Bootstrapping the Operator Norm in High Dimensions: Error Estimation for Covariance Matrices and Sketching

MILES E. LOPES∗, N. BENJAMIN ERICHSON† and MICHAEL W. MAHONEY†

Although the operator (spectral) norm is one of the most widely used metrics for covariance estimation, comparatively little is known about the fluctuations of error in this norm. To be specific, let \( \hat{\Sigma} \) denote the sample covariance matrix of \( n \) i.i.d. observations in \( \mathbb{R}^p \) that arise from a population matrix \( \Sigma \), and let \( T_n = \sqrt{n} \| \hat{\Sigma} - \Sigma \|_{\text{op}} \). In the setting where the eigenvalues of \( \Sigma \) have a decay profile of the form \( \lambda_j(\Sigma) \approx j^{-\beta} \), we analyze how well the bootstrap can approximate the distribution of \( T_n \). Our main result shows that up to factors of \( \log(n) \), the bootstrap can approximate the distribution of \( T_n \) with respect to the Kolmogorov metric at the rate of \( n^{-\beta/2} \), which does not depend on the ambient dimension \( p \). In addition, we offer a supporting result of independent interest that establishes a high-probability upper bound for \( T_n \) based on flexible moment assumptions. More generally, we discuss the consequences of our work beyond covariance matrices, and show how the bootstrap can be used to estimate the errors of sketching algorithms in randomized numerical linear algebra (RandNLA). An illustration of these ideas is also provided with a climate data example.


Keywords: bootstrap, error estimation, high-dimensional statistics, covariance estimation, randomized numerical linear algebra, sketching.

1. Introduction

Within the areas of covariance estimation and principal components analysis, it is of central importance to understand how well a sample covariance matrix \( \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \) approximates its population version \( \Sigma = \mathbb{E}[X_1 X_1^T] \), where \( X_1, \ldots, X_n \in \mathbb{R}^p \) are centered i.i.d. observations. In particular, a major line of research in high-dimensional statistics has focused on the problem of deriving non-asymptotic bounds for the operator (spectral) norm error

\[
T_n = \sqrt{n} \| \hat{\Sigma} - \Sigma \|_{\text{op}},
\]

where the norm is defined as \( \| A \|_{\text{op}} = \sup_{\|u\|_2 = 1} \| Au \|_2 \). A partial overview of work on this problem, as well as some of its extensions, may be found in the papers (Rudelson (1999), Bickel and Levina (2008), Cai, Zhang and Zhou (2010) Adamczak et al.

∗Supported in part by NSF grants DMS-1613218 and DMS-1915786
†Supported in part by ARO, DARPA (FA8750-17-2-0122), NSF, and ONR.

As a whole, this line of work offers many conceptual insights into the ways that error is influenced by model assumptions. However, the literature is less complete with regard to inference, and there are not many guarantees for the problem of constructing confidence intervals for $T_n$, which is equivalent to constructing numerical bounds on the error of $\hat{\Sigma}$, or confidence regions for $\Sigma$. Accordingly, the challenges of inference on high-dimensional covariance matrices have stimulated much recent activity, and there has been a particular interest to understand the limits of the bootstrap in this context (Johnstone and Paul (2018) §X.C, Han, Xu and Zhou (2018); El Karoui and Purdom (2019); Lopes, Blandino and Aue (2019); Naumov, Spokoiny and Ulyanov (2019)).

Simultaneously with these developments, the burgeoning field of randomized numerical linear algebra (RandNLA) has generated many other error estimation problems of a similar nature (Mahoney, 2011; Halko, Martinsson and Tropp, 2011; Woodruff, 2014; Kannan and Vempala, 2017; Drineas and Mahoney, 2018). A prototypical example deals with computing a fast randomized approximation of the product $A^T A$, where $A$ is a very large matrix. Most commonly, the matrix $A$ is randomly “sketched” into a much shorter matrix $\tilde{A}$, which can then be used to quickly compute $\tilde{A}^T \tilde{A}$ as an approximation to $A^T A$. In turn, it is necessary to assess the unknown error $\| \tilde{A}^T \tilde{A} - A^T A \|_{\text{op}}$, which leads to a notable parallel with the statistical literature: There are many existing theoretical error bounds, but very few tools for numerical error estimation (cf. Sections 1.2 and 4).

Furthermore, the operator norm is of special importance, because it governs the accuracy of numerous matrix computations, and it frequently appears in numerical analysis (Golub and Van Loan, 2013).

Motivated by the challenges above, this paper aims to quantify how well the bootstrap can approximate the error distribution $L(T_n)$ for sample covariance matrices, and likewise in the context of RandNLA. Specifically, we consider a setup where $\Sigma$ has low “effective rank” and its ordered eigenvalues satisfy a decay profile of the form

$$\lambda_j(\Sigma) \approx j^{-2\beta},$$

for some parameter $\beta > 1/2$. Variations of this setting have drawn considerable attention in recent years, especially in connection with principal components analysis (e.g., Lounici (2014), Bunea and Xiao (2015), Reiš and Wahl (2020), Koltchinskii and Lounici (2017a,b), Koltchinskii, Löffler and Nickl (2020), Naumov, Spokoiny and Ulyanov (2019), and Jung, Lee and Ahn (2018), among others). Moreover, the condition (1.1) corresponds to problems where sketching algorithms can be highly effective.

1.1. Contributions

To briefly outline our main result, let the Kolmogorov metric be denoted as $d_K(L(U), L(V)) = \sup_{t \in \mathbb{R}} |P(U \leq t) - P(V \leq t)|$ for two generic random variables $U$ and $V$, and let $T^*_n$ denote the bootstrap version of $T_n$, obtained by sampling with replacement from $(X_1, \ldots, X_n)$.
Then, as long as (1.1) is satisfied and the observations have suitable tail behavior, it follows that the bound

\[ d_K(\mathcal{L}(T_n), \mathcal{L}(T_n^*|X)) \leq cn^{-\frac{\beta-1/2}{6\beta+4}} \log(n)^c \]  

(1.2)

holds with probability at least \( 1 - \frac{c}{n} \), where \( \mathcal{L}(T_n^*|X) \) is the conditional distribution of \( T_n^* \) given the observations. (Going forward, we use symbols such as \( c, c_0, c_1 \), etc. to denote positive constants not depending on \( n \) whose values may change at each occurrence.) Most importantly, this non-asymptotic bound does not depend on the ambient dimension \( p \), and explicitly accounts for the structural complexity parameter \( \beta \).

From the standpoint of methodology, our work illustrates new possibilities for applying the bootstrap in the domains of computer science and applied mathematics. At this interface, the bootstrap has a largely untapped potential to make an impact, because error estimation allows randomized computations to be done adaptively, so that “just enough” work is done. More specifically, the estimated error of a rough initial solution can be used to predict how much extra computation is needed to reach a high-quality solution — and this will be demonstrated numerically in Section 4. Lastly, to put this type of application into historical perspective, it is notable that the bootstrap has been traditionally labeled as “computationally intensive”, and so in this respect, it is relatively novel to use the bootstrap in the service of computation.

With regard to theoretical considerations, our work contributes to recent developments on bootstrap methods, as well as covariance estimation. For the bootstrap, we expand upon the progress achieved in the series of papers (Chernozhukov, Chetverikov and Kato, 2013, 2014, 2016, 2017), which address bootstrap approximations for “max statistics” of the form \( M_n = \sup_{f \in \mathcal{F}} \mathbb{G}_n(f) \), where \( \mathcal{F} \) is a class of functions, and \( \mathbb{G}_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i)] \). The basic similarity between \( M_n \) and \( T_n \) is that they can be represented in a common form, due to the variational representation of \( \| \cdot \|_{\text{op}} \). Nevertheless, the statistic \( T_n \) seems to present certain technical obstructions with regard to previous results. First, in order to handle the metric \( d_K \), the mentioned works typically require a “minimum variance condition” such as

\[ \inf_{f \in \mathcal{F}} \text{var}(\mathbb{G}_n(f)) \geq c, \]  

(1.3)

which poses a difficulty in our setting, because the minimum variance may decrease rapidly with \( n \). As a result, a challenge arises in showing that our statistic is well approximated (in \( d_K \)) by the supremum \( \sup_{f \in \mathcal{F}_n} \mathbb{G}_n(f) \), where \( \mathcal{F}_n \subset \mathcal{F} \) is a “nice” subset for which \( \inf_{f \in \mathcal{F}_n} \text{var}(\mathbb{G}_n(f)) \) decreases slowly with \( n \). Second, further challenges are encountered when controlling the discretization error that comes from replacing \( \mathcal{F} \) with a discrete \( \epsilon \)-net. More specifically, this error is significant in our analysis because the relevant class \( \mathcal{F} \) is exponentially larger than VC-type — in the sense that \( \epsilon \)-covering numbers grow exponentially in \( 1/\epsilon \), rather than polynomially. By contrast, previous applications of bootstrap approximation results for max statistics have often been concerned with VC-type function classes, which allow for strong control of the discretization error.

