Journal of Machine Learning Research 27 (2026) 1-61 Submitted 07/25; Published 10/26

Semidefinite programming relaxations and debiasing for MAXCUT-based clustering

Shuheng Zhou SZHOU@UCR.EDU
Department of Statistics
University of California, Riverside
Riverside, CA 92521, USA

Editor: Zhihua Zhang

Abstract

In this paper, we consider the problem of partitioning a small data sample of size n drawn from a mixture of 2 subgaussian distributions in $\mathbb{R}^p$. We consider semidefinite programming relaxations of an integer quadratic program that is formulated essentially as finding the maximum cut on a graph, where edge weights in the cut represent dissimilarity scores between two nodes based on their p features. We define the signal-to-noise ratio (SNR) as $s^2 := \min\{np\gamma^2, \Delta^2\}$, where $\Delta^2 := p\gamma$ denotes the ℓ_2^2 distance between the two cluster centers. For both balanced and unbalanced cases, we allow each population to have distinct covariance structures Σ_1, Σ_2 with diagonal matrices as special cases. Our contributions are twofold. First, we provide a unified framework for analyzing three computationally efficient algorithms: SDP1, BalancedSDP, and Spectral clustering, yielding universal polynomial-rate misclassification guarantees for all three algorithms. Moreover, our framework accommodates a full range of tradeoffs between n, p, and γ, allowing for partial recovery (success rate $< 100\%$) when s^2 is lower bounded by a constant. Second, we prove that the misclassification errors for SDP1 and **BalancedSDP** decay exponentially with respect to the SNR s^2 and the **BalancedSDP** requires no explicit debiasing when the two clusters have equal sizes. To our knowledge, this is the first time such results are obtained for semidefinite relaxations of MAXCUT in population clustering. We provide simulation evidence illuminating the theoretical predictions.

Keywords: Concentration of measure, MAXCUT, mixture of subgaussian distributions, partial recovery, semidefinite programming relaxation, spectral clustering

1. Introduction

We explore a type of classification problem that arises in the context of computational biology. The biological context for this problem is that we are given DNA information from n individuals from k populations of origin and we wish to classify each individual into the correct population. DNA contains a series of markers called SNPs (Single Nucleotide Polymorphisms), each of which has two variants (alleles). We use bit 1 and bit 0 to denote them. The problem is that we are given a small sample of size n, e.g., DNA of n individuals (think of n in the hundreds or thousands), each described by the values of p *features* or *markers*, e.g., SNPs (think of p as an order of magnitude larger than n). Features have slightly different probabilities depending on which population the individual belongs to. For the rest of this section, we focus on $k = 2$. We defer the discussion of the more general

©2026 Shuheng Zhou.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-1670.html.

model with $k > 2$ in Sections 2 and 6. Denote by Δ^2 the ℓ_2^2 distance between two population centers (mean vectors), namely, $\mu^{(1)}, \mu^{(2)} \in \mathbb{R}^p$, where $\mu^{(i)} = (\mu_1^{(i)}, \ldots, \mu_p^{(i)})$, $i = 1, 2$,

$$\Delta^2 := \left\|\mu^{(1)} - \mu^{(2)}\right\|_2^2 = \sum_{k=1}^{p} (\mu_k^{(1)} - \mu_k^{(2)})^2 \quad \text{and} \quad \gamma := \Delta^2/p. \tag{1}$$

Our goal is to use these features to classify the individuals according to their population of origin. The objective we consider is to minimize the total data size $D = np$ needed to correctly classify the individuals in the sample as a function of the "average quality" $\gamma := \Delta^2/p$ of the features; cf. (1). Let $A = (a_{ij})$ be an $n \times n$ symmetric random matrix, where for $1 \le i, j \le n$, $a_{ij} = a_{ji} > 0$ denotes the edge weight, also known as *the dissimilarity score*, between nodes i and j, computed from the individuals' bit vectors.

For ease of discussion, let the edge weight equal the Hamming distance between the bit vectors of the two endpoints. Recall the weight of a cut is the sum of non-negative weights across all edges in the cut. It was previously shown that, in expectation, among all balanced cuts in the complete graph formed among n vertices (sample points), the cut of maximum weight corresponds to the correct partition of the n points according to their distributions in the balanced case with two populations ($n_1 = n_2 = n/2$); see, for example Zhou (2006). Under suitable conditions, the statement above also holds with high probability (w.h.p.) (Chaudhuri et al., 2007; Zhou, 2006).

In other words, in the context of population clustering, Chaudhuri et al. (2007) and Zhou (2006) show that one can use a random instance of the integer quadratic program:

$$(Q) \qquad \text{maximize} \quad x^T A x \quad \text{subject to} \quad x \in \{-1, 1\}^n \tag{2}$$

to identify the correct partition of nodes according to their population of origin w.h.p. in the balanced case, as long as the data size $D = np$ is sufficiently large and the separation metric satisfies $\Delta^2 = \Omega(\log n)$. The integer quadratic program (2) is NP-hard (Karp, 1972). In a groundbreaking paper, Goemans and Williamson (1995) show that one can use a semidefinite program (SDP) as a relaxation to solve (Q) approximately.

1.1 Estimators and convex relaxation

In the present setting, we study three semidefinite programs related to the graph cut problem (2) and provide a unified analysis; cf. Theorems 7, 10, 6 and 17. Formally, suppose we are given a data matrix $X \in \mathbb{R}^{n \times p}$ with samples from two populations $\mathcal{C}_1, \mathcal{C}_2$, where

$$\forall i \in \mathcal{C}_g, \quad \mathbb{E}(X_{ij}) = \mu_j^{(g)},\ j \in [p], \quad \text{where} \ \ \mu^{(g)} = (\mu_1^{(g)}, \ldots, \mu_p^{(g)}) \in \mathbb{R}^p, \forall g \in \{1, 2\}, \tag{3}$$

and the sizes of clusters $|\mathcal{C}_\ell| =: n_\ell, \forall \ell$ may not be the same. Here we denote by $[p]$ the set of integers $\{1, \ldots, p\}$. Our goal is to estimate the group membership vector

$$u_2 \in \{-1, 1\}^n \ \text{ such that } u_{2,j} = 1 \ \text{ for } \ j \in \mathcal{C}_1 \ \text{ and } \ u_{2,j} = -1 \ \text{ for } \ j \in \mathcal{C}_2. \tag{4}$$

Semidefinite relaxations. To estimate the group membership vector $u_2 \in \{-1, 1\}^n$, we consider the following semidefinite optimization problem: for a symmetric matrix A,

$$\text{(SDP)} \qquad \text{maximize} \ \ \langle A, Z \rangle \quad \text{subject to} \quad Z \succeq 0,\ I_n \succeq \operatorname{diag}(Z), \tag{5}$$

where I_n denotes the identity matrix and $A \in \mathbb{R}^{n\times n}$ will be specified in the sequel. Here $A \succeq B$ means that $A - B \succeq 0$ and the inner product of matrices $\langle A, B\rangle = \mathrm{tr}(A^T B)$. Hence $Z \succeq 0$ indicates that the matrix Z is constrained to be positive semidefinite. Let $\mathbf{1}_n^T = [1, \ldots, 1] \in \mathbb{R}^n$ denote a vector of all 1s. Let $E_n := \mathbf{1}_n \mathbf{1}_n^T$. As in many statistical problems, one simple but critical step is to first obtain the centered data Y. Loosely speaking, this procedure is called "global centering" in the statistical literature; see for example Hornstein et al. (2019) and references therein. In the present work, to find a convex relaxation, we will first construct a matrix Y by subtracting the sample mean $\widehat{\mu}_n := \frac{1}{n}\sum_{i=1}^n X_i \in \mathbb{R}^p$ from each row vector X_i of the data matrix X:

$$Y \;=\; X - P_1 X, \quad \text{and} \quad P_1 := \mathbf{1}_n \mathbf{1}_n^T/n = E_n/n \text{ is a projection matrix.} \tag{6}$$

Given matrix Y, with row vectors $Y_1, \ldots, Y_n \in \mathbb{R}^p$, where $Y_i = X_i - \widehat{\mu}_n$, we construct:

$$A \;:=\; YY^T - \lambda(E_n - I_n), \quad \text{where } \lambda = \frac{2}{n(n-1)}\sum_{i<j}\langle Y_i, Y_j\rangle. \tag{7}$$

More explicitly, (SDP) (5) with A as in (7) is equivalent to the optimization problem (SDP1),

$$\text{(SDP1)} \qquad \text{maximize } \langle YY^T - \lambda E_n, Z\rangle \quad \text{subject to} \quad Z \succeq 0, \mathrm{diag}(Z) = I_n. \tag{8}$$

See Proposition 13 for the equivalence of (SDP) and (SDP1). Since all possible solutions of (Q) are feasible for (SDP), the optimal value of (SDP) is an upper bound on the optimal value of (Q). We refer to Goemans and Williamson (1995) for earlier works in the design and analysis of approximation algorithms for the MAX CUT problem.

Variations for balanced partitions. For balanced partitions, we propose the following new estimator **BalancedSDP**, which is a variation of (SDP1):

$$\begin{aligned} \text{(BalancedSDP)}: \qquad & \text{maximize } \langle YY^T, Z\rangle \\ & \text{subject to } \; Z \succeq 0, \;\; I_n = \mathrm{diag}(Z), \;\; \text{and } \langle E_n, Z\rangle = 0. \end{aligned} \tag{9}$$

Here we add one more constraint that $\mathrm{tr}(E_n Z) = \mathbf{1}_n^T Z \mathbf{1}_n = 0$ and hence $\mathbf{1}_n$ is the eigenvector corresponding to the smallest eigenvalue $\lambda_{\min}(Z) = 0$ of $Z \succeq 0$. As we will show, centering the data matrix X plays a crucial role not only in the design and analysis of our convex optimization algorithms, but also in understanding the sample size lower bounds for partial recovery of the clusters for each algorithm.

1.2 Contributions

Our work fills an important gap in the formulation and analysis of the MAXCUT-based SDP (5) as well as in understanding its connection to spectral methods (Blum et al., 2009) and the k-means criterion in clustering. For two numbers a, b, $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$. In this paper, we propose using the matrix $\widetilde{A} := YY^T - \lambda E_n$ (8) (resp. the Gram matrix YY^T) instead of XX^T as the input to (SDP1) (resp. BalancedSDP) as convex relaxations of MAXCUT. We make the following contributions: (a) The global centering of the data matrix X enables sharp concentration of measure bounds to be derived in Theorems 8 and 6, directly leading to the polynomial-rate misclassification guarantees for

(SDP1) and spectral clustering in Theorems 7 and 17; (b) This innovative approach enables a simultaneous, transparent, and unified global and local analysis framework, resulting in an exponentially decaying error bound in s^2 for both aforementioned semidefinite programs in Theorems 10 and 11, respectively, where $s^2 \asymp (p\gamma) \wedge (np\gamma^2)$. This exponentially decaying error bound implies that when $s^2 = \Omega(\log n)$, perfect recovery of the cluster structure is achieved; (c) We demonstrate statistical convergence and computational feasibility even in the low-SNR (signal-to-noise ratio) regime when $s^2 = o(\log n)$; (d) The design and analysis of the semidefinite programs, namely, (SDP1) and (BalancedSDP), also shed light on the important issue of debiasing. In particular, the remarkably good solution to balanced MAX CUT as shown in Theorem 11 is obtained without any explicit debiasing.

To the best of our knowledge, this is the first time such results are obtained for semidefinite relaxations of MAX CUT in population clustering. Previously, the spectral algorithms in Blum et al. (2009) partitioned sample points based on the top k eigenvectors of the Gram matrix XX^T, following an idea that goes back at least to Fiedler (1973). In contrast, the spectral analysis in this paper uses the Gram matrix YY^T based on centered data, and directly improves upon the results in Blum et al. (2009) by removing the lower bound on n. Importantly, through this design choice of global centering of the data matrix, we not only overcome technical obstacles but are also able to make meaningful connections to the semidefinite relaxations for the k-means criterion by Peng and Wei (2007).

Organization. The rest of the paper is organized as follows. Section 2 provides background, describes our global and local theory, and presents preliminary results. In Section 3, we present our main theoretical results in Theorems 7, 10, 11, 16, and 17, respectively. We summarize the relevant related work in Section 3.6. In Section 4, we present a proof sketch for Theorems 10 and 11. In Section 5, we state a reduction principle for proving the new concentration bounds on $\left\|YY^T - \mathbb{E}YY^T\right\|_2$. In Section 6, we highlight the connections and key differences between our approach and convex relaxations based on the k-means criterion. In Section 7, we present numerical examples, where we compare three algorithms based on (SDP1), BalancedSDP, and spectral clustering, and show they have similar trends as predicted by the SNR. We prove Theorems 7, 16, 10, 11, 17, and their corollaries in Sections 3.3, 3.5, and 8 to 12, respectively. We conclude in Section 13. This paper combines results from two arXiv preprints Zhou (2023) and Zhou (2024b). Additional technical proofs are provided in the Appendix. For full proofs omitted from this manuscript, we direct the reader to the extended version of this paper Zhou (2026).

Notation. Let $\mathbf{B}_2^n$ and $\mathbb{S}^{n-1}$ be the unit Euclidean ball and the unit sphere in $\mathbb{R}^n$, respectively. Let $e_1, \ldots, e_n$ be the canonical basis of $\mathbb{R}^n$. For a set $J \subset [n]$, denote $E_J = \mathrm{span}\{e_j : j \in J\}$. For a vector $v \in \mathbb{R}^n$, we use v_J to denote the subvector $(v_j)_{j \in J}$. Let $\|v\|_1 := \sum_j |v_j|$ and $\|v\|_2 := (\sum_j v_j^2)^{1/2}$. For a symmetric matrix M, let $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ be the largest and smallest eigenvalues of M, respectively. The operator norm is defined as $\|M\|_2 = \sqrt{\lambda_{\max}(M^T M)}$. For a matrix $B = (b_{ij})$ of size $n \times n$, let $\mathrm{vec}\{B\}$ be formed by concatenating the columns of B into a vector of size n^2. Denote by $B_1^{n \times n} = \{(b_{ij}) \in \mathbb{R}^{n \times n} : \sum_{i,j} |b_{ij}| \le 1\}$ the unit ℓ_1-ball over $\mathrm{vec}\{B\}$. Let $\|B\|_1 = \sum_{i,j=1}^n |b_{ij}|$ and $\|B\|_F = (\sum_{i,j} b_{ij}^2)^{1/2}$. Let $\|A\|_{\max} = \max_{i,j} |a_{ij}|$. Let $\mathrm{diag}(A)$ and $\mathrm{offd}(A)$ be the diagonal and off-diagonal parts of matrix A, respectively. We write $a \asymp b$ if $ca \le b \le Ca$ for some positive absolute constants c, C that are independent of n, p, and γ. We write $f = O(h)$ or $f \ll h$ if $|f| \le Ch$ for some absolute constant $C < \infty$, and $f = \Omega(h)$ or $f \gg h$ if $h = O(f)$.

We write $f = o(h)$ if $f/h \to 0$ as $n \to \infty$. In this paper, $C, C_1, C_2, C_3, C_4, c, c', c_1, \ldots$ denote various absolute positive constants that may change line by line.

2. Approach and background

Our work is motivated by two threads of research in combinatorial optimization and in community detection to reformulate the MAX CUT problem (2) and its convex relaxations (SDP1 and BalancedSDP) in clustering. The present work is inspired by the original combinatorial analysis in Zhou (2006) and Chaudhuri et al. (2007), as well as the novel ideas in Guedon and Vershynin (2016) to revisit the MAXCUT-based graph partition problem (2). While Chaudhuri et al. (2007) and Zhou (2006) originally proposed the integer quadratic program (2) in the context of population clustering, their focus was on the formulation of (Q) with various choices of A rather than using computationally efficient algorithms such as the SDP relaxation (5) to solve (2). The global analysis framework for the semidefinite relaxation (Guedon and Vershynin, 2016) was set in the context of community detection in sparse networks, where A represents the adjacency matrix of a random graph.

Guedon and Vershynin (2016) study the semidefinite relaxation of the integer program (2), where an $n \times n$ random matrix A (observed) is used to replace the hidden static R in the original integer quadratic problem (10) such that $\mathbb{E}A = R$:

$$\text{maximize } x^T R x \quad \text{subject to } x \in \{-1, 1\}^n. \tag{10}$$

The proof strategy of Guedon and Vershynin (2016) is to apply Grothendieck's inequality to the random error $A - \mathbb{E}A$ rather than to the original matrix A as considered in earlier literature: bounding the cut norm of $A - \mathbb{E}A$ plays an important role in deriving an error bound for $\widehat{Z} - u_2 u_2^T$ in the Frobenius norm, where $\widehat{Z}$ is an optimal solution to (SDP) (5); cf. (4). We refer to this approach as the *global analysis*, following Chrétien et al. (2021).

2.1 Roadmap

In the present work, we focus on the regime where partial recovery (rather than exact recovery) is possible. We first develop a global theory by centering the data and constructing an ideal reference matrix, which via Grothendieck's inequality and Davis–Kahan bounds yields universal polynomial-rate misclassification guarantees for three algorithms (SDP1, BalancedSDP, and spectral clustering) under minimal signal conditions.

The important distinction between the present work and that considered by Guedon and Vershynin (2016) is that to estimate the group membership vector u_2 (4), we replace the random adjacency matrix arising from stochastic block models with an instance of a symmetric matrix A (7) computed from the centered data matrix Y as in (6). However, unlike the setting in Guedon and Vershynin (2016), $\mathbb{E}A \neq R$, resulting in a bias that needs to be corrected. Moreover, with the new concentration of measure bound on $\left\|YY^T - \mathbb{E}YY^T\right\|_2$ in Theorem 6, we can simultaneously analyze (SDP1) in Theorem 7 as well as the spectral algorithm in Theorem 17. We are also able to explain the theoretical underpinning of the observed concentration of measure phenomena in high dimensions (Zhou, 2006; Blum et al., 2009), simultaneously for the spectral method and the semidefinite optimization goals, using global and local analyses, where the input is now based on the Gram matrix YY^T

computed from centered data. With local analyses, we prove an exponentially decaying error in Frobenius norm for $\widehat{Z} - u_2 u_2^T$, where $\widehat{Z}$ is an optimal solution to (SDP1) (resp. BalancedSDP) with respect to the signal-to-noise ratio (SNR) in Theorem 10 (resp. 11). We first present the technical ingredients and important ideas for proving these results, while highlighting connections to and key distinctions from existing work. We will briefly introduce the k-means criterion, where $k \geq 2$, in Section 2.4. We defer further discussion to Sections 3 to 6.

2.2 Technical ingredients

We now elaborate on the construction of a reference matrix and the tools for deriving the large deviation bound on $\left\|YY^T - \mathbb{E}YY^T\right\|_2$, while deferring the discussion of bias and local analyses to Sections 2.3, 3.2, 3.5, 4, and 6. Denote by $w_{\min} := \min_{j=1,2} w_j$ and $w_{\max} := \max_{j=1,2} w_j$. We use $n_{\min} := nw_{\min}$ and $n_{\max} := nw_{\max}$ to represent the size of the smallest and the largest cluster, respectively. We consider $w_{\min}$ *as a constant* throughout this work. Let $\mathcal{M}_G^+ = \{Z : Z \succeq 0, I_n \succeq \mathrm{diag}(Z)\}$ and

$$\mathcal{M}_{\mathrm{opt}} \quad = \quad \{Z : Z \succeq 0, \mathrm{diag}(Z) = I_n\} \subset \mathcal{M}_G^+ \subset [-1,1]^{n\times n}. \tag{11}$$

Construction of a reference matrix. To take this global approach, upon centering, we crucially construct a static reference matrix $R := \mathbb{E}(Y)\mathbb{E}(Y)^T$ with the desired properties. Then, $Z^* = u_2 u_2^T \in \mathcal{M}_{\mathrm{opt}}$, for the convex set $\mathcal{M}_{\mathrm{opt}}$ as in (11) will maximize the reference objective function $\langle R, Z\rangle$ among all $Z \in [-1,1]^{n\times n}$, and naturally among all $Z \in \mathcal{M}_{\mathrm{opt}}$; cf. Lemma 1. Recall the overall goal of convex relaxation is to: (a) estimate the solution of the integer quadratic problem (10) with an appropriately chosen reference matrix R such that solving (10) will recover the cluster *exactly*; and (b) choose the convex set $\mathcal{M}_G^+$ (resp. $\mathcal{M}_{\mathrm{opt}}$ (11)) so that the semidefinite relaxation of the static problem (10) is tight. This means that when we replace A (resp. $\widetilde{A} := YY^T - \lambda E_n$) with R in (SDP) (5) (resp. SDP1 (8)), we obtain a solution $Z^* = u_2 u_2^T$ for the group membership vector $u_2 \in \{-1,1\}^n$ in (4), which can then be used to recover the clusters exactly.

Lemma 1 (The reference matrix). *W.l.o.g., suppose the first n_1 rows of X belong to $\mathcal{C}_1$ and the rest belong to $\mathcal{C}_2$. Let $w_1 := n_1/n$ and $w_2 := n_2/n$. We construct*

$$R \quad := \quad \mathbb{E}(Y)\mathbb{E}(Y)^T = p\gamma \begin{bmatrix} w_2^2 \mathbf{1}_{n_1}\otimes \mathbf{1}_{n_1} & -w_1 w_2 \mathbf{1}_{n_1}\otimes \mathbf{1}_{n_2} \\ -w_1 w_2 \mathbf{1}_{n_2}\otimes \mathbf{1}_{n_1} & w_1^2 \mathbf{1}_{n_2}\otimes \mathbf{1}_{n_2} \end{bmatrix}, \tag{12}$$

for Y as defined in (6). *Then $Z^* = \mathrm{argmax}_{Z\in\mathcal{M}_{opt}}\langle R, Z\rangle = u_2 u_2^T$ for $\mathcal{M}_{opt}$ as in* (11).

Anisotropic subgaussian random vectors as noise. First, recall for a random variable X, the ψ_2-norm of X, denoted by $\|X\|_{\psi_2}$, is $\|X\|_{\psi_2} = \inf\{t > 0 \,:\, \mathbb{E}\exp(X^2/t^2) \leq 2\}$.

Definition 2. *A random vector $W \in \mathbb{R}^m$ is called subgaussian if the one-dimensional marginals $\langle W, h\rangle$ are subgaussian random variables for all $h \in \mathbb{R}^m$: (1) W is called isotropic if for every $h \in \mathbb{S}^{m-1}$, $\mathbb{E}\left|\langle W, h\rangle\right|^2 = 1$; (2) W is ψ_2 with a constant α if for every $h \in \mathbb{S}^{m-1}$, $\left\|\langle W, h\rangle\right\|_{\psi_2} \leq \alpha$. The subgaussian norm of $W \in \mathbb{R}^m$ is denoted by*

$$\|W\|_{\psi_2} := \sup_{h\in\mathbb{S}^{m-1}} \left\|\langle W, h\rangle\right\|_{\psi_2}. \tag{13}$$

Definition 2 is standard; see for example Definition 1.1 Rudelson and Zhou (2017) and Definition 3.4.1 Vershynin (2026). Denote by $\mathbb{Z} = X - \mathbb{E}X = [\mathbb{Z}_1, \ldots, \mathbb{Z}_n]^T$ the noise matrix with independent, mean-zero, subgaussian random vectors $\mathbb{Z}_1, \ldots, \mathbb{Z}_n \in \mathbb{R}^p$. We now state the data generative process for $\mathbb{Z} \in \mathbb{R}^{n\times p}$ in Definition 3 that we use throughout this paper. Crucially, we prove the concentration bound in Theorem 4, which corresponds to the Hanson-Wright inequalities (Hanson and Wright, 1971; Rudelson and Vershynin, 2013), but for anisotropic subgaussian vectors.

Definition 3 (**Data generative process**). *Let $H_1, \ldots, H_n$ be deterministic $p\times m$ matrices, where we assume that $m \geq p$ for simplicity; let $\sigma^2_{\max} = \max_i \|H_i\|_2^2$. Suppose random matrix $\mathbb{W} = [W_1, \ldots, W_n]^T = (w_{jk}) \in \mathbb{R}^{n\times m}$, where $W_1, \ldots, W_n \in \mathbb{R}^m$ are independent, isotropic, subgaussian random vectors, each with independent, mean-zero coordinates such that $\mathsf{Cov}(W_j) := \mathbb{E}W_j W_j^T = I_m, \forall j \in [n]$ and $\mathbb{E}[w_{jk}] = 0, \forall j, k$. Denote by*

$$K := \max_{j,k} \|w_{jk}\|_{\psi_2} \quad \text{and} \quad C_0 := \max_j \|W_j\|_{\psi_2}, \quad \text{where} \quad K \leq C_0 \leq CK \tag{14}$$

for an absolute constant C; see Lemma 3.4.2 (Vershynin, 2026). Suppose

$$\forall j \in [n], \quad \mathbb{Z}_j = H_j W_j, \quad \text{where} \quad H_j \in \mathbb{R}^{p\times m} \quad \text{and} \quad 0 < \|H_j\|_2 \leq \sigma_{\max} < \infty.$$

Theorem 4 (**Hanson-Wright inequality for anisotropic subgaussian vectors**). *Let $H_1, \ldots, H_n$ be deterministic $p \times m$ matrices, where $m \geq p$. Let $\mathbb{Z}_1^T, \ldots, \mathbb{Z}_n^T \in \mathbb{R}^p$ be row vectors of $\mathbb{Z}$. We generate $\mathbb{Z} = [\mathbb{Z}_1, \ldots, \mathbb{Z}_n]^T \in \mathbb{R}^{n\times p}$ according to Definition 3. Let K and C_0 be as defined in (14). Then we have for any $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ and $t > 0$,*

$$\begin{aligned} \mathbb{P}\big(|\sum_{i=1}^n \sum_{j\neq i}^n \langle \mathbb{Z}_i, \mathbb{Z}_j\rangle a_{ij}| > t\big) &\leq 2\exp\big(-c\min\big(\frac{t^2}{p(K\sigma_{\max})^4\|A\|_F^2}, \frac{t}{(K\sigma_{\max})^2\|A\|_2}\big)\big), \\ \text{where } \forall j \in [n],\ \|\mathbb{Z}_j\|_{\psi_2} &:= \sup_{v\in\mathbb{S}^{p-1}} \|\langle \mathbb{Z}_j, v\rangle\|_{\psi_2} \leq \|W_j\|_{\psi_2} \|H_j\|_2 \leq C_0\sigma_{\max}. \end{aligned} \tag{15}$$

Theorem 4 and its proof may be of independent interest; cf. Section A. See Exercise 6.8 by Vershynin (2026) for a related result. In the context of Theorem 4, H_j's are allowed to repeat, e.g., across rows from the same cluster $\mathcal{C}_i$ for $i = 1, 2$. Lemma 5 characterizes the two-group design matrix covariance structures considered in Theorems 6, 7, 10, and 11, using the data generative process in Definition 3. We prove Lemma 5 in Section A.1.

Lemma 5 (**Two-group subgaussian mixture model**). *Denote by X the two-group design matrix considered in (3). Let random matrix $\mathbb{W} = [W_1, \ldots, W_n]^T = (w_{jk}) \in \mathbb{R}^{n\times m}$ and C_0 be the same as in Definition 3 with $CK \geq C_0 \geq K$ where $K := \max_{j,k} \|w_{jk}\|_{\psi_2}$ and C is an absolute constant. Let $H_i \in \mathbb{R}^{p\times m}, i = 1, 2$ be deterministic matrices with $0 < \|H_i\|_2 \leq \sigma_{\max} < \infty, \forall i$. Suppose for $j \in \mathcal{C}_i \subset [n]$, $\mathbb{E}X_j = \mu^{(i)}$ and $\mathbb{Z}_j := X_j - \mathbb{E}X_j = H_i W_j \in \mathbb{R}^p$, where $i = 1, 2$. Then for both $i \in \{1, 2\}$,*

$$\forall j \in \mathcal{C}_i, \quad \mathsf{Cov}(\mathbb{Z}_j) := \mathbb{E}\mathbb{Z}_j\mathbb{Z}_j^T = H_i H_i^T =: \Sigma_i \quad \text{where} \quad \operatorname{tr}(\Sigma_i) =: V_i = \|H_i\|_F^2.$$

Moreover, $\mathbb{Z}_1, \ldots, \mathbb{Z}_n \in \mathbb{R}^p$ are independent, mean zero, subgaussian random vectors with covariance $\mathsf{Cov}(\mathbb{Z}_j)$ satisfying (16) and (17):

$$\forall j \in [n],\ \|\langle \mathbb{Z}_j, x\rangle\|_{\psi_2} \leq C_0 \|\langle \mathbb{Z}_j, x\rangle\|_{L_2} \text{ for any } x \in \mathbb{R}^p, \text{ where} \tag{16}$$

$$\|\langle \mathbb{Z}_j, x\rangle\|_{L_2}^2 := \mathbb{E}\langle Z_j, x\rangle^2 = x^T(\mathbb{E}\mathbb{Z}_j\mathbb{Z}_j^T)x = x^T\mathsf{Cov}(\mathbb{Z}_j)x. \tag{17}$$

2.3 Concentration of measure, local analysis and debiasing

Theorem 6 is an important result in this paper. The concentration bounds on the operator norm of $YY^T - \mathbb{E}YY^T$ imply that, up to a constant factor, the same bounds also hold for $A - \mathbb{E}A$ (and $B - \mathbb{E}B$, cf. (30)). As a result, the error bound in Theorem 6 is crucial in the global analysis leading to Theorem 7, as well as our analysis of a spectral clustering algorithm in Theorem 17.

Theorem 6 (**Concentration for anisotropic design).** *Let X and C_0 be as in Lemma 5 and Y be as in* (6)*. Let $\Delta^2 = \left\|\mu^{(1)} - \mu^{(2)}\right\|_2^2 = p\gamma$. Let*

$$\mu := (\mu^{(1)} - \mu^{(2)})/\sqrt{p\gamma} \in \mathbb{S}^{p-1}. \tag{18}$$

Suppose for some parameter ξ satisfying $w_{\min}^2/16 > \xi = \Omega(1/n_{\min})$ and $\sigma_{\max}^2 := \max_i \|H_i\|_2^2$,

$$p\gamma \;\geq\; C_2 C_0^2 \sigma_{\max}^2/\xi^2 \quad \text{and} \quad np \geq C_3 C_0^4 \sigma_{\max}^4/(\gamma^2\xi^2). \tag{19}$$

Then with probability at least $1 - 2\exp(-c_8 n)$, we have

$$\begin{aligned} \left\|YY^T - \mathbb{E}YY^T\right\|_2 &\leq C_4 C_0 (\max_i \left\|H_i^T \mu\right\|_2) n\sqrt{p\gamma} + C_5 C_0^2 \sigma_{\max}^2 (\sqrt{pn} \vee n) \\ &\leq \xi np\gamma/6. \end{aligned} \tag{20}$$

We mention in passing that a natural and common choice for A in (SDP) in the literature is the Gram matrix XX^T; however, the large deviation bounds on $\left\|XX^T - \mathbb{E}XX^T\right\|_2$ will not be sufficient to apply the global analysis framework as in Guedon and Vershynin (2016). We give a proof sketch of Theorem 6 in Section 5, with full proof in Section C.

Local analysis. In the context of Lemma 5, the notion of SNR introduced in Section 1.2 can now be properly adjusted in view of (16) and (17):

$$\textbf{SNR anisotropic:} \quad s^2 \;=\; \frac{\Delta^2}{C_0^2\sigma_{\max}^2} \wedge \frac{np\gamma^2}{C_0^4\sigma_{\max}^4}, \text{ where } \sigma_{\max}^2 := \max_j \|\mathsf{Cov}(\mathbb{Z}_j)\|_2. \tag{21}$$

Crucially, we also obtain a uniform control for $\langle YY^T - \mathbb{E}YY^T, \widehat{Z} - Z^*\rangle$ over all $\widehat{Z} \in \mathcal{M}_{\text{opt}}$ in the local neighborhood of $Z^* = u_2 u_2^T$ with high probability; cf. Lemma 22. As a result, the misclassification errors are shown to decay exponentially in the SNR parameter s^2 (21) in the settings of Theorems 10, 11, and Corollary 19 for the subgaussian design with dependent features considered in Lemma 5, as long as s^2 is bounded below by a constant as in (19). This gives rise to the notion of a *local analysis*; we refer to Chrétien et al. (2021) and Sections 4, 8, 11, and 12 for further details and references.

