UNIFORM SPECTRAL BOUNDS UNDER DATA-DEPENDENT SAMPLE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Concentration inequalities for sample covariance matrices are fundamental tools in high-dimensional probability. However, data-dependent sample selection prevents direct application of classical covariance concentration bounds. In this paper, we study the extreme eigenvalues of covariance matrices formed from arbitrary, possibly data-dependent subsets of random vectors. For i.i.d. Gaussian vectors, we establish simultaneous lower and upper bounds over all subsets satisfying prescribed cardinality constraints. The bounds depend explicitly on the retained proportion through central and tail-trimmed second moments, with finite-sample slack and complementary constructions establishing limiting tightness. We extend the results to sub-Gaussian vectors under additional distributional assumptions and obtain lower bounds for geometrically strong-mixing Gaussian sequences. For K-subspace clustering under a low-rank Gaussian mixture model, we show that the clustering error of every global minimizer decreases polynomially with the signal-to-noise ratio under explicit sample-size conditions. For Gaussian robust regression, our analysis strengthens the uniformity of covariance and gradient-moment guarantees while improving the sufficient sample bound from $O(\varepsilon^{-4})$ to $O(\varepsilon^{-1})$ at constant confidence.

1 INTRODUCTION

Concentration inequalities for estimating the spectrum of sample covariance matrices of independent random vectors have been extensively studied in the literature and have found wide applications in statistics, signal processing, and machine learning [13, 27, 28]. Typically, these results assume that the random vectors in concentration inequalities are independent and identically distributed (i.i.d.). A classic example of these concentration results is the spectral bound of the sample covariance matrix of i.i.d. Gaussian random vectors [28, Theorem 4.7.1], a consequence of which is as follows[1].

Fact 1. *Suppose that $\{\boldsymbol{a}_i\}_{i=1}^M \subseteq \mathbb{R}^m$ are i.i.d. standard Gaussian random vectors. It holds for $\eta > 0$ with probability at least $1 - 2\exp\left(-2\eta^2\right)$ that*

$$\left\| \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{a}_i \boldsymbol{a}_i^T - \boldsymbol{I}_m \right\| \leq \frac{9(\sqrt{m} + \eta)}{\sqrt{M}} \,. \tag{1}$$

However, the independence assumption may be violated in applications. A typical example is subspace clustering. It seeks low-dimensional subspaces and their associated clusters in high-dimensional data [25, 29]. Under a low-rank Gaussian mixture model (Definition 2), bounding clustering and subspace estimation errors requires controlling the smallest covariance eigenvalue within an estimated cluster. The cluster assignments and subspace bases in Problem (21) both depend on the observations. A useful concentration event must therefore cover all admissible assignments simultaneously, including those chosen by optimization.

Uniform control over selected subsets already appears in robust statistics [3, 16], erasure-robust frames, robust Johnson–Lindenstrauss embeddings, and the strong restricted isometry property [6, 31, 9]. These works address selection-induced dependence and concentration for dependent stochastic processes provides a related line of research [20, 26, 21]. Our focus is the full Gaussian

[1] A proof of Fact 1 can be found in [23].

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

trimmed-moment curves, their finite-sample approximations and complementary constructions, and their use in clustering and regression. A quantitative comparison with prior erasure-robustness bounds and further related work appear in Appendix A.

Contributions. To address the challenges caused by data-dependent sample selection, we study concentration inequalities for sample covariance matrices formed from arbitrary subsets of random vectors. Our contributions are threefold:

- First, we establish concentration inequalities for the minimal and maximal eigenvalues of sample covariance matrices formed from arbitrary subsets of i.i.d. Gaussian random vectors (see Theorems 1 and 2). Our bounds make the dependence on the retained proportion explicit through Gaussian trimmed second moments. We also prove complementary bounds establishing their limiting tightness (see Corollary 1). Moreover, we extend these results to sub-Gaussian random vectors under additional nondegeneracy and uniform tail assumptions (see Theorem 3).
- Second, we extend our concentration framework beyond the independent setting. In particular, we establish a uniform lower spectral bound for arbitrary subsets of stationary, jointly Gaussian observations with geometric strong mixing (see Theorem 4). This extension allows us to handle both dependence among the observations and dependence of the subset selection on the data.
- Third, we apply the established concentration inequalities to analyze the global optimality of the K-subspace clustering problem under a low-rank Gaussian mixture model. Using the concentration result in Theorem 1, we show that the clustering error of every global minimizer decreases polynomially as the signal-to-noise ratio increases, under explicit sample-size and separation conditions (see Theorem 5).
- Finally, we establish uniform covariance and gradient-moment bounds for Gaussian robust regression (see Theorem 6). Compared with the filtering conditions in [11], our bounds hold simultaneously for all parameters and admissible data-dependent weights on the same uncorrupted sample, without parameter-dependent additional trimming. Under Gaussian design and independent Gaussian noise, we improve the sufficient sample-size dependence on the contamination fraction.

Notation. We use $\|\cdot\|$ for the Euclidean or spectral norm, $\|\cdot\|_F$ for the Frobenius norm, $\sigma_{\min}, \sigma_{\max}$ for extreme singular values, and $\lambda_{\min}, \lambda_{\max}$ for extreme eigenvalues of symmetric matrices. Vector and matrix entries are denoted by a_i and a_{ij}, and $\mathrm{diag}(\boldsymbol{a})$ has diagonal $\boldsymbol{a}$. Write $[n] = \{1, \dots, n\}$, $n_{\min} = \min_k n_k$, and $n_{\max} = \max_k n_k$. Let $\mathcal{O}^{n\times d} = \{\boldsymbol{U} : \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_d\}$ and $\mathcal{O}(d) = \mathcal{O}^{d\times d}$. The standard Gaussian density, distribution, and quantile functions are ϕ, Φ, and Φ^{-1}. We use $a_n = o(b_n)$ when $a_n/b_n \to 0$, and $X_n \xrightarrow{d} X$ for convergence in distribution. Covariance sums are uncentered second moments; their normalization is stated explicitly. Constants $c, C > 0$ may change between occurrences; nonabsolute dependence is specified.

2 Main Results

In this section, we establish concentration inequalities for data-dependent subsets of Gaussian random vectors in Section 2.1, and extend them to sub-Gaussian random vectors in Section 2.2.

2.1 Concentration Inequalities

To describe the effect of selection, let $Z \sim \mathcal{N}(0,1)$ and, for $0 < \kappa < 1$, define the central and tail thresholds

$$q_\kappa := \Phi^{-1}((1+\kappa)/2), \qquad r_\kappa := \Phi^{-1}(1-\kappa/2).$$

Their corresponding trimmed second moments are

$$b(\kappa) := \mathbb{E}[Z^2\mathbb{I}\{|Z| \le q_\kappa\}] = \kappa - 2q_\kappa\phi(q_\kappa), \quad u(\kappa) := \mathbb{E}[Z^2\mathbb{I}\{|Z| \ge r_\kappa\}] = \kappa + 2r_\kappa\phi(r_\kappa). \tag{2}$$

Set $q_0 = r_1 = 0$, $b(0) = u(0) = 0$, and $b(1) = u(1) = 1$ by continuity. These moments are unnormalized; dividing by κ gives conditional second moments. As $\kappa \downarrow 0$,

$$b(\kappa) \sim \frac{\pi}{6}\kappa^3, \quad u(\kappa) \sim 2\kappa\log(1/\kappa). \tag{3}$$

We first present a concentration inequality that bounds the minimal eigenvalue of the sample covariance matrix formed from an arbitrary subset of Gaussian random vectors, even when the subset is chosen in a data-dependent manner.

Theorem 1 (Uniform lower bound). *Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M$ be i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$ vectors, with $M \geq m$. Fix $\kappa_1 \in (0, 1]$, $\eta \in (0, 1/2)$, and define*

$$\gamma = q_{(1-\eta)\kappa_1}, \quad \mu = b((1-\eta)\kappa_1) > 0. \tag{4}$$

There is an absolute $c_1 > 0$ such that, with probability at least

$$1 - 2\left(1 + \frac{22}{\eta\sqrt{\mu}}\right)^m \left(\exp\left(-\eta^2\kappa_1 M/2\right) + \exp\left(-c_1\eta^2\mu^2 M\right)\right) - 2\exp\left(-2M\right), \tag{5}$$

simultaneously for every $\Omega \subseteq [M]$ with $|\Omega| \geq \kappa_1 M$,

$$\lambda_{\min}\left(\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right) \geq (1-\eta)^2\mu M. \tag{6}$$

We defer the proof to Appendix F.1. Now, let us make some remarks on this concentration inequality.

- Even though $\{\boldsymbol{a}_i\}_{i=1}^M \subseteq \mathbb{R}^m$ are i.i.d. standard Gaussian random vectors, we should point out that $\Omega \subseteq [M]$ may depend on $\{\boldsymbol{a}_i\}_{i=1}^M$. For example, we define $\Omega = \left\{i \in [M] : \|\boldsymbol{a}_i\| \geq \|\boldsymbol{a}\|_{(\kappa_1 M)}\right\}$, where $\|\boldsymbol{a}\|_{(k)}$ denotes the k-th largest value among $\{\|\boldsymbol{a}_j\|\}_{j=1}^M$. Consequently, $\boldsymbol{a}_i$ for all $i \in \Omega$ depend on both itself and all the other vectors.
- The single event permits arbitrary data-dependent selection. A positive lower bound requires $\lceil \kappa_1 M \rceil \geq m$ by rank. Moreover, (5) is informative only for sufficiently large samples: for fixed η and failure probability $\rho \in (0, 1)$, it suffices that

  $$M \geq C_\eta \kappa_1^{-6} \left[m \log(C_\eta/\kappa_1) + \log(1/\rho)\right]. \tag{7}$$

  Thus arbitrarily small retained fractions require increasing M; the statement does not certify them at a fixed sample size. This sufficient requirement is not sample-optimal.

Comparison with erasure-robustness bounds. Han and Xu [9] already obtain a lower coefficient of order κ^3 for arbitrary retained subsets. Their full-sparsity specialization has sufficient sample size $M \gtrsim \kappa^{-3}[m + \log(1/\rho)]$ at fixed slack, whereas (7) scales as κ^{-6} up to logarithms. Our contribution is the explicit full trimmed-moment curves and complementary limiting constructions, with the extensions and applications below, rather than improved sample dependence. The exact coefficient and confidence comparison appears in Appendix A.

Weakness of simple union bound. One may wonder whether standard concentration inequalities such as Fact 1, combined with a union bound, are sufficient to establish a bound of the form (5). While this approach is indeed possible, as shown in [16], the resulting lower bound is meaningful only when the retained fraction is close to one. In particular, the union-bound argument imposes a stringent requirement on the parameter κ_1, namely $\kappa_1 > 0.78$, below which the resulting bound becomes vacuous in the asymptotic regime.

Consider the case $m = 1$ and a fixed subset Ω with $|\Omega| = \kappa_1 M$. Define $Z_\Omega := \frac{1}{\sqrt{2\kappa_1 M}} \sum_{i\in\Omega}(a_i^2 - 1)$. By the central limit theorem and Mills' ratio, for any fixed $\eta \in (0, 1/2]$ and all sufficiently large M,

$$\mathbb{P}\big(|Z_\Omega| \leq \eta\big) \leq 1 - c\,\frac{\eta}{1+\eta^2}\exp\left(-\frac{\eta^2}{2}\right),$$

where $c > 0$ is an absolute constant. To make the fixed-subset covariance lower bound non-vacuous, one must take $\eta \leq \sqrt{\kappa_1 M/2}$. Taking a union bound over all $C_M^{\kappa_1 M}$ subsets and using Stirling's approximation gives a resulting probability bound of at most

$$1 - \frac{c}{M}\exp\left\{M\left[\kappa_1 \log\frac{1}{\kappa_1} + (1-\kappa_1)\log\frac{1}{1-\kappa_1} - \frac{\kappa_1}{4}\right]\right\}.$$

Hence, for this bound to be non-vacuous as $M \to \infty$, it is necessary $\kappa_1 > e^{-1/4} \approx 0.78$. Please refer to Appendix F.5 for more details. In contrast, our bound allows κ_1 to be arbitrarily small, making it much more flexible.

We next derive an upper bound on the maximum eigenvalue of $\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T$, as stated in the following theorem. The proof is deferred to Appendix F.3.

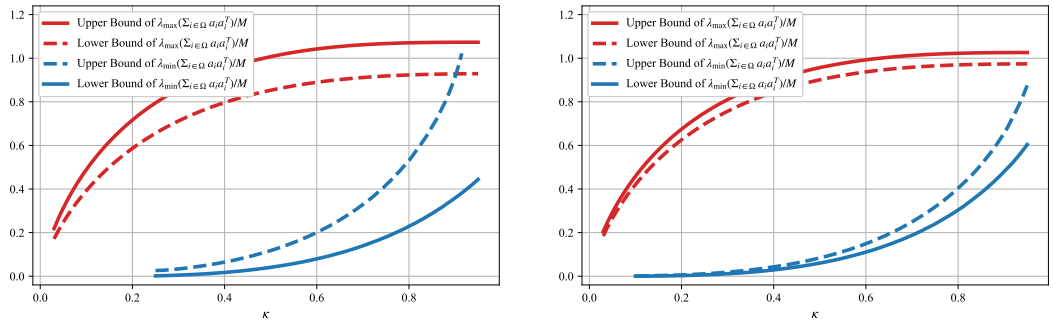


Figure 1: Gaussian subset spectral bounds versus retained fraction κ: lower and complementary upper bounds for $\lambda_{\min}$, and upper and complementary lower bounds for $\lambda_{\max}$. Left: $M/m = 10^4$; right: 10^5. Curves illustrate the coefficients; their nominal 0.999 confidence calibration is not certified because the concentration constants are unspecified.

Theorem 2 (Uniform upper bound)**.** *Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M$ be i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$ vectors. Fix $\kappa_2 \in (0, 1]$ and $\eta \in (0, 1/2)$ with $(1+\eta)\kappa_2 \le 1$, and set $\gamma = r_{(1+\eta)\kappa_2}$ and $\mu = u((1+\eta)\kappa_2)$. There is an absolute $c_2 > 0$ such that, with probability at least*

$$1 - 2\left(1 + \frac{6}{\eta}\right)^m \left(e^{-\eta^2\kappa_2 M/6} + e^{-c_2\eta^2\mu^2 M}\right), \tag{8}$$

simultaneously for every $\Omega \subseteq [M]$ with $|\Omega| \le \kappa_2 M$,

$$\lambda_{\max}\left(\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right) \le (1+\eta)^2 \mu M. \tag{9}$$

For fixed η, failure probability at most ρ follows from $M \ge C_\eta \max\{\kappa_2^{-1}, \mu^{-2}\}[m + \log(1/\rho)]$. The endpoint $(1+\eta)\kappa_2 = 1$ has $\gamma = 0$. In addition, the empty-subset case $\kappa_2 = 0$ is deterministic.

Tightness of lower and upper bounds. A natural question is the tightness of our bounds in Theorem 1 and Theorem 2. To address this concern, we present an upper bound for $\lambda_{\min}\left(\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right)$ and a lower bound for $\lambda_{\max}\left(\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right)$ as follows:

Corollary 1 (Complementary subset constructions)**.** *Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M$ be i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$ vectors and fix $\eta \in (0, 1/2)$.*

(i) For $0 < \kappa_1 < 1/(1+\eta)$, set $\widetilde{\gamma} = q_{(1+\eta)\kappa_1}$ and $\widetilde{\mu} = b((1+\eta)\kappa_1)$. With probability at least

$$1 - 2\exp\left(-\eta^2\kappa_1 M/4\right) - 2\exp\left(-c_1\eta^2\widetilde{\mu}^2 M\right), \tag{10}$$

there exists an $\Omega \subseteq [M]$ with $|\Omega| \ge \kappa_1 M$ such that

$$\lambda_{\min}\left(\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right) \le (1+\eta)\widetilde{\mu} M. \tag{11}$$

(ii) For $\kappa_2 \in (0, 1]$, set $\widetilde{\gamma} = r_{(1-\eta)\kappa_2}$ and $\widetilde{\mu} = u((1-\eta)\kappa_2)$. With probability at least

$$1 - 2\exp\left(-\eta^2\kappa_2 M/2\right) - 2\exp\left(-c_2\eta^2\widetilde{\mu}^2 M\right), \tag{12}$$

there exists an $\Omega \subseteq [M]$ with $|\Omega| \le \kappa_2 M$ such that

$$\lambda_{\max}\left(\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right) \ge (1-\eta)\widetilde{\mu} M, \tag{13}$$

where the constants c_1, c_2 may be chosen as in Theorems 1 and 2.

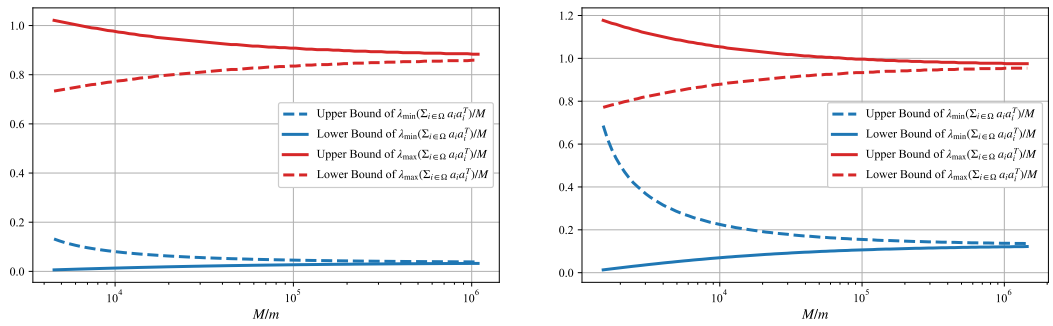


Figure 2: The same spectral bound curves as in Figure 1, versus M/m. Left: $\kappa = 0.4$; right: $\kappa = 0.6$. These coefficient illustrations do not certify the nominal 0.999 confidence level.

In part (i), $\widetilde{\gamma} = q_{(1+\eta)\kappa_1}$ and $\widetilde{\mu} = b((1+\eta)\kappa_1)$; the strict parameter restriction ensures a finite central threshold. In part (ii), $\widetilde{\gamma} = r_{(1-\eta)\kappa_2}$ and $\widetilde{\mu} = u((1-\eta)\kappa_2)$, so no restriction $(1+\eta)\kappa_2 \le 1$ is needed for this converse. If the lower bound, upper bound, and both complementary constructions are required together, their displayed failure probabilities must be added. For a fixed retained fraction $\kappa \in (0,1)$ and a regime in which these failure probabilities vanish as the slack tends to zero, the two sides approach $b(\kappa)M$ and $u(\kappa)M$, respectively. This specifies the limiting sense in which the bounds are tight. The proof is deferred to Appendix F.4.

To further illustrate the tightness of the bounds established in Theorem 1 and Theorem 2, we plot the corresponding lower and upper bounds of $\lambda_{\min}\left(\sum_{i\in\Omega} \boldsymbol{a}_i\boldsymbol{a}_i^T\right)$ given in (6) and (11), as well as those of $\lambda_{\max}\left(\sum_{i\in\Omega} \boldsymbol{a}_i\boldsymbol{a}_i^T\right)$ given in (9) and (13), in Figure 1 and 2. Specifically, we first fix the ratio $M/m = 10^4, 10^5$ and examine how the bounds vary with respect to κ in Figure 1, and then fix $\kappa = 0.4, 0.6$ to investigate their dependence on M/m in Figure 2. These figures are retained to illustrate the dependence of the displayed coefficients on κ and M/m. Their original nominal probability calibration should not be interpreted as a verified 0.999 confidence guarantee, because the numerical constants in the concentration inequalities have not been fixed. The plots compare bound curves rather than independently establishing their accuracy. The tightness claim follows from the complementary subset constructions and the limiting argument above.

2.2 Extension to Sub-Gaussian Random Vectors

In this subsection, we extend the concentration inequalities in Section 2.1 from Gaussian random vectors to sub-Gaussian random vectors [28, Definition 2.5.6]. The proof is deferred to Appendix F.6.

Theorem 3. *Let $\{\boldsymbol{a}_i\}_{i=1}^M \subseteq \mathbb{R}^m$ be i.i.d., centered, isotropic vectors, with $M \ge m$ and $\sup_{\boldsymbol{x}\in\mathbb{S}^{m-1}} \|\boldsymbol{a}_i^T\boldsymbol{x}\|_{\psi_2} \le K_{\mathrm{sg}}$, where $K_{\mathrm{sg}} \ge 1$. For $z \ge 0$, define*

$$\Psi_1(z) := \sup_{\boldsymbol{x}\in\mathbb{S}^{m-1}} \mathbb{P}(|\boldsymbol{a}_i^T\boldsymbol{x}| \le z), \tag{14}$$

$$\Psi_2(z) := \inf_{\boldsymbol{x}\in\mathbb{S}^{m-1}} \mathbb{P}(|\boldsymbol{a}_i^T\boldsymbol{x}| \ge z). \tag{15}$$

Fix $\eta \in (0, 1/2)$. Constants $c_3 \ge 1$ and $c_4 > 0$ below depend only on K_{sg}.

