Bootstrapping Spectral Statistics in High Dimensions

Miles Lopes

UC Davis

Random Matrices and Complex Data Analysis Workshop

Shanghai 2019

Bootstrap for sample covariance matrices

- Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be i.i.d. observations, and let $\widehat{\Sigma}$ be the sample covariance matrix.
- Let $T = \varphi(\widehat{\Sigma})$ denote a statistic of interest.
- We would like to estimate $\mathrm{var}(T)$, or more generally, approximate the sampling distribution of T.
- The non-parametric bootstrap offers a general way to solve these problems.

Bootstrap for sample covariance matrices

- Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be i.i.d. observations, and let $\widehat{\Sigma}$ be the sample covariance matrix.
- Let $T = \varphi(\widehat{\Sigma})$ denote a statistic of interest.
- We would like to estimate $\mathrm{var}(T)$, or more generally, approximate the sampling distribution of T.
- The non-parametric bootstrap offers a general way to solve these problems.

Non-parametric bootstrap.

For: $b = 1, \ldots, B$:

- Sample n points $X_1^*, \ldots, X_n^*$ with replacement from $\{X_1, \ldots, X_n\}$.
- Form the sample covariance matrix $\widehat{\Sigma}^*$ associated with $X_1^*, \ldots, X_n^*$.
- Compute $T_b^* = \varphi(\widehat{\Sigma}^*)$.

Return: the estimate $\frac{1}{B}\sum_{b=1}^{B}(T_b^* - \bar{T}^*)^2$ for $\mathrm{var}(T)$.

Some past work

- In 1985, Beran and Srivastava showed that the standard non-parametric bootstrap generally works for smooth functionals of $\widehat{\Sigma}$ when $p \ll n$. (Exceptions arise for non-smooth functionals, or tied population eigenvalues.)

Some past work

- In 1985, Beran and Srivastava showed that the standard non-parametric bootstrap generally works for smooth functionals of $\widehat{\Sigma}$ when $p \ll n$. (Exceptions arise for non-smooth functionals, or tied population eigenvalues.)

- The paper Hall, Lee, Park, Paul (2009) develops a remedy for tied eigenvalues, as well as a generalization to functional data. (A good literature survey is also provided for many other papers in the $p \ll n$ setting between 1985 and 2009.)

Some past work

- In 1985, Beran and Srivastava showed that the standard non-parametric bootstrap generally works for smooth functionals of $\widehat{\Sigma}$ when $p \ll n$. (Exceptions arise for non-smooth functionals, or tied population eigenvalues.)

- The paper Hall, Lee, Park, Paul (2009) develops a remedy for tied eigenvalues, as well as a generalization to functional data. (A good literature survey is also provided for many other papers in the $p \ll n$ setting between 1985 and 2009.)

- However, $p \asymp n$ or $p \gg n$, relatively little is known about bootstrap consistency.

Some past work

- In 1985, Beran and Srivastava showed that the standard non-parametric bootstrap generally works for smooth functionals of $\widehat{\Sigma}$ when $p \ll n$. (Exceptions arise for non-smooth functionals, or tied population eigenvalues.)

- The paper Hall, Lee, Park, Paul (2009) develops a remedy for tied eigenvalues, as well as a generalization to functional data. (A good literature survey is also provided for many other papers in the $p \ll n$ setting between 1985 and 2009.)

- However, $p \asymp n$ or $p \gg n$, relatively little is known about bootstrap consistency.

★ There are many opportunities for future work on bootstrap methods in high-dimensional inference, especially in connection with random matrix theory.

Some recent developments

- Recently, El Karoui and Purdom (2019) have studied the non-parametric bootstrap, and have demonstrated some negative empirical results for $\lambda_1(\widehat{\Sigma})$. They also prove bootstrap consistency for a fixed number of the largest sample eigenvalues when Σ has low effective rank.

Some recent developments

- Recently, El Karoui and Purdom (2019) have studied the non-parametric bootstrap, and have demonstrated some negative empirical results for $\lambda_1(\widehat{\Sigma})$. They also prove bootstrap consistency for a fixed number of the largest sample eigenvalues when Σ has low effective rank.

- Han, Xu and Zhou (2018) have studied a Gaussian multiplier bootstrap to approximate the distribution of statistics of the form

$$T = \sup_{\|u\|_2 \leq 1, \|u\|_0 \leq s} \sqrt{n} \left| u^\top \left(\widehat{\Sigma} - \Sigma \right) u \right|$$

  and variants thereof, in the case of sparse test vectors with $s \ll n$.

Some recent developments

- Recently, El Karoui and Purdom (2019) have studied the non-parametric bootstrap, and have demonstrated some negative empirical results for $\lambda_1(\widehat{\Sigma})$. They also prove bootstrap consistency for a fixed number of the largest sample eigenvalues when Σ has low effective rank.

- Han, Xu and Zhou (2018) have studied a Gaussian multiplier bootstrap to approximate the distribution of statistics of the form

$$T = \sup_{\|u\|_2 \leq 1, \|u\|_0 \leq s} \sqrt{n} \left| u^\top \left(\widehat{\Sigma} - \Sigma \right) u \right|$$

  and variants thereof, in the case of sparse test vectors with $s \ll n$.

- Naumov, Spokoiny and Ulyanov (2019) have studied multiplier bootstrap methods for approximating the error distribution of spectral projectors, e.g. statistics of the form $T = n \| \widehat{v}_1 \widehat{v}_1^\top - v_1 v_1^\top \|_F^2$, where v_1 and $\widehat{v}_1$ the leading population and sample eigenvectors. Consistency is established when Σ has low effective rank.