Debiasing. Consider the model in Lemma 5. For both balanced and unbalanced cases, we allow each population to have distinct covariance structures with diagonal matrices as special cases. Let $|\mathcal{C}_1| = n_1$ and $|\mathcal{C}_2| = n_2$. Suppose $k = 2$. Let $\mathbb{Z}_j = X_j - \mathbb{E}X_j \in \mathbb{R}^p$, $j \in [n]$. Let V_i denote the trace of the covariance $\mathsf{Cov}(\mathbb{Z}_j) = \mathbb{E}\mathbb{Z}_j\mathbb{Z}_j^T$ for all $j \in \mathcal{C}_i$:

$$V_1 = \mathbb{E}\langle \mathbb{Z}_j, \mathbb{Z}_j\rangle \quad \forall j \in \mathcal{C}_1 \;\text{ and }\; V_2 = \mathbb{E}\langle \mathbb{Z}_j, \mathbb{Z}_j\rangle \quad \forall j \in \mathcal{C}_2. \tag{22}$$

Such a discrepancy in their traces is known to cause a bias, which needs to be explicitly addressed; cf. (57). See, for example Royer (2017). Through the refined *local analysis* in

the present work, we conjecture that for the balanced case ($n_1 = n_2 = n/2$), the bias may have already been *substantially reduced* due to our centering and adjustment steps in the construction of A (7) for (SDP) and (SDP1). This insight leads to the design of a new estimator, namely, BalancedSDP (9). As such, we *eliminate* the bounded trace-difference assumption **(A2)** in Theorem 11 despite not making any explicit adjustment to the Gram matrix YY^T. In this sense, the results obtained in Theorem 11 for balanced partitions are stronger and more robust than those in Giraud and Verzelen (2019) for $k = 2$. See the discussion following Theorem 11. See Sections 6, 11, and 12 for full exposition. We defer the discussion of SNR to Section 3.1 after stating Theorem 10.

2.4 Generalization to the k-means criterion

More generally, one may consider semidefinite relaxations for the following subgaussian mixture model with k centers, where we have n observations $X_1, X_2, \ldots, X_n \in \mathbb{R}^p$:

$$X_i = \mu^{(\psi_i)} + \mathbb{Z}_i, \quad i \in [n], \tag{23}$$

where $\mathbb{Z}_1, \ldots, \mathbb{Z}_n \in \mathbb{R}^p$ are independent, subgaussian, mean-zero random vectors, and $\psi_i \in \{1, \ldots, k\}$ assigns node i to a group $\mathcal{C}_j$ with mean $\mu^{(j)} \in \mathbb{R}^p$ for some $j \in [k]$.

Giraud and Verzelen (2019) establish bounds for a semidefinite program based on the k-means criterion, following the proposal by Peng and Wei (2007), where the positive semidefinite matrix $Z \in \mathbb{R}^{n\times n}$ is constrained to have non-negative entries. Specifically, they consider $\widehat{Z} = \operatorname{argmax}_{Z\in\mathcal{M}_k} \langle XX^T, Z\rangle$, where $\mathcal{M}_k = \{Z \in \mathbb{R}^{n\times n} : Z \succeq 0, \operatorname{tr}(Z) = k, Z\mathbf{1}_n = \mathbf{1}_n, Z_{ij} \geq 0, \forall i, j\}$. Enforcing these non-negativity constraints ($Z_{ij} \geq 0, \forall i, j$) in $\mathcal{M}_k$ regularizes the convex problem to resemble the k-means criterion, but it nonetheless incurs an additional computational cost. In Section 6, we elaborate on k-means relaxations, including SVD-based algorithms that allow for reduced computational complexity by eliminating both the non-negativity and row-sum ($Z\mathbf{1}_n = \mathbf{1}_n$) constraints. We analyze this algorithm for $k = 2$ in Theorem 17 in the present work.

While the theoretical results in Theorem 10 are developed in the same spirit as those of Giraud and Verzelen (2019) for $k = 2$, we prove them for the much simpler MAXCUT-based SDP (5) with the constraint set $\{Z_{ii} \leq 1, \forall i \in [n]\}$, and for (SDP1) with the minimal constraint set $\{Z_{ii} = 1, \forall i \in [n]\}$, both under the condition $Z \succeq 0$. Consequently, we reduce the problem from $O(n^2)$ linear constraints in $\mathcal{M}_k$ to exactly n linear constraints in $\mathcal{M}_G^+$ or $\mathcal{M}_{\text{opt}}$, as defined in (11). The objective of dropping the $O(n^2)$ non-negativity constraints on the elements of Z in $\mathcal{M}_k$ introduced by Peng and Wei (2007) is to reduce the computational complexity *without compromising statistical convergence*. Our theory confirms that the MAXCUT-based semidefinite relaxations, such as (SDP1) (8) and BalancedSDP (9), as well as the spectral clustering algorithm, provide a mathematically rigorous framework with *substantially reduced constraint complexity*, while ensuring statistical convergence.

We emphasize that the reduction techniques used for the concentration of measure bound in Theorem 6, as well as the probabilistic bounds in Theorem 4 (and Lemmas 26 and 42), are derived for the general k-means clustering problem where $\mathbb{Z}_1, \ldots, \mathbb{Z}_n$ are independent but not identically distributed. In particular, the spectral clustering algorithm based on the top $k-1$ eigenvectors of YY^T is directly related to the relaxations for the k-means criterion proposed by Peng and Wei (2007). Consequently, these concentration bounds are applicable to the general k-means model (23), allowing for immediate extensions of our framework.

3. Theory

To set up the context for our analysis, we first state in Section 3.1 the main Theorem 7, under assumptions (A1) and (A2), motivating the consideration of (28) as our baseline scenario. We then state Theorems 10 and 11, followed by a discussion. We construct an oracle SDP to enable the global and local analyses in Section 3.2. We prove Theorem 7 in Section 3.3. In Section 3.4, we state Theorem 17 for the SVD-based algorithm, along with its corollary. In Section 3.5, we provide a proof sketch for Theorem 16, which is an important intermediate result unifying Theorems 7 and 17. We discuss the relevant related work in Section 3.6. Let $C, C', C_1, C_2, c_0, c, c', c_1, \ldots$ be absolute constants.

3.1 Main theorems from the global and local analyses

We first state assumptions (A1) and (A2).
(A1) Let $\mathbb{Z} = X - \mathbb{E}X$. Let $\mathbb{Z}_i = X_i - \mathbb{E}X_i, i \in [n]$ be independent, mean-zero, subgaussian random vectors with independent coordinates such that for all i, j, $\|X_{ij} - \mathbb{E}X_{ij}\|_{\psi_2} \le C_0$.
(A2) The two distributions have variance profile discrepancy bounded:

$$|V_1 - V_2| \le \xi np\gamma/3 \text{ for } V_i \text{ as in (22)}, \quad \text{where} \quad 1/2 > \xi = \Omega(1/n_{\min}). \tag{24}$$

As mentioned, Theorem 8 is an important intermediate result for proving Theorem 7. The anisotropic case is stated in Theorem 6.

Theorem 7. *Suppose for $j \in \mathcal{C}_i$, $\mathbb{E}X_j = \mu^{(i)}$, where $i = 1, 2$. Let $\widehat{Z}$ be a solution of (SDP1). Let $K_G \le \frac{\pi}{2\ln(1+\sqrt{2})} \le 1.783$. Suppose (A1) and (A2) hold, and*

$$p\gamma = \Delta^2 \ge C'C_0^2/\xi^2 \quad \text{and} \quad np \ge CC_0^4/(\xi^2\gamma^2), \tag{25}$$

for some absolute constants C, C', where ξ is the same as in (24). Then with probability at least $1 - 2\exp(-cn)$, for u_2 as in (4),

$$\|\widehat{Z} - u_2u_2^T\|_F^2 \le 2\|\widehat{Z} - u_2u_2^T\|_1 \le 2\delta n^2 \quad \text{where } \delta = 2K_G\xi/w_{\min}^2. \tag{26}$$

Theorem 8 (Design with independent entries (Zhou, 2023))**.** *Suppose that (A1) and (25) hold with $\max_{j,k}\|X_{jk} - \mathbb{E}X_{jk}\|_{\psi_2} \le C_0$. Then, with probability at least $1 - 2\exp(-c_7n)$,*

$$\left\|YY^T - \mathbb{E}YY^T\right\|_2 \quad \le \quad C_2C_0^2(\sqrt{pn} \vee n) + C_3C_0n\sqrt{p\gamma} \le \frac{1}{6}\xi np\gamma.$$

Remarks on the misclassification error rate. The parameter ξ in (A2) is understood to be set as a fraction of the misclassification error rate δ in Theorem 7, which in turn is assumed to be lower bounded: $1 > \delta = \Omega(1/n)$, since we focus on the partial recovery of the group membership vector u_2. To control the misclassification error, the parameters (δ, ξ) in Theorem 7 can be chosen to satisfy the following relations:

$$\delta \quad = \quad 2K_G\xi/w_{\min}^2 \asymp 2K_G/(w_{\min}^2\sqrt{s^2}) \quad \text{and} \quad \xi^2 \asymp 1/s^2, \quad \text{where} \tag{27}$$

$$s^2 \quad := \quad (p\gamma/C_0^2) \wedge (np\gamma^2/C_0^4) \quad (\textbf{SNR isotropic}) \tag{28}$$

in view of (25) and (26). Roughly speaking, ξ^2 in (27) is chosen to be inversely proportional to s^2, resulting in the misclassification error rate $\delta \asymp \xi/w_{\min}^2$ being inversely proportional to the square root of s^2, as we show in Corollary 9.

Corollary 9 ($O(\xi n/w_{\min}^2)$ misclassified vertices)**.** *Let $\widehat{x}$ denote the eigenvector of $\widehat{Z}$ corresponding to the largest eigenvalue, with $\|\widehat{x}\|_2 = \sqrt{n}$. Then in both settings of Theorem 7, we have with probability at least $1 - 2\exp(-cn)$, $\min_{\alpha=\pm 1} \|\alpha\widehat{x} - u_2\|_2^2 \le \delta' n = \delta' \|u_2\|_2^2$, where $\delta' \le 32K_G\xi/w_{\min}^2$. Moreover, the signs of the coefficients of $\widehat{x}$ correctly estimate the partition of the vertices into two clusters, up to at most $\delta' n$ misclassified vertices.*

While (A1) assumes that the random matrix $\mathbb{Z}$ has independent subgaussian entries, matching the separation condition (25) and SNR (28), the conclusions stated in Theorems 7 and 10 hold for the general two-group model considered in Lemma 5, upon adjusting (25). We emphasize that the covariance model under consideration in Theorem 6 is a special case of Theorem 4, cf. Lemma 5, whereas Theorem 8 is a special case of Theorem 6 for an isotropic design matrix under (A1). See Section 5 for a discussion on *covariance estimation*.
Local analysis. Our local analysis builds upon the global analysis; proving such results without the positivity constraints presents greater technical difficulties.

Theorem 10 (**SDP1**)**.** *Let X be as in Lemma 5. Let $\widehat{Z}$ be a solution of (SDP1). Suppose* (19) *holds. Suppose (A2) holds, where the parameter ξ in* (24) *is chosen to be the same as that in* (19)*. Then with probability at least $1 - 2\exp(-c_1 n) - c_2/n^2$, $\left\|\widehat{Z} - u_2 u_2^T\right\|_1 /n^2 \le \exp\big(-c_0 s^2 w_{\min}^4\big)$, where s^2 is as defined in* (21)*.*

Signal-to-noise ratios. Let $\sigma_{\max}^2 := \max_j \|\mathsf{Cov}(\mathbb{Z}_j)\|_2$. The notion of SNR (28) and the associated sample lower bounds appear in Chaudhuri et al. (2007) and Zhou (2006), where $s^2 = \Omega(\log n)$ is required. Our sample size lower bounds (19) and (25) indicate that once the distance between the two centers is bounded below by a constant, namely, $\Delta = \sqrt{p\gamma} = \Omega(\sigma_{\max})$, adding more sample points (individuals) to the clusters reduces the number of features required for partial recovery. This simultaneously mitigates the bias in the sense that **(A2)** holds trivially in the large sample setting where $n > p$, since $|V_1 - V_2| = O(p)$ by Definition 3. See Section 3.5 for further discussion (cf. (42)).

On the other hand, both the previous and current results show that even when sample size n is small, by increasing the dimension p such that the total sample size satisfies (25) or (19), we ensure partial recovery of cluster structures using either the integer quadratic program (10) or its convex relaxation such as (SDP) (5). We mention in passing that this definition of SNR in (28) also coincides with the signal-to-noise ratio used in Giraud and Verzelen (2019) for isotropic design under **(A1)**. Moreover, under **(A1)** and **(A2)**, the exponentially decaying error bounds in both papers are equivalent, despite the use of different SDP relaxations. We next state our main result for balanced partitions.

Theorem 11 (**BalancedSDP**)**.** *Let $X, \mathcal{C}_j \subset [n]$ be as in Lemma 5, with $|\mathcal{C}_1| = |\mathcal{C}_2| = n/2$. Let $\widehat{Z}$ be a solution of* (9)*. Let s^2 be as in* (21)*. Suppose $s^2 > C$ for some absolute constant C. Then with probability at least $1 - c/n^2$, $\left\|\widehat{Z} - u_2 u_2^T\right\|_1 /n^2 \le \exp(-c' s^2)$.*

As mentioned earlier, for balanced partitions, **(A2)** is not required for an exponentially decaying error bound to hold, provided that one additional constraint $\langle E_n, Z\rangle = 0$ is added in the BalancedSDP (9). In this sense, the results that we obtain in Theorem 11 for balanced partitions are stronger and more robust than those in Giraud and Verzelen (2019). In summary, (a) for the balanced case, when $n_1 = n_2$, an arbitrary level of trace difference

$|V_1 - V_2|$ is tolerated through the new estimator BalancedSDP (9). Hence, unlike Royer (2017), Giraud and Verzelen (2019), and Ndaoud (2022), we do not need any explicit debiasing step for balanced partitions; (b) for unbalanced partitions, we impose the bounded trace-difference assumption (A2) in Theorem 10; cf. (24). See Table 1 for conditions for each theorem. When **(A2)** is violated and $n_1 \neq n_2$, we may adopt ideas from Royer (2017), Giraud and Verzelen (2019), and the present work to make adjustments to (SDP1) to further correct the bias, but it comes with an additional computational cost nonetheless. We defer a detailed comparison with Giraud and Verzelen (2019) for anisotropic design, as well as the distinctive SDP relaxations, to Section 6.1. We provide a proof sketch for Theorems 10 and 11 in Section 4, with full proofs of these theorems and their corollaries in Sections 8, 11, 12, and 9, respectively.

Table 1: Conditions of the main theorems

Theorem	**Applicable to isotropic / anisotropic**	**Cluster sizes** (n_1, n_2)	**Needs Trace difference** $\lvert V_1 - V_2\rvert$ **bound in (A2)**	**Lower bounds on SNR** (28)/ (21)
Theorem 6	Yes/Yes	$n_1 \neq n_2$	No	(25) / (19)
Theorem 16	Yes/Yes	$n_1 \neq n_2$	Yes	(25) / (19)
Theorem 7	Yes/Yes	$n_1 \neq n_2$	Yes	(25) / (19)
Theorem 10	Yes/Yes	$n_1 \neq n_2$	Yes	(25) / (19)
Theorem 11	Yes/Yes	$n_1 = n_2$	No	$s^2 > C$
Theorem 17	Yes/Yes	$n_1 \neq n_2$	Yes	(25) / (19)

3.2 Construction of an OracleSDP

A straightforward calculation leads to the expression of the reference matrix $R := \mathbb{E}(Y)\mathbb{E}(Y)^T$ as in (12), in view of (48). However, unlike the setting in Guedon and Vershynin (2016), $\mathbb{E}A \neq R$, resulting in a large bias. This motivates the construction of OracleSDP (29):

$$(\textbf{OracleSDP}) \quad \text{maximize} \quad \langle B, Z\rangle \quad \text{subject to} \quad Z \in \mathcal{M}_{\text{opt}} \text{ as in (11)}, \tag{29}$$

$$\text{where } B := A - \mathbb{E}\tau I_n \text{ for } A \text{ as in (7) and } \tau = \sum_{i=1}^{n} \langle Y_i, Y_i\rangle/n. \tag{30}$$

Here, the adjustment term $\mathbb{E}\tau I_n$ plays no role in optimization, since the extra trace term $\propto \langle I_n, Z\rangle = \text{tr}(Z) = n$ is a constant function of Z on the feasible set $\mathcal{M}_{\text{opt}}$ (11). However, both A (7) and B (30) are defined to bridge the gap between YY^T and $R = \mathbb{E}(Y)\mathbb{E}(Y)^T$ (12). Lemma 12 states that the bias $\mathbb{E}B - R$ is substantially reduced for B as in (30), thanks to the adjustment term $\mathbb{E}\tau I_n$, and is essentially negligible when $V_1 = V_2$.

Lemma 12. *Let X be as in Lemma 5. Suppose (A2) (24) holds. Suppose $\xi \geq \frac{1}{2n}(4 \vee \frac{1}{w_{\min}})$ and $n \geq 4$. Then $\|\mathbb{E}B - R\|_2 \leq 2\,|V_1 - V_2|\,(w_1 \vee w_2) + p\gamma/3 \leq 2\xi np\gamma/3$. When $V_1 = V_2$, we have $\|\mathbb{E}B - R\|_2 \leq p\gamma/3$.*

We prove Lemma 12 in Section D.2. We emphasize that our algorithm solves SDP1 (8) rather than OracleSDP (29). However, formulating the OracleSDP estimator enables both global and local analyses by controlling the operator norm of the bias term, namely, $\|\mathbb{E}B - R\|_2$ as in Lemma 12, and a related quantity to be introduced in Lemma 23. Given the OracleSDP formulation in (29), the bias analysis and the concentration bounds on $\left\|YY^T - \mathbb{E}YY^T\right\|_2$ enable simultaneous analyses of SDP1 and the closely related spectral clustering method; cf. Theorem 17. We show in Proposition 13 that the optimization goals of (SDP1) and (SDP) (5) are equivalent, and both are equivalent to that of OracleSDP (29). Proposition 13 is proved in Section D.4.

Proposition 13 (OracleSDP). *Let $\widetilde{A} := YY^T - \lambda E_n$. The optimal solutions $\widehat{Z}$ as in (5) using A as in (7) must have their diagonals set to I_n. Thus, the set of optimal solutions $\widehat{Z} \in \operatorname{argmax}_{Z \in \mathcal{M}_G^+}\langle A, Z\rangle$ in (5), where $\mathcal{M}_G^+ = \{Z : Z \succeq 0, I_n \succeq \operatorname{diag}(Z)\}$, coincide with those on the convex subset $\mathcal{M}_{opt}$ as in (11),*

$$\operatorname*{argmax}_{Z \in \mathcal{M}_G^+}\langle A, Z\rangle = \operatorname*{argmax}_{Z \in \mathcal{M}_{opt}}\langle \widetilde{A}, Z\rangle = \operatorname*{argmax}_{Z \in \mathcal{M}_{opt}}(\langle A - \mathbb{E}\tau I_n, Z\rangle) \textbf{ (OracleSDP).} \tag{31}$$

3.3 Proof of Theorem 7

For a matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$, we denote by $\|(a_{ij})\|_{\infty\to 1}$ its $\ell_\infty \to \ell_1$ norm, which is

$$\|(a_{ij})\|_{\infty\to 1} = \max_{\|s\|_\infty \leq 1} \|As\|_1 = \max_{s,t\in\{-1,1\}^n} \langle A, st^T\rangle.$$

The concept of the cut norm plays a major role in the work of Frieze and Kannan (1999) on efficient approximation algorithms for dense graph and matrix problems. The cut norm is also essential for the arguments of Guedon and Vershynin (2016) to carry over to the present work. With proper adjustments, we adapt their methodology to our setting and prove the first main result, Theorem 7.

First, following Guedon and Vershynin (2016), we show that $\widehat{Z}$ provides an almost optimal solution to the reference problem if B and R are close; see Lemma 14. Second, we establish that the excess risk $\langle R, Z^* - \widehat{Z}\rangle$ exhibits a non-trivial global curvature over the feasible set $\mathcal{M}_{\text{opt}}$ (11) at the maximizer Z^* in Lemma 15.

Lemma 14 (Lemma 3.3 Guedon and Vershynin (2016))**.** *Let $\mathcal{M}_{opt}$ be as in (11). Let $\widehat{Z} = \operatorname{argmax}_{Z\in\mathcal{M}_{opt}}\langle B, Z\rangle$ for any given B and $Z^* := \operatorname{argmax}_{Z\in\mathcal{M}_{opt}}\langle R, Z\rangle$. Then*

$$0 \leq \langle R, Z^* - \widehat{Z}\rangle \leq 2K_G \|B - R\|_{\infty\to 1}\,, \text{ where } K_G < \frac{\pi}{2\ln(1+\sqrt{2})} \leq 1.783. \tag{32}$$

Lemma 15 (Excess risk lower bound)**.** *Let R be as in (12). Then for every $Z \in \mathcal{M}_{opt}$,*

$$\langle R, Z^* - Z\rangle \;\;\geq\;\; p\gamma w_{\min}^2 \|Z - Z^*\|_1, \text{ where } \;\; Z^* = u_2 u_2^T. \tag{33}$$

Theorem 16 is instrumental in proving Theorem 7 in conjunction with Lemmas 14 and 15. Importantly, computing the operator and cut norms for $B - R$ is also a key technical step unifying Theorems 7 and 17, as we demonstrate in Section 3.4. We prove Theorem 16 in Section 3.5. We prove Lemma 15 in Section 8.1.

Theorem 16 (R **is the leading term).** *Suppose the conditions in Theorem 7 (resp. Theorem 10) hold. Then with probability at least* $1 - 2\exp(-cn)$*, for the oracle* B *in* (30) *and the reference* R *in* (12)*, we have* $\|B - R\|_2 \le \xi np\gamma$ *and* $\|B - R\|_{\infty\to 1} \le \xi n^2 p\gamma$*.*

Proof of Theorem 7. Under the conditions of Lemmas 14 and 15, we have

$$\|\widehat{Z} - Z^*\|_1/n^2 \;\le\; \frac{\langle R, Z^* - \widehat{Z}\rangle}{n^2 p\gamma w_{\min}^2} \le \frac{2K_G \|B - R\|_{\infty\to 1}}{n^2 p\gamma w_{\min}^2} \le \frac{2K_G\xi}{w_{\min}^2} =: \delta,$$

where by Theorem 16, $\|B - R\|_{\infty\to 1} \le \xi n^2 p\gamma$. Finally, the bound on $\|Z^* - \widehat{Z}\|_F^2$ follows from the observation that $\|Z^* - \widehat{Z}\|_F^2 \le \|Z^* - \widehat{Z}\|_{\max}\|Z^* - \widehat{Z}\|_1$. Since all entries of $\widehat{Z}$ and Z^* belong to $[-1, 1]$, we have $\|Z^* - \widehat{Z}\|_{\max} \le 2$, yielding $\|Z^* - \widehat{Z}\|_F^2 \le 2\delta n^2$. □

3.4 Spectral clustering

Before exploring these results in Theorems 7, 10, and 11 further, we briefly discuss the spectral clustering result in Corollary 18 of Theorem 17 and compare it with Corollary 19.

In Theorem 17, we prove convergence results that bound the angle and the ℓ_2 distance between the leading eigenvector of R and that of B (resp., YY^T). Theorem 17 is new to the best of our knowledge.

Theorem 17 (**Spectral clustering).** *Denote by* v_1 *the leading unit-norm eigenvector of* YY^T*, which also coincides with that of* A (7) *and* B (30)*. Let* $\bar{v}_1$ *be the leading unit-norm eigenvector of* R *as in* (12)*. Then* $\langle \bar{v}_1, \mathbf{1}_n\rangle = 0$*, where*

$$\bar{v}_1 \;=\; [w_2\mathbf{1}_{n_1}, -w_1\mathbf{1}_{n_2}]/\sqrt{w_2 w_1 n} = [\sqrt{w_2/w_1}\mathbf{1}_{n_1}, -\sqrt{w_1/w_2}\mathbf{1}_{n_2}]/\sqrt{n}. \tag{34}$$

Let $\theta_1 = \angle(v_1, \bar{v}_1)$ *denote the angle between the two vectors* v_1 *and* $\bar{v}_1$*. Then under the conditions in Theorem 16, we have with probability at least* $1 - 2\exp(-cn)$*,*

$$\sin(\theta_1) := \sin(\angle(v_1, \bar{v}_1)) \le \frac{2\|B - R\|_2}{w_1 w_2 np\gamma} \le \frac{2\xi}{w_1 w_2}, \tag{35}$$

$$\min_{\alpha=\pm 1} \|\alpha v_1 - \bar{v}_1\|_2^2 \le \left(\frac{2^{3/2}\|B - R\|_2}{w_1 w_2 np\gamma}\right)^2 \le \delta', \quad \text{where } \delta' = \frac{8\xi^2}{(w_1^2 w_2^2)}. \tag{36}$$

Moreover, Corollary 18 shows that the signs of the coefficients of the leading eigenvector v_1 of YY^T correctly estimate the partition of the vertices, up to $O(n/s^2)$ misclassified vertices in view of (27). As such, the misclassification error is bounded to be inversely proportional to the SNR parameter s^2 in Theorem 17 and Corollary 18. This should be compared with (SDP1) and BalancedSDP, for which we have up to $O(n\exp(-c_0 s^2 w_{\min}^4))$ misclassified vertices as shown in Theorems 10, 11, and Corollary 19.

Corollary 18. *Denote by v_1 the leading unit-norm eigenvector of YY^T. Under the conditions stated in Theorem 7 or 10, it holds with probability at least $1-2\exp(-cn)$ that the signs of the coefficients of v_1 correctly estimate the partition of the vertices into two clusters, up to $O(\xi^2 n)$ misclassified vertices, where we set $\xi^2 \asymp 1/s^2$ as in (27) for $\xi = \Omega(1/n)$ as in (24).*

Corollary 19 (Exponential decay). *Let $\widehat{x}$ denote the eigenvector of $\widehat{Z}$ corresponding to the largest eigenvalue, with $\|\widehat{x}\|_2 = \sqrt{n}$. Denote by $\theta_{SDP} = \angle(\widehat{x}, u_2)$ the angle between $\widehat{x}$ and u_2. In the settings of Theorems 11 and 10, we have with probability $\geq 1-2\exp(-cn)-2/n^2$,*

$$\begin{aligned} \sin(\theta_{SDP}) &\leq 2\left\|\widehat{Z} - u_2 u_2^T\right\|_2 / n \leq \exp(-c_1 s^2 w_{\min}^4) \quad \textit{and} \qquad (37)\\ \min_{\alpha=\pm 1} \left\|(\alpha\widehat{x} - u_2)/\sqrt{n}\right\|_2 &\leq 2^{3/2}\left\|\widehat{Z} - u_2 u_2^T\right\|_2 / n \leq 4\exp(-c_0 s^2 w_{\min}^4/2). \end{aligned}$$

Under regularity assumptions on the geometry (and spectrum) of the centers (and their Gram matrix), exponentially decaying error bounds were obtained by Abbe et al. (2022) for a variant of spectral clustering on the hollowed Gram matrix $\mathcal{H}(XX^T)$, where $\mathcal{H}(\cdot)$ zeros out the diagonal entries of the Gram matrix. Hence their results apply to a restricted subset of clustering problems. Concurrent with the present work, such error bounds have also been obtained for spectral clustering based on XX^T by Zhang and Zhou (2024) under the large-n and small-p regime. We refer to Zhang and Zhou (2024) for a discussion of other related spectral clustering algorithms. We prove Theorem 17 in Section 10. Corollaries 9, 18, and 19 follow from the Davis-Kahan Perturbation Theorem and Theorem 7, Theorem 17, and Theorem 10, respectively; cf. Section 9. The proof of Corollary 18 follows that of Corollary 9 and is hence omitted.

3.5 Proof of Theorem 16 and debiasing

Our global and local analyses also show the surprising result that Theorems 7, 10, and 17 do not depend on the clusters being balanced, nor do they require identical variance or covariance profiles, provided that Assumption (A2) holds. In this section, we further illuminate the constructions of A and B and discuss bias. Theorem 16 follows immediately from Lemmas 20 and 12, and Theorem 6. We prove Lemmas 20 and 12 in Section D.

Lemma 20. *Let τ be as in (30), λ as in (7), and Y as in (6). Then,*

$$(n-1)\left|\lambda - \mathbb{E}\lambda\right| = \left|\tau - \mathbb{E}\tau\right| \leq \left\|YY^T - \mathbb{E}YY^T\right\|_2. \qquad (38)$$

Proof of Theorem 16. We will only show $\|B - \mathbb{E}B\|_2$ while noting that $\|B - \mathbb{E}B\|_{\infty\to 1} \leq n\|B - \mathbb{E}B\|_2$ by definition. By Lemma 20,

$$\|B - \mathbb{E}B\|_2 \leq \left\|YY^T - \mathbb{E}YY^T\right\|_2 + \left|\lambda - \mathbb{E}\lambda\right| \|E_n - I_n\|_2 \leq 2\left\|YY^T - \mathbb{E}YY^T\right\|_2,$$

where we use the fact that $\|E_n - I_n\|_2 \leq \|E_n - I_n\|_\infty = (n-1)$, and

$$(n-1)\left|\lambda - \mathbb{E}\lambda\right| = \left|\tau - \mathbb{E}\tau\right| = \left|\mathrm{tr}(YY^T - \mathbb{E}YY^T)\right|/n \leq \left\|YY^T - \mathbb{E}YY^T\right\|_2.$$

Moreover, the concentration of measure bounds on $\left\|YY^T - \mathbb{E}YY^T\right\|_2$ imply that, up to a constant factor, the same bounds also hold for $\|B - \mathbb{E}B\|_2$ in view of Lemma 20, since

$$A - \mathbb{E}A = B - \mathbb{E}B \;:=\; YY^T - \mathbb{E}YY^T - (\lambda - \mathbb{E}\lambda)(E_n - I_n), \text{ and} \tag{39}$$

$$\text{moreover } \|B - R\|_2 \;=\; \|B - \mathbb{E}B + \mathbb{E}B - R\|_2 \le \|B - \mathbb{E}B\|_2 + \|\mathbb{E}B - R\|_2 . \tag{40}$$

Theorem 16 holds by Lemma 12, in view of (39), (40), and Theorem 6. □

Bias and variance tradeoff. The upper bound on $\|B - R\|$ crucially depends on the SNR, n, p, γ, and (A2) as shown in Theorem 16. Specifically, the tolerance level on the RHS of (24) is chosen to be of the same order as the upper bound (20) we obtain on $\left\|YY^T - \mathbb{E}YY^T\right\|_2$ in Theorem 6. Roughly speaking, we impose $(A2')$ (42) to ensure (41):

$$\text{Bias and variance tradeoff: } \|\mathbb{E}B - R\|_2 \asymp \|B - \mathbb{E}B\|_2 \le 2\left\|YY^T - \mathbb{E}YY^T\right\|, \tag{41}$$

$$(A2')\ |V_1 - V_2| \le w = O\left(n\Delta \vee \sqrt{np}\right) \quad \text{in view of (20), where } \Delta = \sqrt{p\gamma}. \tag{42}$$

Notice that $\sqrt{np} < n\sqrt{p\gamma}$ when $n > 1/\gamma$. In particular, under **(A2)**, we do not need a separate estimator for the trace of $\Sigma_j = \mathsf{Cov}(\mathbb{Z}_j) = \mathbb{E}\mathbb{Z}_j\mathbb{Z}_j^T, j \in [n]$; cf. (53) and Lemma 5. Clearly, holding all other parameters fixed, a larger separation Δ, or a larger $D := np$, or a larger SNR s^2 will make it easier to satisfy **(A2)**. For ease of discussion, we assume (A1), where we set $C_0 \asymp 1$. Now, by definition of the SNR (28), the lower bound in (25) implies that $|V_1 - V_2| \le \sqrt{s^2}(n \vee (1/\gamma))$ is a sufficient condition for **(A2)** to hold, since

$$|V_1 - V_2| \le [n \vee (1/\gamma)]\sqrt{s^2} \asymp (n/\xi) \vee (1/(\xi\gamma)) \le \xi np\gamma,$$

where $\xi^2 \asymp 1/s^2$. Given a fixed average quality γ, the tolerance for $|V_1 - V_2|$ depends on the sample size n and the separation parameter Δ (and hence the total data size $D := np$).

Bias and the sample size. Since $|V_1 - V_2| = O(p)$ by definition, (42) holds trivially in the large-sample setting where $n \gg p$ and $\Delta = \Omega(1)$. To see this, consider (A1) again. Suppose we generate two clusters according to Lemma 5 with diagonal matrices H_1 and H_2, where $\Sigma_j := H_j H_j^T = \mathrm{diag}([\sigma_{1j}^2, \ldots, \sigma_{pj}^2])$ for all $j \in \{1, 2\}$. Then for each row vector Z_j in $\mathcal{C}_i$, we have by Lemma 5,

$$V_i := \mathrm{tr}(H_i H_i^T) = \mathbb{E}\langle \mathbb{Z}_j, \mathbb{Z}_j \rangle = \sum_{k=1}^{p} \sigma_{ki}^2 = O(p). \tag{43}$$

Hence, given fixed p and γ (and Δ), increasing the sample size n leads to a *higher tolerance* for the variance discrepancy $|V_1 - V_2|$, thus effectively widening the range of models that satisfy (A2). This confirms that the restriction imposed by (A2) is indeed most stringent in the small-sample, high-dimensional regime, and becomes non-restrictive as n grows, consistent with our earlier discussion in Section 3.1.

