Cross-Fitted Clipped Covariance Estimation with a Data-Driven Tail-Energy Criterion

Anonymous authors
Paper under double-blind review

Abstract

Heavy-tailed data make covariance estimation sensitive to the clipping level: stronger clipping reduces variance but increases bias. We study this tuning problem for the centered clipped covariance estimator in the known-mean setting. We propose QTES, a data-driven rule that selects the clipping level by combining a cross-fitted variance certificate with a held-out estimate of the tail energy removed by clipping. The key observation is that, for Euclidean clipping, the operator-norm bias can be upper bounded by a scalar tail-energy quantity. Under a finite L_4 moment condition, together with a mild feasibility condition for the QTES block construction, we prove a uniform finite-sample calibration bound over a positive clipping grid and a finite-sample guarantee for the estimator selected by QTES, relative to the best candidate on that grid under the same bias-variance criterion. This result is a guarantee for data-driven calibration within the clipped family; it does not establish the sharp direct effective-rank rates known for other robust covariance estimators under stronger assumptions. Experiments on clean heavy-tailed and contaminated benchmarks show that QTES improves over uninformed clipping choices and is often close to an oracle-tuned Wei-Minsker benchmark serving as a best-case oracle baseline within the centered clipped family.

1 Introduction

Covariance estimation in operator norm is a basic ingredient in principal component analysis, factor modeling, and high-dimensional spectral inference (Johnstone, 2001; Paul, 2007; Fan et al., 2013). In light-tailed settings, the sample covariance matrix is understood up to the effective rank $r(\Sigma) := \mathrm{tr}(\Sigma)/\|\Sigma\|_{\mathrm{op}}$: for Gaussian and sub-Gaussian data, its benchmark error scale is $\|\Sigma\|_{\mathrm{op}}(\sqrt{r(\Sigma)/n} \vee r(\Sigma)/n)$ (Lounici, 2014; Dirksen, 2015; Koltchinskii & Lounici, 2017). Under weaker $2+\varepsilon$ moments one can still control the empirical covariance, but the behavior deteriorates and sensitivity to rare large observations becomes a dominant practical issue (Srivastava & Vershynin, 2013).

Modern robust covariance estimation therefore asks for two things simultaneously: statistically sharp operator-norm control under weak moments or contamination, and computationally tractable estimators. A substantial nonasymptotic literature now addresses this challenge. PAC-Bayesian Gram/covariance estimators and robust random-matrix mean estimators provide one route (Catoni, 2016; Catoni & Giulini, 2017; Minsker, 2018). Entrywise and contamination-robust procedures provide another (Avella-Medina et al., 2018; Chen et al., 2018). For computationally tractable operator-norm estimation by truncating rank-one contributions, important milestones are Wei & Minsker (2017), the U-statistic refinements of Minsker & Wei (2020), the user-friendly survey and automatic-calibration viewpoint of Ke et al. (2019), and the contamination/low-rank extensions of Minsker & Wang (2024).

The sharp direct theory is also much clearer than it was a few years ago. Under marginal L_4–L_2 norm equivalence, Mendelson & Zhivotovskiy (2020) show that one can attain the same effective-rank order as in the Gaussian benchmark. Oliveira & Rico (2024) and Abdalla & Zhivotovskiy (2026) extend this picture to broader L_p–L_2 regimes and to adversarial contamination, again with Gaussian-type operator-norm rates up

to constants. These results largely settle what is statistically possible for *direct* robust covariance estimation under stronger marginal regularity assumptions.

This paper focuses on a different but practically central question: how should one choose the robustification level inside a simple clipped covariance family when one wants an *observable* and fully implementable calibration rule? In many truncation-based methods, the estimator family is easy to compute once a clipping level is fixed, whereas the hard part is choosing that level from data without a trusted oracle scale range. In the centered/known-mean setting, the rank-one truncation of Wei & Minsker (2017) reduces algebraically to exact Euclidean radial clipping. Because these estimator families share the same structural foundation, evaluating the Wei–Minsker estimator at the analytically optimal scale derived from the unobservable true distribution provides a formidable, theory-guided oracle baseline. By collapsing its candidate grid to this single exact point, we completely disable its Lepski adaptation procedure, yielding a privileged non-data-driven reference. The spectrumwise truncation language of Ke et al. (2019) points to essentially the same computable family. Our contribution is a fully observable, data-driven selector that strives to approach this oracle baseline without requiring any hidden scale knowledge, rather than proposing a new direct covariance estimator outside it.

The key structural point is that Euclidean clipping makes the bias unusually transparent:

$$\Sigma - \mathbb{E}_P[\tilde{Z}(\tau)\tilde{Z}(\tau)^\top] = \mathbb{E}_P\Big[\Big(1 - \min\{1, \tau^2/\|Z\|_2^2\}\Big)ZZ^\top\Big] \succeq 0.$$

Therefore,

$$\big\|\Sigma - \mathbb{E}_P[\tilde{Z}(\tau)\tilde{Z}(\tau)^\top]\big\|_{\mathrm{op}} \le \mathrm{tr}\big(\Sigma - \mathbb{E}_P[\tilde{Z}(\tau)\tilde{Z}(\tau)^\top]\big) = \mathbb{E}_P\big[(\|Z\|_2^2 - \tau^2)_+\big].$$

Thus, the matrix bias is controlled by a scalar tail-energy functional. We exploit this identity by combining a held-out estimator of the tail-energy bias with a cross-fitted variance certificate built from a matrix empirical-Bernstein inequality. The resulting score selects the clipping level from a positive geometric grid without requiring a user-supplied scale range such as $[\sigma_{\min}, \sigma_{\max}]$. The theorem proved below uses the finite fourth-moment condition $\mathbb{E}_P\|Z_1\|_2^4 < \infty$ together with a mild feasibility requirement for the QTES block construction. The price is that the guarantee is selector-relative: it calibrates the score uniformly over the clipping grid, but it does not yet recover the sharp direct effective-rank rates known for other estimators under stronger assumptions.

Contributions.

- We isolate the tail-energy identity for Euclidean clipping and reduce operator-norm bias control to a scalar tail-energy problem.
- We propose the quantile tail-energy surrogate (QTES), which combines a cross-fitted variance certificate with a held-out bias surrogate. As our analysis shows, after accounting for an unobservable concentration slack, this score forms a valid simultaneous upper confidence bound. Under the finite fourth-moment condition $\mathbb{E}_P\|Z_1\|_2^4 < \infty$, together with a mild feasibility requirement for the QTES block construction, the selector satisfies a uniform finite-sample calibration bound and a near-oracle inequality on the positive clipping grid. The procedure is fully implementable and simple: it only requires clipping, sample splitting, and grid search, and it does not rely on rough oracle bounds for an unknown tail scale.
- Empirically, QTES improves over uninformed clipping choices and compares favorably with SCM, OAS, and MCD baselines. More importantly, it tightly tracks an oracle-tuned Wei & Minsker (2017) benchmark. Demonstrating that QTES can closely match this inaccessible baseline confirms that our score successfully resolves the empirical calibration problem within the clipped family.

Data model. Let P denote the underlying distribution. We observe independent samples $Z_1, \ldots, Z_n \sim P$ in $\mathbb{R}^d$ with $\mathbb{E}_P[Z_1] = 0$ and target covariance $\Sigma = \mathbb{E}_P[Z_1 Z_1^\top]$. We assume the finite fourth-moment condition

$$\mathbb{E}_P\|Z_1\|_2^4 < \infty. \tag{1}$$

Related work. Classical robust scatter estimators such as Maronna's estimator, Tyler's estimator, and MCD/MVE-type high-breakdown procedures remain basic references, especially under elliptical or shape-based modeling assumptions (Maronna, 1976; Tyler, 1987; Hubert et al., 2008). For nonasymptotic heavy-tail covariance estimation in operator norm, PAC-Bayesian and robust-mean ideas were developed by Catoni (2016); Catoni & Giulini (2017); Minsker (2018). Computationally tractable truncation/clipping methods include Wei & Minsker (2017), the U-statistic refinements of Minsker & Wei (2020), and the survey and automatic-calibration viewpoint of Ke et al. (2019); complementary entrywise and contamination-focused results include Avella-Medina et al. (2018); Chen et al. (2018); Minsker & Wang (2024). On the sharp statistical front, Mendelson & Zhivotovskiy (2020) obtain Gaussian-type effective-rank performance under L_4–L_2 norm equivalence, while Oliveira & Rico (2024) and Abdalla & Zhivotovskiy (2026) establish broader optimal guarantees under heavy tails and contamination. The robust mean estimators that motivate our bias-UCB construction trace back to Huber (1964); Catoni (2012); Devroye et al. (2016); Lugosi & Mendelson (2019), and the simple Huber-style heuristic instantiation is in the spirit of Sun et al. (2020); Wang et al. (2021). Our variance certificate relies on the matrix empirical-Bernstein inequality of Wang & Ramdas (2025).