Difficulties in high dimensions

- Why do difficulties arise when p is large?

Difficulties in high dimensions

- Why do difficulties arise when p is large?

- If it were possible, we would prefer to draw an i.i.d. sample from the (unknown) distribution $\mathbb{P}$ underlying $\mathcal{D} = \{X_1, \ldots, X_n\}$.

Difficulties in high dimensions

- Why do difficulties arise when p is large?
- If it were possible, we would prefer to draw an i.i.d. sample from the (unknown) distribution $\mathbb{P}$ underlying $\mathcal{D} = \{X_1, \ldots, X_n\}$.
- Instead, the bootstrap uses an i.i.d. sample from the empirical distribution $\widehat{\mathbb{P}}$, which places mass $1/n$ at each point in $\mathcal{D}$.

Difficulties in high dimensions

- Why do difficulties arise when p is large?
- If it were possible, we would prefer to draw an i.i.d. sample from the (unknown) distribution $\mathbb{P}$ underlying $\mathcal{D} = \{X_1, \ldots, X_n\}$.
- Instead, the bootstrap uses an i.i.d. sample from the empirical distribution $\widehat{\mathbb{P}}$, which places mass $1/n$ at each point in $\mathcal{D}$.
- Key difficulty: If p is large, and $\mathbb{P}$ does not have "low-dimensional structure", then $\widehat{\mathbb{P}}$ may be a poor substitute for $\mathbb{P}$.

Possible approaches to bootstrapping in high dimensions

1. When $\mathbb{P}$ does have low-dimensional structure, the non-parametric bootstrap can still succeed.

Possible approaches to bootstrapping in high dimensions

1. When $\mathbb{P}$ does have low-dimensional structure, the non-parametric bootstrap can still succeed.

2. Even if $\mathbb{P}$ does not have such structure, we may instead rely on special invariance properties of the statistic T.

Possible approaches to bootstrapping in high dimensions

1. When $\mathbb{P}$ does have low-dimensional structure, the non-parametric bootstrap can still succeed.

2. Even if $\mathbb{P}$ does not have such structure, we may instead rely on special invariance properties of the statistic T.

    - For instance, universality results may indicate that the fluctuations of T governed by a small set of "relevant" parameters.

Possible approaches to bootstrapping in high dimensions

1. When $\mathbb{P}$ does have low-dimensional structure, the non-parametric bootstrap can still succeed.

2. Even if $\mathbb{P}$ does not have such structure, we may instead rely on special invariance properties of the statistic T.

    - For instance, universality results may indicate that the fluctuations of T governed by a small set of "relevant" parameters.

    - If we can determine the relevant parameters (say θ), then we can sample from a suitable parametric distribution $\mathbb{P}_{\widehat{\theta}}$.

Possible approaches to bootstrapping in high dimensions

1. When $\mathbb{P}$ does have low-dimensional structure, the non-parametric bootstrap can still succeed.

2. Even if $\mathbb{P}$ does not have such structure, we may instead rely on special invariance properties of the statistic T.

   - For instance, universality results may indicate that the fluctuations of T governed by a small set of "relevant" parameters.

   - If we can determine the relevant parameters (say θ), then we can sample from a suitable parametric distribution $\mathbb{P}_{\widehat{\theta}}$.

key point: $\mathcal{L}\big(T(\mathbb{P})\big) \approx \mathcal{L}\big(T(\mathbb{P}_{\widehat{\theta}}) \,|\, \mathcal{D}\big)$ even though $\mathbb{P} \not\approx \mathbb{P}_{\widehat{\theta}}$.

Two parts

Part I: *Bootstrapping spectral statistics in high dimensions*

with A. Blandino, and A. Aue

Biometrika, 2019

Part II: *Bootstrapping the operator norm in high dimensions: Error estimation for covariance matrices and sketching*

with N. B. Erichson and M. W. Mahoney

`https://arxiv.org/abs/1909.06120`

A basic model for studying covariance matrices

Let $\Sigma \in \mathbb{R}^{p \times p}$ be a population covariance matrix.

Suppose $X \in \mathbb{R}^{n \times p}$ is a data matrix with i.i.d. rows generated as

$$X_i = \Sigma^{1/2} Z_i \tag{1}$$

where the vectors $Z_1, \ldots, Z_n \in \mathbb{R}^p$ have i.i.d. entries with $\mathbb{E}[Z_{ij}] = 0$, $\mathbb{E}[Z_{ij}^2] = 1$ and $\mathbb{E}[Z_{ij}^4] =: \kappa > 1$.

Define the sample covariance matrix

$$\widehat{\Sigma} = \frac{1}{n} X^\top X. \tag{2}$$

Linear Spectral Statistics (LSS)

A natural class of prototype statistics for investigating bootstrap consistency are *linear spectral statstics*, which have the form

$$T = \frac{1}{p} \sum_{j=1}^{p} f(\lambda_j(\widehat{\Sigma})), \tag{3}$$

where f is a smooth function on $[0, \infty)$.

Linear Spectral Statistics (LSS)

A natural class of prototype statistics for investigating bootstrap consistency are *linear spectral statstics*, which have the form

$$T = \frac{1}{p} \sum_{j=1}^{p} f(\lambda_j(\widehat{\Sigma})), \tag{3}$$

where f is a smooth function on $[0, \infty)$.