3.6 Related work

Given an undirected graph $G = (V, E)$ and nonnegative weights $w_{ij} = w_{ji}$ on the edges $(i, j) \in E$, the maximum cut problem (MAXCUT) is that of finding the set of vertices $S \subset V$ that maximizes the weight of the edges in the cut(S, S^c), where $S^c := V \setminus S$. See Karp

(1972); Goemans and Williamson (1995). For population clustering, results in Zhou (2006) and Chaudhuri et al. (2007) were among the first to provide a rigorous and intuitive understanding of why proposed MAXCUT-based algorithms and previous methods (Patterson et al., 2006) succeed in low-sample, high-dimensional settings, specifically when $p \gg n$ and np satisfies (25). However, these early results were restricted to balanced partitions and were purely structural, as they did not provide polynomial-time algorithms for finding the MAXCUT.

The primary contribution of Chrétien et al. (2021) is to bridge semidefinite programming (SDP) relaxations with Statistical Learning Theory. Rather than treating SDPs purely as optimization problems, the authors view SDP estimators as special instances of Empirical Risk Minimization (ERM). This allows them to apply advanced statistical tools, specifically fixed-point equations and excess risk curvature, to obtain rates of convergence for various SDP estimators toward their oracles. Through this lens, Chrétien et al. (2021) explicitly revisit the community detection problem on random graphs previously analyzed by Guedon and Vershynin (2016) and Fei and Chen (2019). Notably, they demonstrate how their newly proposed ERM framework systematically captures and compares the global analysis employed by Guedon and Vershynin (2016) with the local analysis introduced by Fei and Chen (2019), casting both bounds under a unified statistical perspective.

While Chrétien et al. (2021) analyze SDP relaxations in the context of masked random graphs, leveraging the framework of Goemans and Williamson (1995) and Guedon and Vershynin (2016), the present work focuses on high-dimensional population clustering under subgaussian designs. Unlike the random graph setting, our framework explicitly addresses the challenges of dependent features by introducing (SDP1) and a debiased BalancedSDP estimator, which allows us to establish exponential error bounds for partial cluster recovery in both low ($n \gg p$) and high-dimensional settings ($p \gg n$).

In the theoretical computer science literature, earlier work focused on learning from mixtures of well-separated Gaussians (component distributions), where one aims to classify each sample according to which component distribution it comes from. In Dasgupta and Schulman (2000); Arora and Kannan (2001), the separation requirement depends on the number of dimensions of each distribution; this has subsequently been reduced to be independent of p, the dimensionality of the distribution for certain classes of distributions; see, for example, Achlioptas and McSherry (2005) and Kannan et al. (2005). While our aim is different from those earlier results, where $n > p$ is almost universal and we focus on cases $p > n$ (a.k.a. high dimensional setting), we do have one common axis for comparison, the ℓ_2-distance between any two centers of the distributions as stated in (44):

$$\Delta^2 = p\gamma \geq C \max_i \|\mathsf{Cov}(\mathbb{Z}_i)\|_2 \left(1/\xi^2 \vee \sqrt{p/(n\xi^2)}\right) \quad \text{for some} \quad 0 < \xi < 1/2, \tag{44}$$

which is essentially optimal; cf. (25) and (19). The main contribution of the present work is: we use the proposed SDP (5) and the related spectral algorithms to find the partition, and prove quantitatively tighter bounds than those obtained by Zhou (2006) and Chaudhuri et al. (2007) by removing these logarithmic factors entirely. Recently, these barriers have also been broken down by a sequence of work (Royer, 2017; Fei and Chen, 2018; Giraud and Verzelen, 2019). However, these recent results still need the SNR to be at the order of $s^2 = \Omega(\log n)$ even for $n > p$, except for the work of Giraud and Verzelen (2019). We compare with Giraud and Verzelen (2019) in Section 6.

Even earlier, a small simulation example (Blum et al., 2009) shows that even when the sample size n is small, by increasing p so that $np = \Omega(1/\gamma^2)$, one can classify a mixture of two product populations in $\{0,1\}^p$ using spectral methods therein with success rate reaching an "oracle" curve, where success rate means the ratio between correctly classified individuals and the sample size n. This behavior was intriguing. To be clear, the "oracle" was computed assuming that distributions were known. Our concentration of measure bounds, namely, Theorems 16 and 6, enable us to bypass the extraneous lower bound on sample size n in Blum et al. (2009) as shown in Theorem 17 and Corollary 18. We reproduce these experiments in Section 7, where we show that a full range of tradeoffs between the sample size and the number of features are feasible so long as s^2 (28) is lower bounded. We also refer to Vempala and Wang (2002), Achlioptas and McSherry (2005), Kalai et al. (2010), Kumar and Kannan (2010), Rohe et al. (2011), Lei and Rinaldo (2015), Guedon and Vershynin (2016), Javanmard et al. (2016), Abbe (2018), Balakrishnan et al. (2017), Iguchi et al. (2017), Zhou and Amini (2019), Li et al. (2020), Chrétien et al. (2021), Löffler et al. (2021), Yurtsever et al. (2021), Abbe et al. (2022), Ndaoud (2022), Doss et al. (2023), Zhang and Zhou (2024), and references therein for more related works on SDP and clustering.

4. Proof sketch of Theorems 10 and 11

The upper bound on the excess risk $\langle R, Z^* - \widehat{Z}\rangle$ in (32) holds for all $\widehat{Z} \in \mathcal{M}_{\rm opt}$, and hence is rather crude. In this section, we first present Lemma 21, through which we derive a tighter upper bound on the excess risk $\langle R, Z^* - \widehat{Z}\rangle$, so as to replace (32). In particular, the upper bound on $\langle R, Z^* - \widehat{Z}\rangle$ will depend on the ℓ_1 distance between $\widehat{Z}$ and Z^*, rather than the cut norm of the matrix $B - R$. All constants such as $1/6, 2/3, \ldots$ are arbitrarily chosen.

Lemma 21 (Elementary inequality). *Let the reference matrix R be constructed as in* (12)*. Let B be as in Theorem 17. By optimality of $\widehat{Z} \in \mathcal{M}_{opt}$ as* (11)*, we have*

$$\begin{aligned} p\gamma w_{\min}^2 \|\widehat{Z} - Z^*\|_1 \;\le\; & \langle R, Z^* - \widehat{Z}\rangle \le \langle YY^T - \mathbb{E}YY^T, \widehat{Z} - Z^*\rangle + \langle \mathbb{E}B - R, \widehat{Z} - Z^*\rangle \\ & - \langle (\lambda - \mathbb{E}\lambda)(E_n - I_n), \widehat{Z} - Z^*\rangle =: F_1 + F_2 + F_3. \end{aligned} \tag{45}$$

With high probability, we obtain a uniform control for $F_1 = \langle YY^T - \mathbb{E}YY^T, \widehat{Z} - Z^*\rangle$ over all $\widehat{Z} \in \mathcal{M}_{\rm opt}$ in an ℓ_1-*neighborhood* of Z^* of radius r_1 in Lemma 22.

Lemma 22. *Suppose all conditions in Theorem 6 hold. Let $r_1 \le 2qn$ for a positive integer $1 \le q < n$. Then on event $\mathcal{G}_1$, where $\mathbb{P}(\mathcal{G}_1) \ge 1 - c/n^2$, we have for $\xi \le w_{\min}^2/16$ and $\sigma_{\max} = \max_i \|H_i\|_2$,*

$$\begin{aligned} & \sup_{\widehat{Z} \in \mathcal{M}_{opt} \cap (Z^* + r_1 B_1^{n\times n})} |\langle YY^T - \mathbb{E}YY^T, \widehat{Z} - Z^*\rangle| \\ & \qquad \le \frac{5}{6}\xi p\gamma \|\widehat{Z} - Z^*\|_1 + C'\big(\sigma_{\max} n\sqrt{p\gamma} + \sigma_{\max}^2\sqrt{np}\big)\left\lceil \frac{r_1}{2n}\right\rceil \sqrt{\log(2en/\lceil \frac{r_1}{2n}\rceil)}. \end{aligned} \tag{46}$$

Such an upper bound (46) now depends on the ℓ_1 radius $r_1 \le 2n(n-1)$. This approach also gives rise to the notion of a *local analysis*, following Fei and Chen (2019), Chrétien et al. (2021), and Mammen and Tsybakov (1999). Therefore, Lemma 22 is one of the main contributions in this paper, which in turn depends on the geometry of the constraint set

$\mathcal{M}_{\rm opt}$ and the sharp concentration bounds on $\|YY^T - \mathbb{E}YY^T\|$ in Theorem 6. Similar to the global analysis, the performance of SDP1 also crucially depends on controlling the bias effect: we now control $\langle \mathbb{E}B - R, \widehat{Z} - Z^*\rangle$ uniformly over all $\widehat{Z} \in \mathcal{M}_{\rm opt}$ in Lemma 23. We emphasize that (A2) is only needed for Lemma 23.

Lemma 23. *Suppose all conditions in Theorem 10 hold. Then for all $\widehat{Z} \in \mathcal{M}_{opt}$,*

$$\left|\langle \mathbb{E}B - R, \widehat{Z} - Z^*\rangle\right| \quad =: \quad |F_2| \le 2p\gamma \|\widehat{Z} - Z^*\|_1\big(\xi + 1/[4(n-1)]\big). \tag{47}$$

Lemma 24. *Under the settings of Theorem 6, with probability at least $1 - \exp(-cn)$,*

$$\forall \widehat{Z} \in \mathcal{M}_{opt}, \quad \left|\langle (\lambda - \mathbb{E}\lambda)(E_n - I_n), \widehat{Z} - Z^*\rangle\right| =: |F_3| \ \le \ \xi p\gamma \|\widehat{Z} - Z^*\|_1/3.$$

Lemmas 23 and 24 show that each term may take out only a small fraction of the signal on the LHS of (45) respectively. We prove Lemmas 21, 22, and 23 in Sections 11.1, 11.2, and 11.3, respectively. Lemma 24 follows from the Lemma 20 and Theorem 6 immediately. Combining these results leads to an exponentially decaying error bound with respect to the SNR s^2 in Theorem 10; cf. Section 8.

Balanced partitions. For BalancedSDP (9), the bias term arising from local analysis is now 0. More precisely, we show $\forall \widehat{Z} \in \mathcal{M}_{\rm opt}, F_2 := \langle \mathbb{E}B - R, \widehat{Z} - Z^*\rangle = 0$, even when $V_1 \neq V_2$. Our design and local analysis on semidefinite relaxations suggest that the original SDP1 may already have a substantially reduced bias F_2 for balanced partitions; cf. Conjecture 37 and discussions in Section 6. See Section 12 for details.

5. Reduction: proof outline for Theorem 6

In this section, we present a unified framework for bounding $\left\|YY^T - \mathbb{E}YY^T\right\|_2$. The decomposition and projection ideas in this section are also crucial in the local analyses for Theorems 11 and 10; cf. proof of Lemma 22 in Section 11.2. Let $c, c', C, C_2, C_3, C_4, \ldots$ be absolute constants. First, by linearity of expectation, we have for Y as in (6), with row vectors $Y_1, \ldots, Y_n \in \mathbb{R}^p$,

$$\mathbb{E}Y_i := \mathbb{E}X_i - \mathbb{E}\widehat{\mu}_n := \begin{cases} w_2(\mu^{(1)} - \mu^{(2)}) & \text{if } \ i \in \mathcal{C}_1; \\ w_1(\mu^{(2)} - \mu^{(1)}) & \text{if } \ i \in \mathcal{C}_2, \end{cases} \quad \text{where } \widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i. \tag{48}$$

Next, we decompose $YY^T - \mathbb{E}YY^T$ in Lemma 25. Controlling the first component, namely, $\widehat{\Sigma}_Y - \Sigma_Y$ in (49) amounts to the problem of covariance estimation, where we assume (implicitly) that $\mathbb{E}X$ is given. In particular, controlling the norm of $\widehat{\Sigma}_Y - \Sigma_Y$ (50) is reduced to bounding that for $\mathbb{Z}\mathbb{Z}^T - \mathbb{E}\mathbb{Z}\mathbb{Z}^T$ as shown in Lemma 26.

Lemma 25 (Decomposition). *Let $Y = (I - P_1)X$ be as in (6). Suppose $\mathbb{Z} = X - \mathbb{E}X$. Let $\widehat{\Sigma}_Y = (Y - \mathbb{E}Y)(Y - \mathbb{E}Y)^T$ and $\Sigma_Y := \mathbb{E}YY^T - \mathbb{E}(Y)\mathbb{E}(Y)^T$. Then*

$$YY^T - \mathbb{E}YY^T \ = \ \widehat{\Sigma}_Y - \Sigma_Y + \mathbb{E}(Y)(Y - \mathbb{E}(Y))^T + (Y - \mathbb{E}(Y))(\mathbb{E}(Y))^T, \tag{49}$$

$$\text{where } \widehat{\Sigma}_Y - \Sigma_Y \ = \ (I - P_1)(\mathbb{Z}\mathbb{Z}^T - \mathbb{E}\mathbb{Z}\mathbb{Z}^T)(I - P_1), \text{ for } \ P_1 = E_n/n. \tag{50}$$

Lemma 26. *Let* $\sigma_{\max}^2 := \max_j \|\mathsf{Cov}(\mathbb{Z}_j)\|_2 = \max_i \|H_i\|_2^2$ *and* C_0 *be as in Theorem 6. In the settings of Theorem 6 (resp. 4) and Lemma 25, we have with probability at least* $1 - 2\exp(-c_6 n)$,

$$\left\|\widehat{\Sigma}_Y - \mathbb{E}\widehat{\Sigma}_Y\right\|_2 \le \left\|\mathbb{Z}\mathbb{Z}^T - \mathbb{E}\mathbb{Z}\mathbb{Z}^T\right\|_2 \le C_2 C_0^2 \sigma_{\max}^2(\sqrt{np} \vee n), \tag{51}$$

where $K \le C_0 = \max_j \|W_j\|_{\psi_2} \le CK$ *for an absolute constant* C.

Next, we state a reduction principle: to control $\left\|\mathbb{E}(Y)(Y - \mathbb{E}(Y))^T + (Y - \mathbb{E}(Y))(\mathbb{E}(Y))^T\right\|$, we need to bound the projection of each mean-zero random vector $\mathbb{Z}_j, \forall j \in [n]$, along the direction of $\mu = (\mu^{(1)} - \mu^{(2)})/\sqrt{p\gamma} \in \mathbb{S}^{p-1}$. In other words, Lemma 27 shows that a particular direction for which we compute the one-dimensional marginals for $\mathbb{Z}_j, j \in [n]$ is the direction between the two centers, namely, $\mu^{(1)}$ and $\mu^{(2)}$ in $\mathbb{R}^p$. Moreover, the bound in (52) is deterministic and independent of the covariance structure of $\mathbb{Z}$. Combining Lemma 27 with the sub-gaussian tail bounds, we have Lemma 28.

Lemma 27 (Reduction: a deterministic comparison lemma). *Let* $\mathbb{Z}_j, j \in [n]$ *be row vectors of* $X - \mathbb{E}X$ *and* $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$. *For* $x_i \in \{-1, 1\}$,

$$\sum_{i=1}^n x_i \langle Y_i - \mathbb{E}Y_i, \mu^{(1)} - \mu^{(2)}\rangle \;\; \le \;\; 2(n-1)/n \sum_{i=1}^n \left|\langle \mathbb{Z}_i, \mu^{(1)} - \mu^{(2)}\rangle\right|. \tag{52}$$

Then we have for $M_Y := \mathbb{E}(Y)(Y - \mathbb{E}(Y))^T + (Y - \mathbb{E}(Y))(\mathbb{E}(Y))^T$,

$$\|M_Y\|_2 \;\; \le \;\; 4\sqrt{n}\sqrt{w_1 w_2} \sup_{q \in S^{n-1}} |\sum_i q_i \langle \mathbb{Z}_i, \mu^{(1)} - \mu^{(2)}\rangle|.$$

Lemma 28 (Projection: probabilistic view). *Let* M_Y *be as in Lemma 27. Let* μ *be as defined in* (18). *Suppose all conditions in Theorem 6 hold. Then with probability at least* $1 - 2\exp(-c'n)$, $\|M_Y\|_2 \le 2C_3C_0(\max_i \left\|H_i^T\mu\right\|_2)n\sqrt{p\gamma} \le \xi np\gamma/12$.

Throughout this paper, it is understood that Theorem 4, Lemma 25, (51), and (52) hold under the covariance model as considered in Definition 3 regardless of the weights or the number of mixture components in data matrix X. Such generalization is useful for the k-component mixture problems (23) for any $k \ge 2$; cf. Section 6. Theorem 6 (for $k = 2$) follows from Lemmas 25 and 28, and (51) immediately, and the probability statements hold upon adjusting the constants.

Remarks on the diagonal covariance assumption. In Theorem 7, each noise vector $\mathbb{Z}_j, \forall j \in [n]$ has independent, mean-zero, subgaussian coordinates with uniformly bounded ψ_2 norms. Suppose now we generate two clusters according to Lemma 5 with diagonal matrices H_1 and H_2, respectively, where $H_j H_j^T = \mathrm{diag}([\sigma_{1j}^2, \ldots, \sigma_{pj}^2])$ for all $j \in \{1, 2\}$. Here, we use $\mathrm{diag}(x)$ to denote the diagonal matrix whose main diagonal entries are the entries of $x = [x_1, \ldots, x_p]$. For each row vector in $\mathcal{C}_i, i = 1, 2$, it follows from Lemma 5 that $\mathsf{Cov}(\mathbb{Z}_j) = H_i H_i^T$ for nodes $j \in \mathcal{C}_i$ and $V_i = \mathrm{tr}(H_i H_i^T) = \sum_{k=1}^p \sigma_{ki}^2$. Let $\sigma_{\max} := \max_i \|H_i\|_2 = (\max_{j,k} \mathbb{E}(z_{jk}^2))^{1/2}$. Then clearly, $\max_j \|\mathsf{Cov}(\mathbb{Z}_j)\|_2 = \sigma_{\max}^2$. Given the independence of the coordinates of $\mathbb{Z}_j$, we have (15) holds with $C_0 \le CK$, since

$$\begin{aligned} \|\mathbb{Z}_j\|_{\psi_2} \;\; &:= \;\; \sup_{h \in \mathbb{S}^{p-1}} \|\langle \mathbb{Z}_j, h\rangle\|_{\psi_2} \le C \max_{k \le p} \|z_{jk}\|_{\psi_2} \le CK\sigma_{\max}, \quad \text{where} \\ \max_{k \le p} \|z_{jk}\|_{\psi_2} \;\; &\le \;\; \sigma_{\max}(\max_{j,k} \|w_{jk}\|_{\psi_2}) = \sigma_{\max} K, \quad \text{for} \quad K := \max_{j,k} \|w_{jk}\|_{\psi_2}. \end{aligned}$$

Remarks on covariance estimation. More generally, we consider the general data generative process in Definition 3, and denote the (sample-by-sample) covariance by

$$\Gamma := \mathbb{E}\mathbb{Z}\mathbb{Z}^T = \mathrm{diag}([\mathrm{tr}(\Sigma_1), \ldots, \mathrm{tr}(\Sigma_n)]), \text{ where } \Sigma_j = \mathsf{Cov}(\mathbb{Z}_j) = H_j H_j^T; \tag{53}$$

Thus, we estimate a diagonal matrix with p dependent features. Essentially, (51) matches the optimal bounds for covariance estimation, where the mean-zero random matrix consists of independent columns $\mathbb{Z}^j, j \in [p]$, which are isotropic, sub-gaussian random vectors in $\mathbb{R}^n$, or columns which can be transformed to be isotropic through a common covariance matrix; see, for example, Theorems 4.6.1 and 4.7.1 (Vershynin, 2018). The differences between (51) and such known results are: (a) we do not assume that columns are independent; and (b) we do not require anisotropic row vectors to share identical covariance matrices. See Zhou (2019, 2024a) and references therein for extensions of the Hanson-Wright inequalities in the complex matrix-variate data. We prove Theorem 6 and Lemmas 26 and 28 in Section C. See Section L Zhou (2026) for the proofs of Lemmas 25 and 27.

6. The k-means criterion of a partition: a comparison

We now discuss the related k-means criterion and its semidefinite relaxations. Denote by $X \in \mathbb{R}^{n\times p}$ the data matrix with row vectors X_i as in (54). The k-means criterion of a partition $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ of sample points $\{1, \ldots, n\}$ is based on the total sum-of-squared Euclidean distances from each point $X_i \in \mathbb{R}^p$ to its assigned cluster centroid $\mathbf{c}_j$, namely,

$$g(X, \mathcal{C}, k) \;:=\; \sum_{j=1}^{k} \sum_{i\in\mathcal{C}_j} \|X_i - \mathbf{c}_j\|_2^2, \quad \text{where} \quad \mathbf{c}_j := \sum_{\ell\in\mathcal{C}_j} X_\ell / \left|\mathcal{C}_j\right| \in \mathbb{R}^p. \tag{54}$$

Getting a global solution to (54) through an integer programming formulation (Peng and Wei, 2007) is NP-hard and it is NP-hard for $k = 2$ (Drineas et al., 2004; Aloise et al., 2009). The partition $\mathcal{C}$ can be represented by a block-diagonal matrix of size $n \times n$, defined as: $\forall i, j \in [k]^2, \forall a, b \in \mathcal{C}_i \times \mathcal{C}_j$, $B^*_{ab} = 1/|\mathcal{C}_j|$ if $i = j$ and $B^*_{ab} = 0$ otherwise. Let Ψ_n denote the linear space of real n by n symmetric matrices. The collection of such matrices can be described by

$$\mathcal{P}_k = \{B \in \Psi_n : B \geq 0, B^2 = B, \mathrm{tr}(B) = k, B\mathbf{1}_n = \mathbf{1}_n\}.$$

Here $B \geq 0$ means that all elements of B are nonnegative. Hence matrices in $\mathcal{P}_k$ are block-diagonal, symmetric, nonnegative projection matrices with $\mathbf{1}_n$ as an eigenvector. Minimizing the k-means objective $g(X, \mathcal{C}, k)$ is equivalent to

$$\text{maximize} \quad \langle XX^T, Z\rangle \quad \text{subject to} \quad Z \in \mathcal{P}_k; \text{ cf. Peng and Wei (2007).} \tag{55}$$

Peng-Wei relaxations. Peng and Wei (2007) first replace the requirement that $Z^2 = Z$, namely, Z is a projection matrix, with $I_n \succeq Z \succeq 0$. Let $\mathcal{M}_k = \{Z \in \Psi_n : Z \geq 0, Z \succeq 0, \mathrm{tr}(Z) = k, Z\mathbf{1}_n = \mathbf{1}_n\}$. Now, Peng and Wei (2007) consider the semidefinite relaxation of the k-means objective (55),

$$\text{maximize} \quad \langle XX^T, Z\rangle \qquad \text{subject to} \qquad Z \in \mathcal{M}_k. \tag{56}$$

Note that in $\mathcal{M}_k$, we have $\|Z\|_2 \le \|Z\|_\infty = 1$, since $Z_{ij} \ge 0, \forall i, j$ and row sum $\|Z_{j,\cdot}\|_1 = 1, \forall j$. The main issue with the k-means relaxation is that the solutions tend to put sample points into groups of the same size, and moreover, the diagonal matrix Γ as in (53) can cause a bias, especially when $\mathrm{tr}(\Sigma_1), \mathrm{tr}(\Sigma_2), \ldots$ differ from each other. To address this bias, building upon the original Peng-Wei SDP relaxation (56), the authors of Bunea et al. (2020), Royer (2017), and Giraud and Verzelen (2019) propose a preliminary estimator of Γ, denoted by $\widehat{\Gamma}$, and consider

$$\text{maximize} \quad \langle XX^T - \widehat{\Gamma}, Z \rangle \quad \text{subject to} \quad Z \in \mathcal{M}_k \ \text{ for } \mathcal{M}_k \quad \text{as in (56).} \tag{57}$$

Variation 2. To speed up computation, one can drop the nonnegative constraint on elements of Z in (56). In particular, Peng and Wei (2007) consider

$$\text{maximize} \quad \langle XX^T, Z \rangle \quad \text{subject to} \quad Z\mathbf{1}_n = \mathbf{1}_n, I_n \succeq Z \succeq 0, \mathrm{tr}(Z) = k. \tag{58}$$

Peng and Wei (2007) show that the set of feasible solutions to (58) has immediate connections to the SVD of YY^T. Recall $YY^T = (I - P_1)XX^T(I - P_1)$. Let $\lambda_1 \ge \ldots \ge \lambda_{n-1}$ be the largest $(n-1)$ eigenvalues of YY^T in descending order. When Z is a feasible solution to (58), $\mathbf{1}_n/\sqrt{n}$ is the unit-norm leading eigenvector of Z and define $Z_1 := (I - P_1)Z = (I - P_1)Z(I - P_1) \succeq 0$. Then, (58) is reduced to

$$\text{maximize} \quad \langle YY^T, Z_1 \rangle \text{ subject to} \quad I_n \succeq Z_1 \succeq 0, \quad \mathrm{tr}(Z_1) = k - 1. \tag{59}$$

The optimal solution to (59) can be achieved if and only if $\langle YY^T, Z_1 \rangle = \sum_{i=1}^{k-1} \lambda_i$ (Overton and Womersley, 1993). Then the SVD-based algorithm for solving (59) and (58) is given as follows:
(a) Use the singular value decomposition method to compute the first $k-1$ largest eigenvalues of YY^T, and their corresponding eigenvectors $v_1, \ldots, v_{k-1}$;
(b) Set $Z_1 = \sum_{j=1}^{k-1} v_j v_j^T$ and return $Z = \mathbf{1}_n\mathbf{1}_n^T/n + Z_1$ as a solution to (58). Now for $k = 2$, we have $Z_1 = v_1 v_1^T$ and Theorem 17.

In summary, the key differences between (57) and SDP (5) are as follows: (a) In the convex set $\mathcal{M}_{\text{opt}}$ (11), we do not enforce entry-wise non-negativity (i.e., $Z_{ij} \ge 0$ for all i, j), nor the trace or row-sum constraints ($Z\mathbf{1}_n = \mathbf{1}_n$). Replacing these $O(n^2)$ constraints with exactly n diagonal linear constraints $\{Z_{jj} = 1, \forall j \in [n]\}$ substantially reduces the theoretical complexity of the semidefinite optimization problem; see, for example, Burer and Monteiro (2003); Monteiro (2003); Yurtsever et al. (2021), and references therein. (b) To derive sufficiently tight concentration of measure bounds, we perform a crucial data-processing step by centering the data according to their column means, following (6). This allows the bias to be substantially reduced, or eliminated entirely for balanced partitions using the BalancedSDP estimator. (c) By enforcing $\langle E_n, Z \rangle = 0$ in BalancedSDP (9), the trace-difference assumption (A2) is eliminated entirely in Theorem 11 for balanced partitions. (d) Consequently, beyond reducing the size of the constraint set, a primary advantage of (SDP1) and its variants is that estimating the individual covariance traces $\mathrm{tr}(\Sigma_j)$ is unnecessary, both for balanced partitions and for general partitions satisfying (A2).

Finally, while the results of Giraud and Verzelen (2019) hold for general $k \ge 2$, the MAXCUT relaxations studied here focus on $k = 2$. Nonetheless, several core components of our analysis extend immediately: our probabilistic bounds (Theorem 4 and Lemma 26)

are established under the general k-means setting (23) for any $k \geq 2$, and the deterministic bounds (Lemmas 27 and 25) can be readily adapted and applied to the rank-k model setting, as mentioned in Section 5.

6.1 The k-means criterion

Denote $\Sigma_j = \mathsf{Cov}(\mathbb{Z}_j)$ and let $\sigma^2 := C_0^2 \max_j \|\Sigma_j\|_2$. Note that both Giraud and Verzelen (2019) and the present work require that $\Delta = \Omega(\sigma)$. In particular, the SNR defined in Giraud and Verzelen (2019), denoted s_{GV}^2, is compared against our SNR, denoted s_Z^2 from (21), as follows:

$$\textbf{(GV):} \quad s_{GV}^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{n\Delta^4}{C_0^4 \max_j \|\Sigma_j\|_F^2}, \quad \text{whereas our } \textbf{SNR (Z):} \quad s_Z^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{n\Delta^4}{p\sigma^4}, \tag{60}$$

where $\Delta^2 = p\gamma$. Theorem 1 in Giraud and Verzelen (2019) states that as long as $s_{GV}^2 = \Omega(k)$, an exponentially decaying misclassification error with respect to the SNR is achieved, a result analogous to Theorem 10 and Corollary 19. Under the assumptions in Definition 3, the SNR s_{GV}^2 in (60) is asymptotically equivalent to s_Z^2 in (21) provided that Σ_j is well-conditioned, in the sense that $0 < \lambda_{\min}(\Sigma_j) \leq \lambda_{\max}(\Sigma_j) < \infty$. This follows from the trace-operator norm relationship:

$$\text{for each } \; j, \quad p\lambda_{\min}(\Sigma_j)^2 \leq \sum_{k=1}^{p} \lambda_k(\Sigma_j)^2 = \|\Sigma_j\|_F^2 \leq p\,\|\Sigma_j\|_2^2\,,$$

which implies that $\|\Sigma_j\|_F^2$ scales linearly with the dimension p. Hence, for isotropic designs or well-conditioned covariance matrices, the error rate in Theorem 10 is essentially equivalent to that obtained by Giraud and Verzelen (2019).

These two definitions diverge primarily when $n \ll p$, as elaborated below; in the large-sample setting where $n \gg p$, both reduce to $s_Z^2 = s_{GV}^2 = \Delta^2/\sigma^2$. Moreover, it is always true that our s_Z^2 in (21) is upper bounded by s_{GV}^2 in (60), that is, $s_Z^2 \leq s_{GV}^2$. Let $r(\Sigma_j) = \frac{\|\Sigma_j\|_F^2}{\|\Sigma_j\|_2^2}$ denote the stable rank. In summary, the error rate $\exp(-cs^2)$ in Theorem 10 will be slightly slower for anisotropic designs when $n < p$ but $r(\Sigma_j) = O(p), \forall j$. On the other hand, when $\max_j r(\Sigma_j) \ll p$, that is, when $\|\Sigma_j\|_F^2$ is significantly smaller than $p\,\|\Sigma_j\|_2^2, \forall j$, the ratio s_Z^2/s_{GV}^2 becomes vanishingly small. We emphasize that such regimes fall outside the scope of our analysis in the present work. More accurately, we consider the following three cases for $\Delta^2 = p\gamma \geq \sigma^2 := C_0^2 \max_j \|\Sigma_j\|_2$,

- **Case 1:** Suppose $n\gamma > \sigma^2$. Then $s_Z^2 = s_{GV}^2 = \Delta^2/\sigma^2$, since

$$n\Delta^2 = np\gamma > p\sigma^2, \quad \text{and consequently,} \quad \frac{n\Delta^4}{C_0^4 \max_j \|\Sigma_j\|_F^2} \geq \frac{n\Delta^4}{p\sigma^4} > \frac{\Delta^2}{\sigma^2}.$$

- **Case 2:** Suppose $p\gamma \geq \sigma^2 \geq n\gamma$, which implies $p\sigma^2 \geq np\gamma$. Here, $np\gamma$ satisfies:

$$p\sigma^2 \geq np\gamma \geq \frac{C_0^4 \max_j \|\Sigma_j\|_F^2}{C_0^2 \max_j \|\Sigma_j\|_2} \quad \text{so that} \quad \frac{n\Delta^4}{C_0^4 \max_j \|\Sigma_j\|_F^2} \geq \frac{\Delta^2}{\sigma^2} \geq \frac{n\Delta^4}{p\sigma^4}.$$

Then $s_{GV}^2 = \Delta^2/\sigma^2$ while $s_Z^2 = n\Delta^4/(p\sigma^4)$. Hence, $s_Z^2/s_{GV}^2 = n(p\gamma)/(p\sigma^2) = n\gamma/\sigma^2$.

- **Case 3:** Suppose $np\gamma \leq \frac{C_0^4 \max_j \|\Sigma_j\|_F^2}{C_0^2 \max_j \|\Sigma_j\|_2}$. Then

$$s_Z^2 \asymp \frac{n\Delta^4}{pC_0^4 \max_j \|\Sigma_j\|_2^2} = \frac{n\Delta^4}{p\sigma^4} \leq \frac{n\Delta^4}{C_0^4 \max_j \|\Sigma_j\|_F^2} \leq \frac{\Delta^2}{C_0^2 \max_j \|\Sigma_j\|_2} := \frac{\Delta^2}{\sigma^2}.$$

Thus, we have $s_Z^2/s_{GV}^2 \approx \frac{\max_j \|\Sigma_j\|_F^2}{p \max_j \|\Sigma_j\|_2^2} \leq 1$ since $\forall j,\ \|\Sigma_j\|_F^2 \leq p\|\Sigma_j\|_2^2 \leq p(\max_j \|\Sigma_j\|_2^2)$.

Discussion. For **Case 2**, the further $n\gamma$ is below σ^2, the larger the performance gap becomes. When $n\gamma \approx \sigma^2$, the two estimators perform nearly identically. When $n\gamma \ll \sigma^2$, the gap widens. For **Case 3**, we have $n \leq \frac{\max_j \|\Sigma_j\|_F^2}{\max_j \|\Sigma_j\|_2^2} \leq p$. The ratio satisfies:

$$\frac{s_Z^2}{s_{GV}^2} = \frac{\max_j \|\Sigma_j\|_F^2}{p \max_j \|\Sigma_j\|_2^2} \leq \max_j \frac{r(\Sigma_j)}{p}.$$

In isotropic cases, $r(\Sigma_j) \approx p, \forall j$, making the ratio ≈ 1. When $\max_j r(\Sigma_j) \ll p$, this ratio reaches its worst-case bound in this small-n, large-p regime.

