F}$ for any $a\ne b$ that

$$\|\hat{\theta}_b-\hat{\theta}_a\|\ge\|\theta^*_b-\theta^*_a\|-\|\hat{\theta}_b-\theta^*_b\|-\|\theta^*_a-\hat{\theta}_a\|\ge\Delta-16\sqrt{2}\sqrt{\beta^{-1}k^2(1+p/n)}.$$

Hence, using also (24), we have on the event $\mathcal{F}$ that

$$\begin{aligned}\min_{a,b\in[k]:a\neq b}\|\hat{\theta}_b^{(1)}-\hat{\theta}_a^{(1)}\| &\geq \min_{a,b\in[k]:a\neq b}(\|\hat{\theta}_b-\hat{\theta}_a\|-\|\hat{\theta}_a^{(2)}\|-\|\hat{\theta}_b^{(2)}\|)\\ &\geq \Delta-(16\sqrt{2}+2\sqrt{2}(k\rho+4))\sqrt{\beta^{-1}k^2\Big(1+\frac{p}{n}\Big)}.\end{aligned}\tag{25}$$

Therefore, by the above, we obtain that

$$\begin{aligned}A_{i,a}\mathbb{I}\{\mathcal{F}\}\leq\mathbb{I}\Bigg\{&\Bigg(\Delta-(16\sqrt{2}+2\sqrt{2}(k\rho+6))\sqrt{\beta^{-1}k^2\Big(1+\frac{p}{n}\Big)}\Bigg)\\ &-\frac{\rho''\Delta^2+\frac{2k^2}{\beta}(1+\frac{p}{n})(k\rho+4)^2}{\Delta-(16\sqrt{2}+2\sqrt{2}(k\rho+4))\sqrt{\beta^{-1}k^2(1+\frac{p}{n})}}\leq 2\|\hat{P}_{\cdot,i}^{(1)}-\hat{\theta}_{z_i^*}^{(1)}\|\Bigg\}\mathbb{I}\{\mathcal{F}\}.\end{aligned}$$

For simplicity, define

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

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

$$A_{i,a}\mathbb{I}\{\mathcal{F}\}\leq\mathbb{I}\Bigg\{\Big(1-c_1\rho''-\frac{c_1k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\Big)\Delta\leq 2\|\hat{P}_{\cdot,i}^{(1)}-\hat{\theta}_{z_i^*}^{(1)}\|\Bigg\}\mathbb{I}\{\mathcal{F}\}.$$

Still working on the event $\mathcal{F}$, we further bound

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

where the last inequality is due to Lemma 4.2. Since $\theta_{z_i^*}^*=P_{\cdot,i}$, we have that $\hat{P}_{\cdot,i}^{(1)}-\hat{U}_{1:r}\hat{U}_{1:r}^T\theta_{z_i^*}^*=(\hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P}-\hat{U}_{1:r}\hat{U}_{1:r}^TP)e_i$. Thus, we obtain that

$$\hat{P}_{\cdot,i}^{(1)}-\hat{U}_{1:r}\hat{U}_{1:r}^T\theta_{z_i^*}^*=\hat{U}_{1:r}\hat{U}_{1:r}^T(\hat{P}-P)VV^Te_i+\hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P}(I-VV^T)e_i.$$

We first bound $\hat{U}_{1:r}\hat{U}_{1:r}^T(\hat{P}-P)VV^Te_i$. Indeed, by Proposition A.1 and Lemma 4.2 we have on the event $\mathcal{F}$ that

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

Thus, there exists some constant $c_2>0$ such that

$$\begin{aligned}A_{i,a}\mathbb{I}\{\mathcal{F}\}&\leq\mathbb{I}\Bigg\{\Big(1-c_1\rho''-\frac{c_2k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\Big)\Delta\leq 2\|\hat{U}_{1:r}\hat{U}_{1:r}^T\hat{P}(I-VV^T)e_i\|\Bigg\}\mathbb{I}\{\mathcal{F}\}\\ &=\mathbb{I}\Bigg\{\Big(1-c_1\rho''-\frac{c_2k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\Big)\Delta\leq 2\|\hat{\Sigma}_{r\times r}\hat{V}_{1:r}^T(I-VV^T)e_i\|\Bigg\}\mathbb{I}\{\mathcal{F}\},\end{aligned}$$

where we define $\hat{\Sigma}_{r\times r}=\operatorname{diag}\{\hat{\sigma}_1,\ldots,\hat{\sigma}_r\}$ and $\hat{V}_{1:r}=(\hat{v}_1,\ldots,\hat{v}_r)$. We define the corresponding population counterparts analogue.