(i) *Let $\kappa_3 \in (0,1]$. Suppose $\gamma > 0$ satisfies $\Psi_1(\gamma) \le (1-\eta)\kappa_3$ and*

$$\mu := \inf_{\boldsymbol{x}\in\mathbb{S}^{m-1}} \mathbb{E}\left[(\boldsymbol{a}_i^T\boldsymbol{x})^2 \mathbf{1}\{|\boldsymbol{a}_i^T\boldsymbol{x}| \le \gamma\}\right] > 0. \tag{16}$$

With probability at least

$$1 - 2\left(\frac{16c_3}{\eta\sqrt{\mu}} + 1\right)^m \left(\exp\left(\frac{-\eta^2\kappa_3 M}{2}\right) + \exp\left(\frac{-\eta^2\mu^2 M}{(2\gamma^4)}\right)\right) - 2\exp(-M),$$

simultaneously for all $\Omega \subseteq [M]$ with $|\Omega| \ge \kappa_3 M$,

$$\lambda_{\min}\left(\sum_{i\in\Omega} \boldsymbol{a}_i\boldsymbol{a}_i^T\right) \ge (1-\eta)^2\mu M. \tag{17}$$

(ii) *Let* $\kappa_4 \in (0, (1+\eta)^{-1}]$. *Suppose* $\gamma \geq 0$ *satisfies* $\Psi_2(\gamma) \geq (1+\eta)\kappa_4$ *and define*

$$\mu := \sup_{\boldsymbol{x} \in \mathbb{S}^{m-1}} \mathbb{E}\left[(\boldsymbol{a}_i^T \boldsymbol{x})^2 \mathbf{1}\{|\boldsymbol{a}_i^T \boldsymbol{x}| \geq \gamma\}\right]. \tag{18}$$

With probability at least

$$1 - 2\left(\frac{6}{\eta} + 1\right)^m \left(\exp\left(\frac{-\eta^2 \kappa_4 M}{6}\right) + \exp\left(-c_4 \eta^2 \mu^2 M\right)\right),$$

simultaneously for all $\Omega \subseteq [M]$ *with* $|\Omega| \leq \kappa_4 M$,

$$\lambda_{\max}\left(\sum_{i \in \Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right) \leq (1+\eta)^2 \mu M. \tag{19}$$

The following example shows that the conditions of Theorem 3 hold for a non-Gaussian isotropic sub-Gaussian distribution whose coordinates need not be independent.

Example 1 (A non-Gaussian isotropic example)**.** *Let* $\boldsymbol{a} = R\boldsymbol{g}$, *where* $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$ *and* R *is independent, nonconstant, and satisfies* $0 < r_- \leq R \leq r_+ < \infty$ *and* $\mathbb{E}[R^2] = 1$. *Then* $\boldsymbol{a}$ *is centered, isotropic, and non-Gaussian, with directional sub-Gaussian norms bounded by* Cr_+ *and possibly dependent coordinates. For example, take* $\mathbb{P}(R = \sqrt{1/2}) = \mathbb{P}(R = \sqrt{3/2}) = 1/2$. *Every unit-direction projection has the distribution of* RZ, *where* $Z \sim \mathcal{N}(0,1)$ *is independent of* R. *Since*

$$\mathbb{P}(|RZ| \leq \gamma) = \mathbb{E}[2\Phi(\gamma/R) - 1]$$

is continuous and strictly increasing from 0 *to* 1, *with a positive central trimmed second moment for every* $\gamma > 0$, *the thresholds required by both parts of Theorem 3 exist. The upper-bound endpoint uses* $\gamma = 0$.

3 Extension to Weakly Dependent Random Vectors

In this section, we discuss extensions beyond the independent setting. The main observation is that independence is not essential to the geometric component of the proof. Instead, the proof of Theorem 1 only requires three probabilistic ingredients: a global operator-norm bound for the full data matrix (39), and two fixed-direction concentration bounds, (46) and (50). Consequently, the same proof strategy extends to weakly dependent sequences $\{\boldsymbol{a}_i\}_{i=1}^M$, as long as these three ingredients can be established.

To make this extension concrete, we consider geometrically strong-mixing sequences [22] as a representative class of weakly dependent data. This class is broad enough to capture many temporally dependent processes while still admitting Bernstein-type concentration inequalities needed in our proof. Let $\{X_t\}_{t \in \mathbb{Z}_+}$ be a stochastic process and define

$$\mathcal{F}_0^\tau := \sigma(X_t : 0 \leq t \leq \tau), \quad \mathcal{F}_\tau^\infty := \sigma(X_t : t \geq \tau).$$

The mixing coefficients of $\{X_t\}_{t \in \mathbb{Z}_+}$ are defined by

$$\alpha(k) := \sup_{t \in \mathbb{Z}_+} \sup_{A \in \mathcal{F}_0^t,\, B \in \mathcal{F}_{t+k}^\infty} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|, \quad k = 1, 2, \ldots.$$

Definition 1. *A sequence* $\{X_t\}_{t \in \mathbb{Z}}$ *is called* geometrically strong mixing *if its mixing coefficients decay geometrically, i.e., there exists some constants* $C > 0$ *and* $\rho \in (0,1)$ *such that* $\alpha(k) \leq C\rho^k$ *for all* $k \geq 1$.

The following example shows that geometrically strong mixing holds for stationary Gaussian processes with sufficiently fast covariance decay; see Proposition 1 in Appendix G.1 for a rigorous statement and proof.

Example 2. *A standard stationary Gaussian sequence with geometrically decaying covariance is geometrically strong mixing. Specifically, suppose there exist constants* $C > 0$ *and* $\rho \in (0,1)$ *such that, for all* $k, t \in \mathbb{Z}_+$,

$$\boldsymbol{a}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m), \quad \mathrm{Cov}(\boldsymbol{a}_0, \boldsymbol{a}_k) = \mathrm{Cov}(\boldsymbol{a}_t, \boldsymbol{a}_{t+k}), \quad \|\mathrm{Cov}(\boldsymbol{a}_0, \boldsymbol{a}_k)\| \leq C\rho^k.$$

Then $\{\boldsymbol{a}_t\}_{t \in \mathbb{Z}_+}$ *is geometrically strong mixing.*

A standard consequence of geometric strong mixing is that bounded functions of the process satisfy Bernstein-type inequalities with an effective sample size loss of logarithmic order. More precisely, throughout this section, we use the notation $L_M := \log(eM)\log\log(e^e M)$ and $M_{\text{eff}} := M/L_M$. The following theorem gives a dependent version of the lower spectral bound, whose proof is deferred to Appendix G.2.

Theorem 4. *Let $\{\boldsymbol{a}_i\}_{i=1}^M \subset \mathbb{R}^m$ be a standard Gaussian sequence satisfying geometrically strong-mixing property with coefficient $C > 0$ and $\rho \in (0,1)$. With probability at least*

$$1 - \exp(-c_1 M) - 2\left(\frac{6\sqrt{B}}{\eta\sqrt{\mu}} + 1\right)^m \left[\exp\left(-c\eta^2\kappa^2 M_{\text{eff}}\right) + \exp\left(-\frac{c\eta^2\mu^2}{\gamma^4} M_{\text{eff}}\right)\right],$$

for some constants c_1, B depending on C and ρ, and other constants same as Theorem 1, it holds simultaneously for every subset $\Omega \subseteq [M]$ satisfying $|\Omega| \geq \kappa M$ that

$$\lambda_{\min}\left(\textstyle\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right) \geq (1-\eta)^2 \mu M.$$

4 Application to K-Subspace Clustering

In this section, we apply the established concentration inequalities to analyze the global optimality of the K-subspace clustering problem under the low-rank Gaussian mixture model (LRGMM).

4.1 Problem Setup and Formulation

Definition 2 (Low-Rank Gaussian Mixture Model)**.** *Let $C_1^\star, \ldots, C_K^\star$ partition $[N]$ into nonempty sets, where $N \geq K \geq 2$, and let $\boldsymbol{U}_k^\star \in \mathcal{O}^{n\times d}$, with $n \geq d \geq 1$. The LRGMM generates*

$$\boldsymbol{z}_i = \boldsymbol{U}_k^\star \boldsymbol{a}_i + \boldsymbol{e}_i, \qquad \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \quad \boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \delta_k^2 \boldsymbol{I}_n), \quad i \in C_k^\star, \tag{20}$$

where $\delta_k \geq 0$ and all latent and noise vectors are mutually independent.

Given unlabeled samples $\{\boldsymbol{z}_i\}_{i=1}^N$ and the number of clusters K, the K-subspace clustering problem aims to recover both the subspace bases $\{\boldsymbol{U}_k^\star\}_{k=1}^K$ and the cluster assignment $\{C_k^\star\}_{k=1}^K$. We consider the following formulation:

$$\min_{\boldsymbol{U},C} \; F(\boldsymbol{U}, C) := \sum_{k=1}^K \sum_{i\in C_k} \|\boldsymbol{z}_i - \boldsymbol{U}_k \boldsymbol{U}_k^T \boldsymbol{z}_i\|^2 \quad \text{s.t.} \quad \boldsymbol{U}_k^T \boldsymbol{U}_k = \boldsymbol{I}_d, \quad k \in [K], \tag{21}$$

where $\boldsymbol{U} = \{\boldsymbol{U}_k\}_{k=1}^K$ denotes the estimated subspaces and $C = \{C_k\}_{k=1}^K$ denotes a partition of $[N]$.

Remark 1. *Any global minimizer $(\hat{\boldsymbol{U}}, \hat{C})$ is a function of the full dataset $\{\boldsymbol{z}_i\}_{i=1}^N$. Hence the selected subsets $\{\boldsymbol{z}_i : i \in \hat{C}_k\}$ are data-dependent, which creates selection-induced statistical dependence and makes the analysis of global optimality nontrivial.*

To evaluate recovery performance, we define the permutation-invariant clustering and subspace errors as

$$\ell(\hat{C}, C^\star) := 1 - \frac{1}{N}\max_{\pi\in\Pi} \sum_{k=1}^K \left|\hat{C}_{\pi(k)} \cap C_k^\star\right|, \quad d_F(\hat{\boldsymbol{U}}, \boldsymbol{U}^\star) := \min_{\pi\in\Pi} \sum_{k=1}^K \left\|\hat{\boldsymbol{U}}_{\pi(k)}\hat{\boldsymbol{U}}_{\pi(k)}^T - \boldsymbol{U}_k^\star \boldsymbol{U}_k^{\star T}\right\|_F,$$

where Π is the set of all permutations of $[K]$ and $\hat{\boldsymbol{U}} := \{\hat{\boldsymbol{U}}_k\}_{k=1}^K$.

Metrics for subspace clustering. The difficulty of subspace clustering is mainly governed by the separation between subspaces, the noise level, and the cluster-size balance. We measure the affinity between the k-th and l-th true subspaces by

$$\mu_{kl} := \|\boldsymbol{U}_k^{\star T} \boldsymbol{U}_l^\star\|_F / \sqrt{d}, \tag{22}$$

and define $\mu_{\max} := \max_{k\neq l} \mu_{kl}$, $\delta_{\max} := \max_{k\in[K]} \delta_k$. A larger subspace separation and a smaller noise level make clustering easier. Accordingly, we define the signal-to-noise ratio as

$$\text{SNR} := (1 - \mu_{\max}^2)/\delta_{\max}. \tag{23}$$

Finally, to quantify cluster-size imbalance, let $N_k^\star := |C_k^\star|$ and $N_{\min}^\star := \min_{k\in[K]} N_k^\star$. We define

$$\kappa_N := N/(N_{\min}^\star K). \tag{24}$$

4.2 Global Optimality of the Loss Function

Assumption 1. *With $\alpha_N := 2\sqrt{\log N} + 2$, the dimensions satisfy*

$$n \geq d \geq \alpha_N^2. \tag{25}$$

Assumption 1 bounds the Gaussian norm fluctuations of order $\sqrt{\log N}$ relative to $\sqrt{d}$ and $\sqrt{n}$ simultaneously over all observations. On the norm event used in the proof, it gives $\|\boldsymbol{a}_i\| \leq 2\sqrt{d}$ and $\|\boldsymbol{e}_i\| \leq 2\delta_k\sqrt{n}$ for every $i \in C_k^\star$, controlling the signal–noise cross terms in the objective comparison. This is a sufficient condition for our proof, not a necessary condition for clustering. The separation condition $\mu_{\max} < 1$ permits nonorthogonal and intersecting subspaces: $1 - \mu_{kl}^2$ is the average squared sine of their principal angles. Thus SNR balances geometric separation against noise, while κ_N records the additional difficulty caused by small clusters.

Theorem 5. *Consider Definition 2 with $\mu_{\max} < 1$ under Assumption 1. Fix $\tau \in (0, 1/K)$ and $\rho \in (0, 1)$. There is an absolute $C > 0$ such that, if*

$$\mathrm{SNR} \geq C\kappa_N K \tau^{-3}\sqrt{n/d}, \tag{26}$$

$$N_{\min}^\star \geq C\tau^{-6}[d\log(C/\tau) + \log(K/\rho)], \quad N \geq 8/\rho, \tag{27}$$

then, with probability at least $1 - \rho$, every global minimizer $(\hat{\boldsymbol{U}}, \hat{C})$ of (21) satisfies

$$\ell(\hat{C}, C^\star) \leq (K-1)\tau. \tag{28}$$

The proof in Appendix H.5 applies Theorem 1 to intersections of true and estimated clusters, ruling out large overlaps between an estimated cluster and two distinct true clusters. The SNR requirement scales as τ^{-3} because the central trimmed energy satisfies $b((1-\eta)\tau) \asymp \tau^3$ for fixed η, while the sufficient sample size scales as τ^{-6} because this trimmed moment enters the concentration exponent quadratically. Consequently, an admissible choice $\tau \asymp (\kappa_N K\sqrt{n/d}/\mathrm{SNR})^{1/3}$ yields

$$\ell(\hat{C}, C^\star) \lesssim (K-1)\left(\kappa_N\sqrt{n/d}\,\frac{K}{\mathrm{SNR}}\right)^{1/3}, \tag{29}$$

provided that $\tau < 1/K$ and (27) hold. This bound quantifies the SNR dependence of the clustering-error guarantee for every global minimizer. Further discussion and an exploratory experiment appear in Appendix D.

5 Application to Gaussian robust regression

Robust linear regression seeks to estimate $\boldsymbol{\theta}^\star$ when an adversary may inspect a clean sample and replace up to an ε fraction of the covariate–response pairs [4]. For the clean Gaussian model specified below, the population squared loss satisfies

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2}\mathbb{E}[(\boldsymbol{x}^T\boldsymbol{\theta} - y)^2] = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^T\boldsymbol{\Sigma}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) + \frac{1}{2}\sigma^2.$$

Thus $\boldsymbol{\theta}^\star$ is the unique population minimizer. Corruption of either covariates or responses can distort empirical least squares. Filtering methods iteratively downweight suspicious observations to control weighted gradients and design second moments [11]. For the weighted squared loss $\mathcal{L}_{\boldsymbol{q}}(\boldsymbol{\theta}) = \frac{1}{2}\sum_i q_i(\boldsymbol{x}_i^T\boldsymbol{\theta} - y_i)^2$, these quantities are

$$\nabla\mathcal{L}_{\boldsymbol{q}}(\boldsymbol{\theta}) = \sum_i q_i(\boldsymbol{x}_i^T\boldsymbol{\theta} - y_i)\boldsymbol{x}_i, \qquad \nabla^2\mathcal{L}_{\boldsymbol{q}}(\boldsymbol{\theta}) = \sum_i q_i\boldsymbol{x}_i\boldsymbol{x}_i^T.$$

Downweights or filtering selects observations based on the observed data, making it a natural application of our framework for data-dependent sample selection. Its analysis requires control of the remaining clean design covariance, gradient bias, and gradient second moments. Since both the filtering weights and the current parameter depend on the same data, we establish a single event controlling these quantities uniformly over all admissible capped weights and parameters.

Theorem 6 (Uniform Gaussian regression moments)**.** *Let $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ independently, where $\boldsymbol{\Sigma} \succ 0$, and $y_i = \boldsymbol{x}_i^T \boldsymbol{\theta}^\star + \xi_i$, where the $\xi_i \sim \mathcal{N}(0, \sigma^2)$ are independent of each other and the design. For $0 < \varepsilon < 1/16$ and $0 < \rho < 1$, there are absolute $C, c > 0$ such that, if*

$$N \geq C\left[(d + \log(2/\rho))^2 + \frac{d + \log(2/\rho)}{\varepsilon}\right], \tag{30}$$

with probability at least $1 - \rho$, for every $\boldsymbol{q} \in \mathcal{Q}_\varepsilon := \left\{\boldsymbol{q} \geq 0 : \sum_i q_i = 1, \quad q_i \leq \frac{1}{(1-4\varepsilon)N}\right\}$ and $\boldsymbol{\theta} \in \mathbb{R}^d$, writing $\boldsymbol{v} = \boldsymbol{\theta} - \boldsymbol{\theta}^\star$, $\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}}^2 = \boldsymbol{v}^T \boldsymbol{\Sigma} \boldsymbol{v}$, and $\boldsymbol{g}_i(\boldsymbol{\theta}) = \boldsymbol{x}_i(\boldsymbol{x}_i^T \boldsymbol{\theta} - y_i)$, we have

$$\left\| \boldsymbol{\Sigma}^{-1/2}\Big(\sum_i q_i \boldsymbol{x}_i \boldsymbol{x}_i^T\Big)\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I}_d \right\| \leq C\sqrt{\varepsilon}, \tag{31}$$

$$\left\| \boldsymbol{\Sigma}^{-1/2}\Big(\sum_i q_i \boldsymbol{g}_i(\boldsymbol{\theta}) - \boldsymbol{\Sigma}\boldsymbol{v}\Big) \right\| \leq C\sqrt{\varepsilon}\sqrt{\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}}^2 + \sigma^2}, \tag{32}$$

$$\sum_i q_i \boldsymbol{g}_i(\boldsymbol{\theta}) \boldsymbol{g}_i(\boldsymbol{\theta})^T \preceq C(\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}}^2 + \sigma^2)\boldsymbol{\Sigma}, \qquad \sum_i q_i \xi_i^2 \leq C\sigma^2. \tag{33}$$

Equation (31) controls the relative error of the weighted design covariance, equivalently the squared-loss Hessian with the weights held fixed, around $\boldsymbol{\Sigma}$. Equation (32) keeps the weighted gradient close to the population gradient $\boldsymbol{\Sigma}\boldsymbol{v}$. The first inequality in Equation (33) bounds the gradient second moment and hence its centered covariance, while the second inequality in (33) controls the weighted noise energy. All four bounds hold on one event for every admissible $\boldsymbol{q}$ and every $\boldsymbol{\theta}$, allowing data-dependent reweighting and parameter updates. Whitening and a uniform Gaussian fourth-moment bound yield this event; see Appendix I.1.

Connection to filtering. Let G denote the uncorrupted indices, with $|G| \geq (1 - \varepsilon)N$. For weights $0 \leq w_i \leq 1/N$ satisfying the ε-saturation condition of [11], the remaining clean mass obeys $\sum_{i \in G} w_i \geq 1 - 3\varepsilon$. Thus the normalized clean weights $q_i = w_i \mathbf{1}_{\{i \in G\}} / \sum_{j \in G} w_j$ belong to $\mathcal{Q}_\varepsilon$, and the theorem supplies the clean-sample moment bounds needed for filtering. Appendix E gives the reduction.

Improvement over existing filtering conditions. At constant confidence, a sufficient sample size is $O(d^2 + d/\varepsilon)$, compared with $O(d \log d/\varepsilon^4 + d^2 \log d \log(1/\varepsilon))$ in Remark 1 of the supplementary material of [11]. Under the stronger assumptions of Gaussian design and independent Gaussian noise, we reduce the contamination dependence of the dimension-linear term from ε^{-4} to ε^{-1} and remove the displayed logarithmic factors from the quadratic term. Moreover, the same uncorrupted set G works for every parameter and admissible weighting, without the parameter-dependent additional trimming allowed by their Assumption 1.2. These gains concern the sample size and uniformity of deterministic filtering conditions; they do not establish a new estimator or an improved estimation-error guarantee.

6 Conclusions

In this work, we established concentration inequalities for the extreme eigenvalues of sample covariance matrices formed from arbitrary subsets of i.i.d. Gaussian random vectors. Our bounds make the dependence on the retained proportion explicit through central and tail trimmed moments, with complementary bounds establishing their limiting tightness. We further extended the results to sub-Gaussian random vectors under additional distributional assumptions and obtained a lower spectral bound for geometrically strong-mixing Gaussian sequences. For applications, we established recovery guarantees for every global minimizer of the K-subspace clustering problem under explicit sample-size and separation conditions, as well as uniform covariance and gradient-moment bounds for Gaussian robust regression. Improving the sample-size dependence, extending the analysis to heavy-tailed observations, and establishing convergence guarantees for noisy subspace clustering algorithms are interesting directions for future research.

AI USE STATEMENT

In this work, we used generative AI tools to assist in identifying relevant prior work and improving the clarity of the manuscript, including phrasing and transitions in both the main text and the proofs. All AI-assisted revisions were carefully reviewed and edited by the authors to ensure their correctness and consistency with the intended content. We also used generative AI tools to assist with writing code for experimental verification and plotting, and all resulting code was checked by the authors for correctness. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, clean and reformat dataset, support qualitative and thematic data analysis, and generate synthetic data sets is not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

[1] Laura Balzano, Arthur Szlam, Benjamin Recht, and Robert Nowak. K-subspaces with missing data. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 612–615. IEEE, 2012.

[2] Paul S Bradley and Olvi L Mangasarian. K-plane clustering. *Journal of Global optimization*, 16(1):23–32, 2000.

[3] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[4] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2745–2754, 2019.

[5] Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[6] Matthew Fickus and Dustin G Mixon. Numerically erasure-robust frames. *Linear Algebra and its Applications*, 437(6):1394–1407, 2012.

