Submitted to Bernoulli
arXiv: arXiv:1810.11180

Hanson-Wright inequality in Hilbert spaces with application to K-means clustering for non-Euclidean data

XIAOHUI CHEN[1,2,*,**] and YUN YANG[1,†]

[1] *Department of Statistics*
University of Illinois at Urbana-Champaign
725 S. Wright Street, Champaign, IL 61820, USA.
E-mail: [*] xhchen@illinois.edu; [†] yy84@illinois.edu

[2] *Institute for Data, Systems, and Society*
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA, 02139-4307, USA.
E-mail: [**] xiaohui@mit.edu
*

We derive a dimension-free Hanson-Wright inequality for quadratic forms of independent sub-gaussian random variables in a separable Hilbert space. Our inequality is an infinite-dimensional generalization of the classical Hanson-Wright inequality for finite-dimensional Euclidean random vectors. We illustrate an application to the generalized K-means clustering problem for non-Euclidean data. Specifically, we establish the exponential rate of convergence for a semidefinite relaxation of the generalized K-means, which together with a simple rounding algorithm imply the exact recovery of the true clustering structure.

Keywords: Hanson-Wright inequality, Hilbert space, K-means, semidefinite relaxation.

1. Introduction

The Hanson-Wright inequality is a fundamental tool for studying the concentration phenomenon for quadratic forms in sub-gaussian random variables [11, 31]. Recently, it has triggered a wide range of statistical applications such as semidefinite programming (SDP) relaxations for K-means clustering [21, 10] and Gaussian approximation bounds for high-dimensional U-statistics (of order two) [6]. Classical form of the Hanson-Wright inequality bounds the tail probability for the quadratic form of a finite-dimensional random vector in a Euclidean space. Below is a version that is frequently cited in literature (cf. Theorem 1.1 in [22]).

[*] *MSC 2010 classification*: 60F10, 62H30.

Theorem 1.1 (Hanson-Wright inequality for quadratic forms of independent sub-gaussian random variables in $\mathbb{R}$)**.** *Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent components X_i such that $\mathbb{E}[X_i] = 0$ and $\|X_i\|_{\psi_2} := \sup_{q \geqslant 1} q^{-1/2} (\mathbb{E}\,|X_i|^q)^{1/q} \leqslant L$. Let A be an $n \times n$ matrix. Then there exists a universal constant $C > 0$ such that for every $t > 0$,*

$$\mathbb{P}(|X^T A X - \mathbb{E}[X^T A X]| \geqslant t) \leqslant 2 \exp\left[-C \min\left(\frac{t^2}{L^4 \|A\|_{HS}^2}, \frac{t}{L^2 \|A\|_{op}}\right)\right], \tag{1.1}$$

where $\|A\|_{HS} = (\sum_{i,j=1}^n a_{ij}^2)^{1/2}$ is the Hilbert-Schmidt (i.e., Frobenius) norm of A and $\|A\|_{op} = \max_{\{x \in \mathbb{R}^n : \|x\|_2 = 1\}} \|Ax\|_2$ is the $\ell_2 \to \ell_2$ operator (i.e., spectral) norm of A.

There are some variants of the finite-dimensional Hanson-Wright inequality. Sharp upper and lower tail inequalities for quadratic forms of independent Gaussian random variables are derived in [15]. [20] and [4] derive the Hanson-Wright inequality for zero-diagonal matrix A with independent Bernoulli and centered sub-gaussian random variables, respectively. [13] establishes an upper tail inequality for positive semidefinite quadratic forms in a sub-gaussian random vector with dependent components. [29] proves a dimension-dependent concentration inequality for a centered random vector under the convex concentration property. [1] further improves the inequality of [29] by removing the dimension dependence in $\mathbb{R}^n$.

In this paper, we first derive an infinite-dimensional analog of the Hanson-Wright inequality (1.1) for sub-gaussian random variables taking values in a Hilbert space, which can be seen as a unified generalization of the aforementioned papers in finite dimensions. Motivation of deriving the dimension-free Hanson-Wright inequality stems from the generalized K-means clustering for non-Euclidean data with non-linear features, which covers the functional data clustering and kernel clustering as special examples. It is well-known that the (classical) Euclidean distance based K-means clustering is computationally NP-hard in the worst case. Various SDP relaxations in literature (cf. [18, 16, 7, 21, 10]) aim to provide exact and partial recovery of the true clustering structure. However, it remains a challenging task to provide strong statistical guarantees for computationally tractable (i.e., polynomial-time) algorithms to cluster non-Euclidean data taking values in a general Hilbert space with non-linear features. As we shall see in Section 3, the Hilbert space version of the Hanson-Wright inequality offers a powerful tool to establish the exponential rate of convergence for an SDP relaxation of the generalized K-means. This partial recovery bound implies the exact recovery of the generalized K-means clustering via a simple rounding algorithm. In contrast to the heuristic greedy algorithms often employed in the kernel clustering setting (cf. [24]), our result provides a principled SDP relaxed kernel clustering algorithm with exact recovery guarantees.

2. Hanson-Wright inequality in Hilbert spaces

To state the Hanson-Wright inequality in a general Hilbert space, we first need to properly specify the sub-gaussian random variables therein.

2.1. Sub-gaussian random variables in Hilbert spaces

Let $\mathbb{H}$ be a real separable Hilbert space and $\mathcal{B}(\mathbb{H})$ be the class of bounded linear operators $\Sigma : \mathbb{H} \to \mathbb{H}$. If the operator $\Sigma \in \mathcal{B}(\mathbb{H})$ is positive definite (i.e., it is self-adjoint $\Sigma^* = \Sigma$ and $\langle \Sigma z, z\rangle \geqslant 0$ for all $z \in \mathbb{H}$), then there is a unique positive definite (and thus self-adjoint) square root operator $\Sigma^{1/2} \in \mathcal{B}(\mathbb{H})$ satisfying $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$ (cf. Theorem 3.4.3 in [12]).

***Definition* 2.1** (Trace class of linear operators on a separable Hilbert space)**.** Let $\Sigma \in \mathcal{B}(\mathbb{H})$. Then Σ is *trace class* if

$$\|\Sigma\|_{\mathrm{tr}} := \sum_{j=1}^{\infty} \langle (\Sigma^*\Sigma)^{1/2} e_j, e_j\rangle < \infty,$$

where $(e_j)_{j=1}^{\infty}$ is a *complete orthonormal system* (CONS) of $\mathbb{H}$. In this case, $\|\Sigma\|_{\mathrm{tr}}$ is the *trace norm* of Σ.

Note that the trace norm does not depend on the choice of the CONS. A self-adjoint and positive definite trace class linear operator Σ is compact and it plays a similar role as a covariance matrix, where the trace norm is simply the trace of the covariance matrix. In particular, if Σ is positive definite trace class, then $\|\Sigma\|_{\mathrm{tr}} = \sum_{j=1}^{\infty} \langle \Sigma e_j, e_j\rangle = \sum_{j=1}^{\infty} \|\Sigma^{1/2} e_j\|^2$. Let $(\Omega, \mathscr{B}, \mathbb{P})$ be a probability space.

***Definition* 2.2** (Hilbert space valued sub-gaussian random variable)**.** Let Z be a random variable in $\mathbb{H}$ and $\Gamma : \mathbb{H} \to \mathbb{H}$ be a positive definite trace class linear operator. Then Z is *sub-gaussian with respect to* Γ (denote as $Z \sim$ sub-gaussian(Γ)) if there exists an $\alpha \geqslant 0$ such that for all $z \in \mathbb{H}$,

$$\mathbb{E}\left[e^{\langle z, Z - \mathbb{E}[Z]\rangle}\right] \leqslant e^{\alpha^2 \langle \Gamma z, z\rangle/2}, \tag{2.1}$$

where the expectation $\mathbb{E}[Z] = \int_\Omega Z d\mathbb{P}$ is defined as a Bochner integral (cf. Chapter 2.6 in [12]). Moreover, if $Z \sim$ sub-gaussian(Γ) with mean $\mu = \mathbb{E}[Z]$, then the ψ_2 *(or sub-gaussian) norm* of Z with respect to Γ is defined as

$$\|Z\|_{\psi_2,\Gamma} = \inf\left\{\alpha \geqslant 0 : \mathbb{E}\left[e^{\langle z, Z-\mu\rangle}\right] \leqslant e^{\alpha^2\langle \Gamma z, z\rangle/2} \quad \forall z \in \mathbb{H}\right\}.$$

Note that Definition 2.2 corresponds to the R-sub-gaussianity in [2], and it is an infinite-dimensional analog of the sub-gaussian random vectors in $\mathbb{R}^p$ (see for example [28] and [13]). Unsurprisingly, the Gaussian random variables in $\mathbb{H}$ is a special case of sub-gaussian random variables in $\mathbb{H}$.

***Definition* 2.3** (Hilbert space valued Gaussian random variable)**.** A random variable Z in $\mathbb{H}$ is *Gaussian with respect to* Γ and with mean $\mu = \mathbb{E}[Z]$ (denote as $Z \sim N(\mu, \Gamma)$) if for all $z \in \mathbb{H}$,

$$\mathbb{E}\left[e^{\langle z, Z-\mu\rangle}\right] = e^{\langle \Gamma z, z\rangle/2}. \tag{2.2}$$

***Lemma* 2.4.** If $Z \sim N(\mu, \Gamma)$, then $\|Z\|_{\psi_2,\Gamma} = 1$ and $\Sigma = \Gamma$, where $\Sigma := \mathbb{E}[(Z-\mu)\otimes(Z-\mu)]$ is the covariance operator of Z. More generally, if $Z \sim$ sub-gaussian(Γ) with mean $\mu = \mathbb{E}[Z]$, then $\Sigma \preceq 4\|Z\|^2_{\psi_2,\Gamma}\Gamma$, i.e., $(4\|Z\|^2_{\psi_2,\Gamma}\Gamma - \Sigma)$ is positive semidefinite.

For $a, b \in \mathbb{H}$, the tensor product $a \otimes b : \mathbb{H} \to \mathbb{H}$ is a linear operator defined as $(a \otimes b)z = \langle b, z\rangle a$ for all $z \in \mathbb{H}$. Lemma 2.4 is proved in Appendix A.2.

Notation. We shall use $c, c_0, c_1, C, C_0, C_1, \ldots$ to denote positive and finite universal constants, whose values may vary from place to place. For $a, b \in \mathbb{R}$, denote $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For $\Sigma \in \mathcal{B}(\mathbb{H})$, the operator norm $\|\Sigma\|_{\mathrm{op}}$ of Σ is defined as the square root of the largest eigenvalue of $\Sigma^*\Sigma$. If $\sum_{j=1}^{\infty}\|\Sigma e_j\|^2 < \infty$, then Σ is a Hilbert-Schmidt (HS) operator and $\|\Sigma\|_{\mathrm{HS}} = (\sum_{j=1}^{\infty}\|\Sigma e_j\|^2)^{1/2}$. For a matrix $Z \in \mathbb{R}^{m\times n}$, $|Z|_1 = \sum_{i=1}^{m}\sum_{j=1}^{n}|Z_{ij}|$.

2.2. Hanson-Wright inequality in Hilbert spaces

Throughout Section 2.2, we assume that $\mathbb{H}$ is a real separable Hilbert space and $\Gamma \in \mathcal{B}(\mathbb{H})$ is a positive definite trace class operator on $\mathbb{H}$. First, we present a Hanson-Wright inequality with zero diagonal in Proposition 2.5.

***Proposition* 2.5** (Hanson-Wright inequality for quadratic forms of sub-gaussian random variables in Hilbert spaces: zero diagonal)**.** Let $X_i, i = 1, \ldots, n$, be a sequence of independent centered sub-gaussian(Γ) random variables in $\mathbb{H}$ and $L_i = \|X_i\|_{\psi_2,\Gamma}$. Let $A = (a_{ij})_{i,j=1}^n$ be an $n \times n$ matrix and $S = \sum_{1\leqslant i\neq j\leqslant n} a_{ij}\langle X_i, X_j\rangle$. Then there exists a universal constant $C > 0$ such that for any $t > 0$,

$$\mathbb{P}\left(S \geqslant t\right) \leqslant \exp\left[-C\min\left(\frac{t^2}{L^4\|\Gamma\|^2_{\mathrm{HS}}\|A\|^2_{\mathrm{HS}}}, \frac{t}{L^2\|\Gamma\|_{\mathrm{op}}\|A\|_{\mathrm{op}}}\right)\right], \tag{2.3}$$

where $L = \max_{1\leqslant i\leqslant n} L_i$.

***Remark* 2.1.** Proposition 2.5 is a dimension-free version of the Hanson-Wright inequality with a zero diagonal weighting matrix for *independent* sub-gaussian random variables in $\mathbb{R}$ [22]. Specifically, Theorem 1.1 (i.e., Theorem 1.1 in [22]) is a special case of Proposition 2.5 with $\mathbb{H} = \mathbb{R}$ and $\langle X_i, X_j\rangle = X_iX_j$. In this case, we may take $\Gamma = 1$ and thus $\|\Gamma\|_{\mathrm{op}} = \|\Gamma\|_{\mathrm{HS}} = 1$. Different from Theorem 1.1, Proposition 2.5 is also able to capture the (component-wise) dependency encoded in Γ for general Hilbert spaces, thus covering certain quadratic forms in a finite-dimensional sub-gaussian random vector with *dependent* components. We emphasize that, although our general proof strategy of decoupling the off-diagonal dependence is based on that of Theorem 1.1 in [22], a key step in our proof to remove the dependency in the Hilbert space valued sub-gaussian random variables is diagonalizing the operator Γ (together with the decoupling). Such diagonalization procedure allows us to perform the calculations in an isometric ℓ^2 space

of $\mathbb{H}$, where linear operators can be conveniently represented by (infinite-dimensional) matrices. This turns out to be the crux to obtain the trade-off between $\|\Gamma\|_{\mathrm{HS}}$ and $\|\Gamma\|_{\mathrm{op}}$ in the tail probability bound for the off-diagonal sum S. ■

Our next result is an upper tail inequality (i.e., one-sided Hanson-Wright inequality) with non-negative diagonal weights in Theorem 2.6 below.

Theorem 2.6 (Upper tail inequality for quadratic forms of sub-gaussian random variables in Hilbert spaces: non-negative diagonal)**.** *Let $X_i, i = 1, \ldots, n$, be a sequence of independent centered sub-gaussian(Γ) random variables in $\mathbb{H}$ and $L_i = \|X_i\|_{\psi_2,\Gamma}$. Let $A = (a_{ij})_{i,j=1}^n$ be an $n \times n$ matrix such that $a_{ii} \geqslant 0$, and $Q = \sum_{i,j=1}^n a_{ij}\langle X_i, X_j\rangle$. Then there exists a universal constant $C > 0$ such that for any $t > 0$,*

$$\mathbb{P}\left(Q \geqslant \sum_{i=1}^n a_{ii}L_i^2\|\Gamma\|_{\mathrm{tr}} + t\right) \leqslant 2\exp\left[-C\min\left(\frac{t^2}{L^4\|\Gamma\|_{HS}^2\|A\|_{HS}^2}, \frac{t}{L^2\|\Gamma\|_{op}\|A\|_{op}}\right)\right], \tag{2.4}$$

where $L = \max_{1\leqslant i\leqslant n} L_i$.

Both Proposition 2.5 and Theorem 2.6 allow $X_i, i = 1, \ldots, n$, to have different covariance operators Σ_i, provided that $\Sigma_i \preceq 4\|X_i\|_{\psi_2,\Gamma}^2\Gamma$ (cf. Lemma 2.4).