For any unit vector $w \in \mathbb{R}^r$, define $w' = \Sigma_{r\times r}^{-1}\hat{\Sigma}_{r\times r} w$. This yields the identity $\Sigma_{r\times r} w' = \hat{\Sigma}_{r\times r} w$. By definition of r and Lemma B.2, we obtain that

$$\max_{j\in[r]} \frac{\hat{\sigma}_j}{\sigma_j} \le \max_{j\in[r]} \frac{\sigma_j + 4\sqrt{n+p}}{\sigma_j} \le 1 + 6\rho^{-1}$$

and, therefore, $\|w'\| \le 1+6\rho^{-1}$. Thus, using the variational characterization of the Euclidean norm, we bound

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

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

We further investigate $e_i^T(I - VV^T)\hat{V}_{1:r}\Sigma_{r\times r} w$. First, we partition the leading $[r]$ singular values. Define s as

$$s := \left|\left\{l \in [r] : \frac{\sigma_l - \sigma_{l+1}}{\sigma_{l+1}} \ge \frac{1}{\rho' k}\right\}\right|, \tag{26}$$

for some $\rho' \to \infty$ whose value will be specified later. We denote its entries by $j_1' < j_2' < \cdots < j_s'$. Due to (20), we have that $j_s' = r$. We define $j_0' = 0$,

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

and split $[r]$ into disjoint sets $\{J_m\}_{m=1}^s$, where $J_m = \{j_m, j_m+1, \ldots, j_m'\}$. This partition has the following properties:

- Defining the singular value gaps $g_m := \min\{\sigma_{j_{m-1}'} - \sigma_{j_m}, \sigma_{j_m'} - \sigma_{j_{m+1}}\}$, $m \in [s]$, with $j_{s+1} = r+1$ and $\sigma_0 = +\infty$, we obtain by (26) for any $m \in [s-1]$ that

$$\sigma_{j_m'} - \sigma_{j_{m+1}} = \sigma_{j_m'} - \sigma_{j_m'+1} \ge \frac{\sigma_{j_m'+1}}{\rho' k} \ge \frac{\sigma_r}{\rho' k} \ge \frac{\rho\sqrt{n+p}}{\rho' k}.$$

  Hence, we obtain that

$$\min_{m\in[s]} g_m \ge \frac{\rho\sqrt{n+p}}{\rho' k}. \tag{27}$$

- The set defined in (26) has an alternative representation, that is,

$$\left\{l \in [r] : \frac{\sigma_l - \sigma_{l+1}}{\sigma_{l+1}} \ge \frac{1}{\rho' k}\right\} = \left\{l \in [r] : \frac{\sigma_l}{\sigma_{l+1}} > 1 + \frac{1}{\rho' k}\right\}.$$

  Therefore, and since $\rho' \to \infty$, we obtain that

$$\max_{m\in[s]} \frac{\sigma_{j_m}}{\sigma_{j_m'}} \le \left(1 + \frac{1}{\rho' k}\right)^{|J_m|} \le \left(1 + \frac{1}{\rho' k}\right)^k \le 1 + \frac{2}{\rho'}. \tag{28}$$

- Due to (26), we have that

$$\max_{m\in[s]} \frac{\sigma_{j_m'}}{\sigma_{j_m'} - \sigma_{j_m'+1}} \le \max_{m\in[s]} \frac{1 + \sigma_{j_m'+1}}{\sigma_{j_m'} - \sigma_{j_m'+1}} \le 1 + \rho' k.$$

  Hence, using also (28) and since $\rho' \to \infty$, we obtain that

$$\max_{m\in[s]} \frac{\sigma_{j_m} - \sigma_{j_m'}}{g_m} \le \frac{2}{\rho'} \max_{m\in[s]} \frac{\sigma_{j_m'}}{\sigma_{j_m'} - \sigma_{j_m'+1}} \le \frac{2}{\rho'}(1 + \rho' k) \le 3k, \tag{29}$$

and

$$\max_{m\in[s]}\frac{\sigma_{j_m}}{g_m}\le\left(1+\frac{2}{\rho'}\right)\max_{m\in[s]}\frac{\sigma_{j'_m}}{\sigma_{j'_m}-\sigma_{j'_m+1}}\le 1+2\rho' k. \tag{30}$$

Now, consider any fixed $w\in\mathbb{R}^r$. For $m\in[s]$, we denote $\hat{V}_{J_m}=(\hat{v}_{j_m},\dots,\hat{v}_{j'_m})$, $V_{J_m}=(v_{j_m},\dots,v_{j'_m})$, $\Sigma_{J_m\times J_m}=\operatorname{diag}\{\sigma_{j_m},\dots,\sigma_{j'_m}\}$, and $w_{J_m}=(w_{j_m},\dots,w_{j'_m})$. Applying this notation, we have that

$$e_i^T(I-VV^T)\hat{V}_{1:r}\Sigma_{r\times r}w=\sum_{m\in[s]}e_i^T(I-VV^T)\hat{V}_{J_m}\Sigma_{J_m\times J_m}w_{J_m}. \tag{31}$$