Examples:

- The choice $f(x) = \log(x)$ leads to $\log(\det(\widehat{\Sigma}))$.
- The choice $f(x) = x^k$, leads to $\mathrm{tr}(\widehat{\Sigma}^k)$
- The normal log-likelihood ratio statistic for testing sphericity is

$$p \log(\mathrm{tr}(\widehat{\Sigma})) - \log(\det(\widehat{\Sigma})).$$

- Even some non-linear spectral statistics are "asymptotically equivalent" to transformations of LSS (cf. Dobriban 2017)

Background ideas for developing a new bootstrap

Beginning with fundamental works of Jonsson (1982) and Bai and Silverstein (2004), a substantial literature has developed increasingly general central limit theorems for LSS:

(e.g. Pan and Zhou (2008), Lytova and Pastur (2009), Bai, Wang and Zhou (2010), Scherbina (2011), Zheng (2012), Wang and Yao (2013), Naijm and Yao (2016), Li, Li and Yao (2018), Hu, Li, Liu and Zhou (2019) among others)

Background ideas for developing a new bootstrap

Beginning with fundamental works of Jonsson (1982) and Bai and Silverstein (2004), a substantial literature has developed increasingly general central limit theorems for LSS:

(e.g. Pan and Zhou (2008), Lytova and Pastur (2009), Bai, Wang and Zhou (2010), Scherbina (2011), Zheng (2012), Wang and Yao (2013), Naijm and Yao (2016), Li, Li and Yao (2018), Hu, Li, Liu and Zhou (2019) among others)

In particular, if $\mathbb{E}[Z_{ij}^4] = 3$, and $p/n \to c \in (0, \infty)$, then under "standard assumptions" we have

$$p(T - \mathbb{E}[T]) \;\Rightarrow\; N(0, \sigma^2),$$

where

$$\sigma^2 = \frac{-1}{2\pi^2} \oiint \frac{f(z_1)f(z_2)}{(\underline{m}(z_1) - \underline{m}(z_2))^2} \frac{d}{dz_1}\underline{m}(z_1) \frac{d}{dz_2}\underline{m}(z_2) dz_1 dz_2$$

Background ideas for developing a new bootstrap

$$p(T - \mathbb{E}[T]) \Rightarrow N(0, \sigma^2)$$

Important property: Under conditions more general than $\mathbb{E}[Z_{ij}^3] = 3$, the variance σ^2 is essentially determined by the limiting spectral distribution of Σ.

Roughly speaking, this means that under certain conditions, the limit laws of LSS are mainly governed by just $(\lambda_1(\Sigma), \ldots, \lambda_p(\Sigma))$, rather than the entire matrix Σ.

This is a major reduction in complexity.

A "parametric bootstrap" approach

More good news: The eigenvalues $\Lambda = \text{diag}(\lambda_1(\Sigma), \ldots, \lambda_p(\Sigma))$ can be estimated well in high dimensions. In particular, for consistent estimation of the population LSD, it is not necessary to use sparsity and/or low-rank conditions.

(cf. El Karoui (2008), Mestre (2008), Li and Yao (2014), Ledoit and Wolf (2015), Kong and Valiant (2017), among others)

A "parametric bootstrap" approach

More good news: The eigenvalues $\Lambda = \text{diag}(\lambda_1(\Sigma), \ldots, \lambda_p(\Sigma))$ can be estimated well in high dimensions. In particular, for consistent estimation of the population LSD, it is not necessary to use sparsity and/or low-rank conditions.

(cf. El Karoui (2008), Mestre (2008), Li and Yao (2014), Ledoit and Wolf (2015), Kong and Valiant (2017), among others)

Intuitive procedure: Generate a "new dataset" X^* that nearly matches the observed data X with respect to Λ.

Then, we compute the statistic T^* arising from the "new data" X^*.

(This is akin to a parametric bootstrap.)

A "parametric bootstrap" approach

More good news: The eigenvalues $\Lambda = \text{diag}(\lambda_1(\Sigma), \ldots, \lambda_p(\Sigma))$ can be estimated well in high dimensions. In particular, for consistent estimation of the population LSD, it is not necessary to use sparsity and/or low-rank conditions.

(cf. El Karoui (2008), Mestre (2008), Li and Yao (2014), Ledoit and Wolf (2015), Kong and Valiant (2017), among others)

Intuitive procedure: Generate a "new dataset" X^* that nearly matches the observed data X with respect to Λ.

Then, we compute the statistic T^* arising from the "new data" X^*.

(This is akin to a parametric bootstrap.)

One extra detail: The kurtosis $\kappa = \mathbb{E}[Z_{ij}^4]$ matters too, but it's estimable.

Proposed method: Spectral Bootstrap

Goal. Approximate the distribution of $T = \frac{1}{p}\sum_{j=1}^{p} f(\lambda_j(\widehat{\Sigma}))$.

Proposed method: Spectral Bootstrap

Goal. Approximate the distribution of $T = \frac{1}{p}\sum_{j=1}^{p} f(\lambda_j(\widehat{\Sigma}))$.

Before resampling, first compute estimates $\widehat{\kappa}$ and $\widehat{\Lambda}$.

Proposed method: Spectral Bootstrap

Goal. Approximate the distribution of $T = \frac{1}{p}\sum_{j=1}^{p} f(\lambda_j(\widehat{\Sigma}))$.

Before resampling, first compute estimates $\widehat{\kappa}$ and $\widehat{\Lambda}$.