***Remark* 2.2** (Connections to the existing upper tail inequality in finite-dimensional Euclidean spaces)**.** First, we mention that the upper tail probability bound (2.4) (also cf. Lemma A.2) is sharper than the one-dimensional Bernstein's inequality for the non-negatively weighted diagonal sum of squared norm of independent sub-gaussian random variables in $\mathbb{H}$. Indeed, if we simply apply Bernstein's inequality (cf. Theorem 2.8.1 in [28]) for the real-valued sub-exponential random variables $\|X_i\|^2$ (cf. Lemma A.4), then the diagonal sum in Q has the following probability bound: for any $t > 0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{i=1}^n a_{ii}(\|X_i\|^2 - \mathbb{E}\,\|X_i\|^2)\right| \geqslant t\right)\\ &\qquad\leqslant 2\exp\left[-C\min\left(\frac{t^2}{L^4\|\Gamma\|_{\mathrm{tr}}^2\sum_{i=1}^n a_{ii}^2}, \frac{t}{L^2\|\Gamma\|_{\mathrm{tr}}\max_{1\leqslant i\leqslant n}|a_{ii}|}\right)\right].\end{aligned} \tag{2.5}$$

Note that the right-hand side of (2.5) is controlled by one parameter $\|\Gamma\|_{\mathrm{tr}}$, which is strictly less sharp than (2.4) since $\|\Gamma\|_{\mathrm{op}} \leqslant \|\Gamma\|_{\mathrm{tr}}$ and $\|\Gamma\|_{\mathrm{HS}}^2 \leqslant \|\Gamma\|_{\mathrm{op}}\|\Gamma\|_{\mathrm{tr}} \leqslant \|\Gamma\|_{\mathrm{tr}}^2$. For instance, if $X_i \in \mathbb{R}^p$, then Γ is often the $p \times p$ covariance matrix of X_i. In the special case for $\Gamma = I_p$, then $\|\Gamma\|_{\mathrm{op}} = 1$, $\|\Gamma\|_{\mathrm{HS}} = p^{1/2}$, and $\|\Gamma\|_{\mathrm{tr}} = p$. Therefore, direct application of the diagonal sum bound (2.5) does not yield the probability bound in Proposition 2.5. In particular, for the generalized K-means clustering problem, this implies that a much more restrictive lower bound condition on the signal-to-noise ratio is required for exact recovery of the true clustering structure for high-dimensional data (more details can be found in the discussion after Theorem 3.3).

Second, for non-negative diagonal weights, Theorem 2.6 is an infinite-dimensional (and thus dimension-free) generalization of the tail inequality for quadratic forms a sub-gaussian random vector with dependent components in $\mathbb{R}^p$ [13]. In particular, if $X = (X_1, \ldots, X_p)$ is a centered sub-gaussian random vector in $\mathbb{R}^p$ (i.e., there exists a $\sigma \geqslant 0$ such that $\mathbb{E}[e^{z^T X}] \leqslant e^{\|z\|_2^2 \sigma^2/2}$ for all $z \in \mathbb{R}^p$), then Theorem 2.1 in [13] states that: for any positive semidefinite matrix Σ and $t > 0$,

$$\mathbb{P}\left(X^T \Gamma X \geqslant \sigma^2(\|\Gamma\|_{\mathrm{tr}} + 2\|\Gamma\|_{\mathrm{HS}}\sqrt{t} + 2\|\Gamma\|_{\mathrm{op}} t)\right) \leq e^{-t}.$$

The last inequality is a special case (up to a universal constant) of (2.4) with $n = 1$, $A = 1$, $\mathbb{H} = \mathbb{R}^p$, $\Gamma^{-1/2} X \sim$ sub-gaussian$(\sigma^2 I_p)$, and $L^2 = \sigma^2$. In addition, we note that the positive semidefinite condition is not needed in our Theorem 2.6. Instead, only a weaker condition on the non-negativity of the diagonal entries in the weighting matrix is required. ■

There are two limitations of Theorem 2.6. First, Q is typically not centered at $\sum_{i=1}^n a_{ii} L_i^2 \|\Gamma\|_{\mathrm{tr}}$. For the generalized K-means application in Section 3, this means that consistency of solutions of the SDP relaxation (3.3) cannot be attained unless $\sum_{i=1}^n a_{ii} L_i^2 \|\Gamma\|_{\mathrm{tr}}$ tends to $\mathbb{E}[Q]$. Second, the non-negativity condition on the diagonal weights $a_{ii} \geqslant 0$ in Theorem 2.6 is not entirely innocuous for obtaining a concentration inequality for Q (i.e., two-sided Hanson-Wright inequality). Without imposing additional assumptions, we cannot expect a lower tail bound for sub-gaussian random variables even in $\mathbb{R}^n$ [1]. To simultaneously fix these two issues and obtain a concentration inequality for $Q - \mathbb{E}[Q]$, we make the following Bernstein-type assumption on the squared norm, in addition to the assumption that $X_1, \ldots, X_n$ are independent sub-gaussian(Γ) with mean zero.

***Assumption* 2.7** (Bernstein condition on the squared norm)**.** There exists a universal constant $C > 0$ such that

$$\mathbb{E}\left|\|X_i\|^2 - \mathbb{E}\|X_i\|^2\right|^k \leqslant C k! L_i^{k-2} \|\Gamma\|_{\mathrm{op}}^{k-2} \|\Sigma_i\|_{\mathrm{HS}}^2 \quad \forall k = 3, 4, \ldots, \tag{2.6}$$

where $\Sigma_i = \mathbb{E}[X_i \otimes X_i]$ is the covariance operator of $X_i, i = 1, \ldots, n$.

***Remark* 2.3** (Comments on Assumption 2.7)**.** Since $\|\Sigma_i\|_{\mathrm{tr}} = \mathbb{E}\|X_i\|^2$, Assumption 2.7 is a mild condition on the sub-exponential tail behavior of $\|X_i\|^2 - \|\Sigma_i\|_{\mathrm{tr}}$. For $\mathbb{H} = \mathbb{R}$, (2.6) is an automatic consequence of the sub-gaussianality (2.1). For $\mathbb{H} = \mathbb{R}^p$, if $X = \Sigma^{1/2} Z$, where $Z = (Z_1, \ldots, Z_p)^T$ has independent components Z_j with bounded sub-gaussian norms, then

$$\mathbb{E}[\|X\|^2 - \mathbb{E}\|X\|^2]^2 = \mathbb{E}[Z^T \Sigma Z - \mathrm{tr}(\Sigma)]^2 \lesssim \|\Sigma\|_{\mathrm{HS}}^2.$$

Such linear transformation of an independent random vector in $\mathbb{R}^p$ with sub-gaussian components is a popular statistical model for the K-means clustering [10, 21]. For the general Hilbert space $\mathbb{H}$, it is easy to verify that Gaussian random variable $Z \sim N(0, \Gamma)$ in $\mathbb{H}$ satisfies (2.6). Comparing with the "centering" term $\sum_{i=1}^n a_{ii} L_i^2 \|\Gamma\|_{\mathrm{tr}}$ in (2.4), we shall see that the correct centering terms $\mathbb{E}\|X_i\|^2$ in (2.6) together with the parameters

$(L_i\|\Gamma\|_{\mathrm{op}}, \|\Sigma_i\|_{\mathrm{HS}})$ are crucial to yield a concentration inequality for $Q-\mathbb{E}[Q]$. By Lemma 2.4, we know that $4L_i^2\|\Gamma\|_{\mathrm{tr}} \geqslant \|\Sigma_i\|_{\mathrm{tr}}$ for any $X_i \sim$ sub-gaussian(Γ). In fact, even in $\mathbb{R}$, it is easy to construct a random variable $X \sim$ sub-gaussian(γ^2) such that $\gamma^2 \gg \sigma^2$ where $\sigma^2 = \mathrm{Var}(X)$ (cf. Example 4.1 and 4.2 in [6]). In particular, here we give a counterexample in $\mathbb{R}$ (so that $L_i = 1$). Let Y_n follow a mixture of Gaussian distributions $F_n = (1-\epsilon_n)N(0,1) + \epsilon_n N(0, a_n^2)$, where $a_n > 1$ and $\epsilon_n = a_n^{-4}$. Then we have $\sigma_n^2 := \mathrm{Var}(Y_n) = 1-a_n^{-4}+a_n^{-2}$ and $Y_n \sim$ sub-gaussian(γ_n^2), where $\gamma_n^2 = Ca_n^2$ for some sufficiently large constant $C > 0$. Thus if $a_n \to \infty$ as $n \to \infty$, then $\sigma_n^2 \asymp 1$ and

$$\mathbb{E}\,|Y_n^2 - \mathbb{E}\,Y_n^2|^k \lesssim a_n^{2k-4}\,\mathbb{E}\,|Z|^{2k} = a_n^{2k-4}(2k-1)!! \leqslant 4k!(2a_n^2)^{k-2} \lesssim k!(\gamma_n^2)^{k-2}(\sigma_n^2)^2,$$

where $Z \sim N(0,1)$. Hence $(Y_n)_{n=1,2,\ldots}$ is a sub-gaussian random variable satisfying Assumption 2.7 and $\sigma_n^2 \ll \gamma_n^2$, provided that $a_n \to \infty$ as $n \to \infty$. ■

Now we are ready to state the Hanson-Wright inequality for the general case.

Theorem 2.8 (Hanson-Wright inequality for quadratic forms of sub-gaussian random variables in Hilbert spaces: general version)**.** *Let $X_i, i = 1,\ldots,n$, be a sequence of independent centered sub-gaussian(Γ) random variables in $\mathbb{H}$ and $L_i = \|X_i\|_{\psi_2,\Gamma}$. Let $A = (a_{ij})_{i,j=1}^n$ be an $n \times n$ matrix and $Q = \sum_{i,j=1}^n a_{ij}\langle X_i, X_j\rangle$. If in addition Assumption 2.7 holds, then there exists a universal constant $C > 0$ such that for any $t > 0$,*

$$\mathbb{P}\left(|Q - \mathbb{E}[Q]| \geqslant t\right) \leqslant 2\exp\left[-C\min\left(\frac{t^2}{L^4\|\Gamma\|_{HS}^2\|A\|_{HS}^2}, \frac{t}{L^2\|\Gamma\|_{op}\|A\|_{op}}\right)\right], \tag{2.7}$$

where $L = \max_{1\leqslant i\leqslant n} L_i$.

[29] and [1] derive Hanson-Wright inequalities under the convex concentration property of a finite-dimensional random vector, which is difficult to verify in general. In contrast, our Theorem 2.8 holds under more transparent conditions (i.e., the sub-gaussian and Bernstein-type assumptions). Note that Theorem 2.8 can be seen as a unified generalization of the finite-dimensional Hanson-Wright inequality to Hilbert spaces for both *independent* sub-gaussian random variables in $\mathbb{R}$ [22] and a sub-gaussian random vector with *dependent* components in $\mathbb{R}^p$ [13].

3. *K*-means clustering in Hilbert spaces and its semidefinite relaxation

In this section, we apply the Hanson-Wright inequality in Section 2.2 (i.e., Theorem 2.8) to the clustering problem of n data points into K clusters such that $K \leqslant n$. Let $X_1,\ldots,X_n$ be a sequence of independent random variables taking values in a measurable space $(\mathbb{X}, \mathcal{X})$ on $(\Omega, \mathscr{B}, \mathbb{P})$. Suppose that there exists a clustering structure $G_1^*,\ldots,G_K^*$ (i.e., a partition on $[n] := \{1,\ldots,n\}$ satisfying $\cup_{k=1}^K G_k^* = \{1,\ldots,n\}$ and $G_k^* \cap G_m^* = \emptyset$ if

$1 \leqslant k \neq m \leqslant K$) on the n data points with $X_i \sim P_k$ for $i \in G_k^*$, where $P_1, \ldots, P_K$ are distinct distributions on $(\mathbb{X}, \mathcal{X})$. We emphasize that $\mathbb{X}$ does not need to be a Euclidean space. Our goal is to develop a statistically correct and computationally tractable algorithm for recovering the true clustering structure based on the similarity of the observations $X_1, \ldots, X_n$.

3.1. *K*-means in Hilbert spaces: 0-1 integer program formulation

Perhaps one of the most widely used clustering methods is the Euclidean distance-based K-means clustering, due to the existence of computationally efficient heuristic algorithms (such as Lloyd's algorithm [17]). This is a particularly attractive feature for large datasets. Given a sequence of observations $X_1, \ldots, X_n \in \mathbb{R}^p$ (i.e., $\mathbb{X} = \mathbb{R}^p$), the (classical) K-means clustering method minimizes the total intra-cluster squared Euclidean distances

$$\min_{G_1,\ldots,G_K} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j \in G_k} \|X_i - X_j\|^2$$

over all possible partitions on $[n]$, where $|G_k|$ is the cardinality of G_k. Dropping the sum of squared norms $\sum_{i=1}^n \|X_i\|^2$, we see that the K-means clustering is equivalent to the maximization of the total intra-cluster correlations

$$\max_{G_1,\ldots,G_K} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j \in G_k} X_i^T X_j.$$

Here, $X_i^T X_j$ can be viewed as a similarity measure specified by the Euclidean space inner product $a_{ij} = \langle X_i, X_j \rangle_{\mathbb{R}^p}$. In general, if space $\mathbb{X}$ is a Hilbert space $\mathbb{H}$, then it is natural to generalize this procedure by replacing $\langle \cdot, \cdot \rangle_{\mathbb{R}^p}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ associated with $\mathbb{H}$, yielding $a_{ij} = \langle X_i, X_j \rangle_{\mathbb{H}}$. Henceforth, we will refer to such a K-means that uses the inner product in a Hilbert space as a generalized K-means.

***Example* 3.1** (Functional data clustering)**.** In many applications, data to be clustered are recorded as curves, surfaces or other things varying over a continuum, such as a time interval and a space span. The random variable underlying data is naturally modelled as a stochastic process $X = \{X(t) : t \in \mathcal{T}\}$ in Hilbert space $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$, where the sequence of observations $X_1, \ldots, X_n \in \mathbb{H}$ is an i.i.d. sample of random variables drawn from the same distribution as X. In clustering problems, the law of X is often assumed to be a mixture distribution over $\mathbb{H}$, with each mixture component as a cluster. When $\mathcal{T} = [0, 1]$ is the unit interval, we can choose $\mathbb{H}$ as the L^2 function space $\mathbb{L}^2[0, 1] = \{f : [0, 1] \to \mathbb{R} : \|f\|_{\mathbb{L}^2}^2 = \int_0^1 |f(t)|^2 \, dt < \infty\}$ with $\mathbb{L}^2$-inner product $\langle f, g \rangle_{\mathbb{L}^2} = \int_0^1 f(t)g(t) \, dt$ for $f, g \in \mathbb{L}^2[0, 1]$. Suppose we have prior information that the observations $\{X_i\}$ are smooth functions, then we can choose a stronger norm to capture the similarity in the (higher-order) derivatives. For example, in [14, 25] and [8], $\mathbb{H}$ are recommended as the Sobolev space with some order $k \in \{1, 2\}$ as $\mathbb{S}^k[0, 1] = \{f : [0, 1] \to \mathbb{R} : \|f^{(k)}\|_{\mathbb{L}^2}^2 = \int_0^1 |f^{(k)}(t)|^2 \, dt < \infty\}$ equipped

with inner product $\langle f, g\rangle_{\mathbb{S}^k} = \sum_{j=0}^{k}\langle f^{(j)}, g^{(j)}\rangle_{\mathbb{L}^2}$, where $f^{(k)}$ denotes the kth derivative of a function $f \in \mathbb{S}^k[0,1]$. As we will see in Section 3.4, a higher smoothness order k in the generalized K-means generally leads to larger separations among cluster centers (between cluster variation) without significantly increasing fluctuations within clusters (within cluster variation), thereby increasing the clustering signal-to-noise ratio (see Theorem 3.3 for a precise definition). ■

