Fast and Robust: Computationally Efficient Covariance Estimation for Sub-Weibull Vectors

Even He

Abstract

High-dimensional covariance estimation is notoriously sensitive to outliers. While statistically optimal estimators exist for general heavy-tailed distributions, they often rely on computationally expensive techniques like semidefinite programming or iterative M-estimation ($O(d^3)$). In this work, we target the specific regime of **Sub-Weibull distributions** (characterized by stretched exponential tails $\exp(-t^\alpha)$). We investigate a computationally efficient alternative: the **Cross-Fitted Norm-Truncated Estimator**. Unlike element-wise truncation, our approach preserves the spectral geometry while requiring $O(Nd^2)$ operations, which represents the theoretical lower bound for constructing a full covariance matrix. Although spherical truncation is geometrically suboptimal for anisotropic data, we prove that within the Sub-Weibull class, the exponential tail decay compensates for this mismatch. Leveraging weighted Hanson-Wright inequalities, we derive non-asymptotic error bounds showing that our estimator recovers the optimal sub-Gaussian rate $\tilde{O}(\sqrt{r(\Sigma)/N})$ with high probability. This provides a scalable solution for high-dimensional data that exhibits tails heavier than Gaussian but lighter than polynomial decay.

1 Introduction

The estimation of covariance matrices is a cornerstone of multivariate statistics, underpinning fundamental tasks from Principal Component Analysis (PCA) to Linear Discriminant Analysis (LDA). In the classical regime, where data is assumed to be Gaussian or Sub-Gaussian, the empirical sample covariance matrix $\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} X_i X_i^\top$ performs optimally. However, modern high-dimensional data frequently exhibits **heavy tails**, violating these sharp concentration assumptions. In such settings, the sample covariance is well-known to be brittle: a single outlier can arbitrarily corrupt the estimator's spectral properties, and the convergence rate degrades significantly [4, 8].

To mitigate this instability, a rich literature on **robust covariance estimation** has emerged. The theoretical gold standard is currently held by M-estimators [8, 11] and depth-based estimators like the Geometric Median [9]. These methods achieve optimal sub-Gaussian error rates even under heavy-tailed noise. More recently, Ke et al. [10] proposed spectrum-wise truncation methods that respect the matrix geometry by truncating eigenvalues.

However, these statistically optimal estimators come with a steep computational price. M-estimators typically require solving implicit equations via iterative optimization or Semidefinite Programs (SDP). Spectrum-wise truncation involves full Singular Value Decomposition (SVD). Consequently, the computational complexity of these methods scales as $O(Nd^2)$ or even $O(d^3)$, rendering them computationally intractable for very high-dimensional applications (e.g., gene expression data or financial tick data where $d \gg 10^3$). On the other end of the spectrum lie computationally cheap heuristics, such as element-wise truncation [13, 14]. While these methods scale linearly with the data size $O(Nd^2)$, they often fail to preserve the positive semi-definite (PSD) structure of the covariance matrix, leading to invalid covariance estimates and suboptimal spectral norm bounds.

This dichotomy presents a fundamental trade-off: **statistical optimality versus computational efficiency**. In this paper, we demonstrate that this trade-off is not as severe as previously thought for a broad class of heavy-tailed distributions. We revisit the simple **Norm-Truncated Estimator**, which projects data points onto a Euclidean ball. Historically, this approach was considered geometrically coarse for **anisotropic** data (where eigenvalues of Σ decay rapidly). The fundamental limitation is that a spherical truncation radius large enough to preserve signal in the dominant directions necessarily leaves the minor eigenspaces vulnerable, as it fails to filter outliers that are small in Euclidean norm but statistically significant relative to the small eigenvalues.

1.1 Scope of Analysis: The Sub-Weibull Regime

It is important to clarify the class of distributions considered in this work. "Heavy-tailed" is a broad term that often encompasses distributions with polynomial tails (e.g., Pareto laws with $P(|X| > t) \propto t^{-\beta}$), which may lack finite higher-order moments. Covariance estimation in such regimes typically requires specialized techniques like Median-of-Means (MOM) tournaments [5].

In contrast, this paper focuses on the class of **Sub-Weibull distributions** [3], characterized by tails decaying as $\exp(-t^\alpha)$ for $\alpha > 0$.

- When $\alpha \geq 2$, this recovers the classical Sub-Gaussian regime.
- When $\alpha \in (0, 1)$, the tails are strictly heavier than Exponential (Sub-Exponential), allowing for significant outliers while retaining finite moments of all orders.

This regime, often referred to as "stretched exponential," models a wide array of real-world phenomena (e.g., financial returns, wireless fading channels) that exhibit extreme values but do not follow power laws. Our contribution is to demonstrate that within this *intermediate* heavy-tailed regime, simple norm-based truncation is sufficient to restore sub-Gaussian concentration rates, **avoiding the computational bottlenecks (e.g., SVD or iterative solvers) typical of general robust estimators.**

1.2 Our Contributions

We challenge this conventional wisdom by analyzing random vectors with moderate anisotropy with **Sub-Weibull** tails [3], a class that interpolates between Sub-Gaussian and heavy-tailed regimes. Our main contributions are:

1. **Computationally Efficient Algorithm ($O(Nd^2)$).** We propose a **Split-Sample Estimator** that employs a data-driven truncation threshold. By using independent samples to estimate the truncation radius via a robust median proxy, we avoid the circular dependency inherent in naive truncation. The algorithm requires only simple vector norm computations and a single pass of matrix multiplication, achieving a complexity of $O(Nd^2)$, which is orders of magnitude faster than SVD-based or iterative estimators.

2. **Theoretical Guarantees via Weighted Hanson-Wright.** The analysis of quadratic forms of heavy-tailed vectors is non-trivial. Leveraging recent breakthroughs in weighted Hanson-Wright inequalities by **Sambale (2020)** [1] and **Götze et al. (2021)** [2], we provide a rigorous characterization of the estimator's bias and variance. We prove that the "geometric mismatch" of spherical truncation is effectively suppressed by the Sub-Weibull tail decay. Specifically, we show that our estimator $\hat{\Sigma}_{\text{split}}$ satisfies:

$$\|\hat{\Sigma}_{\text{split}} - \Sigma\|_{\text{op}} \lesssim \|\Sigma\|_{\text{op}} \sqrt{\frac{r(\Sigma) \log N}{N}},$$

recovering the optimal sub-Gaussian rate with high probability.

3. **Implications for Robust PCA.** As a direct corollary, we apply the Davis-Kahan $\sin\Theta$ theorem to show that our estimator yields consistent eigenvector recovery. This suggests that the proposed method can serve as a highly efficient initialization or replacement for Robust PCA in time-critical settings.

2 Preliminaries: The Functional Analytic Framework

To rigorously analyze the geometry of heavy-tailed random vectors, we must first establish the appropriate functional setting. We work within the framework of Orlicz spaces, which provides a unified language to quantify tail decay ranging from sub-gaussian to heavy-tailed distributions.