[7] Andrew Gitlin, Biaoshuai Tao, Laura Balzano, and John Lipor. Improving K-subspaces via coherence pursuit. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1575–1588, 2018.

[8] Robert D Gordon. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12 (3):364–366, 1941.

[9] Bin Han and ZhiQiang Xu. Robustness properties of dimensionality reduction with gaussian random matrices. *Science China Mathematics*, 60(10):1753–1778, 2017.

[10] Fang Han and Yicheng Li. Moment bounds for large autocovariance matrices under dependence. *Journal of Theoretical Probability*, 33(3):1445–1492, 2020.

[11] Arun Jambulapati, Jerry Li, Tselil Schramm, and Kevin Tian. Robust regression revisited: Acceleration and improved estimation rates. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[12] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

[13] Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001. ISBN 9780821809339.

[14] Gilad Lerman and Teng Zhang. Robust recovery of multiple subspaces by geometric ℓ_p minimization. *The Annals of Statistics*, 39(5):2686–2715, 2011.

[15] Gen Li and Yuantao Gu. Theory of spectral method for union of subspaces-based random geometry graph. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 6337–6345. PMLR, 2021.

[16] Jerry Li. Lecture 6: Stronger spectral signatures for gaussian datasets. Technical report, Robust Machine Learning, Fall 2019, 2019.

[17] John Lipor, David Hong, Yan Shuo Tan, and Laura Balzano. Subspace clustering using ensembles of K-subspaces. *Information and Inference: A Journal of the IMA*, 10(1):73–107, 2021.

[18] Yuanyuan Liu, LC Jiao, and Fanhua Shang. An efficient matrix factorization based low-rank representation for subspace clustering. *Pattern Recognition*, 46(1):284–292, 2013.

[19] Florence Merlevède, Magda Peligrad, and Emmanuel Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High Dimensional Probability V: The Luminy Volume*, pp. 273–292. Institute of Mathematical Statistics, 2009.

[20] Seongoh Park, Xinlei Wang, and Johan Lim. Estimating high-dimensional covariance and precision matrices under general missing dependence. *Electronic Journal of Statistics*, 15(2): 4868–4915, 2021.

[21] Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32, 2015.

[22] Murray Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the national Academy of Sciences*, 42(1):43–47, 1956.

[23] Anthony Man-Cho So. Handout 1: Spectrum of a Gaussian random matrix, 2022. URL `https://www1.se.cuhk.edu.hk/~manchoso/2122/seem5380/1-opnorm`. Lecture notes for SEEM 5380: Optimization Methods for High-Dimensional Statistics at The Chinese University of Hong Kong.

[24] Mahdi Soltanolkotabi and Emmanuel J Candes. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[25] Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J Candes. Robust subspace clustering. *Annals of Statistics*, 42(2):669–699, 2014.

[26] Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

[27] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

[28] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[29] René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

[30] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005.

[31] Vladislav Voroninski and Zhiqiang Xu. A strong restricted isometry property, with an application to phaseless compressed sensing. *Applied and Computational Harmonic Analysis*, 40(2):386–395, 2016.

[32] Peng Wang, Huikang Liu, Anthony Man-Cho So, and Laura Balzano. Convergence and recovery guarantees of the K-subspaces method for subspace clustering. In *International Conference on Machine Learning*, pp. 22884–22918. PMLR, 2022.

[33] Peng Wang, Huijie Zhang, Zekai Zhang, Siyi Chen, Yi Ma, and Qing Qu. Diffusion models learn low-dimensional distributions via subspace clustering. In *2025 IEEE 10th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 211–215. IEEE, 2025.

[34] Yang Wang. Random matrices and erasure robust frames. *Journal of Fourier Analysis and Applications*, 24(1):1–16, 2018.

[35] Yu-Xiang Wang and Huan Xu. Noisy sparse subspace clustering. In *International Conference on Machine Learning*, pp. 89–97. PMLR, 2013.

[36] Alec S Xu, Can Yaras, Peng Wang, and Qing Qu. Linearly separable features in shallow nonlinear networks: Width scales polynomially with intrinsic data dimension. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

[37] Huijie Zhang, Jinfan Zhou, Yifu Lu, Minzhe Guo, Peng Wang, Liyue Shen, and Qing Qu. The emergence of reproducibility and consistency in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 60558–60590. PMLR, 2024.

Appendices

A Related Work

Numerically erasure-robust frames (NERFs). A related line of work studies NERFs, i.e., frames for which every sufficiently large subframe obtained after arbitrary erasures remains uniformly well-conditioned. Fickus and Mixon [6] introduced this notion and showed that random Gaussian frames can tolerate roughly 15% arbitrary erasures. Diakonikolas et al. [3] and the related lecture note [16] use uniform covariance bounds for large retained subsets in robust estimation. Such large-subset statements should be distinguished from the small-retained-fraction regime studied in erasure robustness. Wang [34] established covariance eigenvalue bounds that hold simultaneously for all admissible subsets, although the dependence of the constants on the retained proportion is relatively implicit. Han and Xu [9] derived correct-order estimates for Gaussian matrices under arbitrary row erasures. In comparison, we provide explicit finite-sample bounds expressed through Gaussian quantiles, together with complementary tightness results, and discuss extensions under additional assumptions to sub-Gaussian and strong-mixing observations, with an application to K-subspace clustering.

SRIP and a quantitative retained-fraction comparison. The SRIP of [31] controls every sufficiently large row subset uniformly over sparse directions and also yields an erasure-robust Johnson–Lindenstrauss lemma. Taking the sparsity equal to the ambient dimension reduces the directional part of SRIP to the spectral bounds considered here. Han and Xu [9] allow arbitrary erasure fractions and already obtain the correct small-retained-fraction orders. In their notation, the erasure fraction is $\beta = 1 - \kappa$. We use the retained fraction κ throughout this paper.

For a concrete comparison, specialize Corollary 1.6 of [9] to full sparsity and normalize the covariance sum by M. Set $h_\kappa = \min\{1/4 + \kappa/2, \kappa\}$. For $0 < a < (\pi/12)\kappa^2 h_\kappa$ and covering slack $s \in (0, 1)$, their lower coefficient is

$$(1 - s)\left(\sqrt{\frac{\pi}{6}\kappa^2 h_\kappa} - \sqrt{2a}\right)^2, \tag{34}$$

with failure probability at most $2(24e/s)^m e^{-aM}$, provided $M \geq 1/\kappa$ and $m \log(24e/s) < aM - \log 2$. For small κ, choosing fixed s and a a sufficiently small constant multiple of κ^3 gives a lower coefficient of order κ^3 with

$$M \gtrsim \kappa^{-3}\big[m + \log(1/\rho)\big]$$

for failure probability ρ. By contrast, our displayed probability in Theorem 1 has the sufficient regime

$$M \gtrsim \kappa^{-6}\big[m \log(C/\kappa) + \log(1/\rho)\big]$$

at fixed slack. Thus our result does not improve this sample dependence. Its distinguishing feature is the explicit full curve $b((1 - \eta)\kappa)$, its complementary limiting construction, and the corresponding upper curve $u((1 + \eta)\kappa)$, together with the conditional extensions and applications. Both prior and present results use a single event valid over arbitrary admissible subsets; the choice of subset may therefore depend on the entire sample. The comparison with a direct union bound in Appendix F.5 is restricted to the particular certificate derived there and is not a limitation of all union-bound methods.

Concentration inequalities under dependence. Classical concentration inequalities, such as those by [27, 28], assume independence and are widely applied in random matrix theory and high-dimensional statistics. Recently, several works have extended these inequalities to settings where random vectors have dependent structures, such as martingale differences [26], Markov dependence [21], and mixing-type dependence [19]. In particular, [20] analyzes an inverse probability weighted estimator for high-dimensional covariance and precision matrices in the presence of general dependent missingness. [10] derives moment bounds for the deviation of large autocovariance matrices from their means under weak dependence.

Subspace clustering. Subspace clustering is a central problem in unsupervised learning, which aims to cluster data points that lie in a union of low-dimensional subspaces [1, 25, 24, 29, 32]. Notably, this problem can be solved by different approaches, including sparse subspace clustering [5, 35], spectral clustering [15, 30], and low-rank matrix factorization [18, 32]. To analyze these subspace clustering methods, particularly the non-convex ones, a union-of-subspaces model has become a standard data model for generating data samples [24, 25, 30, 32].

Among objective-based approaches to subspace clustering, a popular one is the K-subspace method [2, 17], which is a generalization of the K-means method to subspace clustering. Recently, its convergence behavior has been extensively studied under a mixture of low-rank Gaussians (see Definition 2). This model is a probabilistic instance of the union-of-subspaces model with Gaussian latent variables; see, e.g., [17, 32]. Beyond subspace clustering, mixtures of low-rank Gaussians have also served as a tractable model for studying other unsupervised learning problems, including deep feature learning [36] and diffusion models [33, 37]. Nevertheless, a complete characterization of the global optimality of the objective function (see Problem (21)) remains an open problem. The main challenge is that the optimal cluster assignments are data-dependent, so standard covariance concentration bounds for fixed subsets cannot be applied directly.

It is useful to distinguish recovery by an objective from convergence of an algorithm. Lerman and Zhang [14] study geometric ℓ_p minimization, proving recovery under their outlier model for $0 < p \leq 1$ and exhibiting failures for $p > 1$; these conclusions do not directly settle finite-sample recovery of the noisy squared-loss objective considered here. Wang et al. [32] establish rapid convergence of a suitably initialized algorithm in a noiseless model. The analysis of EKSS by Lipor et al. [17] includes guarantees for the EKSS-0 construction and noisy no-false-connections conditions. Gitlin et al. [7] establish worst-case approximation hardness and propose robust subspace updates. Our results concern statistical properties of global minimizers under the Gaussian mixture model, without removing computational hardness or asserting that noiseless convergence transfers to positive noise.

Robust regression. Jambulapati et al. [11] formulate deterministic covariance and gradient conditions that support filtering methods for robust regression. Proposition 2 and the sub-Gaussian specialization in Remark 1 of their supplementary material provide sample-size bounds for verifying those conditions. Our application specializes to Gaussian design with independent Gaussian noise and verifies uniform covariance, gradient-mean, and gradient-second-moment conditions with $O(d^2 + d/\varepsilon)$ samples at constant confidence; see Theorem 6. This is a concentration result that can supply deterministic premises for filtering analyses, rather than a new regression algorithm or a guarantee for every sub-Gaussian design.

B Additional Gaussian Results and Illustrations

B.1 Coefficient illustrations

Example 3. *The ratio M/m, the retained fraction κ_1, and the slack η all enter the confidence bound in Theorem 1. To illustrate the coefficient $\mu = b((1-\eta)\kappa_1)$, consider the following parameter values from the original numerical illustrations:*

- *For $M = 10^3 m$, $\kappa_1 = 0.7$, and $\eta = 0.4$, one obtains $\sqrt{\mu} \approx 0.20285$.*
- *For $M = 10^4 m$, $\kappa_1 = 0.3$, and $\eta = 0.26$, one obtains $\sqrt{\mu} \approx 0.07629$.*
- *For $M = 10^6 m$, $\kappa_1 = 0.05$, and $\eta = 0.2$, one obtains $\sqrt{\mu} \approx 0.005790$.*

These are evaluations of the coefficient, not certified confidence levels. In particular, an assertion of probability at least 0.999 would require an explicit numerical choice of the Bernstein constant c_1 and evaluation of (5)*; we do not make that assertion for these examples.*

C Examples and Discussion of the Extensions

Sub-Gaussian threshold conditions. In Theorem 3(i), isotropy does not guarantee a positive central trimmed second moment for the chosen threshold. For instance, an isotropic Rademacher vector has a coordinate projection supported on $\{-1, 1\}$, so this moment vanishes for $\gamma < 1$. In part (ii), Ψ_2 is a *tail* probability: the threshold set must contain at least as many large projections as any admissible Ω. Using infima and suprema avoids assuming that an extremizing direction exists. The scale-mixture family in Example 1 satisfies both sets of conditions.

Dependence assumptions and proof ingredients. The dependent extension requires a global operator-norm bound and two fixed-direction concentration bounds, corresponding in the independent proof to (39), (46), and (50). The same geometric argument then applies. Joint Gaussianity is an assumption on the entire process, not merely its one-time marginals. Likewise, mixing concerns the sigma-fields generated by the entire vectors, not separate coordinates. The process in Example 2 verifies these assumptions directly.

D Additional discussion of K-subspace clustering

Dependence on SNR, cluster count, and sample size. An admissible choice $\tau \asymp (\kappa_N K\sqrt{n/d}/\mathrm{SNR})^{1/3}$ gives

$$\ell(\hat{C}, C^\star) \lesssim (K-1)\left(\kappa_N\sqrt{n/d}\,\frac{K}{\mathrm{SNR}}\right)^{1/3}, \tag{35}$$

provided this choice satisfies $\tau < 1/K$ and (27). In particular, the prefactor $K-1$ cannot be absorbed into an absolute constant when K varies. For balanced clusters, define the dimension-adjusted ratio

$$\overline{\mathrm{SNR}} := \mathrm{SNR}\sqrt{d/n};$$

then $\kappa_N = 1$ and the right-hand side of (35) is at most a constant times $(K^4/\overline{\mathrm{SNR}})^{1/3}$. For a prescribed error $\epsilon \in (0,1)$, the choice $\tau = \epsilon/K$ is admissible and gives error at most $(1-1/K)\epsilon \le \epsilon$ if

$$\overline{\mathrm{SNR}} \ge \frac{CK^4}{\epsilon^3}, \qquad \frac{N}{K} \ge C(K/\epsilon)^6\left[d\log(CK/\epsilon) + \log(K/\rho)\right], \tag{36}$$

along with Assumption 1 and $N \ge 8/\rho$. Thus, for growing K, a sufficient consistency regime must permit $\epsilon \to 0$ while $K^4/\overline{\mathrm{SNR}} \to 0$ and the sample condition in (36) holds. Ignoring logarithmic factors, the latter involves $K^7 d/N \to 0$; the displayed condition specifies the additional dependence on accuracy and confidence. To obtain probability tending to one, one also chooses $\rho \to 0$ compatibly with these conditions. At fixed sample sizes, increasing SNR alone does not make this theorem certify arbitrarily small error. This describes the limitation of the guarantee, and is not an impossibility statement about clustering.

Exploratory algorithmic experiment. To examine the behavior suggested by the recovery bound, we simulate data from the LRGMM and run an initialized K-subspaces algorithm. We denote its output by $\hat{C}_{\mathrm{alg}}$ to distinguish it from a certified global minimizer $\hat{C}$. The rate-motivated ratio corresponding to (35) is

$$R_{\mathrm{alg}} := \frac{\ell(\hat{C}_{\mathrm{alg}}, C^\star)}{(K-1)\left(\sqrt{n/d}\,\kappa_N K/\mathrm{SNR}\right)^{1/3}}. \tag{37}$$

We vary the common noise level $\delta_k = \delta$, and hence the SNR, while fixing $K = 2$, $n = 100$, $N_1^\star = N_2^\star = 500$, and $d = 20$. The original experiment description records $\mu_{\max} = 0.67$. To compute $\hat{C}_{\mathrm{alg}}$, we run the K-subspaces algorithm initialized by the thresholding inner-product spectral method proposed in [32]. For each parameter setting, we generate independent data and repeat the experiment 10 times. The results are summarized in Table 1 in Appendix H.

Parameter reconciliation pending. The reported SNR values equal $0.5/\delta$, which corresponds to $\mu_{\max} = 1/\sqrt{2}$ under (23), whereas $\mu_{\max} = 0.67$ would give $0.5511/\delta$. The actual affinity used in

Table 1: Original reported algorithmic measurements under varying noise levels, retained pending verification of the subspace affinity and SNR normalization. These are not certified global-minimizer errors.

δ	$\mathrm{SNR}_{\mathrm{rep}}$	Mean of R_{rep}	Variance of R_{rep}
0.200	2.500	0.0000	0.00
0.300	1.667	0.0007	0.00
0.400	1.250	0.0034	0.00
0.500	1.000	0.0129	0.00
0.600	0.833	0.0397	0.00
0.700	0.714	0.0862	0.01
0.800	0.625	0.1858	0.03
0.900	0.556	0.2247	0.02
1.000	0.500	0.2226	0.01

the experiment has not yet been verified. We preserve all reported numerical measurements and label them $\mathrm{SNR}_{\mathrm{rep}}$ and R_{rep} until the configuration or raw clustering errors can be checked. Since $K = 2$, the factor $K - 1$ in (37) equals one and does not itself change these reported ratios.

The reported ratios are bounded over the tested settings and have small reported variances. This is an exploratory observation about the initialized algorithm. No certificate of global optimality was obtained. In fact, these settings do not satisfy the sufficient dimensional condition in Assumption 1: for $N = 1000$, $\alpha_N^2 = (2\sqrt{\log 1000} + 2)^2 \approx 52.66 > d = 20$. The sample-size condition in (27) has not been verified either. In particular, these data do not estimate the universal constant in Theorem 5 or verify a noisy algorithmic recovery guarantee. The initialization and noiseless convergence results of [32] do not by themselves establish such a guarantee for the present noisy model.

Finally, for any fixed $\eta \in (0, 1/2)$, the proof of Theorem 5 uses

$$\mu = b((1-\eta)\kappa_1) \gtrsim \kappa_1^3, \qquad 0 < \kappa_1 \le 1, \tag{38}$$

where the implicit constant may depend on the fixed η. We set $\eta = 1/4$ in the theorem's proof, choose $\kappa_1 = \tau$, and verify both the separation condition and the full failure probability in (90). The detailed argument in Appendix H.5 retains the small-κ_1 calculation and makes the sample-size and confidence requirements explicit.

E Regression filtering conditions and comparison

Normalized and unnormalized clean weights. The design condition in the filtering framework can also be stated for unnormalized clean weights. If their total mass is $s \ge 1 - 4\varepsilon$ and each weight is at most $1/N$, their normalization belongs to $\mathcal{Q}_\varepsilon$. Multiplying (31) by s gives a lower bound $s(1 - C\sqrt{\varepsilon})\mathbf{\Sigma}$ and an upper bound $(1 + C\sqrt{\varepsilon})\mathbf{\Sigma}$. For sufficiently small absolute ε, these lie between $\mathbf{\Sigma}/2$ and $3\mathbf{\Sigma}/2$. For the residual condition on all uncorrupted observations, take $q_i = \mathbf{1}_{\{i \in G\}}/|G|$, which also belongs to $\mathcal{Q}_\varepsilon$ because $|G| \ge (1 - \varepsilon)N$.

How the moment bounds complement subset covariance control. For Gaussian design, Theorems 1 and 2 describe the energy that can remain after observations are selected or erased. Regression filtering also requires control of products of a design projection and a residual. We therefore augment the whitened design with the standardized noise and establish a uniform fourth-moment event. On this event, a deterministic comparison between capped weights and the uniform weights controls all reweightings at once. This supplies the additional gradient conditions without treating the selected observations as independent.

The proof in Appendix I.1 gives the fourth-moment estimate and capped-weight calculation; the preceding paragraphs explain the reduction from the strong-contamination model. No assertion about the runtime or convergence of a filtering algorithm is needed for this concentration statement.

Why a uniform $O(\varepsilon)$ covariance perturbation is unavailable. At the population level, deleting $\{|Z| \geq r_\varepsilon\}$ and renormalizing gives second moment $(1-u(\varepsilon))/(1-\varepsilon)$. Its difference from one is

$$\frac{u(\varepsilon)-\varepsilon}{1-\varepsilon} \sim 2\varepsilon \log(1/\varepsilon),$$

which is not $O(\varepsilon)$. This selection is an admissible capped reweighting. The $O(\sqrt{\varepsilon})$ guarantee in Theorem 6 is consistent with this example and is the scale used in the gradient-mean condition.