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

$$\begin{aligned}\|\hat{V}_{J_m}-V_{J_m}O_m\|&\le\sqrt{2}\|\hat{V}_{J_m}\hat{V}_{J_m}^T-V_{J_m}V_{J_m}^T\|\\&\le\frac{4\sqrt{2}\|E\|}{g_m}\le\frac{16\sqrt{2}\rho' k}{\rho},\end{aligned} \tag{32}$$

where we use (27) in the last inequality. Moreover, we have that

$$\|\hat{V}_{J_m}^TV_{J_m}O_m-I\|\le\|\hat{V}_{J_m}-V_{J_m}O_m\|\le\frac{16\sqrt{2}\rho' k}{\rho}, \tag{33}$$

and hence, choosing ρ and ρ' such that $\rho/(\rho' k)>16\sqrt{2}$, we obtain that $V_{J_m}^T\hat{V}_{J_m}$ is invertible and

$$\|(V_{J_m}^T\hat{V}_{J_m})^{-1}\|\le\left(1-\frac{16\sqrt{2}\rho' k}{\rho}\right)^{-1}. \tag{34}$$

Now, for fixed w_{J_m} we define

$$w'_{J_m}=\Sigma_{J_m\times J_m}^{-1}(\hat{V}_{J_m}^TV_{J_m})^{-1}\Sigma_{J_m\times J_m}w_{J_m}\quad\forall m\in[s]. \tag{35}$$

Plugging the above into (31), we obtain that

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

By definition of w'_{J_m}, we have that

$$\begin{aligned}\max_{m\in[s]}\frac{\|w'_{J_m}\|}{\|w_{J_m}\|}&\le\max_{m\in[s]}\|\Sigma_{J_m\times J_m}^{-1}(\hat{V}_{J_m}^TV_{J_m})^{-1}\Sigma_{J_m\times J_m}\|\\&\le(1+2\rho'^{-1})\left(1-\frac{16\sqrt{2}\rho' k}{\rho}\right)^{-1},\end{aligned}$$

where we used in the last inequality that $\max_{m\in[s]}\|\Sigma_{J_m\times J_m}^{-1}\|\|\Sigma_{J_m\times J_m}\|\le 1+2\rho'^{-1}$ by (28) and the upper bound (34). Hence, using also that $(I-VV^T)V_{J_m}=0$, we obtain that

$$\begin{aligned}&\sup_{w:\|w\|\le1}w^T\Sigma_{r\times r}\hat{V}_{1:r}^T(I-VV^T)e_i\\&\le\frac{1+2\rho'^{-1}}{1-16\sqrt{2}\rho'k\rho^{-1}}\sup_{w:\|w\|\le1}\sum_{m\in[s]}e_i^T(I-VV^T)(\hat{V}_{J_m}\hat{V}_{J_m}^T-V_{J_m}V_{J_m}^T)V_{J_m}\Sigma_{J_m\times J_m}w_{J_m}.\end{aligned}$$

We further evaluate the term on the right-hand side above. Applying Lemma 4.3, we obtain that

$$e_i^T(I-VV^T)(\hat{V}_{J_m}\hat{V}_{J_m}^T - V_{J_m}V_{J_m}^T)V_{J_m}\Sigma_{J_m\times J_m}w_{J_m}$$
$$= \sum_{l\in J_m} w_l e_i^T(I-VV^T)E^T u_l + e_i^T\mathbb{E}S_m\Sigma_{J_m\times J_m}w_{J_m} + e_i^T(S_m-\mathbb{E}S_m)\Sigma_{J_m\times J_m}w_{J_m}.$$

We next show that $\mathbb{E}S_m = 0$. Indeed, we have that

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

Applying Lemma 4.4, we obtain that $(I-VV^T)\hat{v}_j/\|(I-VV^T)\hat{v}_j\|$ and $\|(I-VV^T)\hat{v}_j\|\times V_{J_m}^TVV^T\hat{v}_j$ are independent. Hence, using Lemma 4.4 again, we obtain that

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

Hence, we obtain that

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

Summarizing, we obtain that

$$A_{i,a} \le \mathbb{I}\Big\{\frac{1-16\sqrt{2}\rho' k\rho^{-1}}{(1+6\rho^{-1})(1+2\rho'^{-1})}\Big(1-c_1\rho'' - \frac{c_2k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}\Big)\Delta$$
$$\le 2\|U_{1:r}^TE(I-VV^T)e_i\| + \sup_{w\in\mathbb{R}^r:\|w\|\le 1}\sum_{m\in[s]} e_i^T(S_m-\mathbb{E}S_m)\Sigma_{J_m\times J_m}w_{J_m}\Big\}.$$

We next bound the higher order perturbation term on the right-hand side. Applying Lemma B.1 and by construction of the partition, we obtain that $g_m \ge 8\mathbb{E}\|E\|$, and hence we can apply Lemma 4.3. Note that $\|\Sigma_{J_m\times J_m}w_{J_m}e_i^T\|_* = \|\Sigma_{J_m\times J_m}w_{J_m}\|\|e_i^T\| \le \sigma_{j_m}\|w_{J_m}\|$. Together with (27), (29) and (30), for some constant $c_0>0$, we have with probability at least $1-2e^{-(\Delta^2\wedge n)}$ that

$$|e_i^T(S_m-\mathbb{E}S_m)\Sigma_{J_m\times J_m}w_{J_m}|$$
$$\le c_0\Big(1+\frac{\sigma_{j_m}-\sigma_{j_m'}}{g_m}\Big)\frac{\Delta}{g}\Big(\frac{\sqrt{n+p}}{g}\Big)\sigma_{j_m}\|w_{J_m}\|$$
$$\le 16c_0\rho^{-1}k^3\rho'^2\Delta\|w_{J_m}\|.$$