Another segment of our work deals with tail bounds for \( \| \tilde{\Sigma} - \Sigma \|_{\text{op}} \) that do not depend on the ambient dimension \( p \), as studied in (Rudelson and Vershynin, 2007; Oliveira, 2010;
Hsu, Kakade and Zhang, 2012; Koltchinskii and Lounici, 2017a; Minsker, 2017). In the setting of (1.1), this line of work shows that if the observations satisfy $\| X_i \|_2 \leq c$ almost surely, or $\| \langle u, X_i \rangle \|_{\psi_2} \lesssim \| \langle u, X_i \rangle \|_2$ for all unit vectors $u$, then the operator norm error can be bounded as $\| \Sigma - \Sigma \|_{op} \leq c n^{-1/2} \log(n)^e$ with high probability. However, the $\ell_2$-boundedness condition is often restrictive, while the $\psi_2-L_2$ equivalence condition is not well-suited to the discrete distributions that arise from resampling (Vershynin, 2018, §3.4.2). Consequently, as a way to streamline our analysis of both $(X_1, \ldots, X_n)$ and the bootstrap samples $(\hat{X}_1, \ldots, \hat{X}_n)$, it is of interest to develop a bound that can be applied in a more general-purpose way. Indeed, an extension of this type is also suggested briefly in the paper (Rudelson and Vershynin, 2007), but to the best of our knowledge, such a result has not been available without involving dependence on the ambient dimension. Accordingly, one of our secondary main results (Theorem 2.2) serves this purpose by showing that if $q \geq 3$, and if $\xi_1, \ldots, \xi_n$ are i.i.d. random elements of a separable Hilbert space with norm $\| \cdot \|$, then

$$
\left( E \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \otimes \xi_i - E[\xi_1 \otimes \xi_1] \right\|^q_{op} \right)^{1/q} \leq c \cdot \| E[\xi_1 \otimes \xi_1] \|_{op} \cdot \left( \sqrt{\frac{r(q)}{n^{1-\eta}}} + \frac{r(q)}{n^{1-\eta}} \right).
$$

where $r(q)$ is a parameter that plays the role of an effective rank, and satisfies $r(q) < \infty$ as long as $\| \xi_1 \|$ has at least $2q$ moments. In particular, the proof of Theorem 2.2 extends the approach of Rudelson and Vershynin (2007) based on non-commutative Khintchine inequalities. For comparison, the paper (Chen, Gittens and Tropp, 2012) has also extended the approach of Rudelson and Vershynin (2007) to allow for sums of more general types of random matrices, but it differs from our work insofar as the results depend on the ambient dimension.

### 1.2. Related work

The most closely related work to ours is the recent paper (Han, Xu and Zhou, 2018), which studies bootstrap approximations for certain variants of $T_n$. To explain the connection, first recall that $T_n$ may be written in terms of a supremum over the unit sphere $S^{p-1} \subset \mathbb{R}^p$, namely $T_n = \sup_{u \in S^{p-1}} \sqrt{n} \| u \| (\hat{\Sigma} - \Sigma) u$. As an alternative to this, the paper (Han, Xu and Zhou, 2018) analyzes “sparse versions” of $T_n$ obtained by taking the supremum over $\{ u \in S^{p-1} : \| u \|_0 \leq s \}$, where $1 \leq s \leq p$. For these sparse versions of $T_n$, bootstrap approximation results are obtained in the Kolmogorov metric with rates of the form $s^{9/8}/n^{1/8}$, up to logarithmic factors. As this relates to our work, it should be emphasized that the setting in (Han, Xu and Zhou, 2018) is quite different, since the eigenvalues of $\Sigma$ are not assumed to decay. The difference becomes most apparent when $s = p$, so that the sphere $S^{p-1}$ coincides with the set $\{ u \in S^{p-1} : \| u \|_0 \leq s \}$. In this case, the analysis without spectral decay requires $p \ll n^{1/9}$ for bootstrap consistency, whereas our setting places no constraints on $p$. Another work that looks at bootstrapping a structured variant of the operator norm is (Silin and Fan, 2020), but for the different purpose of analyzing the errors of empirical spectral projection matrices. Although this work appeared after
our 2019 preprint (Lopes, Erichson and Mahoney, 2019), it is still worth noting that it shares the approach of using a variational representation of the norm and applying the techniques for bootstrapping suprema of empirical processes mentioned earlier.

Next, the recent paper (El Karoui and Purdom, 2019) looks at both positive and negative results for bootstrapping sample eigenvalues. For the positive results, this work assumes that $\Sigma$ is nearly low-rank, and that the dimension satisfies $p \leq n$. The main result shows that the bootstrap consistently approximates the joint distribution of $L_n = \sqrt{n}(\lambda_j(\bar{\Sigma}) - \lambda_j(\Sigma))_{1 \leq j \leq n}$, where $j_0$ is held fixed as $(n,p) \to \infty$, provided that the eigenvalues $(\lambda_j(\Sigma))_{1 \leq j \leq n}$ each have multiplicity 1. This illustrates a key difference between the statistics $\bar{T}_n$ and $L_n$, since it is known that repeated eigenvalues can interfere with bootstrap approximations for $L_n$ (e.g. Hall et al., 2009), whereas our work will show that the bootstrap can work for $\bar{T}_n$ even in the presence of repeated eigenvalues. Two more papers on bootstrap methods for high-dimensional sample covariance matrices are (Lopes, Blandino and Aue, 2019) and (Naumov, Spokoiny and Ulyanov, 2019). The first of these generalizes the parametric bootstrap for high-dimensional models without spectral decay, and it establishes consistency for linear spectral statistics, while the latter deals with bootstrapping the error of empirical spectral projection matrices. (We refer to (Koltchinskii and Lounici, 2017b,c) for other related distributional approximation results.)

At a more technical level, our work is related to the paper (Lopes, Lin and Müller, 2020), which analyzes rates of bootstrap approximation for max statistics of the form $M_n = \max_{1 \leq j \leq p} \sqrt{n} \bar{X}_j$, where the vector $\bar{X}$ is the sample average of centered i.i.d. observations $X_1, \ldots, X_n \in \mathbb{R}^p$ satisfying a “variance decay” condition. This condition has the form $\sigma^2_{(s)} \leq j^{-2\alpha}$ for some fixed parameter $\alpha > 0$, where $\sigma^2_{(1)} \geq \cdots \geq \sigma^2_{(p)}$ are the sorted versions of $\sigma^2 = \text{var}(X_{1j})$ for $j = 1, \ldots, p$. One of the key steps in the analysis of (Lopes, Lin and Müller, 2020) is to “localize the maximizing index” for $M_n$. That is, if $j \in \{ 1, \ldots, p \}$ is a random index such that $M_n = \sqrt{n} \bar{X}_j$, then the variance decay condition can be used to show that $j$ is likely to fall into a small subset of $\{1, \ldots, p\}$. In the present context, we also use this localization technique. Namely, the eigenvalue decay condition $\lambda_j(\Sigma) \leq j^{-2\alpha}$ implies that if $\bar{u} \in \mathbb{S}^{p-1}$ denotes a maximizing index for $\bar{T}_n = \sup_{u \in \mathbb{S}^{p-1}} \sqrt{n} |u^T (\bar{\Sigma} - \Sigma) u|$, then $\Sigma^{1/2} \bar{u}$ is likely to fall into a suitable subset of an ellipsoid. In carrying out this localization argument, our Proposition A.1 and its supporting Lemmas A.1 and A.2 in this work are parallel versions of Proposition B.2 and Lemmas B.1 and B.2 from (Lopes, Lin and Müller, 2020). (The localization of the maximizer for the bootstrapped statistic $T_n^*$ is similarly handled here in Appendix E.) Apart from this connection, the analyses in the two papers are essentially different. In particular, it should be emphasized that the high probability bound on $|\bar{\Sigma} - \Sigma|_{op}$ in Theorem 2.2, the discretization error bound in Proposition 2.1, as well as the intermediate results in Appendices B, C, D, and F are distinct from the earlier paper. Furthermore, the overall qualitative difference between the analyses is illustrated by the fact that in (Lopes, Lin and Müller, 2020), the rate of bootstrap approximation does not depend on the variance decay parameter $\alpha$, whereas here, the rate does depend on the spectrum decay parameter $\beta$.

Finally, to conclude this section, we describe related work on the estimation of algorithmic error. Here, it is important to note that error estimation has a long history for
**deterministic** algorithms, such as those in numerical partial differential equations and finite-element methods, where it is called a *posteriori error estimation* (Babuška and Rheinboldt, 1978; Verfürth, 1994; Becker and Rannacher, 2001; Jiránek, Strakoš and Vohralík, 2010; Ainsworth and Oden, 2011; Cangiani et al., 2017, among many others). However, in the literature on randomized algorithms, error estimation has received much less attention, and for certain types of computations there are only a few papers addressing error estimation: (low-rank approximation: Liberty et al., 2007; Woolfe et al., 2008; Halko, Martinsson and Tropp, 2011), (least-squares: Lopes, Wang and Mahoney, 2018), (classification: Lopes, 2019), (matrix multiplication: Ar et al., 1993; Sarlós, 2006; Lopes, Wang and Mahoney, 2019). Among these works, the only ones to address error estimation for the operator norm are (Liberty et al., 2007; Woolfe et al., 2008; Halko, Martinsson and Tropp, 2011), but this is done specifically for low-rank approximation, which is complementary to our applications. Also, the approach in these works is quite different from bootstrapping, and is based on the idea of bounding error in terms of random “test vectors”, which is rooted in the classical works (Freivalds, 1979; Dixon, 1983).

In essence, the main difference between the test-vector approach and bootstrapping is which is complementary to our applications. Also, the approach in these works is quite different from bootstrapping, and is based on the idea of bounding error in terms of random “test vectors”, which is rooted in the classical works (Freivalds, 1979; Dixon, 1983).

Lastly, the proof of the main result (Theorem 2.1) and some additional numerical results are deferred to the supplementary material.