F Proofs for Section 2

F.1 Proof of Theorem 1

Proof. In this proof, $\sigma_{\min}(\boldsymbol{A}_\Omega)$ denotes $\min_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\| = \sqrt{\lambda_{\min}(\boldsymbol{A}_\Omega \boldsymbol{A}_\Omega^T)}$, including the value zero when $\boldsymbol{A}_\Omega$ has deficient row rank. For ease of exposition, let $\boldsymbol{A}_\Omega \in \mathbb{R}^{m\times|\Omega|}$ be a matrix whose columns consist of $\boldsymbol{a}_i$ for each $i \in \Omega$ and $\boldsymbol{A} \in \mathbb{R}^{m\times M}$ be a matrix whose columns consist of $\boldsymbol{a}_i$ for all $i \in [M]$. Applying Fact 1 to $\boldsymbol{A}\boldsymbol{A}^T$ with $\eta = \sqrt{M}$ yields that it holds with probability at least $1 - 2\exp(-2M)$ that

$$\left\| \frac{1}{M}\boldsymbol{A}\boldsymbol{A}^T - \boldsymbol{I}_m \right\| \leq 2r + r^2 \leq 15, \qquad r = \frac{\sqrt{m}+2\sqrt{M}}{\sqrt{M}} \leq 3, \tag{39}$$

where the last inequality holds since we assume $M \geq m$. We retain the convenient bound 19 used below (the corrected estimate actually gives 16). We then have

$$\frac{1}{M}\|\boldsymbol{A}\|^2 = \left\| \frac{1}{M}\boldsymbol{A}\boldsymbol{A}^T \right\| \leq \left\| \frac{1}{M}\boldsymbol{A}\boldsymbol{A}^T - \boldsymbol{I}_m \right\| + \|\boldsymbol{I}_m\| \leq 19. \tag{40}$$

Let $\mathcal{N}_\varepsilon \subseteq \mathbb{S}^{m-1}$ be an ε-net of $\mathbb{S}^{m-1}$, i.e., for any $\boldsymbol{x} \in \mathbb{S}^{m-1}$, there exists $\boldsymbol{x}_0 \in \mathcal{N}_\varepsilon$ such that $\|\boldsymbol{x}-\boldsymbol{x}_0\| \leq \varepsilon$. Using [28, Corollary 4.2.13], the cardinality of the ε-net $\mathcal{N}_\varepsilon$ is

$$|\mathcal{N}_\varepsilon| \leq \left(\frac{2}{\varepsilon}+1\right)^m. \tag{41}$$

Let $\boldsymbol{x} \in \mathbb{S}^{m-1}$ be such that $\sigma_{\min}(\boldsymbol{A}_\Omega) = \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\|$ and $\boldsymbol{x}_0 \in \mathcal{N}_\varepsilon$ be such that $\|\boldsymbol{x}-\boldsymbol{x}_0\| \leq \varepsilon$. Using the triangle inequality, we obtain

$$\sigma_{\min}(\boldsymbol{A}_\Omega) \geq \|\boldsymbol{A}_\Omega^T \boldsymbol{x}_0\| - \varepsilon\|\boldsymbol{A}_\Omega\| \geq \min_{\boldsymbol{x}\in\mathcal{N}_\varepsilon} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\| - \varepsilon\|\boldsymbol{A}_\Omega\| \geq \min_{\boldsymbol{x}\in\mathcal{N}_\varepsilon} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\| - \varepsilon\|\boldsymbol{A}\|. \tag{42}$$

We claim and prove below that, on a single event with the probability specified in (51), the following holds simultaneously for all subsets $|\Omega| \geq \kappa_1 M$:

$$\min_{\boldsymbol{x}\in\mathcal{N}_\varepsilon} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\|^2 \geq (1-\eta)\mu M\ , \tag{43}$$

where μ is defined in (4). Substituting this and (40) into (42) yields

$$\sigma_{\min}(\boldsymbol{A}_\Omega) \geq \sqrt{(1-\eta)\mu M} - \varepsilon\sqrt{19M}.$$

From here to get (6), we let $\varepsilon = \sqrt{\mu}\eta/11$, which gives us

$$\sigma_{\min}(\boldsymbol{A}_\Omega) \geq \sqrt{\mu M}\left(\sqrt{1-\eta} - \frac{\sqrt{19}}{11}\eta\right) \geq (1-\eta)\sqrt{\mu M}$$

where the last inequality holds because $\eta \leq 1/2$ implies

$$\sqrt{1-\eta} - (1-\eta) = \frac{(1-\eta)-(1-\eta)^2}{\sqrt{1-\eta}+(1-\eta)} = \eta \cdot \frac{1}{1+1/\sqrt{1-\eta}} \geq \frac{1}{1+\sqrt{2}}\eta \geq \frac{\sqrt{19}}{11}\eta.$$

Adding the failure probability $2e^{-2M}$ of (40) to the failure probability for (43) gives (5). The events are constructed before choosing Ω, so the resulting statement is simultaneous over all admissible subsets.

The rest of the proof is devoted to proving the claim (43). We fix $\boldsymbol{x} \in \mathcal{N}_\varepsilon$ and let

$$\mathcal{A} := \left\{ i \in [M] : \left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \leq \gamma^2 \right\}, \text{ where } \gamma := \Phi^{-1}\left(\frac{1+(1-\eta)\kappa_1}{2}\right). \tag{44}$$

According to $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ for each $i \in [M]$ and $\boldsymbol{x} \in \mathcal{N}_\varepsilon$, we have $\boldsymbol{a}_i^T \boldsymbol{x} \sim \mathcal{N}(0,1)$ for each $i \in [M]$. Therefore, we obtain

$$\mathbb{P}\left(\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \leq \gamma^2\right) = \mathbb{P}\left(-\gamma \leq \boldsymbol{a}_i^T \boldsymbol{x} \leq \gamma\right) = 2\Phi\left(\gamma\right) - 1 = (1-\eta)\kappa_1. \tag{45}$$

It follows from (44) that $|\mathcal{A}| = \sum_{i=1}^M \mathbb{I}\left\{\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \leq \gamma^2\right\}$, where $\mathbb{I}\left\{\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \leq \gamma^2\right\} = 1$ if $\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \leq \gamma^2$ and 0 otherwise. This, together with (45), implies

$$\mathbb{E}\left[|\mathcal{A}|\right] = (1-\eta)\kappa_1 M, \quad \mathrm{Var}\left(|\mathcal{A}|\right) \leq (1-\eta)\kappa_1 M.$$

Applying Bernstein's inequality (see, e.g., [28, Theorem 2.8.4]) to $|\mathcal{A}|$, which is the sum of i.i.d. Bernoulli random variables, yields that

$$\begin{aligned}\mathbb{P}\left(\left||\mathcal{A}| - (1-\eta)\kappa_1 M\right| \geq \eta\kappa_1 M\right) &\leq 2\exp\left(-\frac{(\eta\kappa_1 M)^2}{2\left((1-\eta)\kappa_1 M + \eta\kappa_1 M/3\right)}\right) \\ &\leq 2\exp\left(-\frac{\eta^2\kappa_1 M}{2}\right). \end{aligned} \tag{46}$$

This implies that it holds with probability at least $1 - 2\exp\left(-\eta^2\kappa_1 M/2\right)$ that $|\mathcal{A}| \leq \kappa_1 M$. Since $|\Omega| \geq \kappa_1 M$, we obtain

$$\|\boldsymbol{A}_\Omega^T \boldsymbol{x}\|^2 = \sum_{i\in\Omega}\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \geq \sum_{i\in\mathcal{A}}\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 = \sum_{i=1}^M \left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \mathbb{I}\left\{\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \leq \gamma^2\right\}, \tag{47}$$

where the inequality follows because $\mathcal{A}$ contains the smallest directional squared projections and $|\mathcal{A}| \leq |\Omega|$. It does not require $\mathcal{A} \subseteq \Omega$. The last equality follows from (44). Since $\boldsymbol{a}_i^T \boldsymbol{x} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ for each $i \in [M]$, we compute

$$\begin{aligned}\mu &:= \mathbb{E}\left[\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \mathbb{I}\left\{\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \leq \gamma^2\right\}\right] = \sqrt{\frac{2}{\pi}}\int_0^\gamma x^2 \exp\left(-\frac{x^2}{2}\right) dx \\ &= -\sqrt{\frac{2}{\pi}}\left(x\exp\left(-\frac{x^2}{2}\right)\Big|_0^\gamma - \int_0^\gamma \exp\left(-\frac{x^2}{2}\right) dx\right) = (1-\eta)\kappa_1 - \sqrt{\frac{2\gamma^2}{\pi}}\exp\left(-\frac{\gamma^2}{2}\right), \end{aligned} \tag{48}$$

where the last equality uses the definition of γ in (44). Since $\boldsymbol{a}_i^T \boldsymbol{x}$ is Gaussian, $\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2$ is subexponential. Applying Hoeffding's inequality for bounded distribution [28, Theorem 2.2.6] yields

$$\mathbb{P}\left(\left|\sum_{i=1}^M \left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \mathbb{I}\left\{\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \leq \gamma^2\right\} - M\mu\right| \geq \eta M\mu\right) \leq 2\exp\left(-\frac{2\eta^2\mu^2 M}{\gamma^4}\right). \tag{49}$$

The bounded-variable estimate remains valid for each finite γ but becomes uninformative as γ grows. We therefore also use a subexponential estimate with a constant independent of the truncation threshold. Applying Subexponential Bernstein inequality [28] yields

$$\mathbb{P}\left(\left|\sum_{i=1}^M \left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \mathbb{I}\left\{\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \leq \gamma^2\right\} - M\mu\right| \geq \eta M\mu\right) \leq 2\exp\left(-\frac{c' \cdot \eta^2\mu^2 M}{K_G^2}\right), \tag{50}$$

where $c' > 0$ is an absolute constant and $K_G = \|Z^2\|_{\psi_1} \geq \|Z^2\mathbb{I}\{|Z| \leq \gamma\}\|_{\psi_1}$ for $Z \sim \mathcal{N}(0,1)$. Centering changes the subexponential norm by at most an absolute factor, absorbed in c'. Since $0 < \eta\mu < 1$ and K_G is an absolute constant, the quadratic branch of Bernstein's inequality gives the displayed bound. Let $c = c'/K_G^2$; after decreasing it if needed, it is the constant c_1 in Theorem 1. This implies that it holds with probability at least $1 - 2\exp\left(-c \cdot \eta^2\mu^2 M\right)$ that

$$\sum_{i=1}^M \left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \mathbb{I}\left\{\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \leq \gamma^2\right\} \geq (1-\eta)M\mu,$$

This, together with (47) implies (43). Finally, applying the union bound over the deterministic net in (41) yields that (43) holds simultaneously for every admissible Ω with probability at least

$$1-2\left(\frac{22}{\eta\sqrt{\mu}}+1\right)^{m}\left(\exp\left(-\frac{\eta^2\kappa_1 M}{2}\right)+\min\left\{\exp\left(-c\cdot\eta^2\mu^2 M\right),\exp\left(-\frac{2\eta^2\mu^2 M}{\gamma^4}\right)\right\}\right). \tag{51}$$

□

F.2 Lower Bound of κ_1 in Theorem 1

The Gaussian central and tail curves. The integration-by-parts calculations in (48) and (58) give the formulas for b and u in (2). Since

$$\kappa=\sqrt{\frac{2}{\pi}}\int_0^{q_\kappa} e^{-z^2/2}\,dz,$$

we have $q_\kappa\sim\sqrt{\pi/2}\,\kappa$ as $\kappa\downarrow 0$, and hence

$$b(\kappa)=\sqrt{\frac{2}{\pi}}\int_0^{q_\kappa} z^2e^{-z^2/2}\,dz\sim\sqrt{\frac{2}{\pi}}\,\frac{q_\kappa^3}{3}\sim\frac{\pi}{6}\kappa^3.$$

For the tail, Mills' ratio [8] gives

$$\kappa=2[1-\Phi(r_\kappa)]\sim\frac{2\phi(r_\kappa)}{r_\kappa},\qquad r_\kappa^2\sim 2\log(1/\kappa).$$

Thus $u(\kappa)=\kappa+2r_\kappa\phi(r_\kappa)\sim\kappa r_\kappa^2\sim 2\kappa\log(1/\kappa)$. Dividing by κ yields the orders under selected-cardinality normalization. For each fixed $\eta\in(0,1/2)$, continuity and positivity further give

$$c_\eta\kappa^3\le b((1-\eta)\kappa)\le C_\eta\kappa^3,\qquad 0<\kappa\le 1. \tag{52}$$

When the displayed probability is informative. Formally, Theorem 1 is stated for any $\kappa_1\in(0,1]$. What depends on κ_1 is whether the probability bound in (5) is non-vacuous and quantitatively strong. We write the probability in (5) schematically as

$$1-2\left(1+\frac{22}{\eta\sqrt{\mu}}\right)^{m}\left(\exp\left(-c\,\eta^2\kappa_1 M\right)+\exp\left(-c\,\eta^2\mu^2 M\right)\right)-2\exp\left(-2M\right),$$

At fixed confidence, a convenient sufficient scaling ensuring that the above probability is close to one is

$$m\log\Big(1+\frac{22}{\eta\sqrt{\mu}}\Big)\ll\eta^2\kappa_1 M,\quad m\log\Big(1+\frac{22}{\eta\sqrt{\mu}}\Big)\ll\eta^2\mu^2 M.$$

where $\mu=\mu(\kappa_1,\eta)$ is defined in Theorem 1. Thus, a sufficient regime for this probability certificate is obtained by making these exponents dominate the covering-number term. Using

$$\gamma=\Phi^{-1}\left(\frac{1+(1-\eta)\kappa_1}{2}\right)=\Theta(\kappa_1),\quad \mu=\sqrt{\frac{2}{\pi}}\int_0^{\gamma}x^2e^{-x^2/2}\,dx=\Theta(\kappa_1^3),$$

we obtain the rough scaling

$$\frac{M}{m}\gtrsim\kappa_1^{-6}\log\frac{1}{\kappa_1},$$

up to constants and the dependence on η. Equivalently, for a fixed sufficiently large ratio M/m, solving this sufficient condition gives the rough scale

$$\kappa_1\gtrsim\Big(\frac{m\log(M/m)}{M}\Big)^{1/6}.$$

To keep a target failure probability $\rho\in(0,1)$ explicit, it suffices that

$$\frac{\eta^2\kappa_1 M}{2}\ge m\log\left(1+\frac{22}{\eta\sqrt{\mu}}\right)+\log(6/\rho),$$

$$c_1\eta^2\mu^2 M\ge m\log\left(1+\frac{22}{\eta\sqrt{\mu}}\right)+\log(6/\rho),\quad 2M\ge\log(6/\rho).$$

Each of the three failure terms in (5) is then at most $\rho/3$. Substituting (52) proves the sufficient condition (7), after increasing C_η. The exponent 6 describes the displayed Bernstein certificate and is not asserted to be optimal; see (34) for a prior lower bound with a better small-fraction sample dependence. The numerical coefficients in Example 3 do not certify a particular confidence without specifying c_1.

F.3 Proof of Theorem 2

Proof. Using the ε-net and [28, Exercise 4.4.3], we have

$$\left\|\sum_{i\in\Omega}\boldsymbol{a}_i\boldsymbol{a}_i^T\right\| \le \frac{1}{1-2\varepsilon}\max_{\boldsymbol{x}\in\mathcal{N}_\varepsilon}\sum_{i\in\Omega}\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2. \tag{53}$$

We fix $\boldsymbol{x}\in\mathcal{N}_\varepsilon$ and let

$$\mathcal{A} := \left\{i\in[M] : \left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2 \ge \gamma^2\right\}, \text{ where } \gamma := \Phi^{-1}\left(1-\frac{(1+\eta)\kappa_2}{2}\right). \tag{54}$$

According to $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$ and $\boldsymbol{x}\in\mathcal{N}_\varepsilon$, we have $\boldsymbol{a}_i^T\boldsymbol{x} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ for each $i\in[M]$. Therefore, we obtain for each $i\in[M]$,

$$\mathbb{P}\left(\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2\ge\gamma^2\right) = 2\mathbb{P}\left(\boldsymbol{a}_i^T\boldsymbol{x}\ge\gamma\right) = 2\left(1-\Phi(\gamma)\right) = (1+\eta)\kappa_2. \tag{55}$$

It follows from (54) that $|\mathcal{A}| = \sum_{i=1}^M \mathbb{I}\left\{\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2\ge\gamma^2\right\}$, where $\mathbb{I}\left\{\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2\ge\gamma^2\right\} = 1$ if $\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2\ge\gamma^2$ and 0 otherwise. This, together with (55), implies

$$\mathbb{E}\left[|\mathcal{A}|\right] = (1+\eta)\kappa_2 M, \quad \mathrm{Var}\left(|\mathcal{A}|\right) \le (1+\eta)\kappa_2 M. \tag{56}$$

Applying Bernstein's inequality (see, e.g., [28, Theorem 2.8.4]) to $|\mathcal{A}|$, which is the sum of i.i.d. Bernoulli random variables, yields

$$\begin{aligned}\mathbb{P}\left(\left||\mathcal{A}|-(1+\eta)\kappa_2 M\right|\ge\eta\kappa_2 M\right) &\le 2\exp\left(-\frac{\eta^2\kappa_2^2M^2}{2\left((1+\eta)\kappa_2 M+\eta\kappa_2 M/3\right)}\right)\\ &\le 2\exp\left(-\frac{\eta^2\kappa_2 M}{4}\right),\end{aligned}$$

where the last inequality holds because $\eta\le 1/2$. This implies that it holds with probability at least $1-2\exp\left(-\eta^2\kappa_2 M/4\right)$ that $|\mathcal{A}|\ge\kappa_2 M$. Since $|\Omega|\le\kappa_2 M$, we obtain

$$\|\boldsymbol{A}_\Omega^T\boldsymbol{x}\|^2 = \sum_{i\in\Omega}\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2 \le \sum_{i\in\mathcal{A}}\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2 = \sum_{i=1}^M\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2\mathbb{I}\left\{\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2\ge\gamma^2\right\}, \tag{57}$$

where the inequality uses the ordering of the directional squared projections: $\mathcal{A}$ contains all projections above the threshold and has at least $|\Omega|$ elements. The last equality follows from (54). Since $\boldsymbol{a}_i^T\boldsymbol{x}\overset{i.i.d.}{\sim}\mathcal{N}(0,1)$ for each $i\in[M]$, we compute

$$\begin{aligned}\mu &:= \mathbb{E}\left[\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2\mathbb{I}\left\{\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2\ge\gamma^2\right\}\right] = \sqrt{\frac{2}{\pi}}\int_\gamma^\infty x^2\exp\left(-\frac{x^2}{2}\right)dx\\ &= -\sqrt{\frac{2}{\pi}}\left(x\exp\left(-\frac{x^2}{2}\right)\Big|_\gamma^\infty - \int_\gamma^\infty\exp\left(-\frac{x^2}{2}\right)dx\right) = (1+\eta)\kappa_2 + \sqrt{\frac{2\gamma^2}{\pi}}\exp\left(-\frac{\gamma^2}{2}\right),\end{aligned} \tag{58}$$

where the last equality uses the definition of γ in (54). Similar to (50), Subexponential Bernstein inequality [28] yields

$$\mathbb{P}\left(\left|\sum_{i=1}^M\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2\mathbb{I}\left\{\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)^2\ge\gamma^2\right\} - M\mu\right|\ge\eta M\mu\right)\le 2\exp\left(-c_2\cdot\eta^2\mu^2 M\right), \tag{59}$$

for some absolute constant $c_2 > 0$. Indeed, $0 \le Z^2\mathbb{I}\{|Z| \ge \gamma\} \le Z^2$ gives a uniform subexponential norm, and $\eta\mu < 1$ selects the quadratic branch of Bernstein's inequality. This implies that, with probability at least $1 - 2\exp\left(-c_2 \cdot \eta^2\mu^2 M\right)$,

$$\sum_{i=1}^{M} \left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \mathbb{I}\left\{\left(\boldsymbol{a}_i^T \boldsymbol{x}\right)^2 \ge \gamma^2\right\} \le (1+\eta)\mu M.$$

This, together with (53) and (57), implies

$$\left\|\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right\| \le \frac{(1+\eta)\mu M}{1-2\varepsilon}.$$

Setting $\varepsilon = \eta/3$, since $\eta \in (0, 1/2)$, we have $\frac{1}{1-2\varepsilon} \le 1+\eta$, which together with the above inequality, yields (9). Using (41) and the union bound yields that (9) holds with probability at least

$$1 - 2\left(\frac{6}{\eta} + 1\right)^m \left(\exp\left(-\frac{\eta^2\kappa_2 M}{6}\right) + \exp\left(-c_2 \cdot \eta^2\mu^2 M\right)\right).$$

The same event works for every admissible subset because the ordering argument holds before Ω is chosen. At $(1+\eta)\kappa_2 = 1$, we have $\gamma = 0$ and $\mathcal{A} = [M]$, so the count event is deterministic and the rest of the proof is unchanged. □

F.4 Proof of Corollary 1

Proof of Part (i) in Corollary 1. The assumption $(1+\eta)\kappa_1 < 1$ ensures that the threshold below is finite. We fix some $\boldsymbol{x}_0 \in \mathbb{S}^{m-1}$ and let

$$\widetilde{\mathcal{A}} := \left\{i \in [M] : \left(\boldsymbol{a}_i^T \boldsymbol{x}_0\right)^2 \le \widetilde{\gamma}^2\right\}, \text{ where } \widetilde{\gamma} := \Phi^{-1}\left(\frac{1+(1+\eta)\kappa_1}{2}\right). \tag{60}$$

Similar to (45), we know that

$$\mathbb{P}\left(\left(\boldsymbol{a}_i^T \boldsymbol{x}_0\right)^2 \le \widetilde{\gamma}^2\right) = \mathbb{P}\left(-\widetilde{\gamma} \le \boldsymbol{a}_i^T \boldsymbol{x}_0 \le \widetilde{\gamma}\right) = 2\Phi\left(\widetilde{\gamma}\right) - 1 = (1+\eta)\kappa_1, \tag{61}$$

which implies that $\mathbb{E}\left[|\widetilde{\mathcal{A}}|\right] = (1+\eta)\kappa_1 M$ and $\mathrm{Var}\left(|\widetilde{\mathcal{A}}|\right) \le (1+\eta)\kappa_1 M$. Then, similar to (46), we have

$$\begin{aligned}\mathbb{P}\left(\left||\widetilde{\mathcal{A}}| - (1+\eta)\kappa_1 M\right| \ge \eta\kappa_1 M\right) &\le 2\exp\left(-\frac{(\eta\kappa_1 M)^2}{2\left((1+\eta)\kappa_1 M + \eta\kappa_1 M/3\right)}\right) \\ &\le 2\exp\left(-\frac{\eta^2\kappa_1 M}{4}\right).\end{aligned} \tag{62}$$

