Affine Invariant Covariance Estimation for Heavy-Tailed Distributions

Dmitrii M. Ostrovskii DMITRII.OSTROVSKII@INRIA.FR
Alessandro Rudi ALESSANDRO.RUDI@INRIA.FR
INRIA - Département d'Informatique de l'Ecole Normale Supérieure
PSL Research University, Paris, France

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

In this work we provide an estimator for the covariance matrix of a heavy-tailed multivariate distribution. We prove that the proposed estimator $\widehat{\mathbf{S}}$ admits an *affine-invariant* bound of the form

$$(1-\varepsilon)\mathbf{S} \preccurlyeq \widehat{\mathbf{S}} \preccurlyeq (1+\varepsilon)\mathbf{S}$$

in high probability, where $\mathbf{S}$ is the unknown covariance matrix, and $\preccurlyeq$ is the positive semidefinite order on symmetric matrices. The result only requires the existence of fourth-order moments, and allows for $\varepsilon = O(\sqrt{\kappa^4 d \log(d/\delta)/n})$ where κ^4 is a measure of kurtosis of the distribution, d is the dimensionality of the space, n is the sample size, and $1-\delta$ is the desired confidence level. More generally, we can allow for regularization with level λ, then d gets replaced with the degrees of freedom number. Denoting $\mathrm{cond}(\mathbf{S})$ the condition number of $\mathbf{S}$, the computational cost of the novel estimator is $O(d^2 n + d^3 \log(\mathrm{cond}(\mathbf{S})))$, which is comparable to the cost of the sample covariance estimator in the statistically interesing regime $n \geqslant d$. We consider applications of our estimator to eigenvalue estimation with relative error, and to ridge regression with heavy-tailed random design.

Keywords: Covariance estimation, heavy-tailed distributions, random design linear regression

1. Introduction

We are interested in estimating the covariance matrix $\mathbf{S} = \mathbb{E}[X \otimes X]$ of a zero-mean random vector $X \in \mathbb{R}^d$ from n independent and identically distributed (i.i.d.) copies $X_1, ..., X_n$ of X. This task is crucial – and often arises as a subroutine – in some widely used statistical procedures, such as linear regression, principal component analysis, factor analysis, generalized methods of moments, and mean-variance portfolio selection, to name a few (Friedman et al., 2001; Jolliffe, 2002; Hansen, 1982; Markowitz, 1952). In some of them, the control of $\|\widehat{\mathbf{S}} - \mathbf{S}\|_*$, where $\widehat{\mathbf{S}}$ is a covariance estimator and $\|\cdot\|_*$ is the spectral, Frobenius or trace norm, does not result in sharp theoretical guarantees. Instead, it might be necessary to estimate the eigenvalues of $\mathbf{S}$ in relative scale, ensuring that

$$|\lambda_j(\widehat{\mathbf{S}}) - \lambda_j(\mathbf{S})| \leqslant \varepsilon \lambda_j(\mathbf{S}), \quad j \in \{1, ..., d\},$$

holds for $\varepsilon > 0$ (this task arises in the analysis of the subspace iteration method, see Halko et al. (2011) and Sec. 6). More generally, one may seek to provide affine-invariant bounds of the form

$$(1-\varepsilon)\mathbf{S} \preccurlyeq \widehat{\mathbf{S}} \preccurlyeq (1+\varepsilon)\mathbf{S}, \tag{1}$$

as in the analysis of linear regression with random design (see Hsu et al., 2012, and Sec. 2.3 for more details), where $\preccurlyeq$ is the positive semidefinite partial order for symmetric matrices. In fact, the basic and very natural *sample covariance estimator*

$$\widetilde{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i$$

can be shown to satisfy Eq. (1) with probability at least $1-\delta$, $\delta \in (0,1]$, and accuracy ε scaling as $O((\operatorname{rank}(\mathbf{S})\log(d/\delta)/n)^{1/2})$, provided that X is subgaussian (see Sec. 2.3 for a detailed discussion). However, the assumption of sub-gaussianity might be too strong in the above applications. Going beyond it and similar assumptions is particularly important in mathematical finance, where it is widely accepted that the prices of assets might have heavy-tailed distributions (Kelly and Jiang, 2014; Bradley and Taqqu, 2003).

We propose a simple variation of the sample covariance estimator (see Algorithm 1) in the form

$$\widehat{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n} \alpha_i X_i \otimes X_i,$$

where the coefficients $\alpha_1, \dots, \alpha_n > 0$ are chosen in a data-driven manner. Our main result, stated informally below, shows that the proposed estimator enjoys *high-probability* bounds analogous to those for the sample covariance estimator, under a weak moment assumption on the distribution. Namely, we assume that for some $\kappa \geqslant 1$ it holds

$$\mathbb{E}^{1/4}[\langle X, u\rangle^4] \leqslant \kappa \mathbb{E}^{1/2}[\langle X, u\rangle^2], \quad \forall u \in \mathbb{R}^d. \tag{HT}$$

In other words, the *kurtosis* of X is bounded by κ in all directions.[1] Kurtosis is an affine-invariant and unitless quantity, and it is uniformly bounded from above by a constant for many common families of multivariate distributions: for example, $\kappa = \sqrt[4]{3}$ for any Gaussian distribution, and $\kappa \leqslant \sqrt[4]{9}$ for the multivariate Student-t distribution with at least 5 degrees of freedom.

Now we are ready to informally state our main result.

Theorem 1 (Simplified version of Thm. 7) *Under (HT), there exists an estimator $\widehat{\mathbf{S}}$ that has computational cost $O(d^2 n + d^3)$, and with probability at least $1-\delta$ satisfies (1) with accuracy*

$$\varepsilon \leqslant 48\kappa^2 \sqrt{\frac{\operatorname{rank}(\mathbf{S})\log(4d/\delta)}{n}}. \tag{2}$$

This result shows that the proposed estimator is a valid alternative to the sample covariance estimator: it has comparable accuracy and the same computational complexity, while requiring only boundedness of the fourth moment of X instead of sub-gaussianity. More generally, by allowing a *regularization level* $\lambda > 0$, i.e., using $\widehat{\mathbf{S}}_\lambda := \widehat{\mathbf{S}} + \lambda\mathbf{I}$ instead of $\widehat{\mathbf{S}}$ to estimate $\mathbf{S}_\lambda := \mathbf{S} + \lambda\mathbf{I}$ instead of $\mathbf{S}$ (as required in ridge regression (Hsu et al., 2012), we can replace $\operatorname{rank}(\mathbf{S})$ with the *degrees of freedom* number

$$\mathsf{df}_\lambda(\mathbf{S}) := \operatorname{Tr}(\mathbf{S}\mathbf{S}_\lambda^{-1}). \tag{3}$$

This leads to a better bound, since $\mathsf{df}_\lambda(\mathbf{S})$ is never larger than $\min\{\operatorname{rank}(\mathbf{S}), \operatorname{Tr}(\mathbf{S})/\lambda\}$, and can be way smaller depending on the eigenvalue decay of $\mathbf{S}$: for example, if $\lambda_j(\mathbf{S}) \leqslant j^{-b}$ with $b \geqslant 1$, then $\mathsf{df}_\lambda(\mathbf{S}) \leqslant \lambda^{-1/b}$.

[1] Note that we use a slightly non-standard definition of kurtosis, extracting the corresponding roots from the moments.

Paper Organization In Sec. 2 we recall the known results for the sample covariance matrix estimator under light-tailed assumptions, together with some recent high-probability results for an alternative estimator applicable to heavy-tailed distributions. The novel estimator is presented in Sec. 3 and analyzed and discussed in detail in Sec. 4. In order to achieve the best statistical performance, it requires the knowledge of the distribution parameters κ and $\mathsf{df}_\lambda(\mathbf{S})$ in advance; in Sec. 5 we extend the algorithm, via a variant of Lepskii's method (Lepskii, 1991), to be adaptive to these quantities. Applications to eigenvalue estimation and ridge regression are discussed in Sec. 6.

Notation and Conventions For $X \in \mathbb{R}^d$, $X \otimes X$ denotes the outer product $XX^\top$. W.l.o.g. we assume that $\mathbf{S} = \mathbb{E}[X \otimes X]$ is full-rank (otherwise we can work on its range). To reduce the clutter of parentheses, we convene that powers and multiplication have priority over the expectation, and we denote the $1/p$-th power of expectation, $p \geqslant 1$, with $\mathbb{E}^{1/p}[\cdot]$. We use $\|\cdot\|$ for the spectral norm of a matrix (unless specified otherwise), as well as for the ℓ_2-norm of a vector. We shortand $\min(a, b)$ to $a \wedge b$. We use the $O(\cdot)$ notation in a conventional way, and occasionally replace generic constants with $O(1)$. We use the notation $\mathbf{A}_\lambda := \mathbf{A} + \lambda\mathbf{I}$, where $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{I}$ is the identity matrix.

2. Background and Related Work

In this section we recall some relevant previous work on covariance estimators for light-tailed and heavy-tailed distributions, and provide more intuition about affine-invariant error bounds. Moreover, we introduce basic concepts that will be used later on in the theoretical analysis.

2.1. Relative Error Bounds for the Sample Covariance Estimator

When the estimation error is measured by $\|\widetilde{\mathbf{S}} - \mathbf{S}\|$, where $\|\cdot\|$ is the *spectral norm*, the problem can be reduced, via the Chernoff bounding technique, to the control of the matrix moment generating function, for which one can apply some deep operator-theoretic results such as the Goldon-Thompson inequality (Ahlswede and Winter, 2002; Oliveira, 2010) or Lieb's theorem (Tropp, 2012). Alternatively, one may reduce the task to the control of an underlying empirical process, and exploit advanced tools from empirical process theory such as generic chaining (Koltchinskii and Lounici, 2017). The both families of approaches have focused on the sample covariance estimator and its direct extensions (Vershynin, 2012; Tropp, 2012, 2015), requiring stronger assumptions on the distribution of X than (HT). In particular, consider the *subgaussian moment growth* assumption

$$\mathbb{E}^{1/p}[|\langle X, u\rangle|^p] \leqslant \bar{\kappa}\sqrt{p}\,\mathbb{E}^{1/p}[\langle X, u\rangle^2], \quad \forall u \in \mathbb{R}^d \text{ and } p \geqslant 2, \tag{SG}$$

which implies (HT) with $\kappa = 2\bar{\kappa}$, where κ is defined in Eq. (HT). Define the *effective rank* of $\mathbf{S}$ by

$$\mathtt{r}(\mathbf{S}) := \mathrm{Tr}(\mathbf{S})/\|\mathbf{S}\|. \tag{4}$$

The following result is known.