Notation. For a vector $v \in \mathbb{R}^d$, $\|v\|_2$ denotes its Euclidean norm. For a symmetric matrix $A \in \mathbb{R}^{d\times d}$, $\|A\|_{\mathrm{op}}$ denotes its operator norm, $\mathrm{tr}(A)$ its trace, and $A \succeq 0$ indicates that A is positive semidefinite. For a scalar $x \in \mathbb{R}$, $(x)_+ := \max\{x, 0\}$ denotes the positive part. For $a, b \in \mathbb{R}$, $a \vee b := \max\{a, b\}$. For a finite set S, $|S|$ denotes its cardinality. We use $\mathbb{E}_P$, Var_P, and $\mathbb{P}$ to denote expectation, variance, and probability with respect to the underlying distribution P. The indicator function of an event $\mathcal{E}$ is denoted by $\mathbf{1}\{\mathcal{E}\}$. For a collection of random variables X, $\sigma(X)$ denotes the σ-algebra generated by X.

2 Method and Main Results

The purpose of this section is to turn the clipping bias–variance tradeoff into a computable score. The development has three steps. We first construct a cross-fitted family of clipped estimators together with a simultaneous variance certificate; we then show that Euclidean clipping converts operator-norm bias control into estimation of a scalar tail-energy quantity; finally we combine the two pieces into a selector with a finite-sample near-oracle guarantee.

2.1 Cross-fitted clipping

We first construct the candidate clipped estimators in a way that exposes their variance on held-out data. The role of the split is purely methodological: it lets us choose the clipping radius on one half and certify the stochastic fluctuation on the other half.

Let Π denote a random permutation of the sample indices drawn independently of the data values. We use Π to retain the first $n_{\mathrm{ret}} := 4\lfloor n/4 \rfloor$ indices, and split them into two disjoint test folds I_1 and I_2 of equal size. The corresponding training folds are $J_1 := I_2$ and $J_2 := I_1$. Write $n_k := |I_k|$, so $n_1 = n_2 = n_{\mathrm{ret}}/2$. To account for the split randomness, we define the filtration $\mathcal{F}_k := \sigma(\Pi, \{Z_i\}_{i\in J_k})$. Because the clipping radius on fold k is computed entirely from J_k and Π, it is $\mathcal{F}_k$-measurable; conditional on $\mathcal{F}_k$, the clipped observations on I_k are independent.

We tune over a geometric grid of tail probabilities with ratio $\rho > 1$ and fix

$$\gamma_{\max} := 0.3, \qquad \gamma_{\min} := \min\Big\{0.25, \frac{1}{\min_k |J_k|}\Big\}.$$

The positive grid is

$$\mathcal{G} := \{\gamma_{\max}/\rho^j : j = 0, 1, 2, \dots,\ \gamma_{\max}/\rho^j \geq \gamma_{\min}(1 - 10^{-12})\}.$$

For $\gamma \in \mathcal{G}$, fold k, and the ordered training radii $R^{(k)}_{(1)} \le \cdots \le R^{(k)}_{(|J_k|)}$, define $\tau_k(\gamma) := R^{(k)}_{(|J_k|-q_{k,\gamma})}$ where $q_{k,\gamma} := \lfloor \gamma |J_k| \rfloor$. On the test fold I_k, the clipped covariance is

$$\hat{\Sigma}^{\mathrm{raw}}_k(\gamma) := \frac{1}{n_k} \sum_{i \in I_k} \tilde{Z}^{(k)}_i(\gamma) \tilde{Z}^{(k)}_i(\gamma)^\top, \qquad \tilde{Z}^{(k)}_i(\gamma) := \begin{cases} Z_i \min\{1, \tau_k(\gamma)/\|Z_i\|_2\}, & \|Z_i\|_2 > 0, \\ 0, & \|Z_i\|_2 = 0. \end{cases} \tag{2}$$

with aggregate estimator $\hat{\Sigma}^{\mathrm{raw}}(\gamma) = \sum_{k=1}^{2} \frac{n_k}{n_{\mathrm{ret}}} \hat{\Sigma}^{\mathrm{raw}}_k(\gamma)$.

For variance control, define the normalized matrices $A^{(k)}_i(\gamma) := \tilde{Z}^{(k)}_i(\gamma) \tilde{Z}^{(k)}_i(\gamma)^\top / \tau_k(\gamma)^2$ for $\tau_k(\gamma) > 0$, and set $A^{(k)}_i(\gamma) = 0$ when $\tau_k(\gamma) = 0$. Following Wang & Ramdas (2025), fix any ordering $I_k = \{i_{k,1}, \ldots, i_{k,n_k}\}$ and define the conditional variance proxy

$$V^\star_k(\gamma) := \frac{1}{n_k} \sum_{j=1}^{n_k/2} \Big(A^{(k)}_{i_{k,2j-1}}(\gamma) - A^{(k)}_{i_{k,2j}}(\gamma) \Big)^2, \tag{3}$$

and the matrix empirical-Bernstein radius

$$D_{m,d}(\alpha; V) := \frac{\log\big(\frac{md}{(m-1)\alpha}\big)}{3m} + \sqrt{\frac{2\,\|V\|_{\mathrm{op}} \log\big(\frac{md}{(m-1)\alpha}\big)}{m}} + \Big(\sqrt{\tfrac{5}{3}} + 1\Big) \frac{\sqrt{\log\big(\frac{md}{(m-1)\alpha}\big) \log\big(\frac{2md}{\alpha}\big)}}{m}. \tag{4}$$

Using $\alpha_\gamma := \delta_{\mathrm{var}}/(4|\mathcal{G}|)$, define

$$\Psi^{\mathrm{raw}}_k(\gamma) := \tau_k(\gamma)^2 D_{n_k,d}(\alpha_\gamma; V^\star_k(\gamma)), \qquad \Psi^{\mathrm{raw}}(\gamma) := \sum_{k=1}^{2} \frac{n_k}{n_{\mathrm{ret}}} \Psi^{\mathrm{raw}}_k(\gamma). \tag{5}$$

Over the increasing grid $\gamma^{(1)} < \cdots < \gamma^{(|\mathcal{G}|)}$, define the suffix maximum $\overline{\Psi}^{\mathrm{raw}}(\gamma^{(j)}) := \max_{t \ge j} \Psi^{\mathrm{raw}}(\gamma^{(t)})$.

Theorem 1 (Uniform variance certificate)**.** *Assume $Z_1, \ldots, Z_n$ are i.i.d. from P. Define*

$$\bar{\Sigma}^{\mathrm{raw}}_P(\gamma) := \sum_{k=1}^{2} \frac{n_k}{n_{\mathrm{ret}}} \mathbb{E}_P\big[\hat{\Sigma}^{\mathrm{raw}}_k(\gamma) \mid \mathcal{F}_k\big].$$

With probability at least $1 - \delta_{\mathrm{var}}$, simultaneously for all $\gamma \in \mathcal{G}$,

$$\Big\| \hat{\Sigma}^{\mathrm{raw}}(\gamma) - \bar{\Sigma}^{\mathrm{raw}}_P(\gamma) \Big\|_{\mathrm{op}} \le \overline{\Psi}^{\mathrm{raw}}(\gamma). \tag{6}$$

Theorem 1 says that, once γ is fixed, the deviation of the clipped estimator from its fold-conditional mean is simultaneously observable over the grid.

For the bias analysis, we use a notational device to avoid tying a conditional expectation to a literal sample index. For each fold k and grid point γ, let $Z^{\mathrm{test}}_k \sim P$ be independent of $\mathcal{F}_k$, and define

$$Y^{\mathrm{test}}_k(\gamma) := (\|Z^{\mathrm{test}}_k\|_2^2 - \tau_k(\gamma)^2)_+.$$

Conditional on $\mathcal{F}_k$, this is exactly the law of any $Y^{(k)}_i(\gamma)$ with $i \in I_k$.

2.2 Tail-energy score

Variance is now under control, but the fold-conditional mean still differs from Σ because clipping removes radial mass. The next lemma is the structural step that makes the method possible: for Euclidean clipping, this matrix bias is positive semidefinite and its size is governed by a scalar tail-energy functional.

Lemma 2 (Tail-energy representation of the bias)**.** *Fix a fold k and a grid point γ, and let*

$$B_k(\gamma) := \Sigma - \mathbb{E}_P\big[\hat{\Sigma}_k^{\mathrm{raw}}(\gamma) \mid \mathcal{F}_k\big].$$

Then $B_k(\gamma) \succeq 0$ almost surely and

$$\|B_k(\gamma)\|_{\mathrm{op}} \le \mathrm{tr}(B_k(\gamma)) = \mathbb{E}_P\big[Y_k^{\mathrm{test}}(\gamma) \mid \mathcal{F}_k\big]. \tag{7}$$

Consequently, for

$$b(\gamma) := \sum_{k=1}^{2} \frac{n_k}{n_{\mathrm{ret}}} \mathbb{E}_P\big[Y_k^{\mathrm{test}}(\gamma) \mid \mathcal{F}_k\big],$$

we have $\big\|\Sigma - \bar{\Sigma}_P^{\mathrm{raw}}(\gamma)\big\|_{\mathrm{op}} \le b(\gamma)$.