This implies that it holds with probability at least $1 - 2\exp\left(-\eta^2\kappa_1 M/4\right)$ that $|\widetilde{\mathcal{A}}| \ge \kappa_1 M$. By choosing $\Omega = \widetilde{\mathcal{A}}$, we obtain

$$\lambda_{\min}\left(\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right) \le \|\boldsymbol{A}_\Omega^T \boldsymbol{x}_0\|^2 = \sum_{i\in\widetilde{\mathcal{A}}} \left(\boldsymbol{a}_i^T \boldsymbol{x}_0\right)^2 = \sum_{i=1}^{M} \left(\boldsymbol{a}_i^T \boldsymbol{x}_0\right)^2 \mathbb{I}\left\{\left(\boldsymbol{a}_i^T \boldsymbol{x}_0\right)^2 \le \widetilde{\gamma}^2\right\}, \tag{63}$$

where the last equality follows from (60). Since $\boldsymbol{a}_i^T \boldsymbol{x}_0 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ for each $i \in [M]$, we compute

$$\begin{aligned}\widetilde{\mu} &:= \mathbb{E}\left[\left(\boldsymbol{a}_i^T \boldsymbol{x}_0\right)^2 \mathbb{I}\left\{\left(\boldsymbol{a}_i^T \boldsymbol{x}_0\right)^2 \le \widetilde{\gamma}^2\right\}\right] = \sqrt{\frac{2}{\pi}} \int_0^{\widetilde{\gamma}} x^2 \exp\left(-\frac{x^2}{2}\right) dx \\ &= -\sqrt{\frac{2}{\pi}}\left(x\exp\left(-\frac{x^2}{2}\right)\Big|_0^{\widetilde{\gamma}} - \int_0^{\widetilde{\gamma}} \exp\left(-\frac{x^2}{2}\right) dx\right) = (1+\eta)\kappa_1 - \sqrt{\frac{2\widetilde{\gamma}^2}{\pi}} \exp\left(-\frac{\widetilde{\gamma}^2}{2}\right),\end{aligned} \tag{64}$$

where the last equality uses the definition of $\widetilde{\gamma}$ in (60). The same concentration inequality as (50) implies that it holds with probability at least $1-2\exp\left(-c\cdot\eta^2\widetilde{\mu}^2M\right)$ that

$$\sum_{i=1}^{M}\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\mathbb{I}\left\{\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\le\widetilde{\gamma}^2\right\}\le(1+\eta)M\widetilde{\mu},$$

Thus, the result holds with probability at least

$$1-2\exp\left(-\frac{\eta^2\kappa_1 M}{4}\right)-2\exp\left(-c\cdot\eta^2\widetilde{\mu}^2M\right).$$

Proof of Part (ii) in Corollary 1. For any fixed $\boldsymbol{x}_0\in\mathbb{S}^{m-1}$, let

$$\widetilde{\mathcal{A}}:=\left\{i\in[M]:\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\ge\widetilde{\gamma}^2\right\},\ \text{where}\ \widetilde{\gamma}:=\Phi^{-1}\left(1-\frac{(1-\eta)\kappa_2}{2}\right).\tag{65}$$

According to $\boldsymbol{a}_i\overset{i.i.d.}{\sim}\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$ and $\boldsymbol{x}_0\in\mathbb{S}^{m-1}$, we have $\boldsymbol{a}_i^T\boldsymbol{x}_0\overset{i.i.d.}{\sim}\mathcal{N}(0,1)$ for each $i\in[M]$. Therefore, we obtain for each $i\in[M]$,

$$\mathbb{P}\left(\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\ge\widetilde{\gamma}^2\right)=2\mathbb{P}\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\ge\widetilde{\gamma}\right)=2\left(1-\Phi(\widetilde{\gamma})\right)=(1-\eta)\kappa_2.\tag{66}$$

It follows from (65) that $|\widetilde{\mathcal{A}}|=\sum_{i=1}^{M}\mathbb{I}\left\{\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\ge\widetilde{\gamma}^2\right\}$, where $\mathbb{I}\left\{\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\ge\widetilde{\gamma}^2\right\}=1$ if $\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\ge\widetilde{\gamma}^2$ and 0 otherwise. This, together with (66), implies

$$\mathbb{E}\left[|\widetilde{\mathcal{A}}|\right]=(1-\eta)\kappa_2M,\quad\mathrm{Var}\left(|\widetilde{\mathcal{A}}|\right)\le(1-\eta)\kappa_2M.\tag{67}$$

Applying Bernstein's inequality (see, e.g., [28, Theorem 2.8.4]) to $|\widetilde{\mathcal{A}}|$, which is the sum of i.i.d. Bernoulli random variables, yields

$$\begin{aligned}\mathbb{P}\left(\left||\widetilde{\mathcal{A}}|-(1-\eta)\kappa_2M\right|\ge\eta\kappa_2M\right)&\le2\exp\left(-\frac{\eta^2\kappa_2^2M^2}{2\left((1-\eta)\kappa_2M+\eta\kappa_2M/3\right)}\right)\\&\le2\exp\left(-\frac{\eta^2\kappa_2M}{2}\right).\end{aligned}$$

This implies that it holds with probability at least $1-2\exp\left(-\eta^2\kappa_2M/2\right)$ that $|\widetilde{\mathcal{A}}|\le\kappa_2M$. By choosing $\Omega=\widetilde{\mathcal{A}}$, we obtain

$$\left\|\sum_{i\in\Omega}\boldsymbol{a}_i\boldsymbol{a}_i^T\right\|\ge\|\boldsymbol{A}_\Omega^T\boldsymbol{x}_0\|^2=\sum_{i\in\widetilde{\mathcal{A}}}\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2=\sum_{i=1}^{M}\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\mathbb{I}\left\{\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\ge\widetilde{\gamma}^2\right\},\tag{68}$$

As in (58) and (59), we have

$$\widetilde{\mu}:=\mathbb{E}\left[\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\mathbb{I}\left\{\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\ge\widetilde{\gamma}^2\right\}\right]=(1-\eta)\kappa_2+\sqrt{\frac{2\widetilde{\gamma}^2}{\pi}}\exp\left(-\frac{\widetilde{\gamma}^2}{2}\right).\tag{69}$$

and

$$\mathbb{P}\left(\left|\sum_{i=1}^{M}\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\mathbb{I}\left\{\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)^2\ge\widetilde{\gamma}^2\right\}-M\widetilde{\mu}\right|\ge\eta M\widetilde{\mu}\right)\le2\exp\left(-c_2\cdot\eta^2\widetilde{\mu}^2M\right),\tag{70}$$

Thus, the result holds with probability at least

$$1-2\exp\left(-\frac{\eta^2\kappa_2M}{2}\right)-2\exp\left(-c_2\cdot\eta^2\widetilde{\mu}^2M\right).$$

F.5 Weakness of Simple Union Bound

To address the difficulty caused by the dependence induced by Ω, a natural idea is to use a union bound over all possible choices of Ω based on Fact 1. The resulting certificate can require a large retained fraction. The threshold depends on the chosen fixed-subset estimate; this calculation is not an impossibility result for union bounds in general.

We first analyze a direct use of the corrected covariance inequality, retaining its quadratic term. Assume for simplicity that $s = \kappa_1 M$ is an integer; rounding does not affect the asymptotic comparison. For any fixed set Ω (independent of $\{\boldsymbol{a}_i\}_{i=1}^M$) with cardinality $|\Omega| = \kappa_1 M$, it follows from Fact 1 that

$$\left\| \frac{1}{\kappa_1 M} \sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T - \boldsymbol{I}_m \right\| \le 2r_\Omega + r_\Omega^2, \qquad r_\Omega := \frac{\sqrt{m}+2\eta}{\sqrt{\kappa_1 M}} \tag{71}$$

holds with probability at least $1 - 2\exp\left(-2\eta^2\right)$. For this fixed Ω, applying Weyl's inequality to (71) yields

$$\lambda_{\min}\left(\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right) \ge \kappa_1 M \left(1 - 2r_\Omega - r_\Omega^2\right). \tag{72}$$

For the right-hand side to be positive, one needs $r_\Omega < \sqrt{2} - 1$. In particular,

$$\eta < \frac{(\sqrt{2}-1)\sqrt{\kappa_1 M} - \sqrt{m}}{2} < \frac{\sqrt{2}-1}{2}\sqrt{\kappa_1 M}.$$

Consequently, the exponential rate in this certificate satisfies

$$2\eta^2 < c_0 \kappa_1 M, \qquad c_0 := \frac{(\sqrt{2}-1)^2}{2} \approx 0.08579.$$

This concerns the lower confidence value supplied by the inequality, not the true probability of the event.

On the other hand, the number of Ω with cardinality equal to $\kappa_1 M$ is given by

$$\begin{aligned} C_M^{\kappa_1 M} = \frac{M!}{(\kappa_1 M)!((1-\kappa_1)M)!} \approx & \frac{\sqrt{2\pi M}\left(\frac{M}{e}\right)^M}{\sqrt{2\pi\kappa_1 M}\left(\frac{\kappa_1 M}{e}\right)^{\kappa_1 M}\sqrt{2\pi(1-\kappa_1)M}\left(\frac{(1-\kappa_1)M}{e}\right)^{(1-\kappa_1)M}} \\ = & \frac{\exp\left(M\left(\kappa_1 \log\frac{1}{\kappa_1} + (1-\kappa_1)\log\frac{1}{1-\kappa_1}\right)\right)}{\sqrt{2\pi\kappa_1(1-\kappa_1)M}}, \end{aligned}$$

where the approximation follows from Stirling's formula $n! \sim \sqrt{2\pi n}(n/e)^n$. Applying a union bound gives the lower confidence certificate

$$1 - 2\binom{M}{\kappa_1 M} e^{-2\eta^2}.$$

Writing $H(\kappa) = \kappa\log(1/\kappa) + (1-\kappa)\log(1/(1-\kappa))$, its failure term satisfies

$$2\binom{M}{\kappa_1 M} e^{-2\eta^2} > \frac{\sqrt{2}+o(1)}{\sqrt{\pi\kappa_1(1-\kappa_1)M}} \exp\big(M[H(\kappa_1) - c_0\kappa_1]\big).$$

Thus, for this certificate to remain non-vacuous asymptotically, it is necessary that $H(\kappa_1) \le c_0\kappa_1$. Since $H(\kappa_1) > \kappa_1\log(1/\kappa_1)$, this implies $\kappa_1 > e^{-c_0} \approx 0.918$. This necessary condition for the certificate is not sufficient: the full entropy condition and the dependence on m/M still have to be checked.

A second, scalar chi-square certificate. To separate the issue from constants in Fact 1, consider $m = 1$ and $s = \kappa_1 M \in \mathbb{N}$. For a fixed subset Ω of size s, the sum $\sum_{i\in\Omega} a_i^2$ has the chi-square distribution with s degrees of freedom. The lower-tail estimate of [12] gives

$$\mathbb{P}\left(\sum_{i\in\Omega} a_i^2 \le s - 2\sqrt{sx}\right) \le e^{-x}, \qquad x > 0.$$

For $0 < t < 1$, taking $x = t^2 s/4$ yields

$$\mathbb{P}\left(\sum_{i\in\Omega} a_i^2 \le (1-t)s\right) \le e^{-t^2 s/4}.$$

The union bound over all subsets of this cardinality therefore yields

$$\mathbb{P}\left(\min_{|\Omega|=s} \sum_{i\in\Omega} a_i^2 > (1-t)s\right) \ge 1 - \binom{M}{s} e^{-t^2 s/4}.$$

Using the same Stirling expansion, the failure term has logarithm

$$M\big[H(\kappa_1) - t^2\kappa_1/4\big] + O(\log M).$$

For this particular certificate to be non-vacuous for arbitrarily large M with fixed κ_1 and t, it is necessary that $H(\kappa_1) \le t^2\kappa_1/4$. Since $t < 1$ and $H(\kappa_1) > \kappa_1 \log(1/\kappa_1)$, this implies

$$\kappa_1 > e^{-1/4} \approx 0.7788.$$

Again, this is only a necessary condition for the displayed certificate, not a fundamental threshold for arbitrary-subset covariance bounds. It does not establish tightness of the first calculation. Sharper chi-square large-deviation estimates, order-statistic arguments, and the erasure-robustness results discussed in Appendix A can yield different and much smaller admissible retained fractions. This derivation uses a finite-sample tail inequality throughout; a central limit approximation at a deviation growing with M would not justify the comparison.

Remark 2. *We can also consider the complementary set $\overline{\Omega} = [M] \setminus \Omega$ and compute an upper bound on $\lambda_{\max}\left(\sum_{i\in\overline{\Omega}} \boldsymbol{a}_i \boldsymbol{a}_i^T\right)$ to derive a lower bound for $\lambda_{\min}\left(\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right)$. This gives the valid deterministic inequality $\lambda_{\min}(\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T) \ge \lambda_{\min}(\sum_{i=1}^M \boldsymbol{a}_i \boldsymbol{a}_i^T) - \lambda_{\max}(\sum_{i\in\overline{\Omega}} \boldsymbol{a}_i \boldsymbol{a}_i^T)$. It is useful when the erased portion is small, but its quantitative range depends on the chosen upper-tail estimate; it does not imply a universal retained-fraction threshold.*

F.6 Proof of Theorem 3

Proof. We prove both parts, keeping the directional threshold and the moment explicit throughout. Write $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_M] \in \mathbb{R}^{m\times M}$ and let $\boldsymbol{A}_\Omega$ contain the columns indexed by Ω.

Part (i): global norm and extension from a net. The rows of $\boldsymbol{A}^T$ are independent, centered, isotropic sub-Gaussian vectors. The independent-row singular-value bound [28, Theorem 4.6.1] implies that, for every $t \ge 0$,

$$\mathbb{P}\left(\|\boldsymbol{A}\| > \sqrt{M} + CK_{\mathrm{sg}}^2(\sqrt{m} + t)\right) \le 2e^{-t^2},$$

after an absolute adjustment of C. This statement concerns independent rows of $\boldsymbol{A}^T$ and does not require independent entries. Since $M \ge m$, choosing $t = \sqrt{M}$ and a sufficiently large constant $c_3 \ge 1$ depending only on K_{sg} gives the event

$$\mathcal{E}_{\mathrm{sg}} := \{\|\boldsymbol{A}\| \le 3c_3\sqrt{M}\}, \qquad \mathbb{P}(\mathcal{E}_{\mathrm{sg}}) \ge 1 - 2e^{-M}.$$

Let $\mathcal{N}_\varepsilon$ be an ε-net of $\mathbb{S}^{m-1}$ with cardinality at most $(1 + 2/\varepsilon)^m$. On this event, approximation by the net gives, for every Ω,

$$\begin{aligned}\min_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\| &\ge \min_{\boldsymbol{x}\in\mathcal{N}_\varepsilon} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\| - \varepsilon\|\boldsymbol{A}\| \\ &\ge \min_{\boldsymbol{x}\in\mathcal{N}_\varepsilon} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\| - 3c_3\sqrt{M}\,\varepsilon. \end{aligned} \tag{73}$$

The minimum over $\mathbb{S}^{m-1}$ is used explicitly so that the formula also covers rank-deficient subset matrices. We next prove that the net event

$$\min_{\boldsymbol{x}\in\mathcal{N}_\varepsilon} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\|^2 \ge (1-\eta)\mu M \tag{74}$$

holds simultaneously for all $|\Omega| \ge \kappa_3 M$ with the claimed probability.

Part (i): counting central projections. Fix $\boldsymbol{x} \in \mathcal{N}_\varepsilon$ and use the threshold γ assumed in the theorem. Define

$$\mathcal{A}_{\boldsymbol{x}} := \{i \in [M] : |\boldsymbol{a}_i^T \boldsymbol{x}| \le \gamma\}, \qquad p_{\boldsymbol{x}} := \mathbb{P}(|\boldsymbol{a}_i^T \boldsymbol{x}| \le \gamma). \tag{75}$$

By (14),

$$p_{\boldsymbol{x}} \le \Psi_1(\gamma) \le (1-\eta)\kappa_3. \tag{76}$$

In particular, this step remains valid if $\boldsymbol{a}_i^T \boldsymbol{x}$ has atoms at either threshold endpoint. If it is expressed through a CDF $P_{\boldsymbol{x}}$, the probability is $P_{\boldsymbol{x}}(\gamma) - P_{\boldsymbol{x}}(-\gamma-)$; it is not generally $P_{\boldsymbol{x}}(\gamma) - P_{\boldsymbol{x}}(-\gamma)$.

The cardinality $|\mathcal{A}_{\boldsymbol{x}}|$ is the sum of independent Bernoulli variables, so

$$\mathbb{E}|\mathcal{A}_{\boldsymbol{x}}| = Mp_{\boldsymbol{x}}, \qquad \mathrm{Var}(|\mathcal{A}_{\boldsymbol{x}}|) = Mp_{\boldsymbol{x}}(1-p_{\boldsymbol{x}}) \le (1-\eta)\kappa_3 M.$$

Bernstein's inequality therefore gives

$$\begin{aligned}\mathbb{P}(|\mathcal{A}_{\boldsymbol{x}}| > \kappa_3 M) &\le \mathbb{P}(|\mathcal{A}_{\boldsymbol{x}}| - Mp_{\boldsymbol{x}} > \eta\kappa_3 M) \\ &\le 2\exp\left(-\frac{\eta^2\kappa_3^2 M^2}{2[(1-\eta)\kappa_3 M + \eta\kappa_3 M/3]}\right) \le 2e^{-\eta^2\kappa_3 M/2}.\end{aligned}$$

On the complementary event, $|\mathcal{A}_{\boldsymbol{x}}| \le \kappa_3 M \le |\Omega|$. Every squared projection indexed outside $\mathcal{A}_{\boldsymbol{x}}$ exceeds γ^2, whereas those indexed inside are at most γ^2. Pairing the indices in $\mathcal{A}_{\boldsymbol{x}} \setminus \Omega$ with as many indices in $\Omega \setminus \mathcal{A}_{\boldsymbol{x}}$ thus proves

$$\|\boldsymbol{A}_\Omega^T \boldsymbol{x}\|^2 = \sum_{i\in\Omega} (\boldsymbol{a}_i^T \boldsymbol{x})^2 \ge \sum_{i\in\mathcal{A}_{\boldsymbol{x}}} (\boldsymbol{a}_i^T \boldsymbol{x})^2 = \sum_{i=1}^M (\boldsymbol{a}_i^T \boldsymbol{x})^2 \mathbf{1}\{|\boldsymbol{a}_i^T \boldsymbol{x}| \le \gamma\}. \tag{77}$$

This comparison holds for every admissible subset; it does not require $\mathcal{A}_{\boldsymbol{x}} \subseteq \Omega$.

Part (i): central energy and conclusion. Set

$$T_i^{(\boldsymbol{x})} := (\boldsymbol{a}_i^T \boldsymbol{x})^2 \mathbf{1}\{|\boldsymbol{a}_i^T \boldsymbol{x}| \le \gamma\}, \qquad \mu_{\boldsymbol{x}} := \mathbb{E}T_i^{(\boldsymbol{x})} \ge \mu > 0. \tag{78}$$

The variables $T_i^{(\boldsymbol{x})}$ are independent and take values in $[0, \gamma^2]$. Hoeffding's inequality, with a weakened constant for consistency with the statement, yields

$$\mathbb{P}\left(\left|\sum_{i=1}^M T_i^{(\boldsymbol{x})} - M\mu_{\boldsymbol{x}}\right| \ge \eta\mu M\right) \le 2\exp\left(-\frac{\eta^2\mu^2 M}{2\gamma^4}\right).$$

On the complementary event,

$$\sum_{i=1}^M T_i^{(\boldsymbol{x})} \ge M\mu_{\boldsymbol{x}} - \eta\mu M \ge (1-\eta)\mu M.$$

Together with (77), a union bound over the deterministic net establishes (74) with probability at least

$$1 - 2|\mathcal{N}_\varepsilon|\left[e^{-\eta^2\kappa_3 M/2} + \exp\left(-\frac{\eta^2\mu^2 M}{2\gamma^4}\right)\right]. \tag{79}$$

Choose $\varepsilon = \eta\sqrt{\mu}/(8c_3)$. Isotropy implies $\mu \le 1$, so $0 < \varepsilon < 1$. On the intersection with $\mathcal{E}_{\mathrm{sg}}$, (73) gives, simultaneously for all admissible Ω,

$$\min_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\| \ge \sqrt{(1-\eta)\mu M} - \frac{3}{8}\eta\sqrt{\mu M} \ge (1-\eta)\sqrt{\mu M}.$$

Here the last inequality follows from

$$\frac{\sqrt{1-\eta} - (1-\eta)}{\eta} = \frac{1}{1 + 1/\sqrt{1-\eta}} \ge \frac{1}{1+\sqrt{2}} > \frac{3}{8}.$$

Squaring, using $|\mathcal{N}_\varepsilon| \le (1 + 16c_3/(\eta\sqrt{\mu}))^m$, and including the failure probability of $\mathcal{E}_{\mathrm{sg}}$ proves part (i).