7. Experiments

In this section, we use simulation to illustrate the effectiveness and convergence properties of the three estimators. We implement (i) **SDP1** as described in (8), (ii) **BalancedSDP** as described in (9), with the extra constraint $\mathrm{tr}(E_n Z) = 0$, and (iii) **SVD**, the Peng-Wei spectral method following Peng and Wei (2007). For both SDP1 and BalancedSDP we classify according to the signs of $\widehat{x}$ as prescribed by Corollary 9. We list the SVD procedure in Table 2. The experiment set up is similar to the one used in Blum et al. (2009). The data matrix $X \in \{0,1\}^{n\times p}$ is generated as a mixture of two populations, and consists of independent Bernoulli random variables, where the mean parameters $\mathbb{E}(X_{ij}) := q^j_{\psi(i)}$ for all $i \in [n]$ and $j \in [p]$, where $\psi(i) \in \{1,2\}$ assigns nodes i to a group $\mathcal{C}_1$ or $\mathcal{C}_2$ for each $i \in [n]$. Let $|\mathcal{C}_1| = w_1 n$ and $|\mathcal{C}_2| = w_2 n$. We conduct experiments for both balanced ($w_1 = w_2$) and unbalanced cases. The entrywise expected values are chosen according to Table 3(left) for the two populations, where f_1, f_2, f_3 and m_1, m_2 are configurable. We experiment with two configurations of population variance profiles. (i) Config1: $f_1 = 50\%, f_2 = 50\%, f_3 = 0$, where $V_1 = V_2$. Here $\gamma = 0.0016$ and $1/\gamma = 625$. (ii) Config2: $f_1 = 50\%, f_2 = 30\%, f_3 = 20\%$, and $m_1 = 0.3, m_2 = 0.1$, where $V_1 \neq V_2$, and $\gamma = 0.00928$ and $1/\gamma = 107$. See Table 3(right) for the actual $p\gamma$ values.

Success rate and misclassification rate. For each experiment, we run 100 trials, and for each trial, we first generate a data matrix $X_{n\times p}$ according to the mixture of two Bernoulli distributions with parameters described above, and then feed Y (6) to the three estimators for classification. We measure the success rate and misclassification rate based on P, the output assignment vector. The success rate is computed as the number of correctly classified individuals divided by the sample size n. Hence the misclassification rate is $1-$ success rate. Fig. 1(top left) shows the average success rates of SDP1 and SVD (over 100 trials) for the balanced cases with $V_1 = V_2$ as n increases for different values of p. We observe that SDP1 has higher average success rate for each setting of (n,p) when $np\gamma^2 > 1.5$, despite the

SVD: Spectral method for k-means clustering (Peng and Wei, 2007):

Input: Centered data matrix $Y \in \mathbb{R}^{n\times p}$, $k = 2$

Output: A group assignment vector P

Step 1. Use SVD to obtain the leading eigenvector v_1 of YY^T and let $v := v_1$;

Step 2. Let S be the vector of sorted values of v in descending order. For each index j in $[n]$, compute the two means $\mathbf{c}_1$, $\mathbf{c}_2$, one for each of the two groups, namely, $\mathcal{C}_1 = S_L := \{S_1, \ldots, S_j\}$ and $\mathcal{C}_2 = S_R := \{S_{j+1}, \ldots, S_n\}$ to the left (inclusive) and the right of this index;

Step 3. Compute the total sum-of-squared Euclidean distances from each point within a particular group to the respective mean, according to (54). Let t be the index that gives the minimum total distance, and its corresponding value be S_t;

Step 4. Set $P_i = 1$ if $v_i \geq S_t$, and $P_i = -1$ if $v_i < S_t$.

Table 2: SVD: Spectral method for k-means clustering (Peng and Wei, 2007)

percentage of features	f_1	f_2	f_3
mean parameters (pop. 1)	$\frac{1+\alpha}{2} + \frac{\epsilon}{2}$	$\frac{1-\alpha}{2} + \frac{\epsilon}{2}$	m_1
mean parameters (pop. 2)	$\frac{1-\alpha}{2} + \frac{\epsilon}{2}$	$\frac{1+\alpha}{2} + \frac{\epsilon}{2}$	m_2

p	V_1	V_2	$\Delta^2 = p\gamma$
100	24.17	21.77	0.928
200	48.33	43.54	1.856
400	96.66	87.07	3.712
1000	241.66	217.69	9.280

Table 3: Left: Entrywise expected values. $\alpha = 0.04$, $\epsilon = 0.1\alpha$. Right: Population profiles in Config2, where $V_1 \neq V_2$

exhaustive search in SVD. For $np\gamma^2 < 1.5$, the rates are closer. We also see that when $p < 1/\gamma = 625$, for example, when $p = 500$, the success rate remains flat across n. Note that a success rate of 50% is equivalent to a total failure. In contrast, when $n < 1/\gamma = 625$, we can obtain a higher success rate as p increases, as shown in Fig. 1(top right) for SDP1. In general, $np\gamma^2 > 1$ is indeed necessary to obtain a success rate larger than 60%, when $p \geq 1/\gamma$. When $n < 625$, $np\gamma^2$ plays the role of the SNR, since $np\gamma^2 < p\gamma$. This remains the case throughout our experiments.

BalancedSDP. In the bottom row of Fig. 1, we compare the success rates of all three estimators under two configurations. We observe the average success rates of BalancedSDP and SDP1 are essentially the same for each setting of (n, p) for both configurations. These results are consistent with Conjecture 37, and confirm that the additional constraint in BalancedSDP does not play a major role on the average success rate for the current set of experiments.

Angle and ℓ_2 convergence. Here we take a closer look at the trends of $\widehat{x}$ and $\widehat{Z}$, the solution to SDP1 (8) as n increases, and of v_1, the leading eigenvector of YY^T. In this experiment, we set $p \in \{20000, 50000, 80000\}$, and increase n. In the top panel of Fig. 2, which is for the unbalanced case of $w_1 = 0.7$, and $V_1 = V_2$, we plot $\theta_{\text{SDP 1}} := \angle(\widehat{x}, u_2)$ between $\widehat{x}$ and its reference vector u_2 as defined in Theorem 7 and Corollary 9. For SVD, $\theta_1 := \angle(v_1, \bar{v}_1)$ between v_1 and its reference $\bar{v}_1$, where $\bar{v}_1$ is as defined in Theorem 17. In this case, the angle $\angle(\bar{v}_1, u_2)$ between the two reference vectors is about 22 degrees (blue horizontal dashed line). We observe that as n increases, for both algorithms, the angles

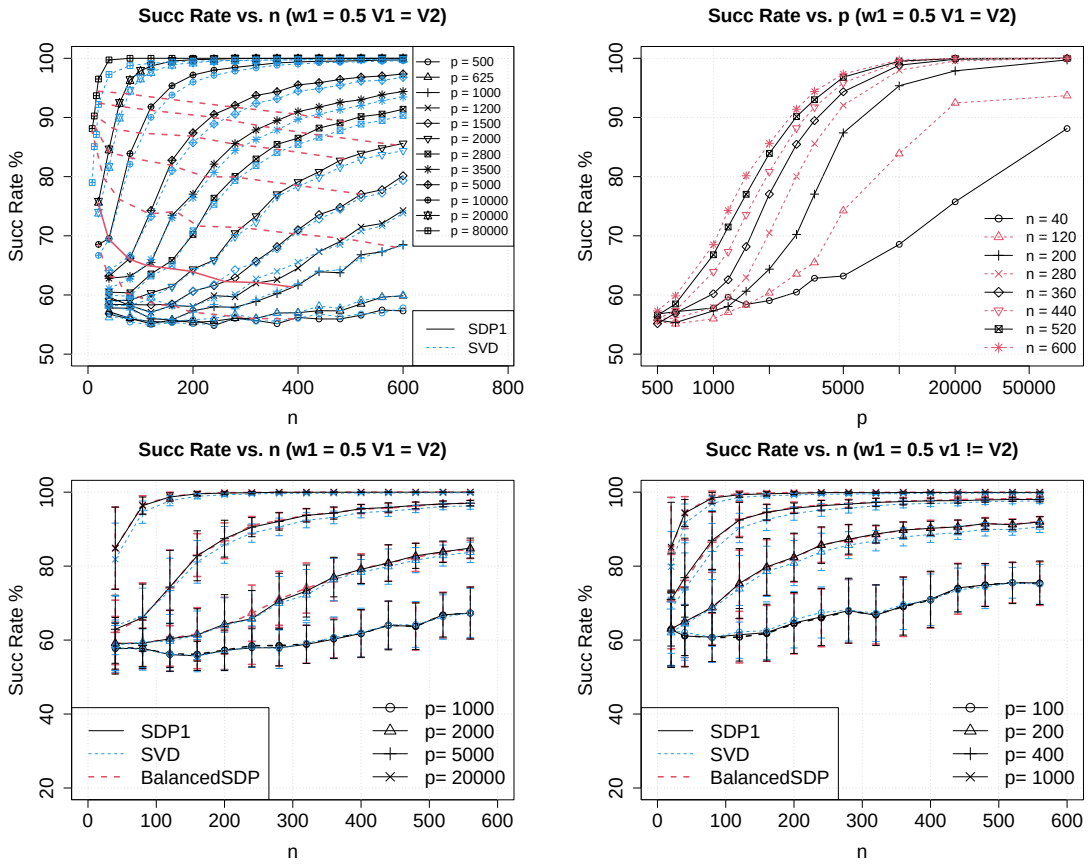


Figure 1: Top left: Success rates of SDP1 and SVD (balanced case and $V_1 = V_2$) for different p as n increases. Lines with the same markers are for the 2 estimators. Red lines highlight the success rates at different levels of $np\gamma^2$ ranging from 0.5 to 3.5 (bottom to top with a step of 0.5). The solid red line is for $np\gamma^2 = 1$. Top right: Same data as the top left plot, showing SDP1's success rate for different n as p increases (x-axis in log scale). Bottom row: Success rates of SDP1, SVD and BalancedSDP for two different variance profiles. Lines with the same markers are for the 3 estimators. Error bars represent 1 standard deviation.

$\theta_{\text{SDP 1}}$ and θ_1 decrease, but $\theta_{\text{SDP 1}}$ drops much faster and decreases to 0 when $n > 200$ for $p = 80,000$. We also show the angle $\phi = \angle(\widehat{x}, v_1)$ between the two leading eigenvectors $\widehat{x}$ and v_1, which largely remains flat across all n. In the bottom panel of Fig. 2, we see that for SDP1, all three metrics, namely $\sin(\theta_{\text{SDP 1}})$, $\|Z^* - \widehat{Z}\|_2/n$, and $\|Z^* - \widehat{Z}\|_F/n$, where $Z^* = u_2 u_2^T$, decrease as n increases, following an exponential decay in n as predicted by our theory in Theorem 10 and Corollary 19, where in each plot, p, γ are fixed. The gaps between the three curves for SDP1 shrink when p, n increase. For SVD, $\sin(\theta_1)$ also decreases as n increases, but at a slower rate of $1/n$, again as predicted by Theorem 17.

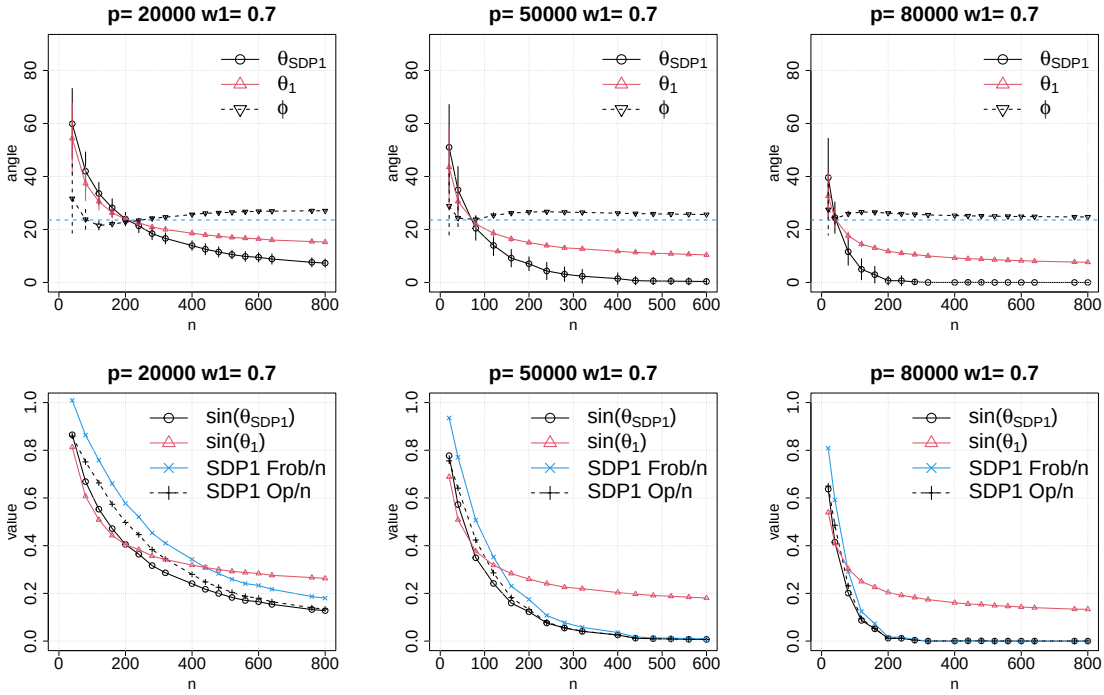


Figure 2: Unbalanced case $w_1 = 0.7$, $p \in \{20000, 50000, 80000\}$.

8. Proof of Theorem 10 and Lemma 15

Let $\lceil r_1/(2n) \rceil =: q < n$ and $r_1 := \left\| \widehat{Z} - Z^* \right\|_1$. Recall $\sigma^2_{\max} := \max_i \|H_i\|_2^2$. Then, we have by Lemma 22, on event $\mathcal{G}_1$,

$$\sup_{\widehat{Z} \in \mathcal{M}_{\text{opt}} \cap (Z^* + r_1 B_1^{n\times n})} \left| \langle YY^T - \mathbb{E}(YY^T), \widehat{Z} - Z^* \rangle \right| \le \frac{5}{6} \xi p\gamma \left\| \widehat{Z} - Z^* \right\|_1 \tag{61}$$
$$+ C' C_0 \sigma_{\max} n \sqrt{p\gamma} + C_1 C_0^2 \sigma^2_{\max} \sqrt{np} \left\lceil \frac{r_1}{2n} \right\rceil \sqrt{\log(2en/\lceil \frac{r_1}{2n} \rceil)}.$$

Let $C, C', C_1, c, c_1, c_2, \ldots$. be absolute positive constants. For the rest of the proof, we assume $\mathcal{G}_1 \cap \mathcal{G}_2$ holds, where $\mathbb{P}(\mathcal{G}_2 \cap \mathcal{G}_1) \ge 1 - \exp(-cn) - c/n^2$. By Lemmas 21, 24, and 15, (47), and (61), we have for all $\widehat{Z} \in \mathcal{M}_{\text{opt}} \cap (Z^* + r_1 B_1^{n\times n})$,

$$\begin{aligned} 0 &\le p\gamma w^2_{\min} \left\| Z^* - \widehat{Z} \right\|_1 \le \langle R, Z^* - \widehat{Z} \rangle \qquad (62) \\ &\le \left| \langle YY^T - \mathbb{E}(YY^T), \widehat{Z} - Z^* \rangle \right| + |F_2| + |F_3| \le (10/3)\xi p\gamma \left\| \widehat{Z} - Z^* \right\|_1 \\ &\quad + C' C_0 \sigma_{\max} q \sqrt{\log(\frac{2en}{q})} \cdot \left(n\sqrt{p\gamma} + C_0 \sigma_{\max} \sqrt{np} \right). \end{aligned}$$

Hence we can control $\xi \le w^2_{\min}/16$ and the lower bound in (19) so that each component F_2, F_3 is only a fraction of the signal at the level of $p\gamma w^2_{\min} \left\| (Z^* - \widehat{Z}) \right\|_1$. By moving

$(10/3)\xi p\gamma \left\|\widehat{Z} - Z^*\right\|_1 < \frac{5}{24} p\gamma w_{\min}^2 r_1$, where $\xi \le w_{\min}^2/16$, to the LHS of (62), we have

$$\begin{aligned} \frac{19}{24} p\gamma w_{\min}^2 r_1 &\le p\gamma w_{\min}^2 \left\|Z^* - \widehat{Z}\right\|_1 - (10/3)\xi p\gamma \left\|\widehat{Z} - Z^*\right\|_1 \\ &\le CC_0\sigma_{\max}\sqrt{\log(2en/q)} \cdot \big(2nq\sqrt{p\gamma} + C_0\sigma_{\max} q\sqrt{np}\big). \end{aligned} \tag{63}$$

Then q must satisfy one of the following two conditions to ensure (63): for $r_1 \le 2nq$,

1. Under the assumption that $\Delta^2 w_{\min}^4 = p\gamma w_{\min}^4 = \Omega(C_0^2\sigma_{\max}^2)$, suppose

$$\sqrt{p\gamma} w_{\min}^2 \le C_1 C_0 \sigma_{\max}\sqrt{\log(2en/q)},$$

   which implies that $\frac{cp\gamma w_{\min}^4}{C_0^2\sigma_{\max}^2} \le \log(2en/q)$ and hence $q \le n\exp\big(-\frac{c_1\Delta^2 w_{\min}^4}{C_0^2\sigma_{\max}^2}\big)$.

2. Alternatively, under the assumption that $np\gamma^2 w_{\min}^4 = \Omega(C_0^4)$, we require

$$p\gamma w_{\min}^2 r_1 \le 2qnp\gamma w_{\min}^2 \le C_2 C_0^2\sigma_{\max}^2\sqrt{\log(2en/q)}\, q\sqrt{np},$$

   and hence $np\gamma^2 w_{\min}^4 \le C_3(C_0\sigma_{\max})^4\log(2en/q)$, which implies that

$$\log(q/2en) \le -\frac{C' np\gamma^2 w_{\min}^4}{C_0^4\sigma_{\max}^4}, \quad \text{resulting in} \quad q \le n\exp\big(-\frac{c_2 np\gamma^2 w_{\min}^4}{C_0^4\sigma_{\max}^4}\big).$$

Putting things together, we have for s^2 as defined in (21),

$$\left\|Z^* - \widehat{Z}\right\|_1 := r_1 \le 2qn \le n^2\exp(-Cs^2 w_{\min}^4), \quad \text{and} \quad q \le n\exp(-Cs^2 w_{\min}^4). \quad \square$$

8.1 Proof of Lemma 15

We will prove that (33) holds for all $Z \in [-1,1]^{n\times n} \supset \mathcal{M}_{\text{opt}}$. Recall we have

$$R = \mathbb{E}(Y)\mathbb{E}(Y)^T = p\gamma \begin{bmatrix} w_2^2 E_{n_1} & -w_1 w_2 E_{n_1\times n_2} \\ -w_1 w_2 E_{n_2\times n_1} & w_1^2 E_{n_2} \end{bmatrix} =: p\gamma \begin{bmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{B}^T & \mathbb{C} \end{bmatrix}.$$

Now all entries of Z^*, Z belong to $[-1,1]$. Clearly, for the upper left and lower right diagonal blocks, denoted by $\mathbb{D} = \{\mathbb{A}, \mathbb{C}\}$, we have $Z^* - Z \ge 0$, since all entries of Z^* on these blocks are 1s. Similarly, for the off-diagonal blocks $\{\mathbb{B}, \mathbb{B}^T\}$, we have $Z - Z^* \ge 0$. Thus we have

$$\langle R, Z^* - Z\rangle/(p\gamma) \ge (w_2^2 \wedge w_1^2 \wedge w_1 w_2)\big(\sum_{(i,j)\in\mathbb{D}} (Z^* - Z)_{ij} + \sum_{(i,j)\in\{\mathbb{B},\mathbb{B}^T\}} (Z - Z^*)_{ij}\big)$$

The lemma thus holds since $\sum_{(i,j)\in\mathbb{D}}(Z^* - Z)_{ij} + \sum_{(i,j)\in\{\mathbb{B},\mathbb{B}^T\}}(Z - Z^*)_{ij} = \|Z - Z^*\|_1$. $\square$

9. Proof of Corollaries 9 and 19

Theorem 29. (Davis-Kahan) *For $A, M \in \mathbb{R}^{n\times n}$ being symmetric matrices and $E = M - A$. Let $\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_n(A)$ be eigenvalues of A with orthonormal eigenvectors $v_1, v_2, \ldots, v_n$, and $\lambda_1(M) \geq \lambda_2(M) \geq \ldots \geq \lambda_n(M)$ be eigenvalues of M with orthonormal eigenvectors $w_1, w_2, \ldots, w_n$ of M, with $\theta_i = \angle(v_i, w_i)$. Then*

$$\theta_i \sim \sin(\theta_i) \leq \frac{2\,\|E\|_2}{\mathit{gap}(i, A)} \quad \textit{where} \quad \mathit{gap}(i, A) = \min_{j\neq i} |\lambda_i(A) - \lambda_j(A)|\,. \tag{64}$$

Theorem 29 is a well-known result in perturbation theory. See Blum et al. (2009) for a proof. See also Theorem 4.5.5 Vershynin (2018) and Corollary 3 in Yu et al. (2015).

9.1 Proof of Corollary 9

The upper bound follows from (26) (Theorem 7) and Corollary 3 (Yu et al., 2015), while noting that the largest eigenvalue of $u_2 u_2^T$ is n while all others are 0, and hence the spectral gap in the sense of Theorem 29 equals n. Then, we have by Theorem 7,

$$\min_{\alpha=\pm 1} \left\|(\alpha\widehat{x} - u_2)/\sqrt{n}\right\|_2^2 \leq \frac{2^3 \left\|\widehat{Z} - u_2 u_2^T\right\|_2^2}{\mathsf{gap}(1, u_2 u_2^T)^2} \leq 2^3 \left\|\widehat{Z} - u_2 u_2^T\right\|_F^2 / n^2 \leq 2^3 4 K_G \xi / w_{\min}^2. \ \square$$

9.2 Proof of Corollary 19

The angle between $\widehat{x}/\sqrt{n}$ and $u_2/\sqrt{n}$ can be expressed as

$$\cos(\theta_{\mathrm{SDP}}) = \cos(\angle(\widehat{x}, u_2)) \quad = \quad \langle \widehat{x}, u_2\rangle / n \tag{65}$$

The upper bound on $\sin(\theta_{\mathrm{SDP}})$ follows from Theorems 10 and 29; by the Davis-Kahan Theorem, cf. Corollary 3 (Yu et al., 2015), we have with probability at least $1 - 2\exp(-cn) - 2/n^2$,

$$\begin{aligned} \min_{\alpha=\pm 1} \left\|(\alpha\widehat{x} - u_2)/\sqrt{n}\right\|_2 \quad &\leq \quad \frac{2^{3/2} \left\|\widehat{Z} - u_2 u_2^T\right\|_2}{\mathsf{gap}(1, u_2 u_2^T)} \leq 2^{3/2} \left\|\widehat{Z} - u_2 u_2^T\right\|_F / n \\ &\leq \quad 2^{3/2} (2 \left\|\widehat{Z} - u_2 u_2^T\right\|_1)^{1/2} / n \leq 4\exp(-c_0 s^2 w_{\min}^4 / 2). \ \square \end{aligned}$$

10. Proof of Theorem 17 and Corollary 19

We first state Propositions 30 and 31.

Proposition 30. *Let $YY^T = \sum_{j=1}^{n-1} \lambda_j v_j v_j^T$. Denote by $\widetilde{A} = YY^T - \lambda E_n$, then*

$$\widetilde{A} \quad := \quad YY^T - \lambda E_n = \sum_{j=1}^{n-1} \lambda_j v_j v_j^T + \frac{n\tau}{n-1} \mathbf{1}\mathbf{1}^T / n \succeq 0. \tag{66}$$

The leading eigenvector of $\widetilde{A}$ (resp. A and B) will coincide with that of YY^T with

$$\lambda_{\max}(\widetilde{A}) \quad = \quad \lambda_{\max}(YY^T) \geq n\tau/(n-1), \tag{67}$$

where strict inequality holds if and only if not all eigenvalues of YY^T are identical. Thus the symmetric matrices $A = \widetilde{A} + \lambda I_n$ and $B = A - \mathbb{E}\tau I_n$ also share the same leading eigenvector v_1 with YY^T, so long as not all eigenvalues of YY^T are identical, with $\lambda_{\max}(A) \geq \tau$.

Proposition 31 is also not surprising, since the sum of all off-diagonal entries of A is 0. We need it for the proof of Lemma 20 and Theorem 17.

Proposition 31. *Suppose that we observe one instance of the Gram matrix $\widehat{S}_n := XX^T$. Recall $\lambda = \frac{1}{\binom{n}{2}} \sum_{i<j} \langle Y_i, Y_j \rangle$ and $\tau = \frac{1}{n} \sum_{i=1}^n \langle Y_i, Y_i \rangle$. Then*

$$\langle YY^T, E_n \rangle = \mathbf{1}_n^T YY^T \mathbf{1}_n = 0, \quad \textit{since } YY^T = (I - P_1) XX^T (I - P_1). \tag{68}$$

Moreover, by construction, we have for A as defined in (7)*,*

$$\begin{aligned} \langle A, E_n \rangle &= \mathbf{1}_n^T YY^T \mathbf{1}_n - \lambda \langle (E_n - I_n), E_n \rangle = -\lambda n(n-1) = \mathrm{tr}(YY^T) \\ \textit{where } \lambda &= \frac{1}{n(n-1)} \sum_{i \neq j} \langle Y_i, Y_j \rangle = -\frac{\mathrm{tr}(YY^T)}{n(n-1)} = -\frac{\tau}{n-1}. \end{aligned}$$

In other words, we have $\langle A, P_1 \rangle = \langle A, \mathbf{1}_n \mathbf{1}_n^T / n \rangle = \mathrm{tr}(YY^T)/n =: \tau$.

Proof of Theorem 17. It is known that for any real symmetric matrix, there exists a set of n orthonormal eigenvectors. Recall that R is rank one with $\lambda_{\max}(R) = \mathrm{tr}(R) = w_1 w_2 np\gamma$ while $u_2 R u_2 / n = (4 w_1 w_2) w_1 w_2 np\gamma \leq \frac{1}{4} np\gamma$. Hence $u_2/\sqrt{n}$ coincides with $\bar{v}_1$ when $w_1 = w_2 = 1/2$. Hence for $\mathsf{gap}(1, R)$, as defined in Theorem 29, $\mathsf{gap}(1, R) = \lambda_{\max}(R) = w_1 w_2 np\gamma$. We check the claim that the leading eigenvector of B coincides with that of YY^T in Proposition 30. Clearly,

$$\cos(\theta_1) = \cos(\angle(v_1, \bar{v}_1)) \quad = \quad \langle v_1, \bar{v}_1 \rangle \tag{69}$$

and hence $\theta_1 = \arccos(\langle v_1, \bar{v}_1 \rangle)$. Hence we can use the first eigenvector of YY^T to partition the two groups of points in $\mathbb{R}^p$. To obtain an upper bound on $\sin(\theta_1)$, we apply the Davis-Kahan perturbation bound as follows. Since $v_1, \bar{v}_1$ are the leading eigenvectors of B and R respectively, (35) holds by Theorems 16 and 29:

$$\sin(\theta_1) \quad := \quad \sin(\angle(v_1, \bar{v}_1)) \leq \frac{2 \left\| B - R \right\|_2}{\lambda_{\max}(R)} = \frac{2 \left\| B - R \right\|_2}{w_1 w_2 np\gamma} \leq \frac{2\xi}{w_1 w_2}.$$

Moreover, we have (36) holds by Corollary 3 (Yu et al., 2015): since

$$\min_{\alpha = \pm 1} \left\| \alpha v_1 - \bar{v}_1 \right\|_2^2 \quad \leq \quad \left(\frac{2^{3/2} \left\| B - R \right\|_2}{w_1 w_2 np\gamma} \right)^2 \leq \frac{2^3 \xi^2}{(w_1 w_2)^2} =: \delta', \text{ where } \delta' = \frac{8\xi^2}{(w_1^2 w_2^2)} \leq \frac{c_2 \xi^2}{w_{\min}^2}.$$

The theorem thus holds. □

Proof of Proposition 30. Clearly, the additional terms involving E_n and I_n are either orthogonal to eigenvectors $v_1, \ldots, v_{n-1}$ of YY^T, or act as an identity map on the subspace spanned by $\{v_1, \ldots, v_{n-1}\}$. Now (66) holds since $\langle v_j, \mathbf{1}_n \rangle = 0$ for all j and hence

$\{v_1, \ldots, v_{n-1}, \mathbf{1}_n/\sqrt{n}\}$ forms the set of orthonormal eigenvectors for $\widetilde{A}$ (resp. A and B), and moreover, in view of Proposition 31,

$$-\lambda E_n \;\; = \;\; \frac{n\tau}{n-1} P_1 = \frac{n\tau}{n-1} \mathbf{1}_n \mathbf{1}_n^T/n, \quad \text{where} \quad \lambda = -\frac{\tau}{n-1}.$$

Since we have at most $n-1$ non-zero eigenvalues and they sum up to be $\mathrm{tr}(YY^T)$, we have

$$\lambda_{\max}(YY^T) \;\; \geq \;\; \mathrm{tr}(YY^T)/(n-1) = n\tau/(n-1),$$

where strict inequality holds when these eigenvalues are not all identical. Finally, (67) holds since $\lambda_1(\widetilde{A}) := \lambda_{\max}(YY^T)$ in view of the eigen-decomposition (66) and the displayed equation immediately above. Now for $A = YY^T - \lambda(E_n - I_n)$, we have $\mathrm{tr}(A) = \mathrm{tr}(YY^T)$, and hence $\lambda_{\max}(A) \geq \tau$. The extra terms $\propto I_n$ in A (resp. B) will not change the order of the sequence of eigenvalues for B (resp. A) with respect to that established for $\widetilde{A}$. Hence all symmetric matrices B, $\widetilde{A}$, and A share the same leading eigenvector v_1 with YY^T. $\square$

11. Proof of key lemmas for Theorems 7 and 10

11.1 Proof of Lemma 21

Proposition 32. *Let $Z^* = u_2 u_2^T$ be as defined in Lemma 1. Then for $P_2 = Z^*/n$,*

$$\mathrm{tr}(P_2(Z^* - \widehat{Z})) = \frac{1}{n}\mathrm{tr}(Z^*(Z^* - \widehat{Z})) = \frac{1}{n}\|Z^* - \widehat{Z}\|_1 \leq 2(n-1). \tag{70}$$

Proof of Lemma 21. Denote by $\widetilde{A} = YY^T - \lambda E_n$. We have for B as in (30),

$$\mathbb{E}B - R \;\; := \;\; \mathbb{E}(YY^T) - \mathbb{E}(Y)\mathbb{E}(Y)^T - \mathbb{E}\lambda(E_n - I_n) - \mathbb{E}\tau I_n, \tag{71}$$

where $R = \mathbb{E}(Y)\mathbb{E}(Y)^T$. Moreover, we have by (71),

$$\begin{aligned} M \;\; &:= \;\; \mathbb{E}(YY^T) - \lambda(E_n - I_n) - \mathbb{E}\tau I_n - \mathbb{E}(Y)\mathbb{E}(Y)^T \\ &= \;\; \big(\mathbb{E}B - R\big) - (\lambda - \mathbb{E}\lambda)(E_n - I_n). \end{aligned} \tag{72}$$

Clearly $\widehat{Z}, Z^* \in \mathcal{M}_{\mathrm{opt}}$ by definition. By optimality of $\widehat{Z} \in \mathcal{M}_{\mathrm{opt}}$, we have by (30) and (5),

$$\begin{aligned} \langle B, \widehat{Z} - Z^* \rangle \;\; &:= \;\; \langle YY^T - \lambda(E_n - I_n) - \mathbb{E}\tau I_n, \widehat{Z} - Z^* \rangle = \langle YY^T - \lambda E_n, \widehat{Z} - Z^* \rangle \geq 0, \\ \langle R, Z^* - \widehat{Z} \rangle \;\; &:= \;\; \langle -\mathbb{E}(Y)\mathbb{E}(Y)^T, \widehat{Z} - Z^* \rangle \\ &\leq \;\; \langle YY^T - \lambda(E_n - I_n) - \mathbb{E}\tau I_n, \widehat{Z} - Z^* \rangle - \langle \mathbb{E}(Y)\mathbb{E}(Y)^T, \widehat{Z} - Z^* \rangle \\ &= \;\; \langle YY^T - \mathbb{E}(YY^T), \widehat{Z} - Z^* \rangle + \langle M, \widehat{Z} - Z^* \rangle, \;\; \text{for } M \text{ as in (72)}. \end{aligned}$$

Thus the RHS of (45) holds. Now the LHS of (45) holds by Lemma 15. $\square$

11.2 Proof sketch of Lemma 22

Let $\Lambda = YY^T - \mathbb{E}YY^T$ and $\Psi := \mathbb{Z}\mathbb{Z}^T - \mathbb{E}\mathbb{Z}\mathbb{Z}^T$. The local analysis on (46) relies on the operator norm bound we obtained in Theorem 6 as already shown in Lemma 23. The key distinction from the global analysis is that we need to analyze the projection operators in

the following sense. Following the proof idea of Fei and Chen (2018), we define the following projection operator: for $P_2 = Z^*/n$,

$$\begin{aligned}\mathcal{P} : M \in \mathbb{R}^{n\times n} &\rightarrow P_2M + MP_2 - P_2MP_2\\ \mathcal{P}^\perp : M \in \mathbb{R}^{n\times n} &\rightarrow M - \mathcal{P}(M) = (I_n - P_2)M(I_n - P_2).\end{aligned}$$

We use the following decomposition to bound for all $\widehat{Z} \in \mathcal{M}_{\rm opt} \cap (Z^* + r_1 B_1^{n\times n})$, where $\mathcal{M}_{\rm opt} := \{Z : Z \succeq 0, \mathrm{diag}(Z) = I_n\} \subset [-1,1]^{n\times n}$:

$$\begin{aligned}\langle \Lambda, \widehat{Z} - Z^*\rangle &= \langle \Lambda, \mathcal{P}(\widehat{Z} - Z^*)\rangle + \langle \Lambda, \mathcal{P}^\perp(\widehat{Z} - Z^*)\rangle =: S_1(\widehat{Z}) + S_2(\widehat{Z}),\\ \text{where}\quad S_1(\widehat{Z}) &= \langle \Lambda, \mathcal{P}(\widehat{Z} - Z^*)\rangle = \langle \mathcal{P}(\Lambda), \widehat{Z} - Z^*\rangle,\\ S_2(\widehat{Z}) &= \langle \Lambda, \mathcal{P}^\perp(\widehat{Z})\rangle \le \|\Lambda\|_2 \|Z^* - \widehat{Z}\|_1/n.\end{aligned}$$

First, we control $S_1(\widehat{Z})$ uniformly for all $\widehat{Z} \in \mathcal{M}_{\rm opt} \cap (Z^* + r_1 B_1^{n\times n})$ in Lemma 33. We then obtain an upper bound for $S_2(\widehat{Z})$ in Lemma 34, uniformly for all $\widehat{Z} \in \mathcal{M}_{\rm opt}$. Lemmas 33 and 34 show that the total *noise* contributed by the second term, namely, $S_2(\widehat{Z})$ and a portion of that by $S_1(\widehat{Z})$ is only a fraction of the signal strength on the LHS of (45). We prove Lemma 33 in Section E. Lemma 22 follows from Lemmas 33 and 34. The proof of Lemma 34 follows from Theorem 6 and arguments in Fei and Chen (2018); we include the proof for self-containment.