Theorem 2 (Simplified version of (Lounici, 2014, Prop. 3)) *Under* (SG)*, the sample covariance estimator* $\widetilde{\mathbf{S}}$ *with probability at least* $1 - \delta$*,* $\delta \in (0, 1]$*, satisfies*

$$\|\widetilde{\mathbf{S}} - \mathbf{S}\| \leqslant O(1)\bar{\kappa}^2\|\mathbf{S}\|\sqrt{\frac{\mathtt{r}(\mathbf{S})\log(2d/\delta)}{n}}, \tag{5}$$

provided that $n \geqslant \widetilde{O}(1)\mathtt{r}(\mathbf{S})$*, where* $\widetilde{O}(1)$ *hides polynomial dependency on* $\log(2d/\delta)$ *and* $\log(2n/\delta)$*.*

It can be shown (Ledoux and Talagrand, 2013, Prop. 6.10) that Eq. (5) nearly optimally depends on $\mathtt{r}(\mathbf{S})$, $\bar{\kappa}$, and $1/\delta$.[2] Another remarkable property of this bound is that it is almost dimension-independent: up to a logarithmic factor, the complexity of estimating $\mathbf{S}$ is independent of the ambient dimension d. Instead, it is controlled by the distribution-dependent quantity $\mathtt{r}(\mathbf{S})$, which always satisfies $\mathtt{r}(\mathbf{S}) \leqslant \mathrm{rank}(\mathbf{S}) \leqslant d$, and can be much smaller than $\mathrm{rank}(\mathbf{S})$ when the distribution of X lies close to a low-dimensional linear subspace, i.e., when $\mathbf{S}$ has only a few relatively large eigenvalues.

2.2. Relative Error Bounds for Heavy-Tailed Distributions

It is possible to obtain relative error bounds of the form $\|\widehat{\mathbf{S}} - \mathbf{S}\|$, including the ones in high probability, under *weak* moment assumptions such as (HT), considering other estimators than the sample covariance matrix. In particular, Wei and Minsker (2017) propose an estimator based on the idea of clipping observations with large norm. Formally, they define the truncation map $\psi_\theta : \mathbb{R} \to \mathbb{R}$,

$$\psi_\theta(x) := (|x| \wedge \theta)\,\mathrm{sign}(x), \tag{6}$$

given a certain threshold $\theta > 0$, and consider the estimator

$$\widehat{\mathbf{S}}^{\mathrm{WM}} := \frac{1}{n}\sum_{i=1}^{n} \rho_\theta(\|X_i\|)\, X_i \otimes X_i, \quad \text{where} \;\; \rho_\theta(x) := \psi_\theta(x^2)/x^2. \tag{7}$$

In other words, one simply truncates observations with squared norm larger than θ prior to averaging. This estimator is a key ingredient in our Algorithm 1, and we now summarize its statistical properties.

Theorem 3 (Minsker and Wei (2017, Lem. 2.1 and Lem. 5.7)) *Define the matrix second moment statistic $W := \|\mathbb{E}[\|X\|^2 X \otimes X]\|$. Let $\overline{W} \geqslant W$, and $\delta \in (0,1]$. Then estimator $\widehat{\mathbf{S}}^{\mathrm{WM}}$, cf. (7), with $\theta = \sqrt{n\overline{W}/\log(2d/\delta)}$ with probability at least $1-\delta$ satisfies $\|\widehat{\mathbf{S}}^{\mathrm{WM}} - \mathbf{S}\| \leqslant 2\sqrt{\overline{W}\log(2d/\delta)/n}$.*

In contrast with Thm. 2, Thm. 3 claims *subgaussian* concentration for the spectral-norm loss under the weak moment assumption (HT). Moreover, we arrive at the relative error bound akin to (5):

$$\left\|\widehat{\mathbf{S}}^{\mathrm{WM}} - \mathbf{S}\right\| \leqslant 2\kappa^2\|\mathbf{S}\|\sqrt{\frac{\mathtt{r}(\mathbf{S})\log(2d/\delta)}{n}}, \tag{8}$$

if we bound the second moment statistic as

$$W \leqslant \left\|\mathbb{E}\|X\|^2 X \otimes X\right\| \leqslant \kappa^4\|\mathbf{S}\|^2\mathtt{r}(\mathbf{S}), \tag{9}$$

see (Wei and Minsker, 2017, Lem. 2.3 and Cor. 5.1), and choose the appropriate truncation level

$$\theta = \kappa^2\|\mathbf{S}\|\sqrt{n\mathtt{r}(\mathbf{S})/\log(2d/\delta)} \geqslant \sqrt{nW/\log(2d/\delta)}.$$

Since this choice depends on the unknown W, one can use a larger value, which will result in the inflation of the right-hand side of Eq. (8). An alternative is to adapt to the unknown W via Lepskii's method (Lepskii, 1991) as described in (Wei and Minsker, 2017, Thm 2.1). To conclude, the estimator $\widehat{\mathbf{S}}^{\mathrm{WM}}$ enjoys subgaussian relative error bounds under the fourth moment assumption (HT), while having essentially the same computation cost as the sample covariance estimator.

[2] In fact, (Koltchinskii and Lounici, 2017, Theorem 9) replaces $\mathtt{r}(\mathbf{S})\log(2d/\delta)$ by $\mathtt{r}(\mathbf{S}) + \log(1/\delta)$, making the bound dimension-independent, but does not specify the dependency on $\bar{\kappa}$. Similar results have been obtained under (HT) for robust estimators (see Sec. 2.2), which, however, are computationally intractable (Mendelson and Zhivotovskiy, 2018).

2.3. Affine-Invariant Bounds for the Sample Covariance Estimator

As we have seen previously, estimator $\widehat{\mathbf{S}}^{\mathrm{WM}}$ has favorable statistical properties compared to $\widetilde{\mathbf{S}}$ when the goal is to estimate $\mathbf{S}$ in relative spectral-norm error as in Eq. (8). However, one can instead be interested in providing affine-invariant bounds in the form of Eq. (1). More generally, one may wish to estimate $\mathbf{S}$ only for the eigenvalues greater than some level $\lambda > 0$, that is, to guarantee that

$$(1-\varepsilon)\mathbf{S}_\lambda \preccurlyeq \widehat{\mathbf{S}}_\lambda \preccurlyeq (1+\varepsilon)\mathbf{S}_\lambda. \tag{10}$$

The need for such bounds arises, in particular, in random-design ridge regression, where the information about inferior eigenvalues is irrelevant, since it is anyway erased by regularization. Note that Eq. (10), for any $\lambda > 0$, can be reformulated in terms of the $\mathbf{S}_\lambda^{-1/2}$-transformed spectral norm:

$$\left\| \mathbf{S}_\lambda^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})\mathbf{S}_\lambda^{-1/2} \right\| \leqslant \varepsilon. \tag{11}$$

The task of obtaining such bounds, with arbitrary regularization level $0 \leqslant \lambda \leqslant \|\mathbf{S}\|$, will be referred to as **calibrated covariance estimation**. Generally, this task is harder than proving relative-error bounds in the spectral norm such as Eq. (8): the latter is equivalent, up to a constant factor loss of accuracy, to proving Eq. (11) with $\lambda = \|\mathbf{S}\|$. On the other hand, calibrated covariance estimation also subsumes Eq. (1) by taking $\lambda = O(\lambda_{\min}(\mathbf{S}))$, where $\lambda_{\min}(\mathbf{S})$ is the smallest eigenvalue of $\mathbf{S}$.

Now, one can make a simple observation that for the sample covariance estimator $\widetilde{\mathbf{S}}$, calibrated bounds of the form (11) "automatically" follow from the dimension-free spectral-norm bounds akin to (5) or (8) due to its affine equivariance. Indeed, $\mathbf{J} := \mathbf{S}_\lambda^{-1/2}\mathbf{S}\mathbf{S}_\lambda^{-1/2}$ is precisely the covariance matrix of the "λ-decorrelated" observations $Z_i = \mathbf{S}_\lambda^{-1/2} X_i$, for which the sample covariance estimator is given by $\widetilde{\mathbf{J}} := \frac{1}{n}\sum_{i=1}^n Z_i \otimes Z_i = \mathbf{S}_\lambda^{-1/2}\widetilde{\mathbf{S}}\mathbf{S}_\lambda^{-1/2}$. Hence, we can apply the spectral-norm bound (5), replacing $\mathbf{S}$ and $\widetilde{\mathbf{S}}$ with $\mathbf{J}$ and $\widetilde{\mathbf{J}}$. Using the fact that $\|\mathbf{J}\| \leqslant 1$ for any $\lambda \geqslant 0$, and that assumptions (HT), (SG) are themselves invariant under (non-singular) linear transforms, we obtain

$$\mathbb{E}^{1/2}\left[\left\| \mathbf{S}_\lambda^{-1/2}(\widetilde{\mathbf{S}} - \mathbf{S})\mathbf{S}_\lambda^{-1/2} \right\|^2 \right] \leqslant O(1)\bar{\kappa}^2 \sqrt{\frac{\mathsf{df}_\lambda(\mathbf{S})\log(2d)}{n}} \tag{12}$$

once $n \geqslant \widetilde{O}(1)\mathsf{df}_\lambda(\mathbf{S})$, where $\mathsf{df}_\lambda(\cdot)$ is defined in Eq. (3) and ranges from $O(\mathtt{r}(\mathbf{S}))$ to $\mathrm{rank}(\mathbf{S}) \leqslant d$ as λ decreases from $\|\mathbf{S}\|$ to zero. In fact, when $\bar{\kappa}$ is a constant, this rate is known to be asymtptocially minimax-optimal over certain natural classes of covariance matrices, e.g., Toeplitz matrices with spectra discretizing those of Hölder-smooth functions (Bickel and Levina, 2008; Cai et al., 2010). It is thus reasonable to ask whether one can extend Eq. (12) in the same manner as Eq. (8) extends Eq. (5). In other words, *can one provide a high-probability guarantee for the calibrated error (cf. Eq.* (11)*) of the estimator* $\widehat{\mathbf{S}}^{\mathrm{WM}}$ *(cf. Eq.* (7)*) under assumption* (HT)*?* The immediate difficulty is that $\widehat{\mathbf{S}}^{\mathrm{WM}}$ – in fact, the only estimator for which finite-sample high-probability guarantees under fourth moment assumptions are known to us – does not allow for the same reasoning as $\widetilde{\mathbf{S}}$ because of the non-linearity introduced by the truncation map. On the other hand, the desired bounds are achieved by the "oracle" estimator that truncates the "λ-decorrelated" vectors $Z_i = \mathbf{S}_\lambda^{-1/2} X_i$ with accordingly adjusted θ:

$$\widehat{\mathbf{S}}^o := \frac{1}{n}\sum_{i=1}^n \rho_\theta(\|\mathbf{S}_\lambda^{-1/2} X_i\|) X_i \otimes X_i = \mathbf{S}_\lambda^{1/2}\left[\frac{1}{n}\sum_{i=1}^n \rho_\theta(\|Z_i\|) Z_i \otimes Z_i\right]\mathbf{S}_\lambda^{1/2}. \tag{13}$$

cf. Eq. (7). Unfortunately, this estimator is unavailable since Z_i's are not observable. In what follows, we present our main methodological contribution: an estimator that achieves the stated goal, and moreover, has a similar complexity of computation and storage as the sample covariance matrix.