Taking a union bound over J_m and since $\sum_m\|w_{J_m}\| \le \sqrt{k}\|w\| = \sqrt{k}$ we obtain with probability at least $1-2k\exp(-(\Delta^2\wedge n))$ that

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

By applying a standard ε-net argument with a union bound, we obtain with probability at least $1 - 2ke^k \exp(-(\Delta^2 \wedge n))$ that

$$\sup_{w\in\mathbb{R}^r:\|w\|\le 1} \sum_{m\in[s]} e_i^T (S_m - \mathbb{E}S_m)\Sigma_{J_m\times J_m} w_{J_m} \le 32c_0\rho^{-1}k^{\frac{7}{2}}\rho'^2\Delta. \tag{36}$$

We denote by $\mathcal{H}_i$ the event where (36) above holds and note that $\mathbb{P}(\mathcal{H}_i) \ge 1 - 2ke^k \times \exp(-(\Delta^2 \wedge n))$. To avoid that $\Delta \wedge n$ (instead of Δ) appears in the convergence rate, we further introduce the global event

$$\mathcal{H}_G := \left\{ \{\Delta > \sqrt{n}\} \bigcap_{i=1}^n \mathcal{H}_i \right\} \cup \{\Delta \le \sqrt{n}\}$$

and note that

$$\mathbb{P}(\mathcal{H}_G) \ge 1 - 2nke^{-n+k}. \tag{37}$$

We are finally ready to bound $A_{i,a}$. Indeed, by the above we obtain that

$$\begin{aligned}\mathbb{E}A_{i,a}\mathbb{I}\{\mathcal{F}\cap\mathcal{H}_G\} \le \mathbb{E}\mathbb{I}\Bigg\{ &\frac{(1-16\sqrt{2}\rho' k\rho^{-1})(1-c_1\rho'' - \frac{c_2k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta})}{(1+6\rho^{-1})(1+2\rho'^{-1})}\Delta \\ &- 32c_0\rho^{-1}k^{\frac{7}{2}}\rho'^2\Delta \le 2\|U_{1:r}^T E(I-VV^T)e_i\|\Bigg\} \\ &+ \mathbb{E}\mathbb{I}\{\mathcal{H}_i \cap \{\Delta < \sqrt{n}\}\}.\end{aligned}$$

We observe that $U_{1:r}^T E(I - VV^T)e_i \sim \mathcal{N}(0, \|(I - VV^T)e_i\|^2 I_{r\times r})$. Moreover, since $\|(I - VV^T)e_i\| \le 1$, we have that

$$\mathbb{P}(\|U_{1:r}^T E(I-VV^T)e_i\|^2 > t) \le \mathbb{P}(\xi_i > t),$$

where by ξ_i we denote a chi-square distributed random variable with k degrees of freedom. Hence, assuming additionally 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

$$\begin{aligned}&\mathbb{E}A_{i,a}\mathbb{I}\{\mathcal{F}\cap\mathcal{H}_G\} \\ &\le \mathbb{I}\left\{\left(1 - c_3\rho'' - \frac{c_3k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta} - \frac{c_3k^{\frac{7}{2}}\rho'^2}{\rho}\right)\Delta \le 2\sqrt{\xi_i}\right\} + 2ke^{-\Delta^2+k} \\ &\le \exp\left(-\frac{1}{8}\left(1 - c_3\rho'' - \frac{c_3k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta} - \frac{c_3k^{\frac{7}{2}}\rho'^2}{\rho} - \frac{2\sqrt{k}}{\Delta}\right)^2\Delta^2\right) + 2ke^{-\Delta^2+k},\end{aligned}$$

where we used Jensen's inequality and Borell's inequality (e.g., Theorem 2.2.7 in [22]) to bound $\mathbb{P}(\sqrt{\xi_i} > t) \le \exp(-(t - \sqrt{k})^2)$.

4.4.3. *Upper bounds on* $\mathbb{E}B_{i,a}$. We now bound

$$B_{i,a} := \mathbb{I}\{\rho''\Delta^2 \le 2\langle \hat{P}_{\cdot,i}^{(2)}, \hat{\theta}_a^{(2)} - \hat{\theta}_{z_i^*}^{(2)}\rangle\}.$$