Lemma 33. *Suppose all conditions in Theorem 6 hold. Let* $\sigma_{\max} := \max_i \|H_i\|_2$ *be as in* (51)*. Then, with probability at least* $1 - c'/n^2$*, for some absolute constants* c', C*,*

$$\sup_{\widehat{Z}\in\mathcal{M}_{opt}\cap(Z^*+r_1B_1^{n\times n})} S_1(\widehat{Z}) \le \frac{2}{3}\xi p\gamma\|\widehat{Z} - Z^*\|_1 + C\sigma_{\max}\lceil\frac{r_1}{2n}\rceil\sqrt{\log(2en/\lceil\frac{r_1}{2n}\rceil)}\big(n\Delta + \sigma_{\max}\sqrt{np}\big).$$

Lemma 34. *Suppose all conditions in Theorem 6 hold. Let* $\xi \le w_{\min}^2/16$*. Then, with probability at least* $1 - \exp(-cn)$*, for all* $\widehat{Z} \in \mathcal{M}_{opt}$*,*

$$S_2(\widehat{Z}) \le \|YY^T - \mathbb{E}(YY^T)\|_2\|Z^* - \widehat{Z}\|_1/n \le \xi p\gamma \left\|\widehat{Z} - Z^*\right\|_1/6.$$

Proof Since $\widehat{Z} \succeq 0$, we have by properties of projection,

$$\mathcal{P}^\perp(\widehat{Z}) = (I_n - P_2)(\widehat{Z})(I_n - P_2) \succeq 0 \quad \text{and} \quad \mathcal{P}^\perp(Z^*) = (I_n - P_2)Z^*(I_n - P_2) = 0,$$

where $P_2 = Z^*/n$. Thus we have for $\mathrm{tr}(\widehat{Z}) = \mathrm{tr}(Z^*) = n$,

$$\begin{aligned}\left\|\mathcal{P}^\perp(\widehat{Z})\right\|_* &:= \left\|(I_n - P_2)(\widehat{Z})(I_n - P_2)\right\|_* = \mathrm{tr}((I_n - P_2)(\widehat{Z})(I_n - P_2))\\ &= \mathrm{tr}(Z^*(Z^* - \widehat{Z}))/n = \left\|Z^* - \widehat{Z}\right\|_1/n,\end{aligned}$$

where $\|\cdot\|_*$ denotes the nuclear norm; thus we have

$$S_2(\widehat{Z}) = \langle \Lambda, \mathcal{P}^\perp(\widehat{Z})\rangle \le \|\Lambda\|_2 \left\|\mathcal{P}^\perp(\widehat{Z})\right\|_* = \left\|YY^T - \mathbb{E}YY^T\right\|_2 \left\|Z^* - \widehat{Z}\right\|_1/n.$$

The lemma thus holds in view of Theorem 6. ■

11.3 Proof of Lemma 23

To be fully transparent, we now decompose the bias term into three components. Intuitively, W_0, W_2 and $\mathbb{W}$ arise due to the imbalance in variance profiles. Hence (A2) is needed to control this bias. We prove Proposition 35 in Section D.5 Zhou (2026).

Proposition 35. (Bias decomposition) *Denote by*

$$W_0 \;=\; (V_1 - V_2)\begin{bmatrix} w_2 I_{n_1} & 0 \\ 0 & -w_1 I_{n_2} \end{bmatrix} \quad \textit{and} \quad W_2 := \frac{V_1 - V_2}{n}\begin{bmatrix} E_{n_1} & 0 \\ 0 & -E_{n_2} \end{bmatrix}. \tag{73}$$

Then for W_0, W_2 as defined in (73),

$$\mathbb{E}B - R \;=\; W_0 - \mathbb{W} - \mathrm{tr}(R)/(n-1)(I_n - E_n/n), \quad \textit{where} \tag{74}$$

$$\mathbb{W} \;:=\; W_2 + \frac{(V_1 - V_2)(w_2 - w_1)}{n} E_n, \quad \textit{and when } V_1 = V_2, W_0 = \mathbb{W} = 0. \tag{75}$$

Proof of Lemma 23. Now we have by Proposition 35, for $P_1 := E_n/n$,

$$\langle \mathbb{E}B - R, \widehat{Z} - Z^* \rangle \;=\; \langle W_0 - \mathbb{W}, \widehat{Z} - Z^* \rangle - \langle I_n - P_1, \widehat{Z} - Z^* \rangle \mathrm{tr}(R)/(n-1),$$

where W_0 is a diagonal matrix as in (73). For the optimal solution $\widehat{Z}$ of SDP1, we have

$$\langle W_0, \widehat{Z} - Z^* \rangle \;=\; \langle W_0, \mathrm{diag}(\widehat{Z} - Z^*) \rangle = 0, \text{ where } \quad \mathrm{diag}(\widehat{Z}) = \mathrm{diag}(Z^*) = I_n, \tag{76}$$

$$\left| \langle \mathbb{W}, \widehat{Z} - Z^* \rangle \right| \;\le\; \|\mathbb{W}\|_{\max} \|\widehat{Z} - Z^*\|_1 \le 2\,|V_1 - V_2| \|(Z^* - \widehat{Z})\|_1 / n. \tag{77}$$

Hence, by the triangle inequality and the fact that $\mathrm{tr}(R)/n = p\gamma w_1 w_2$ and (A2),

$$\begin{aligned} \langle \mathbb{E}B - R, \widehat{Z} - Z^* \rangle \;&\le\; \left| \langle \mathbb{W}, \widehat{Z} - Z^* \rangle \right| + \left| \frac{\mathrm{tr}(R)}{n(n-1)} \langle E_n, \widehat{Z} - Z^* \rangle \right| \\ &\le\; p\gamma \|\widehat{Z} - Z^*\|_1 \big(\xi + 1/(4(n-1))\big) \le 2\xi p\gamma \|\widehat{Z} - Z^*\|_1, \end{aligned} \tag{78}$$

where $2\,|V_1 - V_2|/n \le \xi p\gamma$. The lemma thus holds. □

Remark 36. (Balanced Partitions) *Suppose that $w_1 = w_2$, then we may hope to obtain a tighter bound for $\langle \mathbb{E}B - R, \widehat{Z} - Z^* \rangle$. We conjecture that* (77) *can be substantially tightened for the balanced case, where $\mathbb{W} = W_2$. Now by symmetry of matrix $\widehat{Z} - Z^*$,*

$$\mathbf{1}_n^T (\widehat{Z} - Z^*) u_2 \;=\; \sum_{i \in \mathcal{C}_1} \sum_{j \in \mathcal{C}_1} (\widehat{Z} - Z^*)_{ij} - \sum_{i \in \mathcal{C}_2} \sum_{j \in \mathcal{C}_2} (\widehat{Z} - Z^*)_{ij}, \tag{79}$$

where the two off-diagonal block-wise sums have been cancelled due to symmetry. Hence for $w_1 = w_2$, we have (a) $\frac{1}{n}\mathbf{1}_n^T Z^* u_2 = \mathbf{1}_n^T u_2 = 0$, and (b) for $\mathbb{W} = W_2$,

$$\left| \langle \mathbb{W}, \widehat{Z} - Z^* \rangle \right| = (|V_1 - V_2|/n) \left| \langle \begin{bmatrix} E_{n_1} & 0 \\ 0 & -E_{n_2} \end{bmatrix}, \widehat{Z} - Z^* \rangle \right| \tag{80}$$

$$= \; (|V_1 - V_2|/n) \left| \mathbf{1}_n^T (\widehat{Z} - Z^*) u_2 \right| \le |V_1 - V_2| \, \|Z^* - \widehat{Z}\|_2 / 2, \tag{81}$$

where recall $\mathrm{diag}(\widehat{Z}) = \mathrm{diag}(Z^*) = I_n$ and (81) *holds by symmetry and (a); cf. Lemma 39. Moreover, since $(Z^*_{ij} - \widehat{Z}_{ij}) \ge 0, \forall (i,j) \in \mathcal{C}_k$, the two block-wise sums $S_k := \sum_{i,j \in \mathcal{C}_k} (Z^*_{ij} - \widehat{Z}_{ij}) \ge 0$, $k = 1, 2$, are both nonnegative in* (79)*. Hence we expect the difference between the two block-wise sums $|S_1 - S_2|$ to be small for* (80)*. Formally, we have Conjecture 37.*

Conjecture 37. (Balanced partitions with SDP1) *For the balanced case where $n_1 = n_2$, the bias term* (78) *essentially becomes a small order term compared with the variance.*

12. Analysis of the BalancedSDP for balanced partitions: Bias free

Suppose $w_1 = w_2 = 0.5$. Since $\frac{1}{n}\mathbf{1}_n^T Z^* u_2 = \mathbf{1}_n^T u_2 = 0$, we have by symmetry of $\widehat{Z}$ and (81),

$$\left|\langle \mathbb{W}, \widehat{Z} - Z^* \rangle\right| \;=\; \left|\langle W_2, \widehat{Z}\rangle\right| = [|V_1 - V_2|/n] \left|u_2^T \widehat{Z} \mathbf{1}_n\right|.$$

By Rayleigh's Principle, Theorem 38 holds for the Rayleigh quotient $R(x) = \frac{x^T Z x}{x^T x}$, where $x \neq 0$. Recall the minimum of the Rayleigh quotient is the smallest eigenvalue $\lambda_{\min}(Z)$.

Theorem 38. *Let $Z \succeq 0$ and $\langle E_n, Z\rangle = 0$. Then $R(x)$ reaches the minimum at the $\mathbf{1}_n$, that is, $R(\mathbf{1}_n) := \min_{x\in\mathbb{R}^n, x\neq 0} x^T Z x / x^T x = \lambda_{\min}(Z) = 0$.*

Lemma 39. *Suppose $w_1 = w_2 = 0.5$. Let $\widehat{Z}$ be an optimal solution to (9). Then*

$$\langle \begin{bmatrix} E_{n_1} & 0 \\ 0 & -E_{n_2} \end{bmatrix}, Z^* - \widehat{Z}\rangle = \mathbf{1}_n^T(\widehat{Z} - Z^*)u_2 = \mathbf{1}_n^T \widehat{Z} u_2 = 0. \tag{82}$$

As a consequence, we have $\langle W_2, \widehat{Z} - Z^*\rangle = 0$ and $\langle \mathbb{E}B - R, \widehat{Z} - Z^*\rangle = 0$.

We rewrite (BalancedSDP) (9) as

$$\text{maximize} \qquad \langle YY^T - \lambda E_n, Z\rangle \quad \text{subject to} \quad Z \in \mathcal{M}_{\text{opt}}^+ \;\text{ for} \tag{83}$$

$$\mathcal{M}_{\text{opt}}^+ \;=\; \{Z : Z \succeq 0, \text{diag}(Z) = I_n, \langle Z, E_n\rangle = 0\}, \text{ where } E_n = \mathbf{1}_n\mathbf{1}_n^T. \tag{84}$$

Here the trace term $\langle \lambda E_n, Z\rangle$ in (83) does not play any role in optimization. This eliminates the F_2 and F_3 terms in view of Lemma 39. Hence, we now have the updated Elementary Inequality, replacing Lemma 21. The proof then follows that of Lemma 21, except that now $F_2, F_3 = 0$ due to the new constraint $\langle E_n, Z\rangle = 0$. The rest of the proof for Theorem 11 follows that of Theorem 10 and hence is omitted.

Lemma 40. (Elementary inequality: balanced partitions) *Suppose $w_1 = w_2 = 0.5$. Clearly $Z^* = u_2 u_2^T \in \mathcal{M}_{opt}^+$. By optimality of $\widehat{Z} \in \mathcal{M}_{opt}^+$, we have $\langle \widehat{Z}, I_n\rangle = \langle I_n, Z^*\rangle = n$ and $\langle \widehat{Z}, E_n\rangle = \langle E_n, Z^*\rangle = 0$, and moreover,*

$$p\gamma \|\widehat{Z} - Z^*\|_1/4 \leq \langle R, Z^* - \widehat{Z}\rangle \leq \langle YY^T - \mathbb{E}(YY^T), \widehat{Z} - Z^*\rangle. \tag{85}$$

Proof of Lemma 39. Denote by $\mathbf{1}_{\mathcal{C}_k} \in \{0,1\}^n$ the group indicator vector for $\mathcal{C}_k, k = 1, 2$ and clearly $\sum_k \mathbf{1}_{\mathcal{C}_k} = \mathbf{1}_n$. As a result of Theorem 38, we have $\widehat{Z} v_n = 0 v_n$ for $v_n := \mathbf{1}_n/\sqrt{n}$. Thus $\lambda_n(\widehat{Z}) := \lambda_{\min}(\widehat{Z}) = 0$, and for all other eigenvectors $v_i, i = 1, \ldots, n-1$ of $\widehat{Z}$, we have $\langle v_i, \mathbf{1}_n\rangle = 0$. Now split each v_i into two parts according to $\mathcal{C}_k, k = 1, 2$ to obtain

$$\forall j = 1, \ldots, n-1, \quad \mathbf{1}_{\mathcal{C}_1}^T v_j := \sum_{\ell\in\mathcal{C}_1} v_{j,\ell} = -\sum_{\ell\in\mathcal{C}_2} v_{j,\ell} = -\mathbf{1}_{\mathcal{C}_2}^T v_j.$$

Thus we have for $\widehat{Z} = \sum_{j=1}^n \lambda_j(\widehat{Z}) v_j v_j^T$, where $\lambda_1(\widehat{Z}) \geq \ldots \geq \lambda_n(\widehat{Z}) = 0$, and by symmetry,

$$\mathbf{1}_n^T \widehat{Z} u_2 = \sum_{i\in\mathcal{C}_1}\sum_{j\in\mathcal{C}_1} \widehat{Z}_{ij} - \sum_{i\in\mathcal{C}_2}\sum_{j\in\mathcal{C}_2} \widehat{Z}_{ij} = \mathbf{1}_{\mathcal{C}_1}^T \widehat{Z} \mathbf{1}_{\mathcal{C}_1} - \mathbf{1}_{\mathcal{C}_2}^T \widehat{Z} \mathbf{1}_{\mathcal{C}_2}$$

$$= \mathbf{1}_{\mathcal{C}_1}^T (\sum_{j=1}^{n-1} \lambda_j(\widehat{Z}) v_j v_j^T) \mathbf{1}_{\mathcal{C}_1} - \mathbf{1}_{\mathcal{C}_2}^T (\sum_{j=1}^{n-1} \lambda_j(\widehat{Z}) v_j v_j^T) \mathbf{1}_{\mathcal{C}_2} = 0.$$

Finally, we have by (78), $\langle \mathbb{E}B - R, \widehat{Z} - Z^* \rangle = 0$, where the first component $= 0$ for $\mathbb{W} = W_2$, which follows from (82), while for the second component, we have $\langle E_n, \widehat{Z} \rangle = 0$ by definition of (9) and $\langle E_n, Z^* \rangle = u_2^T E_n u_2 = 0$ by definition of Z^*. □

13. Conclusion

The present work aims to illuminate the geometric and probabilistic features for population clustering through the MAXCUT framework. Centering the data matrix X plays a key role in the statistical analysis and in understanding the roles of sample size lower bounds (on n, p and their tradeoffs) for partial recovery of the clusters using the local and global analyses. With new results on concentration of measure bounds on $\|YY^T - \mathbb{E}YY^T\|$, we can simultaneously analyze SDP1 as well as the spectral algorithm based on the leading eigenvector of YY^T; cf. Theorem 17 and Corollary 18, where we obtain error bounds similar to those in Theorem 7. Importantly, because of these new concentration of measure bounds, our theory works for the small sample and low SNR ($n < p$ and $s^2 = o(\log n)$) cases for both SDP1 (in its local and global analyses) and the spectral algorithm, which is perhaps the more surprising case. Using the local analysis, we discover additional debiasing properties of SDP1 and its variation for balanced partitions, and hence we remove (A2) entirely from Theorem 11. Importantly, we show that BalancedSDP (9) has an inherent tolerance on arbitrary variance discrepancy between the two component distributions in case $n_1 = n_2$, while achieving optimal clustering for balanced partitions. Consequently, for balanced partitions, our result in Theorem 11 for BalancedSDP (9) is substantially stronger than those in the literature in the sense that we now simultaneously eliminate the bounded trace-difference assumption (A2), as well as the need for an explicit debiasing step. Such explicit debiasing steps typically come with an additional computational cost; see for example Royer (2017).

Acknowledgments

The author would like to thank Alan Frieze and Mark Rudelson for helpful discussions on an earlier draft of the paper, and the Simons Institute for the Theory of Computing for the generous support and hospitality. The author thanks the Associate Editor and two anonymous referees for their valuable comments and suggestions.

Appendix A. Proof of Theorem 4 and Lemma 5

In the rest of this section, we prove Theorem 4. The proof may be of independent interest. First, we prove some more general results, from which Lemma 5 follows. We generate $\mathbb{Z}$ according to Definition 3:

$$\mathbb{Z}_j \;=\; H_j W_j \in \mathbb{R}^p, \;\text{ and hence } \mathbb{Z} = \sum_{j=1}^n e_j W_j^T H_j^T = \sum_{j=1}^n \mathrm{diag}(e_j)\mathbb{W} H_j^T,$$

where $W_1^T, \ldots, W_n^T \in \mathbb{R}^m$ are independent, mean-zero, isotropic row vectors of $\mathbb{W} = (w_{jk})$, where we assume that coordinates w_{jk} are also independent with $\max_{j,k} \|w_{jk}\|_{\psi_2} \le K$.

Hence $\mathbb{Z}_1, \ldots, \mathbb{Z}_n \in \mathbb{R}^p$ are independent, mean zero, sub-gaussian random vectors with covariance $\mathsf{Cov}(\mathbb{Z}_j) = \mathbb{E}(\mathbb{Z}_j \mathbb{Z}_j^T)$ for each $j \in [n]$ with the following properties.

First, $\forall j \in [n]$, denote by $\Sigma_j = \mathsf{Cov}(Z_j)$, where

$$\mathsf{Cov}(\mathbb{Z}_j) := \mathbb{E}(\mathbb{Z}_j \mathbb{Z}_j^T) = H_j \mathbb{E}(W_j W_j^T) H_j^T = H_j H_j^T$$

and $\mathrm{tr}(\Sigma_j) = \mathrm{tr}(H_j H_j^T) = \|H_j\|_F^2$. Here by definition of (17), we have for any $v \in \mathbb{R}^p$,

$$\|\langle \mathbb{Z}_j, v \rangle\|_{L_2} = (v^T \mathbb{E}[\mathbb{Z}_j \mathbb{Z}_j^T] v)^{1/2} = (v^T \mathsf{Cov}(\mathbb{Z}_j) v)^{1/2} = \left\|H_j^T v\right\|_2 . \tag{86}$$

Now $\forall h \in \mathbb{S}^{p-1}$, we have by definition of (13) and (86),

$$\begin{aligned} \|\langle \mathbb{Z}_j, h \rangle\|_{\psi_2} &= \|\langle H_j W_j, h \rangle\|_{\psi_2} = \left\|\langle W_j, H_j^T h \rangle\right\|_{\psi_2} \\ &\leq \|W_j\|_{\psi_2} \left\|H_j^T h\right\|_2 = \|W_j\|_{\psi_2} \|\langle \mathbb{Z}_j, h \rangle\|_{L_2} ; \end{aligned} \tag{87}$$

Thus we have shown that (16) and (17) hold for C_0 as in (88) for the independent, sub-gaussian random vectors $\mathbb{Z}_1, \ldots, \mathbb{Z}_n \in \mathbb{R}^p$ generated according to Definition 3, since for some universal constant C,

$$K \leq \max_j \|W_j\|_{\psi_2} \leq CK \ \forall j, \text{ and } \ C_0 = \max_j \|W_j\|_{\psi_2} , \tag{88}$$

by definition of $K = \max_{j,k} \|w_{jk}\|_{\psi_2}$ in (14). See Definition 3.4.1 and Lemma 3.4.2 (Vershynin, 2026) for the proof of (88).

Proof of Theorem 4. Let $e_1, \ldots, e_n$ be the canonical basis of $\mathbb{R}^n$. By Definition 3,

$$\begin{aligned} \mathrm{vec}\{\mathbb{Z}\} &= \sum_{j=1}^n \mathrm{vec}\{\mathrm{diag}(e_j) \mathbb{W} H_j^T\} = \sum_{j=1}^n H_j \otimes \mathrm{diag}(e_j) \mathrm{vec}\{\mathbb{W}\} \\ &=: L \mathrm{vec}\{\mathbb{W}\} \in \mathbb{R}^{np} \quad \text{where} \quad L := \sum_{j=1}^n H_j \otimes \mathrm{diag}(e_j) \in \mathbb{R}^{np \times mn}. \end{aligned} \tag{89}$$

Following Definition 3, we have $\mathsf{Cov}(\mathrm{vec}\{\mathbb{Z}\}) = LL^T := \sum_{i=1}^n H_i H_i^T \otimes \mathrm{diag}(e_i)$. On the other hand, we have for $\mathbb{W} \in \mathbb{R}^{n \times m}$, $\mathbb{Z}_j = H_j W_j$ and hence

$$\begin{aligned} \mathbb{Z}^T &= [\mathbb{Z}_1, \ldots, \mathbb{Z}_n] = \sum_{j=1}^n \mathbb{Z}_j \otimes e_j^T = \sum_{j=1}^n H_j \mathbb{W}^T \mathrm{diag}(e_j), \\ \text{and hence } \mathrm{vec}\{\mathbb{Z}^T\} &= \mathrm{vec}\{\sum_{j=1}^n H_j \mathbb{W}^T \mathrm{diag}(e_j)\} = \sum_{j=1}^n \mathrm{vec}\{H_j \mathbb{W}^T \mathrm{diag}(e_j)\} \\ &= \sum_{j=1}^n (\mathrm{diag}(e_j) \otimes H_j) \mathrm{vec}\{\mathbb{W}^T\} =: R \mathrm{vec}\{\mathbb{W}^T\}. \end{aligned} \tag{90}$$

Then there exist some permutation matrices P, Q such that

$$L = \sum_{i=1}^n H_i \otimes \mathrm{diag}(e_i) = P^T \big(\sum_{i=1}^n \mathrm{diag}(e_i) \otimes H_i \big) Q =: P^T R Q \tag{91}$$

where $H_j H_j^T$ denotes the covariance matrix for each row vector $\mathbb{Z}_j$, $j \in [n]$. We now show (91) with an explicit construction. There exist permutation matrices P, Q such that

$$\begin{aligned} \mathrm{vec}\{\mathbb{Z}^T\} &= P\mathrm{vec}\{\mathbb{Z}\} \text{ and } \mathrm{vec}\{\mathbb{W}^T\} = Q\mathrm{vec}\{\mathbb{W}\} \qquad (92) \\ \text{and hence by (89), } \mathrm{vec}\{\mathbb{Z}^T\} &= P\mathrm{vec}\{\mathbb{Z}\} = PL\mathrm{vec}\{\mathbb{W}\} \end{aligned}$$

On the other hand, we have by (90) and (92)

$$\mathrm{vec}\{\mathbb{Z}^T\} = R\mathrm{vec}\{\mathbb{W}^T\} = RQ\mathrm{vec}\{\mathbb{W}\}.$$

This shows that for $P^T = P^{-1}$, $PL = RQ$ and hence $L = P^T RQ$, and hence (91) holds. See Lemma 4.3.1 and Corollary 4.3.10 Horn and Johnson (1991). First we rewrite the quadratic form as follows: for any matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} \left|\sum_{i=1}^n \sum_{j\neq i}^n \langle \mathbb{Z}_i, \mathbb{Z}_j\rangle a_{ij}\right| &= \left|\sum_{i=1}^n \sum_{j\neq i}^n a_{ij} \sum_{k=1}^p z_{ik} z_{jk}\right| \\ &= \mathrm{vec}\{\mathbb{Z}\}^T \widetilde{A} \mathrm{vec}\{\mathbb{Z}\} = \mathrm{vec}\{\mathbb{W}\}^T L^T \widetilde{A} L \mathrm{vec}\{\mathbb{W}\}, \end{aligned}$$

where $\widetilde{A} \in \mathbb{R}^{np\times np}$ is a block-diagonal matrix with $\mathrm{diag}(\widetilde{A}) = 0$, and p identical blocks $\widetilde{A}^k = \mathrm{offd}(A), \forall k \in [p]$ of size $n \times n$ along the main diagonal, where $\|\mathrm{offd}(A)\|_2 \leq \|A\|_2 + \|\mathrm{diag}(A)\|_2 \leq 2\|A\|_2$. We now compute $\|L\|_2 = \|R\|_2 = \|\sum_{i=1}^n \mathrm{diag}(e_i) \otimes H_i\|_2 \leq \max_i \|H_i\|_2$, and

$$\begin{aligned} \left\|L^T \widetilde{A} L\right\|_2 &\leq \left\|\sum_{i=1}^n H_i \otimes \mathrm{diag}(e_i)\right\|_2^2 \left\|\widetilde{A}\right\|_2 \leq \max_i \|H_i\|_2^2 \left\|\widetilde{A}\right\|_2, \text{ and} \\ \left\|L^T \widetilde{A} L\right\|_F &\leq \|L\|_2^2 \left\|\widetilde{A}\right\|_F \leq \max_i \|H_i\|_2^2 \left\|\widetilde{A}\right\|_F \leq \max_i \|H_i\|_2^2 \sqrt{p}\, \|A\|_F, \end{aligned}$$

where we use the property of block-diagonal matrix for $R = \sum_{i=1}^n \mathrm{diag}(e_i) \otimes H_i$, which is also known as a direct sum over $H_i, i = 1, \ldots, n$. Hence for any $t > 0$, by the Hanson-Wright inequality, cf. Theorem 1.1 Rudelson and Vershynin (2013), for $C_H^2 := K^2(\max_i \|H_i\|_2^2)$,

$$\begin{aligned} &\mathbb{P}\left(\left|\mathrm{vec}\{\mathbb{Z}\}^T \widetilde{A} \mathrm{vec}\{\mathbb{Z}\}\right| > t\right) = \mathbb{P}\left(\left|\mathrm{vec}\{\mathbb{W}\}^T L^T \widetilde{A} L \mathrm{vec}\{\mathbb{W}\}\right| > t\right) \\ &\leq 2\exp\left(-c\min\left(t^2/(C_H^4 \left\|\widetilde{A}\right\|_F^2), t/(C_H^2 \left\|\widetilde{A}\right\|_2)\right)\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{C_H^4 p \|A\|_F^2}, \frac{t}{C_H^2 \|A\|_2}\right)\right). \end{aligned}$$

Finally, for all $j \in [n]$, we have by (86) and (87),

$$\|\mathbb{Z}_j\|_{\psi_2} := \sup_{h\in\mathbb{S}^{p-1}} \|\langle \mathbb{Z}_j, h\rangle\|_{\psi_2} \leq \|W_j\|_{\psi_2} \|H_j\|_2 \leq CK \max_j \|H_j\|_2. \qquad (93)$$

This shows that (15) also holds. The theorem thus holds. □

A.1 Proof of Lemma 5

Suppose that we allow each population to have distinct covariance structures and $Z_j = H_i W_j$ for all $j \in \mathcal{C}_i, \forall i \in \{1, 2\}$. Then $\forall j \in \mathcal{C}_i, i = 1, 2$,

$$\mathsf{Cov}(\mathbb{Z}_j) := \mathbb{E}(\mathbb{Z}_j \mathbb{Z}_j^T) = H_i \mathbb{E}(W_j W_j^T) H_i^T = H_i H_i^T$$

and $V_i := \mathrm{tr}(\Sigma_i) = \mathrm{tr}(H_i H_i^T) = \langle H_i, H_i\rangle = \|H_i\|_F^2$. The rest follows from (87), (93), and (88). □

Appendix B. Preliminary results

We first present preliminary results for the global and local analyses in this section. These results including Theorem 8 are initially proved in Zhou (2023). For completeness, we first define the matrix cut norm as in Frieze and Kannan (1999).

Definition 41. (Matrix cut norm) *Frieze and Kannan (1999) For a matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$, we denote by $\|(a_{ij})\|_{\infty\to 1}$ its $\ell_\infty \to \ell_1$ norm, which is*

$$\|(a_{ij})\|_{\infty\to 1} = \max_{\|s\|_\infty \le 1} \|As\|_1 = \max_{s,t\in\{-1,1\}^n} \langle A, st^T\rangle.$$

This norm is equivalent to the matrix cut norm defined as: for $A \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} \|A\|_\square &= \max_{I\subset[n],J\subset[n]} \left|\sum_{i\in I}\sum_{j\in J} a_{i,j}\right|, \quad \textit{and hence} \\ \|A\|_{\infty\to 1} &= \max_{x,y\in\{-1,1\}^n} \sum_{i=1}^n\sum_{j=1}^n a_{ij}x_iy_i \le \|x\|_2 \|y\|_2 \|A\|_2 \le n\|A\|_2 . \end{aligned}$$

Recall $\sigma^2_{\max} := \max_j \|\mathsf{Cov}(\mathbb{Z}_j)\|_2$. Let $\mu := (\mu^{(1)} - \mu^{(2)})/\sqrt{p\gamma} \in \mathbb{S}^{p-1}$ as in (18). Lemma 42 holds for $\mathbb{Z}$ as in Definition 3, where $\max_i \|R_i\mu\|_2^2$ in (95) and (96) is understood to be replaced by $\max_{j\in[n]} \left\|H_j^T\mu\right\|_2^2$.