**Outline.** Section 2 presents the problem setup and main results, as well as the proofs for some of these results. Section 3 describes numerical results for inference tasks related to covariance matrices. Section 4 introduces the setting of sketching algorithms, and demonstrates the performance of the bootstrap in synthetic problems, as well as in a climate data example. Lastly, the proof of the main result (Theorem 2.1) and some additional numerical results are deferred to the supplementary material.

**Notation and conventions.** For a vector $v \in \mathbb{R}^m$, and a number $q \geq 1$, the $\ell_q$-norm is $|v|_q = (\sum_{j=1}^m |v_j|^q)^{1/q}$. For a real matrix $M$, its Frobenius norm is $\|M\|_F = \sqrt{\text{tr}(M^T M)}$, and its Schatten-$q$ norm is $\|M\|_{S_q} = \text{tr}((M^T M)^{q/2})^{1/q}$. The identity matrix of size $m \times m$ is $I_m$, and the standard basis vectors in $\mathbb{R}^m$ are $\{e_1, \ldots, e_m\}$. The sorted singular values of a real matrix $M$ are written as $\sigma_j(M) \geq \sigma_{j+1}(M)$, and similarly, if $M$ is symmetric, then the sorted eigenvalues are written as $\lambda_j(M) \geq \lambda_{j+1}(M)$. For a random variable $\xi$, the $L_q$ norm is $\|\xi\|_q = (\mathbb{E}[|\xi|^q])^{1/q}$. Also, if $\psi_q(x) = \exp(x^q) - 1$, then the $\psi_q$-Orlicz norm is given by $\|\xi\|_{\psi_q} = \inf\{r > 0 : \mathbb{E}[\psi_{q}(|\xi|/r)] \leq 1\}$. If $\zeta$ is another random variable, then the conditional distribution of $\zeta$ given $\xi$ is denoted as $\mathcal{E}(\zeta|\xi)$. If $a_n$ and $b_n$ are sequences of non-negative real numbers, we write $a_n \preceq b_n$ if there is a constant $c > 0$ not depending on $n$, and integer $n_0 \geq 1$ such that $a_n \leq c b_n$ for all $n \geq n_0$. Likewise, we write $a_n \asymp b_n$ if $a_n \preceq b_n$ and $b_n \preceq a_n$. Lastly, for maxima and minima, we use the notation $a_n \vee b_n = \max\{a_n, b_n\}$ and $a_n \wedge b_n = \min\{a_n, b_n\}$. 
2. Main results

Our setup is based on a sequence of models indexed by \( n \), where all parameters may depend on \( n \), unless stated otherwise. In particular, the dimensions \( p = p(n) \) and \( d = d(n) \) below may vary with \( n \). If a parameter does not depend on \( n \), then it is understood not to depend on \( p \) or \( d \) either.

**Assumption 2.1** (Data-generating model).

(i). There is a deterministic matrix \( A \in \mathbb{R}^{d \times p} \) with \( d \geq p \), and i.i.d. random vectors \( Z_1, \ldots, Z_n \in \mathbb{R}^d \), such that for each \( i \in \{1, \ldots, n\} \), the observation \( X_i \in \mathbb{R}^p \) is generated as

\[
X_i = A^T Z_i. \tag{2.1}
\]

(ii). The random vector \( Z_1 \) has independent entries that satisfy \( E[Z_{1j}] = 0 \), \( E[Z_{1j}^2] = 1 \), and \( \kappa := E[Z_{1j}^4] > 1 \) for all \( j \in \{1, \ldots, d\} \), where \( \kappa \) does not depend on \( n \). In addition, the condition \( \max_{1 \leq j \leq d} \|Z_1\|_{\psi_2} \leq 1 \) holds.

(iii). There is a constant \( \beta > 1/2 \) not depending on \( n \), such that for each \( j \in \{1, \ldots, p\} \), the singular value \( \sigma_j(A) \) satisfies

\[
\sigma_j(A) \asymp j^{-\beta}. \tag{2.2}
\]

**Remarks.** In statistical applications, the matrix \( A \) is typically taken to be the square root \( \sqrt{\Sigma} \), with \( p = d \). However, the extra generality of a rectangular matrix is needed for the application of our work to sketching algorithms in Section 4. To comment on two other aspects of Assumption 2.1, observe that it places no constraints on the relationship between \( n \) and \( p \), and it allows for many eigenvalues of \( \Sigma \) to be repeated.

In order to state our main result, we need to precisely define the statistic \( T_n^* \) that arises from bootstrap sampling. Let \( (X_1^*, \ldots, X_n^*) \) be drawn with replacement from \( (X_1, \ldots, X_n) \), and define the matrix

\[
\hat{\Sigma}^* = \frac{1}{n} \sum_{i=1}^n X_i^*(X_i^*)^T.
\]

Then, the bootstrapped counterpart of \( T_n \) is defined as

\[
T_n^* = \sqrt{n} \|\hat{\Sigma}^* - \hat{\Sigma}\|_{\text{op}}.
\]

The following is our main result.

**Theorem 2.1.** Suppose that Assumption 2.1 holds. Then, there is a constant \( c > 0 \) not depending on \( n \) such that the event

\[
d_K(L(T_n), L(T_n^*|X)) \leq cn^{-\frac{\beta-1/2}{\beta+1/2}} \log(n)^c \tag{2.2}
\]

occurs with probability at least \( 1 - \frac{c}{n} \).
Remarks. A high-level outline of the proof will be given in Section 2.3. To explain how the difference $\beta - 1/2$ arises in the rate of bootstrap approximation, we offer some informal discussion. As preparatory notation, define the ellipsoidal boundary set $E = \{Au | u \in S^{p-1}\}$, as well as its signed version $\Theta = E \times \{\pm 1\}$, whose generic element is denoted by $\theta = (v, s)$. With these items in place, we will consider the following empirical process indexed by $\Theta$,

$$
G_n(\theta) = \sqrt{\frac{n}{\sum_{i=1}^{n} (v, Z_i)^2 - \mathbb{E}[(v, Z_i)^2]},
$$

which allows $T_n$ to be represented as

$$
T_n = \sup_{\theta \in \Theta} G_n(\theta). \tag{2.4}
$$

Given that the set $\Theta$ is uncountable, a standard reduction is to approximate $T_n$ with the supremum of $G_n$ over a discrete $\epsilon$-net for $\Theta$, where the metric is taken to be $\rho(\theta, \tilde{\theta}) = \|v - \tilde{v}\|_2 + |s - \tilde{s}|$. In turn, this requires us to control the discretization error, which leads to bounding the supremum of increments, denoted

$$
\Delta_n(\epsilon) = \sup_{\rho(\theta, \tilde{\theta}) \leq \epsilon} |G_n(\theta) - G_n(\tilde{\theta})|.
$$

In order for the discrete approximation to succeed, the quantity $\mathbb{E}[\Delta_n(\epsilon)]$ should vanish as $\epsilon \to 0$. However, the demonstration of this property depends on the complexity of $\Theta$ through the parameter $\beta$.

We can gain some intuition for the role of $\beta$ by looking at how it affects $\mathbb{E}[\Delta_n(\epsilon)]$ in a much simpler case — where $G_n$ is replaced by a linear Gaussian process indexed by $\Theta$. Namely, consider the process $\tilde{G_n}(\theta) = n^{-1/2} \sum_{i=1}^{n} s(v, \zeta_i)$, where $\zeta_1, \ldots, \zeta_n$ are independent standard Gaussian vectors. In this case, if $\tilde{\Delta}_n(\epsilon)$ denotes the analogue of $\Delta_n(\epsilon)$ for $\tilde{G_n}$ and $p \geq \epsilon^{-1/\beta}$, then the following lower bound can be shown using classical facts about Gaussian processes,

$$
\mathbb{E}[\tilde{\Delta}_n(\epsilon)] \geq \epsilon^{(\beta - 1/2)/\beta}. \tag{2.5}
$$

(See Lemma F.3 in Appendix F.) Thus, the main point to take away here is that even in the simple case of a linear Gaussian process, the condition $\beta > 1/2$ is necessary in high dimensions for the discretization error to vanish as $\epsilon \to 0$.

Another benefit of looking at the linear Gaussian case is that the lower bound (2.5) provides a reference point for assessing our upper bound on the discretization error. For instance, it will follow from Proposition 2.1 that

$$
\mathbb{E}[\Delta_n(\epsilon)] \leq \epsilon^{(\beta - 1/2)/\beta} \log(n), \tag{2.6}
$$

as shown in Section 2.2. Hence, it is notable that the dependence on $\epsilon$ does not change in comparison to the linear Gaussian case, even though the quadratic nature of the process $G_n$ causes it not to be sub-Gaussian with respect to the metric $\rho$. Moreover, it also turns
out that the dependence on $\epsilon$ even remains the same for $L_q$ norms of $\Delta_n(\epsilon)$ when $q$ is large.

One more point of theoretical interest is that the bound (2.6) arises in a situation where standard chaining seems to give a slower dependence on $\epsilon$ than a more problem-specific approach. As an example of a standard approach, one might try to show that $G_n$ is sub-exponential with respect to $\rho$, and then appeal to an entropy integral bound such as in (van der Vaart and Wellner, 2000, Theorem 2.2.4). However, this ultimately leads to an upper bound scaling like $\epsilon^{(\beta-1)/\beta}$, which would require the excessive condition $\beta > 1$ (as opposed to $\beta > 1/2$). Likewise, the development of new techniques for quadratic processes akin to $G_n$ has attracted interest in the literature, as surveyed in (Talagrand, 2014, §9.3-9.4). Nevertheless, it should also be noted that existing results in this direction do not seem to be directly applicable to our analysis of the bootstrap. For instance, the abstract approaches based on Talagrand’s $\gamma_1$ and $\gamma_2$ functionals lead to challenges in connection with the bootstrap, because the discrete process $G_n^*$ (arising from sampling with replacement) induces a random metric on $\Theta$ that does not lend itself to calculations. On the other hand, the approach taken here allows $G_n$ and $G_n^*$ to be treated on nearly equal footing (cf. Proposition D.1).