For each fold k and grid point γ, define the held-out tail-excess variables

$$Y_i^{(k)}(\gamma) := (\|Z_i\|_2^2 - \tau_k(\gamma)^2)_+, \qquad i \in I_k. \tag{8}$$

Conditional on $\mathcal{F}_k$, the variables $\{Y_i^{(k)}(\gamma)\}_{i \in I_k}$ are i.i.d., each with the same conditional law as $Y_k^{\mathrm{test}}(\gamma)$. We split the confidence budget as $\delta_{\mathrm{var}} = \delta_{\mathrm{bias}} = \delta/2$ and define the baseline number of blocks by

$$m_{\mathrm{base}} := \left\lceil 4 \log\left(\frac{4|\mathcal{G}|}{\delta_{\mathrm{bias}}}\right) \right\rceil. \tag{9}$$

Set $m_{\mathrm{blocks}} := \min\{m_{\mathrm{base}}, n_1\}$. For the theorem below we impose the mild block-feasibility condition

$$m_{\mathrm{base}} \le n_1, \tag{10}$$

which, because $n_1 = n_2$ by construction, guarantees that every block on each fold contains at least one sample and makes $m_{\mathrm{blocks}} = m_{\mathrm{base}}$. For each (k, γ), we independently shuffle $\{Y_i^{(k)}(\gamma)\}_{i \in I_k}$, partition them into m_{blocks} groups whose sizes differ by at most one, and let $\widehat{b}_k^{\mathrm{QTES}}(\gamma)$ be the empirical 85% quantile of the resulting block means. We then set

$$\widehat{b}^{\mathrm{QTES}}(\gamma) := \sum_{k=1}^{2} \frac{n_k}{n_{\mathrm{ret}}} \widehat{b}_k^{\mathrm{QTES}}(\gamma), \qquad \widehat{S}_{\mathrm{QTES}}(\gamma) := \overline{\Psi}^{\mathrm{raw}}(\gamma) + \widehat{b}^{\mathrm{QTES}}(\gamma), \qquad \hat{\gamma}_{\mathrm{QTES}} \in \arg\min_{\gamma \in \mathcal{G}} \widehat{S}_{\mathrm{QTES}}(\gamma). \tag{11}$$

The score $\widehat{S}_{\mathrm{QTES}}(\gamma)$ has a direct interpretation: $\overline{\Psi}^{\mathrm{raw}}(\gamma)$ is the observable variance certificate, whereas $\widehat{b}^{\mathrm{QTES}}(\gamma)$ estimates the covariance mass discarded by clipping. Large γ typically decreases the first term and increases the second, while small γ does the reverse. A generic oracle template that replaces $\widehat{b}^{\mathrm{QTES}}$ by any valid upper bound on the conditional tail energy is stated in Appendix B. The theorem below specializes this template to QTES.

2.3 Main result

The next theorem formalizes this interpretation. Uniformly over the grid, the QTES score tracks the ideal bias–variance target up to a residual Δ_n, so minimizing the score yields a selected estimator with near-oracle risk over the grid. It also implies a candidate-level near-optimality statement for the ideal curve $O(\gamma) = \overline{\Psi}^{\mathrm{raw}}(\gamma) + b(\gamma)$, but not for the realized oracle candidate ranked by sample error.

Define

$$s_\star := \left\lfloor \frac{n_1}{m_{\mathrm{blocks}}} \right\rfloor, \qquad M_4 := \big(\mathbb{E}_P \|Z_1\|_2^4\big)^{1/4}, \qquad \Delta_n := 10 M_4^2\, s_\star^{-1/2}.$$

Theorem 3 (Uniform QTES bound and selected-estimator guarantee)**.** *Assume* (1) *and* (10)*. Then with probability at least $1 - \delta$, simultaneously for all $\gamma \in \mathcal{G}$,*

$$\big|\widehat{b}^{\mathrm{QTES}}(\gamma) - b(\gamma)\big| \le \Delta_n, \qquad \Big\|\hat{\Sigma}^{\mathrm{raw}}(\gamma) - \Sigma\Big\|_{\mathrm{op}} \le \widehat{S}_{\mathrm{QTES}}(\gamma) + \Delta_n. \tag{12}$$

Consequently,

$$\left\|\hat{\Sigma}^{\mathrm{raw}}(\hat{\gamma}_{\mathrm{QTES}}) - \Sigma\right\|_{\mathrm{op}} \le \min_{\gamma\in\mathcal{G}} \left\{\overline{\Psi}^{\mathrm{raw}}(\gamma) + b(\gamma)\right\} + 2\Delta_n. \tag{13}$$

The detailed calibration statement and the generic oracle template are collected in Appendix B.

Theorem 3 controls the risk of the selected estimator itself. It is also useful to isolate what the same event implies for the selected *grid point*. The next corollary shows that the clipping level chosen by QTES nearly minimizes the ideal bias–variance curve on $\mathcal{G}$, and therefore coincides with the exact minimizer whenever the gap at the optimum dominates the calibration slack.

Corollary 4 (Near-optimal selected candidate for the ideal curve)**.** *Define*

$$O(\gamma) := \overline{\Psi}^{\mathrm{raw}}(\gamma) + b(\gamma), \qquad \gamma \in \mathcal{G}.$$

Under the same event as Theorem 3,

$$O(\hat{\gamma}_{\mathrm{QTES}}) \le \min_{\gamma\in\mathcal{G}} O(\gamma) + 2\Delta_n. \tag{14}$$

Hence for every $\varepsilon \ge 2\Delta_n$, the selected grid point belongs to the ε-near-minimizer set

$$\Gamma_\varepsilon := \{\gamma \in \mathcal{G} : O(\gamma) \le \min_{t\in\mathcal{G}} O(t) + \varepsilon\}.$$

In particular, if $\gamma_\star$ is a unique minimizer of O and

$$\min_{\gamma\in\mathcal{G}\setminus\{\gamma_\star\}} \big(O(\gamma) - O(\gamma_\star)\big) > 2\Delta_n,$$

then $\hat{\gamma}_{\mathrm{QTES}} = \gamma_\star$.

The proofs of Theorem 3 and Corollary 4 are deferred to Appendix C.

Algorithm 1 summarizes the resulting procedure.

Algorithm 1 Cross-fitted QTES on the positive clipping grid

Require: Centered samples $Z_1, \dots, Z_n \in \mathbb{R}^d$, confidence $\delta \in (0,1)$, grid ratio $\rho = 2.0$.
Ensure: Tuned covariance estimate $\hat{\Sigma}^{\mathrm{raw}}(\hat{\gamma}_{\mathrm{QTES}})$.

```
Set δ_var = δ_bias = δ/2.
Retain n_ret = 4⌊n/4⌋ observations; split evenly into test folds I_1, I_2; define training folds J_1 = I_2, J_2 = I_1.
Set γ_max = 0.3, γ_min = min{0.25, 1/min_k |J_k|}, and build the positive geometric grid G = {γ_max/ρ^j} down to γ_min.
Compute m_base = ⌈4 log(4|G|/δ_bias)⌉ and set m_blocks = min{m_base, n_1}.
for γ ∈ G do
    for k ∈ {1,2} do
        Compute τ_k(γ) from the training-fold radii and assemble Σ̂_k^raw(γ) via (2).
        Compute V_k^⋆(γ) and Ψ_k^raw(γ) via (3)–(5).
        Form Y_i^(k)(γ) = (‖Z_i‖_2^2 − τ_k(γ)^2)_+ for i ∈ I_k.
        Randomly partition I_k into m_blocks blocks and compute the mean of Y on each block.
        Set b̂_k^QTES(γ) to the empirical 85% quantile of these block means.
    end for
    Aggregate Ψ^raw(γ) and b̂^QTES(γ).
end for
Monotonize the variance radius: Ψ̄^raw(γ^(j)) = max_{t≥j} Ψ^raw(γ^(t)).
Select γ̂_QTES ∈ arg min_{γ∈G} Ŝ_QTES(γ).
Return Σ̂^raw(γ̂_QTES).
```

Relation to optimal direct rates. A useful way to interpret Theorem 3 is to compare its remainder with the optimal direct operator-norm rates in the literature. Let $r(\Sigma) := \mathrm{tr}(\Sigma)/\|\Sigma\|_{\mathrm{op}}$. For Gaussian data, and

more generally for light-tailed models of the type studied by Koltchinskii & Lounici (2017), the canonical operator-norm rate is

$$\|\Sigma\|_{\mathrm{op}}\big(\sqrt{r(\Sigma)/n} \vee r(\Sigma)/n\big).$$

Under heavier tails, the same effective-rank order is attainable by direct robust procedures under stronger regularity assumptions: Mendelson & Zhivotovskiy (2020) under L_4–L_2 norm equivalence, and Oliveira & Rico (2024); Abdalla & Zhivotovskiy (2026) under broader L_p–L_2 assumptions and contamination. By contrast, our theorem yields

$$\Big\|\hat{\Sigma}^{\mathrm{raw}}(\hat{\gamma}_{\mathrm{QTES}}) - \Sigma\Big\|_{\mathrm{op}} \le \min_{\gamma\in\mathcal{G}} O(\gamma) + 2\Delta_n, \qquad \Delta_n = 10\big(\mathbb{E}_P\|Z_1\|_2^4\big)^{1/2} s_\star^{-1/2}.$$