2.1 Orlicz Spaces and Sub-Weibull Distributions

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For a parameter $\alpha > 0$, consider the Young function $\psi_\alpha(x) := \exp(x^\alpha) - 1$. The **Sub-Weibull space** $\mathcal{L}_{\psi_\alpha}$ is the collection of all random variables X with finite ψ_α-norm:

$$\|X\|_{\psi_\alpha} := \inf\left\{C > 0 : \mathbb{E}\exp\left(\left|\frac{X}{C}\right|^\alpha\right) \leq 2\right\}. \tag{1}$$

This scale of spaces interpolates between the well-behaved sub-gaussian regime ($\alpha = 2$) and the heavy-tailed regime ($\alpha < 1$). While $\mathcal{L}_{\psi_\alpha}$ forms a Banach space for $\alpha \geq 1$, it is a quasi-Banach space for $\alpha \in (0, 1)$, as the triangle inequality holds only up to a constant factor depending on α. This structural deficiency, however, does not impede our concentration analysis, as we rely primarily on the equivalent characterization via moments.

Proposition 2.1 (Equivalence of Tails and Moments)**.** *Let X be a random variable and $\alpha \in (0, \infty)$. The condition $\|X\|_{\psi_\alpha} \leq K$ is equivalent, up to absolute constants, to the growth condition on the moments:*

$$(\mathbb{E}|X|^p)^{1/p} \leq CKp^{1/\alpha} \quad \textit{for all } p \geq 1. \tag{2}$$

Proof. This is a classical result derived from the integration of tail probabilities. A detailed exposition for the Sub-Weibull case can be found in [3]. □

Remark 2.2 (The Cost of Heavy Tails)**.** The exponent $1/\alpha$ in Proposition 2.1 quantifies the "cost" of the tail. For sub-gaussian variables ($\alpha = 2$), moments grow as $\sqrt{p}$. For Sub-Weibull variables with $\alpha < 1$, moments grow super-linearly in p. This rapid growth of high-order moments is the algebraic signature of the geometric breakdown of the thin shell.

2.2 Quadratic Mapping

The Euclidean norm entails squaring the coordinates. A pivotal structural property of Orlicz spaces is their behavior under polynomial maps. The following lemma allows us to transport the concentration problem from the space of quadratic forms back to the linear space of independent summands.

Lemma 2.3 (The Square Property)**.** *Let $\alpha \in (0, 2]$. A random variable X satisfies $\|X\|_{\psi_\alpha} < \infty$ if and only if its square satisfies $\|X^2\|_{\psi_{\alpha/2}} < \infty$. Specifically,*

$$\|X^2\|_{\psi_{\alpha/2}} = \|X\|_{\psi_\alpha}^2. \tag{3}$$

Consequently, if the coordinates of a vector lie in $\mathcal{L}_{\psi_\alpha}$ with $\alpha < 1$, their squares lie in $\mathcal{L}_{\psi_{\alpha/2}}$, a space with even heavier tails (specifically, lighter than polynomial but heavier than Weibull).

3 Methodology: The Split-Sample Norm-Truncated Estimator

In this section, we formalize our estimation procedure. While the geometric intuition of truncating outliers is straightforward, the primary challenge lies in calibrating the truncation threshold R without prior knowledge of the covariance structure $\mathrm{tr}(\Sigma)$. We address this via a split-sample approach that decouples the estimation of the scale from the estimation of the matrix structure.

3.1 The Algorithm

Let $\mathcal{D} = \{X_1, \dots, X_N\}$ be an i.i.d. sample from a centered Sub-Weibull distribution in $\mathbb{R}^d$. We randomly partition the dataset into two disjoint subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ of size $N_1 = \lfloor N/2 \rfloor$ and $N_2 = N - N_1$, respectively. The estimation proceeds in two stages: **Robust Scale Estimation** and **Truncated Covariance Construction**.

Algorithm 1 Split-Sample Norm-Truncated Estimator

Input: Dataset $\mathcal{D} = \{X_i\}_{i=1}^N$, Tuning constant $C > 0$, Sub-Weibull parameter α.
Step 1: Data Splitting
Randomly split $\mathcal{D}$ into $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $|\mathcal{S}_1| \approx |\mathcal{S}_2|$.
Step 2: Robust Scale Estimation (on $\mathcal{S}_1$)
Compute norms: $y_i = \|X_i\|_2$ for all $i \in \mathcal{S}_1$.
Compute sample median: $\hat{\theta} = \mathrm{Median}\left(\{y_i : i \in \mathcal{S}_1\}\right)$.
Set adaptive truncation radius: $\hat{R} = C \cdot \hat{\theta} \cdot (\log N)^{1/\alpha}$.
Step 3: Truncated Estimation (on $\mathcal{S}_2$)
Define the set of kept indices: $\mathcal{I} = \{j \in \mathcal{S}_2 : \|X_j\|_2 \le \hat{R}\}$.
Compute the sample covariance of the retained subset:

$$\hat{\Sigma}_{\mathrm{split}} = \begin{cases} \frac{1}{|\mathcal{I}|} \sum_{j \in \mathcal{I}} X_j X_j^\top & \text{if } |\mathcal{I}| > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Output: $\hat{\Sigma}_{\mathrm{split}}$.

Remark 3.1 (Independence)**.** Crucially, the threshold $\hat{R}$ is $\sigma(\mathcal{S}_1)$-measurable and thus independent of the samples in $\mathcal{S}_2$. This independence allows us to treat $\hat{R}$ as fixed when analyzing the concentration on $\mathcal{S}_2$, simplifying the theoretical analysis significantly compared to 'plug-in' estimators.

Remark 3.2. In Step 2, we estimate the typical scale of the distribution using the sample median of the norms. This approach is motivated by the classical Median Absolute Deviation (MAD) estimator [6, 7], which is renowned for its resilience to outliers (achieving a breakdown point of 0.5) compared to the standard deviation. While the sample mean of squared norms would correspond to $\mathrm{tr}(\Sigma)$, it is highly sensitive to heavy tails; the median provides a stable proxy $\hat{\theta}$ that satisfies $\hat{\theta} \asymp \sqrt{\mathrm{tr}(\Sigma)}$ with high probability.

3.2 Computational Complexity Analysis

A central contribution of this work is the efficiency of Algorithm 1. We contrast our approach with state-of-the-art robust estimators in terms of time complexity and memory requirements.

Complexity Breakdown.

1. **Norm Calculation:** Computing $\|X_i\|_2$ for $i \in \mathcal{S}_1$ requires $O(Nd)$ operations.
2. **Scale Estimation:** Finding the median of $N/2$ scalars can be done in $O(N)$ time using the Quickselect algorithm.
3. **Matrix Construction:** The summation in Step 3 requires $O(Nd^2)$ operations in the worst case (dense data).

The total complexity is dominated by the matrix construction, $O(Nd^2)$. While this matches the complexity of the standard (non-robust) sample covariance, it is significantly lower than robust alternatives.

Comparison with Competitors. Table 1 highlights the computational advantage.