***Example* 3.2** (Kernel clustering)**.** In pattern recognition and natural language processing, it is often crucial to capture the non-linear similarity for non-Euclidean data (such as images and words). A widely used approach is the kernel method [23], where the similarity a_{ij} between X_i and X_j is characterized by a nonlinear positive semi-definite kernel function $\rho : \mathbb{X}\times\mathbb{X} \to \mathbb{R}$ through $a_{ij} = \rho(X_i, X_j)$. Commonly used kernel functions include polynomial kernels $\rho(x,y) = (\langle x, y\rangle + c)^r$ for some positive integer order r and radial basis function (RBF) kernel $\rho(x,y) = \exp\{-\|x-y\|^2/(2h^2)\}$ for some bandwidth parameter $h > 0$, where $x, y \in \mathbb{R}^p$ are the Euclidean embeddings of the original observations (image pixel level vectorizations or word embeddings). According to the celebrated Mercer's theorem, kernel clustering can also be viewed as K-means in a high-dimensional feature space: there always exists a Hilbert space (feature space) $\mathbb{H}$ equipped with inner product $\langle\cdot,\cdot\rangle_{\mathbb{H}}$ and a feature map $\phi : \mathbb{X} \to \mathbb{H}$, such that

$$\rho(x,y) = \langle\phi(x), \phi(y)\rangle_{\mathbb{H}}, \quad \forall x, y \in \mathbb{X}.$$

More details about a construction of the feature map can be found in Section A.1. From this identity, kernel K-means that uses a nonlinear similarity measure $a_{ij} = \rho(X_i, X_j)$ can be cast into the framework of K-means in Hilbert spaces by identifying X_i as $\phi(X_i)$. On the other hand, explicit representations for the feature map ϕ and the Hilbert space $\mathbb{H}$ are not necessary in order to implement the kernel K-means, which is one of the main practical attractiveness of the method. By choosing a proper kernel ρ, we may capture the non-linear similarity in non-Euclidean spaces through implicitly mapping the original data space $\mathbb{X}$ into a "high-dimensional" feature space, in which linear boundaries can be drawn to separate the data points. For example, the polynomial kernel maps into the space spanned by the products of all monomials up to degree r. In particular, clusters with centers (expectations under P_j's) that are overlapped in the original Euclidean space may have separated centers (expectations under $\phi_\#(P_j)$'s, where $\phi_\#(\mu)$ denotes the pushforward of measure μ defined through $(\phi_\#(\mu))(B) = \mu(\phi^{-1}(B))$ for every measurable subset $B \subset \mathbb{H}$) in the feature space. ■

For a general inner product $\langle\cdot,\cdot\rangle_{\mathbb{H}}$, quadratic sample complexity is needed for the generalized K-means to compute the similarity matrix A [9]. Observe that, for every partition $G_1, \ldots, G_K$, there is a one-to-one $n \times K$ *assignment matrix* $H = (h_{ik}) \in \{0,1\}^{n\times K}$ such that $h_{ij} = 1$ if $i \in G_k$ and $h_{ij} = 0$ if $i \notin G_k$. Thus the K-means clustering problem can be written as a 0-1 integer program:

$$\max\left\{\langle A, HBH^T\rangle : H \in \{0,1\}^{n\times K}, H\mathbf{1}_K = \mathbf{1}_n\right\}, \tag{3.1}$$

where $\mathbf{1}_n$ denotes the $n\times 1$ vector of all ones, $a_{ij} = \langle X_i, X_j\rangle_{\mathbb{H}}$, and $B = \mathrm{diag}(|G_1|^{-1}, \ldots, |G_K|^{-1})$.

The generalized K-means clustering problem (3.1) is typically computationally intractable, namely polynomial-time algorithms with exact solutions only exist in certain cases [24]. For instances, the (classical) K-means clustering is a worst-case NP-hard integer programming problem with a non-linear objective function [18]. Exact and partial recovery properties of various SDP relaxations for the K-means [18, 16, 7, 21, 10] are studied in literature. However, it remains a challenging task to provide statistical guarantees for the generalized K-means clustering to capture the non-linear features of non-Euclidean data taking values in a general Hilbert space.

3.2. SDP relaxation for K-means in Hilbert spaces

We consider the SDP relaxations for the generalized K-means clustering. Note that every partition $G_1, \ldots, G_K$ of $[n]$ can be represented by a partition function $\sigma : [n] \to [K]$ via $G_k = \sigma^{-1}(k), k = 1, \ldots, n$. If we change the variable $Z = HBH^T$ in the 0-1 integer program formulation (3.1) of the generalized K-means, then Z satisfies the following properties:

$$Z^T = Z, \quad Z \succeq 0, \quad \mathrm{tr}(Z) = \sum_{k=1}^{K} |G_k|\, b_{kk}, \quad (Z\mathbf{1}_n)_i = \sum_{k=1}^{K} |G_k|\, b_{\sigma(i)k},\ i = 1, \ldots, n. \tag{3.2}$$

For the generalized K-means $B = \mathrm{diag}(|G_1|^{-1}, \ldots, |G_K|^{-1})$, the last constraint in (3.2) reduces to $Z\mathbf{1}_n = \mathbf{1}_n$, which does not depend on the partition function σ. Thus we can relax the generalized K-means clustering to the SDP problem:

$$\hat{Z} = \mathrm{argmax}\,\{\langle A, Z\rangle : Z \in \mathscr{C}\} \ \text{ with } \mathscr{C} = \{Z^T = Z, Z \succeq 0, \mathrm{tr}(Z) = K, Z\mathbf{1}_n = \mathbf{1}_n, Z \geqslant 0\}, \tag{3.3}$$

where $Z \succeq 0$ means that Z is positive semidefinite and $Z \geqslant 0$ means that all entries of Z are non-negative. We shall use $\hat{Z}$ to estimate the true "membership matrix" Z^*, where

$$Z^*_{ij} = \begin{cases} 1/n_k & \text{if } i, j \in G^*_k \\ 0 & \text{otherwise} \end{cases}, \tag{3.4}$$

where $n_k = |G^*_k|$. Note that $Z^* \in \mathscr{C}$ is a projection matrix such that $Z^*Z^* = Z^*$. If $X_1, \ldots, X_n \in \mathbb{R}^p$ (i.e., $\mathbb{X} = \mathbb{R}^p$) and $a_{ij} = X_i^T X_j$ is the Euclidean space inner product, then (3.3) is the SDP proposed in [18]. Observe that the SDP relaxation (3.3) does not require the knowledge of the cluster sizes other than the number of clusters K. Thus it can handle the general case for unequal cluster sizes.

3.3. Rate of convergence of SDP for K-means in Hilbert spaces

Now we are in the position to state the rate of convergence for the SDP relaxation (3.3) for the generalized K-means clustering. For simplicity, we assume that the trace norms

of the covariance operators for the K-cluster distributions $P_1, \ldots, P_K$ are equal. If the trace norms are not all equal, then a similar de-biased SDP in [5] can be considered. Denote the minimum cluster size as $\underline{n} = \min_{1 \leqslant k \leqslant K} n_k$.

Theorem 3.3 (Exponential rate of convergence of SDP for generalized K-means)**.** *Let $X_1, \ldots, X_n$ be a sample of independent random variables in Hilbert space $\mathbb{H}$ such that $X_i \sim P_k$ for $i \in G_k^*$. Let $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and $\| \cdot \|_{\mathbb{H}}$ be the associated inner product and Hilbert norm with $\mathbb{H}$, and $\mu_k = \mathbb{E}\, X_i$, $\Sigma_k = \mathbb{E}[(X_i - \mu_k) \otimes (X_i - \mu_k)]$ be the covariance operator of $X_i, i \in G_k^*$. Suppose that $\mathbb{H}$ is separable, and $X_i \sim$ sub-gaussian(Σ_k) for $i \in G_k^*$ such that $\|X_i\|_{\psi_2, \Sigma_k} \leqslant L$ and Assumption 2.7 holds with $\Gamma_i = \Sigma_i$ therein being equal to Σ_k. In addition, assume $(\Sigma_k)_{k=1}^K$ to be positive definite trace class, and $\|\Sigma_1\|_{\mathrm{tr}} = \cdots = \|\Sigma_K\|_{\mathrm{tr}}$. Define*

$$\mathsf{SNR}^2 = \frac{\Delta^2}{L^2 \|\Sigma\|_{op}} \wedge \frac{\underline{n} \Delta^4}{L^4 \|\Sigma\|_{HS}^2} \quad \text{with } \Delta = \min_{1 \leqslant i \neq j \leqslant K} \|\mu_i - \mu_j\|_{\mathbb{H}}$$

as the squared signal-to-noise ratio, and suppose $\Sigma \succeq \Sigma_k$ for all $k = 1, \ldots, K$. Then there exist universal constants $c_0, c_0', c, C_1, C_2 > 0$ such that as long as $\mathsf{SNR}^2 \geqslant c_0\, n/\underline{n}$ and $\underline{n}^2 K \geqslant c_0' n$, it holds that

$$|\hat{Z} - Z^*|_1 \leqslant C_1 \exp(-C_2 \mathsf{SNR}^2) \tag{3.5}$$

with probability at least $1 - c/n^2$.

This theorem characterizes the hardness of clustering through the squared signal-to-noise ratio SNR^2 that depends on the ratio of squared between-cluster separation rate Δ^2 to within-clustering variation $L^2\|\Sigma\|_{\mathrm{op}}$ or $L^2\|\Sigma\|_{\mathrm{HS}}$. We postpone its proof to Section 4.2. It turns out that both terms in SNR^2 are necessary depending on different regimes of parameters Δ and Σ. For the optimality of the exponent SNR^2 in the convergence rate for Euclidean space clustering, namely $\mathbb{H} = \mathbb{R}^p$, we refer to Section 3.3 of [10] for a detailed discussion. In particular, if we instead use the weaker version of the concentration inequality (2.5), then an extra p factor will appear in the denominator of each term in SNR^2, which is clearly suboptimal.

Our proof is based on the inequality $\langle A, Z^* \rangle \leqslant \langle A, \hat{Z} \rangle$, which is true due to the optimality of $\hat{Z}$ and the feasibility of Z^*. In particular, in the analysis of $\langle A, \hat{Z} - Z^* \rangle$ by decomposing the similarity matrix A as a sum of its expectation and random fluctuations, one remainder term caused by the random fluctuations involves a quadratic form over Hilbert space $\mathbb{H}$ as the Q in Theorem 2.8. In particular, we prove a uniform version of the Hanson-Wright inequality that leads to the exponential convergence rate (3.5) in Theorem 3.3 by combining our Theorem 2.8 with a careful union bound technique developed in [7] that utilizes the geometric structure of A and improves upon a naive union bound argument via covering.

Theorem 3.3 provides a partial recovery bound for clustering. Next, we show that exact recovery can be achieved by properly rounding the SDP solution $\hat{Z}$. More specifically, we consider the rounding algorithm that proceeds as follows: 1. let $j_1 = 1$ and $\hat{G}_1$ be the set of all indices i such that $\hat{Z}_{j_1 i} \geqslant \frac{1}{2} \hat{Z}_{j_1 j_1}$; 2. let j_2 be the smallest index in

$[n] \setminus \hat{G}_1$ and $\hat{G}_2$ be the set of all indices i such that $\hat{Z}_{j_2 i} \geqslant \frac{1}{2}\hat{Z}_{j_2 j_2}$; ..., end until the remainder index set $[n] \setminus \bigcup_{k=1}^{\hat{K}} \hat{G}_k$ becomes empty for some $\hat{K} \geqslant 1$. Thanks to Theorem 3.3, exact recovery of the true clustering structure is an immediate consequence when $\mathsf{SNR}^2 \gtrsim \max\{n/\underline{n}, \log n\}$.

***Corollary* 3.4** (Exact recovery of SDP for generalized K-means)**.** In the setting of Theorem 3.3, suppose $\mathsf{SNR}^2 \geqslant c_1 \max\{n/\underline{n}, \log n\}$ and $\underline{n}^2 K \geqslant c_2 n$ for some universal constants $c_1, c_2 > 0$, then

$$\mathbb{P}(\hat{K} = K \text{ and } \hat{G}_k = G_k^*, \ \forall k = 1, \ldots, K) \geqslant 1 - Cn^{-2}$$

for some universal constant $C > 0$.

3.4. Implications in functional data clustering

In this subsection, we discuss the consequence of applying Theorem 3.3 to Example 3.1. For simplicity, we assume that for each $k = 1, \ldots, K$, the sampling measure P_k is a Gaussian process (GP) over Hilbert space $\mathbb{L}^2[0,1]$ with inner product $\langle \cdot, \cdot \rangle_{\mathbb{L}^2}$. In particular, we use Theorem 3.3 to study and compare the uses of different inner products (such as Sobolev inner products with different orders) in constructing the similarity matrix A in the generalized K-means for functional data clustering.

Recall the definition of a Gaussian random variable in a Hilbert space in Definition 2.3. When the Hilbert space is a function space, the law $N(\mu, \Sigma)$ of a GP is completely determined by its mean function $\mu : [0,1] \to \mathbb{R} \in \mathbb{L}^2[0,1]$ and covariance function $\Sigma : [0,1]^2 \to \mathbb{L}^2[0,1]$, where $\mu(t) = \mathbb{E}[X(t)]$ and $\Sigma(t,t') = \mathbb{E}[(X(t) - \mu(t))(X(t') - \mu(t'))]$ for any GP realization $X = \{X(t) : t \in [0,1]\}$. The covariance function Σ can be identified with the covariance operator through

$$\Sigma f(t) = \int_0^1 \Sigma(t,t')\, f(t')\, dt', \quad \text{for all } f \in \mathbb{L}^2[0,1] \text{ and } t \in [0,1].$$

Suppose now we have another Hilbert space $\mathbb{H}' \subset \mathbb{H}$, such as the Sobolev space $\mathbb{S}^k[0,1]$ for some $k \geqslant 1$, such that the second moment of $\|X - \mu\|_{\mathbb{H}'}$ is still bounded relative to the stronger norm $\|\cdot\|_{\mathbb{H}'}$ associated with $\mathbb{H}'$, that is $\mathbb{E}[\|X - \mu\|_{\mathbb{H}'}^2] < \infty$. This implies $X - \mu \in \mathbb{H}'$ almost surely, and $\langle h, X - \mu \rangle_{\mathbb{H}'}$ is Gaussian for all $h \in \mathbb{H}'$. As a consequence, $X - \mu$ remains a Gaussian random variable in the new Hilbert space $\mathbb{H}'$ [26], as long as $\mathbb{E}[\|X - \mu\|_{\mathbb{H}'}^2] < \infty$. Here μ may or may not belong to $\mathbb{H}'$ depending on whether $\|\mu\|_{\mathbb{H}'}$ is finite or infinite. We use Σ' to denote its covariance operator as a Gaussian random variable in $\mathbb{H}'$. In cases where Σ has rapid eigenvalue decay (polynomial or exponential), the operator and the Hilbert-Schmidt norms of Σ and Σ' will be dominated by their respective top eigenvalues, henceforth comparable in magnitudes.