Algorithm. (Spectral Bootstrap)

For $b = 1, \ldots, B$:

Proposed method: Spectral Bootstrap

Goal. Approximate the distribution of $T = \frac{1}{p}\sum_{j=1}^{p} f(\lambda_j(\widehat{\Sigma}))$.

Before resampling, first compute estimates $\widehat{\kappa}$ and $\widehat{\Lambda}$.

Algorithm. (Spectral Bootstrap)

For $b = 1, \ldots, B$:

- Generate a random matrix $Z^* \in \mathbb{R}^{n \times p}$ whose entries Z^*_{ij} are drawn i.i.d. from Pearson$(0, 1, 0, \widehat{\kappa})$. (Recall $X_i = \Sigma^{1/2} Z_i$)

Proposed method: Spectral Bootstrap

Goal. Approximate the distribution of $T = \frac{1}{p}\sum_{j=1}^{p} f(\lambda_j(\widehat{\Sigma}))$.

Before resampling, first compute estimates $\widehat{\kappa}$ and $\widehat{\Lambda}$.

Algorithm. (Spectral Bootstrap)

For $b = 1, \ldots, B$:

- Generate a random matrix $Z^* \in \mathbb{R}^{n \times p}$ whose entries Z^*_{ij} are drawn i.i.d. from Pearson$(0, 1, 0, \widehat{\kappa})$. (Recall $X_i = \Sigma^{1/2} Z_i$)
- Compute $\widehat{\Sigma}^* = \frac{1}{n}\widehat{\Lambda}^{1/2}(Z^{*\top} Z^*)\widehat{\Lambda}^{1/2}$. (Note $\widehat{\Sigma} = \frac{1}{n}\Sigma^{1/2} Z^\top Z \Sigma^{1/2}$)

Proposed method: Spectral Bootstrap

Goal. Approximate the distribution of $T = \frac{1}{p}\sum_{j=1}^{p} f(\lambda_j(\widehat{\Sigma}))$.

Before resampling, first compute estimates $\widehat{\kappa}$ and $\widehat{\Lambda}$.

Algorithm. (Spectral Bootstrap)

For $b = 1, \ldots, B$:

- Generate a random matrix $Z^* \in \mathbb{R}^{n \times p}$ whose entries Z^*_{ij} are drawn i.i.d. from Pearson$(0, 1, 0, \widehat{\kappa})$. (Recall $X_i = \Sigma^{1/2} Z_i$)
- Compute $\widehat{\Sigma}^* = \frac{1}{n}\widehat{\Lambda}^{1/2}(Z^{*\top} Z^*)\widehat{\Lambda}^{1/2}$. (Note $\widehat{\Sigma} = \frac{1}{n}\Sigma^{1/2} Z^\top Z \Sigma^{1/2}$)
- Compute the eigenvalues of $\widehat{\Sigma}^*$, and denote them by $(\lambda_1^*, \ldots, \lambda_p^*)$.

Proposed method: Spectral Bootstrap

Goal. Approximate the distribution of $T = \frac{1}{p}\sum_{j=1}^{p} f(\lambda_j(\widehat{\Sigma}))$.

Before resampling, first compute estimates $\widehat{\kappa}$ and $\widehat{\Lambda}$.

Algorithm. (Spectral Bootstrap)

For $b = 1, \ldots, B$:

- Generate a random matrix $Z^* \in \mathbb{R}^{n \times p}$ whose entries Z^*_{ij} are drawn i.i.d. from Pearson$(0, 1, 0, \widehat{\kappa})$. (Recall $X_i = \Sigma^{1/2} Z_i$)
- Compute $\widehat{\Sigma}^* = \frac{1}{n}\widehat{\Lambda}^{1/2}(Z^{*\top} Z^*)\widehat{\Lambda}^{1/2}$. (Note $\widehat{\Sigma} = \frac{1}{n}\Sigma^{1/2} Z^\top Z \Sigma^{1/2}$)
- Compute the eigenvalues of $\widehat{\Sigma}^*$, and denote them by $(\lambda_1^*, \ldots, \lambda_p^*)$.
- Compute the statistic, $T_b^* = \frac{1}{p}\sum_{j=1}^{p} f(\lambda_j^*)$

Return: the empirical distribution of the values $T_1^*, \ldots, T_B^*$.

Generalizing to other spectral statistics

Let $\psi : \mathbb{R}^p \to \mathbb{R}$ be a generic (non-linear) function, and consider the statistic

$$T = \psi(\lambda_1(\widehat{\Sigma}), \ldots, \lambda_p(\widehat{\Sigma})).$$

Key point: To bootstrap T, we only need change the last step.

(This is a distinct benefit of the bootstrap in relation to formulas.)

Generalizing to other spectral statistics

Let $\psi : \mathbb{R}^p \to \mathbb{R}$ be a generic (non-linear) function, and consider the statistic

$$T = \psi(\lambda_1(\widehat{\Sigma}), \ldots, \lambda_p(\widehat{\Sigma})).$$

Key point: To bootstrap T, we only need change the last step.

(This is a distinct benefit of the bootstrap in relation to formulas.)

For $b = 1, \ldots, B$:

- ...
- Compute the eigenvalues of $\widehat{\Sigma}^*$, and denote them by $(\lambda_1^*, \ldots, \lambda_p^*)$.
- Compute the statistic, $T_b^* = \psi(\lambda_1^*, \ldots, \lambda_p^*)$

Return: the empirical distribution of the values $T_1^*, \ldots, T_B^*$.