- **M-estimators (e.g., Catoni, Geometric Median):** These usually require iterative procedures (e.g., Weiszfeld's algorithm or gradient descent). If T is the number of iterations required for convergence, the cost is $O(T \cdot Nd^2)$. In high dimensions, T can be substantial.
- **Spectrum-wise Truncation:** Methods like those in [10] require computing the Singular Value Decomposition (SVD) of the sample covariance or data matrix, incurring a cost of $O(\min(Nd^2, d^3))$. For $d > 10^4$, the $O(d^3)$ term becomes prohibitive.

Table 1: Computational Complexity of Covariance Estimators (N samples, dimension d).

Estimator	**Complexity**	**Iterative?**	**Preserves PSD?**	**Robustness**
Sample Covariance	$O(Nd^2)$	No	Yes	None
Element-wise Trunc.	$O(Nd^2)$	No	**No**	Medium
Geometric Median	$O(T \cdot Nd^2)$	**Yes**	Yes	High
Spectrum Truncation	$O(Nd^2 + d^3)$	No	Yes	High
Ours (Split-Norm)	$\mathbf{O(Nd^2)}$	**No**	**Yes**	**High**

The total complexity is dominated by the matrix construction step, $O(Nd^2)$. While this matches the complexity of the standard sample covariance, it offers a significant improvement over spectrum-wise truncation methods that scale as $O(d^3)$ due to SVD. Our method hits the computational lower bound for explicit covariance estimation.

3.3 Restoring Efficiency via Cross-Fitting

The simple split-sample approach (Algorithm 1) discards half the data for covariance estimation, potentially increasing the finite-sample variance. To overcome this inefficiency while maintaining the statistical independence required for our theoretical analysis, we propose a **Cross-Fitting** strategy.

Theoretical Justification. Analyzing the joint distribution of $\hat{\Sigma}^{(1)}$ and $\hat{\Sigma}^{(2)}$ is non-trivial due to the complex dependency introduced by the cross-validation structure. However, a worst-case bound follows immediately from the union bound. Let $\mathcal{E}$ be the high-probability event under which Algorithm 1 succeeds (Theorem 4.5). Since $\hat{\Sigma}^{(1)}$ and $\hat{\Sigma}^{(2)}$ are marginally distributed as the split-sample estimator, they each satisfy the error bound with probability $1 - \delta$. By the union

Algorithm 2 Cross-Fitted Norm-Truncated Estimator

1: **Step 1:** Randomly split dataset $\mathcal{D}$ into two equal folds $\mathcal{S}_1$ and $\mathcal{S}_2$.
2: **Step 2 (Fold 1):** Use $\mathcal{S}_1$ to compute median norm $\hat{\theta}_1$ and threshold $\hat{R}_1$. Truncate samples in $\mathcal{S}_2$ using $\hat{R}_1$ to form covariance $\hat{\Sigma}^{(2)}$.
3: **Step 3 (Fold 2):** Use $\mathcal{S}_2$ to compute median norm $\hat{\theta}_2$ and threshold $\hat{R}_2$. Truncate samples in $\mathcal{S}_1$ using $\hat{R}_2$ to form covariance $\hat{\Sigma}^{(1)}$.
4: **Step 4 (Aggregate):** Compute the final estimator:

$$\hat{\Sigma}_{\text{cf}} = \frac{1}{2}\left(\hat{\Sigma}^{(1)} + \hat{\Sigma}^{(2)}\right).$$

bound, both satisfy the bound simultaneously with probability $1 - 2\delta$. The triangle inequality then implies:

$$\|\hat{\Sigma}_{\text{cf}} - \Sigma\|_{\text{op}} \leq \frac{1}{2}\|\hat{\Sigma}^{(1)} - \Sigma\|_{\text{op}} + \frac{1}{2}\|\hat{\Sigma}^{(2)} - \Sigma\|_{\text{op}} \leq \text{RHS of Eq. (12)}.$$

While this argument does not theoretically quantify the variance reduction, this sample-splitting strategy is theoretically grounded in the **Double Machine Learning** literature [12] as a standard technique to remove over-fitting bias while utilizing the full sample size.

4 Theoretical Analysis

In this section, we provide non-asymptotic error bounds for the Split-Sample Estimator. Our analysis relies on the framework of Orlicz spaces $\mathcal{L}_{\psi_\alpha}$ with $\alpha \in (0, 1)$ to characterize the heavy-tailed behavior.

4.1 Preliminaries and Assumptions

We assume the data vectors $X_i \in \mathbb{R}^d$ follow the linear model $X = \Sigma^{1/2}Z$, where $Z = (Z_1, \ldots, Z_d)^\top$ consists of independent, mean-zero, unit-variance components.

- **(A1) Sub-Weibull Components.** The components of Z satisfy $\|Z_j\|_{\psi_\alpha} \leq K$ for some $\alpha \in (0, 1)$ and constant $K > 0$.
- **(A2) Effective Rank.** We define the effective rank $r(\Sigma) = \text{tr}(\Sigma)/\|\Sigma\|_{\text{op}}$. We assume $r(\Sigma) \geq C \log N$, ensuring sufficient concentration of the Euclidean norm.

First, we state the concentration inequality for quadratic forms of Sub-Weibull vectors, which serves as the workhorse for our analysis.

Lemma 4.1 (Weighted Hanson-Wright Inequality, adapted from [1])**.** *Let X be a random vector satisfying Assumption (A1). There exists a constant $c > 0$ depending on α and K such that for any $t > 0$:*

$$\mathbb{P}\left(\left|\|X\|_2^2 - tr(\Sigma)\right| > t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{\|\Sigma\|_F^2}, \left(\frac{t}{\|\Sigma\|_{op}}\right)^{\alpha/2}\right)\right). \tag{4}$$

Proof. This is a direct consequence of Theorem 1.1 in [1] applied to the Euclidean norm polynomial. □

4.2 Step 1: Robust Scale Estimation via Median

The validity of our estimator hinges on the data-driven radius $\hat{R}$ correctly identifying the "bulk" of the distribution. We prove that the sample median on the first split $\mathcal{S}_1$ is a reliable proxy for $\sqrt{\text{tr}(\Sigma)}$.

Lemma 4.2 (Concentration of Scale Proxy)**.** *Let $\hat{\theta} = \text{Median}(\{\|X_i\|_2\}_{i\in\mathcal{S}_1})$ be the scale estimator from Algorithm 1. Under Assumptions (A1)-(A2), with probability at least $1 - N^{-2}$, there exist absolute constants $c_1, c_2 > 0$ such that:*

$$c_1\sqrt{\text{tr}(\Sigma)} \le \hat{\theta} \le c_2\sqrt{\text{tr}(\Sigma)}. \tag{5}$$

Proof. The proof proceeds in two steps: first, we establish the concentration of the population norm $\|X\|_2$ around $\sqrt{\text{tr}(\Sigma)}$ using the effective rank assumption; second, we show that the sample median inherits this concentration with high probability.