Lemma 42 (Projection for anisotropic subgaussian random vectors). *Let X, C_0 and K be as in Lemma 5. In particular, $K \le C_0 = \max_j \|W_j\|_{\psi_2} \le CK$ for an absolute constant C, by definition of $K = \max_{j,k}\|w_{jk}\|_{\psi_2}$ in (14). Let Y be as in (6). Suppose all conditions in Lemma 28 hold. Then for $R_i = H_i^T$,*

$$\begin{aligned} \|\langle \mathbb{Z}_j, \mu\rangle\|_{\psi_2} &\le C_0 \|\langle \mathbb{Z}_j, \mu\rangle\|_{L_2} = C_0\sqrt{\mu^T \mathsf{Cov}(\mathbb{Z}_j)\mu} \qquad (94) \\ \textit{where } \sqrt{\mu^T \mathsf{Cov}(\mathbb{Z}_j)\mu} &= \sqrt{\mu^T H_i H_i^T \mu} = \|R_i\mu\|_2 \quad \textit{for each } j \in \mathcal{C}_i, i = 1, 2, \end{aligned}$$

where $R_i = H_i^T$. Thus for any $t > 0$, for some absolute constants c, c', we have for each $v \in \mathbb{S}^{n-1}$ and $u = (u_1, \ldots, u_n) \in \{-1, 1\}^n$ the following tail bounds:

$$\mathbb{P}\left(\left|\sum_{i=1}^n v_i\langle \mathbb{Z}_i, \mu\rangle\right| \ge t\right) \le 2\exp\left(-\frac{ct^2}{(C_0 \max_i \|R_i\mu\|_2)^2}\right), \quad \textit{and} \qquad (95)$$

$$\mathbb{P}\left(\left|\sum_{i=1}^n u_i\langle \mathbb{Z}_i, \mu\rangle\right| \ge t\right) \le 2\exp\left(-\frac{c't^2}{n(C_0 \max_i \|R_i\mu\|_2)^2}\right). \qquad (96)$$

Proof First, $\mathbb{Z}_j = H_iW_j$ for each $j \in \mathcal{C}_i, i = 1, 2$, is a subgaussian random vector with its marginal ψ_2 norm bounded in the sense of (16) and (17); hence for $R_i = H_i^T$,

$$\|\langle \mathbb{Z}_j, \mu\rangle\|_{\psi_2} = \|\langle H_iW_j, \mu\rangle\|_{\psi_2} \le \|W_j\|_{\psi_2} \|R_i\mu\|_2 = \|W_j\|_{\psi_2} \|\langle \mathbb{Z}_j, \mu\rangle\|_{L_2}, \qquad (97)$$

where $\|\langle \mathbb{Z}_j, \mu\rangle\|_{L_2}^2 = \mu^T \mathsf{Cov}(\mathbb{Z}_j)\mu = \mu^T H_i H_i^T \mu = \|R_i\mu\|_2^2$ by Lemma 5. Hence, by (97) and definition of K as in (14), we have for $i = 1, 2$ and $C_0 = \max_j \|W_j\|_{\psi_2}$,

$$\forall j \in \mathcal{C}_i, \quad \|\langle \mathbb{Z}_j, \mu\rangle\|_{\psi_2} \le \|W_j\|_{\psi_2} \|R_i\mu\|_2 \le C_0 \|R_i\mu\|_2 = C_0 \|\langle \mathbb{Z}_j, \mu\rangle\|_{L_2}.$$

Second, we have by independence of $\mathbb{Z}_j, \forall j$ and for $u = (u_1, \ldots, u_n) \in \{-1, 1\}^n$,

$$\begin{aligned}
\left\| \sum_{i=1}^n u_i \langle \mathbb{Z}_i, \mu\rangle \right\|_{\psi_2}^2 &\le C \sum_{i=1}^n \|\langle \mathbb{Z}_i, \mu\rangle\|_{\psi_2}^2 \le Cn(C_0 \max_i \|R_i\mu\|_2)^2, \text{ and} \\
\left\| \sum_{i=1}^n v_i \langle \mathbb{Z}_i, \mu\rangle \right\|_{\psi_2}^2 &\le C \sum_{i=1}^n v_i^2 \|\langle \mathbb{Z}_i, \mu\rangle\|_{\psi_2}^2 \le C(C_0 \max_i \|R_i\mu\|_2)^2, \ \forall v \in \mathbb{S}^{n-1}.
\end{aligned}$$

Then (95) and (96) follow from the subgaussian tail bound; see, for example, Propositions 2.6.6 and 2.7.1, and Theorem 2.7.3 (Vershynin, 2026). ∎

Corollary 43. *Let X, C_0 and K be as in Lemma 5. Let Y be as in* (6)*. Under the conditions in Lemma 42, we have with probability at least $1 - 2\exp(-cn)$, for μ as in* (18)*,*

$$\left| \langle \mu, \sum_{j \in \mathcal{C}_1} (Y_j - \mathbb{E}(Y_j)) - \sum_{j \in \mathcal{C}_2} (Y_j - \mathbb{E}(Y_j)) \rangle \right| \le 2 \sum_{i=1}^n |\langle \mu, \mathbb{Z}_i\rangle| \le CC_0 n \max_i \|R_i\mu\|_2.$$

Proof Now for $t = CC_0 n \max_i \|R_i\mu\|_2 / 2$, we have by (96) and the union bound,

$$\mathbb{P}\left(\max_{u \in \{-1,1\}^n} \sum_{j=1}^n u_j \langle \mu, \mathbb{Z}_j\rangle \ge t \right) \le 2^n 2 \cdot \exp\left(-\frac{c't^2}{n(C_0 \max_i \|R_i\mu\|_2)^2} \right) \le 2\exp(-cn).$$

Hence, with probability $\ge 1 - 2\exp(-cn)$, we have for $x = (x_1, \ldots, x_n) \in \{-1, 1\}^n$,

$$\begin{aligned}
\left| \langle \mu, \sum_{j \in \mathcal{C}_1} (Y_j - \mathbb{E}(Y_j)) - \sum_{j \in \mathcal{C}_2} (Y_j - \mathbb{E}(Y_j)) \rangle \right| &\le \max_{x \in \{-1,1\}^n} \sum_{i=1}^n x_i \langle \mu, Y_i - \mathbb{E}(Y_i)\rangle \\
&\le 2 \max_{u \in \{-1,1\}^n} \sum_{i=1}^n u_i \langle \mu, \mathbb{Z}_i\rangle \le 2t,
\end{aligned}$$

by Lemma 27. ∎

Appendix C. Proof of Theorem 6, and Lemmas 26 and 28

Theorem 6 follows from (49), Lemmas 25, 26, and 28 immediately, and the probability statements hold upon adjusting the constants. We prove Lemmas 26 and 28 in the supplementary Sections C.1 and C.2 respectively. First, we state Proposition 44, which holds regardless of the weights or the number of mixture components. This justifies (50) and Lemma 25.

Proposition 44 (Covariance projection). *Let $Y = X - P_1 X$, where $P_1 = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$. Let $\mathbb{Z} = X - \mathbb{E}X$. We first rewrite $\widehat{\Sigma}_Y = (Y - \mathbb{E}(Y))(Y - \mathbb{E}(Y))^T$ as follows:*

$$\begin{aligned}\widehat{\Sigma}_Y &:= (I - P_1)\mathbb{Z}\mathbb{Z}^T(I - P_1) = M_1 - (M_2 - M_3), \quad \textit{where}\\ M_1 &:= \widehat{\Sigma}_X = (X - \mathbb{E}(X))(X - \mathbb{E}(X))^T = \mathbb{Z}\mathbb{Z}^T,\\ M_2 &= P_1\mathbb{Z}\mathbb{Z}^T + \mathbb{Z}\mathbb{Z}^T P_1 \ \textit{and}\ M_3 = P_1\mathbb{Z}\mathbb{Z}^T P_1.\end{aligned}$$

Hence $\Sigma_Y := \mathbb{E}\widehat{\Sigma}_Y := (I - P_1)\mathbb{E}(\mathbb{Z}\mathbb{Z}^T)(I - P_1)$. Moreover, we have (49).

C.1 Proof of Lemma 26

We prove the lemma with the full generality by allowing each vector $Z_j = H_j W_j \in \mathbb{R}^p$ to have its own covariance $\Sigma_j = H_j H_j^T$, where $\Sigma_j \in \mathbb{R}^{p\times p}$, as in Definition 3. That is, we consider the general data generative process in Definition 3, and denote the (sample-by-sample) covariance by

$$\Gamma := \mathbb{E}\mathbb{Z}\mathbb{Z}^T = \mathrm{diag}([\mathrm{tr}(\Sigma_1), \ldots, \mathrm{tr}(\Sigma_n)]), \text{ where } \Sigma_j = \mathsf{Cov}(\mathbb{Z}_j) = H_j H_j^T; \tag{98}$$

The model under consideration in Lemma 5 is a special case of the general model in (98), since we let $H_j = H_i$ for all $j \in \mathcal{C}_i$, for $i = 1, 2$.

Lemma 45. *Let W_j be an isotropic, sub-gaussian random vector with independent, mean-zero coordinates as in Definition 3 where $K := \max_{j,k} \|w_{jk}\|_{\psi_2}$. Let $\{\mathbb{Z}_j^T, j \in [n]\}$ be row vectors of $\mathbb{Z}$, where $\mathbb{Z}_j = H_j W_j$ for $j \in [n]$. Then for each $j \in [n]$, we have*

$$\begin{aligned}\mathbb{P}\left(\left|\|\mathbb{Z}_j\|_2^2 - \mathbb{E}\|\mathbb{Z}_j\|_2^2\right| > t\right) &= \mathbb{P}\left(\left|\|H_j W_j\|_2^2 - \|H_j\|_F^2\right| > t\right)\\ &\le 2\exp\left(-c\min\left(\frac{t^2}{(K^2\|H_j\|_2\|H_j\|_F)^2}, \frac{t}{(K\|H_j\|_2)^2}\right)\right).\end{aligned}$$

Proof Fix j. First, we introduce the positive semidefinite matrix $M := H_j^T H_j \succeq 0 \in \mathbb{R}^{m\times m}$. Now, we bound for each anisotropic vector $\mathbb{Z}_j = H_j W_j \in \mathbb{R}^p$, where $W_j \in \mathbb{R}^m$, and

$$\begin{aligned}\forall j \in [n], \quad \|\mathbb{Z}_j\|_2^2 &= \langle H_j W_j, H_j W_j\rangle = W_j^T H_j^T H_j W_j =: W_j^T M W_j, \ \text{and hence} \quad (99)\\ \mathbb{E}\|\mathbb{Z}_j\|_2^2 &= \mathbb{E}(W_j^T(H_j^T H_j)W_j) = \mathrm{tr}(H_j^T H_j) = \|H_j\|_F^2, \quad (100)\end{aligned}$$

where $\mathrm{tr}(H_j H_j^T) = \mathrm{tr}(M) = \|H_j\|_F^2 \le (m \wedge p)\|H_j\|_2^2$. Now clearly,

$$\|M\|_F = \left\|H_j^T H_j\right\|_F \le \|H_j\|_2\|H_j\|_F \quad \text{and } \|M\|_2 = \|H_j\|_2^2. \tag{101}$$

Thus we have for any $t > 0$, by the Hanson-Wright inequality, cf. Theorem 1.1 Rudelson and Vershynin (2013), for $\|H_j W_j\|_2^2 - \mathbb{E}\|H_j W_j\|_2^2 = W_j^T M W_j - \mathrm{tr}(M)$,

$$\begin{aligned}\mathbb{P}\left(\left|\|\mathbb{Z}_j\|_2^2 - \mathbb{E}\|\mathbb{Z}_j\|_2^2\right| > t\right) &= \mathbb{P}\left(\left|W_j^T M W_j - \mathrm{tr}(M)\right| > t\right)\\ &\le 2\exp\left(-c\min\left(\frac{t^2}{K^4\|M\|_F^2}, \frac{t}{K^2\|M\|_2}\right)\right).\end{aligned}$$

The lemma thus holds in view of (101). ■

Proof of Lemma 26. Recall $\sigma_{\max}^2 := \max_j \|\mathsf{Cov}(\mathbb{Z}_j)\|_2$ and $C_0 = \max_j \|W_j\|_{\psi_2} \geq K$ by definition of C_0 and $K = \max_{j,k} \|w_{jk}\|_{\psi_2}$ as in (14). First, we have by Lemma 45 and for $t_{\text{diag}} = C_{\text{diag}} C_0^2 (\sqrt{n} \max_j (\|H_j\|_2 \|H_j\|_F)) \vee (n\sigma_{\max}^2)$, where $\max_j (\|H_j\|_2 \|H_j\|_F) \leq \sqrt{p}\sigma_{\max}^2$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{j\in[n]} \left| \|\mathbb{Z}_j\|_2^2 - \mathbb{E}\|\mathbb{Z}_j\|_2^2 \right| > t_{\text{diag}} \right) =: \mathbb{P}(\mathbb{E}_0) = \qquad (102)\\
&\leq \sum_{j=1}^n 2\exp\left(-c_1 \min\left(\frac{(C_{\text{diag}}^2 C_0^4 (\max_j \|H_j\|_2 \|H_j\|_F)^2) n}{K^4 \|H_j\|_2^2 \|H_j\|_F^2}, \frac{(C_{\text{diag}} C_0^2 \max_j \|H_j\|_2^2) n}{K^2 \|H_j\|_2^2} \right) \right)\\
&\leq 2n\exp(-c_1 \min(C_{\text{diag}}^2, C_{\text{diag}}) n) \leq 2\exp(-c'n),
\end{aligned}$$

where for the $p \times m$ matrix H_j, we have $\|H_j\|_F \leq \sqrt{p \wedge m}\, \|H_j\|_2$ for all $j \in [n]$. We use Theorem 4 to bound the off-diagonal part. Hence for all $q, h \in \mathbb{S}^{n-1}$, $A(q,h) := \text{offd}(q \otimes h) = (a_{ij})$, $\|A(q,h)\|_2 \leq \|A(q,h)\|_F \leq 1$. Let $t_{\text{offd}} = C_1 C_0^2 \sigma_{\max}^2 (\sqrt{pn} \vee n) \geq C_1 K^2 \sigma_{\max}^2 (\sqrt{pn} \vee n)$. For a particular realization of $q, h \in \mathbb{S}^{n-1}$ and $A(q,h) = (a_{ij})$ as defined above,

$$\begin{aligned}
&\mathbb{P}\left(\left| \sum_{i=1}^n \sum_{j\neq i}^n \langle \mathbb{Z}_i, \mathbb{Z}_j \rangle q_i h_j \right| > t_{\text{offd}} \right) = \mathbb{P}\left(\left| \sum_{i=1}^n \sum_{j\neq i}^n \langle \mathbb{Z}_i, \mathbb{Z}_j \rangle a_{ij} \right| > t_{\text{offd}} \right)\\
&\leq 2\exp\left(-c_2 \min\left(\frac{C_1^2 (C_0\sigma_{\max})^4 (\sqrt{pn} \vee n)^2}{(K\sigma_{\max})^4 p \|A(q,h)\|_F^2}, \frac{C_1 (C_0\sigma_{\max})^2 (\sqrt{pn} \vee n)}{(K\sigma_{\max})^2 \|A(q,h)\|_2} \right) \right)\\
&\leq 2\exp\left(-c_2 n \min(C_1^2, C_1) \right) \leq 2\exp(-c_2 n)
\end{aligned}$$

for some sufficiently large constants $C_1 \geq 1$ and $c_2 > 4\ln 9$, by Theorem 4. Taking a union bound over all $|\mathcal{N}|^2 \leq 9^{2n}$ pairs $q, h \in \mathcal{N}$, the ε-net of $\mathbb{S}^{n-1}$, we conclude that

$$\begin{aligned}
&\mathbb{P}\left(\max_{q,h\in\mathcal{N}} \left| \sum_{i=1}^n \sum_{j\neq i}^n q_i h_j \langle \mathbb{Z}_i, \mathbb{Z}_j \rangle \right| > t_{\text{offd}} \right) =: \mathbb{P}(\mathbb{E}_8) \qquad (103)\\
&\leq |\mathcal{N}|^2 \cdot 2\exp(-c_2 n \min(C_1^2, C_1)) \leq 2\exp\left(-c_2 n + 2n\ln 9\right) = 2\exp\left(-c_3 n\right).
\end{aligned}$$

One can show that (51) holds by a standard approximation argument under $\mathbb{E}_8^c$; see, for example Exercise 4.4.3 Vershynin (2018). Thus we have on event $\mathbb{E}_8^c \cap \mathbb{E}_0^c$,

$$\begin{aligned}
\left\| \mathbb{Z}\mathbb{Z}^T - \mathbb{E}(\mathbb{Z}\mathbb{Z}^T) \right\|_2 &\leq \left\| \text{diag}(\mathbb{Z}\mathbb{Z}^T) - \mathbb{E}\text{diag}(\mathbb{Z}\mathbb{Z}^T) \right\|_{\max} + \left\| \text{offd}(\mathbb{Z}\mathbb{Z}^T) \right\|_2\\
&\leq C_2 C_0^2 \sigma_{\max}^2 (\sqrt{np} \vee n); \text{ cf. (102) and (103)}.
\end{aligned}$$

Moreover, we have $\mathbb{P}(\mathbb{E}_8^c \cap \mathbb{E}_0^c) \geq 1 - 2\exp(-cn)$ upon adjusting the constants. □

C.2 Proof of Lemma 28

Let c, c', C_3, C_4 be some absolute constants. Let $M_Y := \mathbb{E}(Y)(Y - \mathbb{E}(Y))^T + (Y - \mathbb{E}(Y))\mathbb{E}(Y)^T$. Clearly, vectors $\mathbb{Z}_1, \mathbb{Z}_2, \ldots, \mathbb{Z}_n \in \mathbb{R}^p$ are independent. Let μ be as in (18). Let $R_i = H_i^T$.

Construct an ε-net Π_n of $\mathbb{S}^{n-1}$, where $\varepsilon = 1/3$ and $|\Pi_n| \le (1+2/\varepsilon)^n$. For a suitably chosen constant C_3, we have by Lemma 42,

$$\begin{aligned}\mathbb{P}(\mathbb{E}_3) &:= \mathbb{P}\left(\exists q \in \Pi_n, \left|\sum_{i=1}^n q_i\langle \mathbb{Z}_i, \mu\rangle\right| \ge \frac{1}{2}C_3(C_0 \max_i \|R_i\mu\|_2)\sqrt{n}\right)\\ &\le 9^n \cdot 2\exp\left(-\frac{cn(C_0\max_i \|R_i\mu\|_2)^2}{C(C_0\max_i \|R_i\mu\|_2)^2}\right) \le 2\exp(-c'n).\end{aligned}$$

By a standard approximation argument and Lemma 27, on event $\mathbb{E}_3^c$

$$\begin{aligned}\sup_{q\in\mathbb{S}^{n-1}}\sum_{i=1}^n q_i\langle \mathbb{Z}_i,\mu\rangle &\le \frac{1}{1-\varepsilon}\sup_{q\in\Pi_n}\sum_{i=1}^n q_i\langle \mathbb{Z}_i,\mu\rangle \le C_3C_0(\max_i \|R_i\mu\|_2)\sqrt{n},\\ \text{and } \|M_Y\|_2/\sqrt{p\gamma} &\le 4\sqrt{w_1w_2n}\sup_{q\in\mathbb{S}^{n-1}}\left|\sum_{i=1}^n q_i\langle \mathbb{Z}_i,\mu\rangle\right| \le 2C_3(C_0\max_i\|R_i\mu\|_2)n. \ \square\end{aligned}$$

Appendix D. Proof of Lemmas 12 and 20

We have the following facts about R as defined in (12).

Proposition 46. *Recall $R = \mathbb{E}(Y)\mathbb{E}(Y)^T$. Clearly $\mathbf{1}_n^T R\mathbf{1}_n = 0$, and hence*

$$\begin{aligned}\mathbf{1}_n^T\mathrm{offd}(R)\mathbf{1}_n &= \mathbf{1}_n^TR\mathbf{1}_n - \mathrm{tr}(R) = -\mathrm{tr}(R) = -np\gamma w_2w_1, \quad \textit{where}\\ \mathrm{tr}(R)/(p\gamma) &= nw_2^2w_1 + nw_1^2w_2 = w_1w_2n \quad \textit{and} \quad \|R\|_2 = \mathrm{tr}(R) = w_1w_2np\gamma.\end{aligned} \tag{104}$$

Thus $\mathrm{tr}(R)/n = p\gamma w_1w_2$ and for $n \ge 4$,

$$\left\|\frac{\mathrm{tr}(R)}{n-1}(I_n - E_n/n)\right\|_2 \le \frac{n}{n-1}p\gamma w_1w_2 \le p\gamma/3, \tag{105}$$

where $\|I_n - E_n/n\|_2 = 1$ since $I_n - E_n/n$ is a projection matrix.

Next, we state the following fact about W_2. Let M_1, M_2, M_3, W_0, W_2, V_1, V_2... be the same as in Propositions 44 and 35.

Proposition 47. *Denote by*

$$\mathbb{W} := W_2 - \frac{1}{n}\mathrm{tr}(W_2)I_n - \frac{\mathbf{1}_n^T\mathrm{offd}(W_2)\mathbf{1}_n}{n(n-1)}(E_n - I_n)$$

Then $\mathbb{W}$ coincides with (75). Moreover, we have for $\delta_V := (V_2 - V_1)(w_2 - w_1)$

$$\begin{aligned}\mathbb{W} &= W_2 - \frac{\delta_V E_n}{n} = \frac{V_1 - V_2}{n}\begin{bmatrix} 2w_2E_{n_1} & (w_2-w_1)E_{n_1\times n_2}\\ (w_2-w_1)E_{n_2\times n_1} & -2w_1E_{n_2}\end{bmatrix},\\ \|\mathbb{W}\|_2 &\le |V_1 - V_2|(w_1\vee w_2) \quad \textit{and} \quad \|\mathbb{W}\|_{\infty\to 1} \le n|V_1-V_2|(w_1\vee w_2).\end{aligned} \tag{106}$$

Proof By definition of W_2 as in (73), we have

$$\begin{aligned} \mathrm{tr}(W_2) &= \frac{V_2 - V_1}{n}(w_2 n - w_1 n) = (V_2 - V_1)(w_2 - w_1) \text{ and} \\ \mathbf{1}_n^T W_2 \mathbf{1}_n &= \frac{V_2 - V_1}{n}(w_2^2 n^2 - w_1^2 n^2) = n(V_2 - V_1)(w_2 - w_1). \end{aligned}$$

Thus

$$\begin{aligned} \frac{\mathbf{1}_n^T \mathrm{offd}(W_2)\mathbf{1}_n}{n(n-1)} &= \frac{\mathbf{1}_n^T W_2 \mathbf{1}_n - \mathrm{tr}(W_2)}{n(n-1)} = \frac{(V_2 - V_1)(w_2 - w_1)}{n}; \text{ and hence} \\ \mathbb{W} &= W_2 - \frac{1}{n}\mathrm{tr}(W_2) I_n - \frac{\mathbf{1}_n^T \mathrm{offd}(W_2)\mathbf{1}_n}{n(n-1)}(E_n - I_n) \\ &= W_2 - \frac{(V_2 - V_1)(w_2 - w_1)}{n} E_n. \end{aligned}$$

Hence (75) holds. Moreover, by symmetry, $\|\mathbb{W}\|_2 \le \|\mathbb{W}\|_\infty \le |V_1 - V_2|\,(w_1 \vee w_2)$. ■

D.1 Some useful propositions

Next, we compute the mean values in Proposition 48, and we obtain an expression on $\mathbb{E}B - R$ in Proposition 35. Proposition 48 is proved in Section D.4 in Zhou (2026).

Proposition 48. (Covariance projection: two groups) *Let $N_j = w_j n$ for $j \in \{1, 2\}$. W.l.o.g., suppose that the first n_1 rows in X are in $\mathcal{C}_1$ and the following n_2 rows are in $\mathcal{C}_2$. Let M_1, M_2, M_3 be defined as in Proposition 44. Let V_1 and V_2 be the same as in* (24)*:*

$$V_1 := \mathbb{E}\langle \mathbb{Z}_j, \mathbb{Z}_j \rangle \ \ \forall j \in \mathcal{C}_1 \quad \text{and} \quad V_2 := \mathbb{E}\langle \mathbb{Z}_j, \mathbb{Z}_j \rangle \ \ \forall j \in \mathcal{C}_2. \tag{107}$$

Let $V_m := w_1 V_1 + w_2 V_2 = W_n/n$. Let $\widehat{\Sigma}_Y$ be as in (98)*. Let W_n be defined as in* (108)*:*

$$W_n := \mathbb{E}\sum_{i=1}^{n}\langle \mathbb{Z}_i, \mathbb{Z}_i \rangle = \sum_{i=1}^{n}\sum_{k=1}^{p}\mathbb{E}z_{ik}^2, \tag{108}$$

$$\mathbb{E}M_1 := \begin{bmatrix} V_1 I_{n_1} & 0 \\ 0 & V_2 I_{n_2} \end{bmatrix} =: (w_1 V_1 + w_2 V_2) I_n + W_0, \tag{109}$$

$$\mathbb{E}M_2 = \frac{V_1 + V_2}{n} E_n + W_2, \text{ and } \mathbb{E}M_3 = \frac{W_n}{n^2} E_n, \tag{110}$$

$$\text{where} \quad W_n/n^2 := (|\mathcal{C}_1|\, V_1 + |\mathcal{C}_2|\, V_2)/n^2 = (w_1 V_1 + w_2 V_2)/n =: \frac{V_w}{n}, \tag{111}$$

$\mathrm{tr}(\mathbb{E}M_1)/n = V_m$, *and W_0, W_2 are as defined in* (73)*. Now putting things together, we obtain the expression for covariance of Y:*

$$\begin{aligned} \Sigma_Y &:= \mathbb{E}(YY^T) - \mathbb{E}(Y)\mathbb{E}(Y)^T = \mathbb{E}M_1 + \mathbb{E}M_3 - \mathbb{E}M_2 \\ &= \begin{bmatrix} V_1 I_{n_1} & 0 \\ 0 & V_2 I_{n_2} \end{bmatrix} - \frac{w_2 V_1 + w_1 V_2}{n} E_n - W_2 \end{aligned}$$

which simplifies to $\Sigma_Y = V(I_n - E_n/n)$ in case $V_1 = V_2 = V$.

D.2 Proof of Lemma 12

We have by Proposition 35,

$$\begin{aligned}\|\mathbb{E}B - R\| &= \left\|W_0 - \mathbb{W} - \frac{\mathrm{tr}(R)}{(n-1)}(I_n - \frac{E_n}{n})\right\| \\ &\leq \|W_0\| + \|\mathbb{W}\| + \left\|\frac{\mathrm{tr}(R)}{(n-1)}(I_n - \frac{E_n}{n})\right\|, \end{aligned} \tag{112}$$

where the $\|\cdot\|$ is understood to be either the operator or the cut norm. Clearly, we have by Proposition 47,

$$\|\mathbb{W}\|_2 \leq |V_1 - V_2|\,(w_1 \vee w_2) \;\text{ and }\; \|W_0\|_2 \leq |V_1 - V_2|\,(w_2 \vee w_1). \tag{113}$$

Combining (105), (112), and (113), we have for $w_{\min} := w_1 \wedge w_2$ and $n\xi \geq \frac{1}{2w_{\min}}$,

$$\begin{aligned}\|\mathbb{E}B - R\|_2 &\leq \|W_0\|_2 + \|\mathbb{W}\|_2 + \left\|\frac{\mathrm{tr}(R)}{(n-1)}(I_n - \frac{E_n}{n})\right\|_2 \\ &\leq \frac{2}{3}\xi np\gamma(1 - w_{\min}) + p\gamma/3 \leq \frac{2}{3}\xi np\gamma,\end{aligned}$$

where we use the fact that $\frac{2}{3}\xi p\gamma n w_{\min} \geq p\gamma/3$ since $2\xi n w_{\min} \geq 1$. Now the bound on the cut norm follows since $\|\mathbb{E}B - R\|_{\infty\to 1} \leq n\,\|\mathbb{E}B - R\|_2 \leq \frac{2}{3}\xi n^2 p\gamma$. The lemma thus holds for the general setting; when $V_1 = V_2$, we leave this as an exercise. □

D.3 Proof of Lemma 20

Proof of Lemma 20. Now (38) holds since by Proposition 31,

$$\begin{aligned}2\binom{n}{2}|\lambda - \mathbb{E}\lambda| &= |n(\tau - \mathbb{E}\tau)| \leq n\left\|YY^T - \mathbb{E}(YY^T)\right\|_2, \\ \text{where } n\,|(\tau - \mathbb{E}\tau)| &\leq n\max_i |\langle Y_i, Y_i\rangle - \mathbb{E}\langle Y_i, Y_i\rangle| \leq n\left\|YY^T - \mathbb{E}(YY^T)\right\|_2.\end{aligned}$$

□

Corollary 49 follows from the proof of Lemma 12, which we state to prove a bound for the balanced cases. The proof is given in Section D.6 in Zhou (2026).

Corollary 49. *For general cases, we have by definition,*

$$W_2 := \frac{V_1 - V_2}{n}\begin{bmatrix} E_{n_1} & 0E_{n_1\times n_2} \\ 0E_{n_2\times n_1} & -E_{n_2}\end{bmatrix}$$

Moreover we have the following term which depends on the weights,

$$\begin{aligned}\left\|\frac{(V_2 - V_1)(w_2 - w_1)}{n}E_n\right\|_{\infty\to 1} &\leq n\,|(V_2 - V_1)(w_2 - w_1)| \leq \xi n^2 p\gamma\,|w_2 - w_1| \\ \|W_2\|_{\infty\to 1} &:= n\,|V_1 - V_2|\,(w_1^2 + w_2^2) \\ \textit{and hence}\quad \|W_0 - W_2\|_{\infty\to 1} &\leq n\,\|W_0 - W_2\|_2 < n\,|V_1 - V_2| \leq \xi p n^2\gamma\end{aligned}$$

Lemma 50. (Reductions) *Suppose $n \geq 4$. Let W_0, W_2, V_1, V_2 be the same as in Proposition 35. Recall that $R = \mathbb{E}(Y)\mathbb{E}(Y)^T$.* **When** *$V_1 = V_2$, we have*

$$\begin{aligned} \mathbb{E}B - R &= -\frac{\mathrm{tr}(R)}{(n-1)}(I_n - \frac{E_n}{n}) = -p\gamma w_2 w_1 \frac{n}{n-1}(I_n - \frac{E_n}{n}), \\ \|\mathbb{E}B - R\|_{\infty\to 1} &= \left\| p\gamma w_2 w_1 \frac{n}{n-1}(I_n - \frac{E_n}{n}) \right\|_{\infty\to 1} \leq np\gamma/3, \quad \textit{for} \quad n \geq 4. \end{aligned}$$

For balanced clusters ($w_1 = w_2$), *$\mathbb{E}B - R = W_0 - W_2 - \frac{\mathrm{tr}(R)}{(n-1)}(I_n - \frac{E_n}{n})$ and for $n \geq 4$,*

$$\|\mathbb{E}B - R\|_{\infty\to 1} \leq n\,\|\mathbb{E}B - R\|_2 \leq \frac{np\gamma}{3} + |V_1 - V_2|\, n \leq \frac{1}{3}(1 + o(1))\xi p n^2 \gamma.$$

Proof Recall

$$W_0 - W_2 = (V_1 - V_2) \begin{bmatrix} w_2 I_{n_1} - E_{n_1}/n & 0 \\ 0 & -(w_1 I_{n_2} - E_{n_2}/n) \end{bmatrix}$$

The case where $V_1 = V_2$ follows from Proposition 46 and (74). We now show the balanced case where $w_1 = w_2$. Under the conditions of Lemma 12, we have for $\xi = \Omega(1/n)$, by (105) and (74),

$$\begin{aligned} \|\mathbb{E}B - R\|_2 &\leq \|W_0 - W_2\|_2 + \left\| \frac{(V_1 - V_2)(w_1 - w_2)}{n} E_n \right\|_2 + \left\| \frac{\mathrm{tr}(R)}{n-1}\,(I_n - E_n/n) \right\|_2 \\ &\leq |V_1 - V_2| + p\gamma/3 \leq \frac{1}{3}\xi np\gamma + p\gamma/3 \end{aligned}$$

Similarly, the upper bound on $\|\mathbb{E}B - R\|_{\infty\to 1}$ follows from Corollary 49 immediately. ■

D.4 Proof of Proposition 13

The optimization goals of SDP1 and SDP (5) are equivalent to that of **Oracle SDP** (29) in view of Proposition 13. Denote by

$$\mathcal{M}_G^+ := \{Z : Z \succeq 0, I_n \succeq \mathrm{diag}(Z)\} \subset [-1,1]^{n\times n}. \tag{114}$$

Proof of Proposition 13. Recall for $\mathcal{M}_G^+$ as in (114),

$$\mathcal{M}_{\mathrm{opt}} := \{Z : Z \succeq 0, \mathrm{diag}(Z) = I_n\} \subset \mathcal{M}_G^+ \subset [-1,1]^{n\times n};$$

Notice that for the second term in (7), we have $\langle (E_n - I_n), Z\rangle = \langle (E_n - I_n), \mathrm{offd}(Z)\rangle$ in the objective function (5), which does not depend on $\mathrm{diag}(Z)$. Hence, to maximize

$$\begin{aligned} \langle A, Z\rangle &= \langle A, \mathrm{offd}(Z)\rangle + \langle A, \mathrm{diag}(Z)\rangle \\ &= \langle \mathrm{offd}(A), \mathrm{offd}(Z)\rangle + \langle \mathrm{diag}(YY^T), \mathrm{diag}(Z)\rangle, \end{aligned}$$

over all $Z \in \mathcal{M}_G^+$, one must set the diagonal $Z_{jj} \in [0,1]$ to be 1, since (a) $\mathrm{diag}(YY^T) \geq 0$ and moreover, (b) increasing $\mathrm{diag}(Z)$ will only make it easier to satisfy $Z \succeq 0$ and hence to maximize $\langle \mathrm{offd}(A), \mathrm{offd}(Z)\rangle$. Thus, the set of optimizers $\widehat{Z}$ as in (5) must satisfy $\mathrm{diag}(\widehat{Z}) = I_n$. Thus (31) holds by definition of $\mathcal{M}_{\mathrm{opt}}$ as above. Moreover, (31) holds due to the fact that $\langle I_n, Z\rangle = \mathrm{tr}(Z) = n$ for all Z in the feasible set $\mathcal{M}_{\mathrm{opt}}$. □

Appendix E. Proof outline for Lemma 33

Let $\widehat{\delta}_Z := \widehat{Z} - Z^*$ and $S_1 := \langle \mathcal{P}(\Lambda), \widehat{\delta}_Z \rangle$. We need to first present Lemmas 51 and 52, which are proved in Sections E.1 and E.2 respectively. Throughout this section, we consider $\widehat{Z} \in \mathcal{M}_{\text{opt}} \cap (Z^* + r_1 B_1^{n\times n})$, where $r_1 \le 2qn$ for some positive integer $q \in [n]$.