Returning to the functional data clustering, we assume $X_i \sim N(\mu_k, \Sigma_k)$ for $i \in G_k^*$ as Gaussian random variables in $\mathbb{H}$. Consider two choices $a_{ij} = \langle X_i, X_j \rangle_{\mathbb{H}}$ and $a'_{ij} =$

$\langle X_i, X_j\rangle_{\mathbb{H}'}$ for constructing the similarity matrix A in the SDP for the generalized K-means clustering. From our previous discussion, we know that $X_i - \mu_k$ remains Gaussian in $\mathbb{H}'$ as long as $\mathbb{E}[\|X_i - \mu_k\|_{\mathbb{H}'}^2] < \infty$. We use Σ_k' to denote the covariance operator of $X_i - \mu_k$ as a Gaussian random variable in $\mathbb{H}'$. We can then apply Theorem 3.3 with Hilbert space $\mathbb{H}$ and $\mathbb{H}'$ to obtain the signal-to-noise ratios under these two choices,

$$\mathsf{SNR}^2 = \frac{\Delta^2}{L^2\|\Sigma\|_{\mathrm{op}}} \wedge \frac{n\Delta^4}{L^4\|\Sigma\|_{\mathrm{HS}}^2} \quad \text{with } \Delta = \min_{1\leqslant i\neq j\leqslant K} \|\mu_i - \mu_j\|_{\mathbb{H}}, \quad \text{and}$$
$$(\mathsf{SNR}')^2 = \frac{(\Delta')^2}{L^2\|\Sigma'\|_{\mathrm{op}}} \wedge \frac{n(\Delta')^4}{L^4\|\Sigma'\|_{\mathrm{HS}}^2} \quad \text{with } \Delta' = \min_{1\leqslant i\neq j\leqslant K} \|\mu_i - \mu_j\|_{\mathbb{H}'},$$

where $\Sigma \succeq \Sigma_k$ and $\Sigma' \succeq \Sigma_k'$ for each k. The denominators of SNR^2 and $(\mathsf{SNR}')^2$ are comparable when Σ and Σ' have rapid eigenvalue decay, while the signal strength Δ' can be much larger than Δ, making the overall $(\mathsf{SNR}')^2$ larger as well. For functional data with $\mathbb{H} = \mathbb{L}^2[0,1]$, faster eigenvalue decay in the covariance operator corresponds to a higher smoothness order of the sample path. For example, if $\gamma_1 \geqslant \gamma_2 \geqslant \dots$ are ordered eigenvalues of Σ with $\gamma_j \approx j^{-2\beta-1}$ for $j = 1, 2, \dots$ and some $\beta > 0$, then sample paths from $N(0,\Sigma)$ are at least β times differentiable [19] almost surely. If we choose $\mathbb{H}'$ to be $\mathbb{S}^k[0,1]$ for any $0 \leqslant k \leqslant \lfloor\beta\rfloor$, where $\lfloor\beta\rfloor$ denotes the largest integer smaller than β, then $\mathbb{E}[\|X_i - \mu_k\|_{\mathbb{H}'}^2] < \infty$. On the other hand side, Δ' can be much larger than Δ when the difference $\{\mu_i - \mu_j : 1 \leqslant i \neq j \leqslant K\}$ has smoothness order (characterized via the decay rate of coefficients with respect to eigenfunctions $\{e_i\}$ of Σ) lower than k. In such scenarios, using the inner product induced by a stronger norm in constructing the similarity matrix A may increase the signal-to-noise ratio and reduce the SDP error $|\hat{Z} - Z^*|_1$.

4. Proof of main results

4.1. Proof of main results in Section 2.2

In this subsection, we prove Proposition 2.5, Theorem 2.6, and 2.8.

Proof of Proposition 2.5. By Markov's inequality, we have for any $\lambda > 0$ and $t > 0$,

$$\mathbb{P}(S \geqslant t) \leqslant e^{-\lambda t}\,\mathbb{E}[e^{\lambda S}].$$

Step 1: decoupling. Let $\delta_1, \dots, \delta_n \in \{0,1\}$ be i.i.d. symmetric Bernoulli random variables (i.e., $\mathbb{P}(\delta_i = 0) = \mathbb{P}(\delta_i = 1) = 1/2$) that are independent of $X_1, \dots, X_n$. Since

$$\mathbb{E}[\delta_i(1-\delta_j)] = \begin{cases} 0 & \text{if } i = j \\ 1/4 & \text{if } i \neq j \end{cases},$$

we have $S = 4\,\mathbb{E}_\delta[S_\delta]$, where $S_\delta = \sum_{i,j=1}^n \delta_i(1-\delta_j)a_{ij}\langle X_i, X_j\rangle$ and $\mathbb{E}_\delta[\cdot]$ is the expectation taken with respect to the random variables δ_i. Below, $\mathbb{E}_X[\cdot]$ is similarly defined. By Jensen's inequality, we get

$$\mathbb{E}[e^{\lambda S}] \leqslant E_{X,\delta}[e^{4\lambda S_\delta}].$$

Let $\Lambda_\delta = \{i \in [n] : \delta_i = 1\}$. Then we can write

$$S_\delta = \sum_{i\in\Lambda_\delta}\sum_{j\in\Lambda_\delta^c} a_{ij}\langle X_i, X_j\rangle = \sum_{j\in\Lambda_\delta^c}\langle \sum_{i\in\Lambda_\delta} a_{ij}X_i, X_j\rangle.$$

Taking the expectation with respect to $(X_j)_{j\in\Lambda_\delta^c}$ (i.e., conditioning on $(\delta_i)_{i=1,\dots,n}$ and $(X_i)_{i\in\Lambda_\delta}$), it follows from the assumption X_i are independent sub-gaussian(Γ) with mean zero that

$$\mathbb{E}_{(X_j)_{j\in\Lambda_\delta^c}}[e^{4\lambda S_\delta}] \leqslant e^{8\lambda^2\sigma_\delta^2},$$

where $\sigma_\delta^2 = \sum_{j\in\Lambda_\delta^c} L_j^2\langle\Gamma(\sum_{i\in\Lambda_\delta} a_{ij}X_i), (\sum_{i\in\Lambda_\delta} a_{ij}X_i)\rangle$. Thus we get

$$\mathbb{E}_X[e^{4\lambda S_\delta}] \leqslant \mathbb{E}_X\left[e^{8\lambda^2\sigma_\delta^2}\right].$$

Step 2: reduction to Gaussian random variables. For $j = 1,\dots,n$, let g_j be independent $N(0, 16L_j^2\Gamma)$ random variables in $\mathbb{H}$ that are independent of $X_1,\dots,X_n$ and $\delta_1,\dots,\delta_n$. Define

$$T := \sum_{j\in\Lambda_\delta^c}\langle g_j, \sum_{i\in\Lambda_\delta} a_{ij}X_i\rangle.$$

Then, by the definition of Gaussian random variables in $\mathbb{H}$, we have

$$\begin{aligned}\mathbb{E}_g[e^{\lambda T}] &= \prod_{j\in\Lambda_\delta^c}\mathbb{E}_g\left[e^{\langle g_j, \lambda\sum_{i\in\Lambda_\delta} a_{ij}X_i\rangle}\right]\\ &= \exp\left(8\lambda^2\sum_{j\in\Lambda_\delta^c} L_j^2\langle\Gamma(\sum_{i\in\Lambda_\delta} a_{ij}X_i), (\sum_{i\in\Lambda_\delta} a_{ij}X_i)\rangle\right) = \exp\left(8\lambda^2\sigma_\delta^2\right).\end{aligned}$$

So it follows that

$$\mathbb{E}_X[e^{4\lambda S_\delta}] \leqslant \mathbb{E}_{X,g}[e^{\lambda T}].$$

Since $T = \sum_{i\in\Lambda_\delta}\langle\sum_{j\in\Lambda_\delta^c} a_{ij}g_j, X_i\rangle$, we have

$$\mathbb{E}_{(X_i)_{i\in\Lambda_\delta}}[e^{\lambda T}] \leqslant \exp\left(\frac{\lambda^2}{2}\sum_{i\in\Lambda_\delta} L_i^2\langle\Gamma(\sum_{j\in\Lambda_\delta^c} a_{ij}g_j), (\sum_{j\in\Lambda_\delta^c} a_{ij}g_j)\rangle\right),$$

which implies that

$$\mathbb{E}_X[e^{4\lambda S_\delta}] \leqslant \mathbb{E}_g\left[\exp\left(\lambda^2\tau_\delta^2/2\right)\right], \tag{4.1}$$

where $\tau_\delta^2 = \sum_{i\in\Lambda_\delta} L_i^2\langle\Gamma(\sum_{j\in\Lambda_\delta^c} a_{ij}g_j), (\sum_{j\in\Lambda_\delta^c} a_{ij}g_j)\rangle$.

Step 3: diagonalization. Since $\Gamma \in \mathcal{B}(\mathbb{H})$ is trace class (thus compact) and positive definite, it follows from Theorem 4.2.4 in [12] that the eigendecomposition of Γ is given by

$$\Gamma = \sum_{k=1}^\infty \gamma_k(e_k \otimes e_k),$$

where $\gamma_k \geqslant 0$ are eigenvalues of Γ and $(e_k)_{k=1}^{\infty}$ are eigenfunctions forming a CONS of $\overline{\mathrm{Im}(\Gamma)}$; namely $\Gamma h = \sum_{k=1}^{\infty} \gamma_k \langle h, e_k\rangle e_k$ for every $h \in \mathbb{H}$. Here, $\otimes$ denotes the tensor product and $\overline{\mathrm{Im}(\Gamma)}$ denotes the closure of the image of Γ. In addition, there exists a unique positive definite square root operator $\Gamma^{1/2} \in \mathcal{B}(\mathbb{H})$ such that $\Gamma^{1/2}\Gamma^{1/2} = \Gamma$ (cf. Theorem 3.4.3 in [12]). Then we have $\Gamma^{1/2} g_j = \sum_{k=1}^{\infty} \gamma_k^{1/2} \langle g_j, e_k\rangle e_k$ and

$$\begin{aligned}
\tau_\delta^2 &= \sum_{i\in\Lambda_\delta} L_i^2 \langle \Gamma^{1/2}(\sum_{j\in\Lambda_\delta^c} a_{ij} g_j), \Gamma^{1/2}(\sum_{j\in\Lambda_\delta^c} a_{ij} g_j)\rangle = \sum_{i\in\Lambda_\delta} L_i^2 \|\Gamma^{1/2}(\sum_{j\in\Lambda_\delta^c} a_{ij} g_j)\|^2 \\
&= \sum_{i\in\Lambda_\delta} L_i^2 \|\sum_{j\in\Lambda_\delta^c} a_{ij} \Gamma^{1/2} g_j\|^2 = \sum_{i\in\Lambda_\delta} L_i^2 \|\sum_{k=1}^{\infty} \gamma_k^{1/2} (\sum_{j\in\Lambda_\delta^c} a_{ij} \langle g_j, e_k\rangle) e_k\|^2 \\
&= \sum_{k=1}^{\infty} \gamma_k \sum_{i\in\Lambda_\delta} \left(\sum_{j\in\Lambda_\delta^c} L_i a_{ij} \langle g_j, e_k\rangle \right)^2,
\end{aligned}$$

where the last step follows from Parseval's identity. Note that

$$\|\Gamma^{1/2} e_k\|^2 = \langle \Gamma e_k, e_k\rangle = \langle \gamma_k e_k, e_k\rangle = \gamma_k.$$

Thus for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\, e^{\lambda \langle g_j, e_k\rangle} = e^{8L_j^2\lambda^2 \langle \Gamma e_k, e_k\rangle} = e^{8L_j^2\lambda^2 \|\Gamma^{1/2} e_k\|^2} = e^{8L_j^2\lambda^2 \gamma_k},$$

which implies that $G_{jk} := \langle g_j, e_k\rangle, j = 1, \dots, n$, are independent $N(0, 16L_j^2\gamma_k)$ random variables. Now let $f = (\sqrt{\gamma_1} f_1^T, \sqrt{\gamma_2} f_2^T, \dots)^T$, where $f_k = (G_{1k}, \dots, G_{nk})^T$ for $k = 1, 2, \dots$. Then $f \sim N(0, \widetilde{\Gamma})$, where $\widetilde{\Gamma} = (\widetilde{\Gamma}_{km})_{k,m=1}^{\infty}$ with $\widetilde{\Gamma}_{km} = \mathrm{diag}(E_{km,11}, \dots, E_{km,nn})$ and $E_{km,jj} = \sqrt{\gamma_k \gamma_m}\, \mathbb{E}[G_{jk} G_{jm}]$. Note that

$$\begin{aligned}
\mathbb{E}[G_{jk} G_{jm}] &= \mathbb{E}[\langle \langle g_j, e_k\rangle g_j, e_m\rangle] = \langle (\mathbb{E}\langle g_j \otimes g_j\rangle) e_k, e_m\rangle \\
&= 16L_j^2 \langle \Gamma e_k, e_m\rangle = 16L_j^2 \langle \gamma_k e_k, e_m\rangle = 16L_j^2 \gamma_k \mathbf{1}(k = m).
\end{aligned}$$

Thus $\widetilde{\Gamma}_{km}$ is an $n \times n$ matrix of all zeros if $k \neq m$, and $\widetilde{\Gamma}_{kk} = 16\gamma_k^2 \mathrm{diag}(L_1^2, \dots, L_n^2)$.

Step 4: bound the eigenvalues. Let $P_\delta : \mathbb{R}^n \to \mathbb{R}^n$ be the restriction matrix such that $P_{\delta,ii} = 1$ if $i \in \Lambda_\delta$ and $P_{\delta,ij} = 0$ otherwise. Let further $R_\delta = \mathrm{diag}(P_\delta \widetilde{A}(I_n - P_\delta), P_\delta \widetilde{A}(I_n - P_\delta), \dots)$ and $Z = (Z_1, Z_2, \dots)^T$, where $\widetilde{A} = (\widetilde{a}_{ij})_{i,j=1}^n$ with $\widetilde{a}_{ij} = L_i a_{ij}$ and Z_i are i.i.d. standard Gaussian random variables in $\mathbb{R}$. By the rotational invariance of Gaussian distributions, we have

$$\tau_\delta^2 = \|R_\delta f\|^2 \overset{d}{=} \left\| R_\delta \widetilde{\Gamma}^{1/2} Z \right\|^2 = Z^T \widetilde{\Gamma}^{1/2} R_\delta^T R_\delta \widetilde{\Gamma}^{1/2} Z \overset{d}{=} \sum_{k=1}^{\infty} s_k^2 Z_k^2,$$

where $(s_k^2)_{k=1}^{\infty}$ are the eigenvalues of $\widetilde{\Gamma}^{1/2} R_\delta^T R_\delta \widetilde{\Gamma}^{1/2}$. So it follows that

$$\max_k s_k^2 \leqslant \|R_\delta\|_{\mathrm{op}}^2 \|\widetilde{\Gamma}\|_{\mathrm{op}} \leqslant \|\widetilde{A}\|_{\mathrm{op}}^2 \|\widetilde{\Gamma}\|_{\mathrm{op}} \leqslant L^2 \|A\|_{\mathrm{op}}^2 \|\widetilde{\Gamma}\|_{\mathrm{op}},$$

where

$$\|\widetilde{\Gamma}\|_{\mathrm{op}} \leqslant 16(\max_{1\leqslant j\leqslant n}\|X_j\|_{\psi_2}^2)(\max_k \gamma_k^2) \leqslant 16L^2\|\Gamma\|_{\mathrm{op}}^2.$$

In addition, we also have

$$\begin{aligned}\sum_k s_k^2 &= \mathrm{tr}(\widetilde{\Gamma}^{1/2}R_\delta^T R_\delta \widetilde{\Gamma}^{1/2}) = \mathrm{tr}(R_\delta \widetilde{\Gamma} R_\delta^T) = \sum_{k=1}^{\infty}\mathrm{tr}([P_\delta \widetilde{A}(I_n - P_\delta)]\widetilde{\Gamma}_{kk}[P_\delta \widetilde{A}(I_n - P_\delta)]^T)\\ &\leqslant \sum_{k=1}^{\infty} 16L^2\gamma_k^2\|P_\delta \widetilde{A}(I_n - P_\delta)\|_{\mathrm{HS}}^2 \leqslant \sum_{k=1}^{\infty} 16L^2\gamma_k^2\|\widetilde{A}\|_{\mathrm{HS}}^2 \leqslant 16L^4\|\Gamma\|_{\mathrm{HS}}^2\|A\|_{\mathrm{HS}}^2.\end{aligned}$$

Invoking (4.1), we get

$$\mathbb{E}_X[e^{4\lambda S_\delta}] \leqslant \prod_{k=1}^{\infty}\mathbb{E}_Z[\exp(\lambda^2 s_k^2 Z_k^2/2)].$$

Since Z_k^2 are i.i.d. χ_1^2 random variables with the moment generating function $\mathbb{E}[e^{tZ_k^2}] = (1-2t)^{-1/2}$ for $t < 1/2$, we have

$$\mathbb{E}_X[e^{4\lambda S_\delta}] \leqslant \prod_{k=1}^{\infty}\frac{1}{\sqrt{1-\lambda^2 s_k^2}}, \qquad \text{if } \max_k \lambda^2 s_k^2 < 1.$$

Using $(1-z)^{-1/2} \leqslant e^z$ for $z \in [0, 1/2]$, we get that if $16L^4\|A\|_{\mathrm{op}}^2\|\Gamma\|_{\mathrm{op}}^2\lambda^2 < 1$, then

$$\mathbb{E}_X[e^{4\lambda S_\delta}] \leqslant \exp(\lambda^2\sum_{k=1}^{\infty} s_k^2) \leqslant \exp(16\lambda^2 L^4\|\Gamma\|_{\mathrm{HS}}^2\|A\|_{\mathrm{HS}}^2).$$

Note that the last inequality is uniform in δ. Taking expectation with respect to δ, we obtain that

$$\mathbb{E}_X[e^{\lambda S}] \leqslant \mathbb{E}_{X,\delta}[e^{4\lambda S_\delta}] \leqslant \exp(16\lambda^2 L^4\|\Gamma\|_{\mathrm{HS}}^2\|A\|_{\mathrm{HS}}^2),$$

whenever $0 < \lambda < (4L^2\|A\|_{\mathrm{op}}\|\Gamma\|_{\mathrm{op}})^{-1}$.