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

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

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 (22) and (23), we have that

$$|\langle\hat{P}_{\cdot,i}^{(2)},\hat{\theta}_a^{(2)}-\hat{\theta}_{z_i^*}^{(2)}\rangle| \le 2(k\rho+4)^2\sqrt{\frac{2nk}{\beta}\Big(1+\frac{p}{n}\Big)^2}\sum_{l=r+1}^{k}|\hat{V}_{i,l}|, \tag{38}$$

and hence we bound

$$\begin{aligned}\sum_{a\neq z_i^*}B_{i,a}\mathbb{I}\{\mathcal{F}\} &\le k\mathbb{I}\Big\{\rho''\Delta^2 \le 4(k\rho+4)^2\sqrt{\frac{2nk}{\beta}\Big(1+\frac{p}{n}\Big)^2}\sum_{l=r+1}^{k}|\hat{V}_{i,l}|\Big\}\\ &\le k\sum_{l=r+1}^{k}\mathbb{I}\Big\{\rho''\Delta^2 \le 4k(k\rho+4)^2\sqrt{\frac{2nk}{\beta}\Big(1+\frac{p}{n}\Big)^2}|\hat{V}_{i,l}|\Big\} =: k\sum_{l=r+1}^{k}C_{i,l}.\end{aligned}$$

We bound each $C_{i,l}$ separately, by showing that $\hat{V}_{i,l}$ is, approximately, univariate Gaussian with variance $1/n$. We first apply Proposition A.2 to obtain that

$$|\hat{V}_{i,l}| \le \|V^Te_i\| + |e_i^T(I-VV^T)\hat{v}_l| \le \sqrt{\beta^{-1}k/n} + |e_i^T(I-VV^T)\hat{v}_l|.$$

Hence, assuming that $\Delta^2\rho''/(k^4\rho^2\beta^{-1}(1+p/n))$ is large enough and afterwards applying Lemma 4.4, we obtain that for some constant $c_4>0$,

$$\begin{aligned}C_{i,l} &\le \mathbb{I}\Big\{c_4\frac{\rho''\Delta^2}{k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}(1+\frac{p}{n})} \le \sqrt{n}\frac{|e_i^T(I-VV^T)\hat{v}_l|}{\|(I-VV^T)\hat{v}_l\|}\Big\}\\ &\overset{d}{=} \mathbb{I}\Big\{c_4\frac{\rho''\Delta^2}{k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}(1+\frac{p}{n})} \le \sqrt{n}\frac{|e_i^T(I-VV^T)\zeta_{i,l}|}{\|(I-VV^T)\zeta_{i,l}\|}\Big\},\end{aligned}$$

where $\overset{d}{=}$ denotes equality in distribution and where $\zeta_{i,l}\sim\mathcal{N}(0,I_n)$. We next provide a lower bound for the denominator above. Indeed, since $(I-VV^T)\zeta_{i,l}\sim\mathcal{N}(0,(I-VV^T))$, we see that $\|(I-VV^T)\zeta_{i,l}\|^2$ is chi-square distributed with at least $n-k$ degrees of freedom. Hence, using tail-bounds for the lower tail of chi-square distributed random variables (e.g., Lemma 1 in [39]), we obtain that the event $\mathcal{T}$ defined below occurs with high probability, that is,

$$\mathbb{P}(\mathcal{T}) := \mathbb{P}\Big(\bigcap_{i,l}\Big\{\|(I-VV^T)\zeta_{i,l}\|^2 \ge \frac{(n-k)}{3}\Big\}\Big) \ge 1-nk\exp\Big(-\frac{(n-k)}{9}\Big). \tag{39}$$

Hence, working on the event $\mathcal{T}\cap\mathcal{F}$, we bound

$$\begin{aligned}\mathbb{E}C_{i,l}\mathbb{I}\{\mathcal{T}\cap\mathcal{F}\} &\le \mathbb{E}\mathbb{I}\Big\{c_4\frac{\rho''\Delta^2}{k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}(1+\frac{p}{n})}\sqrt{\frac{n-k}{3n}} \le |e_i^T(I-VV^T)\zeta_{i,l}|\Big\}\\ &\le 2\exp\Big(-\frac{1}{2}\Big(c_4\frac{\rho''\Delta}{k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}(1+\frac{p}{n})}\sqrt{\frac{n-k}{3n}}\Big)^2\Delta^2\Big),\end{aligned}$$

where we used that $e_i^T(I-VV^T\zeta_{i,l})$ is univariate Gaussian with variance bounded by 1.

4.4.4. *Obtaining the final result.* Combining the above upper bounds together, we have that

$$\begin{aligned}&\mathbb{E}\ell(\hat{z},z^*)\mathbb{I}\{\mathcal{F}\cap\mathcal{H}_G\cap\mathcal{T}\}\\ &\le \frac{1}{n}\sum_{i=1}^{n}\sum_{a\neq z_i^*}\mathbb{E}A_{i,a}\mathbb{I}\{\mathcal{F}\cap\mathcal{H}_G\} + \frac{k}{n}\sum_{i=1}^{n}\sum_{l=r+1}^{k}\mathbb{E}C_{i,l}\mathbb{I}\{\mathcal{F}\cap\mathcal{T}\}\end{aligned}$$

$$\leq k\exp\biggl(-\frac{1}{8}\biggl(1-c_3\rho''-\frac{c_3k^2\rho\beta^{-\frac{1}{2}}\eta}{\Delta}-\frac{c_3k^{\frac{7}{2}}\rho'^2}{\rho}-\frac{2\sqrt{k}}{\Delta}\biggr)^2\Delta^2\biggr)+2k^2e^{-\Delta^2+k}$$
$$+2k^2\exp\biggl(-\frac{1}{2}\biggl(c_4\frac{\rho''\Delta}{k^{\frac{7}{2}}\rho^2\beta^{-\frac{1}{2}}\eta^2}\sqrt{\frac{n-k}{3n}}\biggr)^2\Delta^2\biggr).$$