Since $s_\star$ is of order $n/\log(|\mathcal{G}|/\delta)$, the selector remainder is essentially

$$\big(\mathbb{E}\|Z_1\|_2^4\big)^{1/2}\sqrt{\frac{\log(|\mathcal{G}|/\delta)}{n}}.$$

Thus the proof controls the cost of choosing γ through a global fourth-moment term, rather than directly at the effective-rank scale. If one additionally assumes the usual κ-L_4–L_2 comparison from the optimal-rate papers, then the crude bound

$$\big(\mathbb{E}\|Z_1\|_2^4\big)^{1/2} \le \kappa^2\,\mathrm{tr}(\Sigma) = \kappa^2 r(\Sigma)\|\Sigma\|_{\mathrm{op}}$$

implies that our selector remainder is at most a constant multiple of

$$\kappa^2 r(\Sigma)\|\Sigma\|_{\mathrm{op}}\sqrt{\frac{\log(|\mathcal{G}|/\delta)}{n}},$$

which is typically looser than the optimal direct rate by roughly a factor $\sqrt{r(\Sigma)}$ in the dominant regime. The present theorem should therefore be read as a finite-sample calibration guarantee for tuning, not yet as a sharp direct covariance-estimation rate.

3 Experiments

The experiments are organized around some questions. First, does QTES rank candidate clipping levels better than an uninformed choice, and how close does it come to the in-grid oracle? Second, at $(n, d) = (400, 200)$, does the selected estimator remain competitive across both clean heavy-tailed samples and gross contamination? Third, how does the data-driven selector behave asymptotically as the sample size increases relative to the ambient dimension? Unless stated otherwise, all reported table entries are averages over repeated Monte Carlo trials. The primary metric is the relative spectral error

$$\mathrm{RelSpecErr}(\hat{\Sigma}, \Sigma) := \frac{\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}}.$$

For eigenspace recovery we also report

$$\mathrm{SubspaceErr}(\hat{\Sigma}, \Sigma; s) := \|P_s(\hat{\Sigma}) - P_s(\Sigma)\|_{\mathrm{op}},$$

where $P_s(A)$ is the orthogonal projector onto the leading s-dimensional eigenspace of A. By the Davis–Kahan theorem (Yu et al., 2015), this subspace distance is fundamentally controlled by the operator-norm error of the covariance estimate provided there is a sufficient eigengap. Runtime is mean wall-clock time per fit. We fix $\gamma_{\max} = 0.3$ and evaluate the positive grid $\mathcal{G}$ used in the theory.

3.1 Experimental setup

We keep the signal model fixed across experiments so that changes in performance can be attributed to the distributional regime and the tuning problem rather than to a moving target covariance.

The target covariance is a spiked model

$$\Sigma = Q\,\mathrm{diag}(1+\theta,\ldots,1+\theta,1,\ldots,1)\,Q^\top$$

with rank $s = 5$, spike strength $\theta = 10$, and a single random orthogonal matrix Q fixed across trials within each experiment. Data is generated as $Z_i = \Sigma^{1/2}\xi_i$, where $\Sigma^{1/2}$ denotes the symmetric positive semidefinite square root of Σ, and the coordinates of ξ_i are standardized and independent. We consider three coordinate laws: Gaussian, Student-$t(5)$ rescaled to unit variance, and standardized lognormal. All synthetic distributions are mean zero, and throughout the experiments we treat this mean as known: the raw draws are passed to every estimator without empirical recentering.

The selector study and the benchmark comparison both use $(n, d) = (400, 200)$. The selector study uses 100 trials per distribution. In the benchmark comparison we use 100 clean trials per distribution and 100 trials for each contamination level. The empirical comparison is intentionally restricted to QTES, SCM, OAS, MCD, and an oracle-tuned Wei–Minsker baseline (Wei & Minsker, 2017). Methods with an explicit centering option are run in their centered/known-mean mode. For the Wei–Minsker baseline we deliberately give the method oracle access to a single scale σ_0 by setting $\sigma_{\min} = \sigma_{\max} = \sigma_0$, so its Lepski grid collapses to a single oracle-tuned truncation level. On the clean benchmarks σ_0 is computed from the known data-generating distribution. In the contamination studies the target remains the uncontaminated covariance Σ, so we again use the clean inlier lognormal scale rather than the contaminated-mixture scale. This is intentional data leakage: WM-oracle should be interpreted as a best-case upper benchmark within the centered clipped family, not as a fair data-driven competitor or as the theorem-optimal σ_0 for the observed mixture law.

3.2 Selector quality

This experiment isolates the tuning problem at $(n, d) = (400, 200)$, using 100 independent trials for each distribution. The theory controls the ideal criterion

$$O(\gamma) = \overline{\Psi}^{\mathrm{raw}}(\gamma) + b(\gamma),$$

whereas Table 1 evaluates the clipping level selected from the finite grid $\mathcal{G}$ against the realized in-grid oracle under relative spectral error.

The columns are interpreted as follows. *QTES* is the mean relative spectral error achieved by the clipping level selected by our procedure. *Random-γ* is the expected error of a uniformly random choice from $\mathcal{G}$, implemented trialwise as the grid-average error. *Oracle* is the mean error of the best grid point in hindsight on each trial. *Sel./Oracle* is the ratio between the selected error and the oracle error, so values closer to 1 are better. *Top-2 Hit* is the fraction of trials in which the selected γ lies among the two best grid points, so larger values are better.

Table 1: Selector diagnostics at $(n, d) = (400, 200)$ over 100 trials per distribution

Distribution	QTES	Random-γ	Oracle	Sel./Oracle	Top-2 Hit
Gaussian	0.3330	0.3384	0.3122	1.0660	0.5000
Student-$t(5)$	0.3335	0.3410	0.3130	1.0657	0.5000
Lognormal	0.3492	0.3713	0.3296	1.0585	0.5800

Overall, QTES consistently improves on the random baseline and stays close to the in-grid oracle across all three distributions: the selected/oracle ratio ranges from 1.0585 to 1.0660, and the Top-2 hit rate ranges from 0.50 to 0.58. The selector should therefore be interpreted as a stable oracle-approximation rule rather than an exact oracle-recovery mechanism.

3.3 Benchmark

We now evaluate end-to-end covariance estimation at $(n, d) = (400, 200)$. All results in this subsection are averages over 100 independent trials per setting. The goal is to compare QTES with standard baselines under heavy-tailed sampling.

Table 2 compares QTES with SCM, OAS, MCD, and WM-oracle under clean sampling. Within each distribution block, *RelSpecErr* is the relative operator-norm error, *SubspaceErr* is the operator-norm difference between the leading-s eigenspace projectors, and *Time (s)* is the mean wall-clock time per fit. The first two columns measure statistical accuracy, while the last column records computational cost.

Table 2: Clean-data benchmark at $(n, d) = (400, 200)$ over 100 trials per distribution

	Gaussian			Student-$t(5)$			Lognormal		
Method	RelSpecErr	SubspaceErr	Time (s)	RelSpecErr	SubspaceErr	Time (s)	RelSpecErr	SubspaceErr	Time (s)
QTES	**0.3330**	0.2625	0.0300	**0.3335**	**0.2618**	0.0303	0.3492	**0.2768**	0.0269
SCM	0.3534	**0.2621**	0.0004	0.3618	0.2640	0.0004	0.5994	0.3665	0.0004
OAS	0.3919	**0.2621**	0.0081	0.3909	0.2640	0.0081	0.5045	0.3665	0.0075
MCD	0.4101	0.3021	0.6514	0.4138	0.3045	0.6474	0.4139	0.3421	0.7239
WM-oracle	0.3487	0.2622	0.0007	0.3451	0.2621	0.0007	**0.3417**	0.2797	0.0007

The benchmark shows the expected trade-off, but with sharper separation under 100 trials. QTES attains the smallest relative spectral error on Gaussian and Student-$t(5)$, and on lognormal it remains within 0.0075 of WM-oracle while delivering the best subspace recovery. SCM and OAS remain competitive on the Gaussian subspace metric and are substantially faster, while MCD is uniformly much slower and does not provide a compensating accuracy advantage on this benchmark.

3.4 Asymptotic Behavior and Empirical Calibration

To investigate the large-sample behavior of the fully data-driven QTES method and contextualize it against the theoretically tuned benchmark (WM-oracle), we conduct an expanding-sample experiment. We fix the dimension at $d = 200$ and vary the sample size $n \in \{100, 200, 400, 800, 1600, 3200, 5000\}$. The target covariance and data generating mechanisms remain identical to those in Section 3.1. Each (distribution, n, method) setting is averaged over 100 independent trials. The WM baseline is provided with the exact true scale σ_0 of the clean distribution, serving as an ideal theoretical benchmark for this family of radial clipping estimators.