Step 5: conclusion. Now we have

$$\mathbb{P}(S \geqslant t) \leqslant \exp(-\lambda t + 16\lambda^2 L^4\|\Gamma\|_{\mathrm{HS}}^2\|A\|_{\mathrm{HS}}^2) \quad \text{for } 0 < \lambda \leqslant (8L^2\|A\|_{\mathrm{op}}\|\Gamma\|_{\mathrm{op}})^{-1}.$$

Optimizing in λ, we deduce that there exists a universal constant $C > 0$ such that

$$\mathbb{P}(S \geqslant t) \leqslant \exp\left[-C\min\left(\frac{t^2}{L^4\|\Gamma\|_{\mathrm{HS}}^2\|A\|_{\mathrm{HS}}^2}, \frac{t}{L^2\|\Gamma\|_{\mathrm{op}}\|A\|_{\mathrm{op}}}\right)\right],$$

as desired in (2.3). ■

Proof of Theorem 2.6. Decompose $Q = \sum_{i=1}^n a_{ii}\|X_i\|^2 + S$, where $S = \sum_{1\leqslant i\neq j\leqslant n} a_{ij}\langle X_i, X_j\rangle$. In view of the off-diagonal sum bound for S in Proposition 2.5, it suffices to show the following inequality for the diagonal sum: for any $t > 0$,

$$\begin{aligned}&\mathbb{P}\left(\sum_{i=1}^n a_{ii}\|X_i\|^2 \geqslant \sum_{i=1}^n a_{ii}L_i^2\|\Gamma\|_{\mathrm{tr}} + t\right)\\ &\qquad\leqslant \exp\left[-C\min\left(\frac{t^2}{L^4\|\Gamma\|_{\mathrm{HS}}^2\sum_{i=1}^n a_{ii}^2}, \frac{t}{L^2\|\Gamma\|_{\mathrm{op}}\max_{1\leqslant i\leqslant n} a_{ii}}\right)\right],\end{aligned} \tag{4.2}$$

since $\sum_{i=1}^n a_{ii}^2 \leqslant \|A\|_{\mathrm{HS}}^2$ and $\overline{a} := \max_{1\leqslant i\leqslant n} a_{ii} \leqslant \|A\|_{\mathrm{op}}$. By Markov's inequality and Lemma A.3, we have for any $\lambda > 0$ and $t > 0$,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n a_{ii}(\|X_i\|^2 - L_i^2\|\Gamma\|_{\mathrm{tr}}) \geqslant t\right) &\leqslant e^{-\lambda t}\prod_{i=1}^n \mathbb{E}[e^{\lambda a_{ii}(\|X_i\|^2 - L_i^2\|\Gamma\|_{\mathrm{tr}})}]\\ &\leqslant e^{-\lambda t}\prod_{i=1}^n e^{2\lambda^2 a_{ii}^2 L_i^4\|\Gamma\|_{\mathrm{HS}}^2} \leqslant \exp\left(-\lambda t + 2\lambda^2(\sum_{i=1}^n a_{ii}^2)L^4\|\Gamma\|_{\mathrm{HS}}^2\right)\end{aligned}$$

holds for all $0 \leqslant \lambda < (4L^2\|\Gamma\|_{\mathrm{op}}\overline{a})^{-1}$. Choosing

$$\lambda = \frac{t}{4(\sum_{i=1}^n a_{ii}^2)L^4\|\Gamma\|_{\mathrm{HS}}^2} \wedge \frac{1}{8\overline{a}L^2\|\Gamma\|_{\mathrm{op}}},$$

we get (4.2). ∎

Proof of Theorem 2.8. Under Assumption 2.7, we have the following standard moment generating function bound

$$\mathbb{E}\left[e^{\lambda(\|X_i\|^2 - \mathbb{E}\|X_i\|^2)}\right] \leqslant e^{\frac{C\lambda^2\|\Gamma\|_{\mathrm{HS}}^2}{2}} \quad \forall|\lambda| < \frac{1}{2\|\Gamma\|_{\mathrm{op}}}.$$

See for example Chapter 2 in [30]. Then we have for any $\lambda > 0$ and $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n a_{ii}(\|X_i\|^2 - \mathbb{E}\|X_i\|^2) \geqslant t\right) \leqslant \exp\left(-\lambda t + C\lambda^2(\sum_{i=1}^n a_{ii}^2)\|\Gamma\|_{\mathrm{HS}}^2\right) \quad \forall|\lambda| < \frac{1}{2\overline{a}\|\Gamma\|_{\mathrm{op}}},$$

where $\overline{a} := \max_{1\leqslant i\leqslant n}|a_{ii}|$. Note that $\sum_{i=1}^n a_{ii}^2 \leqslant \|A\|_{\mathrm{HS}}^2$ and $\overline{a} \leqslant \|A\|_{\mathrm{op}}$. Optimizing over λ and combining with Proposition 2.5, we get

$$\mathbb{P}\left(Q - \mathbb{E}[Q] \geqslant t\right) \leqslant 2\exp\left[-C\min\left(\frac{t^2}{L^4\|\Gamma\|_{\mathrm{HS}}^2\|A\|_{\mathrm{HS}}^2}, \frac{t}{L^2\|\Gamma\|_{\mathrm{op}}\|A\|_{\mathrm{op}}}\right)\right].$$

Applying the same argument by replacing Q with $-Q$, we obtain (2.7) with constant 4, which can be reduced to 2 by adjusting the value of constant C. ∎

4.2. Proof of main results in Section 3

In this subsection, we prove Theorem 3.3 and Corollary 3.4.

Theorem 3.3. Recall that $\mathscr{C} = \{Z_{n\times n} : Z^T = Z, Z \succeq 0, \mathrm{tr}(Z) = K, Z\mathbf{1}_n = \mathbf{1}_n, Z \geqslant 0\}$ is the SDP constraint set for the generalized K-means in (3.3). For $i \in G_k^*$, let $\mu_k = \mathbb{E}[X_i]$ and $\delta_i = X_i - \mu_k$. For notation simplicity, we will omit in the proof the subscript $\mathbb{H}$ in the Hilbert space inner product $\langle\cdot,\cdot\rangle_{\mathbb{H}}$ and norm $\|\cdot\|_{\mathbb{H}}$.

Step 1: a generic bound. For any $Z \in \mathscr{C}$, consider $\langle A, Z - Z^*\rangle = \sum_{i,j=1}^n a_{ij}(Z_{ij} - Z_{ij}^*)$. Note that if $i \in G_k^*$ and $j \in G_m^*$, then

$$\begin{aligned} a_{ij} =& \langle \mu_k + \delta_i, \mu_m + \delta_j\rangle = \langle \mu_k, \mu_m\rangle + \langle \mu_k, \delta_j\rangle + \langle \delta_i, \mu_m\rangle + \langle \delta_i, \delta_j\rangle \\ =& \langle \mu_k, \mu_m\rangle + \langle \mu_k - \mu_m, \delta_j - \delta_i\rangle + \langle \mu_k, \delta_i\rangle + \langle \delta_j, \mu_m\rangle + \langle \delta_i, \delta_j\rangle \\ =& -\frac{1}{2}\|\mu_k - \mu_m\|^2 + \frac{1}{2}(\|\mu_k\|^2 + \|\mu_m\|^2) + \langle \mu_k - \mu_m, \delta_j - \delta_i\rangle + \langle \mu_k, \delta_i\rangle + \langle \delta_j, \mu_m\rangle + \langle \delta_i, \delta_j\rangle. \end{aligned}$$

Since $\sum_{j=1}^n Z_{ij} = (Z\mathbf{1}_n)_i = 1$ for all $Z \in \mathscr{C}$ and Z^* is feasible for $\mathscr{C}$, we have

$$\sum_{i,j=1}^n \sum_{k,m=1}^K \|\mu_k\|^2 \mathbf{1}(i \in G_k^*, j \in G_m^*)(Z_{ij} - Z_{ij}^*) = \sum_{i=1}^n \sum_{k=1}^K \|\mu_k\|^2 \mathbf{1}(i \in G_k^*) \sum_{j=1}^n (Z_{ij} - Z_{ij}^*) = 0$$

and

$$\sum_{i,j=1}^n \sum_{k,m=1}^K \langle \mu_k, \delta_i\rangle \mathbf{1}(i \in G_k^*, j \in G_m^*)(Z_{ij} - Z_{ij}^*) = \sum_{i=1}^n \sum_{k=1}^K \langle \mu_k, \delta_i\rangle \mathbf{1}(i \in G_k^*) \sum_{j=1}^n (Z_{ij} - Z_{ij}^*) = 0.$$

Then by the symmetry of Z (i.e., $Z^T = Z$), we have

$$\langle A, Z - Z^*\rangle = \langle T_1 + T_2 + T_3 + T_4, Z - Z^*\rangle,$$

where for $i \in G_k^*$ and $j \in G_m^*$,

$$\begin{aligned} T_{1,ij} &= -\frac{1}{2}\|\mu_k - \mu_m\|^2, \qquad & T_{2,ij} &= \langle \mu_k - \mu_m, \delta_j - \delta_i\rangle, \\ T_{3,ij} &= \langle \delta_i, \delta_j\rangle - \mathbb{E}\langle \delta_i, \delta_j\rangle, \qquad & T_{4,ij} &= \mathbb{E}\langle \delta_i, \delta_j\rangle. \end{aligned}$$

Observe that

$$\begin{aligned} \langle T_1, Z - Z^*\rangle &= -\frac{1}{2} \sum_{1\leqslant k\neq m\leqslant K} \|\mu_k - \mu_m\|^2 \sum_{i\in G_k^*, j\in G_m^*} (Z_{ij} - Z_{ij}^*) && (4.3) \\ &= -\frac{1}{2} \sum_{1\leqslant k\neq m\leqslant K} \|\mu_k - \mu_m\|^2 |Z_{G_k^* G_m^*}|_1, && (4.4) \end{aligned}$$

where the last step follows from $Z \geqslant 0$ and $Z^*_{ij} = 0$ if $i \in G^*_k, j \in G^*_m$ for $k \neq m$. Here, $|Z_{G^*_k G^*_m}|_1 = \sum_{i \in G^*_k, j \in G^*_m} |Z_{ij}|$. By definition, we have $\langle A, Z^* \rangle \leqslant \langle A, \hat{Z} \rangle$, which implies that $0 \leqslant \langle A, \hat{Z} - Z^* \rangle$. Thus we have

$$0 \leqslant \frac{1}{2} \sum_{1 \leqslant k \neq m \leqslant K} \|\mu_k - \mu_m\|^2 |\hat{Z}_{G^*_k G^*_m}|_1 = \langle T_1, Z^* - \hat{Z} \rangle \leqslant \langle T_2 + T_3 + T_4, \hat{Z} - Z^* \rangle. \quad (4.5)$$

Let $\Delta = \min_{1 \leqslant k \neq m \leqslant K} \|\mu_k - \mu_m\|$. By (A.5) and (A.3) in Lemma A.6, we have

$$|\hat{Z} - Z^*|_1 \leqslant \frac{2n}{\underline{n}} |Z^* - Z^* \hat{Z}|_1 = \frac{4n}{\underline{n}} \sum_{1 \leqslant k \neq m \leqslant K} |\hat{Z}_{G^*_k G^*_m}|_1,$$

where $\underline{n} = \min_{1 \leqslant k \leqslant K} n_k$. Then we get

$$|\hat{Z} - Z^*|_1 \leqslant \frac{8n}{\Delta^2 \underline{n}} \langle T_2 + T_3 + T_4, \hat{Z} - Z^* \rangle. \quad (4.6)$$

Step 2: bound $\langle T_4, \hat{Z} - Z^* \rangle$. Since $\delta_1, \ldots, \delta_n$ are independent with mean zero, we have

$$\langle T_4, \hat{Z} - Z^* \rangle = \sum_{i=1}^n \mathbb{E} \|\delta_i\|^2 (\hat{Z}_{ii} - Z^*_{ii}).$$

Since $\mathbb{E} \|\delta_i\|^2 = \|\mathbb{E}[\delta_i \otimes \delta_i]\|_{\mathrm{tr}} = \|\Sigma_k\|_{\mathrm{tr}}$ if $i \in G^*_k$, and $\|\Sigma_k\|_{\mathrm{tr}}, k = 1, \ldots, K$ are all equal, it follows that

$$\langle T_4, \hat{Z} - Z^* \rangle = \|\Sigma_1\|_{\mathrm{tr}} \operatorname{tr}(\hat{Z} - Z^*) = 0,$$

where the last step is due to $\operatorname{tr}(\hat{Z}) = \operatorname{tr}(Z^*) = K$ since both $\hat{Z}, Z^* \in \mathscr{C}$.

Step 3: bound $\langle T_2, \hat{Z} - Z^* \rangle$. Consider

$$\begin{aligned}
\langle T_2, \hat{Z} - Z^* \rangle &= \sum_{1 \leqslant k \neq m \leqslant K} \sum_{i,j=1}^n \langle \mu_k - \mu_m, \delta_j - \delta_i \rangle \mathbf{1}(i \in G^*_k, j \in G^*_m)(\hat{Z}_{ij} - Z^*_{ij}) \\
&= \sum_{1 \leqslant k \neq m \leqslant K} \sum_{i \in G^*_k, j \in G^*_m} \langle \mu_k - \mu_m, \delta_j - \delta_i \rangle \hat{Z}_{ij} \\
&= 2 \sum_{1 \leqslant k \neq m \leqslant K} \sum_{i \in G^*_k, j \in G^*_m} \langle \mu_k - \mu_m, \delta_i \rangle \hat{Z}_{ij} \\
&= 2 \sum_{1 \leqslant k \neq m \leqslant K} \sum_{i \in G^*_k} \langle \mu_k - \mu_m, \delta_i \rangle |\hat{Z}_{i G^*_m}|_1,
\end{aligned}$$

where the third equality is due to symmetry. For each $k \neq m$, let $\epsilon^{(k,m)}_i = \langle \mu_k - \mu_m, \delta_i \rangle$ and $s_{k,m} = \sum_{i \in G^*_k} |\hat{Z}_{i G^*_m}|_1$. Since $|\hat{Z}_{i G^*_m}|_1 \leqslant 1$, by Lemma A.5,

$$\sum_{i \in G^*_k} \langle \mu_k - \mu_m, \delta_i \rangle |\hat{Z}_{i G^*_m}|_1 \leqslant \sum_{i=1}^{s_{k,m}} \epsilon^{(k,m)}_{(i)},$$

where $\epsilon_{(1)}^{(k,m)} \geqslant \ldots \geqslant \epsilon_{(n)}^{(k,m)}$ are the order statistics of $\epsilon_1^{(k,m)}, \ldots, \epsilon_n^{(k,m)}$. Note that $(\epsilon_i^{(k,m)})_{i=1}^n$ are i.i.d. mean-zero sub-gaussian random variables in $\mathbb{R}$ with respect to $\tau_{k,m}^2 := L^2 \langle \Sigma(\mu_k - \mu_m), \mu_k - \mu_m \rangle$ (recall that $\Sigma \succeq \Sigma_k$ for all $k = 1, \ldots, K$). Thus for any $s = 1, \ldots, n$, we have $\sum_{i=1}^s \epsilon_i^{(k,m)}$ is a mean-zero sub-gaussian random variable with respect to $s\tau_{k,m}^2$. By the union bound, we get for all $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{s} \epsilon_{(i)}^{(k,m)} \geqslant t\right) \leqslant \binom{n}{s} \exp\left(-\frac{t^2}{2s\tau_{k,m}^2}\right) \leqslant \left(\frac{en}{s}\right)^s \exp\left(-\frac{t^2}{2s\tau_{k,m}^2}\right).$$