Let $\sigma_{\max} := \max_i \|H_i\|_2$. As predicted, the decomposition and reduction ideas for the global analysis in Section 5 are useful for the local analysis as well. Let

$$\widehat{\Sigma}_Y = (Y - \mathbb{E}(Y))(Y - \mathbb{E}(Y))^T, \quad \widehat{\Sigma}_Y - \Sigma_Y = (I - P_1)(\mathbb{Z}\mathbb{Z}^T - \mathbb{E}(\mathbb{Z}\mathbb{Z}^T))(I - P_1). \quad (115)$$

Then we have by the triangle inequality,

$$S_1 \le \left|\langle \mathcal{P}((Y - \mathbb{E}(Y))\mathbb{E}(Y)^T), \widehat{\delta}_Z\rangle\right| + \left|\langle \mathcal{P}(\mathbb{E}Y(Y - \mathbb{E}(Y))^T), \widehat{\delta}_Z\rangle\right| + \left|\langle \mathcal{P}(\widehat{\Sigma}_Y - \Sigma_Y), \widehat{\delta}_Z\rangle\right|.$$

Lemma 51. *Suppose all conditions in Theorem 6 hold. Then* $\left|\langle \mathcal{P}(\mathbb{E}(Y)(Y - \mathbb{E}(Y))^T), \widehat{Z} - Z^*\rangle\right| = \left|\langle \mathcal{P}((Y - \mathbb{E}(Y))\mathbb{E}(Y)^T), \widehat{Z} - Z^*\rangle\right|$*, and with probability at least* $1 - \frac{c}{n^2}$*,*

$$\sup_{\widehat{Z}\in\mathcal{M}_{opt}\cap(Z^*+r_1B_1^{n\times n})} \left|\langle \mathcal{P}((Y - \mathbb{E}(Y))\mathbb{E}(Y)^T), \widehat{Z} - Z^*\rangle\right| \le C_5C_0(\max_i \left\|H_i^T\mu\right\|_2)\sqrt{p\gamma}r_1\sqrt{\log(2en/\lceil\frac{r_1}{2n}\rceil)}.$$

Lemma 52. *Under the conditions in Theorem 6, we have with probability at least* $1 - \frac{c'}{n^2}$*,*

$$\begin{aligned}
&\sup_{\widehat{Z}\in\mathcal{M}_{opt}\cap(Z^*+r_1B_1^{n\times n})} \left|\langle \mathcal{P}(\widehat{\Sigma}_Y - \Sigma_Y), \widehat{Z} - Z^*\rangle\right| \\
&\le \frac{2}{3}\xi p\gamma \left\|\widehat{Z} - Z^*\right\|_1 + C_{10}(C_0\sigma_{\max})^2\lceil\frac{r_1}{2n}\rceil\left(\sqrt{np\log(2en/\lceil\frac{r_1}{2n}\rceil)} + \sqrt{r_1}\log(2en/\lceil\frac{r_1}{2n}\rceil)\right).
\end{aligned}$$

Proof of Lemma 33. Combining Lemmas 51 and 52, we have with probability at least $1 - \frac{c}{n^2}$, for $\Delta^2 = p\gamma$ and $\max_i \left\|H_i^T\mu\right\|_2 \le \max_i \|H_i\|_2 = \sigma_{\max}$,

$$\begin{aligned}
\sup_{\widehat{Z}\in\mathcal{M}_{\text{opt}}\cap(Z^*+r_1B_1^{n\times n})} S_1(\widehat{Z}) \le& \sup_{\widehat{Z}\in\mathcal{M}_{\text{opt}}\cap(Z^*+r_1B_1^{n\times n})} \left|\langle \mathcal{P}(\widehat{\Sigma}_Y - \Sigma_Y), \widehat{Z} - Z^*\rangle\right| \\
&+ \sup_{\widehat{Z}\in\mathcal{M}_{\text{opt}}\cap(Z^*+r_1B_1^{n\times n})} 2\left|\langle \mathcal{P}((Y - \mathbb{E}(Y))\mathbb{E}(Y)^T), \widehat{Z} - Z^*\rangle\right| \\
\le& \frac{2}{3}\xi p\gamma \left\|\widehat{Z} - Z^*\right\|_1 + 2C_5C_0(\max_i \left\|H_i^T\mu\right\|_2)\lceil\frac{r_1}{2n}\rceil\sqrt{\log(2en/\lceil\frac{r_1}{2n}\rceil)}n\Delta + \\
&C_{10}(C_0\sigma_{\max})\lceil\frac{r_1}{2n}\rceil\sqrt{\log(2en/\lceil\frac{r_1}{2n}\rceil)}\left(C_0\sigma_{\max}\sqrt{r_1}\sqrt{\log(2en/\lceil\frac{r_1}{2n}\rceil)} + C_0\sigma_{\max}\sqrt{np}\right) \\
\le& \frac{2}{3}\xi p\gamma \left\|\widehat{Z} - Z^*\right\|_1 + C'C_0\sigma_{\max}\lceil\frac{r_1}{2n}\rceil\sqrt{\log(2en/\lceil\frac{r_1}{2n}\rceil)}\left(n\Delta + C_0\sigma_{\max}\sqrt{np}\right),
\end{aligned}$$

where in the last inequality, we have by (19), for $r_1 \le 2qn$ and $q = \lceil\frac{r_1}{2n}\rceil$,

$$\begin{aligned}
C_0\sigma_{\max}\sqrt{r_1}\sqrt{\log(2en/\lceil\frac{r_1}{2n}\rceil)} &\le C_0\sigma_{\max}\sqrt{2nq\log(2en/q)} \\
&\le C_0\sigma_{\max}n\sqrt{2\log(2e)} = O(n\sqrt{p\gamma}),
\end{aligned}$$

where the first inequality holds since $\max_{q\in[n]} q\log(2en/q) = n\log(2e)$, given that $q\log(2en/q)$ is a monotonically increasing function of q, for $1 \le q < n$, and the second inequality holds since $p\gamma = \Omega((C_0 \max_i \|H_i\|_2)^2)$. Lemma 33 thus holds. □

E.1 Proof of Lemma 51

First, we have

$$\begin{aligned} \langle (Y - \mathbb{E}(Y))\mathbb{E}(Y)^T P_2, \widehat{Z} - Z^* \rangle &= \langle P_2 \mathbb{E}(Y)(Y - \mathbb{E}(Y))^T, \widehat{Z} - Z^* \rangle, \quad (116) \\ \langle P_2(Y - \mathbb{E}(Y))\mathbb{E}(Y)^T, \widehat{Z} - Z^* \rangle &= \langle \mathbb{E}(Y)(Y - \mathbb{E}(Y))^T P_2, \widehat{Z} - Z^* \rangle. \quad (117) \end{aligned}$$

We will bound the two terms T_1, T_2 in Lemmas 53 and 54 respectively.

Lemma 53. *Under the conditions in Theorem 6, we have for all $\widehat{Z} \in \mathcal{M}_{opt}$, with probability at least $1 - 2\exp(-cn)$,*

$$|T_1| := \left| \langle (I - P_2)\mathbb{E}(Y)(Y - \mathbb{E}(Y))^T P_2, \widehat{Z} - Z^* \rangle \right| \le C_4 C_0 (\max_i \left\| H_i^T \mu \right\|_2) \sqrt{p\gamma} \left\| \widehat{Z} - Z^* \right\|_1 .$$

Moreover, when $w_1 = w_2$, $T_1 = 0$.

Lemma 54. *Suppose all conditions in Theorem 6 hold. Suppose $r_1 \le 2qn$ for some positive integer $1 \le q < n$. Let $\Delta = \sqrt{p\gamma}$. Then, with probability at least $1 - \frac{c}{n^2}$,*

$$\sup_{\widehat{Z} \in \mathcal{M}_{opt} \cap (Z^* + r_1 B_1^{n\times n})} \langle P_2 \mathbb{E}(Y)(Y - \mathbb{E}(Y))^T, \widehat{Z} - Z^* \rangle \le C_3 C_0 (\max_i \left\| H_i^T \mu \right\|_2) \Delta r_1 \sqrt{\log(2en/\left\lceil \frac{r_1}{2n} \right\rceil)}.$$

Then we have by the triangle inequality, (117) and (116),

$$\begin{aligned} &\left| \langle \mathcal{P}((Y - \mathbb{E}(Y))\mathbb{E}(Y)^T), \widehat{Z} - Z^* \rangle \right| = \left| \langle \mathcal{P}(\mathbb{E}(Y)(Y - \mathbb{E}(Y))^T), \widehat{Z} - Z^* \rangle \right| \quad (118) \\ \le\ & \left| \langle (\mathbb{E}(Y)(Y - \mathbb{E}(Y))^T) P_2, \widehat{Z} - Z^* \rangle - \langle P_2 \mathbb{E}(Y)(Y - \mathbb{E}(Y))^T P_2, \widehat{Z} - Z^* \rangle \right| \\ & + \left| \langle P_2 \mathbb{E}(Y)(Y - \mathbb{E}(Y))^T, \widehat{Z} - Z^* \rangle \right| =: |T_1| + |T_2| . \end{aligned}$$

Putting things together, we have a uniform upper bound on $\langle \mathcal{P}(\mathbb{E}(Y)(Y - \mathbb{E}(Y))^T), \widehat{Z} - Z^* \rangle$. We prove Lemma 53 in Section F.1. We prove Lemma 54 in Section E.3. □

E.2 Proof of Lemma 52

Upon obtaining the bounds in Lemma 51, there are only two unique terms left, which we bound in Lemmas 56 and 55 respectively. Notice that by the triangle inequality, we have for symmetric matrices $\widehat{\Sigma}_Y - \Sigma_Y = (I - P_1)\Psi(I - P_1)$ and $\widehat{Z} - Z^*$,

$$\left| \langle \mathcal{P}(\widehat{\Sigma}_Y - \Sigma_Y), \widehat{Z} - Z^* \rangle \right| \ \le\ 2 \left| \langle P_2(\widehat{\Sigma}_Y - \Sigma_Y), \widehat{Z} - Z^* \rangle \right| + \left| \langle P_2(\widehat{\Sigma}_Y - \Sigma_Y)P_2, \widehat{Z} - Z^* \rangle \right| .$$

As the first step, we first obtain in Lemma 55 a deterministic bound on $\langle P_2(\widehat{\Sigma}_Y - \Sigma_Y)P_2, \widehat{Z} - Z^* \rangle$. Then, in combination with Theorem 6, we obtain the high probability bound as well. We prove Lemma 55 in Section F.2. Next we state Lemma 56, which we prove in Section E.4.

Lemma 55. *Denote by* $\Psi := (\mathbb{Z}\mathbb{Z}^T - \mathbb{E}(\mathbb{Z}\mathbb{Z}^T))$. *Then*

$$\left|\langle P_2(\widehat{\Sigma}_Y - \Sigma_Y)P_2, \widehat{Z} - Z^*\rangle\right| \;\leq\; (1+|w_1-w_2|)^2 \|\Psi\|_2 \left\|\widehat{Z} - Z^*\right\|_1 /n.$$

Moreover, under the conditions in Theorem 6, we have with probability at least $1-\exp(-cn)$,

$$\sup_{\widehat{Z}\in\mathcal{M}_{opt}} \left|\langle P_2(\widehat{\Sigma}_Y - \Sigma_Y)P_2, \widehat{Z} - Z^*\rangle\right| \;\leq\; \frac{1}{3}\xi p\gamma \left\|\widehat{Z} - Z^*\right\|_1.$$

Lemma 56. *Suppose* $r_1 \leq 2qn$ *for some positive integer* $1 \leq q < n$. *Suppose all conditions in Theorem 10 hold. Then, with probability at least* $1 - \frac{c}{n^2}$,

$$\sup_{\widehat{Z}\in\mathcal{M}_{opt}\cap(Z^*+r_1B_1^{n\times n})} 2\left|\langle P_2(\widehat{\Sigma}_Y - \Sigma_Y), \widehat{Z} - Z^*\rangle\right| \leq \frac{1}{3}\xi p\gamma \left\|\widehat{Z} - Z^*\right\|_1 +$$
$$C_{10}(C_0 \max_i \|H_i\|_2)^2 \lceil\frac{r_1}{2n}\rceil \left(\sqrt{np\log(2en/\lceil\frac{r_1}{2n}\rceil)} + \sqrt{r_1}\log(2en/\lceil\frac{r_1}{2n}\rceil)\right).$$

Lemma 52 follows from Lemmas 55 and 56 immediately. Lemma 55 follows immediately from Lemma 57 and Lemma 26, which controls $\left\|\mathbb{Z}\mathbb{Z}^T - \mathbb{E}(\mathbb{Z}\mathbb{Z}^T)\right\|_2$. See Section F.4 Zhou (2026) for the proof of Lemma 57.

Lemma 57. (Deterministic bounds) *Let* $\Psi = \mathbb{Z}\mathbb{Z}^T - \mathbb{E}(\mathbb{Z}\mathbb{Z}^T)$. *Then*

$$\left|\langle P_2\Psi P_1, \widehat{Z} - Z^*\rangle\right| \;\leq\; \|\Psi\|_2 \left\|\widehat{Z} - Z^*\right\|_1 /n, \tag{119}$$
$$\left|\langle P_2P_1\Psi P_1, \widehat{Z} - Z^*\rangle\right| \;\leq\; |w_1-w_2| \|\Psi\|_2 \left\|\widehat{Z} - Z^*\right\|_1 /n, \tag{120}$$
$$\text{and}\quad \left|\langle P_2P_1\Psi P_1P_2, \widehat{Z} - Z^*\rangle\right| \;\leq\; |w_1-w_2|^2 \|\Psi\|_2 \left\|\widehat{Z} - Z^*\right\|_1 /n. \tag{121}$$

Finally, we will show the following bounds: for $\widehat{\delta}_Z := \widehat{Z} - Z^*$,

$$\langle P_2\Psi P_2, \widehat{Z} - Z^*\rangle \leq \|\Psi\|_2 \left\|\widehat{Z} - Z^*\right\|_1 /n, \tag{122}$$
$$\left|\langle P_2\Psi P_1P_2, \widehat{\delta}_Z\rangle\right| = \left|\langle P_2P_1\Psi P_2, \widehat{\delta}_Z\rangle\right| \leq \|\Psi\|_2 \frac{|w_1-w_2|}{n} \left\|\widehat{Z} - Z^*\right\|_1. \tag{123}$$

When $w_1 = w_2$, $P_2P_1 = 0$, *and the RHS of* (120), (121) *and* (123) *all become 0.*

E.3 Proof of Lemma 54

We state in Lemmas 58 and 59 two reduction steps. For all $\widehat{Z} \in M_{\text{opt}} \cap (Z^* + r_1B_1^{n\times n})$, we have $\widehat{w} \in B_\infty^n \cap \lceil r_1/(2n)\rceil B_1^n$, where

$$\widehat{w}_j \;:=\; \frac{1}{2n}\left\|(\widehat{Z} - Z^*)_{\cdot j}\right\|_1 \leq 1 \text{ and } \sum_j \widehat{w}_j \leq \lceil r_1/(2n)\rceil, \tag{124}$$

where B_∞^n and B_1^n denote the unit ℓ_∞ ball and ℓ_1 ball respectively. Denote by $L_i := \langle \mu^{(1)} - \mu^{(2)}, \mathbb{Z}_i\rangle$ throughout this section. Denote by $L_1^* \geq L_2^* \geq \ldots \geq L_n^*$ (resp. $\widehat{w}_1^* \geq \widehat{w}_2^* \geq$

$\ldots \geq \widehat{w}_n^*$) the non-decreasing arrangement of $|L_j|$ (resp. $\widehat{w}_j$). We then have the reduction as in Lemma 59. It remains to bound the sum on the RHS of (126), namely, $\sum_{j=1}^{\lceil r_1/(2n)\rceil} L_j^*$. To do so, we state in Proposition 60 a high probability bound on $\sum_{j=1}^{q} L_j^*$, which holds simultaneously for all $q \in [n]$. Denote by $\widehat{\delta}_Z := \widehat{Z} - Z^*$.

Lemma 58. (Deterministic bounds) *Let $\widehat{w}_j$ be as in (124). Then for $\widehat{\delta}_Z := \widehat{Z} - Z^*$,*

$$\left|\langle P_2(\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T), \widehat{\delta}_Z\rangle\right| \leq 4w_1w_2n\sum_{j=1}^{n}|L_j|\,\widehat{w}_j + \frac{2w_1w_2}{n}\left|\sum_{i=1}^{n}L_i\right|\left\|\widehat{\delta}_Z\right\|_1. \tag{125}$$

Lemma 59. (Reduction to order statistics) *Under the settings of Lemma 58, we have*

$$\sup_{\widehat{Z}\in\mathcal{M}_{opt}\cap(Z^*+r_1B_1^{n\times n})}\left|\langle P_2(\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T), \widehat{Z}-Z^*\rangle\right| \leq 3n\sum_{j=1}^{\lceil r_1/2n\rceil} L_j^*. \tag{126}$$

Proposition 60. *Suppose all conditions in Lemma 5 hold. Let random matrix $\mathbb{Z}$ satisfy the conditions as stated therein. Let H_i and $\mu = (\mu^{(1)}-\mu^{(2)})/\sqrt{p\gamma}$ be as defined in Theorem 6. Let $1 \leq q \leq n$ denote a positive integer. Then for some absolute constants C_5, c, we have with probability at least $1-\frac{c}{n^2}$, simultaneously for all $q \in [n]$,*

$$\sum_{j=1}^{q} L_j^* \leq C_5C_0(\max_i \left\|H_i^T\mu\right\|_2)\sqrt{p\gamma}q\sqrt{\log(2en/q)}.$$

It remains to prove Lemma 54. Indeed, we have with probability at least $1-\frac{c}{n^2}$, by (125) and (126), and Proposition 60, for $w_1w_2 \leq 1/4$,

$$\begin{aligned}&\sup_{\widehat{Z}\in\mathcal{M}_{\rm opt}\cap(Z^*+r_1B_1^{n\times n})}\left|\langle P_2(\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T), \widehat{Z}-Z^*\rangle\right| \\ \leq\ & 3n\sum_{j=1}^{\lceil r_1/(2n)\rceil} L_j^* \leq C_3C_0(\max_i \left\|H_i^T\mu\right\|_2)\Delta r_1\sqrt{\log(2en/\lceil\frac{r_1}{2n}\rceil)},\end{aligned}$$

where it is understood that $q = \lceil\frac{r_1}{2n}\rceil \in [n]$ for some $r_1 > 0$. We prove Lemmas 58 and 59, in Sections F.4 and G respectively. We prove Proposition 60 in Section H. □

E.4 Proof of Lemma 56

First we note that by (115), for $\widehat{\delta}_Z := \widehat{Z} - Z^*$,

$$\begin{aligned}&\langle(\widehat{\Sigma}_Y-\Sigma_Y)P_2, \widehat{\delta}_Z\rangle = \langle P_2(\widehat{\Sigma}_Y-\Sigma_Y), \widehat{\delta}_Z\rangle \\ =\ & \langle P_2\Psi, \widehat{\delta}_Z\rangle - \langle P_2P_1\Psi, \widehat{\delta}_Z\rangle + \langle P_2P_1\Psi P_1, \widehat{\delta}_Z\rangle - \langle P_2\Psi P_1, \widehat{\delta}_Z\rangle.\end{aligned} \tag{127}$$

Lemma 61. (Deterministic bounds) *Denote by $\Psi = \mathbb{Z}\mathbb{Z}^T - \mathbb{E}(\mathbb{Z}\mathbb{Z}^T)$ and $\mathrm{offd}(\Psi)$ its off-diagonal component, where the diagonal elements are set to be 0. Denote by*

$$\begin{aligned}Q_j &:= \sum_{t\in\mathcal{C}_1, t\neq j}\Psi_{jt} - \sum_{t\in\mathcal{C}_2, t\neq j}\Psi_{jt} = \sum_{t\in\mathcal{C}_1}\mathrm{offd}(\Psi)_{jt} - \sum_{t\in\mathcal{C}_2}\mathrm{offd}(\Psi)_{jt}, \\ S_j &:= \sum_{t\in[n], t\neq j}\Psi_{jt} = \sum_{t\in[n]}\mathrm{offd}(\Psi)_{jt}.\end{aligned} \tag{128}$$

Let $\widehat{w}_i$ be the same as in (124). *Then for* $\|\mathrm{diag}(\Psi)\|_{\max} := \max_j |\Psi_{jj}|$,

$$\begin{aligned}
\left|\langle P_2 \mathrm{diag}(\Psi), \widehat{Z} - Z^* \rangle\right| &\leq \|\mathrm{diag}(\Psi)\|_{\max} \left\|\widehat{Z} - Z^*\right\|_1 / n & (129)\\
\left|\langle P_2 P_1 \mathrm{diag}(\Psi), \widehat{Z} - Z^* \rangle\right| &\leq 2\,|w_1 - w_2|\,\|\mathrm{diag}(\Psi)\|_{\max} \left\|\widehat{Z} - Z^*\right\|_1 / n & (130)\\
\left|\langle P_2 \mathrm{offd}(\Psi), \widehat{Z} - Z^* \rangle\right| &\leq 2 \sum_{j\in[n]} |Q_j|\,\widehat{w}_j, \quad \textit{and} & (131)\\
\left|\langle P_2 P_1 \mathrm{offd}(\Psi), \widehat{Z} - Z^* \rangle\right| &\leq 2\,|w_1 - w_2| \sum_k |S_k| \cdot \widehat{w}_k. & (132)
\end{aligned}$$

We prove Lemmas 61 and 56 in Sections I and F.3. Denote by $Q_1^* \geq Q_2^* \geq \ldots \geq Q_n^*$ (resp. $\widehat{w}_1^* \geq \widehat{w}_2^* \geq \ldots \geq \widehat{w}_n^*$) the non-decreasing arrangement of $|Q_i|$ (resp. $\widehat{w}_j$). Denote by $S_1^* \geq S_2^* \geq \ldots \geq S_n^*$ the non-decreasing arrangement of $|S_i|$ (resp. $\widehat{w}_j$). The RHS of (132) and (131) are bounded to be at the same order; cf. Proposition 62. Lemma 56 follows from Lemma 61, Proposition 62, and Corollary 63. We prove Proposition 62 and Corollary 63 in Sections J and K respectively. Let $C_4, C_5, c, \ldots$ be absolute constants.

Proposition 62. *Under the conditions in Theorem 6, we have with probability at least* $1 - \frac{c}{n^2}$, *simultaneously for all positive integers* $q \in [n]$,

$$\max\big(\sum_{j=1}^{q} Q_j^*, \sum_{j=1}^{q} S_j^*\big) \leq C_4 (C_0 \max_i \|H_i\|_2)^2 q\big(\sqrt{np\log(2en/q)} + \sqrt{nq}\log(2en/q)\big). \quad (133)$$

Corollary 63. *Under the settings in Proposition 62, we have with probability at least* $1 - \frac{c}{n^2}$,

$$\begin{aligned}
\sup_{\widehat{Z}\in\mathcal{M}_{opt}\cap(Z^*+r_1 B_1^{n\times n})} \left|\langle P_2 \mathrm{offd}(\Psi), \widehat{Z} - Z^* \rangle\right| &\leq 4 \sum_{j=1}^{\lceil r_1/(2n)\rceil} Q_j^* \quad \textit{and} & (134)\\
\sup_{\widehat{Z}\in\mathcal{M}_{opt}\cap(Z^*+r_1 B_1^{n\times n})} \left|\langle P_2 P_1 \mathrm{offd}(\Psi), \widehat{Z} - Z^* \rangle\right| &\leq 4\,|w_1 - w_2| \sum_{j=1}^{\lceil r_1/(2n)\rceil} S_j^*. & (135)
\end{aligned}$$

Appendix F. Additional proofs for Theorem 10

First we state Propositions 64, 65 and 66.

Proposition 64. *Let $Z_{\cdot j}$ represent the j^{th} column of Z. Let $\widehat{Z} \in \mathcal{M}_{opt}$ and $\left\|\widehat{Z} - Z^*\right\|_1 \leq r_1 \leq 2qn$, where $q = \lceil \frac{r_1}{2n} \rceil < n$ is a positive integer and Z^* is as defined in* (70). *Let*

$$\widehat{w}_j := \frac{1}{2n} \sum_{i\in[n]} \left|(\widehat{Z} - Z^*)_{ij}\right| < 1 \textit{ and hence } \sum_{j=1}^{n} \widehat{w}_j := \left\|\widehat{Z} - Z^*\right\|_1 / (2n) \leq q < n. \quad (136)$$

Hence the weight vector $\widehat{w} = (\widehat{w}_1, \ldots, \widehat{w}_n) \in B_\infty^n \cap qB_1^n$, where $B_\infty^n = \{y \in \mathbb{R}^n : |y_j| \leq 1\}$ and $B_1^n = \{y \in \mathbb{R}^n : \sum_j |y_j| \leq 1\}$ denote the unit ℓ_∞ ball and ℓ_1 ball respectively.

Proposition 65. *Under the settings in Proposition 64, we have for $\widehat{w}_j$ as in* (136)*,*

$$\begin{aligned}\forall j \in \mathcal{C}_1,\ \ \widehat{w}_j &= -\frac{1}{2n}\Big(\sum_{i\in\mathcal{C}_1}(\widehat{Z}-Z^*)_{ij} - \sum_{i\in\mathcal{C}_2}(\widehat{Z}-Z^*)_{ij}\Big) \quad \textit{and}\\ \forall j \in \mathcal{C}_2,\ \ \widehat{w}_j &= \frac{1}{2n}\Big(\sum_{i\in\mathcal{C}_1}(\widehat{Z}-Z^*)_{ij} - \sum_{i\in\mathcal{C}_2}(\widehat{Z}-Z^*)_{ij}\Big).\end{aligned}$$

Proof of Proposition 65. By definition of Z^*, we have $\forall j \in \mathcal{C}_1$,

$$\forall i \in \mathcal{C}_2, \quad (\widehat{Z}-Z^*)_{ij} \geq 0 \quad \text{and} \quad \forall i \in \mathcal{C}_1,\ (\widehat{Z}-Z^*)_{ij} \leq 0,$$

and hence

$$\widehat{w}_j = \frac{1}{2n}\Big(\sum_{i\in\mathcal{C}_2}(\widehat{Z}-Z^*)_{ij} - \sum_{i\in\mathcal{C}_1}(\widehat{Z}-Z^*)_{ij}\Big) = \Big\|(\widehat{Z}-Z^*)_{\cdot j}\Big\|_1/(2n) < 1.$$

Similarly, we have $\forall j \in \mathcal{C}_2$,

$$\forall i \in \mathcal{C}_1, \quad (\widehat{Z}-Z^*)_{ij} \geq 0 \text{ and } \quad \forall i \in \mathcal{C}_2, (\widehat{Z}-Z^*)_{ij} \leq 0,$$

and hence

$$\widehat{w}_j = \frac{1}{2n}\Big(\sum_{i\in\mathcal{C}_1}(\widehat{Z}-Z^*)_{ij} - \sum_{i\in\mathcal{C}_2}(\widehat{Z}-Z^*)_{ij}\Big) = \Big\|(\widehat{Z}-Z^*)_{\cdot j}\Big\|_1/(2n) < 1.$$

□

Proposition 66 follows from Propositions 64 and 65 immediately.

Proposition 66. *We have by symmetry of $\widehat{Z}-Z^*$, under the settings in Proposition 64,*

$$\begin{aligned}\frac{1}{2n}\left|\mathbf{1}_n^T(\widehat{Z}-Z^*)u_2\right| &= \left|\frac{1}{2n}\sum_{i=1}^n\Big(\sum_{j\in\mathcal{C}_1}(\widehat{Z}-Z^*)_{ij} - \sum_{j\in\mathcal{C}_2}(\widehat{Z}-Z^*)_{ij}\Big)\right|\\ &\leq \sum_{i=1}^n|\widehat{w}_i| = \frac{1}{2n}\Big\|(\widehat{Z}-Z^*)\Big\|_1 = \frac{1}{2n}\sum_j\Big\|(\widehat{Z}-Z^*)_{j\cdot}\Big\|_1.\end{aligned}$$

F.1 Proof outline of Lemma 53

First, we show that (116) and (117) hold. To see this, we have

$$\begin{aligned}\langle (Y-\mathbb{E}(Y))\mathbb{E}(Y)^TP_2, \widehat{Z}-Z^*\rangle &= \mathrm{tr}((\widehat{Z}-Z^*)(Y-\mathbb{E}(Y))\mathbb{E}(Y)^TP_2)\\ &= \langle P_2\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T, \widehat{Z}-Z^*\rangle. \qquad (137)\\ \langle P_2(Y-\mathbb{E}(Y))\mathbb{E}(Y)^T, \widehat{Z}-Z^*\rangle &= \mathrm{tr}((\widehat{Z}-Z^*)P_2(Y-\mathbb{E}(Y))\mathbb{E}(Y)^T)\\ &= \langle \mathbb{E}(Y)(Y-\mathbb{E}(Y))^TP_2, \widehat{Z}-Z^*\rangle. \qquad (138)\end{aligned}$$

Denote by $V = (\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T)u_2 = (V_1,\ldots,V_n)$ such that

$$\begin{aligned}\forall i \in \mathcal{C}_1, \quad V_i &= -w_2\langle \mu^{(2)}-\mu^{(1)}, \sum_{j\in C_1}(Y_j-\mathbb{E}(Y_j)) - \sum_{j\in C_2}(Y_j-\mathbb{E}(Y_j))\rangle,\\ \forall i \in \mathcal{C}_2, \quad V_i &= w_1\langle \mu^{(2)}-\mu^{(1)}, \sum_{j\in C_1}(Y_j-\mathbb{E}(Y_j)) - \sum_{j\in C_2}(Y_j-\mathbb{E}(Y_j))\rangle,\end{aligned}$$

where $Y_j \in \mathbb{R}^p$ represents the j^{th} row vector of matrix Y. Similarly, we have by (138),

$$\begin{aligned}\langle P_2(Y-\mathbb{E}(Y))\mathbb{E}(Y)^T, \widehat{Z}-Z^*\rangle &= \mathrm{tr}((\widehat{Z}-Z^*)\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T P_2)\\ &= \frac{1}{n}\sum_k V_k\big(\sum_{i\in\mathcal{C}_1}(\widehat{Z}-Z^*)_{ik} - \sum_{i\in\mathcal{C}_2}(\widehat{Z}-Z^*)_{ik}\big) =: S.\end{aligned}$$

Hence by (136) and Proposition 64, with some algebra, one obtains

$$\begin{aligned} S &= 2C_S\cdot\langle \mu^{(2)}-\mu^{(1)}, \sum_{j\in C_1}(Y_j-\mathbb{E}(Y_j)) - \sum_{j\in C_2}(Y_j-\mathbb{E}(Y_j))\rangle, \qquad (139)\\ \text{where}\quad & C_S := w_2\sum_{k\in\mathcal{C}_1}\widehat{w}_k + w_1\sum_{k\in\mathcal{C}_2}\widehat{w}_k \le \sum_{k\in[n]}\widehat{w}_k = \left\|\widehat{Z}-Z^*\right\|_1/(2n).\end{aligned}$$