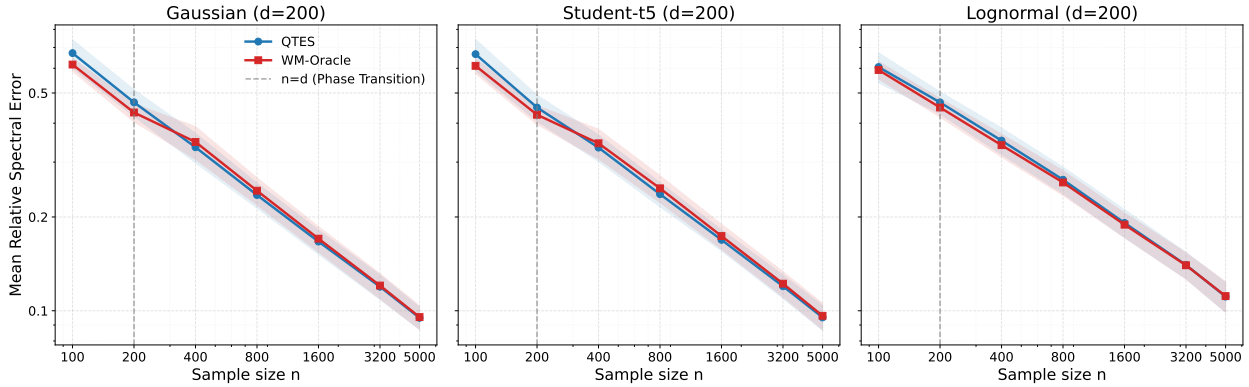

Figure 1: Asymptotic relative spectral error of QTES and the WM-oracle at a fixed dimension $d = 200$. The vertical dashed line indicates the phase transition threshold $n = d$. Shaded regions represent ± 1 standard deviation over 100 independent trials. Both axes are on a logarithmic scale.

The results, illustrated in Figure 1, reveal a clear sample-complexity phase transition around $n = d$, highlighting the tradeoff between analytical relaxation and empirical calibration.

In the sample-starved regime ($n < d$), QTES is consistently more conservative than the WM-oracle under all three distributions. At $n = 100$, the mean relative spectral errors are 0.6714 versus 0.6169 for Gaussian data, 0.6664 versus 0.6111 for Student-$t(5)$, and 0.6057 versus 0.5927 for lognormal data. This behavior matches the mathematical cost of the trace relaxation ($\|B\|_{\mathrm{op}} \leq \mathrm{tr}(B)$) used in Lemma 2: when the dimension is

large relative to the sample size, the trace inflation makes the QTES score overestimate the clipping bias and therefore encourages conservative under-clipping.

Once n reaches the ambient dimension, however, the gap collapses rapidly. At $n = 400$, QTES already slightly outperforms WM-oracle for Gaussian data (0.3348 versus 0.3482) and Student-$t(5)$ data (0.3339 versus 0.3448), while remaining close on lognormal data (0.3517 versus 0.3401). By $n = 5000$, the two methods are nearly indistinguishable across all three distributions; for example, on lognormal data the mean errors are 0.1110 for QTES and 0.1114 for WM-oracle.

3.5 Empirical Contamination Stress Test

To evaluate the resilience of QTES beyond its theoretical comfort zone, we conduct an empirical stress test under gross contamination.

We start from standardized lognormal inliers and replace 5% of the sample by Gaussian outliers drawn from $\mathcal{N}(0, c\Sigma)$ with inflation factor $c \in \{1, 25, 100, 400\}$. Each contamination level is averaged over 100 independent trials. Figure 2 summarizes the resulting relative spectral errors along this contamination path.

The main qualitative conclusion is unchanged, but the 100-trial averages make the separation especially clear. Across $c \in \{1, 25, 100, 400\}$, QTES attains mean relative spectral errors 0.3484, 0.3833, 0.3911, and 0.3873, compared with 0.3406, 0.3563, 0.3549, and 0.3547 for WM-oracle and 0.4116–0.4270 for MCD. By contrast, SCM and OAS deteriorate catastrophically as the outliers are inflated, rising from 0.5867/0.4824 at $c = 1$ to 56.1519/53.5103 at $c = 400$. QTES therefore remains the strongest fully data-driven method in this experiment, while the WM-oracle curve remains the best reference point because it is given the clean inlier scale.

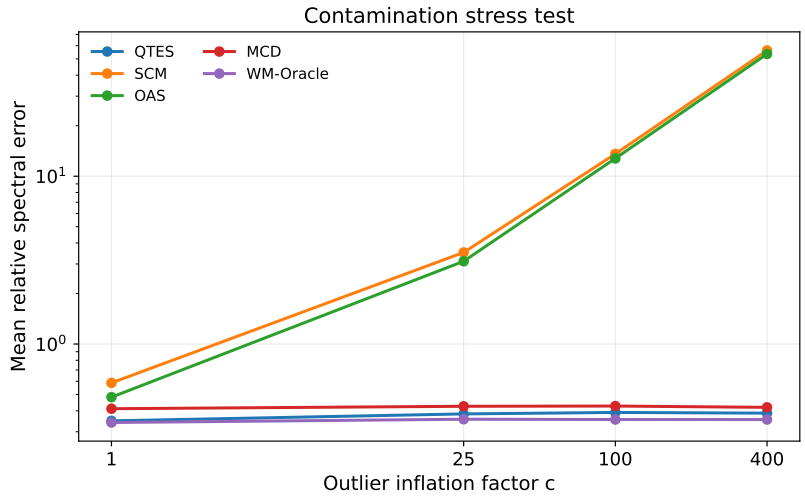

Figure 2: Gross-contamination stress test at $(n, d) = (400, 200)$ over 100 independent trials per contamination level.

4 Discussion and conclusion

The central contribution of this paper is a simple and observable way to tune clipping in robust covariance estimation. Once the sample split is fixed, QTES combines two quantities that are both computable from the data: a variance certificate built from held-out clipped matrices and a scalar tail-energy proxy for the bias introduced by clipping. This makes the bias–variance tradeoff operational without requiring a user-supplied oracle range for an unknown scale. Empirically, the resulting calibration principle performs well across the regimes we study. On the clean benchmarks, QTES is best or near-best among the fully data-driven procedures; under contamination, it remains the strongest fully data-driven method in our experiments.

The experiments also clarify what should and should not be concluded from the comparison with WM-oracle. Once that benchmark is supplied with the clean oracle scale, it becomes an intentionally favorable reference point for this centered clipped family. Accordingly, the remaining gap between QTES and WM-oracle is most naturally interpreted as a calibration gap within the family, not as evidence that the clipped family itself is fundamentally inadequate. In that sense, the contamination results support the main methodological claim of the paper: the dominant difficulty is selecting the clipping level from the data, and QTES addresses that problem in a transparent and implementable way.

On the theoretical side, the contribution is more specific. For a simple computable clipped family, Euclidean clipping reduces operator-norm bias control to a scalar tail-energy term, and QTES yields a uniform finite-sample near-oracle calibration guarantee on a positive grid under a finite fourth-moment condition together with the mild block-feasibility requirement used by the procedure. What the current analysis does not yet provide is a sharp direct rate for the final selected estimator itself. The theorem is selector-relative: it shows that the data-driven score tracks the best candidate on the grid up to calibration slack, but it does not yet place the selected estimator on the effective-rank scale achieved by sharper direct robust covariance estimators under stronger assumptions.

This gap points to the main open problem. To approach the sharp direct theory while remaining fully observable, one would need a selector analysis driven by the local tail behavior near the oracle clipping radius rather than by a global fourth-moment remainder. Put differently, the missing ingredient is not obviously a more elaborate clipped family, but a sharper way to analyze data-driven calibration itself.

There are also two practical limitations. First, the current analysis is carried out in the centered, known-mean setting. Extending the same observable calibration logic to unknown-mean procedures, ideally without sacrificing computational simplicity, remains an important next step. Second, the randomized split introduces visible variability under contamination. A simple empirical stabilization is to rerun QTES with several independent cross-fitting permutations and aggregate the resulting covariance estimates by an operator-norm medoid, in the spirit of the metric-space median ideas of Hsu & Sabato (2016). We view this as a practical recommendation rather than a theorem-backed amplification device, because the repeated outputs are coupled through the common sample and the present theory does not provide the conditional success probabilities needed for a formal exponential-improvement argument.

A related extension is to decouple selection from final estimation. From a practical standpoint, splitting is arguably only needed for choosing the clipping fraction. This suggests a natural "select-then-refit" variant: use QTES to select $\hat{\gamma}_{\mathrm{QTES}}$, recompute a single clipping radius from all n observations, and then refit the covariance on the full clipped sample. We did not benchmark this variant systematically in the reported tables or figures. In untabulated exploratory runs, however, the post-selection full-sample refit was often slightly more accurate than the cross-fitted estimator. This also suggests that QTES may be useful as a generic heavy-tail preprocessing step before downstream structured procedures such as shrinkage or sparse precision-matrix estimation. The obstacle is theoretical rather than algorithmic: once the full sample is reused for both calibration and estimation, the conditional independence exploited by our proof disappears. Establishing guarantees of the same flavor would likely require genuinely uniform empirical-process control over a data-dependent clipping family, which appears substantially more delicate.