Part (ii): counting tail projections. Now use the threshold and moment from part (ii). For a fixed $\boldsymbol{x} \in \mathbb{S}^{m-1}$, define

$$\mathcal{B}_{\boldsymbol{x}} := \{i \in [M] : |\boldsymbol{a}_i^T \boldsymbol{x}| \geq \gamma\}, \qquad p_{\boldsymbol{x}} := \mathbb{P}(|\boldsymbol{a}_i^T \boldsymbol{x}| \geq \gamma) \geq (1+\eta)\kappa_4.$$

For a Bernoulli sum, the lower-tail Chernoff bound gives

$$\mathbb{P}(|\mathcal{B}_{\boldsymbol{x}}| < \kappa_4 M) \leq \exp\left(-\frac{(p_{\boldsymbol{x}} - \kappa_4)^2 M}{2p_{\boldsymbol{x}}}\right) \leq 2e^{-\eta^2 \kappa_4 M/6}.$$

Indeed, $(p-\kappa_4)^2/p$ is increasing for $p \geq \kappa_4$, and its value at $(1+\eta)\kappa_4$ is $\eta^2\kappa_4/(1+\eta)$. This also explains why the condition in part (ii) must concern the tail probability. On the event $|\mathcal{B}_{\boldsymbol{x}}| \geq \kappa_4 M$, the same ordering argument, now applied to the largest projected magnitudes, yields for every $|\Omega| \leq \kappa_4 M$,

$$\sum_{i \in \Omega} (\boldsymbol{a}_i^T \boldsymbol{x})^2 \leq \sum_{i \in \mathcal{B}_{\boldsymbol{x}}} (\boldsymbol{a}_i^T \boldsymbol{x})^2 = \sum_{i=1}^{M} W_i^{(\boldsymbol{x})}, \tag{80}$$

where $W_i^{(\boldsymbol{x})} := (\boldsymbol{a}_i^T \boldsymbol{x})^2 \mathbf{1}\{|\boldsymbol{a}_i^T \boldsymbol{x}| \geq \gamma\}$.

Part (ii): tail energy and extension from a net. Write $\nu_{\boldsymbol{x}} := \mathbb{E} W_i^{(\boldsymbol{x})} \leq \mu \leq 1$, where the last inequality follows from isotropy. Moreover, $\mu > 0$: if $\gamma > 0$, the stipulated tail probability gives $\nu_{\boldsymbol{x}} \geq \gamma^2(1+\eta)\kappa_4 > 0$; if $\gamma = 0$, isotropy gives $\mu = 1$. Since $0 \leq W_i^{(\boldsymbol{x})} \leq (\boldsymbol{a}_i^T \boldsymbol{x})^2$, the centered variables have uniformly bounded subexponential norms,

$$\|W_i^{(\boldsymbol{x})} - \nu_{\boldsymbol{x}}\|_{\psi_1} \leq C K_{\mathrm{sg}}^2.$$

Bernstein's inequality for independent subexponential variables gives

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{M} W_i^{(\boldsymbol{x})} > (1+\eta)\mu M\right) &\leq \mathbb{P}\left(\sum_{i=1}^{M} (W_i^{(\boldsymbol{x})} - \nu_{\boldsymbol{x}}) > \eta\mu M\right) \\ &\leq 2\exp\left[-cM \min\left\{\frac{\eta^2\mu^2}{K_{\mathrm{sg}}^4}, \frac{\eta\mu}{K_{\mathrm{sg}}^2}\right\}\right] \leq 2e^{-c_4 \eta^2 \mu^2 M},\end{aligned}$$

where c_4 depends only on K_{sg}. The quadratic branch suffices because $K_{\mathrm{sg}} \geq 1$ and $0 < \eta\mu < 1$.

Take an ε-net with $\varepsilon = \eta/3$. A union bound together with (80) shows that, outside an event of probability at most

$$2\left(1 + \frac{6}{\eta}\right)^m \left[e^{-\eta^2 \kappa_4 M/6} + e^{-c_4 \eta^2 \mu^2 M}\right],$$

every admissible Ω satisfies

$$\max_{\boldsymbol{x} \in \mathcal{N}_\varepsilon} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\|^2 \leq (1+\eta)\mu M.$$

For any matrix, the vector-net inequality is

$$\|\boldsymbol{A}_\Omega\| \leq \frac{1}{1-\varepsilon} \max_{\boldsymbol{x} \in \mathcal{N}_\varepsilon} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\|.$$

Consequently,

$$\lambda_{\max}(\boldsymbol{A}_\Omega \boldsymbol{A}_\Omega^T) \leq \frac{1+\eta}{(1-\eta/3)^2} \mu M \leq (1+\eta)^2 \mu M.$$

The last inequality holds for $0 < \eta < 1/2$, since $(1+\eta)(1-\eta/3)^2 \geq 1$. This proves part (ii), including simultaneous validity over all admissible subsets. □

G Proofs for Section 3

G.1 Covariance Decay Implied by Gaussian Strong Mixing

Proposition 1. *Let $\{\boldsymbol{a}_i\}_{i \in \mathbb{Z}} \subseteq \mathbb{R}^m$ be a centered, stationary, jointly Gaussian process with marginal law $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$. If $\alpha(k) \leq C_\alpha \rho_0^k$, then*

$$\|\operatorname{Cov}(\boldsymbol{a}_i, \boldsymbol{a}_j)\| \leq 2\pi C_\alpha \rho_0^{|i-j|}, \qquad i \neq j.$$

The constant is independent of the dimension when C_α and ρ_0 are dimension-independent.

Proof. Fix $i < j$ and $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{S}^{m-1}$. By joint Gaussianity,

$$X := \boldsymbol{u}^T \boldsymbol{a}_i, \qquad Y := \boldsymbol{v}^T \boldsymbol{a}_j$$

form a centered bivariate Gaussian vector with unit variances and correlation

$$r := \mathbb{E}[XY] = \boldsymbol{u}^T \operatorname{Cov}(\boldsymbol{a}_i, \boldsymbol{a}_j)\boldsymbol{v}.$$

The Gaussian quadrant identity states that

$$\mathbb{P}(X \geq 0, Y \geq 0) - \frac{1}{4} = \frac{1}{2\pi} \arcsin r.$$

For completeness, when $|r| < 1$, write $X = G_1$ and $Y = rG_1 + \sqrt{1-r^2}G_2$ with independent standard Gaussian G_1, G_2. Their direction is uniform on the circle. The intersection of the two positive half-planes has angle $\pi/2 + \arcsin r$, giving the displayed probability after division by 2π. The endpoint cases follow by continuity.

The events $\{X \geq 0\}$ and $\{Y \geq 0\}$ belong to the past and future sigma-fields separated by $j - i$. Therefore,

$$\frac{|\arcsin r|}{2\pi} = |\mathbb{P}(X \geq 0, Y \geq 0) - \mathbb{P}(X \geq 0)\mathbb{P}(Y \geq 0)| \leq \alpha(j-i) \leq C_\alpha \rho_0^{j-i}.$$

Since $|r| \leq |\arcsin r|$, we obtain $|r| \leq 2\pi C_\alpha \rho_0^{j-i}$. Taking the supremum over $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{S}^{m-1}$ proves the claim. □

This proposition provides the direction of implication actually used below. We make no converse deduction from pairwise covariance decay to strong mixing of the entire past and future: such a deduction needs additional conditions controlling the joint covariance structure. In particular, a bound on unnormalized lag covariance blocks does not by itself justify a bound on dependence between arbitrary past and future events.

G.2 Proof of Theorem 4

We establish the three probabilistic ingredients separately and then retain the same threshold-set and net argument as in the independent proof.

Bernstein inequalities for the bounded projected quantities. For each fixed $\boldsymbol{x} \in \mathbb{S}^{m-1}$, define

$$Y_i^{(\boldsymbol{x},\gamma)} := \mathbf{1}\{|\boldsymbol{a}_i^T \boldsymbol{x}| \leq \gamma\}, \qquad Z_i^{(\boldsymbol{x},\gamma)} := (\boldsymbol{a}_i^T \boldsymbol{x})^2 \mathbf{1}\{|\boldsymbol{a}_i^T \boldsymbol{x}| \leq \gamma\}.$$

Measurable transformations cannot increase the strong-mixing coefficients, because the sigma-fields they generate are subfields of those generated by the original vectors. Thus, both scalar sequences inherit the bound $C_\alpha \rho_0^k$. Their centered versions are bounded by 1 and γ^2, respectively.

Theorem 1 of [19] gives a Bernstein inequality for centered bounded sequences with geometrically decaying strong-mixing coefficients. Its exponential-mixing formulation also covers $\alpha(k) \leq C_\alpha \rho_0^k$: using $\alpha(k) \leq 1/4$, one can choose a sufficiently small positive exponential rate, depending only on C_α, ρ_0, to absorb the prefactor for all $k \geq 1$. Consequently, for $M \geq 4$ and $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{M} (Y_i^{(\boldsymbol{x},\gamma)} - \mathbb{E}Y_i^{(\boldsymbol{x},\gamma)})\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{M + tL_M}\right), \tag{81}$$

$$\mathbb{P}\left(\left|\sum_{i=1}^{M} (Z_i^{(\boldsymbol{x},\gamma)} - \mathbb{E}Z_i^{(\boldsymbol{x},\gamma)})\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{M\gamma^4 + \gamma^2 tL_M}\right). \tag{82}$$

Here and below the positive constant c can change from one inequality to another and depends only on the mixing parameters. Replacing the original logarithmic factors by $L_M = \log(eM)\log\log(e^e M)$ only weakens the bounds and avoids small-sample logarithmic conventions.

Operator norm of the full data matrix. Joint Gaussianity allows us to control this unbounded quantity using its covariance structure. We cannot apply the bounded-variable Bernstein inequalities directly to the untruncated squares.

Lemma 1. *Let $\{\boldsymbol{a}_i\}_{i\in\mathbb{Z}}$ be a centered, stationary, jointly Gaussian process in $\mathbb{R}^m$ with marginal law $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$ and $\alpha(k) \le C_\alpha \rho_0^k$. There exist $C_0, B \ge 1$ and $c_{\mathrm{op}} > 0$, depending only on C_α, ρ_0, such that, whenever $M \ge C_0 m$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^{M} \boldsymbol{a}_i \boldsymbol{a}_i^T\right\| \le BM\right) \ge 1 - e^{-c_{\mathrm{op}} M}.$$

Proof. By Proposition 1,

$$\|\operatorname{Cov}(\boldsymbol{a}_i, \boldsymbol{a}_j)\| \le 2\pi C_\alpha \rho_0^{|i-j|}, \qquad i \ne j.$$

This uses the mixing assumption in the stated direction and the joint Gaussian law of each pair. The explicit quadrant calculation is given in the preceding subsection.

Fix $\boldsymbol{x} \in \mathbb{S}^{m-1}$ and define

$$\boldsymbol{b}_{\boldsymbol{x}} := (\boldsymbol{a}_1^T \boldsymbol{x}, \dots, \boldsymbol{a}_M^T \boldsymbol{x})^T \in \mathbb{R}^M.$$

Then $\boldsymbol{b}_{\boldsymbol{x}}$ is a centered Gaussian vector with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}} \in \mathbb{R}^{M\times M}$ satisfying

$$(\boldsymbol{\Sigma}_{\boldsymbol{x}})_{ij} = \boldsymbol{x}^T \operatorname{Cov}(\boldsymbol{a}_i, \boldsymbol{a}_j)\boldsymbol{x}.$$

Since $\boldsymbol{a}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$, we have $\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) = M$. Moreover, by the covariance decay just proved,

$$\|\boldsymbol{\Sigma}_{\boldsymbol{x}}\| \le \max_{i\in[M]} \sum_{j=1}^{M} |(\boldsymbol{\Sigma}_{\boldsymbol{x}})_{ij}| \le 1 + 4\pi C_\alpha \sum_{h=1}^{\infty} \rho_0^h =: L.$$

Therefore,

$$\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{x}}^2) \le \|\boldsymbol{\Sigma}_{\boldsymbol{x}}\| \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \le LM.$$

By the Gaussian quadratic-form concentration inequality, for every $t > 0$,

$$\mathbb{P}\left(\|\boldsymbol{b}_{\boldsymbol{x}}\|^2 \ge M + 2\sqrt{LMt} + 2Lt\right) \le e^{-t}.$$

Equivalently,

$$\mathbb{P}\left(\sum_{i=1}^{M} (\boldsymbol{a}_i^T \boldsymbol{x})^2 \ge M + 2\sqrt{LMt} + 2Lt\right) \le e^{-t}.$$

Let $\mathcal{N}$ be a $1/4$-net of $\mathbb{S}^{m-1}$. Then $|\mathcal{N}| \le 9^m$. Taking $t = 2M$ and applying the union bound, we obtain

$$\max_{\boldsymbol{x}\in\mathcal{N}} \sum_{i=1}^{M} (\boldsymbol{a}_i^T \boldsymbol{x})^2 \le C_L M$$

with probability at least $1 - 9^m e^{-2M}$, where $C_L := 1 + 2\sqrt{2L} + 4L$ depends only on C_α, ρ_0. If $M \ge C_0 m$ with $C_0 \ge \log 9$, then $9^m e^{-2M} \le e^{-M}$. Thus we may take $c_{\mathrm{op}} = 1$, or any smaller positive constant.

Finally, by the standard net argument for symmetric matrices,

$$\left\|\sum_{i=1}^{M} \boldsymbol{a}_i \boldsymbol{a}_i^T\right\| \le 2 \max_{\boldsymbol{x}\in\mathcal{N}} \boldsymbol{x}^T \left(\sum_{i=1}^{M} \boldsymbol{a}_i \boldsymbol{a}_i^T\right) \boldsymbol{x}.$$

Therefore,

$$\left\|\sum_{i=1}^{M} \boldsymbol{a}_i \boldsymbol{a}_i^T\right\| \le 2C_L M.$$

Setting $B := 2C_L$ completes the proof. □

Proof of Theorem 4. Let $\boldsymbol{A} \in \mathbb{R}^{m\times M}$ be the matrix whose columns are $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M$, and let $\boldsymbol{A}_\Omega$ be the submatrix whose columns are $\{\boldsymbol{a}_i : i \in \Omega\}$. Define

$$s_\Omega := \min_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\|, \qquad \mathcal{E}_{\text{op}} := \{\|\boldsymbol{A}\|^2 \le BM\}.$$

Then $s_\Omega^2 = \lambda_{\min}(\boldsymbol{A}_\Omega \boldsymbol{A}_\Omega^T)$, including when the subset matrix is rank deficient. By Lemma 1, $\mathbb{P}(\mathcal{E}_{\text{op}}) \ge 1 - e^{-c_{\text{op}} M}$.

Let $\mathcal{N}_\varepsilon$ be an ε-net of $\mathbb{S}^{m-1}$ with

$$|\mathcal{N}_\varepsilon| \le \left(\frac{2}{\varepsilon} + 1\right)^m.$$

For any $\boldsymbol{x} \in \mathbb{S}^{m-1}$, choose $\boldsymbol{x}_0 \in \mathcal{N}_\varepsilon$ such that $\|\boldsymbol{x} - \boldsymbol{x}_0\| \le \varepsilon$. On the event $\mathcal{E}_{\text{op}}$, we have

$$\|\boldsymbol{A}_\Omega^T \boldsymbol{x}\| \ge \|\boldsymbol{A}_\Omega^T \boldsymbol{x}_0\| - \varepsilon \|\boldsymbol{A}_\Omega\| \ge \|\boldsymbol{A}_\Omega^T \boldsymbol{x}_0\| - \varepsilon \|\boldsymbol{A}\| \ge \|\boldsymbol{A}_\Omega^T \boldsymbol{x}_0\| - \varepsilon\sqrt{BM},$$

where the last inequality follows from Lemma 1, under the event $\mathcal{E}_{\text{op}}$. Therefore,

$$s_\Omega \ge \min_{\boldsymbol{x}\in\mathcal{N}_\varepsilon} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\| - \varepsilon\sqrt{BM}.$$

We next prove a uniform lower bound over the net. Fix $\boldsymbol{x} \in \mathcal{N}_\varepsilon$ and define

$$\mathcal{A}_{\boldsymbol{x}} := \left\{i \in [M] : |\boldsymbol{a}_i^T \boldsymbol{x}| \le \gamma\right\}.$$

Since $\boldsymbol{a}_i^T \boldsymbol{x} \sim \mathcal{N}(0,1)$, we have

$$\mathbb{E}|\mathcal{A}_{\boldsymbol{x}}| = M\mathbb{P}(|g| \le \gamma) = (1-\eta)\kappa M.$$

Applying (81) with $t = \eta\kappa M$ yields the next bound, since $L_M \ge 1$ and

$$M + \eta\kappa M L_M \le 2ML_M.$$

Thus,

$$\mathbb{P}\left(|\mathcal{A}_{\boldsymbol{x}}| > \kappa M\right) \le 2\exp\left(-c\eta^2\kappa^2 M_{\text{eff}}\right),$$

where the constant $c > 0$ depends only on C_α, ρ_0. Hence, on the complementary event, $|\mathcal{A}_{\boldsymbol{x}}| \le \kappa M$.

Since $|\Omega| \ge \kappa M$, and $\mathcal{A}_{\boldsymbol{x}}$ consists of the indices with the smallest projected magnitudes below the threshold γ, we have

$$\sum_{i\in\Omega} (\boldsymbol{a}_i^T \boldsymbol{x})^2 \ge \sum_{i\in\mathcal{A}_{\boldsymbol{x}}} (\boldsymbol{a}_i^T \boldsymbol{x})^2 = \sum_{i=1}^M (\boldsymbol{a}_i^T \boldsymbol{x})^2 \mathbf{1}\{|\boldsymbol{a}_i^T \boldsymbol{x}| \le \gamma\}.$$

Moreover,

$$\mathbb{E}\left[(\boldsymbol{a}_i^T \boldsymbol{x})^2 \mathbf{1}\{|\boldsymbol{a}_i^T \boldsymbol{x}| \le \gamma\}\right] = \mu.$$

Applying (82) with $t = \eta\mu M$ gives the next bound. Indeed, $0 < \mu = \mathbb{E}Z_i^{(\boldsymbol{x},\gamma)} \le \gamma^2$, so

$$M\gamma^4 + \gamma^2 \eta\mu M L_M \le 2M\gamma^4 L_M.$$

Therefore,

$$\mathbb{P}\left(\sum_{i=1}^M (\boldsymbol{a}_i^T \boldsymbol{x})^2 \mathbf{1}\{|\boldsymbol{a}_i^T \boldsymbol{x}| \le \gamma\} \le (1-\eta)\mu M\right) \le 2\exp\left(-\frac{c\eta^2\mu^2}{\gamma^4} M_{\text{eff}}\right).$$

Thus, for this fixed $\boldsymbol{x} \in \mathcal{N}_\varepsilon$, with probability at least

$$1 - 2\exp\left(-c\eta^2\kappa^2 M_{\text{eff}}\right) - 2\exp\left(-\frac{c\eta^2\mu^2}{\gamma^4} M_{\text{eff}}\right),$$

we have

$$\|\boldsymbol{A}_\Omega^T \boldsymbol{x}\|^2 = \sum_{i\in\Omega} (\boldsymbol{a}_i^T \boldsymbol{x})^2 \ge (1-\eta)\mu M.$$

Taking a union bound over all $\boldsymbol{x} \in \mathcal{N}_\varepsilon$ gives

$$\min_{\boldsymbol{x}\in\mathcal{N}_\varepsilon} \|\boldsymbol{A}_\Omega^T \boldsymbol{x}\|^2 \ge (1-\eta)\mu M$$

with probability at least

$$1 - 2|\mathcal{N}_\varepsilon| \left[\exp\left(-c\eta^2\kappa^2 M_{\text{eff}}\right) + \exp\left(-\frac{c\eta^2\mu^2}{\gamma^4} M_{\text{eff}}\right)\right].$$

It remains to choose the net radius. Let $\varepsilon := \eta\sqrt{\mu}/(3\sqrt{B})$. Then, we have

$$|\mathcal{N}_\varepsilon| \le \left(\frac{6\sqrt{B}}{\eta\sqrt{\mu}} + 1\right)^m.$$

On the intersection of the above event and $\mathcal{E}_{\text{op}}$, we obtain

$$s_\Omega \ge \sqrt{(1-\eta)\mu M} - \frac{\eta\sqrt{\mu M}}{3}.$$

Since $\eta \in (0, 1/2)$, we have

$$\sqrt{1-\eta} - (1-\eta) = \frac{\eta}{1 + 1/\sqrt{1-\eta}} \ge \frac{\eta}{1+\sqrt{2}} > \frac{\eta}{3}.$$

Therefore,

$$s_\Omega \ge (1-\eta)\sqrt{\mu M}.$$

Squaring both sides yields

$$\lambda_{\min}\left(\sum_{i\in\Omega} \boldsymbol{a}_i \boldsymbol{a}_i^T\right) = s_\Omega^2 \ge (1-\eta)^2 \mu M.$$

The desired probability follows by combining the net event with $\mathbb{P}(\mathcal{E}_{\text{op}}) \ge 1 - e^{-c_{\text{op}} M}$. No independence between these events is needed. All subset comparisons on the net are deterministic once the directional count and energy events hold, so the final event is simultaneous over all admissible Ω. □

H Proofs for Section 4

H.1 Proof Outline

In this subsection, we outline the proof of Theorem 5. We first provide necessary conditions for a global minimizer of (21); their proof is deferred to Appendix H.2.

Proposition 2. *Let $(\hat{\boldsymbol{U}}, \hat{C})$ be a global minimizer of* (21)*. For each $k \in [K]$,*

$$i \in \hat{C}_k \quad \Longrightarrow \quad \|\hat{\boldsymbol{U}}_k^T \boldsymbol{z}_i\| \ge \|\hat{\boldsymbol{U}}_l^T \boldsymbol{z}_i\|, \qquad \forall l \in [K]. \tag{83}$$

The columns of $\hat{\boldsymbol{U}}_k$ span a top-d eigenspace of $\sum_{i\in\hat{C}_k} \boldsymbol{z}_i \boldsymbol{z}_i^T$. If several clusters maximize the projected norm of a point, its label in $\hat{C}$ specifies one maximizing cluster; the maximizing sets themselves need not be disjoint. Thus (83) *is a membership implication, not an equality of each cluster with the full set of maximizers.*

Based on global optimality, we have the following objective comparison; its proof is deferred to Appendix H.3. Write $\hat{\boldsymbol{P}}_k := \hat{\boldsymbol{U}}_k \hat{\boldsymbol{U}}_k^T$ and $\boldsymbol{P}_l^\star := \boldsymbol{U}_l^\star \boldsymbol{U}_l^{\star T}$ for the orthogonal projectors.