Remark 4 *Some robust covariance estimators, such as MCD or MVE (Campbell, 1980; Lopuhaa and Rousseeuw, 1991; Rousseeuw and Driessen, 1999), are affine equivariant, but to the best of our knowledge, the desired bounds are not known for them. On the other hand, Oliveira (2016) shows that if one only seeks for the left-hand side bound in (10), the sample covariance estimator suffices under (HT).*

3. Proposed Estimator

Our goal can be summarized as follows: given $\delta \in (0,1]$, and $\lambda \geqslant 0$, provide an estimate $\widehat{\mathbf{S}}$ satisfying

$$\left\| \mathbf{S}_\lambda^{-1/2} (\widehat{\mathbf{S}} - \mathbf{S}) \mathbf{S}_\lambda^{-1/2} \right\| \leqslant O(1)\kappa^2 \sqrt{\frac{\mathsf{df}_\lambda(\mathbf{S}) \log(2d/\delta)}{n}},$$

where κ is the kurtosis parameter of X (cf. (HT)). Moreover, we can restrict ourselves to the case $\lambda \leqslant \|\mathbf{S}\|$, since otherwise the task is resolved by the estimator $\widehat{\mathbf{S}}^{\mathrm{WM}}$ as can be seen from Eq. (8).

As we have seen before, the oracle estimator $\widehat{\mathbf{S}}^o$ introduced in the previous section (cf. Eq. (13)) achieves the stated goal, but is unavailable since it depends explicitly on $\mathbf{S}$. The key idea of our construction is to approximate $\widehat{\mathbf{S}}^o$ in an iterative fashion – roughly, to start with $\widehat{\mathbf{S}}^{(0)} = \widehat{\mathbf{S}}^{\mathrm{WM}}$, which is already a good estimate for $\mathbf{S}_\lambda$ with the crudest regularization level $\lambda = \|\mathbf{S}\|$ due to Eq. (8), and then iteratively refine the estimate by computing

$$\widehat{\mathbf{S}}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \rho_\theta(\|[\widehat{\mathbf{S}}_\lambda^{(t)}]^{-1/2} X_i\|)\, X_i \otimes X_i. \tag{14}$$

To make this simple idea work, we need to adjust it in two ways. Firstly, $\widehat{\mathbf{S}}^{(t)}$ depends on the observations X_i, hence the random vectors $[\widehat{\mathbf{S}}_\lambda^{(t)}]^{-1/2} X_i$ are not independent. To simplify the analysis, we split the sample into batches corresponding to different iterations, and at each iteration use observations of the new batch instead of X_i's in Eq. (14). Since $\widehat{\mathbf{S}}^{(t)}$ is independent from the new observations, we can apply Thm. 3 conditionally at each step.

Secondly, as discussed before, the estimator $\widehat{\mathbf{S}}^{\mathrm{WM}}$ given by (7) already solves the problem for $\lambda = \|\mathbf{S}\|$. To achieve Eq. (11) for a given λ, the idea is to start with $\lambda_0 = \|\mathbf{S}\|$, and reduce λ_t by a constant factor at each iteration, so that the error $\|\mathbf{S}_\lambda^{-1/2}(\widehat{\mathbf{S}}^{(t)} - \mathbf{S})\mathbf{S}_\lambda^{-1/2}\|$ remains controlled for $\lambda = \lambda_t$ at each step. This way we also ensure that the total number of iterations is logarithmic in $\|\mathbf{S}\|/\lambda$, and, in particular, logarithmic in the condition number $\|\mathbf{S}\|/\lambda_{\min}(\mathbf{S})$ when $\lambda \geqslant \lambda_{\min}(\mathbf{S})$.

Algorithm 1 presented below implements these ideas. Note that the final batch of observations takes a half of the overall sample: this is needed to achieve the best possible accuracy (up to a constant factor) for the final regularization level, while at the previous levels it suffices to maintain the accuracy $\varepsilon = 1/2$, and one can use smaller batches taking up a half of the sample in total, see Lem. 6 in Sec. 4 for details. Once the coefficients $\alpha_i^{(t)}$ at the given step have been computed, the new estimate $\widehat{\mathbf{S}}^{(t+1)}$ reduces to the sample covariance matrix of the weighted observations, which can be computed in time $O(d^2 m)$ where m is the size of the batch. The total cost of these computations in the course of the algorithm is thus $O(d^2 n)$. As for $\alpha_i^{(t)}$, they are obtained by first performing the Cholesky decomposition (Golub and Van Loan, 2012) of $\widehat{\mathbf{S}}_{\lambda_t}^{(t)}$, i.e., finding the unique lower-triangular matrix $\mathbf{R}_t$ such that $\mathbf{R}_t \mathbf{R}_t^\top = \widehat{\mathbf{S}}_{\lambda_t}^{(t)}$ which requires $O(d^3)$ in time and $O(d^2)$ in space,[3]and then computing each product $\mathbf{R}_t^{-1} X_i^{(t+1)}$ by solving the corresponding linear system in $O(d^2)$. The total complexity of Algorithm 1 is thus

$$O\left(d^2 n + d^3 \log(L/\lambda)\right) \text{ in time,} \quad O(d^2) \text{ in space.}$$

Moreover, the time complexity becomes $O(d^2 n)$ when $n \gg d \log(L/\lambda)$; as we show next, this is anyway required to obtain a statistical performance guarantee. Note moreover that it is possible to obtain the non-regularized version of Eq. (1) by choosing $\lambda = O(\lambda_{\min}(\mathbf{S}))$, in time $O(d^2 n + d^3 \log(\mathrm{cond}(\mathbf{S})))$, where $\mathrm{cond}(\mathbf{S}) = \|\mathbf{S}\|/\lambda_{\min}$ is the condition number of $\mathbf{S}$.

Remark 5 *In practice, sample splitting in Algorithm 1 could be avoided, and iterations could be performed on the same sample. We expect our statistical guarantees in Sec. 4 to extend to this setup.*

Next we present a statistical guarantee for Algorithm 1, and suggest a way to select the parameters.

Algorithm 1: Robust Calibrated Covariance Estimation

```
Input: $X_1, ..., X_n \in \mathbb{R}^d$, $\delta \in (0,1]$, regularization level $\lambda \leqslant \|\mathbf{S}\|$, $L \geqslant \|\mathbf{S}\|$, truncation level $\theta > 0$
$\lambda_0 = L,\ T = \lceil \log_2(L/\lambda) \rceil,\ m = \lfloor n/(2(T+1)) \rfloor$
$\alpha_i^{(0)} = L\rho_\theta(\|X_i\|/\sqrt{L})$, for $i \in \{1, \dots, m\}$, with $\rho_\theta(x) = \psi_\theta(x^2)/x^2$ and $\psi_\theta(\cdot)$ as in Eq. (6)
$\widehat{\mathbf{S}}^{(0)} = \frac{1}{m}\sum_{i=1}^m \alpha_i^{(0)} X_i \otimes X_i$
for $t = 0$ to $T-1$ do
    $(X_1^{(t+1)} : X_m^{(t+1)}) = (X_{m(t+1)+1} : X_{m(t+1)+m})$ {Obtain a new batch}
    $\mathbf{R}_t = \text{Cholesky}(\widehat{\mathbf{S}}^{(t)}_{\lambda_t})$
    $\alpha_i^{(t+1)} = \rho_\theta(\|\mathbf{R}_t^{-1} X_i^{(t+1)}\|)$, for $i \in \{1, \dots, m\}$
    $\widehat{\mathbf{S}}^{(t+1)} = \frac{1}{m}\sum_{i=1}^m \alpha_i^{(t+1)} X_i^{(t+1)} \otimes X_i^{(t+1)}$
    $\lambda_{t+1} = \lambda_t/2$
end for
$r = n - m(T+1)$ {Size of the remaining sample, roughly $n/2$}
$\theta_T = 2\theta(T+1)^{1/2}\left(1 + \log(T+1)/\log(4d/\delta)\right)^{1/2}$ {Final truncation level}
$(X_1^\star : X_r^\star) = (X_{mT+1} : X_n)$ {Remaining sample}
$\mathbf{R}_T = \text{Cholesky}(\widehat{\mathbf{S}}^{(T)}_{\lambda_T})$
$\alpha_i^\star = \rho_{\theta_T}(\|\mathbf{R}_T^{-1} X_i^\star\|)$, for $i \in \{1, \dots, r\}$
$\widehat{\mathbf{S}}^\star = \frac{1}{r}\sum_{i=1}^r \alpha_i^\star X_i^\star \otimes X_i^\star$ {Final estimate}
Output: $\widehat{\mathbf{S}}^\star$
```

4. Statistical Guarantee

In Theorem 7 below, we show that the estimator produced by Algorithm 1 achieves a high-probability bound of the type (11) requiring only the existence of the fourth-order moments of X, and the correct choice of the truncation level θ. We begin with the lemma that justifies the proposed update rule.

Lemma 6 *Let* $\widehat{\mathbf{S}}$ *be a symmetric estimate of* $\mathbf{S}$ *such that, for some* $\lambda > 0$,

$$\|\mathbf{S}_\lambda^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})\mathbf{S}_\lambda^{-1/2}\| \leqslant 1/2. \tag{15}$$

Conditioned on $\widehat{\mathbf{S}}$, let $X_1, \dots, X_m$ be i.i.d., have zero mean, covariance $\mathbf{S}$, and finite fourth-order moments. Let κ be the associated (conditional) kurtosis as in Eq. (HT). Define $\mathbf{S}^{(+)}$ as

$$\widehat{\mathbf{S}}^{(+)} := \frac{1}{m}\sum_{j=1}^{m} \rho_\theta(\|\widehat{\mathbf{S}}_\lambda^{-1/2} X_j\|) X_j \otimes X_j, \tag{16}$$

with ρ_θ defined in Eqs. (6)–(7). Choose $\theta \geqslant 2\sqrt{2}\kappa^2\sqrt{m\mathsf{df}_\lambda(\mathbf{S})/\log(2d/\delta)}$, where $\mathsf{df}_\lambda(\mathbf{S})$ is defined by Eq. (3). Then with conditional probability at least $1-\delta$ over $(X_1, ..., X_m)$ it holds

$$\left\|\mathbf{S}_{\lambda/2}^{-1/2}(\widehat{\mathbf{S}}^{(+)} - \mathbf{S})\mathbf{S}_{\lambda/2}^{-1/2}\right\| \leqslant \frac{6\theta\log(2d/\delta)}{m}. \tag{17}$$

Lemma 6 is proved in Appendix B. Its role is to guarantee the stability of the iterative process in Algorithm 1 when we pass to the next regularization level by $\lambda^{(t)} \leftarrow \lambda^{(t-1)}/2$. Indeed, if the size m of the new batch is large enough, the right-hand side of Eq. (17) can be made smaller than $1/2$, which allows to apply Lemma 6 sequentially. We are now ready to present the guarantee for Algorithm 1.