Now it follows that

$$\begin{aligned}
&\mathbb{P}\left(\exists 1 \leqslant k \neq m \leqslant K \text{ such that } \sum_{i=1}^{s_{k,m}} \epsilon_{(i)}^{(k,m)} \geqslant C_1 \tau_{k,m} s_{k,m} \sqrt{\log\left(\frac{nK}{s_{k,m}}\right)}\right) \\
\leqslant &\sum_{1 \leqslant k \neq m \leqslant K} \sum_{1 \leqslant s \leqslant n} \mathbb{P}\left(\sum_{i=1}^{s} \epsilon_{(i)}^{(k,m)} \geqslant C_1 \tau_{k,m} s \sqrt{\log\left(\frac{nK}{s}\right)}\right) \\
\leqslant &\sum_{1 \leqslant k \neq m \leqslant K} \sum_{s=1}^{n} \left(\frac{en}{s}\right)^s \exp\left(-\frac{C_1^2}{2} s \log\left(\frac{nK}{s}\right)\right) \\
\leqslant &K^2 \sum_{s=1}^{n} \exp\left(-C_2 s \log\left(\frac{nK}{s}\right)\right) \leqslant \frac{C_3 K^2}{(nK)^2} = \frac{C_3}{n^2}.
\end{aligned}$$

Thus we have $\mathbb{P}(\mathcal{G}_1) \geqslant 1 - C_3 n^{-2}$, where

$$\mathcal{G}_1 = \left\{ \sum_{i=1}^{s_{k,m}} \epsilon_{(i)}^{(k,m)} \leqslant C_1 \tau_{k,m} s_{k,m} \sqrt{\log\left(\frac{nK}{s_{k,m}}\right)} \quad \forall 1 \leqslant k \neq m \leqslant K \right\}.$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\langle T_2, \hat{Z} - Z^* \rangle \leqslant &2C_1 \sum_{1 \leqslant k \neq m \leqslant K} \tau_{k,m} s_{k,m} \sqrt{\log\left(\frac{nK}{s_{k,m}}\right)} \\
\leqslant &2C_1 \sqrt{\sum_{1 \leqslant k \neq m \leqslant K} \tau_{k,m}^2 s_{k,m}} \sqrt{\sum_{1 \leqslant k \neq m \leqslant K} s_{k,m} \log\left(\frac{nK}{s_{k,m}}\right)}
\end{aligned}$$

on the event $\mathcal{G}_1$. Since $s_{k,m} = |\hat{Z}_{G_k^* G_m^*}|_1$ and

$$\tau_{k,m} \leqslant L \|\Sigma^{1/2}(\mu_k - \mu_m)\| \leqslant L \|\Sigma^{1/2}\|_{\mathrm{op}} \|\mu_k - \mu_m\| = L \|\Sigma\|_{\mathrm{op}}^{1/2} \|\mu_k - \mu_m\|,$$

it follows from the first equality in (4.5) that

$$\sum_{1 \leqslant k \neq m \leqslant K} \tau_{k,m}^2 s_{k,m} \leqslant \sum_{1 \leqslant k \neq m \leqslant K} L^2 \|\Sigma\|_{\mathrm{op}} \|\mu_k - \mu_m\|^2 |\hat{Z}_{G_k^* G_m^*}|_1 = 2L^2 \|\Sigma\|_{\mathrm{op}} \langle T_1, Z^* - \hat{Z} \rangle.$$

By (A.3) in Lemma A.6, $S := |Z^*(\hat{Z} - Z^*)|_1 = 2\sum_{1 \leqslant k \neq m \leqslant K} s_{k,m}$. Then it follows from Jensen's inequality that

$$\sum_{1 \leqslant k \neq m \leqslant K} s_{k,m} \log\left(\frac{nK}{s_{k,m}}\right) \leqslant \frac{S}{2} \log\left(\frac{2nK^3}{S}\right).$$

Thus we get

$$\langle T_2, \hat{Z} - Z^* \rangle \leqslant 2C_1 L \sqrt{\|\Sigma\|_{\mathrm{op}} \langle T_1, Z^* - \hat{Z} \rangle} \sqrt{S \log\left(\frac{2nK^3}{S}\right)} \tag{4.7}$$

on the event $\mathcal{G}_1$.

Step 4: bound $\langle T_3, \hat{Z} - Z^* \rangle$. Decompose

$$\langle T_3, \hat{Z} - Z^* \rangle = \langle (I - Z^*) T_3 (I - Z^*), \hat{Z} - Z^* \rangle + \langle Z^* T_3, \hat{Z} - Z^* \rangle + \langle T_3 Z^*, \hat{Z} - Z^* \rangle - \langle Z^* T_3 Z^*, \hat{Z} - Z^* \rangle.$$

Note that

$$\begin{aligned}
\langle (I - Z^*) T_3 (I - Z^*), \hat{Z} - Z^* \rangle &=_{(1)} \langle T_3, (I - Z^*)(\hat{Z} - Z^*)(I - Z^*) \rangle \\
&=_{(2)} \langle T_3, (I - Z^*) \hat{Z} (I - Z^*) \rangle \\
&\leqslant_{(3)} \|T_3\|_{\mathrm{op}} \|(I - Z^*) \hat{Z} (I - Z^*)\|_{\mathrm{tr}} \\
&\leqslant_{(4)} \|T_3\|_{\mathrm{op}} \frac{|Z^* - Z^* \hat{Z}|_1}{2n},
\end{aligned}$$

where (1) follows from the symmetry of Z^*, (2) from the idempotence of Z^* (recall that Z^* is a projection matrix such that $Z^* Z^* = Z^*$), (3) from the duality of the operator and trace norms, and (4) from (A.4) in Lemma A.6. Let $\mathbb{S}^{n-1}$ be the (compact) unit sphere in $\mathbb{R}^n$ and $\mathcal{N}$ be a 1/4-net for $\mathbb{S}^{n-1}$. By Lemma 5.2 and 5.4 in [27], we have $|\mathcal{N}| \leqslant 9^n$ and $\|T_3\|_{\mathrm{op}} \leqslant 2 \max_{x \in \mathcal{N}} x^T T_3 x$. Thus, by the union bound, we have for any $t > 0$,

$$\mathbb{P}(\|T_3\|_{\mathrm{op}} \geqslant t) \leqslant \sum_{x \in \mathcal{N}} \mathbb{P}(x^T T_3 x \geqslant t/2). \tag{4.8}$$

Fix an $x \in \mathcal{N}$. Note that $\|xx^T\|_{\mathrm{HS}}^2 = \|x\|_2^4 = 1$ and $\|xx^T\|_{\mathrm{op}} \leqslant 1$. Since $\Sigma \succeq \Sigma_k$ for all $k = 1, \ldots, K$, we have $\delta_i \sim$ sub-gaussian(Σ) such that $\mathbb{E}[\delta_i] = 0$ and $\|\delta_i\|_{\psi_2, \Sigma} \leqslant L$. By Theorem 2.8 with $A = xx^T$, we get for all $t > 0$,

$$\mathbb{P}(x^T T_3 x \geqslant t/2) = \mathbb{P}\Big(\sum_{i,j=1}^{n} x_i x_j T_{3,ij} \geqslant t/2\Big) \leqslant 2 \exp\left[-C \min\left(\frac{t^2}{L^4 \|\Sigma\|_{\mathrm{HS}}^2}, \frac{t}{L^2 \|\Sigma\|_{\mathrm{op}}}\right)\right].$$

Combining the last inequality with (4.8), we obtain that with probability at least $1 - cn^{-2}$,

$$\|T_3\|_{\mathrm{op}} \leqslant C_5 L^2 (\sqrt{n} \|\Sigma\|_{\mathrm{HS}} + n \|\Sigma\|_{\mathrm{op}}).$$

Then,

$$\begin{aligned}\langle (I-Z^*)T_3(I-Z^*), \hat{Z}-Z^*\rangle &\leqslant C_5 L^2 \frac{\sqrt{n}\|\Sigma\|_{\mathrm{HS}} + n\|\Sigma\|_{\mathrm{op}}}{2\underline{n}} |Z^* - Z^*\hat{Z}|_1 \\ &\leqslant_{(1)} C_5 \frac{\Delta^2}{2}(c_0^{-1} + c_0^{-1/2})|Z^* - Z^*\hat{Z}|_1 \\ &=_{(2)} C_5 \frac{\Delta^2}{2}(c_0^{-1} + c_0^{-1/2}) 2 \sum_{1\leqslant k\neq m\leqslant K} |\hat{Z}_{G_k^* G_m^*}|_1 \\ &\leqslant_{(3)} 2C_5(c_0^{-1} + c_0^{-1/2})\langle T_1, Z^* - \hat{Z}\rangle \\ &\leqslant_{(4)} \frac{1}{2}\langle T_1, Z^* - \hat{Z}\rangle,\end{aligned}$$

where (1) follows from the definition of SNR^2 and the condition that $\mathsf{SNR}^2 \geqslant c_0\, n/\underline{n}$, (2) from (A.3) in Lemma A.6, (3) from the definition of Δ^2 and (4.5), and (4) from choosing c_0 sufficiently large.

Next, we consider $\langle Z^*T_3, \hat{Z} - Z^*\rangle = \langle Z^*T_3, Z^*\hat{Z} - Z^*\rangle$. By (3.4), we have

$$\begin{aligned}\langle Z^*T_3, Z^*\hat{Z} - Z^*\rangle &= \sum_{i,j=1}^n (Z^*T_3)_{ij}(Z^*\hat{Z} - Z^*)_{ij} \\ &= \sum_{k,m=1}^K \sum_{i\in G_k^*} \sum_{j\in G_m^*} \left(\sum_{\ell=1}^n Z^*_{i\ell} T_{3,\ell j}\right)\left(\sum_{\ell=1}^n Z^*_{i\ell}\hat{Z}_{\ell j} - Z^*_{ij}\right) \\ &= \sum_{k,m=1}^K \sum_{i\in G_k^*} \sum_{j\in G_m^*} \left(\frac{1}{n_k}\sum_{\ell\in G_k^*} T_{3,\ell j}\right)\frac{1}{n_k}\left(\sum_{\ell\in G_k^*}\hat{Z}_{\ell j} - \mathbf{1}(k=m)\right) \\ &= \sum_{k,m=1}^K \sum_{j\in G_m^*} \underbrace{\left(\frac{(-1)^{\mathbf{1}(k=m)}}{n_k}\sum_{\ell\in G_k^*} T_{3,\ell j}\right)}_{=:B_{kj}} \underbrace{\left|(Z^* - Z^*\hat{Z})_{G_k^* j}\right|_1}_{=:\beta_{kj}}.\end{aligned}$$

Note that $\beta_{kj} \in [0,1]$. By Lemma A.5, we have

$$\langle Z^*T_3, Z^*\hat{Z} - Z^*\rangle \leqslant \sum_{k,m=1}^K \sum_{j=1}^{b_{km}} B_{(j)}^{(k,m)},$$

where $b_{km} = \sum_{j\in G_m^*}\beta_{kj} = |(Z^* - Z^*\hat{Z})_{G_k^* G_m^*}|_1$ and $B_{(1)}^{(k,m)} \geqslant B_{(2)}^{(k,m)} \geqslant \cdots$ is the ordered sequence of $(B_{kj})_{j\in G_m^*}$. Now fix a (k,m). For any $E \subset G_m^*$ with $1 \leqslant q := |E| \leqslant n_m$, we can write

$$\sum_{j\in E} B_{kj} = \sum_{j,\ell=1}^n d_{\ell j}^{(k,m)}\left(\langle \delta_\ell, \delta_j\rangle - \mathbb{E}\langle \delta_\ell, \delta_j\rangle\right),$$

where $D^{(k,m)} = (d_{\ell j}^{(k,m)})_{\ell,j=1}^{n}$ and $d_{\ell j}^{(k,m)} = -n_m^{-1}\mathbf{1}(j \in E)\mathbf{1}(\ell \in G_k^*)$. By Theorem 2.8 (one-sided version) and the union bound, we have $t > 0$,

$$\mathbb{P}\left(\sum_{j=1}^{q} B_{(j)}^{(k,m)} \geqslant t\right) \leqslant \binom{n_m}{q} \exp\left[-C\min\left(\frac{t^2}{L^4\|\Sigma\|_{\mathrm{HS}}^2\|D^{(k,m)}\|_{\mathrm{HS}}^2}, \frac{t}{L^2\|\Sigma\|_{\mathrm{op}}\|D^{(k,m)}\|_{\mathrm{op}}}\right)\right].$$

Since $\|D^{(k,m)}\|_{\mathrm{HS}} = \|D^{(k,m)}\|_{\mathrm{op}} = \sqrt{q/n_m}$, we deduce that

$$\begin{aligned}
&\mathbb{P}\Bigg(\exists 1 \leqslant k,\, m \leqslant K \text{ such that } \sum_{i=1}^{b_{km}} B_{(j)}^{(k,m)} \geqslant \\
&\qquad\qquad C_6\, L^2\left(\|\Sigma\|_{\mathrm{HS}}\frac{b_{km}}{\sqrt{n_m}}\sqrt{\log\frac{n_m K}{b_{km}}} + \|\Sigma\|_{\mathrm{op}}\frac{b_{km}^{3/2}}{\sqrt{n_m}}\log\frac{n_m K}{b_{km}}\right)\Bigg) \\
&\leqslant \sum_{k,m=1}^{K}\sum_{1\leqslant q\leqslant n_m} \mathbb{P}\left(\sum_{j=1}^{q} B_{(j)}^{(k,m)} \geqslant C_6\, L^2\left(\|\Sigma\|_{\mathrm{HS}}\frac{q}{\sqrt{n_m}}\sqrt{\log\frac{n_m K}{q}} + \|\Sigma\|_{\mathrm{op}}\frac{q^{3/2}}{\sqrt{n_m}}\log\frac{n_m K}{q}\right)\right) \\
&\leqslant \sum_{k,m=1}^{K}\sum_{q=1}^{n_m}\left(\frac{en_m}{q}\right)^q \exp\left(-C_6^2\, q\log\left(\frac{n_m K}{q}\right)\right) \\
&\leqslant K^2 \min_m \sum_{q=1}^{n_m} \exp\left(-C_7\, q\log\left(\frac{n_m K}{q}\right)\right) \leqslant \frac{C_8 K^2}{(\underline{n}K)^4} \leqslant \frac{C_8}{{c_0'}^2 n^2},
\end{aligned}$$

where the last inequality is due to $\underline{n}^2 K \geqslant c_0' n$. Thus, we obtain that with probability at least $1 - C_8/n^2$ that

$$\langle Z^* T_3, \hat{Z} - Z^*\rangle \leqslant C_6\, L^2 \sum_{k,m=1}^{K}\left(\|\Sigma\|_{\mathrm{HS}}\frac{b_{km}}{\sqrt{n_m}}\sqrt{\log\frac{n_m K}{b_{km}}} + \|\Sigma\|_{\mathrm{op}}\frac{b_{km}^{3/2}}{\sqrt{n_m}}\log\frac{n_m K}{b_{km}}\right).$$

Recall that $\sum_{k,m=1}^{K} b_{km} = |Z^* - Z^*\hat{Z}|_1 = S$. Since functions $x^{-1/2}\log x$ and $x^{-1/2}\sqrt{\log x}$ are monotonically decreasing for $x \geqslant e^2$, we obtain from Jensen's inequality that

$$\langle Z^* T_3, \hat{Z} - Z^*\rangle \leqslant C_6\, L^2\, \frac{S}{\sqrt{\underline{n}}}\left(\|\Sigma\|_{\mathrm{HS}}\sqrt{\log\frac{\underline{n}K^3}{S}} + \|\Sigma\|_{\mathrm{op}}\sqrt{S}\log\frac{\underline{n}K^3}{S}\right).$$

By the cyclic invariance of trace and the symmetry of T_3 and $\hat{Z} - Z^*$, the same bound holds for $\langle T_3 Z^*, \hat{Z} - Z^*\rangle = \langle Z^* T_3, \hat{Z} - Z^*\rangle$. In addition, the term $\langle Z^* T_3 Z^*, \hat{Z} - Z^*\rangle = \langle Z^* T_3, Z^*(\hat{Z} - Z^*)Z^*\rangle$ can be handled in the same way as $\langle Z^* T_3, \hat{Z} - Z^*\rangle$, by noticing that $|Z^*(\hat{Z} - Z^*)Z^*|_1 = |Z^*(\hat{Z} - Z^*)|_1$ according to Lemma A.6.