Lemma 2. *Under Definition 2 and assumption 1, let $\mathcal{E}_{\text{norm}}$ be the simultaneous event in* (106)*, which has probability at least $1 - 4/N$. On this event, every global minimizer $(\hat{\boldsymbol{U}}, \hat{C})$ of* (21) *satisfies*

$$\sum_{k=1}^K \sum_{l=1}^K \sum_{i\in\hat{C}_k\cap C_l^\star} \left(\|\boldsymbol{a}_i\|^2 - \|\hat{\boldsymbol{U}}_k^T \boldsymbol{U}_l^\star \boldsymbol{a}_i\|^2\right) \le E_0, \tag{84}$$

where $E_0 := 9N(\delta_{\max}\sqrt{nd} + \delta_{\max}^2 n)$ is a deterministic bound.

An important observation is that

$$\sum_{i\in\hat{C}_k\cap C_l^\star}\left(\|\boldsymbol{a}_i\|^2-\|\hat{\boldsymbol{U}}_k^T\boldsymbol{U}_l^\star\boldsymbol{a}_i\|^2\right)=\left\langle\sum_{i\in\hat{C}_k\cap C_l^\star}\boldsymbol{a}_i\boldsymbol{a}_i^T,\boldsymbol{I}_d-\boldsymbol{U}_l^{\star T}\hat{\boldsymbol{P}}_k\boldsymbol{U}_l^\star\right\rangle$$
$$\geq\lambda_{\min}\left(\sum_{i\in\hat{C}_k\cap C_l^\star}\boldsymbol{a}_i\boldsymbol{a}_i^T\right)\left(d-\|\hat{\boldsymbol{U}}_k^T\boldsymbol{U}_l^\star\|_F^2\right). \tag{85}$$

The matrix $\boldsymbol{I}_d-\boldsymbol{U}_l^{\star T}\hat{\boldsymbol{P}}_k\boldsymbol{U}_l^\star$ is positive semidefinite, which justifies the last inequality. As noted in Remark 1, the optimal solution depends on all latent and noise vectors. Consequently, we cannot apply the fixed-subset covariance bound in Fact 1 directly to $\sum_{i\in\hat{C}_k\cap C_l^\star}\boldsymbol{a}_i\boldsymbol{a}_i^T$. This motivates Theorem 1, whose simultaneous event accommodates the data-dependent intersections.

Suppose that, for some $k\in[K]$, distinct $l_1,l_2\in[K]$, and $\kappa_1\in(0,1/K)$,

$$|\hat{C}_k\cap C_{l_1}^\star|\geq\kappa_1 N_{l_1}^\star,\qquad|\hat{C}_k\cap C_{l_2}^\star|\geq\kappa_1 N_{l_2}^\star. \tag{86}$$

Assuming $N_{\min}^\star\geq d$ as in Proposition 3, we apply Theorem 1 within every true cluster. A union bound over the K true clusters gives one event on which

$$\lambda_{\min}\left(\sum_{i\in\hat{C}_k\cap C_l^\star}\boldsymbol{a}_i\boldsymbol{a}_i^T\right)\geq(1-\eta)^2\mu N_l^\star,\qquad l\in\{l_1,l_2\}, \tag{87}$$

whenever (86) holds. Here $\eta\in(0,1/2)$ and $\mu=b((1-\eta)\kappa_1)$ is defined by (4). The total failure probability, including the norm event, is stated in Proposition 3 below. Combining (84), (85), and (87), and using nonnegativity of all terms in (84), yields

$$d-\|\hat{\boldsymbol{U}}_k^T\boldsymbol{U}_{l_1}^\star\|_F^2+d-\|\hat{\boldsymbol{U}}_k^T\boldsymbol{U}_{l_2}^\star\|_F^2\leq\frac{E_0}{(1-\eta)^2\mu N_{\min}^\star}.$$

The orthogonal Procrustes comparison detailed in Appendix H.4 gives

$$d-\|\boldsymbol{U}_{l_1}^{\star T}\boldsymbol{U}_{l_2}^\star\|_F^2\leq 4\left(d-\|\hat{\boldsymbol{U}}_k^T\boldsymbol{U}_{l_1}^\star\|_F^2+d-\|\hat{\boldsymbol{U}}_k^T\boldsymbol{U}_{l_2}^\star\|_F^2\right)\leq\frac{4E_0}{(1-\eta)^2\mu N_{\min}^\star}. \tag{88}$$

Since $d-\|\boldsymbol{U}_{l_1}^{\star T}\boldsymbol{U}_{l_2}^\star\|_F^2\geq(1-\mu_{\max}^2)d$, the scenario in (86) is impossible when the noise level, and thus E_0, is sufficiently small. The following proposition turns this observation into a matching and an error bound.

Proposition 3. *Consider Definition 2 and assumption 1, with $\mu_{\max}<1$ and $N_{\min}^\star\geq d$. Fix $\eta\in(0,1/2)$ and $\kappa_1\in(0,1/K)$, and let $\mu=b((1-\eta)\kappa_1)$ be as in* (4)*. Suppose*

$$\delta_{\max}\leq\frac{\mu(1-\eta)^2(1-\mu_{\max}^2)}{40\kappa_N K}\sqrt{d/n}. \tag{89}$$

With probability at least

$$\begin{aligned}&1-\frac{4}{N}-2K\exp(-2N_{\min}^\star)\\&\quad-2K\left(\frac{22}{\eta\sqrt{\mu}}+1\right)^d\left[\exp\left(-\frac{\eta^2\kappa_1 N_{\min}^\star}{2}\right)+\exp\left(-c_1\eta^2\mu^2 N_{\min}^\star\right)\right],\end{aligned} \tag{90}$$

every global minimizer $(\hat{\boldsymbol{U}},\hat{C})$ admits a permutation $\pi:[K]\to[K]$ such that

$$|\hat{C}_{\pi(k)}\cap C_k^\star|\geq(1-(K-1)\kappa_1)\,N_k^\star,\qquad k\in[K]. \tag{91}$$

This permutation is the unique matching whose paired intersections have size at least $\kappa_1 N_k^\star$. In particular, $\ell(\hat{C},C^\star)\leq(K-1)\kappa_1$.

The proof is given in Appendix H.4. The event depends only on the generated sample, rather than on the choice of a global minimizer or a tie-breaking rule.

H.2 Proof of Proposition 2

Proof. For fixed C, minimizing $F(\boldsymbol{U}, C)$ over each $\boldsymbol{U}_k \in \mathcal{O}^{n\times d}$ is equivalent to maximizing

$$\sum_{i\in C_k} \|\boldsymbol{U}_k^T \boldsymbol{z}_i\|^2 = \left\langle \boldsymbol{U}_k \boldsymbol{U}_k^T, \sum_{i\in C_k} \boldsymbol{z}_i \boldsymbol{z}_i^T \right\rangle.$$

The variational characterization of the sum of the largest d eigenvalues therefore shows that the columns of $\hat{\boldsymbol{U}}_k$ span a top-d eigenspace of the cluster covariance sum. Eigenvalue multiplicities need not make that eigenspace unique.

For fixed $\hat{\boldsymbol{U}}$, suppose $i \in \hat{C}_k$ and another cluster has a strictly larger projected norm for $\boldsymbol{z}_i$. If $|\hat{C}_k| \geq 2$, moving i to that cluster preserves nonemptiness of the partition and strictly decreases the objective, contradicting global optimality. If $\hat{C}_k = \{i\}$, the eigenspace condition and $d \geq 1$ imply $\|\hat{\boldsymbol{U}}_k^T \boldsymbol{z}_i\| = \|\boldsymbol{z}_i\|$: when $\boldsymbol{z}_i \neq \boldsymbol{0}$ its direction is a top eigenvector, and the zero-vector case is immediate. No orthogonal projection can have a larger norm. Thus the same membership implication holds for singleton clusters. Ties cause no difficulty, because the given partition assigns each point to just one maximizing cluster. This proves (83) without changing the feasible set of (21). □

H.3 Proof of Lemma 2

Proof. Let $\mathcal{E}_{\mathrm{norm}}$ denote the simultaneous event in (106). By Lemma 7, $\mathbb{P}(\mathcal{E}_{\mathrm{norm}}) \geq 1 - 4/N$. On this event, Assumption 1 gives, for every $i \in C_l^\star$,

$$\|\boldsymbol{a}_i\| \leq \sqrt{d} + \alpha_N \leq 2\sqrt{d}, \qquad \|\boldsymbol{e}_i\| \leq \delta_l(\sqrt{n} + \alpha_N) \leq 2\delta_l\sqrt{n}. \tag{92}$$

Since $\boldsymbol{U}_k^T \boldsymbol{U}_k = \boldsymbol{I}_d$, we have

$$F(\boldsymbol{U}, C) = \sum_{k=1}^K \sum_{i\in C_k} \left(\|\boldsymbol{z}_i\|^2 - \|\boldsymbol{U}_k^T \boldsymbol{z}_i\|^2 \right).$$

For any global minimizer, $F(\hat{\boldsymbol{U}}, \hat{C}) \leq F(\boldsymbol{U}^\star, C^\star)$. Substituting $\boldsymbol{z}_i = \boldsymbol{U}_l^\star \boldsymbol{a}_i + \boldsymbol{e}_i$ for $i \in C_l^\star$ yields

$$\begin{aligned}
F(\hat{\boldsymbol{U}}, \hat{C}) - F(\boldsymbol{U}^\star, C^\star) &= \sum_{l=1}^K \sum_{i\in C_l^\star} \|\boldsymbol{U}_l^{\star T} \boldsymbol{z}_i\|^2 - \sum_{k=1}^K \sum_{i\in \hat{C}_k} \|\hat{\boldsymbol{U}}_k^T \boldsymbol{z}_i\|^2 \\
&= \sum_{l=1}^K \sum_{i\in C_l^\star} \|\boldsymbol{a}_i + \boldsymbol{U}_l^{\star T} \boldsymbol{e}_i\|^2 \\
&\quad - \sum_{k=1}^K \sum_{l=1}^K \sum_{i\in \hat{C}_k \cap C_l^\star} \|\hat{\boldsymbol{U}}_k^T (\boldsymbol{U}_l^\star \boldsymbol{a}_i + \boldsymbol{e}_i)\|^2 \\
&= \sum_{i=1}^N \|\boldsymbol{a}_i\|^2 - \sum_{k=1}^K \sum_{l=1}^K \sum_{i\in \hat{C}_k \cap C_l^\star} \|\hat{\boldsymbol{U}}_k^T \boldsymbol{U}_l^\star \boldsymbol{a}_i\|^2 - \mathcal{R}(\hat{\boldsymbol{U}}, \hat{C}),
\end{aligned}$$

where the random noise contribution is

$$\begin{aligned}
\mathcal{R}(\hat{\boldsymbol{U}}, \hat{C}) := 2 \sum_{k=1}^K \sum_{l=1}^K \sum_{i\in \hat{C}_k \cap C_l^\star} \langle \boldsymbol{a}_i, \boldsymbol{U}_l^{\star T} \hat{\boldsymbol{P}}_k \boldsymbol{e}_i \rangle - 2 \sum_{l=1}^K \sum_{i\in C_l^\star} \langle \boldsymbol{a}_i, \boldsymbol{U}_l^{\star T} \boldsymbol{e}_i \rangle \\
+ \sum_{k=1}^K \sum_{i\in \hat{C}_k} \|\hat{\boldsymbol{U}}_k^T \boldsymbol{e}_i\|^2 - \sum_{l=1}^K \sum_{i\in C_l^\star} \|\boldsymbol{U}_l^{\star T} \boldsymbol{e}_i\|^2.
\end{aligned} \tag{93}$$

The preceding expansion and global optimality imply

$$\sum_{k=1}^{K}\sum_{l=1}^{K}\left\langle \boldsymbol{I}_d - \boldsymbol{U}_l^{\star T}\hat{\boldsymbol{P}}_k\boldsymbol{U}_l^{\star}, \sum_{i\in\hat{C}_k\cap C_l^{\star}} \boldsymbol{a}_i\boldsymbol{a}_i^T \right\rangle$$
$$= \sum_{i=1}^{N}\|\boldsymbol{a}_i\|^2 - \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i\in\hat{C}_k\cap C_l^{\star}} \|\hat{\boldsymbol{U}}_k^T\boldsymbol{U}_l^{\star}\boldsymbol{a}_i\|^2 \le \mathcal{R}(\hat{\boldsymbol{U}},\hat{C}). \tag{94}$$

To bound this random term uniformly over the candidate subspaces and assignments, combine the terms associated with the same observation. Because $\boldsymbol{U}_l^{\star T}\boldsymbol{P}_l^{\star} = \boldsymbol{U}_l^{\star T}$,

$$\mathcal{R}(\hat{\boldsymbol{U}},\hat{C}) = \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i\in\hat{C}_k\cap C_l^{\star}} \left[2\boldsymbol{a}_i^T\boldsymbol{U}_l^{\star T}(\hat{\boldsymbol{P}}_k - \boldsymbol{P}_l^{\star})\boldsymbol{e}_i + \boldsymbol{e}_i^T(\hat{\boldsymbol{P}}_k - \boldsymbol{P}_l^{\star})\boldsymbol{e}_i\right].$$

For orthogonal projectors, $-\boldsymbol{I}_n \preceq \hat{\boldsymbol{P}}_k - \boldsymbol{P}_l^{\star} \preceq \boldsymbol{I}_n$, so $\|\hat{\boldsymbol{P}}_k - \boldsymbol{P}_l^{\star}\| \le 1$. Hence, on $\mathcal{E}_{\mathrm{norm}}$,

$$\mathcal{R}(\hat{\boldsymbol{U}},\hat{C}) \le 2\sum_{i=1}^{N}\|\boldsymbol{a}_i\|\,\|\boldsymbol{e}_i\| + \sum_{i=1}^{N}\|\boldsymbol{e}_i\|^2$$
$$\le 8N\delta_{\max}\sqrt{nd} + 4N\delta_{\max}^2 n \le 9N(\delta_{\max}\sqrt{nd} + \delta_{\max}^2 n) = E_0. \tag{95}$$

This bound uses only (92), so it holds on the same event for every global minimizer. Combining it with (94) proves the lemma. □

H.4 Proof of Proposition 3

Proof. We first justify the orthogonal Procrustes comparison in (88). For any $\boldsymbol{V},\boldsymbol{W}\in\mathcal{O}^{n\times d}$, write $\sigma_i = \sigma_i(\boldsymbol{V}^T\boldsymbol{W}) \in [0,1]$. The singular-value decomposition gives

$$d - \|\boldsymbol{V}^T\boldsymbol{W}\|_F^2 = \sum_{i=1}^{d}(1-\sigma_i^2),$$
$$\min_{\boldsymbol{Q}\in\mathcal{O}(d)}\|\boldsymbol{V} - \boldsymbol{W}\boldsymbol{Q}\|_F^2 = 2\sum_{i=1}^{d}(1-\sigma_i). \tag{96}$$

Both equalities involve Frobenius norms. Since $1-\sigma_i \le 1-\sigma_i^2 \le 2(1-\sigma_i)$,

$$d - \|\boldsymbol{U}_{l_1}^{\star T}\boldsymbol{U}_{l_2}^{\star}\|_F^2 \le \min_{\boldsymbol{Q}\in\mathcal{O}(d)}\|\boldsymbol{U}_{l_1}^{\star} - \boldsymbol{U}_{l_2}^{\star}\boldsymbol{Q}\|_F^2$$
$$\le 2\min_{\boldsymbol{Q},\boldsymbol{Q}'\in\mathcal{O}(d)}\left(\|\boldsymbol{U}_{l_1}^{\star} - \hat{\boldsymbol{U}}_k\boldsymbol{Q}'\|_F^2 + \|\hat{\boldsymbol{U}}_k\boldsymbol{Q}' - \boldsymbol{U}_{l_2}^{\star}\boldsymbol{Q}\|_F^2\right)$$
$$= 2\min_{\boldsymbol{Q}'\in\mathcal{O}(d)}\|\boldsymbol{U}_{l_1}^{\star} - \hat{\boldsymbol{U}}_k\boldsymbol{Q}'\|_F^2 + 2\min_{\boldsymbol{Q}\in\mathcal{O}(d)}\|\hat{\boldsymbol{U}}_k - \boldsymbol{U}_{l_2}^{\star}\boldsymbol{Q}\|_F^2$$
$$= 4\sum_{i=1}^{d}\left[1-\sigma_i(\boldsymbol{U}_{l_1}^{\star T}\hat{\boldsymbol{U}}_k)\right] + 4\sum_{i=1}^{d}\left[1-\sigma_i(\hat{\boldsymbol{U}}_k^T\boldsymbol{U}_{l_2}^{\star})\right]$$
$$\le 4\left(d - \|\hat{\boldsymbol{U}}_k^T\boldsymbol{U}_{l_1}^{\star}\|_F^2 + d - \|\hat{\boldsymbol{U}}_k^T\boldsymbol{U}_{l_2}^{\star}\|_F^2\right).$$

The second inequality follows by inserting $\hat{\boldsymbol{U}}_k\boldsymbol{Q}'$ and using $\|\boldsymbol{A}+\boldsymbol{B}\|_F^2 \le 2\|\boldsymbol{A}\|_F^2 + 2\|\boldsymbol{B}\|_F^2$. For the next equality, right-orthogonal invariance allows $\boldsymbol{Q}\boldsymbol{Q}'^T$ to range independently over $\mathcal{O}(d)$. This proves the first inequality in (88).

For each $l\in[K]$, apply Theorem 1 to the $N_l^{\star}$ latent vectors in $C_l^{\star}$, in dimension d, using the common parameters η,κ_1. Let $\mathcal{E}_l$ be its simultaneous subset event. On $\cap_{l=1}^{K}\mathcal{E}_l$, every intersection with at least $\kappa_1 N_l^{\star}$ points obeys (87), including intersections chosen after seeing the noise. Intersect this event with $\mathcal{E}_{\mathrm{norm}}$. A union bound, using $N_l^{\star} \ge N_{\min}^{\star}$, gives exactly the probability lower bound (90). No independence between the subset events and the norm event is required.

We now work deterministically on this common event. Suppose the two large intersections in (86) exist. The objective comparison and (85) give the second inequality in (88). Since $\mu \le 1$, $1-\mu_{\max}^2 \le 1$, and $\kappa_N K = N/N^\star_{\min} \ge 1$, (89) implies $\delta_{\max} \le \frac{1}{40}\sqrt{d/n}$. If $\delta_{\max} > 0$, then

$$\begin{aligned} E_0 &= 9N(\delta_{\max}\sqrt{nd} + \delta_{\max}^2 n) \\ &\le 9\left(1+\frac{1}{40}\right) N\delta_{\max}\sqrt{nd} < 10N\delta_{\max}\sqrt{nd} \\ &\le \frac{\mu(1-\eta)^2(1-\mu_{\max}^2)N^\star_{\min} d}{4}. \end{aligned}$$

Consequently,

$$d - \|\boldsymbol{U}_{l_1}^{\star T}\boldsymbol{U}_{l_2}^\star\|_F^2 \le \frac{4E_0}{(1-\eta)^2\mu N^\star_{\min}} < (1-\mu_{\max}^2)d,$$

contradicting the definition of $\mu_{\max}$. If $\delta_{\max} = 0$, then $E_0 = 0$ and the contradiction follows directly from $\mu_{\max} < 1$. Thus every estimated cluster has a large intersection with at most one true cluster.

For each true cluster $C_l^\star$, the disjoint intersections with the K estimated clusters sum to $N_l^\star$. At least one therefore has size at least $N_l^\star/K > \kappa_1 N_l^\star$. Form a bipartite graph with true and estimated clusters as its vertices and an edge whenever $|\hat{C}_k \cap C_l^\star| \ge \kappa_1 N_l^\star$. Every true vertex has degree at least one, while every estimated vertex has degree at most one. There are K vertices on each side, so there are exactly K edges and all vertices have degree one. This gives a unique permutation π such that

$$|\hat{C}_{\pi(l)} \cap C_l^\star| \ge \kappa_1 N_l^\star, \qquad l \in [K].$$

For each $k \ne \pi(l)$, the absence of an edge implies $|\hat{C}_k \cap C_l^\star| < \kappa_1 N_l^\star$. Subtracting these $K-1$ intersections from $N_l^\star$ yields

$$|\hat{C}_{\pi(l)} \cap C_l^\star| \ge [1-(K-1)\kappa_1]N_l^\star.$$

Because $1-(K-1)\kappa_1 > \kappa_1$, this pairing is also the only one satisfying the displayed large-intersection bound. Finally, summing over l and using the definition of ℓ gives $\ell(\hat{C}, C^\star) \le (K-1)\kappa_1$. Every step holds for an arbitrary global minimizer on the same sample event, establishing the simultaneous statement. □

H.5 Proof of Theorem 5

Proof. We fix $\eta = 1/4$, so all constants below are absolute, and choose $\kappa_1 = \tau$. For a sufficiently large C, (27) implies $N^\star_{\min} \ge d$, as needed to apply Theorem 1 within each true cluster. We first retain the small-fraction computation underlying (38). With γ and μ as in (4),

$$\mu = \sqrt{\frac{2}{\pi}}\int_0^\gamma x^2 e^{-x^2/2}\,dx = \frac{1}{3}\sqrt{\frac{2}{\pi}}\gamma^3 + o(\gamma^3),$$

whereas

$$\frac{(1-\eta)\tau}{2} = \frac{1}{\sqrt{2\pi}}\int_0^\gamma e^{-x^2/2}\,dx = \frac{\gamma}{\sqrt{2\pi}} + o(\gamma).$$

It follows that

$$\lim_{\tau\downarrow 0}\frac{\mu}{\tau^3} = \lim_{\gamma\downarrow 0}\frac{\frac{1}{3}\sqrt{2/\pi}\,\gamma^3}{\left(\sqrt{2/\pi}\,\gamma/(1-\eta)\right)^3} = \frac{(1-\eta)^3\pi}{6}. \tag{97}$$

The function $b((1-\eta)\tau)/\tau^3$ is positive and continuous for $\tau \in (0,1]$ and has the positive finite limit in (97). Hence there is an absolute $c_0 > 0$ such that $\mu \ge c_0\tau^3$ throughout this interval. This establishes (38) with a constant uniform over the choices of τ needed here.