Theorem 7 *Let $X_1, \dots, X_n$ be i.i.d. zero-mean random vectors in $\mathbb{R}^d$ satisfying $\mathbb{E}[X_i \otimes X_i] = \mathbf{S}$ and the kurtosis assumption (HT). Let Algorithm 1 be run with $\delta \in (0,1]$, $0 < \lambda \leqslant \|\mathbf{S}\| \leqslant L$, and*

$$\theta \geqslant \theta_* := 2\kappa^2\sqrt{\frac{n\mathsf{df}_\lambda(\mathbf{S})}{\mathsf{q}\log(4\mathsf{q}d/\delta)}}, \quad \textit{where} \quad \mathsf{q} := \lceil \log_2(L/\lambda)\rceil + 1. \tag{18}$$

Whenever the sample size satisfies

$$n \geqslant 48\mathsf{q}\theta\log(4\mathsf{q}d/\delta), \tag{19}$$

the resulting estimator $\widehat{\mathbf{S}}$ with probability at least $1-\delta$ satisfies

$$\left\|\mathbf{S}_\lambda^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})\mathbf{S}_\lambda^{-1/2}\right\| \leqslant \frac{24\theta\sqrt{\mathsf{q}\log(4\mathsf{q}d/\delta)\log(4d/\delta)}}{n}. \tag{20}$$

The above theorem shows that when the conditions on θ and n are met, the proposed estimator satisfies an affine-invariant error bound with accuracy of the same order as the one available for the sample covariance estimator (cf. Eq. (12)) under the more stringent sub-gaussian assumption. This is made explicit in the next corollary, where we simply put $\theta = \theta_*$ as suggested by Eq. (18), obtaining the bound (cf. Eq. (22)) that matches Eq. (12) up to a constant factor and the replacement of $\bar{\kappa}$ with κ.

Corollary 8 *Under the premise of Theorem 7, assume that*

$$n \geqslant 96^2\kappa^4\mathsf{q}\mathsf{df}_\lambda(\mathbf{S})\log(4\mathsf{q}d/\delta), \tag{21}$$

and choose $\theta = \theta_*$, cf. Eq. (18). Then the estimator given by Algorithm 1 w.p. at least $1-\delta$ satisfies

$$\left\|\mathbf{S}_\lambda^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})\mathbf{S}_\lambda^{-1/2}\right\| \leqslant 48\kappa^2\sqrt{\frac{\mathsf{df}_\lambda(\mathbf{S})\log(4d/\delta)}{n}}. \tag{22}$$

We conclude with a remark on choosing L, while in Sec. 5 we will provide an adaptive version of the estimator based on a version of Lepskii's method Lepskii (1991) in which θ is tuned automatically.

Remark 9 (Choosing L) *One can simply put $L = 2\|\widehat{\mathbf{S}}^{\mathrm{WM}}\|$, requiring that $\lambda \leqslant \frac{2}{3}\|\widehat{\mathbf{S}}^{\mathrm{WM}}\|$, with $\widehat{\mathbf{S}}^{\mathrm{WM}}$ defined in Eq. (7) and using an independent subsample of size $n/(T+1)$. Indeed, under Eq. (19), the result of Theorem 3 ensures that $\frac{2}{3}\|\widehat{\mathbf{S}}^{\mathrm{WM}}\| \leqslant \|\mathbf{S}\| \leqslant 2\|\widehat{\mathbf{S}}^{\mathrm{WM}}\|$ with probability $\geqslant 1-\delta$.*

4.1. Proof of Theorem 7

Note that $\mathsf{q} = T + 1$ is the number of batches processed by the end of the for-loop in Algorithm 1. Thus, using that $m = \lfloor n/(2\mathsf{q}) \rfloor$ and $n \geqslant 4\mathsf{q}$ (see Eq. (18)–(19) and use that $\kappa \geqslant 1$, $\mathsf{df}_\lambda \geqslant 1$), we get

$$n/(4\mathsf{q}) \leqslant m \leqslant n/(2\mathsf{q}). \tag{23}$$

We will proceed by induction over the steps $0 \leqslant t \leqslant T$, showing that

$$\left\| \mathbf{S}_{\lambda_t}^{-1/2} (\widehat{\mathbf{S}}^{(t)} - \mathbf{S}) \mathbf{S}_{\lambda_t}^{-1/2} \right\| \leqslant 1/2 \tag{24}$$

holds for $t = 0, ..., T$ with probability $\geqslant (1 - \delta/(2\mathsf{q}))^{t+1}$. Then we will derive Eq. (20) as a corollary.

$\mathbf{1^o}$. For the base, we can apply Thm. 3, exploiting that $\widehat{\mathbf{S}}^{(0)} = L\widehat{\mathbf{S}}^{\mathrm{WM}}$ for the (renormalized) initial batch $\frac{1}{\sqrt{L}}(X_1, ..., X_m)$. Thus, with probability at least $1 - \delta/(2\mathsf{q})$ over this batch, it holds

$$\frac{1}{L}\left\| \widehat{\mathbf{S}}^{(0)} - \mathbf{S} \right\| \leqslant \frac{2\theta \log(4\mathsf{q}d/\delta)}{m} \leqslant \frac{8\mathsf{q}\theta \log(4\mathsf{q}d/\delta)}{n}, \tag{25}$$

provided that (recall the condition $\theta \geqslant \sqrt{nW/\log(2d/\delta)}$ in Thm. 3 and combine it with Eq. (9)):

$$\theta \geqslant \kappa^2 \|\mathbf{S}\|/L \cdot \sqrt{m\mathtt{r}(\mathbf{S})/\log(4\mathsf{q}d/\delta)}.$$

But this follows from Eq. (18), since $\|\mathbf{S}\| \leqslant L$, $m \leqslant n/(2\mathsf{q})$, and $\mathtt{r}(\mathbf{S}) \leqslant 2\mathsf{df}_{\|\mathbf{S}\|}(\mathbf{S}) \leqslant 2\mathsf{df}_\lambda(\mathbf{S})$. Noting that $\|\mathbf{S}_L^{-1}\| \leqslant 1/L$, from Eqs. (19) and (25) we get

$$\left\| \mathbf{S}_L^{-1/2} (\widehat{\mathbf{S}}^{(0)} - \mathbf{S}) \mathbf{S}_L^{-1/2} \right\| \leqslant \frac{8\mathsf{q}\theta \log(4\mathsf{q}d/\delta)}{n} \leqslant \frac{1}{6}.$$

Since $\lambda_0 = L$, the induction base is proved. Note that when $T = 0$, this already results in Eq. (24).

$\mathbf{2^o}$. Let $T \geqslant 1$ and $0 \leqslant t \leqslant T - 1$. For the induction step, we apply Lemma 6 conditionally on the first t iterations, with $\widehat{\mathbf{S}}^{(t)}$ in the role of the current estimate, λ_t in the role of the current regularization level, and $(X_1^{(t+1)}, ..., X_m^{(t+1)})$ as the new batch (which is independent from $\widehat{\mathbf{S}}^{(t)}$ by construction). By the induction hypothesis, we have Eq. (24) with conditional probability $\geqslant (1 - \delta/(2\mathsf{q}))^{t+1}$. By Lemma 6, since $\lambda_t \geqslant \lambda$ (and thus $\mathsf{df}_{\lambda_t}(\mathbf{S}) \leqslant \mathsf{df}_\lambda(\mathbf{S})$), Eq. (18), when combined with the upper bound in Eq. (23), guarantees that with conditional probability $\geqslant 1 - \delta/(2\mathsf{q})$ over the new batch,

$$\left\| \mathbf{S}_{\lambda_{t+1}}^{-1/2} (\widehat{\mathbf{S}}^{(t+1)} - \mathbf{S}) \mathbf{S}_{\lambda_{t+1}}^{-1/2} \right\| \leqslant 6\theta \log(4\mathsf{q}d/\delta)/m \leqslant 24\mathsf{q}\theta \log(4\mathsf{q}d/\delta)/n \leqslant 1/2.$$

Here in the second transition we used the lower bound of Eq. (23), and in the last transition we used Eq. (19). Thus, the induction claim is proved. In particular, we have obtained that the bound

$$\left\| \mathbf{S}_{\lambda_T}^{-1/2} (\widehat{\mathbf{S}}^{(T)} - \mathbf{S}) \mathbf{S}_{\lambda_T}^{-1/2} \right\| \leqslant 1/2$$

holds with probability at least $(1 - \delta/(2\mathsf{q}))^{\mathsf{q}} \geqslant 1 - \delta/2$ over the first $\mathsf{q} = T + 1$ batches.

$\mathbf{3^o}$. Finally, to obtain Eq. (20), we apply Lemma 6 once again, this time conditioning on $\widehat{\mathbf{S}}^{(T)}$, and using the last batch $(X_{n-r+1}, ..., X_n)$ with the final estimator $\widehat{\mathbf{S}}$ in the role of $\widehat{\mathbf{S}}^{(+)}$. Note that the first condition Eq. (15) of Lemma 6 follows from the just proved induction claim. On the other hand, by Eq. (18) the final truncation level θ_T, cf. line 12 of Algorithm 1, satisfies

$$\theta_T = 2\theta\sqrt{\mathsf{q}\log(4\mathsf{q}d/\delta)/\log(4d/\delta)} \geqslant 4\kappa^2\sqrt{n\mathsf{df}_\lambda(\mathbf{S})/\log(4d/\delta)}.$$

The number of degrees of freedom is a stable quantity: we can easily prove (see Lem. 14 in Appendix) that $\mathsf{df}_{\lambda/2}(\mathbf{S}) \leqslant 2\mathsf{df}_{\lambda}(\mathbf{S})$. On the other hand, $\mathsf{df}_{\lambda_T} \leqslant \mathsf{df}_{\lambda/2}$ since $\lambda_T \geqslant \lambda/2$. Using that, we have

$$\theta_T \geqslant 2\sqrt{2}\kappa^2\sqrt{n\mathsf{df}_{\lambda_T}(\mathbf{S})/\log(4d/\delta)} \geqslant 2\sqrt{2}\kappa^2\sqrt{r\mathsf{df}_{\lambda_T}(\mathbf{S})/\log(4d/\delta)},$$

meeting the requirement on the truncation level imposed in Lemma 6. Applying the lemma, and using that $r \geqslant n/2$, we get that with conditional probability $\geqslant 1-\delta/2$ over the last batch $(X_{n-r+1}, ..., X_n)$,

$$\left\|\mathbf{S}_{\lambda_T/2}^{-1/2}(\widehat{\mathbf{S}}-\mathbf{S})\mathbf{S}_{\lambda_T/2}^{-1/2}\right\| \leqslant \frac{6\theta_T\log(4d/\delta)}{r} \leqslant \frac{24\theta\sqrt{\mathsf{q}\log(4\mathsf{q}d/\delta)\log(4d/\delta)}}{n}.$$

Since $\lambda_T \leqslant \lambda$ implies $\|\mathbf{S}_{\lambda_T/2}\mathbf{S}_{\lambda}^{-1}\| \leqslant 1$, by the union bound we arrive at Eq. (20). ■