Then

$$\begin{aligned}\langle P_2\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T P_2, \widehat{Z}-Z^*\rangle &= \mathrm{tr}(P_2(\widehat{Z}-Z^*)P_2\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T)\\ &= \frac{1}{2n}u_2^T(\widehat{Z}-Z^*)u_2\big(2u_2^T\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T u_2/n\big) =: -W\sum_k\widehat{w}_k, \text{ where } (140)\\ W &:= 2u_2^T\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T u_2/n = 2u_2^T V/n = \frac{2}{n}(\sum_{i\in\mathcal{C}_1}V_i - \sum_{i\in\mathcal{C}_2}V_i) \qquad (141)\\ &= \langle \mu^{(1)}-\mu^{(2)}, \sum_{j\in\mathcal{C}_1}(Y_j-\mathbb{E}Y_j) - \sum_{j\in\mathcal{C}_2}(Y_j-\mathbb{E}Y_j)\rangle 4w_1w_2.\end{aligned}$$

where (140) holds by (141) and Proposition 32, since for $P_2 = u_2u_2^T/n = Z^*/n$, $\frac{u_2^T}{2n}(\widehat{Z}-Z^*)u_2 = \mathrm{tr}(P_2(\widehat{Z}-Z^*))/2 = -\left\|\widehat{Z}-Z^*\right\|_1/(2n) = -\sum_k\widehat{w}_k$. Putting things together, we have by (136), (139) and (140),

$$\begin{aligned} T_1 &:= \langle \mathbb{E}(Y)(Y-\mathbb{E}(Y))^T P_2, \widehat{Z}-Z^*\rangle - \langle P_2\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T P_2, \widehat{Z}-Z^*\rangle\\ &= S + W\sum_k\widehat{w}_k = 2\langle \mu^{(2)}-\mu^{(1)}, \sum_{j\in C_1}(Y_j-\mathbb{E}(Y_j)) - \sum_{j\in C_2}(Y_j-\mathbb{E}(Y_j))\rangle\cdot\\ &\quad \big(w_2\sum_{k\in\mathcal{C}_1}\widehat{w}_k + w_1\sum_{k\in\mathcal{C}_2}\widehat{w}_k - 2w_1w_2\sum_{k\in[n]}\widehat{w}_k\big),\\ \text{where}\quad & \left|w_2\sum_{k\in\mathcal{C}_1}\widehat{w}_k + w_1\sum_{k\in\mathcal{C}_2}\widehat{w}_k - 2w_1w_2\sum_{k\in[n]}\widehat{w}_k\right| \le \frac{1}{2n}\left\|\widehat{Z}-Z^*\right\|_1 \le q < n.\end{aligned}$$

Now, for $w_1 = w_2 = 1/2$, the above sum is 0 and $T_1 = 0$. Let $r_1 \le 2qn$. Then, by Corollary 43, we have with probability $\ge 1-2\exp(-cn)$, for all $\widehat{Z}\in\mathcal{M}_{\mathrm{opt}}\cap(Z^*+r_1B_1^{n\times n})$,

$$\begin{aligned}|T_1| &\le \left|\langle \mu^{(2)}-\mu^{(1)}, \sum_{j\in C_1}(Y_j-\mathbb{E}(Y_j)) - \sum_{j\in C_2}(Y_j-\mathbb{E}(Y_j))\rangle\right|\left\|\widehat{Z}-Z^*\right\|_1/n\\ &\le C_4C_0(\max_i\|R_i\mu\|_2)\sqrt{p\gamma}\left\|\widehat{Z}-Z^*\right\|_1 \le C_4C_0(\max_i\|R_i\mu\|_2)r_1\sqrt{p\gamma}. \quad \square\end{aligned}$$

F.2 Proof of Lemma 55

The deterministic statement in Lemma 55 holds in view of (121), (122), and (123), since

$$\langle P_2(\widehat{\Sigma}_Y - \Sigma_Y)P_2, \widehat{Z} - Z^*\rangle = \langle P_2(I-P_1)\Psi(I-P_1)P_2, \widehat{Z} - Z^*\rangle, \text{ and hence}$$
$$\left|\langle P_2(\widehat{\Sigma}_Y - \Sigma_Y)P_2, \widehat{Z} - Z^*\rangle\right| \le \left|\langle P_2\Psi P_2, \widehat{Z} - Z^*\rangle\right| + \left|\langle P_2P_1\Psi P_1P_2, \widehat{Z} - Z^*\rangle\right| +$$
$$2\left|\langle P_2P_1\Psi P_2, \widehat{Z} - Z^*\rangle\right| \le \frac{1}{n}\left\|\widehat{Z} - Z^*\right\|_1 \|\Psi\|_2 (1+|w_1-w_2|)^2.$$

Thus, we have with probability at least $1-\exp(-c'n)$, for all $\widehat{Z} \in \mathcal{M}_{\rm opt}$,

$$\left|\langle P_2(\widehat{\Sigma}_Y - \Sigma_Y)P_2, \widehat{Z} - Z^*\rangle\right| \;\le\; 8\|\Psi\|_2 \left\|\widehat{Z} - Z^*\right\|_1/(2n) \le \frac{1}{3}\xi p\gamma \left\|\widehat{Z} - Z^*\right\|_1,$$

by Lemma 26. The lemma is thus proved. □

F.3 Proof of Lemma 56

Let $\widehat{Z} \in \mathcal{M}_{\rm opt} \cap (Z^* + r_1 B_1^{n\times n})$, where $r_1 \le 2qn$ for some positive integer $1 \le q \le n$. By Lemma 57 (cf. (119) and (120)), we have for all $\widehat{Z} \in \mathcal{M}_{\rm opt}$,

$$W_1 := \left|\langle P_2\Psi P_1, \widehat{Z} - Z^*\rangle\right| + \left|\langle P_2P_1\Psi P_1, \widehat{Z} - Z^*\rangle\right| \le (1+|w_1-w_2|)\|\Psi\|_2 \frac{1}{n}\left\|\widehat{Z} - Z^*\right\|_1.$$

Thus we have by the triangle inequality, by (127),

$$\left|\langle P_2(\widehat{\Sigma}_Y - \Sigma_Y), \widehat{Z} - Z^*\rangle\right| \le W_1 + W_2, \text{ where } W_2 := \left|\langle P_2\Psi, \widehat{Z} - Z^*\rangle\right| + \left|\langle P_2P_1\Psi, \widehat{Z} - Z^*\rangle\right|,$$

where we decompose for $\widehat{\delta}_Z := \widehat{Z} - Z^*$,

$$\begin{aligned} W_2 \;\le\; & \left|\langle P_2{\rm diag}(\Psi), \widehat{\delta}_Z\rangle\right| + \left|\langle P_2P_1{\rm diag}(\Psi), \widehat{\delta}_Z\rangle\right| \quad (W_2({\rm diag})) \\ & + \left|\langle P_2{\rm offd}(\Psi), \widehat{\delta}_Z\rangle\right| + \left|\langle P_2P_1{\rm offd}(\Psi), \widehat{\delta}_Z\rangle\right| \quad (W_2({\rm offd})). \end{aligned}$$

Then, we have for all $\widehat{Z} \in Z^* + r_1 B_1^{n\times n}$, by (120), (119), (129), and (130),

$$\begin{aligned} & W_2({\rm diag}) + W_1 = \left|\langle P_2{\rm diag}(\Psi), \widehat{Z} - Z^*\rangle\right| + \left|\langle P_2P_1{\rm diag}(\Psi), \widehat{Z} - Z^*\rangle\right| + W_1 \\ & \le \; (1+|w_1-w_2|)\max_k(|\Psi_{kk}| + \|\Psi\|_2)\left\|\widehat{Z} - Z^*\right\|_1/n \le 4\|\Psi\|_2\left\|\widehat{Z} - Z^*\right\|_1/n, \quad (142) \end{aligned}$$

and by (134) and (135), for $W_2({\rm offd}) = \left|\langle P_2{\rm offd}(\Psi), \widehat{Z} - Z^*\rangle\right| + \left|\langle P_2P_1{\rm offd}(\Psi), \widehat{Z} - Z^*\rangle\right|$,

$$\sup_{\widehat{Z}\in\mathcal{M}_{\rm opt}\cap(Z^*+r_1B_1^{n\times n})} W_2({\rm offd}) \;\le\; 4 \sup_{\widehat{w}\in B_\infty^n \cap \lceil \frac{r_1}{2n}\rceil B_1^n} \Big(\sum_{j=1}^{\lceil \frac{r_1}{2n}\rceil} Q_j^* + |w_1-w_2| \sum_{j=1}^{\lceil \frac{r_1}{2n}\rceil} S_j^*\Big). \quad (143)$$

Let $\Psi = \mathbb{Z}\mathbb{Z}^T - \mathbb{E}(\mathbb{Z}\mathbb{Z}^T)$ and $q = \lceil r_1/(2n)\rceil$. The lemma thus holds by Proposition 62 and Corollary 63, (142) and (143), since

$$\sup_{\widehat{Z}\in\mathcal{M}_{\rm opt}\cap(Z^*+r_1B_1^{n\times n})} \left|\langle P_2(\widehat{\Sigma}_Y - \Sigma_Y), \widehat{Z} - Z^*\rangle\right| \le \sup_{\widehat{Z}\in\mathcal{M}_{\rm opt}\cap(Z^*+r_1B_1^{n\times n})} 4\|\Psi\|_2\left\|\widehat{Z} - Z^*\right\|_1/n$$
$$+ \;\; 4 \sup_{\widehat{w}\in B_\infty^n \cap \lceil r_1/(2n)\rceil B_1^n} \Big(\sum_{j=1}^{\lceil r_1/(2n)\rceil} Q_j^* + |w_1-w_2| \sum_{j=1}^{\lceil r_1/(2n)\rceil} S_j^*\Big). \quad \square$$

F.4 Proof of Lemma 58

The proof of Lemma 58 is entirely deterministic. Recall by (48), we have

$$\frac{1}{n}\mathbb{E}(Y)^T u_2 = \frac{1}{n}(\sum_{i\in\mathcal{C}_1}\mathbb{E}(Y_i) - \sum_{i\in\mathcal{C}_2}\mathbb{E}(Y_i)) = 2w_1w_2(\mu^{(1)}-\mu^{(2)}).$$

Throughout this proof, we denote by $\mu := (\mu^{(1)}-\mu^{(2)})/\sqrt{p\gamma}$. Let Z^* be as defined in (70). Thus we have by symmetry of $\widehat{Z}-Z^*$ and definition of $\widehat{w}_j$ in (136) and $P_2 := u_2u_2^T/n = Z^*/n$,

$$\begin{aligned}\langle P_2(\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T), \widehat{Z}-Z^*\rangle &= \frac{1}{n}u_2^T(\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T)(\widehat{Z}-Z^*)u_2\\ &= 4nw_1w_2\sqrt{p\gamma}\big(\sum_{j\in\mathcal{C}_2}\widehat{w}_j\langle\mu, Y_j-\mathbb{E}(Y_j)\rangle - \sum_{j\in\mathcal{C}_1}\widehat{w}_j\langle\mu, Y_j-\mathbb{E}(Y_j)\rangle\big),\end{aligned}$$

Moreover, we have for each $j\in[n]$, $n\langle\mu, Y_j-\mathbb{E}(Y_j)\rangle = n\langle\mu,\mathbb{Z}_j\rangle - \sum_{i=1}^n\langle\mu,\mathbb{Z}_i\rangle$. Hence we have by (136), and $L_j = \sqrt{p\gamma}\langle\mu, Z_j\rangle$,

$$\langle P_2(\mathbb{E}(Y)(Y-\mathbb{E}(Y))^T), \widehat{Z}-Z^*\rangle \le 4w_1w_2n\sum_{j\in[n]}|L_j|\,\widehat{w}_j + \frac{4w_1w_2}{2n}\left|\sum_{i=1}^n L_i\right|\left\|(\widehat{Z}-Z^*)\right\|_1. \qquad \square$$

Appendix G. Proof of Lemma 59

Let $1\le q<n$ denote a positive integer. By Proposition 64, we have $\widehat{w}=(\widehat{w}_1,\dots,\widehat{w}_n)\in B_\infty^n\cap\lceil\frac{r_1}{2n}\rceil B_1^n$, for all $\widehat{Z}\in\mathcal{M}_{\rm opt}\cap(Z^*+r_1B_1^{n\times n})$. Then for the first component on the RHS of (125) in Lemma 58, we have

$$\begin{aligned}4w_1w_2\sum_{j\in[n]}|L_j|\,\widehat{w}_j &\le \sum_{j\in[n]}L_j^*\widehat{w}_j^* = \sum_{j=1}^q L_j^*\widehat{w}_j^* + \sum_{j=q+1}^n L_j^*\widehat{w}_j^*, \text{ and}\\ \sup_{\widehat{w}\in B_\infty^n\cap qB_1^n}\sum_{j\in[n]}|L_j|\,\widehat{w}_j &\le \sup_{\widehat{w}\in B_\infty^n}\sum_{j=1}^q L_j^*\widehat{w}_j^* + \sup_{\widehat{w}\in qB_1^n}\sum_{j=q+1}^n L_j^*\widehat{w}_j^*\\ &\le \sum_{j=1}^q L_j^* + qL_{q+1}^* \le 2\sum_{j=1}^q L_j^*. \qquad (144)\end{aligned}$$

Moreover, for the second component on the RHS of (125), $W_L := \left|\sum_{j=1}^n L_j\right|\left\|(\widehat{Z}-Z^*)\right\|_1/(2n) \le \sum_{j=1}^n|L_j|\left|\sum_j\widehat{w}_j\right| =: U_L$, where by (136), $\left\|(\widehat{Z}-Z^*)\right\|_1/(2n) = \left|\sum_j\widehat{w}_j\right|$, and hence

$$\sup_{\widehat{Z}\in\mathcal{M}_{\rm opt}\cap Z^*+r_1B_1^{n\times n}} W_L/n \le \sup_{\widehat{w}\in B_\infty^n\cap\lceil\frac{r_1}{2n}\rceil B_1^n} U_L/n \le \frac{\lceil r_1/2n\rceil}{n}\sum_{j=1}^n L_j^* \le \sum_{j=1}^{\lceil r_1/(2n)\rceil} L_j^*. \qquad (145)$$

where (145) holds since in the second sum $\sum_{j=1}^q L_j^*$, where $q=\lceil\frac{r_1}{2n}\rceil$, we give a unit weight to each of the q largest components $L_j^*, j\in[q]$, while in $\frac{q}{n}\sum_{i=1}^n L_j^*$, we give the same

weight $q/n \leq 1$ to each of n components $L_j^*, j \in [n]$. Putting things together, we have by (125), (145), and (144), where we set $q = \lceil \frac{r_1}{2n} \rceil \in [n]$,

$$\sup_{\widehat{Z} \in \mathcal{M}_{\text{opt}} \cap (Z^* + r_1 B_1^{n \times n})} \left| \langle P_2(\mathbb{E}(Y)(Y - \mathbb{E}(Y))^T), \widehat{Z} - Z^* \rangle \right|$$

$$\leq \sup_{\widehat{w} \in B_\infty^n \cap \lceil \frac{r_1}{2n} \rceil B_1^n} n \sum_{j \in [n]} |L_j| \, \widehat{w}_j + \sup_{\widehat{w} \in B_\infty^n \cap \lceil \frac{r_1}{2n} \rceil B_1^n} \left| \sum_{j=1}^n L_j \right| \sum_j \widehat{w}_j \leq 3 \sum_{j=1}^{\lceil \frac{r_1}{2n} \rceil} L_j^*. \quad \square$$

Appendix H. Proof of Proposition 60

First, we bound for any integer $q \in [n]$, and $\tau > 0$,

$$\mathbb{P}\left(\sum_{j=1}^q L_j^* > \tau \right) \leq \mathbb{P}\left(\exists J \subset [n], |J| = q, \sum_{j \in J} |L_j| > \tau \right) \tag{146}$$

$$\leq \sum_{J : |J| = q} \sum_{u \in \{-1,1\}^q} \mathbb{P}\left(\sum_{j \in J} u_j L_j / \Delta > \tau / \Delta \right) \leq \binom{n}{q} 2^q \exp\left(- \frac{(\tau/\Delta)^2}{C_1 C_0^2 (\max_j \|R_j \mu\|_2^2) q} \right),$$

where (146) holds by the union bound and the subgaussian tail bound; to see this, notice that for each fixed index set $J \subset [n]$ and vector $u \in \{-1, 1\}^q$, and $\Delta = \sqrt{p\gamma}$,

$$\left\| \sum_{j \in J, |J| = q} \frac{u_j L_j}{\Delta} \right\|_{\psi_2}^2 \leq \sum_{j \in J, |J| = q} \|L_j / \Delta\|_{\psi_2}^2 =: \sum_{j \in J, |J| = q} \|\langle \mu, \mathbb{Z}_j \rangle\|_{\psi_2}^2 \leq C_1 q (C_0 \max_j \|R_j \mu\|_2)^2,$$

following the proof of Lemma 42; hence by Lemma 42, upon replacing n with an integer $1 \leq q < n$, we have for each fixed index set $J \subset [n]$, $\mu \in \mathbb{S}^{p-1}$,

$$\mathbb{P}\left(\sum_{j \in J : |J| = q} u_j L_j / \Delta > \tau / \Delta \right) \leq \exp\left(- \frac{(\tau/\Delta)^2}{q C_1 (C_0 \max_j \|R_j \mu\|_2)^2} \right). \tag{147}$$

Set $\tau_q = C_3 C_0 (\max_j \|R_j \mu\|_2) \Delta q \sqrt{\log(2en/q)}$ for some absolute constant $C_3 > c\sqrt{C_1}$ for $c \geq 2$. Then by (146) and (147), we have

$$\begin{aligned} \mathbb{P}\left(\sum_{j=1}^q L_j^* > \tau_q \right) &\leq \binom{n}{q} 2^q \exp\left(- \frac{C_3^2 (C_0 \max_j \|R_j \mu\|_2)^2 q^2 \log(2en/q)}{C_1 (C_0 \max_j \|R_j \mu\|_2)^2 q} \right) \\ &\leq e^{q \log(2en/q)} e^{-cq \log(2en/q)} \leq e^{-c' q \log(2en/q)}, \end{aligned}$$

where $\binom{n}{q} 2^q \leq (en/q)^q 2^q = (2en/q)^q = \exp(q \log(2en/q))$. Following the calculation as done in Giraud and Verzelen (2019), and the inequality immediately above, we have

$$\sum_{q=1}^n \binom{n}{q} 2^q e^{-\tau_q^2 / (C_1 C_0^2 q \Delta^2)} \leq \sum_{q=1}^n e^{-c' q \log(2en/q)} \leq \frac{c}{n^2}. \qquad \square$$

Appendix I. Proof of Lemma 61

Denote by $\widehat{\delta}_Z := \widehat{Z} - Z^*$. First, we have $\langle P_2P_1\Psi, \widehat{\delta}_Z\rangle = \mathrm{tr}(\widehat{\delta}_Z P_2P_1\Psi) = (w_1-w_2)\mathbf{1}_n^T\Psi\widehat{\delta}_Z\frac{u_2}{n}$, where $\frac{\widehat{\delta}_Z u_2}{n} = (-\widehat{w}_1,\ldots,-\widehat{w}_{n_1},\widehat{w}_{n_1+1},\ldots,\widehat{w}_n)$ for u_2 as in (4). Now

$$\begin{aligned}
\left|\langle P_2P_1\mathrm{offd}(\Psi), \widehat{\delta}_Z\rangle\right| &= |w_1-w_2|\left|\mathbf{1}_n^T\mathrm{offd}(\Psi)\widehat{\delta}_Z\frac{u_2}{n}\right| \le 2\,|w_1-w_2|\sum_k |S_k|\,\widehat{w}_k;\\
\left|\langle P_2P_1\mathrm{diag}(\Psi), \widehat{\delta}_Z\rangle\right| &= |S_{\mathrm{diag}}| \le |w_1-w_2|\sum_{k\in[n]}|\Psi_{kk}|\,\widehat{w}_k \le |w_1-w_2|\max_k|\Psi_{kk}|\left\|\widehat{\delta}_Z\right\|_1/n;
\end{aligned}$$

Hence (132) and (130) hold. Now (129) and (131) hold since

$$\begin{aligned}
&\langle P_2\Psi, \widehat{Z}-Z^*\rangle = \mathrm{tr}((\widehat{Z}-Z^*)P_2\Psi) = \mathrm{tr}(P_2\Psi(\widehat{Z}-Z^*))\\
&= 2\Big(\sum_{k\in\mathcal{C}_1}\Psi_{kk}(-\widehat{w}_k) - \sum_{t\in\mathcal{C}_2}\Psi_{tt}\widehat{w}_t\Big) + 2\Big(\sum_{j\in\mathcal{C}_1}Q_j(-\widehat{w}_j) + \sum_{j\in\mathcal{C}_2}Q_j\widehat{w}_j\Big), \quad \text{where}\\
&\left|\langle P_2\mathrm{diag}(\Psi), \widehat{Z}-Z^*\rangle\right| = 2\left|\sum_{k\in C_1}\Psi_{kk}(-\widehat{w}_k) + \sum_{k\in C_2}(-\Psi_{kk})\widehat{w}_k\right| \le 2\sum_{k\in[n]}|\Psi_{kk}|\,\widehat{w}_k \quad \text{and}\\
&\left|\langle P_2\mathrm{offd}(\Psi), \widehat{Z}-Z^*\rangle\right| = 2\left|\sum_{j\in\mathcal{C}_1}Q_j(-\widehat{w}_j) + \sum_{j\in\mathcal{C}_2}Q_j\widehat{w}_j\right| \le 2\sum_{j\in[n]}|Q_j|\,\widehat{w}_j. \quad \square
\end{aligned}$$

Appendix J. Proof of Proposition 62

Denote by $\Psi = \mathbb{Z}\mathbb{Z}^T - \mathbb{E}(\mathbb{Z}\mathbb{Z}^T)$. The two bounds follow identical steps, and hence we will only prove (133). First, we bound for any positive integer $1 \le q < n$ and any $\tau > 0$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{j=1}^q Q_j^* > \tau\right) &= \mathbb{P}\left(\exists J\subset[n], |J|=q, \sum_{j\in J}|Q_j| > \tau\right)\\
&\le \sum_{J:|J|=q}\sum_{y\in\{-1,1\}^q}\mathbb{P}\left(\sum_{j\in J}y_jQ_j > \tau\right) = q_0(\tau). \qquad (148)
\end{aligned}$$

We now bound $q_0(\tau)$. Denote by $w\in\{0,-1,1\}^n$ the 0-extended vector of $y\in\{-1,1\}^{|J|}$, that is, for the fixed index set $J\subset[n]$, we have $w_j = y_j, \forall j\in J$ and $w_j = 0$ otherwise. Now we can write for $A = (a_{tj}) = w\otimes u_2$, where u_2 is the group membership vector: $\forall t\in\mathcal{C}_1, j\in J,\quad a_{jt} = u_{2,t}y_j = y_j$ and $\forall t\in\mathcal{C}_2, j\in J,\quad a_{jt} = u_{2,t}y_j = -y_j$; otherwise, $a_{jt} = 0$. In other words, we have nq non-zero entries in A and each entry has absolute value 1. Thus $\|A\|_2 \le \|A\|_F = \sqrt{qn}$ and Q_j as in (128),

$$\sum_{j\in J}y_jQ_j = \sum_{j\in J}y_j\left(\sum_{t\in\mathcal{C}_1,t\neq j}\langle\mathbb{Z}_t,\mathbb{Z}_j\rangle - \sum_{t\in\mathcal{C}_2,t\neq j}\langle\mathbb{Z}_t,\mathbb{Z}_j\rangle\right) = \sum_{j\in[n]}\sum_{t\neq j}a_{jt}\langle\mathbb{Z}_t,\mathbb{Z}_j\rangle,$$

since each row of offd(A) in J has $n-1$ nonzero entries and $|J|=q$. Now for each fixed index set $J\subset[n]$ with $|J|=q$, and a fixed vector $y\in\{-1,1\}^q$,

$$\mathbb{P}\left(\sum_{j\in J} y_j Q_j > \tau_q\right) \;=\; \mathbb{P}\left(\sum_j \sum_{t\neq j} a_{jt}\langle \mathbb{Z}_t, \mathbb{Z}_j\rangle > \tau_q\right).$$

By Theorem 4, we have for $q\in[n]$, $A=(a_{ij})\in\mathbb{R}^{n\times n}$, $\sigma_{\max}:=\max_i \|H_i\|_2$, and $\tau_q = C_4(C_0\sigma_{\max})^2 q(\sqrt{np\log(2en/q)}+\sqrt{nq}\log(2en/q))$, for $C_0\asymp K$ by definition of K in (14),

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^n\sum_{j\neq i}^n \langle \mathbb{Z}_i,\mathbb{Z}_j\rangle a_{ij}\right| > \tau_q\right) &\le 2\exp\left(-c\min\left(\frac{\tau_q^2}{(K\sigma_{\max})^4 p\,\|A\|_F^2}, \frac{\tau_q}{(K\sigma_{\max})^2\,\|A\|_2}\right)\right)\\ &\le 2\exp\left(-c(C_4^2\wedge C_4) q\log(2en/q)\right), \qquad (149)\end{aligned}$$

where $\max_i \|Z_i\|_{\psi_2}\le CK\sigma_{\max}$ in the sense of (93). Now by (148) and (149),

$$q_0(\tau_q) \;\le\; \binom{n}{q} 2^q \exp\left(-c(C_4^2\wedge C_4) q\log(2en/q)\right) \le 2\exp\left(-c'(C_4^2\wedge C_4) q\log(2en/q)\right),$$

where $\binom{n}{q}2^q \le (en/q)^q 2^q = (2en/q)^q = \exp(q\log(2en/q))$. Hence following the calculation as done in Giraud and Verzelen (2019), we have

$$\sum_{q=1}^n q_0(\tau_q) \le \sum_{q=1}^n \binom{n}{q} 2^q \exp\left(-c(C_4^2\wedge C_4) q\log(2en/q)\right) \le \sum_{q=1}^n e^{-c' q\log(2en/q)} \le \frac{c'}{n^2}.$$

The proposition thus holds. □

Appendix K. Proof of Corollary 63

For any positive integer $q\in[n]$ and $\widehat{w}$ as in (124),

$$\begin{aligned}\sup_{\widehat{w}\in B_\infty^n\cap qB_1^n} \sum_{j\subset[n]} |Q_j|\,\widehat{w}_j &\le \sup_{\widehat{w}\in B_\infty^n\cap qB_1^n} \sum_{j\in[n]} Q_j^*\widehat{w}_j^* = \sup_{\widehat{w}\in B_\infty^n\cap qB_1^n} \Big(\sum_{j=1}^q Q_j^*\widehat{w}_j^* + \sum_{j=q+1}^n Q_j^*\widehat{w}_j^*\Big)\\ &\le \sum_{j=1}^q Q_j^* + qQ_{q+1}^* \le 2\sum_{j=1}^q Q_j^*. \qquad (150)\end{aligned}$$

Following the same arguments in Lemma 54, we have by (131) and (124),

$$\sup_{\widehat{Z}\in M_{\rm opt}\cap Z^*+r_1B_1^{n\times n}} \left|\langle P_2\mathrm{offd}(\Psi), \widehat{Z}-Z^*\rangle\right| \;\le\; 2\sup_{\widehat{w}\in B_\infty^n\cap\lceil r_1/(2n)\rceil B_1^n} \sum_{j\in[n]} |Q_j|\,\widehat{w}_j$$

Now (134) follows from (131), (150), and (133), where we set $q=\lceil r_1/(2n)\rceil$. Notice that for any positive integer $q\in[n]$, following the same argument, we have

$$\sup_{\widehat{w}\in B_\infty^n\cap qB_1^n} \sum_{j\in[n]} |S_j|\,\widehat{w}_j \;\le\; \sup_{\widehat{w}\in B_\infty^n\cap qB_1^n} \sum_{j\in[n]} S_j^*\widehat{w}_j^* \le 2\sum_{j=1}^q S_j^*. \qquad (151)$$

Hence (135) follows from (132), (151), (133), and Proposition 62. □

References

E. Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

E. Abbe, J. Fan, and K. Wang. An ℓ_p theory of PCA and spectral clustering. *Ann. Statist.*, 50(4):2359–2385, 2022.

D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the 18th Annual COLT*, pages 458–469, 2005.

D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75:245–248, 2009.

S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of 33rd ACM Symposium on Theory of Computing*, pages 247–257, 2001.

S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45:77–120, 2017.

A. Blum, A. Coja-Oghlan, A. Frieze, and S. Zhou. Separating populations with wide data: a spectral analysis. *Electronic Journal of Statistics*, 3:76–113, 2009.

F. Bunea, C. Giraud, X. Luo, M. Royer, and N. Verzelen. Model assisted variable clustering: minimax-optimal recovery and algorithms. *The Annals of Statistics*, 48(1):111–137, 2020.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

K. Chaudhuri, E. Halperin, S. Rao, and S. Zhou. A rigorous analysis of population stratification with limited data. In *Proceedings of the 18th ACM-SIAM SODA*, 2007.

S. Chrétien, M. Cucuringu, G. Lecué, and L. Neirac. Learning with semi-definite programming: statistical bounds based on fixed point analysis and excess risk curvature. *Journal of Machine Learning Research*, 22(230), 2021.

S. Dasgupta and L. J. Schulman. A two-round variant of EM for Gaussian mixtures. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000.

N. Doss, Y. Wu, P. Yang, and H. H. Zhou. Optimal estimation of high-dimensional Gaussian location mixtures. *The Annals of Statistics*, 51(1):62–95, 2023.

P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56:9–33, 2004.

Y. Fei and Y. Chen. Hidden integrality of SDP relaxations for sub-Gaussian mixture models. In *Conference On Learning Theory*, pages 1931–1965, 2018.

Y. Fei and Y. Chen. Exponential error rates of SDP for block models: beyond Grothendieck's inequality. *IEEE Trans. Inform. Theory*, 65(1):551–571, 2019.

Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23 (2):298–305, 1973. URL `http://eudml.org/doc/12723`.

A. M. Frieze and R. Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–200, 1999.

C. Giraud and N. Verzelen. Partial recovery bounds for clustering with the relaxed k-means. *Mathematical Statistics and Learning*, 1:317–374, 2019.

M. Goemans and D. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *JACM*, 42(6):1115–1145, 1995.

O. Guedon and R. Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165:1025–1049, 2016.

D. L. Hanson and E. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42:1079–1083, 1971.

R. Horn and C. Johnson. *Topics in Matrix Analysis*. Cambridge University Press; Reprint edition, 1991.

M. Hornstein, R. Fan, K. Shedden, and S. Zhou. Joint mean and covariance estimation for unreplicated matrix-variate data. *Journal of the American Statistical Association (Theory and Methods)*, 114(526):682–696, 2019.

T. Iguchi, D. G. Mixon, J. Peterson, and S. Villar. Probably certifiably correct k-means clustering. *Mathematical Programming*, 165:605–642, 2017.

A. Javanmard, A. Montanari, and F. Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *Proc. Natl. Acad. Sci. USA*, 113(16):E2218–E2223, 2016.

A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, pages 553–562. ACM, 2010.

R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proc. of the 18th Annual COLT*, 2005.

R. M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller and James W. Thatcher, editors, *Proceedings of a symposium on the Complexity of Computer Computations, held March 20-22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA*, The IBM Research Symposia Series, pages 85–103. Plenum Press, New York, 1972.

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

J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

X. Li, Y. Li, S. Ling, T. Strohmer, and K. Wei. When do birds of a feather flock together? k-means, proximity, and conic programming. *Mathematical Programming*, 179(1-2, Ser. A):295–341, 2020.

M. Löffler, A. Zhang, and H. Zhou. Optimality of spectral clustering in the Gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530, 2021.

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27:1808–1829, 1999.

R. D. C. Monteiro. First- and second-order methods for semidefinite programming. *Mathematical Programming*, 97(1–2):209–244, 2003. doi: 10.1007/s10107-003-0451-1.

M. Ndaoud. Sharp optimal recovery in the two component Gaussian mixture model. *Ann. Statist.*, 50(4):2096–2126, 2022.

M.L. Overton and R.S. Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62: 321–357, 1993.

N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12), 2006. doi:10.1371/journal.pgen.0020190.

J. Peng and Y. Wei. Approximating K-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18:186–205, 2007.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, 39(4):1878–1915, 2011.

M. Royer. Adaptive clustering through semidefinite programming. *Advances in Neural Information Processing Systems*, pages 1795–1803, 2017.

M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

M. Rudelson and S. Zhou. Errors-in-variables models with dependent measurements. *Electron. J. Statist.*, 11(1):1699–1797, 2017.

V. Vempala and G. Wang. A spectral algorithm of learning mixtures of distributions. In *Proceedings of the 43rd IEEE FOCS*, pages 113–123, 2002.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science (Second Edition)*. Cambridge University Press, 2026.

Y. Yu, T. Wang, and R.J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102:315–323, 2015.

A. Yurtsever, J. A. Tropp, O. Fercoq, M. Udell, and V. Cevher. Scalable semidefinite programming. *SIAM Journal on Mathematics of Data Science*, 3(1):171–200, 2021. doi: 10.1137/19M1305045.

A. Y. Zhang and H. Y. Zhou. Leave-one-out singular subspace perturbation analysis for spectral clustering. *The Annals of Statistics*, 52(5):2004–2033, 2024.

S. Zhou. Routing, disjoint paths, and classification. *PhD Thesis, Carnegie Mellon University*, 2006.

S. Zhou. Sparse Hanson-Wright inequalities for subgaussian quadratic forms. *Bernoulli*, 25 (3):1603–1639, 2019.

S. Zhou. Semidefinite programming on population clustering: a global analysis. *arXiv Preprint*, 2023. https://arxiv.org/pdf/2301.00344.pdf.

S. Zhou. Concentration of measure bounds for matrix-variate data with missing values. *Bernoulli*, 30(1):198–226, 2024a.

S. Zhou. Debiasing and a local analysis for population clustering using semidefinite programming. *arXiv Preprint*, 2024b. https://arxiv.org/pdf/2401.10927v1.

S. Zhou. Semidefinite programming relaxations and debiasing for MAXCUT-based clustering. *arXiv Preprint*, 2026. https://arxiv.org/pdf/2401.10927v3.

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