Overall, QTES contributes a practical calibration principle to the robust covariance toolbox. It is simple to implement, requires no rough oracle range for a hidden tail scale, and comes with a finite-sample selector guarantee under a finite fourth-moment assumption. The most promising next steps are to sharpen the selector remainder from a global to a local oracle-scale quantity, extend the method beyond the known-mean setting, and understand when stabilization or full-sample refitting can be added without losing theoretical control.

References

Pedro Abdalla and Nikita Zhivotovskiy. Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *Journal of the European Mathematical Society*, 28(4):1809–1847, 2026. doi: 10.4171/JEMS/1505. Published online in 2024.

Marco Avella-Medina, Heather S. Battey, Jianqing Fan, and Quefeng Li. Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, 105(2):271–284, 2018. doi: 10.1093/biomet/asy011.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 48(4):1148–1185, 2012. doi: 10.1214/11-AIHP454.

Olivier Catoni. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016. URL `https://arxiv.org/abs/1603.05229`.

Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017. URL `https://arxiv.org/abs/1712.02747`.

Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018. doi: 10.1214/17-AOS1607.

Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016. doi: 10.1214/16-AOS1440.

Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20(none):1 – 29, 2015. doi: 10.1214/EJP.v20-3760. URL `https://doi.org/10.1214/EJP.v20-3760`.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4): 603–680, 2013. doi: 10.1111/rssb.12016.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016. URL `https://jmlr.org/papers/v17/14-273.html`.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1): 73–101, 1964. doi: 10.1214/aoms/1177703732.

Mia Hubert, Peter J. Rousseeuw, and Stefan Van Aelst. High-breakdown robust multivariate methods. *Statistical Science*, 23(1):92–119, 2008. doi: 10.1214/088342307000000087.

Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001. doi: 10.1214/aos/1009210544.

Yuan Ke, Stanislav Minsker, Zhao Ren, Qiang Sun, and Wen-Xin Zhou. User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, 34(3):454–471, 2019. doi: 10.1214/19-STS711.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017. doi: 10.3150/15-BEJ730.

Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3): 1029 – 1058, 2014. doi: 10.3150/12-BEJ487. URL `https://doi.org/10.3150/12-BEJ487`.

Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019. doi: 10.1214/17-AOS1639.

Ricardo A. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4 (1):51–67, 1976. doi: 10.1214/aos/1176343347.

Shahar Mendelson and Nikita Zhivotovskiy. Robust covariance estimation under L_4–L_2 norm equivalence. *The Annals of Statistics*, 48(3):1648–1664, 2020. doi: 10.1214/19-AOS1862.

Stanislav Minsker. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018. doi: 10.1214/17-AOS1642.

Stanislav Minsker and Lang Wang. Robust estimation of covariance matrices: adversarial contamination and beyond. *Statistica Sinica*, 34(3):1565–1583, 2024. doi: 10.5705/ss.202021.0388.

Stanislav Minsker and Xiaohan Wei. Robust modifications of U-statistics and applications to covariance estimation problems. *Bernoulli*, 26(1):694–727, 2020. doi: 10.3150/19-BEJ1149.

Roberto I. Oliveira and Zoraida F. Rico. Improved covariance estimation: Optimal robustness and sub-Gaussian guarantees under heavy tails. *The Annals of Statistics*, 52(5):1953–1977, 2024. doi: 10.1214/24-AOS2407.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007. URL `http://www.jstor.org/stable/24307692`.

Nikhil Srivastava and Roman Vershynin. Covariance estimation for distributions with $2+\varepsilon$ moments. *The Annals of Probability*, 41(5):3081–3111, 2013. doi: 10.1214/12-AOP760.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020. doi: 10.1080/01621459.2018.1543124.

David E. Tyler. A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, 15(1): 234–251, 1987. doi: 10.1214/aos/1176350263.

Hongjian Wang and Aaditya Ramdas. Sharp matrix empirical bernstein inequalities. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=Mgug9GUNy7`.

Lili Wang, Chao Zheng, Wen Zhou, and Wen-Xin Zhou. A new principle for tuning-free Huber regression. *Statistica Sinica*, 31(4):2153–2177, 2021. doi: 10.5705/ss.202019.0045.

Xiaohan Wei and Stanislav Minsker. Estimation of the covariance structure of heavy-tailed distributions. In *Advances in Neural Information Processing Systems 30*, pp. 2859–2868, 2017.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008.

A Proofs of the Main Variance and Bias Ingredients

This appendix groups each auxiliary statement with its proof. As in Section 2, for each fold k and grid point γ, let $Z_k^{\mathrm{test}} \sim P$ be independent of $\mathcal{F}_k$, and write

$$Y_k^{\mathrm{test}}(\gamma) := (\|Z_k^{\mathrm{test}}\|_2^2 - \tau_k(\gamma)^2)_+.$$

Conditional on $\mathcal{F}_k$, this is the same law as any held-out variable $Y_i^{(k)}(\gamma)$ with $i \in I_k$.

A.1 Proof of Theorem 1

Proof. Fix $\gamma \in \mathcal{G}$. For fold $k \in \{1,2\}$, condition on $\mathcal{F}_k := \sigma(\{Z_i\}_{i\in J_k})$. If $\tau_k(\gamma) = 0$, then $A_i^{(k)}(\gamma) = 0$, $V_k^\star(\gamma) = 0$, and $\Psi_k^{\mathrm{raw}}(\gamma) = 0$. Otherwise, the matrices $\{A_i^{(k)}(\gamma)\}_{i\in I_k}$ are conditionally i.i.d. self-adjoint with eigenvalues in $[0,1]$. For any fixed $i \in I_k$, define

$$M_k(\gamma) := \mathbb{E}_P[A_i^{(k)}(\gamma) \mid \mathcal{F}_k], \qquad \hat{M}_k(\gamma) := \frac{1}{n_k}\sum_{i\in I_k} A_i^{(k)}(\gamma).$$

Then $\hat{\Sigma}_k^{\mathrm{raw}}(\gamma) = \tau_k(\gamma)^2\,\hat{M}_k(\gamma)$ and $\mathbb{E}_P[\hat{\Sigma}_k^{\mathrm{raw}}(\gamma) \mid \mathcal{F}_k] = \tau_k(\gamma)^2\,M_k(\gamma)$. Applying Wang & Ramdas (2025, Theorem 3.1) to $\{A_i^{(k)}(\gamma)\}_{i\in I_k}$ yields the one-sided deviation bound with radius $D_{n_k,d}(\alpha_\gamma; V_k^\star(\gamma))$. Applying the same inequality to $\{I_d - A_i^{(k)}(\gamma)\}_{i\in I_k}$ gives the lower tail. A union bound over both tails, both folds, and all $|\mathcal{G}|$ points guarantees the envelope with probability at least $1-\delta_{\mathrm{var}}$. Summing with weights n_k/n_{ret} proves (6). The suffix maximum inequality is deterministic. □

A.2 Proof of Lemma 2

Proof. Fix k and γ, and define

$$\alpha(z) := \begin{cases} \min\{1, \tau_k(\gamma)/\|z\|_2\}, & \|z\|_2 > 0, \\ 1, & \|z\|_2 = 0. \end{cases}$$

Then $\alpha(z)z$ is exactly the clipped vector appearing in (2). Since Z_k^{test} is independent of $\mathcal{F}_k$ and distributed as P,

$$\mathbb{E}_P[\hat{\Sigma}_k^{\text{raw}}(\gamma) \mid \mathcal{F}_k] = \mathbb{E}_P\big[\alpha(Z_k^{\text{test}})^2 Z_k^{\text{test}}(Z_k^{\text{test}})^\top \mid \mathcal{F}_k\big].$$

Hence

$$B_k(\gamma) = \mathbb{E}_P\big[(1 - \alpha(Z_k^{\text{test}})^2) Z_k^{\text{test}}(Z_k^{\text{test}})^\top \mid \mathcal{F}_k\big].$$

Since $0 \le 1 - \alpha(Z_k^{\text{test}})^2 \le 1$, the integrand is positive semidefinite, so $B_k(\gamma) \succeq 0$. Therefore $\|B_k(\gamma)\|_{\text{op}} \le \text{tr}(B_k(\gamma))$. Moreover,

$$\text{tr}(B_k(\gamma)) = \mathbb{E}_P\big[(1 - \alpha(Z_k^{\text{test}})^2)\|Z_k^{\text{test}}\|_2^2 \mid \mathcal{F}_k\big].$$

If $\|Z_k^{\text{test}}\|_2 \le \tau_k(\gamma)$, then $\alpha(Z_k^{\text{test}}) = 1$ and the contribution is zero. If $\|Z_k^{\text{test}}\|_2 > \tau_k(\gamma)$, then $\alpha(Z_k^{\text{test}})^2 = \tau_k(\gamma)^2/\|Z_k^{\text{test}}\|_2^2$, so

$$(1 - \alpha(Z_k^{\text{test}})^2)\|Z_k^{\text{test}}\|_2^2 = \|Z_k^{\text{test}}\|_2^2 - \tau_k(\gamma)^2.$$

Thus

$$\text{tr}(B_k(\gamma)) = \mathbb{E}_P\big[Y_k^{\text{test}}(\gamma) \mid \mathcal{F}_k\big].$$

□

B Auxiliary Results and Proofs

We next collect the auxiliary theorem-proof pairs used to prove Theorem 3.