5. Adaptive Estimator

One limitation of Algorithm 1 is that the truncation level θ has to be chosen in advance in order to obtain the optimal statistical performance (see Theorem 7), and the optimal choice θ_* (see Eq. (18)) depends on κ, $\mathsf{df}_{\lambda}(\mathbf{S})$ that are usually unknown. To address this, we propose an *adaptive estimator* (see Algorithm 2), in which Algorithm 1 is combined with a Lepskii-type procedure (Lepskii, 1991), resulting in a near-optimal guarantee without the knowledge of θ_*. Namely, let us be given a range $0 < \theta_{\min} \leqslant \theta_{\max}$ known to contain θ_* but possibly very loose, and define the logarithmic grid

$$\theta_j = 2^j\theta_{\min}, \quad \text{where} \;\; j \in \mathcal{J} := \{j \in \mathbb{Z} : \theta_{\min} \leqslant \theta_j \leqslant 2\theta_{\max}\}. \tag{26}$$

Define also

$$\varepsilon_j := \frac{24\theta_j\sqrt{\mathsf{q}\cdot\log(4\mathsf{q}d|\mathcal{J}|/\delta)\cdot\log(4d|\mathcal{J}|/\delta)}}{n}. \tag{27}$$

Later on we will we show (see Thm. 7) that ε_j is the error bound, with probability at least $1-\delta$, for the estimator produced by Algorithm 1 with truncation level $\theta = \theta_j$. In Algorithm 2, we first compute ε_j and the basic estimators $\widehat{\mathbf{S}}_j := \widehat{\mathbf{S}}[\theta_j]$ for all truncation levels θ_j, then select

$$\widehat{\jmath} = \min\left\{j \in \mathcal{J} : \;\; \forall j' \geqslant j \;\; \text{s.t.} \;\; j' \in \mathcal{J} \;\; \text{it holds} \;\; \left\|\widehat{\mathbf{S}}_{j',\lambda}^{-1/2}(\widehat{\mathbf{S}}_{j'} - \widehat{\mathbf{S}}_j)\widehat{\mathbf{S}}_{j',\lambda}^{-1/2}\right\| \leqslant 2(\varepsilon_{j'} + \varepsilon_j)\right\}, \tag{28}$$

and output $\widehat{\mathbf{S}}_{\widehat{\jmath}}$ as the final estimator. In Thm. 10 below, we show that this estimator admits essentially the same statistical guarantee (in the sense of Eq. (11)) as the "ideal" estimator which uses $\theta = \theta_*$.

Algorithm 2: Robust Calibrated Covariance Estimation with Adaptive Truncation Level

```
Input: $X_1, ..., X_n \in \mathbb{R}^d$, $\delta \in (0,1]$, regularization level $\lambda \leqslant \|\mathbf{S}\|$, $L \geqslant \|\mathbf{S}\|$, range $[\theta_{\min}, \theta_{\max}]$
Form the grid $\mathcal{J} := \{j \in \mathbb{Z} : \theta_{\min} \leqslant \theta_j \leqslant 2\theta_{\max}\}$
for $j \in \mathcal{J}$ do
    Compute the output $\widehat{\mathbf{S}}_j$ of Algorithm 1 with truncation level $\theta_j = 2^j\theta_{\min}$; compute $\varepsilon_j$ by (27)
end for
Output: $\widehat{\mathbf{S}}_{\widehat{\jmath}}$ with $\widehat{\jmath} \in \mathcal{J}$ selected according to (28)
```

Next we present a statistical performance guarantee for Algorithm 2. Its proof, given in Appendix C, hinges upon the observation that the matrix $\widehat{\mathbf{S}}_{j',\lambda}$ in the error bound of Eq. (28) can essentially be replaced with its unobservable counterpart $\mathbf{S}_{\lambda}$; this makes the analyzed errors *additive*, so that the usual argument for Lepskii's method could be applied.

Theorem 10 *Assume* (HT)*, and let Algorithm* 2 *be initialized with* $\lambda \leqslant \|\mathbf{S}\|$, $L \geqslant \|\mathbf{S}\|$, $\delta \in (0,1]$, *and a range* $[\theta_{\min}, \theta_{\max}]$ *containing the optimal truncation level* θ_* *given by Eq.* (18)*. Moreover, let*

$$n \geqslant 96\mathsf{q}\theta_{\max} \log(4\mathsf{q}d|\mathcal{J}|/\delta), \tag{29}$$

where q *is defined in Eq.* (18)*, and* $|\mathcal{J}| \leqslant 1 + \log_2(\theta_{\max}/\theta_{\min})$ *is the cardinality of the grid defined in Eq.* (26)*. Then the estimator* $\widehat{\mathbf{S}}_{\widehat{\jmath}}$ *produced by Algorithm* 2 *with probability at least* $1-\delta$ *satisfies*

$$\left\|\mathbf{S}_\lambda^{-1/2}(\widehat{\mathbf{S}}_{\widehat{\jmath}} - \mathbf{S})\mathbf{S}_\lambda^{-1/2}\right\| \leqslant 720\kappa^2\sqrt{\frac{(1+\rho)\mathsf{df}_\lambda(\mathbf{S})\log(4d|\mathcal{J}|/\delta)}{n}}, \quad \textit{where} \quad \rho := \frac{\log|\mathcal{J}|}{\log(4\mathsf{q}d/\delta)}.$$

From the result of Thm. 10, we see that the adaptive estimator nearly attains the best possible stastistical guarantee that corresponds to the optimal value of the truncation level, up to the iterated logarithm of the ratio $\theta_{\max}/\theta_{\min}$. However, the premise of Thm. 10 requires $\theta_{\max}$ to be bounded both from above and below (cf. Eqs. (18) and (29)), and the two bounds are compatible only starting from a certain sample size. Next we state a corollary of Thm. 10 that explicitly specifies the required sample size (the requirement is similar to Eq. (21)), and provides a reasonables choice of $[\theta_{\min}, \theta_{\max}]$.

Corollary 11 *Assume that we have*

$$n \geqslant 192^2(1+\rho_J)\kappa^4\mathsf{q}\mathsf{df}_\lambda(\mathbf{S})\log(4\mathsf{q}dJ/\delta), \tag{30}$$

where $J := 1 + \left\lceil \frac{1}{2}\log_2(n/96\mathsf{q})\right\rceil$ *and* $\rho_J := \log(J)/\log(4\mathsf{q}d/\delta)$*. Then the premise of Theorem* 10 *holds for the grid* $\mathcal{J}$ *with cardinality* J *defined by* $\theta_{\max} = n/96\mathsf{q}\log(4\mathsf{q}dJ/\delta)$, $\theta_{\min} = 2^{1-J}\theta_{\max}$.

6. Applications

6.1. Relative-Scale Bounds for Eigenvalues

Recall that the bounds obtained in Thms. 7 and 10 for the estimators $\widehat{\mathbf{S}}$ given by Algorithms 1–2 read

$$(1-\varepsilon)\mathbf{S}_\lambda \preccurlyeq \widehat{\mathbf{S}}_\lambda \preccurlyeq (1+\varepsilon)\mathbf{S}_\lambda \tag{31}$$

for certain accuracy $\varepsilon < 1/2$ and regularization level $\lambda \geqslant 0$, provided that the sample size is large enough. Using that the positive-semidefinite order preserves the order of eigenvalues, we obtain the corollary of Thm. 7 for eigenvalue estimation (Thm. 10 has a similar corollary, which we omit).

Corollary 12 *Assume that* n *satisfies Eq.* (21)*, and let* $\widehat{\mathbf{S}}$ *be given by Algorithm* 1 *with the optimal choice of the truncation level* $\theta = \theta_*$*, cf. Eq.* (18)*. Let also* $\|\mathbf{S}\| = \lambda_1 \geqslant ... \geqslant \lambda_d = \lambda_{\min}$ *be the ordered eigenvalues of* $\mathbf{S}$*, and* $\widehat{\lambda}_1 \geqslant ... \geqslant \widehat{\lambda}_d$ *those of* $\widehat{\mathbf{S}}$*. Finally, assume that the regularization level in Algorithm* 1 *satisfies* $\lambda \leqslant \lambda_k$ *for some* $1 \leqslant k \leqslant d$*. Then, with probability at least* $1-\delta$ *it holds*

$$(1-2\varepsilon)\lambda_i \leqslant \widehat{\lambda}_i \leqslant (1+2\varepsilon)\lambda_i, \quad \textit{for any} \;\; 1 \leqslant i \leqslant k, \tag{32}$$

with ε *given by* (22)*. As a consequence, we have* $(1-2\varepsilon)^2 \cdot \lambda_i/\lambda_k \leqslant \widehat{\lambda}_i/\widehat{\lambda}_k \leqslant (1-2\varepsilon)^{-2} \cdot \lambda_i/\lambda_k$.

As a simple application of this result, consider the task of “noisy” principal component analysis (PCA), i.e., performing PCA for the unknown covariance matrix $\mathbf{S}$ from the observations $X_1, ..., X_n$. A common way to approach it is by performing *subspace iteration* (Halko et al., 2011; Mitliagkas et al., 2013; Hardt and Price, 2014; Balcan et al., 2016) with the estimated covariance $\widehat{\mathbf{S}}$: randomly choose $U^{(0)} \in \mathbb{R}^{d\times k}$, and then iteratively multiply $U^{(t)}$ by $\widehat{\mathbf{S}}$ and orthonormalize the result until convergence. The iterate converges to the projector on the subspace of the top k eigenvalues of $\widehat{\mathbf{S}}$ (providing an estimate of the corresponding subspace for $\mathbf{S}$), and its rate of convergence is known to be controlled by the ratio $\widehat{\lambda}_k/\widehat{\lambda}_{k+1}$.[4] Hence, if we use the estimate $\widehat{\mathbf{S}}$ produced by Algorithm 1 or Algorithm 2, and if n is sufficient to guarantee that $\varepsilon < 1/2$, the convergence rate to the top-k eigenspace of $\widehat{\mathbf{S}}$ will essentially be the same as that for the exact method and the target subspace of $\mathbf{S}$.

6.2. Ridge Regression with Heavy-Tailed Observations

In random design linear regression (Hsu et al., 2012), one wants to fit the linear model $Y = X^\top w$ from i.i.d. observations $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \leqslant i \leqslant n$. More previsely, the goal is to find a minimizer $w^* \in \mathbb{R}^d$ of the quadratic risk $L(w) := \mathbb{E}(Y - X^\top w)^2$, where the expectation is over the test pair (X, Y) independent of the sample and coming from the same distribution. In ordinary ridge regression, one fixes the regularization level $\lambda \geqslant 0$, and estimates w^* with the minimizer of the regularized empirical risk $\widetilde{L}_\lambda(w) := \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^\top w)^2 + \lambda\|w\|^2$. The case $\lambda = 0$ corresponds to the ordinary least-squares estimator, while $\lambda > 0$ allows for some bias.