2.1. A general-purpose bound for sample covariance matrices

Below, we provide a general-purpose high-probability bound for the operator norm error of sample covariance matrices. Notably, the result only requires control on the moments of the norm of a random vector. Since the result may be of independent interest in other problems, we have stated it in the general context of a separable Hilbert space $H$, with inner product and norm denoted as $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$. In addition, if $x, y, z \in H$, then $x \otimes y$ denotes the linear operator from $H$ to $H$ satisfying $(x \otimes y)z = \langle y, z \rangle x$.

**Theorem 2.2.** Let $q \geq 3$, and let $\xi_1, \ldots, \xi_n \in H$ be i.i.d. random elements. Also, define the quantity

$$r(q) = q \left( \frac{\mathbb{E}[\|\xi_1\|^{2q}]}{\|\mathbb{E}[\xi_1 \otimes \xi_1]\|_{\text{op}}} \right)^{\frac{1}{q}}.$$  

(2.7)

Then, there is an absolute constant $c > 0$ such that

$$\left( \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \xi_i \otimes \xi_i - \mathbb{E}[\xi_i \otimes \xi_i] \right]^{q} \right)^{1/q} \leq c \left( \mathbb{E}[\xi_1 \otimes \xi_1] \right)_{\text{op}} \left( \sqrt{\frac{r(q)}{n^{1-3/q}}} \vee \frac{r(q)}{n^{1-3/q}} \right).$$  

(2.8)

**Remarks.** It will sometimes be useful to consider the special case where the random variable $\|\xi_1\|$ can be described in terms of its $\psi_2$ Orlicz norm. This gives

$$r(q) \leq c q^2 \frac{\|\xi_1\|_{\psi_2}^2}{\|\mathbb{E}[\xi_1 \otimes \xi_1]\|_{\text{op}}},$$  

(2.9)
for an absolute constant $c > 0$, which can be obtained from the facts about Orlicz norms summarized in Lemmas G.1 and G.2. Before proceeding directly to the proof of Theorem 2.2, we need a preparatory lemma, which is a slightly relaxed version of a result from (Rudelson, 1999, p.63). As a matter of notation, the Schatten-$q$ norm of an operator $M$ will be denoted as $\|M\|_{S_q}$.

Lemma 2.1. Let $x_1, \ldots, x_n \in \mathcal{H}$ be fixed, and let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables. Then, there is an absolute constant $c > 0$ such that for any $q \geq 2$,

$$\left( \mathbb{E} \left\| \sum_{i=1}^{n} \varepsilon_i x_i \otimes x_i \right\|_{S_q}^{q} \right)^{1/q} \leq c \cdot n^{1/q} \cdot \sqrt{q} \cdot \left( \max_{1 \leq i \leq n} \|x_i\| \right) \cdot \left\| \sum_{i=1}^{n} x_i \otimes x_i \right\|_{op}^{1/2}. \quad (2.10)$$

Proof. First, we make use of a non-commutative Khinchine inequality originating from (Lust-Piquard, 1986). Specifically, we use the version from (Pisier, 2016, Theorem 14.6), which implies

$$\left( \mathbb{E} \left\| \sum_{i=1}^{n} \varepsilon_i x_i \otimes x_i \right\|_{S_q}^{q} \right)^{1/q} \leq c \sqrt{q} \left( \sum_{i=1}^{n} \|x_i\|^{2} \right)^{1/2}. \quad \text{(2.11)}$$

Next, observe that any operator $M$ with rank at most $r$ satisfies $\|M\|_{S_q} \leq r^{1/q} \|M\|_{op}$, and that the operator $\sum_{i=1}^{n} \|x_i\|^{2} x_i \otimes x_i$ has rank at most $n$. Combining this with the previous bound leads to the stated result. \hfill \Box

Proof of Theorem 2.2. The proof extends the approach developed in (Rudelson and Vershynin, 2007) to the case of unbounded random vectors. Using a standard symmetrization argument, we have

$$\left( \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \otimes \xi_i - \mathbb{E}[\xi_i \otimes \xi_i] \right\|_{op}^{q} \right)^{1/q} \leq c \left( \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \xi_i \otimes \xi_i \right\|_{op}^{q} \right)^{1/q}, \quad (2.11)$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher variables that are also independent of $\xi_1, \ldots, \xi_n$, and $c > 0$ is an absolute constant. Next, since $\| \cdot \|_{op} \leq \| \cdot \|_{S_q}$, it follows from Lemma 2.1 and the Cauchy-Schwarz inequality that

$$\left( \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \xi_i \otimes \xi_i \right\|_{op}^{q} \right)^{1/q} \leq c \cdot n^{1/q} \cdot \sqrt{\mathbb{E} \left[ \max_{1 \leq i \leq n} \|\xi_i\|^{2q} \right]} \cdot \left( \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \otimes \xi_i \right\|_{op}^{q} \right)^{1/q}. \quad (2.12)$$

To bound the $L_{2q}$ norm of $\max_{1 \leq i \leq n} \|\xi_i\|$, we have

$$\mathbb{E} \left[ \max_{1 \leq i \leq n} \|\xi_i\|^{2q} \right]^{\frac{1}{2q}} \leq n^{\frac{1}{2q}} \left( \mathbb{E} \left[ \|\xi_1\|^{2q} \right] \right)^{\frac{1}{2q}}.$$

Also, for the last factor in the bound (2.12), we have

$$\left( \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \otimes \xi_i \right\|_{op}^{q} \right)^{1/q} \leq \left( \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \otimes \xi_i - \mathbb{E}[\xi_i \otimes \xi_i] \right\|_{op}^{q} \right)^{1/q} + \left\| \mathbb{E}[\xi_1 \otimes \xi_1] \right\|_{op}.$$
Proposition 2.1. Let \( q = \log(n) \vee 3 \), let \( \epsilon = \epsilon(n) \in (0, 1) \), and suppose that Assumption 2.1 holds. Then,

\[
\left\| \sup_{\rho(v, \theta) \leq \epsilon} |G_n(\theta) - G_n(\hat{\theta})| \right\|_q \leq \epsilon^{1 - \frac{1}{\bar{q}}} \log(n). \tag{2.13}
\]

Proof. As an initial observation, note that the condition \( \rho(\theta, \hat{\theta}) \leq \epsilon < 1 \) implies that the signs \( s \) and \( \hat{s} \) must be equal. This leads to the algebraic identity

\[
|G_n(\theta) - G_n(\hat{\theta})| = \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} (v + \hat{v}, Z_i)(v - \hat{v}, Z_i) - \mathbb{E}[\{(v + \hat{v}, Z_i)(v - \hat{v}, Z_i)\}] \right|.
\]

To rewrite the quadratic forms in terms of a symmetric matrix, let

\[
Q = \frac{1}{2} \left( (v + \hat{v})(v - \hat{v})^\top + (v - \hat{v})(v + \hat{v})^\top \right),
\]

so that

\[
|G_n(\theta) - G_n(\hat{\theta})| = \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} Z_i^\top Q Z_i - \mathbb{E}[Z_i^\top Q Z_i] \right|.
\]

Next, let \( t > 0 \) denote a free parameter to be chosen later, and define the vectors

\[
\omega(t) = \frac{1}{2}t(v + \hat{v}) + \frac{1}{t}(v - \hat{v}) \quad \text{and} \quad \bar{\omega}(t) = \frac{1}{2}t(v + \hat{v}) - \frac{1}{t}(v - \hat{v}).
\]

In turn, it can be checked that these vectors give the following representation of \( Q \),

\[
Q = \omega(t)\omega(t)^\top - \bar{\omega}(t)\bar{\omega}(t)^\top, \tag{2.14}
\]
which has a certain invariance property, insofar as it holds for every $t > 0$, while $Q$ itself does not depend on $t$. The utility of this representation is that it will allow us to work with sums of squares, and also, to optimize with respect to the choice of $t$.