Put all pieces together, we obtain that with probability at least $1 - c/n^2$ that

$$\langle T_3, \hat{Z} - Z^*\rangle \leqslant \frac{1}{2}\langle T_1, Z^* - \hat{Z}\rangle + 3C_6\, L^2\, S\, \frac{1}{\sqrt{\underline{n}}}\left(\|\Sigma\|_{\mathrm{HS}}\sqrt{\log\frac{\underline{n}K^3}{S}} + \|\Sigma\|_{\mathrm{op}}\sqrt{S}\log\frac{\underline{n}K^3}{S}\right).$$

Step 5: conclude. Now we combine the bounds in Step 1 – 4 to obtain that

$$\frac{1}{2}\langle T_1, Z^* - \hat{Z}\rangle \leqslant 2C_1 L \sqrt{\langle T_1, Z^* - \hat{Z}\rangle}\sqrt{\|\Sigma\|_{\mathrm{op}} S \log\left(\frac{2nK^3}{S}\right)} + 3C_6 L^2 S \frac{1}{\sqrt{\underline{n}}}\left(\|\Sigma\|_{\mathrm{HS}}\sqrt{\log\frac{\underline{n}K^3}{S}} + \|\Sigma\|_{\mathrm{op}}\sqrt{S}\log\frac{\underline{n}K^3}{S}\right). \tag{4.9}$$

holds with probability at least $1 - c/n^2$, where recall that $S = |Z^* - Z^*\hat{Z}|_1$. According to equation (4.5) in Step 1 and equation (A.5) in Lemma A.6, we have $\langle T_1, Z^* - \hat{Z}\rangle \geqslant \Delta^2 S/4 \geqslant 0$. Then solution of the quadratic inequality (4.9) for $\sqrt{\langle T_1, Z^* - \hat{Z}\rangle}$ implies

$$\Delta^2 \leqslant C_9 L^2 \|\Sigma\|_{\mathrm{op}} \log\left(\frac{2nK^3}{S}\right) + C_9 L^2 \frac{1}{\sqrt{\underline{n}}}\left(\|\Sigma\|_{\mathrm{HS}}\sqrt{\log\frac{\underline{n}K^3}{S}} + \|\Sigma\|_{\mathrm{op}}\sqrt{S}\log\frac{\underline{n}K^3}{S}\right). \tag{4.10}$$

This inequality combined with $S \leqslant |Z^* - \hat{Z}|_1$ due to (A.5) and the trivial upper bound $|Z^* - \hat{Z}|_1 \leqslant 2n$ imply

$$\Delta^2 \leqslant 3C_9 L^2 \|\Sigma\|_{\mathrm{op}}\sqrt{\frac{n}{\underline{n}}}\log\left(\frac{2nK^3}{S}\right) + C_9 L^2 \frac{1}{\sqrt{\underline{n}}}\|\Sigma\|_{\mathrm{HS}}\sqrt{\log\frac{\underline{n}K^3}{S}}.$$

As a consequence, we have

$$S \leqslant 2nK^3 \exp\left(-C_{10}\left(\sqrt{\frac{\underline{n}}{n}}\frac{\Delta^2}{L^2\|\Sigma\|_{\mathrm{op}}} \wedge \frac{\underline{n}\,\Delta^4}{L^4\|\Sigma\|_{\mathrm{HS}}^2}\right)\right) \leqslant 2nK^3\exp(-C_{11}\sqrt{n/\underline{n}}) \leqslant \underline{n},$$

where we have used in the second last step our condition that $\mathsf{SNR}^2 \geqslant c_0\, n/\underline{n} \geqslant c_0\, K$ for sufficiently large constant c_0. Now combining the preceding display with inequality (4.10), we obtain

$$\Delta^2 \leqslant 3C_9 L^2 \|\Sigma\|_{\mathrm{op}} \log\left(\frac{2nK^3}{S}\right) + C_9 L^2 \frac{1}{\sqrt{\underline{n}}}\|\Sigma\|_{\mathrm{HS}}\sqrt{\log\frac{\underline{n}K^3}{S}}.$$

Finally, this inequality combined with equation (A.5) in Lemma A.6 implies the desired bound

$$|\hat{Z} - Z^*|_1 \leqslant \frac{2n}{\underline{n}}S \leqslant C_{12}\, n^2\, K^3/\underline{n}\, \exp(-C_{10}\,\mathsf{SNR}^2) \leqslant C_{12}\, \exp(-C_{13}\,\mathsf{SNR}^2),$$

where the last step is due to the lower bound condition $\mathsf{SNR}^2 \geqslant c_0\, n/\underline{n}$. ■

Proof of Corollary 3.4. For easy presentation, we consider the equal-size clusters case where $n_1 = \ldots = n_K = \underline{n}$ and $G_k^* = \{(k-1)\underline{n}, (k-1)\underline{n}+1, \ldots, k\underline{n}\}$ for $k = 1, \ldots, K$ by reordering the indices. Under this setup, we have

$$Z_{ij}^* = \begin{cases} 1/\underline{n} & \text{if } i,j \in G_k^* \\ 0 & \text{otherwise} \end{cases}.$$

Take c_1 large enough so that the upper bound in Theorem 3.3 satisfies $C_1 \exp(-C_2 \mathsf{SNR}^2) \leqslant \frac{1}{3\underline{n}}$. We use induction to prove that $\hat{G}_k = G_k^*$ at each step for each $k = 1, \ldots, K$, which also implies $\hat{K} = K$. In fact, at $k = 1$, since $\max_i |\hat{Z}_{1i} - Z_{1i}^*| \leqslant |\hat{Z} - Z^*|_1 \leqslant \frac{1}{3\underline{n}}$, we must have $\hat{Z}_{1i} \in \left[\frac{2}{3\underline{n}}, \frac{4}{3\underline{n}}\right]$ for $i \in G_1^*$ and $\hat{Z}_{1i} \leqslant \frac{1}{3\underline{n}}$ for $i \notin G_1^*$ according to the definition of Z^*. This implies $\hat{G}_1 = G_1^*$ according to the choice of $\hat{G}_1$ in the algorithm. Similarly, assume $\hat{G}_l = G_l^*$ for all $l \leqslant k$, then $[n] \setminus \bigcup_{l=1}^k \hat{G}_l = \{k\underline{n} + 1, k\underline{n} + 2, \ldots, n\}$ and $j_{k+1} = k\underline{n} + 1$ by definition. Then the fact that $\max_i |\hat{Z}_{j_{k+1}i} - Z^*_{j_{k+1}i}| \leqslant |\hat{Z} - Z^*| \leqslant \frac{1}{3\underline{n}}$ and the definition of Z^* imply $\hat{Z}_{j_{k+1}i} \in \left[\frac{2}{3\underline{n}}, \frac{4}{3\underline{n}}\right]$ for $i \in G_{k+1}^*$ and $\hat{Z}_{1i} \leqslant \frac{1}{3\underline{n}}$ for $i \notin G_{k+1}^*$. Consequently, we must have $\hat{G}_{k+1} = G_{k+1}^*$ according to the choice of $\hat{G}_{k+1}$ in the algorithm. This completes the proof by induction. ■

Appendix A: Auxiliary results

In this section, we collect and prove all auxiliary results in the paper.

A.1. Feature maps in reproducing kernel Hilbert spaces

In this subsection, we provide a concrete construction of the feature map in kernel clustering. To this end, we invoke the theory of reproducing kernel Hilbert space (RKHS). For a detailed survey of linear operators on Hilbert spaces with statistical applications, we refer to the text [12] as an excellent monograph.

Let the bivariate function $\rho : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a symmetric and positive definite kernel; namely, $\sum_{i,j=1}^m c_i c_j \rho(x_i, x_j) \geqslant 0$ for all $m \geqslant 1, x_1, \ldots, x_m \in \mathbb{X}$, and $c_1, \ldots, c_m \in \mathbb{R}$. By the Moore-Aronszajn Theorem (cf. Theorem 2.7.4 in [12]), there exists a unique Hilbert space $\mathbb{H} := \mathbb{H}(\rho)$ of real-valued functions on $\mathbb{X}$ with ρ as its reproducing kernel, i.e.,

(i) for every $x \in \mathbb{X}$, $\rho(\cdot, x) \in \mathbb{H}$;
(ii) for every $f \in \mathbb{H}$ and $x \in \mathbb{X}$, $f(x) = \langle f, \rho(\cdot, x)\rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product of $\mathbb{H}$.

Property (i) defines a feature map $\phi : \mathbb{X} \to \mathbb{H}$ via $x \mapsto \rho(\cdot, x)$, which is known in literature as the RKHS map [3]. Property (ii) shows that ρ satisfies the reproducing kernel property for all functions in the Hilbert space $\mathbb{H}$. Thus $\mathbb{H}$ is the RKHS associated with ρ. It is immediate from these two properties that

$$\rho(x, y) = \langle \rho(\cdot, y), \rho(\cdot, x)\rangle = \langle \phi(x), \phi(y)\rangle \quad \forall x, y \in \mathbb{X}.$$

Then the similarity matrix A is chosen $a_{ij} = \rho(X_i, X_j) = \langle \phi(X_i), \phi(X_j)\rangle$. Statistical properties of the SDP solution $\hat{Z}$ for (3.3) rely on the distribution of the feature vectors $\phi(X_i)$ in $\mathbb{H}$, which is a special case of Theorem 3.3.

A.2. Auxiliary proofs and lemmas

In this subsection, we provide additional proofs of the technical results used in the paper.

Proof of Lemma 2.4. Without loss of generality, we may assume $\mu = 0$. Suppose that $Z \sim N(0, \Gamma)$. Then $\|Z\|_{\psi_2,\Gamma} = 1$ is obvious from Definition 2.2 and 2.3. Let $M(t) = \mathbb{E}[e^{t\langle z,Z\rangle}], t \in \mathbb{R}$, be the moment generating function of $\langle z, Z\rangle$. Then Taylor's expansion yields that

$$\left.\frac{d^2M(t)}{dt^2}\right|_{t=0} = \mathbb{E}\langle z, Z\rangle^2 = \mathbb{E}\langle z, \langle z, Z\rangle Z\rangle = \mathbb{E}\langle z, (Z \otimes Z)z\rangle = \langle z, \mathbb{E}(Z \otimes Z)z\rangle = \langle z, \Sigma z\rangle.$$

On the other hand, since $Z \sim N(0, \Gamma)$, we have

$$\frac{d^2M(t)}{dt^2} = (1 + t^2)\langle \Gamma z, z\rangle e^{t^2\langle \Gamma z,z\rangle/2}.$$

Thus it follows that

$$\langle (\Sigma - \Gamma)z, z\rangle = 0 \quad \text{for all } z \in \mathbb{H},$$

which implies that $\Sigma = \Gamma$. Suppose that $Z \sim$ sub-gaussian(Γ). By Markov's inequality and Definition 2.2, we have

$$\mathbb{P}(\langle z, Z\rangle \geqslant t) \leqslant \inf_{\lambda>0} e^{-\lambda t}\,\mathbb{E}[e^{\lambda\langle z,Z\rangle}] \leqslant \inf_{\lambda>0} e^{-\lambda t + \frac{\alpha^2\lambda^2}{2}\langle \Gamma z,z\rangle} = e^{-\frac{t^2}{2\alpha^2\langle \Gamma z,z\rangle}},$$

where $\alpha^2 = \|Z\|^2_{\psi_2,\Gamma}$. Then,

$$\langle \Sigma z, z\rangle = \mathbb{E}\langle z, Z\rangle^2 = \int_0^\infty \mathbb{P}(|\langle z, Z\rangle| \geqslant \sqrt{t})dt \leqslant 2\int_0^\infty e^{-\frac{t}{2\alpha^2\langle \Gamma z,z\rangle}}\,dt = 4\alpha^2\langle \Gamma z, z\rangle.$$

Thus it is immediate that $\langle (4\alpha^2\Gamma - \Sigma)z, z\rangle \geqslant 0$ for all $z \in \mathbb{H}$, i.e., $\Sigma \preceq 4\|Z\|^2_{\psi_2,\Gamma}\Gamma$. ■

***Lemma* A.1** (Moment generating function bound for squared norm of a sub-gaussian random variable in $\mathbb{R}^n$)**.** Let Γ be an $n \times n$ positive semidefinite matrix and X be a random variable in $\mathbb{R}^n$ such that $\mathbb{E}[X] = 0$ and $\mathbb{E}[e^{z^TX}] \leqslant e^{z^T\Gamma z/2}$ for all $z \in \mathbb{R}^n$. Let $Z \sim N(0, \Gamma)$. Then,

$$\mathbb{E}\left[e^{\frac{t\|X\|_2^2}{2}}\right] \leqslant \mathbb{E}\left[e^{\frac{t\|Z\|_2^2}{2}}\right] \quad \forall\, 0 \leqslant t < \|\Gamma\|_{\mathrm{op}}^{-1}.$$

Proof of Lemma A.1. The case for $t = 0$ is obvious. Without loss of generality, we may assume Γ is (strictly) positive definite since otherwise we can consider $\Gamma + \delta I_n$ for $\delta > 0$ and then let $\delta \to 0$. Consider $t \in (0, \|\Gamma\|_{\mathrm{op}}^{-1})$. Denote the determinant of Γ as $|\Gamma|$.