Here we propose a couterpart of this estimator with a favorable statistical guarantee under fourth-moment assumptions on the design and response. Given the sample $X_1, ..., X_{2n}$, we first compute the covariance estimator $\widehat{\mathbf{S}}$ by feeding the hold-out sample $X_{n+1}, ..., X_{2n}$ to Algorithm 1 with $\theta = \theta_*$ (one could also use Algorithm 2). Then, using the first half of the observations, we construct the "pseudo-decorrelated" observations $(\widehat{Z}_1, ..., \widehat{Z}_n)$ with $\widehat{Z}_i = \widehat{\mathbf{S}}_\lambda^{-1/2} X_i Y_i$, and compute the estimator

$$\bar{w}_\lambda = \widehat{\mathbf{S}}_\lambda^{-1/2}\bar{Z}, \quad \text{where} \quad \bar{Z} = \frac{1}{n}\sum_{i=1}^n \rho_{\bar{\theta}}(\|\widehat{Z}_i\|^{1/2})\widehat{Z}_i. \tag{33}$$

Here, $\rho_{\bar{\theta}}(\cdot)$ is as in Eq. (6)–(7), and $\bar{\theta}$ is defined later. We prove the following result (see Appendix D).

Theorem 13 *In the above setting, assume that X satisfies $\mathbb{E}[X] = 0$, $\mathbb{E}[X \otimes X] = \mathbf{S}$, and assumption (HT), and that Y has finite second and fourth moments: $\mathbb{E}[Y^2] \leqslant v^2$, $\mathbb{E}[Y^4] \leqslant \varkappa^4 v^4$. Assume also that n satisfies Eq. (21) from the premise of Thm. 7. Then, the estimator $\bar{w}_\lambda$ given by (33) with $\bar{\theta} = \sqrt{n\kappa^2\varkappa^2 v^2 \mathsf{df}_\lambda(\mathbf{S})/\log(1/\delta)}$ with probability at least $1-\delta$ satisfies*

$$L(\bar{w}_\lambda) - L(w^*) \leqslant O(1)\left[(\kappa^4 + \kappa^2\varkappa^2)\frac{v^2\mathsf{df}_\lambda(\mathbf{S})\log(2d/\delta)}{n} + \lambda^2\left\|\mathbf{S}_\lambda^{-1/2}w^*\right\|^2\right]. \tag{34}$$

In the above result, the bias term is correct (leading to the minimax-optimal rates in the fixed design setting), and the stochastic term has the asymptotically optimal scaling $O(d_{\mathrm{eff}}\log(1/\delta)/n)$, see, e.g., Caponnetto and De Vito (2007). However, the obtained bound depends on the second moment v^2 of the response instead of its variance. We believe that this problem could be fixed, leading to the fully optimal result, by replacing the truncated estimator $\bar{Z}$ in Eq. (33) with median-of-means.

Acknowledgments

The first author has been supported by the ERCIM Alain Bensoussan Fellowship. The second author acknowledges support from the European Research Council (grant SEQUOIA 724063). We thank Zaid Harchaoui and Anatoli Juditsky for fruitful discussions, and thank anonymous reviewers and Nikita Zhivotovskiy for insightful remarks, as well as for pointing out relevant results in the literature.

[4] One can use $U^{(t)} \in \mathbb{R}^{d\times r}$, $r \geqslant k$; the convergence rate is then controlled by the ratio of *non-sequential* eigenvalues.

References

Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

Maria-Florina Balcan, Simon S. Du, Yining Wang, and Adams W. Yu. An improved gap-dependency analysis of the noisy power method. In *Conference on Learning Theory*, pages 284–309, 2016.

Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

Brendan O. Bradley and Murad S. Taqqu. Financial risk and heavy tails. *Handbook of Heavy-Tailed Distributions in Finance, ST Rachev, ed. Elsevier, Amsterdam*, pages 35–103, 2003.

T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38(4):2118–2144, 2010. doi: 10.1214/09-AOS752.

Norm A. Campbell. Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied statistics*, pages 231–237, 1980.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

Gene H. Golub and Charles F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2): 217–288, 2011.

Lars P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *The Journal of Machine Learning Research*, 23(9):1–24, 2012.

Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2002.

Bryan Kelly and Hao Jiang. Tail risk and asset prices. *The Review of Financial Studies*, 27(10): 2841–2871, 2014.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 02 2017.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Oleg V. Lepskii. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

Hendrik P. Lopuhaa and Peter J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19(1):229–248, 1991.

Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 08 2014.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Shahar Mendelson and Nikita Zhivotovskiy. Robust covariance estimation under $L_4 - L_2$ norm equivalence. *arXiv:1809.10462*, 2018.

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.

Stanislav Minsker and Xiaohan Wei. Estimation of the covariance structure of heavy-tailed distributions. *arXiv:1708.00502*, 2017.

Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. Memory limited, streaming pca. In *Advances in Neural Information Processing Systems*, pages 2886–2894, 2013.

Roberto I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electron. Commun. Probab.*, 15(26):203–212, 2010.

Roberto I. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3-4):1175–1194, 2016.

Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

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

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

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

Appendix A. Degrees of Freedom Lemma

Lemma 14 *For any* $\mathbf{S} \succcurlyeq 0$ *and* $\lambda \geqslant 0$, *define* $\mathsf{df}_\lambda(\mathbf{S}) := \mathrm{Tr}(\mathbf{S}\mathbf{S}_\lambda^{-1})$. *Then,* $\mathsf{df}_{\lambda/2}(\mathbf{S}) \leqslant 2\mathsf{df}_\lambda(\mathbf{S})$.

Proof We have

$$\begin{aligned}
|\mathsf{df}_{\lambda/2}(\mathbf{S}) - \mathsf{df}_\lambda(\mathbf{S})| &= \left|\mathrm{Tr}\left[\mathbf{S}\left(\mathbf{S}_\lambda^{-1} - \mathbf{S}_{\lambda/2}^{-1}\right)\right]\right| \\
&= \left|\mathrm{Tr}\left[\mathbf{S}_\lambda^{-1/2}\mathbf{S}\mathbf{S}_\lambda^{-1/2}\left(\mathbf{I} - \mathbf{S}_\lambda^{1/2}\mathbf{S}_{\lambda/2}^{-1}\mathbf{S}_\lambda^{1/2}\right)\right]\right| \\
&\leqslant \mathsf{df}_\lambda(\mathbf{S})\left\|\mathbf{I} - \mathbf{S}_\lambda^{1/2}\mathbf{S}_{\lambda/2}^{-1}\mathbf{S}_\lambda^{1/2}\right\| \leqslant \mathsf{df}_\lambda(\mathbf{S}),
\end{aligned}$$

where we first used commutativity of the trace, and then the fact (following from the trace Hölder inequality) that $|\mathrm{Tr}(\mathbf{A}\mathbf{B})| \leqslant \|\mathbf{B}\|\,\mathrm{Tr}(\mathbf{A})$ for $\mathbf{A} \succcurlyeq 0$ and $\mathbf{B}$ with compatible dimensions. The claim follows. ■

Appendix B. Proof of Lemma 6

$\mathbf{1^o}$. We start by deriving the consequences of (15). First, note that

$$\mathbf{S}_\lambda^{1/2}\widehat{\mathbf{S}}_\lambda^{-1}\mathbf{S}_\lambda^{1/2} = \mathbf{S}_\lambda^{1/2}\left[\mathbf{S}_\lambda - (\mathbf{S} - \widehat{\mathbf{S}})\right]^{-1}\mathbf{S}_\lambda^{1/2} = \left[\mathbf{I} - \mathbf{S}_\lambda^{-1/2}(\mathbf{S} - \widehat{\mathbf{S}})\mathbf{S}_\lambda^{-1/2}\right]^{-1}.$$

Whence, using (15) and the similarity rules,

$$\frac{2}{3}\mathbf{I} \preccurlyeq \mathbf{S}_\lambda^{1/2}\widehat{\mathbf{S}}_\lambda^{-1}\mathbf{S}_\lambda^{1/2} \preccurlyeq 2\mathbf{I}. \tag{35}$$

By the properties of the spectral norm, this implies

$$\begin{aligned}
\left\|\mathbf{S}_\lambda^{1/2}\widehat{\mathbf{S}}_\lambda^{-1/2}\right\|^2 &= \left\|\widehat{\mathbf{S}}_\lambda^{-1/2}\mathbf{S}_\lambda^{1/2}\right\|^2 \leqslant 2; \\
\left\|\mathbf{S}_\lambda^{-1/2}\widehat{\mathbf{S}}_\lambda^{1/2}\right\|^2 &= \left\|\widehat{\mathbf{S}}_\lambda^{1/2}\mathbf{S}_\lambda^{-1/2}\right\|^2 \leqslant \frac{3}{2}.
\end{aligned} \tag{36}$$

Using that, and proceding as in the proof of Lemma 14, we can bound the degrees of freedom surrogate $\mathrm{Tr}(\mathbf{S}\widehat{\mathbf{S}}_\lambda^{-1})$ in terms of the true quantity $\mathsf{df}_\lambda(\mathbf{S}) = \mathrm{Tr}(\mathbf{S}\mathbf{S}_\lambda^{-1})$:

$$\begin{aligned}
\mathrm{Tr}(\mathbf{S}\widehat{\mathbf{S}}_\lambda^{-1}) - \mathsf{df}_\lambda(\mathbf{S}) &= \mathrm{Tr}\left[\mathbf{S}(\mathbf{S}_\lambda^{-1} - \widehat{\mathbf{S}}_\lambda^{-1})\right] \\
&= \mathrm{Tr}\left[\mathbf{S}\mathbf{S}_\lambda^{-1/2}\left(\mathbf{I} - \mathbf{S}_\lambda^{1/2}\widehat{\mathbf{S}}_\lambda^{-1}\mathbf{S}_\lambda^{1/2}\right)\mathbf{S}_\lambda^{-1/2}\right] \\
&= \mathrm{Tr}\left[\mathbf{S}_\lambda^{-1/2}\mathbf{S}\mathbf{S}_\lambda^{-1/2}\left(\mathbf{I} - \mathbf{S}_\lambda^{1/2}\widehat{\mathbf{S}}_\lambda^{-1}\mathbf{S}_\lambda^{1/2}\right)\right],
\end{aligned}$$

where in the third line we used commutativity of the trace. Applying the trace Hölder inequality as in the proof of Lemma 14, we obtain

$$\left|\mathrm{Tr}(\mathbf{S}\widehat{\mathbf{S}}_\lambda^{-1}) - \mathsf{df}_\lambda(\mathbf{S})\right| \leqslant \left\|\mathbf{I} - \mathbf{S}_\lambda^{1/2}\widehat{\mathbf{S}}_\lambda^{-1}\mathbf{S}_\lambda^{1/2}\right\|\mathsf{df}_\lambda(\mathbf{S}) \leqslant 3\mathsf{df}_\lambda(\mathbf{S}), \tag{37}$$

where we combined the triangle inequality with the right-hand side of (35).