For positive noise, (26) and (23) imply

$$\delta_{\max} \le \frac{(1-\mu_{\max}^2)\tau^3}{C\kappa_N K}\sqrt{d/n}.$$

Taking $C \geq 40/[(1-\eta)^2 c_0]$ ensures (89); it is immediate in the zero-noise case. The requirement $\tau < 1/K$ is part of the theorem and is therefore available for the matching argument.

It remains to verify the probability, which cannot be inferred from the SNR condition alone. Since $0 < \tau < 1$ and $\mu \geq c_0 \tau^3$, there are absolute constants $C_1, c_{\mathrm{KSS}} > 0$ such that

$$\log\left(\frac{22}{\eta\sqrt{\mu}} + 1\right) \leq C_1 \log(C_1/\tau), \qquad \min\left\{\frac{\eta^2\tau}{2},\, c_1\eta^2\mu^2\right\} \geq c_{\mathrm{KSS}}\tau^6.$$

Thus the subset failure probability in (90), excluding $4/N$, is at most

$$4K \exp\left[C_1 d \log(C_1/\tau) - c_{\mathrm{KSS}}\tau^6 N^\star_{\min}\right] + 2K\exp(-2N^\star_{\min}). \tag{98}$$

For example, it is enough to require

$$c_{\mathrm{KSS}}\tau^6 N^\star_{\min} \geq C_1 d \log(C_1/\tau) + \log(16K/\rho),$$
$$2N^\star_{\min} \geq \log(8K/\rho).$$

With a sufficiently large absolute C, both follow from (27). Each term in (98) is then at most $\rho/4$. Also $N \geq 8/\rho$ ensures $4/N \leq \rho/2$. The total failure probability in (90) is therefore at most ρ. On its complementary event, Proposition 3 gives $\ell(\hat{C}, C^\star) \leq (K-1)\tau$ for every global minimizer, completing the proof. □

I Proofs For Section 5

I.1 Proof of the robust-regression application

Lemma 3 (Uniform Gaussian fourth moments)**.** *Let $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N$ be independent $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ vectors. If $N \geq C[p + \log(2/\rho)]^2$, then with probability at least $1-\rho$,*

$$\sup_{\|\boldsymbol{v}\|=1} \frac{1}{N}\sum_i (\boldsymbol{v}^T\boldsymbol{h}_i)^4 \leq C. \tag{99}$$

Proof. For a fixed unit $\boldsymbol{v}$, let $\boldsymbol{H}$ have rows $\boldsymbol{h}_i^T$. The map $\boldsymbol{H} \mapsto \|\boldsymbol{H}\boldsymbol{v}\|_4$ is 1-Lipschitz in the Frobenius norm, since the ℓ_4 norm is bounded by the ℓ_2 norm. Jensen's inequality gives $\mathbb{E}\|\boldsymbol{H}\boldsymbol{v}\|_4 \leq (3N)^{1/4}$. Gaussian concentration [13] consequently yields

$$\mathbb{P}\{\|\boldsymbol{H}\boldsymbol{v}\|_4 > (3N)^{1/4} + N^{1/4}\} \leq e^{-\sqrt{N}/2}.$$

Take a union bound over a $1/4$-net of size at most 9^p. For the linear operator norm from ℓ_2^p to ℓ_4^N, $\|\boldsymbol{H}\|_{2\to 4} \leq (1-1/4)^{-1}\max_{\boldsymbol{v}\in\mathcal{N}_{1/4}}\|\boldsymbol{H}\boldsymbol{v}\|_4$. The claimed sample requirement makes $9^p e^{-\sqrt{N}/2} \leq \rho$ and proves (99). □

Proof of Theorem 6. Represent $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$ and $\xi_i = \sigma\zeta_i$, where $\boldsymbol{h}_i = (\boldsymbol{z}_i, \zeta_i) \in \mathbb{R}^{d+1}$ are independent standard Gaussians; this representation also covers $\sigma = 0$. Under (30), Lemma 3 and fact 1 and a union bound give, with probability at least $1-\rho$,

$$\sup_{\|\boldsymbol{v}\|=1} N^{-1}\sum_i (\boldsymbol{v}^T\boldsymbol{h}_i)^4 \leq C, \qquad \left\| N^{-1}\sum_i \boldsymbol{h}_i\boldsymbol{h}_i^T - \boldsymbol{I}_{d+1}\right\| \leq C\sqrt{\varepsilon}. \tag{100}$$

For the covariance estimate, take $\eta = \sqrt{\frac{1}{2}\log(4/\rho)}$ in Fact 1. The term $(d + \log(2/\rho))/\varepsilon$ controls both $2r$ and r^2 in that fact.

Fix any $\boldsymbol{q} \in \mathcal{Q}_\varepsilon$, and put $\Delta_i = q_i - 1/N$. With $a = 4\varepsilon < 1/4$, the cap and $\sum_i q_i = 1$ imply

$$\sum_i |\Delta_i| = 2\sum_{\Delta_i > 0}\Delta_i \leq 2a = 8\varepsilon, \qquad |\Delta_i| \leq 1/N.$$

For the first bound, if $S = \{i : q_i > 1/N\}$ and $t = |S|/N$, its positive excess is at most $\min\{at/(1-a), 1-t\} \le a$. For unit vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^{d+1}$, two applications of Cauchy–Schwarz and (100) now give

$$\left|\sum_i \Delta_i (\boldsymbol{v}^T \boldsymbol{h}_i)(\boldsymbol{w}^T \boldsymbol{h}_i)\right| \le \left(\sum_i |\Delta_i|\right)^{1/2} \left(\sum_i |\Delta_i| (\boldsymbol{v}^T \boldsymbol{h}_i)^2 (\boldsymbol{w}^T \boldsymbol{h}_i)^2\right)^{1/2}$$
$$\le \sqrt{8\varepsilon} \left[\left(N^{-1} \sum_i (\boldsymbol{v}^T \boldsymbol{h}_i)^4\right)\left(N^{-1} \sum_i (\boldsymbol{w}^T \boldsymbol{h}_i)^4\right)\right]^{1/4} \le C\sqrt{\varepsilon}.$$

Thus, simultaneously for all admissible weights,

$$\left\| \sum_i q_i \boldsymbol{h}_i \boldsymbol{h}_i^T - \boldsymbol{I}_{d+1} \right\| \le C\sqrt{\varepsilon}. \tag{101}$$

The upper-left block gives (31), and the last diagonal entry gives (33).

For any $\boldsymbol{v} = \boldsymbol{\theta} - \boldsymbol{\theta}^\star$, set $\boldsymbol{w} = (\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}, -\sigma)$. Then $\boldsymbol{x}_i^T \boldsymbol{v} - \xi_i = \boldsymbol{h}_i^T \boldsymbol{w}$ and $\boldsymbol{\Sigma}^{-1/2} \boldsymbol{g}_i(\boldsymbol{\theta}) = \boldsymbol{z}_i (\boldsymbol{h}_i^T \boldsymbol{w})$. Taking the first d coordinates of (101) applied to $\boldsymbol{w}$ proves (32). For any $\boldsymbol{t} \in \mathbb{R}^d$, write $\boldsymbol{a} = (\boldsymbol{\Sigma}^{1/2}\boldsymbol{t}, 0)$ and use the cap and (99):

$$\sum_i q_i (\boldsymbol{t}^T \boldsymbol{g}_i(\boldsymbol{\theta}))^2 = \sum_i q_i (\boldsymbol{a}^T \boldsymbol{h}_i)^2 (\boldsymbol{w}^T \boldsymbol{h}_i)^2$$
$$\le \frac{1}{1-4\varepsilon} \left[N^{-1} \sum_i (\boldsymbol{a}^T \boldsymbol{h}_i)^4\right]^{1/2} \left[N^{-1} \sum_i (\boldsymbol{w}^T \boldsymbol{h}_i)^4\right]^{1/2}$$
$$\le C(\boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t})(\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}}^2 + \sigma^2).$$

This proves the first inequality in (33); subtracting the positive semidefinite outer product of the weighted mean gives the centered covariance claim.

Finally, for the saturation statement, let $r_G = \sum_{i \in G}(1/N - w_i)$ and $r_B = \sum_{i \in B}(1/N - w_i)$. Since $r_G \le r_B + \varepsilon \le 2\varepsilon$, $\sum_{i \in G} w_i = |G|/N - r_G \ge 1 - 3\varepsilon$. The normalized clean weights therefore belong to $\mathcal{Q}_\varepsilon$. The event above is uniform over these weights and all parameters, so adversarial choice of G creates no additional independence requirement. □

J AUXILIARY LEMMAS

In this section, we present some concentration inequalities for random vectors, which play an important role in our analysis. We first present a tail bound for the weighted sum of squared Gaussian random variables from its mean, which is proved in [12, Lemma 1].

Lemma 4. *Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ be a Gaussian random vector and $\lambda_1, \ldots, \lambda_d \ge 0$ be constants, with $\lambda_{\max} := \max_i \lambda_i > 0$. It holds for any $t > 0$ that*

$$\mathbb{P}\left(\left|\sum_{i=1}^d \lambda_i^2 (x_i^2 - 1)\right| \ge 2\sqrt{\sum_{i=1}^d \lambda_i^4 t + 2\lambda_{\max}^2 t}\right) \le 2\exp(-t).$$

Then, we present a lemma that provides a concentration inequality on the sum of Lipschitz functions of Gaussian random variables, which can be found in [13].

Lemma 5. *Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ be a Gaussian random vector and $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then, we have*

$$\mathbb{P}(|f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{x})]| \ge t) \le 2\exp\left(-\frac{t^2}{2L^2}\right), \ \forall t \ge 0.$$

Based on the above lemma, we present a probabilistic bound on the deviation of the norm of a weighted sum of squared Gaussian random variables from its mean. This is an extension of [15, Lemma 7].

Lemma 6. *Let $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ be a Gaussian random vector and $\lambda_1, \ldots, \lambda_d \geq 0$ be constants, with $\lambda_{\max} := \max_i \lambda_i > 0$. It holds for any $t > 0$ that*

$$\mathbb{P}\left(\left|\sqrt{\sum_{i=1}^{d} \lambda_i^2 x_i^2} - \sqrt{\sum_{i=1}^{d} \lambda_i^2}\right| \geq t + 2\lambda_{\max}\right) \leq 2\exp\left(-\frac{t^2}{2\lambda_{\max}^2}\right), \tag{102}$$

where $\lambda_{\max} = \max\{\lambda_i : i \in [d]\}$. If all weights are zero, the norm is identically zero and the claimed deviation event is empty for $t > 0$.

Proof. We define $f(\boldsymbol{x}) = \sqrt{\sum_{i=1}^d \lambda_i^2 x_i^2}/\sqrt{\sum_{i=1}^d \lambda_i^2}$. The triangle inequality first shows that this weighted norm is Lipschitz. At points where it is differentiable, direct calculation gives

$$\|\nabla f(\boldsymbol{x})\| = \frac{1}{\sqrt{\sum_{i=1}^d \lambda_i^2}} \sqrt{\frac{\sum_{i=1}^d \lambda_i^4 x_i^2}{\sum_{i=1}^d \lambda_i^2 x_i^2}} \leq \frac{\lambda_{\max}}{\sqrt{\sum_{i=1}^d \lambda_i^2}},$$

where $\lambda_{\max} = \max\{\lambda_i : i \in [d]\}$. Applying Lemma 5 to $f(\boldsymbol{x})$ yields that

$$\mathbb{P}\left(|f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{x})]| \geq \frac{t}{\sqrt{\sum_{i=1}^d \lambda_i^2}}\right) \leq 2\exp\left(-\frac{t^2}{2\lambda_{\max}^2}\right).$$

This implies

$$\mathbb{P}\left(\left|\sqrt{\sum_{i=1}^{d} \lambda_i^2 x_i^2} - \mathbb{E}\left[\sqrt{\sum_{i=1}^{d} \lambda_i^2 x_i^2}\right]\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\lambda_{\max}^2}\right). \tag{103}$$

We first note that

$$\mathbb{E}\left[\sqrt{\sum_{i=1}^{d} \lambda_i^2 x_i^2}\right] \leq \sqrt{\mathbb{E}\left[\sum_{i=1}^{d} \lambda_i^2 x_i^2\right]} = \sqrt{\sum_{i=1}^{d} \lambda_i^2}. \tag{104}$$

By letting $X := \sqrt{\sum_{i=1}^d \lambda_i^2 x_i^2} \geq 0$ and $\mu_X := \mathbb{E}[X]$, we compute

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}\left[(X - \mu_X)^2\right] &= \int_0^\infty x^2 d\mathbb{P}\left(|X - \mu_X| \leq x\right) = -\int_0^\infty x^2 d\mathbb{P}\left(|X - \mu_X| > x\right) \\ &= \int_0^\infty 2x\mathbb{P}\left(|X - \mu_X| > x\right) dx \leq \int_0^\infty 4x\exp\left(-\frac{x^2}{2\lambda_{\max}^2}\right) dx = 4\lambda_{\max}^2,\end{aligned}$$

where the fourth equality and the last one follow from integration by parts and the inequality is due to (103). Thus, we have

$$\mu_X^2 = \mathbb{E}\left[X^2\right] - \mathrm{Var}(X) = \mathbb{E}\left[\sum_{i=1}^{d} \lambda_i^2 x_i^2\right] - \mathrm{Var}(X) \geq \sum_{i=1}^{d} \lambda_i^2 - 4\lambda_{\max}^2.$$

Suppose that $\sum_{i=1}^d \lambda_i^2 \geq 4\lambda_{\max}^2$. This implies

$$\mathbb{E}[X] \geq \sqrt{\sum_{i=1}^{d} \lambda_i^2 - 4\lambda_{\max}^2} \geq \sqrt{\sum_{i=1}^{d} \lambda_i^2} - 2\lambda_{\max}. \tag{105}$$

Plugging (104) and (105) into (103) yields that (102) holds when $\sum_{i=1}^d \lambda_i^2 \geq 4\lambda_{\max}^2$.

On the other hand, suppose that $\sum_{i=1}^d \lambda_i^2 < 4\lambda_{\max}^2$. Write $s = (\sum_{i=1}^d \lambda_i^2)^{1/2}$. Since $X \geq 0$, Jensen's inequality in (104) gives

$$0 \leq s - \mathbb{E}[X] \leq s < 2\lambda_{\max}.$$

Consequently,

$$\{|X - s| \geq t + 2\lambda_{\max}\} \subseteq \{|X - \mathbb{E}[X]| \geq t\}.$$

Applying (103) proves (102) in this remaining case. This direct argument does not rescale the weights: a common rescaling leaves the ratio $\sum_i \lambda_i^2/\lambda_{\max}^2$ unchanged and therefore cannot move this case into the preceding one. □

To begin, we present a lemma that estimates the norms of Gaussian random vectors, the inner product of Gaussian random vectors, and the product of a matrix by a Gaussian random vector.

Lemma 7. *Consider the setting of the LRGMM in Definition 2. Let* $\alpha_N := 2\sqrt{\log N} + 2$. *The following statements hold:*
(i) For all $i \in C_k^\star$ *and all* $k \in [K]$, *it holds with probability at least* $1 - 4N^{-1}$ *that*

$$\left| \|\boldsymbol{a}_i\| - \sqrt{d} \right| \leq \alpha_N, \quad \left| \|\boldsymbol{e}_i\| - \delta_k \sqrt{n} \right| \leq \delta_k \alpha_N. \tag{106}$$

(ii) For all $i \in C_k^\star$ *and all* $k \in [K]$, *it holds with probability at least* $1 - 4N^{-1}$ *that*

$$\left\| \boldsymbol{U}_k^\star \boldsymbol{a}_i \boldsymbol{e}_i^T \right\| \leq \delta_k (\sqrt{d} + \alpha_N)(\sqrt{n} + \alpha_N). \tag{107}$$

(iii) Fix a deterministic $\boldsymbol{V} \in \mathcal{O}^{n \times d}$, *independent of the observations. For all* $i \in C_k^\star$ *and all* $k \in [K]$, *it holds with probability at least* $1 - 2N^{-1}$ *that*

$$\left| \|\boldsymbol{V}^T \boldsymbol{U}_k^\star \boldsymbol{a}_i\| - \|\boldsymbol{V}^T \boldsymbol{U}_k^\star\|_F \right| \leq \alpha_N. \tag{108}$$

Proof. According to Definition 2 and $i \in C_k^\star$, we have $\boldsymbol{z}_i = \boldsymbol{U}_k^\star \boldsymbol{a}_i + \boldsymbol{e}_i$, where $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ is independent of $\boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \delta_k^2 \boldsymbol{I}_n)$.

(i) Applying Lemma 6 to $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, together with setting $t = 2\sqrt{\log N}$ and $\lambda_j = 1$ for all $j \in [d]$, yields

$$\mathbb{P}\left(\left| \|\boldsymbol{a}_i\| - \sqrt{d} \right| \geq 2\sqrt{\log N} + 2 \right) \leq 2N^{-2}.$$

This implies that $\left| \|\boldsymbol{a}_i\| - \sqrt{d} \right| \leq \alpha_N$ holds for all $i \in C_k^\star$ and $k \in [K]$ with probability at least $1 - 2N^{-1}$. Similarly, we obtain that $|\|\boldsymbol{e}_i\| - \delta_k \sqrt{n}| \leq \delta_k \alpha_N$ holds for all $i \in C_k^\star$ and $k \in [K]$ with probability at least $1 - 2N^{-1}$. Finally, using the union bound, we prove (106).

(ii) On the event in part (i), since $\boldsymbol{U}_k^\star$ has orthonormal columns,

$$\left\| \boldsymbol{U}_k^\star \boldsymbol{a}_i \boldsymbol{e}_i^T \right\| \leq \|\boldsymbol{U}_k^\star \boldsymbol{a}_i\| \, \|\boldsymbol{e}_i\| = \|\boldsymbol{a}_i\| \, \|\boldsymbol{e}_i\| \leq \delta_k (\sqrt{d} + \alpha_N)(\sqrt{n} + \alpha_N).$$

The same union-bound probability as in part (i) gives the displayed claim.

(iii) Let $\boldsymbol{V}^T \boldsymbol{U}_k^\star = \boldsymbol{P} \boldsymbol{\Sigma} \boldsymbol{Q}^T$ be a singular value decomposition of $\boldsymbol{V}^T \boldsymbol{U}_k^\star$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ with the diagonal elements $0 \leq \sigma_d \leq \ldots \sigma_1 \leq 1$ being the singular values of $\boldsymbol{V}^T \boldsymbol{U}_k^\star$, $\boldsymbol{P} \in \mathcal{O}^{d \times d}$, and $\boldsymbol{Q} \in \mathcal{O}^{d \times d}$. This, together with the orthogonal invariance of the Gaussian distribution, yields

$$\|\boldsymbol{V}^T \boldsymbol{U}_k^\star \boldsymbol{a}_i\| = \|\boldsymbol{\Sigma} \boldsymbol{Q}^T \boldsymbol{a}_i\| \overset{d}{=} \|\boldsymbol{\Sigma} \boldsymbol{a}_i\| = \sqrt{\sum_{j=1}^d \sigma_j^2 a_{ij}^2}. \tag{109}$$

If $\sigma_1 = 0$, both norms in (108) vanish and the claim is deterministic. Otherwise, using Lemma 6 with $t = 2\sigma_1 \sqrt{\log N}$ and $\lambda_j = \sigma_j \leq 1$ for all j yields

$$\mathbb{P}\left(\left| \|\boldsymbol{V}^T \boldsymbol{U}_k^\star \boldsymbol{a}_i\| - \|\boldsymbol{V}^T \boldsymbol{U}_k^\star\|_F \right| \geq \sigma_1 \alpha_N \right) = \mathbb{P}\left(\left| \sqrt{\sum_{j=1}^d \sigma_j^2 a_{ij}^2} - \sqrt{\sum_{j=1}^d \sigma_j^2} \right| \geq \sigma_1 \alpha_N \right) \leq 2N^{-2}.$$

This, together with $\sigma_1 \leq 1$ and the union bound over the N observations, yields (108) for the fixed matrix $\boldsymbol{V}$. It does not assert uniformity over data-dependent choices of $\boldsymbol{V}$; the uniform subset theorem is used for those choices in the clustering proof. □