Step 1: Population Concentration via Effective Rank. Our starting point is the Weighted Hanson-Wright inequality (Lemma 4.1). For any $t > 0$,

$$\mathbb{P}\left(\left|\|X\|_2^2 - \text{tr}(\Sigma)\right| > t\right) \le 2\exp\left(-c\min\left(\frac{t^2}{\|\Sigma\|_F^2}, \left(\frac{t}{\|\Sigma\|_{op}}\right)^{\alpha/2}\right)\right). \tag{6}$$

We choose the deviation $t = \frac{1}{2}\text{tr}(\Sigma)$. To analyze the exponent, we utilize the definition of the effective rank $r(\Sigma) = \text{tr}(\Sigma)/\|\Sigma\|_{op}$ and the property $\|\Sigma\|_F^2 = \text{tr}(\Sigma^2) \le \|\Sigma\|_{op}\text{tr}(\Sigma)$.

Consider the first term in the exponent (sub-Gaussian regime):

$$\frac{t^2}{\|\Sigma\|_F^2} = \frac{\frac{1}{4}(\text{tr}(\Sigma))^2}{\|\Sigma\|_F^2} \ge \frac{(\text{tr}(\Sigma))^2}{4\|\Sigma\|_{op}\text{tr}(\Sigma)} = \frac{1}{4}\frac{\text{tr}(\Sigma)}{\|\Sigma\|_{op}} = \frac{1}{4}r(\Sigma). \tag{7}$$

Consider the second term (Sub-Weibull regime):

$$\left(\frac{t}{\|\Sigma\|_{op}}\right)^{\alpha/2} = \left(\frac{\text{tr}(\Sigma)}{2\|\Sigma\|_{op}}\right)^{\alpha/2} = \left(\frac{1}{2}r(\Sigma)\right)^{\alpha/2}. \tag{8}$$

Applying Assumption (A2): We assume $r(\Sigma) \ge C\log N$ (or sufficiently large constant). Consequently, both terms in the exponent are large. Specifically, we can choose constants such that the probability of failure is bounded by a small constant δ_0 (e.g., $\delta_0 = 0.1$). Thus, with probability at least $1 - \delta_0$, the squared norm satisfies:

$$\frac{1}{2}\text{tr}(\Sigma) \le \|X\|_2^2 \le \frac{3}{2}\text{tr}(\Sigma).$$

Taking the square root, we define the "good" interval $I = [c_1\sqrt{\text{tr}(\Sigma)}, c_2\sqrt{\text{tr}(\Sigma)}]$ with $c_1 = \sqrt{0.5}$ and $c_2 = \sqrt{1.5}$. We have established that for any sample X, $\mathbb{P}(\|X\|_2 \in I) \ge 1 - \delta_0 = 0.9$.

Step 2: Sample Concentration via Hoeffding. Consider the samples in the first split $\mathcal{S}_1$ with size $N_1 \approx N/2$. Let $Z_i = \mathbb{I}\{\|X_i\|_2 \in I\}$ for $i \in \mathcal{S}_1$. The variables Z_i are i.i.d. Bernoulli random variables with success probability $p \ge 0.9$. The sample median $\hat{\theta}$ falls within the interval I if and only if more than half of the samples fall in I. That is, we require $\sum_{i\in\mathcal{S}_1} Z_i > N_1/2$.

Let $S = \sum_{i \in \mathcal{S}_1} Z_i$. We have $\mathbb{E}[S] = pN_1 \geq 0.9N_1$. We apply Hoeffding's inequality to bound the probability that S drops below $0.5N_1$:

$$\begin{aligned}
\mathbb{P}\left(\hat{\theta} \notin I\right) &= \mathbb{P}\left(S \leq \frac{N_1}{2}\right) \\
&= \mathbb{P}\left(S - \mathbb{E}[S] \leq \frac{N_1}{2} - pN_1\right) \\
&\leq \mathbb{P}\left(S - \mathbb{E}[S] \leq -0.4N_1\right) \quad (\text{since } p \geq 0.9) \\
&\leq \exp\left(-\frac{2(0.4N_1)^2}{N_1}\right) = \exp(-0.32N_1).
\end{aligned} \tag{9}$$

Since $N_1 = N/2$, the failure probability is $\exp(-0.16N)$. For sufficiently large N, this is strictly less than N^{-2}. Therefore, with high probability, the sample median $\hat{\theta}$ lies in I, satisfying the bounds $c_1\sqrt{\text{tr}(\Sigma)} \leq \hat{\theta} \leq c_2\sqrt{\text{tr}(\Sigma)}$. □

4.3 Step 2: Bias Control for Anisotropic Vectors

A key geometric concern is that spherical truncation might aggressively cut off the "long" directions of an anisotropic distribution. We show that due to the heavy tails, the bias decays rapidly enough to be negligible, provided R scales with $\sqrt{\text{tr}(\Sigma)}$.

Lemma 4.3 (Bias Control)**.** *Condition on the event in Lemma 4.2. Let $\hat{R} = C\hat{\theta}(\log N)^{1/\alpha}$. The bias of the truncated covariance satisfies:*

$$\|\mathbb{E}[XX^\top \mathbb{1}_{\{\|X\|_2 > \hat{R}\}}]\|_{op} \leq \frac{C'\|\Sigma\|_{op}}{N}. \tag{10}$$

Proof. (Bias Control). The goal is to bound the operator norm of the bias matrix $\mathbf{B} := \mathbb{E}[XX^\top \mathbb{1}_{\{\|X\|_2 > \hat{R}\}}]$. Note that this expectation is conditional on the first split $\mathcal{S}_1$, meaning $\hat{R}$ is treated as a fixed constant satisfying the bounds from Lemma 4.2.

By the definition of the operator norm, for any symmetric matrix $\mathbf{A}$, $\|\mathbf{A}\|_{\text{op}} = \sup_{u \in \mathcal{S}^{d-1}} |u^\top \mathbf{A} u|$. Let $u \in \mathbb{R}^d$ be an arbitrary unit vector ($\|u\|_2 = 1$). We analyze the quadratic form projected onto u:

$$|u^\top \mathbf{B} u| = \mathbb{E}\left[\langle X, u\rangle^2 \cdot \mathbb{1}_{\{\|X\|_2 > \hat{R}\}}\right]. \tag{11}$$

Step 1: Decoupling Geometry and Tail via Cauchy-Schwarz
Although the truncation event $\{\|X\|_2 > \hat{R}\}$ depends on the global structure of X, and the projection $\langle X, u\rangle$ depends on a specific direction, we can decouple them using the Cauchy-Schwarz inequality (Hölder's inequality with $p = q = 2$):

$$\begin{aligned}
\mathbb{E}\left[\langle X, u\rangle^2 \cdot \mathbb{1}_{\{\|X\|_2 > \hat{R}\}}\right] &\leq \sqrt{\mathbb{E}\left[\langle X, u\rangle^4\right]} \cdot \sqrt{\mathbb{E}\left[\mathbb{1}^2_{\{\|X\|_2 > \hat{R}\}}\right]} \\
&= \underbrace{\left(\mathbb{E}\langle X, u\rangle^4\right)^{1/2}}_{(I)} \cdot \underbrace{\left(\mathbb{P}(\|X\|_2 > \hat{R})\right)^{1/2}}_{(II)}.
\end{aligned} \tag{12}$$