2^o. We now invoke the results of Wei and Minsker (2017) (in what follows, the expectation is conditioned on $\widehat{\mathbf{S}}_\lambda$). Note that conditionally on $\widehat{\mathbf{S}}$, the random vectors $Z_j = \widehat{\mathbf{S}}_\lambda^{-1/2} X_j$ are i.i.d. with mean zero and covariance $\widehat{\mathbf{J}} = \widehat{\mathbf{S}}_\lambda^{-1/2} \mathbf{S} \widehat{\mathbf{S}}_\lambda^{-1/2}$. By (9), and using the linear invariance of (HT),

$$\left\| \mathbb{E}\left[\|Z_j\|^2 Z_j \otimes Z_j \right] \right\| \leqslant \kappa^4 \|\widehat{\mathbf{J}}\|^2 \mathtt{r}(\widehat{\mathbf{J}}) = \kappa^4 \|\widehat{\mathbf{J}}\| \operatorname{Tr}(\widehat{\mathbf{J}}).$$

Using (35) and (37) to bound $\|\widehat{\mathbf{J}}\|$ and $\operatorname{Tr}(\widehat{\mathbf{J}})$ correspondingly, this results in

$$\|\mathbb{E}\|Z_j\|^2 Z_j \otimes Z_j\| \leqslant 8\kappa^4 \mathsf{df}_\lambda(\mathbf{S}). \tag{38}$$

On the other hand, the estimator $\widehat{\mathbf{S}}^{(+)}$ defined in (16) satisfies

$$\widehat{\mathbf{S}}_\lambda^{-1/2} \widehat{\mathbf{S}}^{(+)} \widehat{\mathbf{S}}_\lambda^{-1/2} = \frac{1}{m} \sum_{j=1}^{m} \rho_\theta(\|Z_j\|) Z_j \otimes Z_j,$$

that is, $\widehat{\mathbf{S}}_\lambda^{-1/2} \widehat{\mathbf{S}}^{(+)} \widehat{\mathbf{S}}_\lambda^{-1/2}$ is precisely the Wei-Misnker estimator (cf. (7)) of $\widehat{\mathbf{J}}$, computed from the sample $(Z_1, ..., Z_m)$. Hence, combining the result of Theorem 3 with (38), we see that whenever

$$\theta \geqslant 2\sqrt{2}\kappa^2 \sqrt{\frac{m\mathsf{df}_\lambda(\mathbf{S})}{\log(2d/\delta)}},$$

with conditional probability at least $1 - \delta$ it holds

$$\left\| \widehat{\mathbf{S}}_\lambda^{-1/2} (\widehat{\mathbf{S}}^{(+)} - \mathbf{S}) \widehat{\mathbf{S}}_\lambda^{-1/2} \right\| \leqslant \frac{2\theta \log(2d/\delta)}{m}.$$

Finally, we arrive at (17) by writing

$$\left\| \mathbf{S}_{\lambda/2}^{-1/2} (\widehat{\mathbf{S}}^{(+)} - \mathbf{S}) \mathbf{S}_{\lambda/2}^{-1/2} \right\| \leqslant \left\| \mathbf{S}_{\lambda/2}^{-1/2} \mathbf{S}_\lambda^{1/2} \right\|^2 \left\| \mathbf{S}_\lambda^{-1/2} \widehat{\mathbf{S}}_\lambda^{1/2} \right\|^2 \left\| \widehat{\mathbf{S}}_\lambda^{-1/2} (\widehat{\mathbf{S}}^{(+)} - \mathbf{S}) \widehat{\mathbf{S}}_\lambda^{-1/2} \right\|,$$

noting that $\left\| \mathbf{S}_{\lambda/2}^{-1/2} \mathbf{S}_\lambda^{1/2} \right\|^2 = \left\| \mathbf{S}_{\lambda/2}^{-1} \mathbf{S}_\lambda \right\| \leqslant 2$, and bounding $\left\| \mathbf{S}_\lambda^{-1/2} \widehat{\mathbf{S}}_\lambda^{1/2} \right\|^2 \leqslant 3/2$ via (36) ■

Appendix C. Proof of Theorem 10 and Corollary 11

Let us call the truncation level $\theta_j = \theta_{\min} 2^j$, with $j \in \mathcal{J}$, *admissible* if it satisfies the condition in (28), so that $\theta_{\hat{\jmath}}$ is the smallest such level. Let $j^* \in \mathcal{J}$ be the minimal $j \in \mathcal{J}$ such that $\theta_{j^*} \geqslant \theta^*$ for θ_* defined in Eq. (18); note that this is always possible by the definition (26), and we have

$$\theta_{j^*} \leqslant 2\theta_*. \tag{39}$$

1^o. Let us prove that θ_{j^*} is admissible with probability at least $1 - \delta$. Indeed, due to (29), the premise (19) of Theorem 7 holds for any $\theta = \theta_j$ with $j \in \mathcal{J}$ (recall that $\theta_j \leqslant 2\theta_{\max}$). On the other hand, the premise (18) of Theorem 7 holds whenever $\theta_j \geqslant \theta_*$. Hence the bound (20) of Theorem 7 holds for all $\theta = \theta_j$ with $j \geqslant j^*$, and by the union bound we get that with probability at least $1 - \delta$,

$$\left\| \mathbf{S}_\lambda^{-1/2} (\widehat{\mathbf{S}}_j - \mathbf{S}) \mathbf{S}_\lambda^{-1/2} \right\| \leqslant \varepsilon_j = \frac{24\theta_j \sqrt{\mathtt{q} \log(4\mathtt{q}d|\mathcal{J}|/\delta) \log(4d|\mathcal{J}|/\delta)}}{n}, \quad \forall j \geqslant j_*, \;\; j \in \mathcal{J}. \tag{40}$$

Moreover, from (29) we also obtain

$$\varepsilon_j \leqslant \frac{1}{2}\sqrt{\frac{\log(4d|\mathcal{J}|/\delta)}{\mathsf{q}\log(4\mathsf{q}d|\mathcal{J}|/\delta)}} \leqslant \frac{1}{2}, \quad j \in \mathcal{J}.$$

Whence for any $j \in \mathcal{J}$ such that $j \geqslant j_*$ we have, under the event (40), and denoting $\widehat{\mathbf{S}}_{j,\lambda} = \widehat{\mathbf{S}}_j + \lambda \mathbf{I}$,

$$\begin{aligned}\left\|\widehat{\mathbf{S}}_{j,\lambda}^{-1/2}(\widehat{\mathbf{S}}_j - \widehat{\mathbf{S}}_{j_*})\widehat{\mathbf{S}}_{j,\lambda}^{-1/2}\right\| &\leqslant \left\|\widehat{\mathbf{S}}_{j,\lambda}^{-1/2}\mathbf{S}_\lambda^{1/2}\right\|^2 \cdot \left\|\mathbf{S}_\lambda^{-1/2}(\widehat{\mathbf{S}}_j - \widehat{\mathbf{S}}_{j_*})\mathbf{S}_\lambda^{-1/2}\right\| \\ &\leqslant \left\|\widehat{\mathbf{S}}_{j,\lambda}^{-1/2}\mathbf{S}_\lambda^{1/2}\right\|^2 \left(\left\|\mathbf{S}_\lambda^{-1/2}(\widehat{\mathbf{S}}_j - \mathbf{S})\mathbf{S}_\lambda^{-1/2}\right\| + \left\|\mathbf{S}_\lambda^{-1/2}(\widehat{\mathbf{S}}_{j_*} - \mathbf{S})\mathbf{S}_\lambda^{-1/2}\right\|\right) \\ &\leqslant 2(\varepsilon_j + \varepsilon_{j_*})\end{aligned}$$

where we used (40) to bound the terms in the parentheses, and also used (cf. (36) in Appendix B):

$$\left\|\widehat{\mathbf{S}}_{j,\lambda}^{-1/2}\mathbf{S}_\lambda^{1/2}\right\|^2 \leqslant \frac{1}{1-\varepsilon_j} \leqslant 2, \quad \forall j \geqslant j_*.$$

Thus, j_* is indeed admissible with probability $\geqslant 1-\delta$.

2^o. Whenever j_* is admissible, we have $\widehat{\jmath} \leqslant j_*$, whence $\varepsilon_{\widehat{\jmath}} \leqslant \varepsilon_{j_*}$ using that ε_j increases in j. Thus, with probability at least $1-\delta$ it holds

$$\begin{aligned}&\left\|\mathbf{S}_\lambda^{-1/2}(\widehat{\mathbf{S}}_{\widehat{\jmath}} - \mathbf{S})\mathbf{S}_\lambda^{-1/2}\right\| \\ &\quad\leqslant \left\|\mathbf{S}_\lambda^{-1/2}\widehat{\mathbf{S}}_{j_*,\lambda}^{1/2}\right\|^2 \cdot \left\|\widehat{\mathbf{S}}_{j_*,\lambda}^{-1/2}(\widehat{\mathbf{S}}_{\widehat{\jmath}} - \mathbf{S})\widehat{\mathbf{S}}_{j_*,\lambda}^{-1/2}\right\| \\ &\quad\leqslant \left\|\mathbf{S}_\lambda^{-1/2}\widehat{\mathbf{S}}_{j_*,\lambda}^{1/2}\right\|^2 \cdot \left(\left\|\widehat{\mathbf{S}}_{j_*,\lambda}^{-1/2}(\widehat{\mathbf{S}}_{j_*} - \widehat{\mathbf{S}}_{\widehat{\jmath}})\widehat{\mathbf{S}}_{j_*,\lambda}^{-1/2}\right\| + \left\|\widehat{\mathbf{S}}_{j_*,\lambda}^{-1/2}(\widehat{\mathbf{S}}_{j_*} - \mathbf{S})\widehat{\mathbf{S}}_{j_*,\lambda}^{-1/2}\right\|\right) \\ &\quad\leqslant \frac{3}{2}(2(\varepsilon_{\widehat{\jmath}} + \varepsilon_{j_*}) + \varepsilon_{j_*}) \leqslant \frac{15}{2}\varepsilon_{j_*},\end{aligned} \tag{41}$$

where in order to obtain the last line we used (cf. (36) in Appendix B) that

$$\left\|\mathbf{S}_\lambda^{-1/2}\widehat{\mathbf{S}}_{j_*,\lambda}^{1/2}\right\|^2 \leqslant 1 + \varepsilon_{j_*} \leqslant 3/2.$$

Finally, combining this with the expression for θ_* in (18), and using (39)–(41), we arrive at the claimed bound. ■

Proof of Corollary 11 First, $\theta_{\max}$ defined in the premise satisfies Eq. (29) by construction; moreover, Eq. (29) is satisfied as an equality. On the other hand, by simple algebra Eq. (30) guarantees that $\theta_{\max} \geqslant \theta_*$ for θ_* defined in Eq. (18). Finally, verifying that $\theta_{\min} \leqslant \theta_*$ is trivial using that $\kappa \geqslant 1$ and $\mathsf{df}_\lambda(\mathbf{S}) \geqslant 1$. ■

Appendix D. Proof of Theorem 13

1^o. First of all, note that n satisfying (21) suffices to guarantee that

$$\left\|\mathbf{S}_\lambda^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})\mathbf{S}_\lambda^{-1/2}\right\| \leqslant 48\kappa^2\sqrt{\frac{\mathsf{df}_\lambda(\mathbf{S})\log(2d/\delta)}{n}} \leqslant \frac{1}{2} \tag{42}$$

holds with probability at least $1-\delta/2$, cf. (22). Note also that $\widehat{\mathbf{S}}_\lambda$ is independent from $(X_1, ..., X_n)$, hence the vectors $\widehat{Z}_i = \widehat{\mathbf{S}}_\lambda^{-1/2} X_i Y_i$, $1 \leqslant i \leqslant n$, are independent when conditioned on $(X_{n+1}, ..., X_{2n})$. Finally, the conditional to the hold-out sample $X_{n+1}, ..., X_{2n}$ expectation of $\widehat{Z}_i$ is

$$\widehat{\mathbb{E}}[\widehat{Z}_i] = \widehat{\mathbf{S}}_\lambda^{-1/2} \mathbf{S} w_*, \tag{43}$$

where we used that the residual $\xi = Y - X^\top w_*$ satisfies $\mathbb{E}[\xi X] = 0$, which follows from the fact that w_* minimizes $L(w)$.