Since, by assumption, $\frac{\Delta}{k^{10.5}\beta^{-0.5}\eta^2(\frac{n-k}{n})^{-0.5}}\to\infty$, recalling that $\eta=\sqrt{1+p/n}$ and denoting $\lambda=(\frac{n-k}{n})^{-0.5}$ we can choose

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

to obtain that

$$\mathbb{E}\ell(\hat{z},z^*)\mathbb{I}\{\mathcal{F}\cap\mathcal{H}_G\cap\mathcal{T}\}\leq\exp\biggl(-\biggl(1-\frac{1}{2}\biggl(\frac{\Delta}{k^{10.5}\beta^{-0.5}\eta^2\lambda}\biggr)^{-0.1}\biggr)\frac{\Delta^2}{8}\biggr).$$

Applying Markov's inequality, we obtain that

$$\ell(\hat{z},z^*)\mathbb{I}\{\mathcal{F}\cap\mathcal{H}_G\cap\mathcal{T}\}\leq\exp\biggl(-\biggl(1-\biggl(\frac{\Delta}{k^{10.5}\beta^{-0.5}\eta^2\lambda}\biggr)^{-0.1}\biggr)\frac{\Delta^2}{8}\biggr),$$

with probability at least $1-\exp(-\Delta)$. Finally, the proof is completed by using a union bound accounting for the events $\mathcal{F}$, $\mathcal{H}_G$ and $\mathcal{T}$, where $\mathbb{P}(\mathcal{F}\cap\mathcal{H}_G\cap\mathcal{T})\geq 1-\exp(-0.08n)-3nk\exp(-0.08(n-k))$ by (37), (39) and Lemma B.1.

To obtain the in-expectation result (8), there is no need to apply Markov's inequality and a union bound is sufficient:

$$\begin{aligned}\mathbb{E}\ell(\hat{z},z^*)&\leq\mathbb{E}\ell(\hat{z},z^*)\mathbb{I}\{\mathcal{F}\cap\mathcal{H}_G\cap\mathcal{T}\}+\bigl(1-\mathbb{P}(\mathcal{F}\cap\mathcal{H}_G\cap\mathcal{T})\bigr)\\&\leq\exp\biggl(-\biggl(1-\frac{1}{2}\biggl(\frac{\Delta}{k^{10.5}\beta^{-0.5}\eta^2\lambda}\biggr)^{-0.1}\biggr)\frac{\Delta^2}{8}\biggr)\\&\quad+\exp(-0.08n)+3nk\exp\bigl(-0.08(n-k)\bigr)\\&=\exp\biggl(-\bigl(1-o(1)\bigr)\frac{\Delta^2}{8}\biggr)+\exp\bigl(-\bigl(1-o(1)\bigr)0.08n\bigr).\end{aligned}$$

Acknowledgments. The authors would like to thank Zhou Fan from Yale University for pointing out the references [32, 52]. The authors are further grateful to the Co-Editor, Ming Yuan, an anonymous Associate Editor and three anonymous referees for careful reading of the manuscript and their valuable remarks and suggestions.

Funding. M. Löffler gratefully acknowledges financial support of ERC grant UQMSI/647812 and EPSRC grant EP/L016516/1, which funded a research visit to Yale University, where parts of this work were completed. These grants also funded M. Löffler during his PhD studies at the University of Cambridge.

SUPPLEMENTARY MATERIAL

Supplement to "Optimality of spectral clustering in the Gaussian mixture model" (DOI: 10.1214/20-AOS2044SUPP; .pdf). In the Supplementary Material [42], 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 4.1, Lemma 4.2 and Lemma 4.4. We give an extension of Proposition 2.1 in Appendix D and prove Theorem 2.2 in Appendix E. The proof of Lemma 4.3 is given in Appendix F.

REFERENCES

[1] ABBE, E., FAN, J. and WANG, K. (2020). An ℓ_p-theory of PCA and spectral clustering. Preprint.

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

[3] ALPERT, C. J. and YAO, S. (1995). Spectral partitioning: The more eigenvectors, the better. In 32*nd Design Automation Conference* 195–200. IEEE, New York.

[4] ANANDKUMAR, A., GE, R., HSU, D. and KAKADE, S. M. (2014). A tensor approach to learning mixed membership community models. *J. Mach. Learn. Res.* **15** 2239–2312. MR3231594

[5] BACH, F. R. and JORDAN, M. I. (2006). Learning spectral clustering, with application to speech separation. *J. Mach. Learn. Res.* **7** 1963–2001. MR2274430

[6] BALAKRISHNAN, S., XU, M., KRISHNAMURTHY, A. and SINGH, A. (2011). Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems* 954–962.

[7] BELKIN, M. and NIYOGI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15** 1373–1396.

[8] CHAUDHURI, K., CHUNG, F. and TSIATAS, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory* 35.1–35.23.