Estimating kurtosis

Recall $\kappa = \mathbb{E}[Z_{ij}^4]$, and all row vectors satisfy $X_i = \Sigma^{1/2} Z_i$.

Our estimate of κ is based on a general formula for the variance of a quadratic form

$$\kappa = 3 + \frac{\mathrm{Var}(\|X_1\|_2^2) - 2\|\Sigma\|_F^2}{\sum_{j=1}^p \sigma_j^4}.$$

Estimating kurtosis

Recall $\kappa = \mathbb{E}[Z_{ij}^4]$, and all row vectors satisfy $X_i = \Sigma^{1/2} Z_i$.

Our estimate of κ is based on a general formula for the variance of a quadratic form

$$\kappa = 3 + \frac{\mathrm{Var}(\|X_1\|_2^2) - 2\|\Sigma\|_F^2}{\sum_{j=1}^p \sigma_j^4}.$$

All the quantities on the right side have ratio-consistent estimators when $p \asymp n$ under standard conditions.

Estimating kurtosis

Recall $\kappa = \mathbb{E}[Z_{ij}^4]$, and all row vectors satisfy $X_i = \Sigma^{1/2} Z_i$.

Our estimate of κ is based on a general formula for the variance of a quadratic form

$$\kappa = 3 + \frac{\mathrm{Var}(\|X_1\|_2^2) - 2\|\Sigma\|_F^2}{\sum_{j=1}^p \sigma_j^4}.$$

All the quantities on the right side have ratio-consistent estimators when $p \asymp n$ under standard conditions.

The estimation of $\|\Sigma\|_F^2$ was handled previously in Bai and Saranadasa (1996), but it seems that a consistent estimate for κ has not been available in the high-dimensional setting.

Estimating kurtosis

Recall $\kappa = \mathbb{E}[Z_{ij}^4]$, and all row vectors satisfy $X_i = \Sigma^{1/2} Z_i$.

Our estimate of κ is based on a general formula for the variance of a quadratic form

$$\kappa = 3 + \frac{\mathrm{Var}(\|X_1\|_2^2) - 2\|\Sigma\|_F^2}{\sum_{j=1}^p \sigma_j^4}.$$

All the quantities on the right side have ratio-consistent estimators when $p \asymp n$ under standard conditions.

The estimation of $\|\Sigma\|_F^2$ was handled previously in Bai and Saranadasa (1996), but it seems that a consistent estimate for κ has not been available in the high-dimensional setting.

An estimate of κ may also be of independent interest as a diagnostic tool for checking if data are approximately Gaussian.

Estimating eigenvalues

We use the QUEST method (Ledoit and Wolf, 2015).

For bootstrapping LSS, the essential issue is to use eigenvalue estimates $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p$ that lead to a consistent estimate of the population LSD.

Let H_p denote the spectral the distribution function associated with $\lambda_1(\Sigma), \ldots, \lambda_p(\Sigma)$,

$$H_p(t) = \frac{1}{p} \sum_{j=1}^{p} 1\{\lambda_j(\Sigma) \leq t\}.$$

Then, an estimate $\widehat{H}_p$ may be formed by taking the estimates $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p$ as the quantiles.

Estimating eigenvalues

We use the QUEST method (Ledoit and Wolf, 2015).

For bootstrapping LSS, the essential issue is to use eigenvalue estimates $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p$ that lead to a consistent estimate of the population LSD.

Let H_p denote the spectral the distribution function associated with $\lambda_1(\Sigma), \ldots, \lambda_p(\Sigma)$,

$$H_p(t) = \frac{1}{p} \sum_{j=1}^{p} 1\{\lambda_j(\Sigma) \leq t\}.$$

Then, an estimate $\widehat{H}_p$ may be formed by taking the estimates $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p$ as the quantiles.

Consistency. If there is a population LSD H satisfying

$$H_p \Rightarrow H,$$

then the QUEST estimator $\widehat{H}_p$ satisfies the following limit under standard assumptions,

$$\widehat{H}_p \Rightarrow H \quad \text{almost surely.}$$

Estimating eigenvalues

We use the QUEST method (Ledoit and Wolf, 2015).

For bootstrapping LSS, the essential issue is to use eigenvalue estimates $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p$ that lead to a consistent estimate of the population LSD.

Let H_p denote the spectral the distribution function associated with $\lambda_1(\Sigma), \ldots, \lambda_p(\Sigma)$,

$$H_p(t) = \frac{1}{p} \sum_{j=1}^{p} 1\{\lambda_j(\Sigma) \leq t\}.$$

Then, an estimate $\widehat{H}_p$ may be formed by taking the estimates $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p$ as the quantiles.

Consistency. If there is a population LSD H satisfying

$$H_p \Rightarrow H,$$

then the QUEST estimator $\widehat{H}_p$ satisfies the following limit under standard assumptions,

$$\widehat{H}_p \Rightarrow H \quad \text{almost surely.}$$

Note: Other spectrum estimation methods are also compatible with the proposed bootstrap — provided that the above limit holds.