B.1 A Generic Oracle Template

We begin with a generic template. Its role is to separate variance control from bias calibration: once a simultaneous upper bound for the bias term is available, the statement below converts it into a uniform risk bound over the entire clipping grid.

Theorem 5 (Generic oracle inequality from a valid bias upper bound)**.** *Suppose a data-dependent quantity $\widehat{b}_k^{\text{UCB}}(\gamma)$ is constructed from $\{Y_i^{(k)}(\gamma) : i \in I_k\}$ such that with probability at least $1 - \delta_{\text{bias}}$,*

$$\mathbb{E}_P\big[Y_k^{\text{test}}(\gamma) \mid \mathcal{F}_k\big] \le \widehat{b}_k^{\text{UCB}}(\gamma) \qquad \textit{simultaneously for all } k \in \{1,2\},\ \gamma \in \mathcal{G}. \tag{15}$$

Define

$$\widehat{U}_{\text{BiasUCB}}(\gamma) := \overline{\Psi}^{\text{raw}}(\gamma) + \sum_{k=1}^{2} \frac{n_k}{n_{\text{ret}}} \widehat{b}_k^{\text{UCB}}(\gamma). \tag{16}$$

If the event (6) *of Theorem 1 also holds, then with probability at least $1 - \delta_{\text{var}} - \delta_{\text{bias}}$, simultaneously for all $\gamma \in \mathcal{G}$,*

$$\left\|\hat{\Sigma}^{\text{raw}}(\gamma) - \Sigma\right\|_{\text{op}} \le \widehat{U}_{\text{BiasUCB}}(\gamma). \tag{17}$$

Consequently, for $\hat{\gamma}_{\text{BiasUCB}} \in \arg\min_{\gamma \in \mathcal{G}} \widehat{U}_{\text{BiasUCB}}(\gamma)$,

$$\left\|\hat{\Sigma}^{\text{raw}}(\hat{\gamma}_{\text{BiasUCB}}) - \Sigma\right\|_{\text{op}} \le \min_{\gamma \in \mathcal{G}} \widehat{U}_{\text{BiasUCB}}(\gamma). \tag{18}$$

Proof. On the intersection of the events (6) and (15), for every $\gamma \in \mathcal{G}$,

$$\left\|\hat{\Sigma}^{\text{raw}}(\gamma) - \Sigma\right\|_{\text{op}} \le \left\|\hat{\Sigma}^{\text{raw}}(\gamma) - \bar{\Sigma}_P^{\text{raw}}(\gamma)\right\|_{\text{op}} + \left\|\bar{\Sigma}_P^{\text{raw}}(\gamma) - \Sigma\right\|_{\text{op}}.$$

The first term is bounded by $\overline{\Psi}^{\text{raw}}(\gamma)$ from Theorem 1. The second term is bounded by $b(\gamma)$ from Lemma 2, which is in turn bounded by $\sum_{k=1}^{2}(n_k/n_{\text{ret}})\widehat{b}_k^{\text{UCB}}(\gamma)$ on the event (15). This proves (17). Taking the minimum over γ proves (18). □

B.2 Calibration of the QTES Bias Score

The next theorem verifies that the QTES bias score satisfies exactly the type of simultaneous calibration required by Theorem 5. This is the main technical input behind the bias part of Theorem 3.

Theorem 6 (Uniform calibration of the QTES bias score)**.** *Assume* (1) *and* (10)*, and define*

$$s_\star := \left\lfloor \frac{n_1}{m_{\text{blocks}}} \right\rfloor.$$

For each fold k and grid point γ, let

$$\mu_k(\gamma) := \mathbb{E}_P\big[Y_k^{\text{test}}(\gamma) \mid \mathcal{F}_k\big], \qquad v_k(\gamma) := \mathrm{Var}_P\left(Y_k^{\text{test}}(\gamma) \mid \mathcal{F}_k\right).$$

Then with probability at least $1 - \delta_{\text{bias}}$, simultaneously for all $k \in \{1, 2\}$ and $\gamma \in \mathcal{G}$,

$$\left|\widehat{b}_k^{\text{QTES}}(\gamma) - \mu_k(\gamma)\right| \le 10\sqrt{\frac{v_k(\gamma)}{s_\star}}. \tag{19}$$

In particular, since $Y_i^{(k)}(\gamma) \le \|Z_i\|_2^2$, the same event implies

$$\left|\widehat{b}_k^{\text{QTES}}(\gamma) - \mu_k(\gamma)\right| \le \Delta_n, \qquad \Delta_n := 10\sqrt{\frac{\mathbb{E}_P\|Z_1\|_2^4}{s_\star}}. \tag{20}$$

Consequently, with the same probability,

$$\left|\widehat{b}^{\text{QTES}}(\gamma) - b(\gamma)\right| \le \Delta_n \qquad \textit{simultaneously for all } \gamma \in \mathcal{G}. \tag{21}$$

Proof. Fix a fold $k \in \{1, 2\}$ and a grid point $\gamma \in \mathcal{G}$. Let $\Pi_{k,\gamma}$ denote the independent random partition of I_k into $m := m_{\text{blocks}}$ blocks whose sizes differ by at most one, and define

$$\mathcal{H}_{k,\gamma} := \sigma(\mathcal{F}_k, \Pi_{k,\gamma}).$$

Write the blocks as $G_1, \ldots, G_m$, let $s_j := |G_j|$, and note that $s_j \ge s_\star = \lfloor n_1/m \rfloor$ for all j. Define the block means

$$U_j^{(k)}(\gamma) := \frac{1}{s_j} \sum_{i \in G_j} Y_i^{(k)}(\gamma), \qquad j = 1, \ldots, m.$$

Conditionally on $\mathcal{H}_{k,\gamma}$, the variables $U_1^{(k)}(\gamma), \ldots, U_m^{(k)}(\gamma)$ are independent and each has mean $\mu_k(\gamma)$. Moreover,

$$\mathrm{Var}_P\left(U_j^{(k)}(\gamma) \mid \mathcal{H}_{k,\gamma}\right) \le \frac{v_k(\gamma)}{s_j} \le \frac{v_k(\gamma)}{s_\star}.$$

Set

$$t_k(\gamma) := 10\sqrt{\frac{v_k(\gamma)}{s_\star}}.$$

By Chebyshev's inequality, for every j,

$$\mathbb{P}\Big(U_j^{(k)}(\gamma) > \mu_k(\gamma) + t_k(\gamma) \mid \mathcal{H}_{k,\gamma}\Big) \le 10^{-2},$$

and similarly

$$\mathbb{P}\Big(U_j^{(k)}(\gamma) < \mu_k(\gamma) - t_k(\gamma) \mid \mathcal{H}_{k,\gamma}\Big) \le 10^{-2}.$$

Let $U_{(1)}^{(k)}(\gamma) \le \cdots \le U_{(m)}^{(k)}(\gamma)$ denote the order statistics, and let

$$r := \lceil 0.85m \rceil, \qquad \widehat{b}_k^{\text{QTES}}(\gamma) = U_{(r)}^{(k)}(\gamma).$$

If $\widehat{b}_k^{\mathrm{QTES}}(\gamma) > \mu_k(\gamma) + t_k(\gamma)$, then at least $m - r + 1 \geq 0.15m$ block means exceed $\mu_k(\gamma) + t_k(\gamma)$. Hence, with

$$N_+ := \sum_{j=1}^{m} \mathbf{1}\Big\{U_j^{(k)}(\gamma) > \mu_k(\gamma) + t_k(\gamma)\Big\},$$

we have, since N_+ is a sum of independent Bernoulli variables with conditional success probabilities at most 0.01,

$$\mathbb{P}\Big(\widehat{b}_k^{\mathrm{QTES}}(\gamma) > \mu_k(\gamma) + t_k(\gamma) \mid \mathcal{H}_{k,\gamma}\Big) \leq \mathbb{P}\big(N_+ \geq 0.15m \mid \mathcal{H}_{k,\gamma}\big) \leq \exp\big(-mD(0.15\|0.01)\big),$$

where $D(a\|b) := a\log(a/b) + (1-a)\log((1-a)/(1-b))$ is the Bernoulli relative entropy. Likewise, if $\widehat{b}_k^{\mathrm{QTES}}(\gamma) < \mu_k(\gamma) - t_k(\gamma)$, then at least $r \geq 0.85m$ block means fall below $\mu_k(\gamma) - t_k(\gamma)$, so with

$$N_- := \sum_{j=1}^{m} \mathbf{1}\Big\{U_j^{(k)}(\gamma) < \mu_k(\gamma) - t_k(\gamma)\Big\},$$

we obtain by the same Poisson-binomial Chernoff bound

$$\mathbb{P}\Big(\widehat{b}_k^{\mathrm{QTES}}(\gamma) < \mu_k(\gamma) - t_k(\gamma) \mid \mathcal{H}_{k,\gamma}\Big) \leq \mathbb{P}\big(N_- \geq 0.85m \mid \mathcal{H}_{k,\gamma}\big) \leq \exp\big(-mD(0.85\|0.01)\big).$$