[9] CHEN, X. and YANG, Y. (2020). Cutoff for exact recovery of Gaussian mixture models. Preprint.

[10] COJA-OGHLAN, A. (2010). Graph partitioning via adaptive spectral techniques. *Combin. Probab. Comput.* **19** 227–284. MR2593622 https://doi.org/10.1017/S0963548309990514

[11] DHILLON, I. S. (2001). 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* 269–274. ACM, New York.

[12] DING, C., HE, X. and SIMON, H. D. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the* 2005 *SIAM International Conference on Data Mining* 606–610. SIAM, Philadelphia.

[13] DING, C. H. Q., HE, X., ZHA, H., GU, M. and SIMON, H. D. (2001). A min-max cut algorithm for graph partitioning and data clustering. In *Proceedings* 2001 *IEEE International Conference on Data Mining* 107–114. IEEE, New York.

[14] DING, X. (2020). High dimensional deformed rectangular matrices with applications in matrix denoising. *Bernoulli* **26** 387–417. MR4036038 https://doi.org/10.3150/19-BEJ1129

[15] DONATH, W. E. and HOFFMAN, A. J. (2003). Lower bounds for the partitioning of graphs. In *Selected Papers of Alan J. Hoffman*: *With Commentary* 437–442. World Scientific, Singapore.

[16] FEI, Y. and CHEN, Y. (2018). Hidden integrality of SDP relaxations for sub-Gaussian mixture models. In *Conference on Learning Theory* 1931–1965.

[17] FIEDLER, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Math. J.* **23** 298–305. MR0318007

[18] FISHKIND, D. E., SUSSMAN, D. L., TANG, M., VOGELSTEIN, J. T. and PRIEBE, C. E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* **34** 23–39. MR3032990 https://doi.org/10.1137/120875600

[19] FURUI, S. (1989). Unsupervised speaker adaptation based on hierarchical spectral clustering. *IEEE Trans. Acoust. Speech Signal Process.* **37** 1923–1930.

[20] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2018). Community detection in degree-corrected block models. *Ann. Statist.* **46** 2153–2185. MR3845014 https://doi.org/10.1214/17-AOS1615

[21] GINÉ, E. and KOLTCHINSKII, V. (2006). Empirical graph Laplacian approximation of Laplace–Beltrami operators: Large sample results. In *High Dimensional Probability*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **51** 238–259. IMS, Beachwood, OH. MR2387773 https://doi.org/10.1214/074921706000000888

[22] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **40**. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862

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

[24] GUATTERY, S. and MILLER, G. L. (1998). On the quality of spectral separators. *SIAM J. Matrix Anal. Appl.* **19** 701–719. MR1611179 https://doi.org/10.1137/S0895479896312262

[25] HALL, K. M. (1970). An r-dimensional quadratic placement algorithm. *Manage. Sci.* **17** 219–229.

[26] HAN, X., TONG, X. and FAN, Y. (2020). Eigen selection in spectral clustering: A theory guided practice. Preprint.

[27] HEIN, M. (2006). Uniform convergence of adaptive graph-based regularization. In *Learning Theory. Lecture Notes in Computer Science* **4005** 50–64. Springer, Berlin. MR2277918 https://doi.org/10.1007/11776420_7

[28] HEIN, M., AUDIBERT, J.-Y. and VON LUXBURG, U. (2005). From graphs to manifolds—Weak and strong pointwise consistency of graph Laplacians. In *Learning Theory. Lecture Notes in Computer Science* **3559** 470–485. Springer, Berlin. MR2203281 https://doi.org/10.1007/11503415_32

[29] HENDRICKSON, B. and LELAND, R. (1995). An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM J. Sci. Comput.* **16** 452–469. MR1317066 https://doi.org/10.1137/0916028

[30] INABA, M., KATOH, N. and IMAI, H. (1994). Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. In *Proceedings of* 10*th ACM Symposium on Computational Geometry* 332–339.

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

[32] JOHNSTONE, I. M. and PAUL, D. (2018). PCA in high dimensions: An orientation. *Proc. IEEE Inst. Electr. Electron. Eng.* **106** 1277–1292. https://doi.org/10.1109/JPROC.2018.2846730

[33] KANNAN, R. and VEMPALA, S. (2009). Spectral algorithms. *Found. Trends Theor. Comput. Sci.* **4** 157–288. MR2558901 https://doi.org/10.1561/0400000025

[34] KANNAN, R., VEMPALA, S. and VETTA, A. (2004). On clusterings: Good, bad and spectral. *J. ACM* **51** 497–515. MR2145863 https://doi.org/10.1145/990308.990313

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

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

[37] KUMAR, A. and KANNAN, R. (2010). Clustering with spectral norm and the k-means algorithm. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science—FOCS* 2010 299–308. IEEE Computer Soc., Los Alamitos, CA. MR3025203

[38] KUMAR, A., SABHARWAL, Y. and SEN, S. (2004). A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In 45*th Annual IEEE Symposium on Foundations of Computer Science* 454–462.