$\mathbf{2^o}$. We now decompose the excess risk of $\bar{w}_\lambda$ as follows:

$$\sqrt{L(\bar{w}_\lambda) - L(w^*)} \leqslant \underbrace{\|\mathbf{S}^{1/2}(\bar{w}_\lambda - \widehat{w}_\lambda)\|}_{E_1} + \underbrace{\|\mathbf{S}^{1/2}(\widehat{w}_\lambda - w_\lambda)\|}_{E_2} + \underbrace{\|\mathbf{S}^{1/2}(w_\lambda - w^*)\|}_{E_3}, \tag{44}$$

where w_λ, given by

$$w_\lambda = \mathbf{S}_\lambda^{-1} \mathbf{S} w^*, \tag{45}$$

is the minimizer of $L_\lambda(w) = L(w) + \lambda \|w\|^2$, and $\widehat{w}_\lambda := \widehat{\mathbb{E}}[\widehat{\mathbf{S}}_\lambda^{-1/2} \widehat{Z}_1]$ can be calculated using (43):

$$\widehat{w}_\lambda = \widehat{\mathbf{S}}_\lambda^{-1} \mathbf{S} w^*. \tag{46}$$

The easiest to control in (44) is the term E_3 corresponding to the squared bias in the fixed-design setting:

$$\|\mathbf{S}^{1/2}(w_\lambda - w^*)\| \leqslant \|\mathbf{S}_\lambda^{1/2}(w_\lambda - w^*)\| = \lambda \|\mathbf{S}_\lambda^{-1/2} w^*\|, \tag{47}$$

resulting in the second term in the brackets in (34).

$\mathbf{3^o}$. On the other hand, using (45)–(46) we have

$$\begin{aligned} E_2 &\leqslant \left\|\mathbf{S}_\lambda^{1/2}(\mathbf{S}_\lambda^{-1} - \widehat{\mathbf{S}}_\lambda^{-1})\mathbf{S} w_*\right\| \\ &= \left\|\mathbf{S}_\lambda^{1/2} \widehat{\mathbf{S}}_\lambda^{-1} (\mathbf{S} - \widehat{\mathbf{S}}) \mathbf{S}_\lambda^{-1} \mathbf{S} w_*\right\| \\ &\leqslant \left\|\mathbf{S}_\lambda^{1/2} \widehat{\mathbf{S}}_\lambda^{-1/2}\right\|^2 \cdot \left\|\mathbf{S}_\lambda^{-1/2}(\mathbf{S} - \widehat{\mathbf{S}})\mathbf{S}_\lambda^{-1/2}\right\| \cdot \left\|\mathbf{S}_\lambda^{-1/2}\mathbf{S}^{1/2}\right\| \cdot \left\|\mathbf{S}^{1/2} w^*\right\|, \end{aligned} \tag{48}$$

where the last inequality can be verified by removing the norms. Under the event (42), we can bound the first term by a constant (see the proof of Lemma B in Appendix B), and the second term by

$$O(1)\kappa^2 \sqrt{\frac{\mathsf{df}_\lambda(\mathbf{S}) \log(2d/\delta)}{n}},$$

cf. (42). The third term is at most one. Finally, we have $\|\mathbf{S}^{1/2} w^*\|^2 = \mathbb{E}[(X^\top w^*)^2] \leqslant \mathbb{E}[Y^2] = v^2$. Collecting the above, under the event (42) we have

$$E_2 \leqslant O(1)\kappa^2 \sqrt{\frac{v^2 \mathsf{df}_\lambda(\mathbf{S}) \log(2d/\delta)}{n}}.$$

$\mathbf{4^o}$. Finally, let us estimate the term E_1 which corresponds to the additive noise, and delivers the first term in the brackets in (34). Note that we can bound

$$\begin{aligned} E_1 &= \|\mathbf{S}^{1/2}(\bar{w}_\lambda - \widehat{w}_\lambda)\| \\ &\leqslant \|\mathbf{S}_\lambda^{1/2}(\bar{w}_\lambda - \widehat{w}_\lambda)\| \\ &\leqslant \left\|\mathbf{S}_\lambda^{1/2} \widehat{\mathbf{S}}_\lambda^{-1/2}\right\| \cdot \left\|\widehat{\mathbf{S}}_\lambda^{1/2}(\bar{w}_\lambda - \widehat{w}_\lambda)\right\| \\ &= \left\|\mathbf{S}_\lambda^{1/2} \widehat{\mathbf{S}}_\lambda^{-1/2}\right\| \cdot \left\|\bar{Z} - \widehat{\mathbb{E}}[\widehat{Z}]\right\|, \quad \text{where } \widehat{Z} = \widehat{\mathbf{S}}_\lambda^{-1/2} XY, \end{aligned}$$

cf. (33) and (43). Recall that under the event (42), the first term in the product is bounded by a constant, and it remains to control the deviations of the estimator $\bar{Z}$ of $\widehat{Z}$ from the (conditional) average $\widehat{\mathbb{E}}[\widehat{Z}]$. To this end, consider the following construction due to (Minsker, 2018, Sec. 3.3). For the general matrix $A \in \mathbb{R}^{d_1 \times d_2}$, define its Hermitian dilation

$$\mathcal{H}(A) = \begin{pmatrix} 0_{d_1 \times d_1} & A \\ A^\top & 0_{d_2 \times d_2} \end{pmatrix}, \tag{49}$$

and for any $\varphi : \mathbb{R} \to \mathbb{R}$, define the map on the space of $(d_1 + d_2) \times (d_1 + d_2)$ Hermitian matrices:

$$\varphi(\mathbf{A}) := Q\varphi(\mathbf{\Lambda})Q^\top = Q\,\mathbf{diag}(\varphi(\lambda_1) \cdots \varphi(\lambda_d))Q^\top,$$

where $Q\mathbf{\Lambda}Q^\top$ is the eigendecomposition of $\mathbf{A}$. In this notation, consider the following estimator of $\mathbb{E}[A] \in \mathbb{R}^{d_1 \times d_2}$ from i.i.d. copies $A_1, ..., A_n$ of A: compute the $(d_1 + d_2) \times (d_1 + d_2)$ Hermitian matrix

$$\widehat{\mathbf{T}} = \frac{1}{n}\sum_{i=1}^{n} \psi_{\bar{\theta}}(\mathcal{H}(A_i)), \tag{50}$$

where $\psi_{\bar{\theta}}(\cdot)$ is the matrix map corresponding to (6) with truncation level $\bar{\theta}$, and then output the top right block of $\widehat{\mathbf{T}}$ (i.e., the one corresponding to A in $\mathcal{H}(A)$) as the final estimate. As proved in (Minsker, 2018, Cor. 3.1), the resulting estimate satisfies

$$\left\|\bar{A} - \mathbb{E}[A]\right\| \leqslant O(1)\frac{\bar{\theta}\log(1/\delta)}{n}$$

with probability at least $1 - \delta$, provided that $\delta \leqslant 1/2$, and

$$\bar{\theta} = \sqrt{\frac{n\overline{w}}{\log(1/\delta)}} \quad \text{for some} \;\; \overline{w} \geqslant \operatorname{Tr}\mathbb{E}[A \otimes A].$$

On the other hand, one can verify that this construction reduces to $\bar{Z}$ when estimating $\widehat{\mathbb{E}}[\widehat{Z}]$ from $\widehat{Z}_1, ..., \widehat{Z}_n$ with the same $\bar{\theta}$. Thus, with (conditional) probability $\geqslant 1 - \delta/2$ it holds

$$\left\|\bar{Z} - \widehat{\mathbb{E}}[\widehat{Z}]\right\| \leqslant O(1)\frac{\bar{\theta}\log(2/\delta)}{n}$$

whenever $\bar{\theta}$ is taken to be

$$\bar{\theta} = \sqrt{\frac{n\overline{w}}{\log(1/\delta)}} \quad \text{for some} \;\; \overline{w} \geqslant \operatorname{Tr}\widehat{\mathbb{E}}[\widehat{Z} \otimes \widehat{Z}],$$

which then results in the bound

$$\left\|\bar{Z} - \widehat{\mathbb{E}}[\widehat{Z}]\right\| \leqslant O(1)\sqrt{\frac{\overline{w}\log(2/\delta)}{n}}.$$

It remains to bound $\bar{w} = \operatorname{Tr}\widehat{\mathbb{E}}[\widehat{Z} \otimes \widehat{Z}]$. Using the trace Hölder inequality, we have

$$\begin{aligned}
\operatorname{Tr}\left[\widehat{\mathbb{E}}[\widehat{Z} \otimes \widehat{Z}]\right] &= \operatorname{Tr}\left[\widehat{\mathbf{S}}_\lambda^{-1}\mathbb{E}[Y^2XX^\top]\right] \\
&\leqslant \left\|\mathbf{S}_\lambda^{1/2}\widehat{\mathbf{S}}_\lambda^{-1/2}\right\| \cdot \operatorname{Tr}\left[\mathbf{S}_\lambda^{-1}\mathbb{E}[Y^2XX^\top]\right], \\
&= \left\|\mathbf{S}_\lambda^{1/2}\widehat{\mathbf{S}}_\lambda^{-1/2}\right\| \cdot \mathbb{E}\left[\left\|Y\mathbf{S}_\lambda^{-1/2}X\right\|^2\right],
\end{aligned}$$

where the first term on the right is at most a constant under (42). Finally, under the fourth-moment assumptions in the premise of the theorem, we can bound the last term coordinatewise, using that each coordinate of $\mathbf{S}_{\lambda}^{-1/2}X$ is simply the projection of $\mathbf{S}_{\lambda}^{-1/2}X$ onto the corresponding coordinate vector, and proceeding via Cauchy-Schwarz:

$$\mathbb{E}\left[\left\|Y\mathbf{S}_{\lambda}^{-1/2}X\right\|^2\right] \leqslant \varkappa^2\kappa^2 v^2 \mathbb{E}\left[\left\|\mathbf{S}_{\lambda}^{-1/2}X\right\|^2\right] = \varkappa^2\kappa^2 v^2 \mathsf{df}_{\lambda}(\mathbf{S}).$$

Combining the previous steps, we obtain the claimed result. ■