To proceed, we will define a particular ellipsoid that contains the vectors $\omega(t)$ and $\tilde{\omega}(t)$, and then take a supremum over this ellipsoid to derive a stochastic upper bound on $\sup_{\rho \leq \theta} \left| G_n(\theta) - G_n(\tilde{\theta}) \right|$. For this purpose let $A(\epsilon) \in \mathbb{R}^{d \times p}$ be the matrix with the same s.v.d. as $A$, except that the singular value $\sigma_j(A)$ is replaced with $\sqrt{2} \min\{\sigma_j(A), \epsilon/2\}$ for every $j \in \{1, \ldots, p\}$. Also, define $A(t, \epsilon) \in \mathbb{R}^{d \times 2p}$ as the column concatenation

$$A(t, \epsilon) = \left[ tA, \frac{1}{t} A(\epsilon) \right].$$

With this matrix in hand, it can be shown that both vectors $\omega(t)$ and $\tilde{\omega}(t)$ lie in the ellipsoid $A(t, \epsilon)(\mathbb{B}_{2p}(2))$, where $\mathbb{B}_{2p}(2)$ denotes the $\ell_2$-ball of radius 2 in $\mathbb{R}^{2p}$. (For the details, see Lemma F.2.) In particular, this ellipsoid does not depend on the indices $\theta$ and $\tilde{\theta}$ underlying $\omega(t)$ and $\tilde{\omega}(t)$. As a result, we have

$$\sup_{\rho \leq \theta} \left| G_n(\theta) - G_n(\tilde{\theta}) \right| \leq \sup_{w \in \mathbb{B}_{2p}(2)} \frac{2}{\sqrt{n}} \\left| \sum_{i=1}^{\infty} (Z_i, A(t, \epsilon)w)^2 - \mathbb{E}[(Z_i, A(t, \epsilon)w)^2] \right|. \tag{2.15}$$

We now apply Theorem 2.2 with $\xi_i = (A(t, \epsilon)^T Z_i)_i$, which gives

$$\left\| \sup_{\rho \leq \theta} \left| G_n(\theta) - G_n(\tilde{\theta}) \right| \right\|_q \leq \sqrt{n} \cdot \| A(t, \epsilon) \|_{2,op}^2 \cdot \left( \frac{\sqrt{\varphi(q)}}{n^{1-3/q}} \vee \frac{\varphi(q)}{n^{1-3/q}} \right). \tag{2.15}$$

Due to the choice $q = \log(n) \vee 3$, we have $n^{1-3/q} \geq n$, and also, the bound (2.9) implies

$$\varphi(q) \leq q^2 \frac{\| A(t, \epsilon)^T Z_1 \|_2^2}{\| A(t, \epsilon) \|_{2,op}^2}.$$

Furthermore, Lemma G.2 gives

$$\left\| A(t, \epsilon)^T Z_1 \right\|_2 \leq \left\| A(t, \epsilon) \right\|_F, \tag{2.16}$$

and then combining with (2.15) leads to

$$\left\| \sup_{\rho \leq \theta} \left| G_n(\theta) - G_n(\tilde{\theta}) \right| \right\|_q \leq \left( q \cdot \| A(t, \epsilon) \|_{2,op} \cdot \| A(t, \epsilon) \|_F \right) \vee \left( \frac{q^2}{n} \| A(t, \epsilon) \|_{2,op}^2 \right).$$

Hence, to complete the proof, it remains to bound the norms of $A(t, \epsilon)$ and then specify a value of $t$. From the definition of $A(t, \epsilon)$ and a short calculation, we have

$$\| A(t, \epsilon) \|_F \leq \| tA \|_F + \frac{1}{t} \| A(\epsilon) \|_F \leq t + \frac{1}{t} \epsilon \sqrt{\frac{1}{4}}, \tag{2.17}$$

as well as

$$\| A(t, \epsilon) \|_{2,op} \leq \| tA \|_{2,op} + \frac{1}{t} \| A(\epsilon) \|_{2,op} \leq t + \frac{1}{t} \epsilon. \tag{2.18}$$

Taking $t = \epsilon^{\frac{1}{2} - \frac{1}{24}}$ leads to the stated result. \qed
2.3. Outline for the proof of Theorem 2.1

In this subsection, we define several objects that will recur in our arguments, and then explain how the main components of the proof fit together.

Bootstrap and Gaussian processes. Let \((Z_1^*, \ldots, Z_n^*)\) be sampled with replacement from \((Z_1, \ldots, Z_n)\), and define the bootstrap counterpart of \(G_n\) as

\[
G_n^*(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (v, Z_i^*)^2 - \mathbb{E}[(v, Z_i^*)^2 | X],
\]

where \(\mathbb{E}[(v, Z_i^*)^2 | X]\) refers to expectation that is conditional on \(X_1, \ldots, X_n\), and we note that \(\mathbb{E}[(v, Z_i^*)^2 | X] = \frac{1}{n} \sum_{i=1}^{n} (v, Z_i)^2\). This definition of \(G_n^*\) allows \(T_n^*\) to be expressed as

\[
T_n^* = \sup_{\theta \in \Theta} G_n^*(\theta).
\]

In addition, we define \(G_n\) as the centered Gaussian process on \(\Theta\) whose covariance structure matches that of \(G_n\),

\[
\text{cov}(G_n(\theta), G_n(\tilde{\theta})) = \text{cov}(G_n(\theta), G_n(\tilde{\theta})) \quad \text{for all} \quad \theta, \tilde{\theta} \in \Theta.
\]

Subsets of indices. In order to define some special subsets of \(E\) and \(\Theta\), let

\[
\ell_n = \left\lfloor (1 \lor \log(n)^3) \land p \right\rfloor \quad \text{and} \quad k_n = \left\lfloor (\ell_n \lor \log(n)^{\frac{6\beta+4}{3\beta+2}}) \land p \right\rfloor,
\]

which always satisfy \(1 \leq \ell_n \leq k_n \leq p\). Also, let the columns of \(V_{k_n} \in \mathbb{R}^{p \times k_n}\) contain the leading \(k_n\) right singular vectors of \(A\). Based on these items, we define \(E_n^\ell\) as a subset of \(E\) arising from vectors in \(S^{p-1}\) that are “partially aligned” with the columns of \(V_{k_n}\),

\[
E_n^\ell = \left\{ Au \mid u \in S^{p-1} \text{ and } \|V_{k_n}^T u\|_2 > \frac{1}{2} k_n^{-\beta+1/2} \right\}.
\]

Likewise, by analogy with the definition of \(\Theta\), let

\[
\Theta_n^\ell = E_n^\ell \times \{ \pm 1 \}.
\]

The next piece of notation is an \(\epsilon\)-net for \(\Theta_n^\ell\) with respect to the metric \(\rho\). This net is denoted as \(\Theta_n^\ell(\epsilon) \subset \Theta_n^\ell\) and has the defining property that for any \(\theta \in \Theta_n^\ell\), there is at least one point \(\theta' \in \Theta_n^\ell(\epsilon)\) with \(\rho(\theta, \theta') \leq \epsilon\). Throughout the proofs, we will mostly use the particular choice \(\epsilon = \epsilon_n\) with

\[
\epsilon_n = n^{-\beta/(6\beta+4)}.
\]

(2.19)

Lastly, due to classical bounds on the metric entropy of ellipsoids (as recorded in Lemma G.3), it is possible to choose an \(\epsilon_n\)-net for \(\Theta_n^\ell(\epsilon_n)\) with respect to \(\rho\) so that its cardinality satisfies \(\text{log card}(\Theta_n^\ell(\epsilon_n)) \leq \epsilon_n^{-1/\beta}\).
Decomposition into six main terms. We will bound the Kolmogorov distance between $\mathcal{L}(T_n)$ and $\mathcal{L}(T^*_n|X)$ with six terms,

$$d_K(\mathcal{L}(T_n), \mathcal{L}(T^*_n|X)) \leq I_n + II_n + III_n + \tilde{III}_n + \tilde{II}_n + \tilde{I}_n,$$

which are defined below. The essential novelty of the proof deals with the four terms ($I_n, II_n, \tilde{II}_n, \tilde{I}_n$), and almost all of the effort will be focused on these.

1. Localizing the maximizer of $G_n$:

$$I_n = d_K\left( \mathcal{L}\left( \sup_{\theta \in \Theta} G_n(\theta) \right), \mathcal{L}\left( \sup_{\theta \in \Theta} G_n(\theta) \right) \right).$$

(We use the phrase “localizing the maximizer of $G_n$”, because the problem of showing that $I_n$ is small amounts to showing that the maximizing index for $G_n$ is likely to fall in $\Theta_n^*$.)

2. Discrete approximation of $G_n$:

$$II_n = d_K\left( \mathcal{L}\left( \sup_{\theta \in \Theta} G_n^*(\theta) \right), \mathcal{L}\left( \sup_{\theta \in \Theta} G_n^*(\theta) \right) \right).$$

3. Gaussian approximation:

$$III_n = d_K\left( \mathcal{L}\left( \sup_{\theta \in \Theta} G_n^*(\theta) \right), \mathcal{L}\left( \sup_{\theta \in \Theta} G_n^*(\theta) \right) \right).$$

4. Bootstrap approximation:

$$\tilde{III}_n = d_K\left( \mathcal{L}\left( \sup_{\theta \in \Theta} G_n^*(\theta) \right), \mathcal{L}\left( \sup_{\theta \in \Theta} G_n^*(\theta) \right) \right).$$

5. Discrete approximation of $G^*_n$:

$$\tilde{II}_n = d_K\left( \mathcal{L}\left( \sup_{\theta \in \Theta} G_n^*(\theta) \right), \mathcal{L}\left( \sup_{\theta \in \Theta} G_n^*(\theta) \right) \right).$$

6. Localizing the maximizer of $G^*_n$:

$$\tilde{I}_n = d_K\left( \mathcal{L}\left( \sup_{\theta \in \Theta} G_n^*(\theta) \right), \mathcal{L}\left( \sup_{\theta \in \Theta} G_n^*(\theta) \right) \right).$$

Altogether, the six terms are handled consecutively in Appendices A through E, with each appendix corresponding to a different term (except for $III_n$ and $\tilde{III}_n$, which are handled together).