Main result: bootstrap consistency

Assumptions in brief:

- $p/n \to c \in (0, \infty)$
- $\lambda_p(\Sigma)$ and $\lambda_1(\Sigma)$ bounded away from 0 and ∞
- Finite 8th moment: $\mathbb{E}[Z_{11}^8] < \infty$.
- $H_p \Rightarrow H$.
- Asymptotic "regularity" of population eigenvectors (more later)

Main result: bootstrap consistency

Assumptions in brief:

- $p/n \to c \in (0, \infty)$
- $\lambda_p(\Sigma)$ and $\lambda_1(\Sigma)$ bounded away from 0 and ∞
- Finite 8th moment: $\mathbb{E}[Z_{11}^8] < \infty$.
- $H_p \Rightarrow H$.
- Asymptotic "regularity" of population eigenvectors (more later)

Theorem 1 (LBA 2019)

Let d_{LP} denote the Lévy-Prohorov metric. Then, under the stated assumptions, the following limit holds as $(n, p) \to \infty$,

$$d_{\mathrm{LP}}\Big(\mathcal{L}\big(p(T^* - \mathbb{E}[T^*|X])|X\big)\,,\,\mathcal{L}\big(p(T - \mathbb{E}[T])\big)\Big) \to 0 \quad \textit{in probability.}$$

High-level comments on the proof

The proof draws substantially from the arguments in Najim and Yao (2016), based on the Helffer-Sjöstrand formula. This formula allows for the following distributional approximation as $(n, p) \to \infty$,

$$\mathcal{L}\big\{p(T - \mathbb{E}[T])\big\} \;\approx\; \mathcal{L}\{\phi_f(G_n)\},$$

where ϕ_f is a linear functional, and $G_n = G_n(z)$ is a centered Gaussian process that arises from the empirical Stieltjes transform $\frac{1}{p}\mathrm{tr}\big((\widehat{\Sigma} - zI_p)^{-1}\big)$.

High-level comments on the proof

The proof draws substantially from the arguments in Najim and Yao (2016), based on the Helffer-Sjöstrand formula. This formula allows for the following distributional approximation as $(n, p) \to \infty$,

$$\mathcal{L}\{p(T - \mathbb{E}[T])\} \approx \mathcal{L}\{\phi_f(G_n)\},$$

where ϕ_f is a linear functional, and $G_n = G_n(z)$ is a centered Gaussian process that arises from the empirical Stieltjes transform $\frac{1}{p}\mathrm{tr}\big((\widehat{\Sigma} - zI_p)^{-1}\big)$.

In the "bootstrap world", there is a corresponding conditional approximation,

$$\mathcal{L}\{p(T^* - \mathbb{E}[T^*|X])|X\} \approx \mathcal{L}\{\phi_f(G_n^*)|X\}.$$

High-level comments on the proof

The proof draws substantially from the arguments in Najim and Yao (2016), based on the Helffer-Sjöstrand formula. This formula allows for the following distributional approximation as $(n, p) \to \infty$,

$$\mathcal{L}\{p(T - \mathbb{E}[T])\} \;\approx\; \mathcal{L}\{\phi_f(G_n)\},$$

where ϕ_f is a linear functional, and $G_n = G_n(z)$ is a centered Gaussian process that arises from the empirical Stieltjes transform $\frac{1}{p}\mathrm{tr}\big((\widehat{\Sigma} - zI_p)^{-1}\big)$.

In the "bootstrap world", there is a corresponding conditional approximation,

$$\mathcal{L}\{p(T^* - \mathbb{E}[T^*|X])|X\} \;\approx\; \mathcal{L}\{\phi_f(G_n^*)|X\}.$$

Finally, the consistency of $\widehat{H}$ and $\widehat{\kappa}$ are used to obtain the conditional approximation

$$\mathcal{L}\{\phi_f(G_n^*)|X\} \;\approx\; \mathcal{L}\{\phi_f(G_n)\},$$

by comparing the covariance functions of G_n^* and G_n.

Regularity of eigenvectors

Let U be the matrix of eigenvectors of Σ, and consider the non-random quantity

$$K_p(z_1, z_2) := \frac{1}{p} \sum_{j=1}^{p} \left[U D_n(z_1) U^\top \right]_{jj} \left[U D_n(z_2) U^\top \right]_{jj}$$

where $z_1, z_2 \in \mathbb{C} \setminus \mathbb{R}$, and $D_n(\cdot) \in \mathbb{C}^{p \times p}$ is a diagonal matrix that only depends on the spectrum of Σ.

Regularity of eigenvectors

Let U be the matrix of eigenvectors of Σ, and consider the non-random quantity

$$K_p(z_1, z_2) := \frac{1}{p} \sum_{j=1}^{p} \left[U D_n(z_1) U^\top \right]_{jj} \left[U D_n(z_2) U^\top \right]_{jj}$$

where $z_1, z_2 \in \mathbb{C} \setminus \mathbb{R}$, and $D_n(\cdot) \in \mathbb{C}^{p \times p}$ is a diagonal matrix that only depends on the spectrum of Σ.

Also let

$$K_p'(z_1, z_2) := \frac{1}{p} \sum_{j=1}^{p} \left[D_n(z_1) \right]_{jj} \left[D_n(z_2) \right]_{jj}.$$

Regularity of eigenvectors

Let U be the matrix of eigenvectors of Σ, and consider the non-random quantity

$$K_p(z_1, z_2) := \frac{1}{p} \sum_{j=1}^{p} \left[U D_n(z_1) U^\top \right]_{jj} \left[U D_n(z_2) U^\top \right]_{jj}$$

where $z_1, z_2 \in \mathbb{C} \setminus \mathbb{R}$, and $D_n(\cdot) \in \mathbb{C}^{p \times p}$ is a diagonal matrix that only depends on the spectrum of Σ.