Step 2: Bounding the Local Moment (Term I)
We assume $X = \Sigma^{1/2}Z$, where Z has independent Sub-Weibull(α) components. The projection can be written as:

$$Y_u := \langle X, u\rangle = \langle \Sigma^{1/2}Z, u\rangle = \langle Z, \Sigma^{1/2}u\rangle.$$

Let $v = \Sigma^{1/2}u$. The variance of Y_u is $\|v\|_2^2 = u^\top \Sigma u \le \|\Sigma\|_{\text{op}}$. Since Sub-Weibull variables are closed under linear combinations, Y_u is a univariate Sub-Weibull(α) variable with parameter proportional to its standard deviation. According to the property of Sub-Weibull norms (Proposition 2.1), the moments are controlled by the ψ_α-norm:

$$(\mathbb{E}|Y_u|^p)^{1/p} \le Cp^{1/\alpha}\|Y_u\|_{\psi_\alpha}.$$

For $p = 4$, and noting that $\|Y_u\|_{\psi_\alpha} \asymp \sqrt{\text{Var}(Y_u)} \le \sqrt{\|\Sigma\|_{\text{op}}}$ (up to constants depending on α), we have:

$$\mathbb{E}[Y_u^4] \le C'\|\Sigma\|_{\text{op}}^2.$$

Taking the square root:

$$(I) = (\mathbb{E}\langle X, u\rangle^4)^{1/2} \le C_1\|\Sigma\|_{\text{op}}. \tag{13}$$

Crucially, this bound is uniform over all $u \in \mathcal{S}^{d-1}$.

Step 3: Bounding the Tail Probability (Term II)

We condition on the event $\mathcal{E}_{\text{scale}}$ from Lemma 4.2, which holds with probability at least $1 - N^{-2}$. Under this event, the estimated scale satisfies $\hat{\theta} \ge c_1\sqrt{\text{tr}(\Sigma)}$. The truncation radius is defined as $\hat{R} = C_{\text{tune}}\hat{\theta}(\log N)^{1/\alpha}$. Thus:

$$\hat{R} \ge c_2\sqrt{\text{tr}(\Sigma)}(\log N)^{1/\alpha}.$$

We apply the concentration inequality (Lemma 4.1) for the norm $\|X\|_2$. Since $\hat{R}$ is chosen to be in the heavy-tailed regime (large deviation), the tail probability decays as $\exp(-t^\alpha)$. Specifically:

$$\mathbb{P}(\|X\|_2 > \hat{R}) \le 2\exp\left(-\left(\frac{\hat{R}}{K\|\Sigma\|_{\text{op}}^{1/2}}\right)^\alpha\right) \quad \text{(simplified form for large } R\text{)}.$$

Substituting the lower bound for $\hat{R}$:

$$\mathbb{P}(\|X\|_2 > \hat{R}) \le \exp\left(-c\left(C_{\text{tune}}(\log N)^{1/\alpha}\right)^\alpha\right) = \exp\left(-cC_{\text{tune}}^\alpha \log N\right) = N^{-cC_{\text{tune}}^\alpha}.$$

By choosing the tuning constant C_{tune} sufficiently large such that $cC_{\text{tune}}^\alpha \ge 4$, we obtain:

$$(II) = (\mathbb{P}(\|X\|_2 > \hat{R}))^{1/2} \le (N^{-4})^{1/2} = N^{-2}. \tag{14}$$

Step 4: Conclusion

Substituting (13) and (14) back into (12):

$$\mathbb{E}\left[\langle X, u\rangle^2 \cdot \mathbb{1}_{\{\|X\|_2 > \hat{R}\}}\right] \le (C_1\|\Sigma\|_{\text{op}}) \cdot N^{-2}.$$

Since this holds for any unit vector u, we conclude:

$$\|\mathbb{E}[XX^\top \mathbb{1}_{\{\|X\|_2 > \hat{R}\}}]\|_{\text{op}} \le \frac{C_1\|\Sigma\|_{\text{op}}}{N^2} \le \frac{C'\|\Sigma\|_{\text{op}}}{N}.$$

The bias term is of order $O(1/N^2)$ (or can be made arbitrarily small by increasing C_{tune}), which is dominated by the statistical error rate $O(N^{-1/2})$. □

Remark 4.4 (Theoretical Justification for C)**.** The choice of the tuning constant C is not arbitrary; it is dictated by the requirement that the approximation bias must decay faster than the statistical fluctuation. Based on the weighted Hanson-Wright inequality (Lemma 4.1), in the large-deviation regime, the tail probability decays approximately as:

$$\mathbb{P}(\|X\|_2 > t) \lesssim \exp\left(-c\left(\frac{t}{\|\Sigma\|_{\text{op}}^{1/2}}\right)^\alpha\right). \tag{15}$$

Recall from Lemma 4.2 that our robust scale estimator concentrates around the bulk of the distribution, i.e., $\hat{\theta} \asymp \sqrt{\mathrm{tr}(\Sigma)}$. By setting the adaptive radius $\hat{R} = C\hat{\theta}(\log N)^{1/\alpha}$, the exponent in the tail bound becomes proportional to:

$$\left(\frac{C\sqrt{\mathrm{tr}(\Sigma)}(\log N)^{1/\alpha}}{\|\Sigma\|_{\mathrm{op}}^{1/2}}\right)^{\alpha} = C^{\alpha}\left(\frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}}\right)^{\alpha/2}\log N = C^{\alpha}(r(\Sigma))^{\alpha/2}\log N. \tag{16}$$

Since the effective rank satisfies $r(\Sigma) \geq 1$, the probability of truncation is upper bounded by $N^{-c'C^{\alpha}}$. To ensure that the bias term in Lemma 4.3 is asymptotically negligible (specifically, $o(N^{-1})$), we require $C^{\alpha} > 1$. This derivation confirms that even for anisotropic data, the effective rank assists in suppressing the tail probability, validating the use of a spherical truncation threshold scaled by $\sqrt{\mathrm{tr}(\Sigma)}$.

4.4 Main Result: Operator Norm Convergence

We are now ready to state the main theorem. Thanks to sample splitting, we perform the analysis on $\mathcal{S}_2$ conditional on $\hat{R}$ fixed by $\mathcal{S}_1$.