To comment on some of the techniques used in the various stages of the proof, the bound on $I_n$ builds on a lower-tail bound for Gaussian maxima developed in (Lopes, Lin and Müller, 2020) and requires fine-grained control on the covariances $\text{cov}(G_n(\theta), G_n(\theta))$
Bootstrapping the Operator Norm

for certain choices of \( \theta \) and \( \tilde{\theta} \). Next, the analysis of \( \Pi_n \) uses both the result on sample covariance matrices in Theorem 2.2 and the increment bound in Proposition 2.1. With regard to the terms \( \Omega_n \) and \( \tilde{\Omega}_n \), the localization and discrete approximation steps make it possible to invoke the Gaussian and bootstrap approximation results from (Chernozhukov, Chetverikov and Kato, 2017). Lastly, the terms \( \tilde{\Pi}_n \) and \( \tilde{I}_n \) are analyzed in correspondence with \( \Pi_n \) and \( I_n \), and here the general-purpose nature of Theorem 2.2 is especially helpful, since it allows the bootstrapped process \( \mathbb{G}^*n \) to be treated in a similar manner to the original process \( \mathbb{G}_n \).

3. Application to inference on covariance matrices

To illustrate the numerical performance of the bootstrap, this section looks at the coverage probabilities of bootstrap confidence regions for \( \Sigma \). An additional set of numerical results dealing with simultaneous confidence intervals for the eigenvalues of \( \Sigma \) are presented in Appendix H of the supplementary material. It should also be noted that both sets of numerical results were obtained in a situation where the leading eigenvalue \( \lambda_1(\Sigma) \) has high multiplicity.

Simulation settings. Simulations were based on the model described in Assumption 2.1, with \( n \in \{300, 500, 700\} \) and \( d = p = 1,000 \), giving \( n < p \) in every case. Also, the matrix \( A \) was constructed to be symmetric so that it can be interpreted as \( A = \Sigma^{1/2} \). To specify \( A \) in more detail, its singular values (equivalently eigenvalues) were chosen as

\[
\sigma_1(A) = \cdots = \sigma_5(A) = 1 \quad \text{and} \quad \sigma_j(A) = j^{-\beta} \quad \text{for } j \in \{6, \ldots, p\},
\]

with decay parameter values \( \beta \in \{0.25, 0.50, 0.75, 1.0, 1.25\} \), and its eigenvectors were taken as the orthogonal factor from a QR decomposition of a \( p \times p \) matrix with independent \( N(0,1) \) entries. Next, for each pair \( (n, \beta) \), we conducted 5,000 trials in which the \( n \times p \) data matrix \( X = ZA \) was generated by filling \( Z \in \mathbb{R}^{n \times p} \) with independent random variables drawn from \( N(0,1) \) or a standardized \( t_{20} \) distribution. Lastly, for each trial, we generated 500 bootstrap samples \( T_n^* \) by sampling the rows of \( X \) with replacement, as described in Section 2.

Error estimation and confidence regions. A natural way to formulate the problem of error estimation for \( \hat{\Sigma} \) is in terms of the \( 1 - \alpha \) quantile of \( T_n \), denoted by \( q_{1-\alpha} \). By definition, this quantity gives the tightest bound of the form

\[
\| \hat{\Sigma} - \Sigma \|_\text{op} \leq \frac{q_{1-\alpha}}{\sqrt{n}}
\]

that holds with probability at least \( 1 - \alpha \). Likewise, if we let \( \bar{q}_{1-\alpha} \) denote the empirical \( (1 - \alpha) \)-quantile of the bootstrap samples \( T_n^* \), then we may regard \( \bar{q}_{1-\alpha}/\sqrt{n} \) as an error estimate for \( \hat{\Sigma} \).
Alternatively, the estimate $\tilde{q}_{1-\alpha}$ can be viewed as specifying an approximate $(1 - \alpha)$-confidence region for $\Sigma$. That is, if we let $B_{\text{op}}(r; \tilde{\Sigma}) \subset \mathbb{R}^{p \times p}$ denote the operator-norm ball of radius $r > 0$ centered at $\tilde{\Sigma}$, then $q_{1-\alpha}/\sqrt{n}$ is the smallest value of $r$ such that

$$P\left(\Sigma \in B_{\text{op}}(r; \tilde{\Sigma})\right) \geq 1 - \alpha.$$ 

Hence, the ideal confidence region may be approximated with $B_{\text{op}}(\tilde{\Sigma}/\sqrt{n}; \tilde{\Sigma})$.

### Table 1. Observed coverage probabilities for $\tilde{q}_{1-\alpha}$, with $\alpha = 0.1$ and $p = 1,000$.  

<table>
<thead>
<tr>
<th>decay param. $\beta$</th>
<th>sample size $n$</th>
<th>300</th>
<th>500</th>
<th>700</th>
<th>300</th>
<th>500</th>
<th>700</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td></td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>0.50</td>
<td></td>
<td>94.3%</td>
<td>92.0%</td>
<td>93.6%</td>
<td>95.4%</td>
<td>93.0%</td>
<td>92.2%</td>
</tr>
<tr>
<td>0.75</td>
<td></td>
<td>92.8%</td>
<td>92.3%</td>
<td>92.0%</td>
<td>92.9%</td>
<td>92.0%</td>
<td>91.9%</td>
</tr>
<tr>
<td>1.00</td>
<td></td>
<td>92.7%</td>
<td>91.7%</td>
<td>91.2%</td>
<td>92.5%</td>
<td>91.8%</td>
<td>91.6%</td>
</tr>
<tr>
<td>1.25</td>
<td></td>
<td>92.4%</td>
<td>91.5%</td>
<td>91.2%</td>
<td>92.5%</td>
<td>91.7%</td>
<td>91.5%</td>
</tr>
</tbody>
</table>

(a) $N(0, 1)$ distribution  
(b) standardized $t_{20}$ distribution

The observed coverage probabilities based on $\tilde{q}_{1-\alpha}$ have been listed in Table 1, with the nominal coverage level always being set to 90%. (Note that these probabilities can be interpreted either with respect to the coverage of the error bound or the confidence region.) Looking at the results, we see that they generally conform with our theoretical analysis, in the sense that the performance of the bootstrap depends substantially on whether $\beta$ is above or below 0.50. Namely, for $\beta > 0.50$ and sufficiently large sample sizes, the observed coverage comes nearly within 1% of the desired level. Furthermore, this holds in spite of the fact that $\lambda_1(\Sigma)$ has high multiplicity. On the other hand, when $\beta < 0.50$, the bootstrap breaks down even at large sample sizes, with the observed coverage of 100% being far higher than the desired 90%. Apart from this dichotomy, it should be noted that the bootstrap has the favorable property that it errs reliably in the conservative direction, with the observed coverage never falling below the nominal level. More generally, we observed similar patterns in the context of simultaneous confidence intervals for the eigenvalues of $\Sigma$, as discussed in Appendix H.

## 4. Application to randomized numerical linear algebra

Over the past decade, RandNLA has become the focus of intense activity in many fields related to large-scale computation (Mahoney, 2011; Halko, Martinsson and Tropp, 2011; Woodruff, 2014; Kannan and Vempala, 2017; Drineas and Mahoney, 2018). Broadly speaking, this new direction of research has stemmed from the principle that randomization is a very general mechanism for scaling up algorithms. However, in exchange for scalability, randomized sketching algorithms are typically less accurate than their deterministic predecessors. Therefore, in order to use sketching reliably, it is crucial to verify...
that the algorithmic error is small, which motivates new applications of the bootstrap beyond its traditional domains.

The purpose of this section is to illustrate how the bootstrap can be applied to estimate operator-norm error for randomized matrix multiplication, which has been a prominent topic in the RandNLA literature (e.g., Drineas and Kannan, 2001; Drineas, Kannan and Mahoney, 2006; Magen and Zouzias, 2011; Pagh, 2013; Holodnak and Ipsen, 2015; Cohen, Nelson and Woodruff, 2016; Gupta et al., 2018). A related study of the bootstrap for this application can also be found in (Lopes, Wang and Mahoney, 2019), which differs from the current work insofar as it deals exclusively with the entrywise $\ell_\infty$-norm and does not focus the role of spectrum decay.

To proceed, we will first provide a brief review of the algorithmic setting (Section 4.1), followed by an efficient implementation of the bootstrap using an extrapolation technique (Section 4.2). In turn, we will present numerical results for synthetic matrices (Section 4.3), as well as an example concerning spatial modes of temperature variation (Section 4.4).

### 4.1. Rudiments of sketching

Consider a situation involving a very large deterministic matrix $A \in \mathbb{R}^{d \times p}$ with $p \ll d$, where the product $A^T A$ is too expensive to compute to high precision. For instance, this often occurs when $A$ must be stored on disk because it exceeds the constraints of fast memory. Alternatively, even when memory is not a bottleneck, matrix products can become too expensive if they must be computed frequently as a subroutine of a larger pipeline.

The basic idea of sketching is to work with a shorter version of $A$, referred to as a “sketch of $A$”, and denoted as $\tilde{A} \in \mathbb{R}^{n \times p}$ where $n \ll d$. This matrix is defined as

$$\tilde{A} = SA,$$

where $S \in \mathbb{R}^{n \times d}$ is a random “sketching matrix” that is generated by the user. In particular, the user must choose the “sketch size” $n$. Intuitively, the matrix $S$ is intended to shorten $A$ in a way that retains most of the information, so that the inexpensive product $\tilde{A}^T \tilde{A}$ will provide a good approximation to $A^T A$.