Also let

$$K'_p(z_1, z_2) := \frac{1}{p} \sum_{j=1}^{p} \left[D_n(z_1) \right]_{jj} \left[D_n(z_2) \right]_{jj}.$$

Regularity of eigenvectors. We say that the eigenvectors of Σ are regular if for any $z_1, z_2 \in \mathbb{C} \setminus \mathbb{R}$, as $(n, p) \to \infty$

$$K_p(z_1, z_2) = K'_p(z_1, z_2) + o(1).$$

Regularity of eigenvectors

Let U be the matrix of eigenvectors of Σ, and consider the non-random quantity

$$K_p(z_1, z_2) := \frac{1}{p} \sum_{j=1}^{p} \left[U D_n(z_1) U^\top \right]_{jj} \left[U D_n(z_2) U^\top \right]_{jj}$$

where $z_1, z_2 \in \mathbb{C} \setminus \mathbb{R}$, and $D_n(\cdot) \in \mathbb{C}^{p \times p}$ is a diagonal matrix that only depends on the spectrum of Σ.

Also let

$$K'_p(z_1, z_2) := \frac{1}{p} \sum_{j=1}^{p} \left[D_n(z_1) \right]_{jj} \left[D_n(z_2) \right]_{jj}.$$

Regularity of eigenvectors. We say that the eigenvectors of Σ are regular if for any $z_1, z_2 \in \mathbb{C} \setminus \mathbb{R}$, as $(n, p) \to \infty$

$$K_p(z_1, z_2) = K'_p(z_1, z_2) + o(1).$$

Remarks. The papers Pan and Zhou (2008) and Najim and Yao (2016) show that unless $\kappa = 3$, a limit law for standardized LSS may not exist unless U has some regularity. Nevertheless, empirical results suggest that such regularity may be "typical".

Regularity of eigenvectors (cont.)

Example 1. (Rank k perturbations, $k \to \infty$).

Suppose $\lambda_1(\Sigma)$ is bounded away from ∞, and let Λ be otherwise unrestricted.

If U is of the form

$$U = I_{p \times p} - 2\Pi,$$

where Π is any orthogonal projection matrix of rank k, and $k = o(p)$, then the eigenvectors are regular.

This is a fairly substantial perturbation from the diagonal case.

Regularity of eigenvectors (cont.)

Example 2. (Spiked covariance models).

Suppose Λ is of the form

$$\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k, 1, \ldots, 1)$$

where $k = o(p)$, and $\lambda_1 = \lambda_1(\Sigma)$ is bounded away from infinity.

Then, any choice of U will be regular.

Simulations for LSS ($\kappa > 3$)

Recall $X_i = \Sigma^{1/2} Z_i$.

- Z_i generated with standardized i.i.d. t-dist (df=20)
- kurtosis $\kappa \approx 3.4$
- decaying population spectrum is $\lambda_j = j^{-1/2}$
- population eigenvectors uniformly drawn from Haar measure

We tabulate the std. dev., 95th percentile, and 99th percentile of $p(T - \mathbb{E}[T])$.

Simulations for LSS ($\kappa > 3$)

Recall $X_i = \Sigma^{1/2} Z_i$.

- Z_i generated with standardized i.i.d. t-dist (df=20)
- kurtosis $\kappa \approx 3.4$
- decaying population spectrum is $\lambda_j = j^{-1/2}$
- population eigenvectors uniformly drawn from Haar measure

We tabulate the std. dev., 95th percentile, and 99th percentile of $p(T - \mathbb{E}[T])$.

	$f(x) = x$						$f(x) = \log(x)$					
(n,p)	std. dev.		95th		99th		std. dev.		95th		99th	
(500,200)	0.16		0.27		0.36		1.07		1.82		2.41	
	0.17	**(0.01)**	**0.28**	**(0.03)**	**0.39**	**(0.06)**	**1.08**	**(0.08)**	**1.76**	**(0.20)**	**2.51**	**(0.35)**
(500,400)	0.18		0.29		0.41		4.41		7.03		9.77	
	0.18	**(0.02)**	**0.30**	**(0.04)**	**0.42**	**(0.06)**	**4.27**	**(0.33)**	**6.72**	**(0.70)**	**9.29**	**(1.18)**
(500,600)	0.17		0.29		0.40							
	0.18	**(0.02)**	**0.30**	**(0.04)**	**0.43**	**(0.07)**	-		-		-	

Simulations for LSS ($\kappa < 3$)

Recall $X_i = \Sigma^{1/2} Z_i$.

- Z_i generated with standardized i.i.d. Beta(6,6)
- kurtosis $\kappa = 2.6$
- decaying population spectrum is $\lambda_j = j^{-1/2}$
- population eigenvectors uniformly drawn from Haar measure

	$f(x) = x$						$f(x) = \log(x)$					
(n,p)	std. dev.		95th		99th		std. dev.		95th		99th	
(500,200)	0.14		0.23		0.33		0.93		1.51		1.92	
	0.14	**(0.01)**	**0.23**	**(0.03)**	**0.32**	**(0.05)**	**0.93**	**(0.08)**	**1.52**	**(0.17)**	**2.15**	**(0.31)**
(500,400)	0.15		0.25		0.34		1.65		2.64		3.64	
	0.14	**(0.01)**	**0.24**	**(0.03)**	**0.34**	**(0.05)**	**1.70**	**(0.13)**	**2.81**	**(0.31)**	**3.97**	**(0.56)**
(500,600)	0.16		0.26		0.34							
	0.15	**(0.01)**	**0.25**	**(0.03)**	**0.35**	**(0.05)**	-		-		-	

What about other spectral statistics?