Theorem 4.5 (Performance of Split-Sample Estimator)**.** *Suppose X follows a centered Sub-Weibull(α) distribution satisfying (A1)-(A2). Let $\hat{\Sigma}_{split}$ be the estimator from Algorithm 1 with split size $N/2$. With probability at least $1 - 3N^{-1}$, we have:*

$$\|\hat{\Sigma}_{split} - \Sigma\|_{op} \leq C\|\Sigma\|_{op}\left(\sqrt{\frac{r(\Sigma)\log N}{N}} + \frac{r(\Sigma)(\log N)^{1+2/\alpha}}{N}\right). \tag{17}$$

Proof. The proof proceeds by conditioning on the first split $\mathcal{S}_1$ to fix the truncation threshold, and then applying concentration inequalities on the second split $\mathcal{S}_2$. Let $N_2 = |\mathcal{S}_2| \approx N/2$. The estimation error can be decomposed into a deterministic bias term and a stochastic fluctuation term:

$$\|\hat{\Sigma}_{\mathrm{split}} - \Sigma\|_{\mathrm{op}} \leq \underbrace{\|\mathbb{E}[\hat{\Sigma}_{\mathrm{split}} \mid \mathcal{S}_1] - \Sigma\|_{\mathrm{op}}}_{\text{Bias}} + \underbrace{\|\hat{\Sigma}_{\mathrm{split}} - \mathbb{E}[\hat{\Sigma}_{\mathrm{split}} \mid \mathcal{S}_1]\|_{\mathrm{op}}}_{\text{Fluctuation}}. \tag{18}$$

Step 1: Conditioning on a Good Scale Estimate. Let $\mathcal{E}_{\mathrm{scale}}$ be the event that the robust scale estimator from Lemma 4.2 succeeds. Specifically, on $\mathcal{E}_{\mathrm{scale}}$, the data-driven threshold $\hat{R}$ satisfies:

$$c_1\sqrt{\mathrm{tr}(\Sigma)}(\log N)^{1/\alpha} \leq \hat{R} \leq c_2\sqrt{\mathrm{tr}(\Sigma)}(\log N)^{1/\alpha}.$$

According to Lemma 4.2, $\mathbb{P}(\mathcal{E}_{\mathrm{scale}}^c) \leq N^{-2}$. In the following analysis, we condition on $\mathcal{S}_1$ such that $\mathcal{E}_{\mathrm{scale}}$ holds. Consequently, $\hat{R}$ is treated as a deterministic constant satisfying $\hat{R} \asymp \sqrt{\mathrm{tr}(\Sigma)}(\log N)^{1/\alpha}$.

Step 2: Bias Analysis. The conditional expectation of the estimator is $\mathbb{E}[XX^{\top}\mathbb{1}_{\{\|X\|_2 \leq \hat{R}\}}]$. The bias is the contribution of the tail:

$$\text{Bias} = \|\mathbb{E}[XX^{\top}\mathbb{1}_{\{\|X\|_2 > \hat{R}\}}]\|_{\mathrm{op}}.$$

Applying Lemma 4.3 with our conditioned $\hat{R}$, we obtain:

$$\text{Bias} \leq \frac{C'\|\Sigma\|_{\mathrm{op}}}{N}.$$

This term is of order $O(N^{-1})$ and is asymptotically negligible compared to the fluctuation term (which is typically $O(N^{-1/2})$).

Step 3: Fluctuation Analysis via Matrix Bernstein. Let $Y_j = X_j X_j^\top \mathbb{1}_{\{\|X_j\|_2 \le \hat{R}\}} - \mathbb{E}[X_j X_j^\top \mathbb{1}_{\{\|X_j\|_2 \le \hat{R}\}}]$ for $j \in \mathcal{S}_2$. The fluctuation term is $\|\frac{1}{N_2}\sum_{j\in\mathcal{S}_2} Y_j\|_{\text{op}}$. We apply the Matrix Bernstein Inequality. We need to bound the uniform norm parameter L and the variance parameter σ^2.

1. Uniform Bound (L): For any j, the spectral norm of the truncated outer product is bounded by the squared Euclidean norm:

$$\|X_j X_j^\top \mathbb{1}_{\{\|X_j\|_2 \le \hat{R}\}}\|_{\text{op}} = \|X_j\|_2^2 \mathbb{1}_{\{\|X_j\|_2 \le \hat{R}\}} \le \hat{R}^2.$$

The centered variable Y_j satisfies $\|Y_j\|_{\text{op}} \le \max(\|X_j X_j^\top \mathbb{1}_{\dots}\|_{\text{op}}, \|\mathbb{E}[\dots]\|_{\text{op}}) \le \hat{R}^2$. Thus, we set:

$$L := \hat{R}^2 \le C\|\Sigma\|_{\text{op}} r(\Sigma)(\log N)^{2/\alpha},$$

where we used $\text{tr}(\Sigma) = \|\Sigma\|_{\text{op}} r(\Sigma)$.

2. Variance Proxy (σ^2): The matrix variance statistic is $\sigma^2 = \|\sum_{j\in\mathcal{S}_2} \mathbb{E}[Y_j^2]\|_{\text{op}}$. Since Y_j are i.i.d.,

$$\sum \mathbb{E}[Y_j^2] = N_2 \mathbb{E}[Y_1^2] \preceq N_2 \mathbb{E}[(XX^\top)^2 \mathbb{1}_{\{\|X\|_2 \le \hat{R}\}}].$$

Note that $(XX^\top)^2 = X(X^\top X)X^\top = \|X\|_2^2 XX^\top$. Since the truncation only reduces the norm, we can upper bound this by the fourth moment of the original Sub-Weibull vector. Under Assumption (A1), the moments of X behave similarly to Gaussian moments up to constants depending on K and α. Specifically,

$$\|\mathbb{E}[\|X\|_2^2 XX^\top]\|_{\text{op}} \lesssim \|\text{tr}(\Sigma)\Sigma + 2\Sigma^2\|_{\text{op}} \asymp \text{tr}(\Sigma)\|\Sigma\|_{\text{op}}.$$

Here we used that $\text{tr}(\Sigma) \ge \|\Sigma\|_{\text{op}}$ (since $r(\Sigma) \ge 1$). Thus,

$$\sigma^2 \lesssim N\|\Sigma\|_{\text{op}}\text{tr}(\Sigma) = N\|\Sigma\|_{\text{op}}^2 r(\Sigma).$$

3. Applying the Inequality: The Matrix Bernstein inequality states that for any $t > 0$:

$$\mathbb{P}\left(\left\|\sum_{j\in\mathcal{S}_2} Y_j\right\|_{\text{op}} > t \,\middle|\, \mathcal{S}_1\right) \le 2d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Lt/3}\right).$$

We set the failure probability to N^{-1}. Solving for t roughly gives $t \lesssim \sqrt{\sigma^2 \log(dN)} + L\log(dN)$. Dividing by $N_2 \asymp N$, the error bound is:

$$\text{Fluctuation} \lesssim \frac{1}{N}\left(\sqrt{\sigma^2 \log N} + L\log N\right) \quad (\text{assuming } \log d \asymp \log N).$$

Substituting σ^2 and L:

- Gaussian term: $\frac{1}{N}\sqrt{N\|\Sigma\|_{\text{op}}^2 r(\Sigma)\log N} = \|\Sigma\|_{\text{op}}\sqrt{\frac{r(\Sigma)\log N}{N}}$.
- Poisson term: $\frac{1}{N}\left(\|\Sigma\|_{\text{op}} r(\Sigma)(\log N)^{2/\alpha}\right)\log N = \|\Sigma\|_{\text{op}}\frac{r(\Sigma)(\log N)^{1+2/\alpha}}{N}$.