The sketching matrix. Typically, the action of $S$ upon $A$ is interpreted in either of two ways: randomly projecting columns from $\mathbb{R}^d$ into $\mathbb{R}^n$, or discretely sampling $n$ among $d$ rows. In addition, the matrix $S$ is commonly generated by the user so that its rows are i.i.d., and that it satisfies $\mathbb{E}[S^T S] = I_n$, which implies that $\tilde{A}^T \tilde{A}$ is unbiased with respect to $A^T A$. At a high level, these basic properties are sufficient to understand all of our work below, but numerous types of sketching matrices have been studied in the literature. For instance, two of the most well-known are the Gaussian random projection and uniform row sampling types, where the former has i.i.d. rows drawn from $N(0, \frac{1}{n} I_d)$, and the latter has i.i.d. rows drawn uniformly from $\{\sqrt{\frac{d}{n}} e_1, \ldots, \sqrt{\frac{d}{n}} e_d\}$. More elaborate examples may be found in the references above.
**Cost versus accuracy.** Whenever sketching is implemented, the choice of the sketch size $n$ plays a pivotal role in a tradeoff between computational cost and accuracy. To see this, note that on one hand, the cost to compute $\tilde{A}^\top \tilde{A}$ is generally proportional to $n$, with the number of operations being $O(np^2)$. On the other hand, the operator-norm error of $\tilde{A}^\top \tilde{A} - A^\top A$ tends to decrease stochastically like $1/\sqrt{n}$, because the difference $\tilde{A}^\top \tilde{A} - A^\top A$ can be expressed as a sample average of $n$ centered random matrices (i.e., in the same way as $\tilde{\Sigma} - \Sigma$).

**The need for error estimation.** Although the choice of the sketch size $n$ has clear importance, this choice also involves practical difficulties that expose a major gap between the theory and practice of sketching. Specifically, these difficulties arise because the value of the sketching error $\|\tilde{A}^\top \tilde{A} - A^\top A\|_{\text{op}}$ is unknown in practice, as it depends on the unknown product $A^\top A$. Hence, it is hard for the user to know if any given choice of $n$ will achieve a desired level of accuracy.

As a way to handle this dilemma, one option is to consult the RandNLA literature on theoretical error bounds for $\|\tilde{A}^\top \tilde{A} - A^\top A\|_{\text{op}}$, as surveyed in the references above. However, much like in the setting of covariance estimation, these results usually only provide qualitative guidance, and they rarely offer an explicit numerical bound. Most often, this occurs because of unspecified theoretical constants, but there is also a second key limitation: Theoretical error bounds are generally formulated to hold in a worst-case sense, and so they often fail to account for special structure. Due to these issues, we propose instead to directly estimate the error via a computationally efficient bootstrap method. This has the twofold benefit of providing a numerical bound and adapting automatically to the structure of the problem at hand.

**Comparison of sketching and covariance estimation.** To clarify the relationship between the sketching error $\|\tilde{A}^\top \tilde{A} - A^\top A\|_{\text{op}}$ and the covariance estimation error $\|\tilde{\Sigma} - \Sigma\|_{\text{op}}$, let $A$, $\Sigma$, and $\tilde{\Sigma}$ be understood as in the context of the model 2.1 with $Z \in \mathbb{R}^{n \times d}$ having rows $Z_1, \ldots, Z_n$, and let

$$S = \frac{1}{\sqrt{n}} Z.$$

Under these conditions, the matrix $S$ has the desired properties of a sketching matrix mentioned earlier, and furthermore $\tilde{A}^\top \tilde{A} - A^\top A = \tilde{\Sigma} - \Sigma$. However, it is worth highlighting that this formal similarity conceals some operational differences. For instance, the matrices $Z$ and $A$ are unobservable to the user in covariance estimation, whereas the user does have access to $S$ and $A$ in sketching. Secondly, in covariance estimation, the user often does not have the option to increase $n$, but in sketching, it is possible to construct a rough initial sketch of $A$ for inspection, and then take a second sketch to improve performance. Later on, we will show how this second point has an important link with our error estimation method, because it will enable the user to dynamically predict the total sketch size needed to reach a given level of accuracy.
4.2. Error estimation with an extrapolated bootstrap

The intuition for applying the bootstrap to sketching comes from thinking of the matrix \( \tilde{A} \) as a “dataset” whose rows are “observations”. In particular, this interpretation is supported by the fact that many types of sketching matrices \( S \) cause the rows of \( \tilde{A} \) to be i.i.d. Therefore, we may expect that sampling from the rows of \( \tilde{A} \) with replacement will faithfully mimic the process that generated \( \tilde{A} \).

To fix some notation for describing the bootstrap method, let \( q_{1-\alpha} \) denote the \((1-\alpha)\)-quantile of the sketching error variable \( \| \hat{A}^\top \hat{A} - A^\top A \|_\text{op} \), which is the minimal value such that the event

\[
\| \hat{A}^\top \hat{A} - A^\top A \|_\text{op} \leq q_{1-\alpha}
\]

holds with probability at least \( 1 - \alpha \). Our main goal is to construct an estimate \( \hat{q}_{1-\alpha} \) using only the sketch \( \tilde{A} \) as a source of information. Below, we state a basic version of the bootstrap method in Algorithm 1, which will later be accelerated via an extrapolation technique in Section 4.2.1.

**Algorithm 1.** (Bootstrap estimate of sketching error).

<table>
<thead>
<tr>
<th>Input</th>
<th>The number of bootstrap samples ( B ), and the sketch ( \tilde{A} \in \mathbb{R}^{n \times p} ).</th>
</tr>
</thead>
<tbody>
<tr>
<td>For ( b = 1, \ldots, B ) do</td>
<td></td>
</tr>
<tr>
<td>1. Form ( \hat{A}^* \in \mathbb{R}^{n \times p} ) by drawing ( n ) rows from ( \tilde{A} ) with replacement.</td>
<td></td>
</tr>
<tr>
<td>2. Compute the bootstrap sample ( \varepsilon_b^* := | (\hat{A}^<em>)^\top (\hat{A}^</em>) - \hat{A}^\top \hat{A} |_\text{op} ).</td>
<td></td>
</tr>
<tr>
<td>Return: ( \hat{q}_{1-\alpha} \leftarrow ) the ((1-\alpha))-quantile of the values ( \varepsilon_1^<em>, \ldots, \varepsilon_B^</em> ).</td>
<td></td>
</tr>
</tbody>
</table>

**Remark.** Given that the construction of \( \hat{A} \) is fully controlled by the user, one might ask why bootstrapping is preferable to carrying out many repetitions of the actual sketching process. The answer comes down to the fact that constructing \( \hat{A} \) requires a computation involving the full matrix \( A \), which often incurs high communication costs. In fact, this issue is one of the primary motivations for the whole subject of RandNLA, which usually deals with situations where it is only feasible to access \( A \) at most a handful of times. In contrast to the task of constructing \( \hat{A} \) from \( A \), Algorithm 1 only requires inexpensive access to the much smaller matrix \( \tilde{A} \), and it requires no access to \( A \) whatsoever.

4.2.1. Extrapolation

Because the user has the option to increase the sketch size \( n \) by performing an extra round of sketching, it becomes possible to accelerate the bootstrap with an extrapolation technique that is often not applicable in covariance estimation. To develop the idea, we should first recall that the fluctuations of \( \| \hat{A}^\top \hat{A} - A^\top A \|_\text{op} \) tend to scale like \( 1/\sqrt{n} \) as a function of \( n \), because the difference \( \hat{A}^\top \hat{A} - A^\top A \) can be written as a centered sample average of \( n \) random matrices. Therefore, if we view the sketching error quantile as
a function of \( n \), say \( q_{1-\alpha} = q_{1-\alpha}(n) \), then we may expect the following approximate relationship between a small “initial” sketch size \( n_0 \), and a larger “final” sketch size \( n_1 \),

\[
q_{1-\alpha}(n_1) \approx \sqrt{\frac{n_0}{n_1}} q_{1-\alpha}(n_0).
\] (4.1)

The significance of this approximation is that \( q_{1-\alpha}(n_0) \) is computationally much easier to estimate than \( q_{1-\alpha}(n_1) \), since the former involves bootstrapping a matrix of size \( n_0 \times p \), rather than \( n_1 \times p \). More general background on the connections between extrapolation and resampling methods can be found in (Bickel and Yahav, 1988; Bertail et al., 1997; Bertail and Politis, 2001; Bickel and Sakov, 2002; Lopes, 2019), among others.

Based on the heuristic approximation (4.1), we can obtain an inexpensive estimate of \( q_{1-\alpha}(n_1) \) for any \( n_1 > n_0 \) by using

\[
\tilde{q}_{1-\alpha}^{\text{ext}}(n_1) := \sqrt{\frac{n_0}{n_1}} \tilde{q}_{1-\alpha}(n_0),
\] (4.2)

where \( \tilde{q}_{1-\alpha}(n_0) \) is obtained from Algorithm 1. More concretely, if the user has the ultimate intention of achieving \( \| \hat{A}^\top \hat{A} - A^\top A \|_{\text{op}} \leq \epsilon_{\text{tol}} \) for some tolerance \( \epsilon_{\text{tol}} \), then extrapolation may be applied in the following way: First, the user should check the condition \( \tilde{q}_{1-\alpha}(n_0) \leq \epsilon_{\text{tol}} \) to see if \( n_0 \) is already large enough. Second, if \( n_0 \) is too small, then the rule (4.2) instructs the user to obtain a final sketch size \( n_1 \) satisfying \( \tilde{q}_{1-\alpha}^{\text{ext}}(n_1) \leq \epsilon_{\text{tol}} \), which is equivalent to \( n_1 \geq \frac{n_0}{\epsilon_{\text{tol}}} \tilde{q}_{1-\alpha}(n_0)^2 \). Furthermore, our numerical results will demonstrate that this simple technique remains highly effective even when \( n_1 \) is much larger than \( n_0 \), such as by an order of magnitude (cf. Sections 4.3 and 4.4).