Observe that

$$\begin{aligned}
A :=& \frac{1}{(2\pi)^{n/2}|\Gamma|^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{\|z\|_2^2}{2t}} \mathbb{E}[e^{z^T X}] dz \\
=_{(1)}& \mathbb{E}\left[\frac{1}{(2\pi)^{n/2}|\Gamma|^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{\|z-tX\|_2^2}{2t}} dz \, e^{\frac{t\|X\|_2^2}{2}} \right] \\
=_{(2)}& \mathbb{E}\left[e^{\frac{t\|X\|_2^2}{2}} \right] \frac{1}{(2\pi)^{n/2}|\Gamma|^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{\|z\|_2^2}{2t}} dz \\
=_{(3)}& \mathbb{E}\left[e^{\frac{t\|X\|_2^2}{2}} \right] \frac{1}{|t^{-1}\Gamma|^{1/2}},
\end{aligned}$$

where (1) follows from Fubini's theorem, (2) from the translational invariance of the Gaussian density integral, and (3) from that the integration of the standard Gaussian distribution $N(0, I_n)$ equals to one. Thus we get

$$\mathbb{E}\left[e^{\frac{t\|X\|_2^2}{2}} \right] = |t^{-1}\Gamma|^{1/2} A.$$

Since $\mathbb{E}[e^{z^T X}] \leqslant e^{z^T \Gamma z/2}$ for all $z \in \mathbb{R}^n$, we have for $t \in (0, \|\Gamma\|_{\mathrm{op}}^{-1})$,

$$\begin{aligned}
A \leqslant& \frac{1}{(2\pi)^{n/2}|\Gamma|^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{z^T z}{2t}} e^{\frac{z^T \Gamma z}{2}} dz \\
=& \frac{1}{(2\pi)^{n/2}|\Gamma|^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2} z^T (t^{-1} I_n - \Gamma) z} dz \\
=& \frac{1}{|\Gamma|^{1/2}|t^{-1} I_n - \Gamma|^{1/2}} \left[\frac{1}{(2\pi)^{n/2}|(t^{-1} I_n - \Gamma)^{-1}|^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2} z^T (t^{-1} I_n - \Gamma) z} dz \right] \\
=& \frac{1}{|\Gamma|^{1/2}|t^{-1} I_n - \Gamma|^{1/2}}.
\end{aligned}$$

Then we have

$$\mathbb{E}\left[e^{\frac{t\|X\|_2^2}{2}} \right] \leqslant \frac{|t^{-1}\Gamma|^{1/2}}{|\Gamma|^{1/2}|t^{-1} I_n - \Gamma|^{1/2}} = \frac{1}{|I_n - t\Gamma|^{1/2}} \quad \forall\, 0 \leqslant t < \|\Gamma\|_{\mathrm{op}}^{-1}.$$

On the other hand, for $Z \sim N(0, \Gamma)$, similar calculations show that

$$\begin{aligned}
\mathbb{E}\left[e^{\frac{s\|Z\|_2^2}{2}} \right] =& \frac{1}{(2\pi)^{n/2}|\Gamma|^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2} z^T \Gamma^{-1} z} e^{\frac{s}{2} z^T z} dz \\
=& \frac{1}{(2\pi)^{n/2}|\Gamma|^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2} z^T (\Gamma^{-1} - s I_n) z} dz \\
=& \frac{|\Gamma^{-1}(I_n - s\Gamma)|^{-1/2}}{|\Gamma|^{1/2}} = \frac{1}{|I_n - s\Gamma|^{1/2}} \quad \forall\, s < \|\Gamma\|_{\mathrm{op}}^{-1},
\end{aligned}$$

from which Lemma A.1 is immediate. ■

***Lemma* A.2** (Upper bound for squared norm of a sub-gaussian random variable in $\mathbb{R}^n$)**.** In the setting of Lemma A.1, we have

$$\mathbb{E}\left[e^{\frac{t}{2}(\|X\|_2^2-\mathrm{tr}(\Gamma))}\right] \leqslant e^{\frac{t^2}{2}\|\Gamma\|_{\mathrm{HS}}^2} \quad \forall\, 0 \leqslant t < (2\|\Gamma\|_{\mathrm{op}})^{-1}. \tag{A.1}$$

Consequently, we have for any $u > 0$,

$$\mathbb{P}\left(\|X\|_2^2 - \mathrm{tr}(\Gamma) \geqslant u\right) \leqslant \exp\left[-\frac{1}{8}\min\left(\frac{u^2}{\|\Gamma\|_{\mathrm{HS}}^2}, \frac{u}{\|\Gamma\|_{\mathrm{op}}}\right)\right]. \tag{A.2}$$

Proof of Lemma A.2. Let $Z \sim N(0,\Gamma)$. By the calculations in Lemma A.1, we have for all $t < \|\Gamma\|_{\mathrm{op}}^{-1}$,

$$\mathbb{E}\left[e^{\frac{t}{2}(\|Z\|_2^2-\mathrm{tr}(\Gamma))}\right] = \frac{e^{-\frac{t}{2}\mathrm{tr}(\Gamma)}}{|I_n - t\Gamma|^{1/2}} = \prod_{i=1}^n \frac{e^{-t\gamma_i/2}}{\sqrt{1-t\gamma_i}},$$

where $(\gamma_i)_{i=1}^n$ are eigenvalues of Γ. Using the inequality

$$\frac{e^{-t}}{\sqrt{1-2t}} \leqslant e^{2t^2} \quad \forall |t| < 1/4,$$

we have

$$\mathbb{E}\left[e^{\frac{t}{2}(\|Z\|_2^2-\mathrm{tr}(\Gamma))}\right] \leqslant \prod_{i=1}^n e^{\frac{t^2\gamma_i^2}{2}} = e^{\frac{t^2\|\Gamma\|_{\mathrm{HS}}^2}{2}} \quad \forall |t| < (2\|\Gamma\|_{\mathrm{op}})^{-1}.$$

Combining the last inequality with Lemma A.1, we get (A.1). By Markov's inequality, we have for any $u > 0$ and $0 \leqslant t < (2\|\Gamma\|_{\mathrm{op}})^{-1}$,

$$\mathbb{P}\left(\|X\|_2^2 - \mathrm{tr}(\Gamma) \geqslant u\right) \leqslant e^{-\frac{tu}{2}+\frac{t^2}{2}\|\Gamma\|_{\mathrm{HS}}^2}.$$

Choosing $t = t^* := \frac{u}{2\|\Gamma\|_{\mathrm{HS}}^2} \wedge \frac{1}{2\|\Gamma\|_{\mathrm{op}}}$, we get

$$\mathbb{P}\left(\|X\|_2^2 - \mathrm{tr}(\Gamma) \geqslant u\right) \leqslant \exp\left(-\frac{ut^*}{4}\right) = \exp\left[-\frac{1}{8}\min\left(\frac{u^2}{\|\Gamma\|_{\mathrm{HS}}^2}, \frac{u}{\|\Gamma\|_{\mathrm{op}}}\right)\right].$$

■

***Lemma* A.3** (Moment generating function bound for centered squared norm of a sub–gaussian random variable in $\mathbb{H}$)**.** Let $\Gamma \in \mathcal{B}(\mathbb{H})$ be a positive definite trace class operator on $\mathbb{H}$. Let X be a centered sub-gaussian random variable in $\mathbb{H}$ with respect to Γ and $L = \|X\|_{\psi_2}$. Then,

$$\mathbb{E}\left[e^{\frac{t}{2}(\|X\|^2-L^2\|\Gamma\|_{\mathrm{tr}})}\right] \leqslant e^{\frac{t^2L^4}{2}\|\Gamma\|_{\mathrm{HS}}^2} \quad \forall\, 0 \leqslant t < \frac{1}{2L^2\|\Gamma\|_{\mathrm{op}}}.$$

Proof of Lemma A.3. The proof is a standard approximation argument combined with Lemma A.2. Let $(e_k)_{k=1}^{\infty}$ be a CONS of $\mathbb{H}$. By Parseval's identity, $\|X\|^2 = \sum_{k=1}^{\infty}\langle X, e_k\rangle^2$, where convergence of the sum is made in the ℓ^2 sense. Let $K > 0$ be a finite integer. Put $X_K = (\langle X, e_1\rangle, \ldots, \langle X, e_K\rangle)^T$. Then $X_K \sim$ sub-gaussian$(L^2\Gamma_K)$ is a mean-zero random variable in $\mathbb{R}^n$ with $\Gamma_{K,jk} = \langle \Gamma e_j, e_k\rangle$ for $j, k = 1, \ldots, K$. Since $\|\Gamma_K\|_{\mathrm{op}} \leqslant \|\Gamma\|_{\mathrm{op}}$, it follows from Lemma A.2 that

$$\mathbb{E}\left[e^{\frac{t}{2}(\|X_K\|^2 - L^2\|\Gamma_K\|_{\mathrm{tr}})}\right] \leqslant e^{\frac{t^2L^4}{2}\|\Gamma_K\|_{\mathrm{HS}}^2} \quad \forall\, 0 \leqslant t < \frac{1}{L^2\|\Gamma\|_{\mathrm{op}}}.$$

Letting $K \to \infty$, we have $\|X_K\|_2^2 \nearrow \|X\|^2$, $\mathrm{tr}(\Gamma_K) = \|\Gamma_K\|_{\mathrm{tr}} \nearrow \|\Gamma\|_{\mathrm{tr}}$, and $\|\Gamma_K\|_{\mathrm{HS}}^2 \nearrow \|\Gamma\|_{\mathrm{HS}}^2$. Then Lemma A.3 follows from the monotone convergence theorem. ■

***Lemma* A.4** (Squared norm of a sub-gaussian random variable in $\mathbb{H}$ is sub-exponential)**.** Let $\Gamma \in \mathcal{B}(\mathbb{H})$ be a positive definite trace class operator on $\mathbb{H}$ and X be a centered sub-gaussian(Γ) random variable in $\mathbb{H}$. Then there exists a universal constant $C > 0$ such that

$$\left\|\|X\|^2\right\|_{\psi_1} \leqslant C\|X\|_{\psi_2}^2\|\Gamma\|_{\mathrm{tr}}.$$

Thus $\|X\|^2$ is a sub-exponential random variable in $\mathbb{R}$.

Proof of Lemma A.4. Let $(e_k)_{k=1}^{\infty}$ be a CONS of $\mathbb{H}$. By Parseval's identity, $\|X\|^2 = \sum_{k=1}^{\infty}\langle X, e_k\rangle^2$. Since $\|\cdot\|_{\psi_1}$ for real-valued random variables is a norm, we have by triangle inequality that

$$\left\|\|X\|^2\right\|_{\psi_1} \leqslant \sum_{k=1}^{\infty}\left\|\langle X, e_k\rangle^2\right\|_{\psi_1} = \sum_{k=1}^{\infty}\|\langle X, e_k\rangle\|_{\psi_2}^2,$$

where the last step follows from Lemma 2.7.6 in [28]. Since $X \sim$ sub-gaussian(Γ) with mean zero, we have for any $\lambda > 0$,

$$\mathbb{E}\left[e^{\lambda\langle X, e_k\rangle}\right] \leqslant e^{\frac{\lambda^2}{2}\|X\|_{\psi_2}^2\langle \Gamma e_k, e_k\rangle},$$

which implies that there exists a universal constant $C > 0$ such that

$$\|\langle X, e_k\rangle\|_{\psi_2} \leqslant C\|X\|_{\psi_2}\sqrt{\langle \Gamma e_k, e_k\rangle}.$$

Then,

$$\left\|\|X\|^2\right\|_{\psi_1} \leqslant \sum_{k=1}^{\infty} C^2\|X\|_{\psi_2}^2\langle \Gamma e_k, e_k\rangle = C^2\|X\|_{\psi_2}^2\|\Gamma\|_{\mathrm{tr}}.$$

■

Let r be a non-negative integer and $0 \leqslant f < 1$. For $s = r + f \geqslant 0$, we define the (generalized) sum

$$\sum_{i=1}^{s} a_i := \sum_{i=1}^{r} a_i + f a_{r+1}.$$

***Lemma* A.5** (Monotone rearrangement)**.** For any $a_1, \ldots, a_n \in \mathbb{R}$ and $b_1, \ldots, b_n \in [0,1]$, we have

$$\sum_{i=1}^{n} a_i b_i \leqslant \sum_{i=1}^{s} a_{(i)},$$

where $a_{(1)} \geqslant \cdots \geqslant a_{(n)}$ and $s = \sum_{i=1}^{n} b_i$.

By definition, we clearly have $\sum_{i=1}^{s} a_i \leqslant \max\{\sum_{i=1}^{r} a_i, \sum_{i=1}^{r+1} a_i\}$, and for $0 \leqslant s \leqslant 1$, $\sum_{i=1}^{s} = sa_1 \leqslant sa_{(1)}$. Moreover, Lemma A.5 is tighter than the classical inequality $\sum_{i=1}^{n} a_i b_i \leqslant |a|_\infty |b|_1$ because $a_{(i)} \leqslant |a|_\infty$. Using the order statistics structure, we are able to obtain the exponential decay of error result in the K-means SDP clustering problem in Section 3.3.

Proof of Lemma A.5. Write $s = r + f$, where r is a non-negative integer and $f \in [0,1)$. Let X be a random variable taking values in $\{a_1, \ldots, a_n\}$ with the probability mass function $\mathbb{P}(X = a_i) = b_i/s$. Let Y be another random variable taking values in $\{a_{(1)}, \ldots, a_{(n)}\} = \{a_1, \ldots, a_n\}$ with the probability mass function $\mathbb{P}(Y = a_{(j)}) = 1/s$ for $1 \leqslant j \leqslant r$, $\mathbb{P}(Y = a_{(r+1)}) = f/s$, and $\mathbb{P}(Y = a_{(j)}) = 0$ for $r+2 \leqslant j \leqslant n$. Since $b_i \in [0,1]$, we can always shift a non-negative proportion of mass from X to Y. Thus we have $\mathbb{E}[X] \leqslant \mathbb{E}[Y]$ and the lemma follows. ■

***Lemma* A.6.** Let Z^* be defined in (3.4). Then for any $Z \in \mathscr{C}$ defined in (3.3), we have

$$|Z^* - Z^* Z Z^*|_1 = |Z^* - Z^* Z|_1 = 2 \sum_{1 \leqslant k \neq m \leqslant K} |Z_{G_k^* G_m^*}|_1, \tag{A.3}$$

$$\|(I - Z^*) Z (I - Z^*)\|_{\mathrm{tr}} \leqslant \frac{|Z^* - Z^* Z|_1}{2\underline{n}}, \tag{A.4}$$

$$|Z^* - Z^* Z|_1 \leqslant |Z^* - Z|_1 \leqslant \frac{2n}{\underline{n}} |Z^* - Z^* Z|_1. \tag{A.5}$$

Proof of Lemma A.6. See Lemma 1 in [10]. ■

Acknowledgements

The authors would like to thank an anonymous referee and an Associate Editor for their many careful comments that improved the quality of this paper. X. Chen's research was supported in part by NSF DMS-1404891, NSF CAREER Award DMS-1752614, UIUC Research Board Awards (RB17092, RB18099), and a Simons Fellowship. Y. Yang's research was supported in part by NSF DMS-1810831.

References

[1] Radosław Adamczak. A note on the hanson-wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20(72):1–13, 2015.

[2] Rita Giuliano Antonini. Subgaussian random variables in Hilbert spaces. *Rend. Sem. Mat. Univ. Padova*, 98:89–99, 1997.

[3] N Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[4] Franck Barthe and Emanuel Milman. Transference principles for log-sobolev and spectral-gap with applications to conservative spin systems. *Communications in Mathematical Physics*, 323(2):575–625, 2013.

[5] Florentina Bunea, Christophe Giraud, Martin Royer, and Nicolas Verzelen. PECOK: a convex optimization approach to variable clustering. *arXiv:1606.05100*, 2016.

[6] Xiaohui Chen. Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *Annals of Statistics*, 46(2):642–678, 2018.

[7] Yingjie Fei and Yudong Chen. Hidden integrality of sdp relaxation for sub-gaussian mixture models. *arXiv:1803.06510*, 2018.

[8] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[9] Maurizio Filippone, Francesco Camastra, Franscesco Masulli, and Stefano Rovetta. A survey of kernel and spectral methods for clustering. *Pattern Recognition*, 41:176–190, 2008.

[10] Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed $k$means. *arXiv:1807.07547*, 2018.

[11] D.L. Hanson and E.T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42:1079–1083, 1971.

[12] Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 2015.

[13] Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012.

[14] Francesca Ieva, Anna M Paganoni, Davide Pigoli, and Valeria Vitelli. Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62:401–418, 2013.

[15] B. Laurent and Massart P. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[16] Xiaodong Li, Yang Li, Shuyang Ling, Thomas Stohmer, and Ke Wei. When do birds of a feather flock together? k-means, proximity, and conic programming. *arXiv:1710.06008*, 2017.

[17] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[18] Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM J. OPTIM*, 18(1):186–205, 2007.

[19] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[20] Holger Rauhut and Simon Foucart. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. BirkHäuser, 2013.

[21] Martin Royer. Adaptive clustering through semidefinite programming. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1795–1803. Curran Associates, Inc., 2017.

[22] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

[23] Bernhard Schölkopf and Alexander Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 1 edition, 2001.

[24] Le Song, Alex Smola, Arthur Gretton, and Karsten Borgwardt. A dependence maximization view of clustering. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[25] Thaddeus Tarpey and Kimberly KJ Kinateder. Clustering functional data. *Journal of classification*, 20:093–114, 2003.

[26] Aad W van der Vaart, J Harry van Zanten, et al. Reproducing kernel hilbert spaces of gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 200–222. Institute of Mathematical Statistics, 2008.

[27] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, pages 210–268. Compressed sensing. Cambridge University Press, 2012.

[28] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[29] Van Vu and Ke Wang. Random weighted projections, random quadratic forms and random eigenvectors. *Random Structures & Algorithms*, 47(4):792–821, 2015.

[30] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[31] E.T. Wright. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *The Annals of Probability*, 1:1068–1070, 1973.