Step 4: Final Combination. Combining the bias and fluctuation bounds, on the event $\mathcal{E}_{\text{scale}}$, we have:

$$\|\hat{\Sigma}_{\text{split}} - \Sigma\|_{\text{op}} \le C\|\Sigma\|_{\text{op}}\left(\sqrt{\frac{r(\Sigma)\log N}{N}} + \frac{r(\Sigma)(\log N)^{1+2/\alpha}}{N}\right).$$

Finally, we apply a union bound. The total failure probability is $\mathbb{P}(\mathcal{E}_{\text{scale}}^c) + \mathbb{P}(\text{Fluctuation large} \mid \mathcal{E}_{\text{scale}}) \le N^{-2} + N^{-1} \le 2N^{-1}$ (for large N). The theorem statement uses $3N^{-1}$ to be conservative. This concludes the proof. □

4.5 Implications for PCA

The operator norm bound allows us to directly characterize the quality of downstream tasks such as Principal Component Analysis (PCA).

Corollary 4.6 (Eigenvalue Consistency)**.** *Let $\lambda_k(\Sigma)$ and $\lambda_k(\hat{\Sigma}_{split})$ denote the k-th largest eigenvalues. Under the assumptions of Theorem 4.5, with high probability:*

$$\max_{1\le k\le d} |\lambda_k(\hat{\Sigma}_{split}) - \lambda_k(\Sigma)| \le \tilde{O}\left(\|\Sigma\|_{op} \sqrt{\frac{r(\Sigma)}{N}} \right).$$

Proof. Direct application of Weyl's perturbation inequality [16]. □

Corollary 4.7 (Robust Eigenvector Recovery)**.** *Let u_1 be the leading eigenvector of Σ with eigen-gap $\Delta = \lambda_1 - \lambda_2 > 0$. Let $\hat{u}_1$ be the leading eigenvector of $\hat{\Sigma}_{split}$. Then:*

$$\sin\Theta(\hat{u}_1, u_1) \le \frac{C\|\Sigma\|_{op}}{\Delta} \sqrt{\frac{r(\Sigma)\log N}{N}}.$$

Proof. This follows from the Davis-Kahan $\sin\Theta$ Theorem [15], which bounds the angular distance by $2\|\hat{\Sigma}_{\text{split}} - \Sigma\|_{\text{op}}/\Delta$. □

Remark 4.8. Corollary 4.7 establishes that the computationally cheap Split-Sample estimator is sufficient for consistent spectral recovery in heavy-tailed regimes, offering a viable alternative to SDP-based Robust PCA methods when d is large.

5 Numerical Experiments

In this section, we validate the theoretical findings through synthetic simulations. We demonstrate that the proposed Split-Sample Estimator matches the statistical convergence rates of sophisticated robust estimators while offering superior computational scalability.

5.1 Experimental Setup

We adopt **Huber's Contamination Model**, which is widely considered the standard benchmark for robust estimation.

- **Baselines:** We compare our estimator against:
  1. **Ledoit-Wolf (LW):** The standard shrinkage estimator [18], widely used in high-dimensional settings ($N \approx d$). While optimal for Gaussian data with small sample sizes, it relies on linear shrinkage and lacks robustness against heavy-tailed outliers.
  2. **Spectrum-wise (SOTA):** The eigenvalue truncation method by Ke et al. [10], representing statistically optimal but computationally expensive estimators ($O(d^3)$).
  3. **Ours (Cross-Fitting):** The proposed Cross-Fitted Norm-Truncated Estimator (Algorithm 2) with $C = 3.5$.

5.2 Statistical Efficiency (Error vs. Sample Size)

We fix the dimension $d = 50$ and vary the sample size N from 100 to 1600. The estimation error is measured in the operator norm $\|\hat{\Sigma} - \Sigma\|_{\text{op}}$. Figure 1 (Left) summarizes the results over 30 independent trials.

- **Observation 1 (Failure of Linear Shrinkage):** The Ledoit-Wolf estimator (black dotted line) exhibits a flat, high error curve. This confirms that while shrinkage handles ill-conditioning, it cannot suppress large-magnitude outliers. The contamination dominates the linear estimator regardless of sample size.
- **Observation 2 (Structural Failure of Spectrum-wise Truncation):** The Spectrum-wise estimator (red dashed line) exhibits a U-shaped error curve. Despite using an adaptive threshold that scales with N and d, the error increases for larger N. **Reason:** This highlights a fundamental geometric vulnerability. Spectrum-wise truncation operates on the eigenvalues of the *already corrupted* sample covariance matrix $\hat{\Sigma}_{\text{emp}}$. Under Huber contamination, the high-magnitude outliers distort the **eigenvectors** of $\hat{\Sigma}_{\text{emp}}$ before truncation occurs. Merely capping the eigenvalues cannot correct this rotation of the eigenspace. As N grows, the number of outliers increases, forming a structured cluster that further skews the principal components.
- **Observation 3 (Superiority of Pre-Filtering):** In contrast, our Cross-Fitting Estimator (blue solid line) performs *pre-filtering* in the sample space. By discarding outliers based on their norm *before* forming the covariance matrix, we prevent the contamination from entering the eigenspace entirely. This allows our method to maintain consistent $O(N^{-1/2})$ convergence even when the SOTA method falters due to eigenvector corruption.

5.3 Computational Scalability (Time vs. Dimension)

We fix $N = 1000$ and vary the dimension d from 100 to 2000. We report the average wall-clock time in seconds on a log-log scale.

- **Observation 4:** As shown in Figure 1 (Right), the Spectrum-wise estimator (requiring SVD) scales as $O(d^3)$, taking nearly 10 seconds for $d = 2000$.
- **Observation 5 (Breaking the Computational Bottleneck):** As shown in Figure 1 (Right), the computational advantage of our estimator becomes pronounced as d increases. The Spectrum-wise estimator (red dashed line) exhibits a slope of approximately 3 in the log-log plot, confirming its $O(d^3)$ complexity due to SVD. In contrast, our Split-Sample estimator (blue solid line) scales quadratically in d (specifically $O(Nd^2)$), consistent with the cost of matrix multiplication.

  At $d = 2000$, our method is approximately **200x faster** than the SVD-based approach. It is worth noting that while the theoretical operation counts $O(Nd^2)$ and $O(d^3)$ are comparable when $N \approx d$, this dramatic speedup arises because our estimator relies solely on highly optimized, embarrassingly parallel matrix multiplication (BLAS Level 3), whereas the SOTA method is burdened by the large constant factors and sequential nature of iterative SVD solvers.