### 4.2.2. Assessment of cost

Since the overall purpose of sketching is to reduce computation, it is important to explain why the added cost of the bootstrap is manageable. In particular, the added cost should not be much higher than the cost of sketching itself. As a simple point of reference, the cost to construct \( A \) and then compute \( \hat{A}^\top \hat{A} \) with most state-of-the-art sketching algorithms is at least \( C_{\text{sketch}} = \Omega(dp + n_1p^2) \), where \( n_1 \) refers to the “final” sketch size described above.

Next, to assess the cost of the bootstrap, we can take advantage of a small initial sketch size \( n_0 \) by using extrapolation, as well as the fact that the bootstrap samples can be trivially computed in parallel, with say \( m \) processors. When these basic factors are taken into account, the cost of the bootstrap turns out to be at most \( C_{\text{boot}} = \mathcal{O}(Bn_0p^2/m) \).

From this discussion of cost, perhaps the most essential point to emphasize is that \( C_{\text{sketch}} \) grows linearly with \( d \), whereas \( C_{\text{boot}} \) is independent of \( d \). Indeed, this is of great importance for scalability, because randomized matrix multiplication is of primary interest in situations where \( d \) is extremely large. Beyond this general observation, we can also take a more detailed look to see that the condition \( C_{\text{boot}} = \mathcal{O}(C_{\text{sketch}}) \) occurs when \( B = \mathcal{O}((\frac{n_1}{n_0} + \frac{d}{p n_0})m) \). Furthermore, such a condition on \( B \) can be considered realistic in light of our experiments, since the modest choice of \( B = 50 \) is shown to yield good results.
4.3. Numerical results for synthetic matrices

We now demonstrate the performance of the bootstrap estimate $\hat{q}_{1-\alpha}$ in a range of conditions, both with and without extrapolation. Most notably, the numerical results for extrapolation are quite encouraging.

**Simulation settings.** The choices for the matrix $A \in \mathbb{R}^{d \times p}$ were developed in analogy with those in Section 3, except that in this context, the matrix is very tall with $d = 10,000$ and $p = 1,000$. If we let $A = UDV^T$ denote the singular value decomposition, then the singular vectors were specified by taking $U \in \mathbb{R}^{d \times p}$ and $V \in \mathbb{R}^{p \times p}$ to be orthonormal factors from QR decompositions of matrices filled with independent $N(0,1)$ entries. In addition, the singular values were chosen as $\sigma_1(A) = \cdots = \sigma_5(A) = 1$ and $\sigma_j(A) = j^{-\beta}$ for $j \in \{6, \ldots, p\}$, with decay parameters $\beta \in \{0.75, 1.0, 1.25\}$. These values were chosen in order to show that the bootstrap can work even when there are no gaps among the leading singular values.

**Design of simulations.** The design of the simulations can be understood in terms of Figure 1. For each value of $\beta$, and sketch size $n \in \{300, \ldots, 2,100\}$, we performed $1,000$ trials of sketching to compute independent copies $\tilde{A} \in \mathbb{R}^{n \times p}$ using two different types of sketching matrices: Gaussian random projection, and uniform row sampling, as defined in Section 4.1. In turn, the actual values of $\|\tilde{A}^T \tilde{A} - A^T A\|_{op}$ in these trials yielded a high quality surrogate for the true 90% quantile $q_{0.9} = q_{0.9}(n)$, plotted as a function of $n$ with the black dashed line.

With regard to Algorithm 1, it was applied during each trial to compute $\hat{q}_{1-\alpha}$ using $B = 50$ bootstrap samples. The average of these estimates is plotted as a function of $n$ with the solid blue line. In addition, the performance of the extrapolation rule (4.2) was studied by applying it to each estimate $\hat{q}_{0.9}(n_0)$ computed at $n_0 = 300$. The average of the extrapolated curves is plotted in solid red, with the pink envelope signifying $\pm 1$ standard deviation.

**Comments on results.** Figure 1 shows that on average, the bootstrap estimates are nearly equal to the true quantile over the entire range of sketch sizes $n \in \{300, \ldots, 2,100\}$, both with and without extrapolation. Indeed, the performance of the extrapolated estimate is especially striking, because it shows that bootstrapping a rough initial sketch $\tilde{A}$ of size $300 \times 1,000$ can be used to accurately predict the error of a much larger sketch of size $2,100 \times 1,000$. To put this into context, we should also remember that the original matrix $A$ is of size $10,000 \times 1,000$, and hence the initial sketch is able to provide quite a bit of information about the sketching task for a small computational price. Moreover, the fact that the extrapolation works up to the larger sketch size of $2,100$ means that a 7-fold speedup can be obtained in comparison to naively applying Algorithm 1 to the larger sketch. (In fact, the plots seem to suggest that the extrapolation would remain accurate for sketch sizes beyond 2,100, and that even larger speedups are attainable.) Lastly, it is worth noting that even though a small choice of $B = 50$ bootstrap samples was used, the standard deviation of the extrapolated estimate is rather well-behaved, as
indicated by the pink envelope.

\[ \beta = 0.75 \]

\[ \beta = 1.0 \]

\[ \beta = 1.25 \]

(a) Sketching with Gaussian random projections.

(b) Sketching with uniform row sampling.

Figure 1: Bootstrap estimates for the 90% quantile of the error \( \| \hat{A}^\top \hat{A} - A^\top A \|_{op} \).

### 4.4. Sea surface temperature measurements

Large-scale dynamical systems are ubiquitous in the physical sciences, and advances in technology for measuring these systems have led to rapidly increasing volumes of data. Consequently, it is often too costly to apply standard tools of exploratory data analysis in a direct manner, and there has been growing interest to use sketching as a data-reduction strategy that preserves the essential information (e.g. Brunton et al., 2015; Erichson et al., 2019; Ribeiro, Yeh and Taira, 2019; Bai et al., 2019; Saibaba, 2019; Bjarkason, 2019; Tropp et al., 2019, among others).

This type of situation is especially common in fields such as climate science and fluid dynamics, where we may be presented with a very large matrix \( A \in \mathbb{R}^{d \times p} \) whose rows form a long sequence of “snapshots” that represent a dynamical system at time points 1, \ldots, \( d \). As a concrete example, we consider satellite recordings of sea surface temperature that have been collected over the time period 1981-2018, and are available from the National Oceanic and Atmospheric Administration (NOAA) (cf. Reynolds et al., 2002). More specifically, we deal with a particular subset of the data corresponding to \( d = 13,271 \) temporal snapshots at \( p = 3,944 \) spatial grid points in the eastern Pacific Ocean, shown in Figure (2a). From the standpoint of climate science, this region is important for studying the phenomenon known as the El Niño Southern Oscillation (ENSO).
This example is relevant to our discussion of sketching for several reasons. First, the matrix product $A^T A$ is of interest because it describes spatial modes of temperature variation through its eigenstructure. In particular, the fourth eigenvector (mode) of $A^T A$ identifies the intermittent El Niño and La Niña warming events that are influential global weather patterns, as displayed in Figure (2b) (Erichson et al., 2018). Second, the singular values of $A$ have a natural decay profile, which is illustrated in Figure (4a). Lastly, the example demonstrates the need for error estimation in order to guide the choice of sketch size. This can be seen in Figures (3a) and (3b) below, where it is shown that an insufficient sketch size can heavily distort the ENSO mode in comparison to the exact form given in Figure (2b).

To conclude this example, we present numerical results for the bootstrap error estimates. Analogously to Section 4.3, we consider the task of estimating the 90% quantile $q_{0.9}(n)$ of the sketching error, viewed as a function of $n$. The full matrix $A$ is of size $13,271 \times 3,944$, as described earlier, except that it was normalized to satisfy $\sigma_1(A) = 1$, so that the results here can be easily compared on the same scale with the previous results in Section 4.3. Also, the results shown here in Figure (4b) are plotted in the same format, with the number of trials being 1,000, the number of bootstrap samples being $B = 50$, and the sketching matrices being Gaussian random projections.

From looking at Figure (4b), we see that the performance of the bootstrap in the case of the naturally generated matrix $A$ is very similar to that in the previous cases of synthetic matrices. Namely, the averages of both the extrapolated and non-extrapolated estimates virtually overlap with the true curve, and furthermore, the fluctuations of the extrapolated estimates are well controlled. Lastly, the extrapolation rule accurately estimates the quantile value $q_{0.9}(n_1)$ at a final sketch size $n_1 = 5,000$ that is 10 times larger than the initial sketch size $n_0 = 500$, which shows the potential of this rule to accelerate computations without sacrificing the quality of estimation.
Figure 3: The left and right panels show approximations to the ENSO mode based on the approximate product $\tilde{A}^T \tilde{A}$, obtained from Gaussian random projections with sketch sizes $n = 500$ and $n = 3,000$. A comparison with the exact ENSO mode in Figure (2b) above shows that an insufficient sketch size can lead to a substantial distortion.

Figure 4: The left panel displays the decaying eigenvalues of $A^T A$, where the x-axis is logarithmic. The right panel demonstrates that the extrapolated and non-extrapolated bootstrap methods accurately estimate the 90% quantile of the sketching error $\|\tilde{A}^T \tilde{A} - A^T A\|_{op}$ over a wide range of sketch sizes. In particular, the extrapolation rule gives accurate results at a final sketch size $n_1 = 5,000$ that is 10 times larger than the initial sketch size $n_0 = 500$.

References


Lust-Piquard, F. (1986). Inégalités de Khintchine dans $C_p$ ($1 < p < \infty$). *Comptes


Processes, Springer.