Numerically, $D(0.15\|0.01) \approx 0.2766$ and $D(0.85\|0.01) > 0.2766$. Therefore, for each pair (k,γ),

$$\mathbb{P}\Big(\big|\widehat{b}_k^{\mathrm{QTES}}(\gamma) - \mu_k(\gamma)\big| > t_k(\gamma)\Big) \leq 2\exp(-0.2766\, m).$$

By definition, $m = m_{\mathrm{base}} \geq 4\log(4|\mathcal{G}|/\delta_{\mathrm{bias}})$. A union bound over the $2|\mathcal{G}|$ choices of (k,γ) yields total failure probability at most

$$4|\mathcal{G}|\exp(-0.2766\, m) \leq 4|\mathcal{G}|\left(\frac{\delta_{\mathrm{bias}}}{4|\mathcal{G}|}\right)^{4\cdot 0.2766} \leq \delta_{\mathrm{bias}},$$

because $4 \cdot 0.2766 > 1$ and $\delta_{\mathrm{bias}}/(4|\mathcal{G}|) \leq 1$. This proves (19).

For (20), note that $0 \leq Y_i^{(k)}(\gamma) \leq \|Z_i\|_2^2$, so

$$v_k(\gamma) \leq \mathbb{E}_P\big[(Y_i^{(k)}(\gamma))^2 \mid \mathcal{F}_k\big] \leq \mathbb{E}_P\|Z_1\|_2^4.$$

Finally,

$$\big|\widehat{b}^{\mathrm{QTES}}(\gamma) - b(\gamma)\big| = \left|\sum_{k=1}^{2} \frac{n_k}{n_{\mathrm{ret}}}\big(\widehat{b}_k^{\mathrm{QTES}}(\gamma) - \mu_k(\gamma)\big)\right| \leq \sum_{k=1}^{2} \frac{n_k}{n_{\mathrm{ret}}}\big|\widehat{b}_k^{\mathrm{QTES}}(\gamma) - \mu_k(\gamma)\big| \leq \Delta_n,$$

which proves (21). □

B.3 Consequences for the QTES Score

Applying Theorem 5 with the calibration supplied by Theorem 6 shows that the QTES score is, up to an additive constant independent of γ, a strict simultaneous upper confidence family over $\mathcal{G}$. We record this consequence separately because it is the form used in the proof of Theorem 3.

Corollary 7 (Simultaneous UCB form)**.** *With probability at least $1-\delta$, the following statements hold simultaneously for all $\gamma \in \mathcal{G}$:*

$$\Big\|\hat{\Sigma}^{\mathrm{raw}}(\gamma) - \Sigma\Big\|_{\mathrm{op}} \leq \widehat{S}_{\mathrm{QTES}}(\gamma) + \Delta_n. \tag{22}$$

Hence the family

$$\widetilde{U}_{\mathrm{QTES}}(\gamma) := \widehat{S}_{\mathrm{QTES}}(\gamma) + \Delta_n$$

is a simultaneous upper confidence family over $\mathcal{G}$. Because Δ_n is independent of γ, the argmin sets coincide:

$$\arg\min_{\gamma\in\mathcal{G}} \widehat{S}_{\mathrm{QTES}}(\gamma) = \arg\min_{\gamma\in\mathcal{G}} \widetilde{U}_{\mathrm{QTES}}(\gamma). \tag{23}$$

In particular, $\hat{\gamma}_{\mathrm{QTES}}$ exactly minimizes a strict simultaneous upper confidence family.

Proof. Fix $\gamma \in \mathcal{G}$ and work on the intersection of the events from Theorem 1 and Theorem 6. By the triangle inequality,

$$\left\|\hat{\Sigma}^{\mathrm{raw}}(\gamma) - \Sigma\right\|_{\mathrm{op}} \leq \left\|\hat{\Sigma}^{\mathrm{raw}}(\gamma) - \bar{\Sigma}_P^{\mathrm{raw}}(\gamma)\right\|_{\mathrm{op}} + \left\|\bar{\Sigma}_P^{\mathrm{raw}}(\gamma) - \Sigma\right\|_{\mathrm{op}}.$$

The first term is bounded by $\overline{\Psi}^{\mathrm{raw}}(\gamma)$ from Theorem 1. The second term is bounded by $b(\gamma)$ from Lemma 2. By (21),

$$b(\gamma) \leq \widehat{b}^{\mathrm{QTES}}(\gamma) + \Delta_n.$$

Combining these inequalities gives (22). Since adding the same constant Δ_n to every score does not change the minimizers, (23) follows immediately. □

Because the additive slack in Corollary 7 does not depend on γ, minimizing the QTES score is equivalent to minimizing the associated upper confidence family. The next corollary turns this observation into the selected-estimator guarantee stated in Theorem 3.

Corollary 8 (Near-oracle inequality for the selected QTES estimator)**.** *Under the same event as Corollary 7,*

$$\left\|\hat{\Sigma}^{\mathrm{raw}}(\hat{\gamma}_{\mathrm{QTES}}) - \Sigma\right\|_{\mathrm{op}} \leq \min_{\gamma \in \mathcal{G}} \left\{\overline{\Psi}^{\mathrm{raw}}(\gamma) + b(\gamma)\right\} + 2\Delta_n. \tag{24}$$

Proof. Define the ideal observable-plus-bias curve

$$O(\gamma) := \overline{\Psi}^{\mathrm{raw}}(\gamma) + b(\gamma), \qquad \gamma \in \mathcal{G}.$$

On the event of Corollary 7, (21) implies

$$\left|\widehat{S}_{\mathrm{QTES}}(\gamma) - O(\gamma)\right| = \left|\widehat{b}^{\mathrm{QTES}}(\gamma) - b(\gamma)\right| \leq \Delta_n \qquad \text{for all } \gamma \in \mathcal{G}.$$

Let $\gamma_\star \in \arg\min_{\gamma \in \mathcal{G}} O(\gamma)$. Because $\hat{\gamma}_{\mathrm{QTES}}$ minimizes $\widehat{S}_{\mathrm{QTES}}$, we have

$$\widehat{S}_{\mathrm{QTES}}(\hat{\gamma}_{\mathrm{QTES}}) \leq \widehat{S}_{\mathrm{QTES}}(\gamma_\star) \leq O(\gamma_\star) + \Delta_n.$$

Applying (22) at $\hat{\gamma}_{\mathrm{QTES}}$ yields

$$\left\|\hat{\Sigma}^{\mathrm{raw}}(\hat{\gamma}_{\mathrm{QTES}}) - \Sigma\right\|_{\mathrm{op}} \leq \widehat{S}_{\mathrm{QTES}}(\hat{\gamma}_{\mathrm{QTES}}) + \Delta_n \leq O(\gamma_\star) + 2\Delta_n,$$

which is exactly (24). □

C Proofs of the Remaining Main-Text Results

C.1 Proof of Theorem 3

Proof. Work on the intersection of the events from Theorem 1 and Theorem 6. On this event, the aggregate calibration bound (21) holds for all $\gamma \in \mathcal{G}$, and Corollary 7 yields the simultaneous upper bound (22) for all $\gamma \in \mathcal{G}$. These are exactly the two claims in (12). The selected-estimator inequality (13) is Corollary 8. Since $\delta_{\mathrm{var}} = \delta_{\mathrm{bias}} = \delta/2$, the intersection event has probability at least $1 - \delta$. □

C.2 Proof of Corollary 4

Proof. Define

$$O(\gamma) := \overline{\Psi}^{\mathrm{raw}}(\gamma) + b(\gamma), \qquad \gamma \in \mathcal{G},$$

and let $\gamma_\star \in \arg\min_{\gamma \in \mathcal{G}} O(\gamma)$. On the event of Corollary 7, the calibration bound (21) gives

$$\left|\widehat{S}_{\mathrm{QTES}}(\gamma) - O(\gamma)\right| \leq \Delta_n \qquad \text{for all } \gamma \in \mathcal{G}.$$

Because $\hat{\gamma}_{\mathrm{QTES}}$ minimizes $\widehat{S}_{\mathrm{QTES}}$,

$$O(\hat{\gamma}_{\mathrm{QTES}}) \leq \widehat{S}_{\mathrm{QTES}}(\hat{\gamma}_{\mathrm{QTES}}) + \Delta_n \leq \widehat{S}_{\mathrm{QTES}}(\gamma_\star) + \Delta_n \leq O(\gamma_\star) + 2\Delta_n,$$

which proves (14). The Γ_ε claim is immediate when $\varepsilon \geq 2\Delta_n$. If the minimizer $\gamma_\star$ is unique and every other grid point lies more than $2\Delta_n$ above $O(\gamma_\star)$, then (14) forces $\hat{\gamma}_{\mathrm{QTES}} = \gamma_\star$. $\square$