[39] LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785 https://doi.org/10.1214/aos/1015957395

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

[41] LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28** 129–137. MR0651807 https://doi.org/10.1109/TIT.1982.1056489

[42] LÖFFLER, M., ZHANG, A. Y. and ZHOU, H. H. (2021). Supplement to "Optimality of spectral clustering in the Gaussian mixture model." https://doi.org/10.1214/20-AOS2044SUPP

[43] LU, Y. and ZHOU, H. H. (2016). Statistical and computational guarantees of Lloyd's algorithm and its variants. Preprint.

[44] MAHAJAN, M., NIMBHORKAR, P. and VARADARAJAN, K. (2009). The planar k-means problem is NP-hard. In *WALCOM—Algorithms and Computation. Lecture Notes in Computer Science* **5431** 274–285. Springer, Berlin. MR2540919 https://doi.org/10.1007/978-3-642-00202-1_24

[45] MCSHERRY, F. (2001). Spectral partitioning of random graphs. In 42*nd IEEE Symposium on Foundations of Computer Science* (*Las Vegas*, *NV*, 2001) 529–537. IEEE Computer Soc., Los Alamitos, CA. MR1948742

[46] MEILA, M. and SHI, J. (2001). Learning segmentation by random walks. In *Advances in Neural Information Processing Systems* 873–879.

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

[48] NDAOUD, M. (2019). Sharp optimal recovery in the two component Gaussian mixture model. Preprint.

[49] NG, A. Y., JORDAN, M. I. and WEISS, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* 849–856.
[50] OTTO, F. and VILLANI, C. (2000). Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.* **173** 361–400. MR1760620 https://doi.org/10.1006/jfan.1999.3557
[51] PAN, S. J., NI, X., SUN, J., YANG, Q. and CHEN, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the* 19*th International Conference on World Wide Web* 751–760. ACM, New York.
[52] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865
[53] PENG, J. and WEI, Y. (2007). Approximating k-means-type clustering via semidefinite programming. *SIAM J. Optim.* **18** 186–205. MR2299680 https://doi.org/10.1137/050641983
[54] QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems* 3120–3128.
[55] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887
[56] ROYER, M. (2017). Adaptive clustering through semidefinite programming. *Adv. Neural Inf. Process. Syst.* 1795–1803.
[57] SARKAR, P. and BICKEL, P. J. (2015). Role of normalization in spectral clustering for stochastic blockmodels. *Ann. Statist.* **43** 962–990. MR3346694 https://doi.org/10.1214/14-AOS1285
[58] SHI, J. and MALIK, J. (2000). Normalized cuts and image segmentation. Departmental Papers (CIS) 107.
[59] SIMON, H. D. (1991). Partitioning of unstructured problems for parallel processing. *Comput. Syst. Eng.* **2** 135–148.
[60] SPIELMAN, D. A. and TENG, S.-H. (1996). Spectral partitioning works: Planar graphs and finite element meshes. In 37*th Annual Symposium on Foundations of Computer Science* (*Burlington, VT*, 1996) 96–105. IEEE Comput. Soc. Press, Los Alamitos, CA. MR1450607 https://doi.org/10.1109/SFCS.1996.548468
[61] SRIVASTAVA, P. R., PURNAMRITA, S. and HANASUSANTO, G. A. (2020). A robust spectral clustering algorithm for sub-Gaussian mixture models with outliers. Preprint.
[62] TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 411–423. MR1841503 https://doi.org/10.1111/1467-9868.00293
[63] VAN DRIESSCHE, R. and ROOSE, D. (1995). An improved spectral bisection algorithm and its application to dynamic load balancing. *Parallel Comput.* **21** 29–48. MR1314376 https://doi.org/10.1016/0167-8191(94)00059-J
[64] VEMPALA, S. and WANG, G. (2004). A spectral algorithm for learning mixture models. *J. Comput. System Sci.* **68** 841–860. MR2059647 https://doi.org/10.1016/j.jcss.2003.11.008
[65] VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. MR2409803 https://doi.org/10.1007/s11222-007-9033-z
[66] VON LUXBURG, U., BELKIN, M. and BOUSQUET, O. (2008). Consistency of spectral clustering. *Ann. Statist.* **36** 555–586. MR2396807 https://doi.org/10.1214/009053607000000640
[67] VU, V. (2018). A simple SVD algorithm for finding hidden partitions. *Combin. Probab. Comput.* **27** 124–140. MR3734334 https://doi.org/10.1017/S0963548317000463
[68] WANG, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika* **97** 893–904. MR2746159 https://doi.org/10.1093/biomet/asq061
[69] YU, S. and SHI, J. (2003). Multiclass spectral clustering. In *Proceedings Ninth IEEE International Conference on Computer Vision* 313–319.
[70] ZHANG, A. Y. and ZHOU, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.* **44** 2252–2280. MR3546450 https://doi.org/10.1214/15-AOS1428
[71] ZHOU, Z. and AMINI, A. A. (2019). Analysis of spectral clustering algorithms for community detection: The general bipartite setting. *J. Mach. Learn. Res.* **20** Paper No. 47, 47. MR3948087