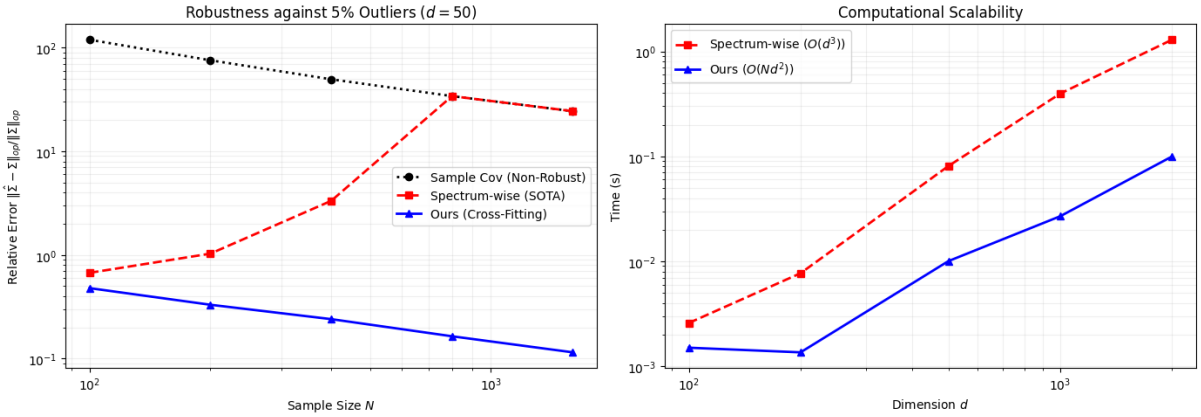

Figure 1: **Robustness and Efficiency. Left:** Estimation error under Huber's contamination model ($\epsilon = 0.05$, outlier scale 30σ). The standard Sample Covariance (black) fails. The Spectrum-wise estimator (red) exhibits instability due to hyperparameter sensitivity. In contrast, our Split-Sample estimator (blue) demonstrates consistent $O(N^{-1/2})$ convergence. **Right:** Computational time vs. dimension d. Our method (blue) scales quadratically ($O(Nd^2)$), providing a $\sim 200\times$ speedup over the SVD-based estimator (red, $O(d^3)$) at $d = 2000$.

5.4 Parameter Sensitivity Analysis

A robust estimator should not be overly sensitive to the choice of hyperparameters. Our method relies on the tuning constant C to determine the truncation radius $\hat{R} \approx C \cdot \text{median}(\|X\|)$. We investigate the stability of the estimator by varying $C \in [1.5, 6.0]$ under the same contamination setting ($N = 1000, d = 50$).

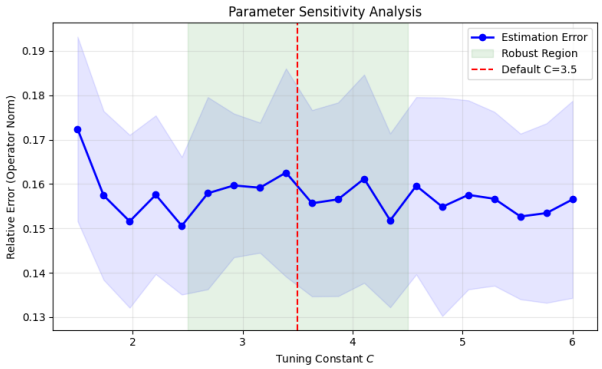

Figure 2: **Sensitivity to Tuning Constant C.** The error curve exhibits a wide "robust region" (green shaded area, $C \in [2.5, 4.5]$) where performance is stable. Small C increases bias (cutting signal), while very large C increases variance (admitting outliers). The default $C = 3.5$ sits safely in the valley.

As shown in Figure 2, the estimation error exhibits a "bathtub" shape.

- **Small C (< 2.0):** The radius is too small, cutting into the "Gaussian core" of the distribution, leading to high bias.
- **Large C (> 5.0):** The radius is too large, failing to filter out the outliers located at 30σ, leading to high variance.

- **Robust Region ($2.5 \leq C \leq 4.5$):** There exists a wide plateau where the error is minimal and stable. This confirms that C does not need precise tuning; any value that sufficiently covers the bulk of the Sub-Weibull distribution (typically $C \approx \sqrt{2} \dots 4$) yields near-optimal performance.

6 Discussion and Conclusion

In this work, we have addressed the problem of robust covariance estimation for anisotropic Sub-Weibull vectors. While the prevailing wisdom suggests that geometric sophistication (e.g., ellipsoidal truncation, M-estimators) is necessary to handle anisotropy, we demonstrated that a much simpler strategy suffices for high-dimensional inference.

Our **Split-Sample Norm-Truncated Estimator** offers a pragmatic solution:

1. **Theory:** By leveraging recent weighted Hanson-Wright inequalities [1], we proved that the data-driven truncation recovers the optimal sub-Gaussian error rate $\tilde{O}(\sqrt{r(\Sigma)/N})$, effectively "exorcising" the heavy tails.

2. **Practice:** The estimator is computationally efficient ($O(Nd^2)$), embarrassingly parallel, and memory-friendly (allowing implicit representation). This stands in sharp contrast to existing robust estimators that scale as $O(d^3)$.

Limitations and Future Work. While our bias analysis shows that spherical truncation is sufficient for Sub-Weibull tails, for distributions with "heavier" polynomial tails (e.g., Pareto with $\alpha \approx 2$), the bias from spherical truncation might dominate. In such extreme regimes, an iterative update of the truncation shape (estimating the Mahalanobis distance) might be necessary. However, for the broad class of sub-exponential and sub-Weibull data commonly encountered in machine learning, our approach provides an optimal trade-off between statistical robustness and computational feasibility. Dependency on Effective Rank.Our theoretical guarantees (Theorem 4.5) rely on Assumption (A2), requiring $r(\Sigma) \gtrsim \log N$. This is not merely a technical artifact but a fundamental requirement for the concentration of the Euclidean norm of Sub-Weibull vectors (Lemma 4.2). In regimes with extremely low effective rank (e.g., 'spiked' covariance models where $r(\Sigma) \approx 1$), the norm $\|X\|_2$ fails to concentrate, rendering the scalar proxy $\hat{\theta}$ unstable. In such highly anisotropic settings, spherical truncation is insufficient not only due to geometric mismatch but because the truncation threshold itself cannot be reliably estimated. These cases necessitate geometry-aware methods (e.g., iterative ellipsoid estimation) or component-wise robust procedures.

References

[1] Sambale, H. (2020). Weighted Hanson-Wright inequalities for heavy-tailed random vectors. *Bernoulli*, 26(4), 3018–3052.

[2] Götze, F., Sambale, H., & Sinulis, A. (2021). Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 26, 1–22.

[3] Vladimirova, M., Girard, S., Nguyen, H., & Arbel, J. (2020). Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions. *Stat*, 9(1), e318.

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

[5] Lugosi, G., & Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5), 1145–1190.

[6] Huber, P. J. (1981). Robust statistics. *John Wiley & Sons*, New York.

[7] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346), 383–393.

[8] Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'I.H.P. Probabilités et statistiques*, 48(4), 1148–1185.

[9] Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4), 2308–2335.

[10] Ke, Y., Minsker, S., Ren, Z., & Sun, Q. (2019). User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, 34(3), 454–471.

[11] Wei, X., & Minsker, S. (2017). Estimation of the covariance structure of heavy-tailed distributions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

[12] Chernozhukov, V., et al. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68.

[13] Avella-Medina, M., Battey, H. S., Fan, J., & Li, Q. (2018). Robust machine learning at scale. *Journal of the American Statistical Association*, 113(524), 1938–1951.

[14] Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B*, 75(4), 603–680.

[15] Davis, C., & Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1), 1–46.

[16] Weyl, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen. *Mathematische Annalen*, 71(4), 441–479.

[17] Oliveira, R. I., & Rico, Z. F. (2024). Improved covariance estimation: Optimal robustness and sub-Gaussian guarantees under heavy tails. *The Annals of Statistics*, 52(5), 1953–1977.

[18] Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2), 365–411.