In principle, the proposed method can be applied to any spectral statistic.

Below, we present some simulation results for some *non-linear* statistics:

- $T_{\text{max}} = \lambda_1(\widehat{\Sigma})$.
- $T_{\text{gap}} = \lambda_1(\widehat{\Sigma}) - \lambda_2(\widehat{\Sigma})$

Simulations for non-linear statistics ($\kappa < 3$)

Recall $X_i = \Sigma^{1/2} Z_i$.

- Z_i generated with standardized i.i.d. Beta(6,6)
- kurtosis $\kappa = 2.6$
- decaying population spectrum is $\lambda_j = j^{-1/2}$
- population eigenvectors uniformly drawn from Haar measure

Simulations for non-linear statistics ($\kappa < 3$)

Recall $X_i = \Sigma^{1/2} Z_i$.

- Z_i generated with standardized i.i.d. Beta(6,6)
- kurtosis $\kappa = 2.6$
- decaying population spectrum is $\lambda_j = j^{-1/2}$
- population eigenvectors uniformly drawn from Haar measure

	$T_{\max} - \mathbb{E}[T_{\max}]$						$T_{\text{gap}} - \mathbb{E}[T_{\text{gap}}]$					
(n,p)	std. dev.		95th		99th		std. dev.		95th		99th	
(500,200)	0.06		0.11		0.15		0.08		0.13		0.17	
	0.06	**(0.01)**	**0.09**	**(0.01)**	**0.13**	**(0.02)**	**0.07**	**(0.01)**	**0.11**	**(0.01)**	**0.16**	**(0.03)**
(500,400)	0.06		0.10		0.15		0.08		0.13		0.18	
	0.06	**(0.01)**	**0.09**	**(0.01)**	**0.13**	**(0.02)**	**0.07**	**(0.01)**	**0.11**	**(0.01)**	**0.16**	**(0.03)**
(500,600)	0.06		0.11		0.14		0.07		0.13		0.17	
	0.06	**(0.01)**	**0.09**	**(0.01)**	**0.13**	**(0.02)**	**0.07**	**(0.01)**	**0.11**	**(0.02)**	**0.16**	**(0.03)**

Simulations for non-linear statistics ($\kappa > 3$)

Recall $X_i = \Sigma^{1/2} Z_i$.

- Z_i generated with standardized i.i.d. t-dist (df=20)
- kurtosis $\kappa \approx 3.4$
- decaying population spectrum is $\lambda_j = j^{-1/2}$
- population eigenvectors uniformly drawn from Haar measure

	$T_{\max} - \mathbb{E}[T_{\max}]$						$T_{\text{gap}} - \mathbb{E}[T_{\text{gap}}]$					
(n,p)	std. dev.		95th		99th		std. dev.		95th		99th	
(500,200)	0.06		0.10		0.15		0.07		0.12		0.17	
	0.07	**(0.01)**	**0.11**	**(0.02)**	**0.17**	**(0.03)**	**0.08**	**(0.01)**	**0.13**	**(0.02)**	**0.19**	**(0.03)**
(500,400)	0.06		0.10		0.14		0.07		0.13		0.17	
	0.07	**(0.01)**	**0.11**	**(0.02)**	**0.17**	**(0.03)**	**0.08**	**(0.01)**	**0.13**	**(0.02)**	**0.19**	**(0.03)**
(500,600)	0.06		0.11		0.16		0.08		0.13		0.18	
	0.07	**(0.01)**	**0.11**	**(0.02)**	**0.16**	**(0.03)**	**0.08**	**(0.01)**	**0.13**	**(0.02)**	**0.19**	**(0.03)**

Part I summary: Bootstrap for spectral statistics

- LSS are a general class of statistics for which bootstrapping can succeed in high dimensions.
- This offers general-purpose way to approximate the laws of LSS without relying on asymptotic formulas.
- The method is akin to the parametric bootstrap — using the fact that spectral statistics may depend on relatively few parameters of the full data-generating distribution.
- Numerical results are encouraging.
- The method appears to extend to some non-linear spectral statistics — for which asymptotic formulas are often unavailable.
- Further work on non-linear statistics is underway ...

Part II: Bootstrapping the operator norm in high dimensions: Error estimation for covariance matrices and sketching

Motivations and background

Let $X_1, \ldots, X_n$ are centered i.i.d. observations in $\mathbb{R}^p$ with $\Sigma = \mathbb{E}[X_1 X_1^\top]$, and let $\widehat{\Sigma}$ denote the sample covariance matrix.

When $p \asymp n$ or $p \gg n$, there is a large literature on the problem of deriving high-probabilty non-asymptotic bounds on the operator norm error

$$T = \sqrt{n}\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}.$$

Motivations and background

Let $X_1, \ldots, X_n$ are centered i.i.d. observations in $\mathbb{R}^p$ with $\Sigma = \mathbb{E}[X_1 X_1^\top]$, and let $\widehat{\Sigma}$ denote the sample covariance matrix.

When $p \asymp n$ or $p \gg n$, there is a large literature on the problem of deriving high-probabilty non-asymptotic bounds on the operator norm error

$$T = \sqrt{n}\|\widehat{\Sigma} - \Sigma\|_{\text{op}}.$$

However, such bounds are typically only given up to unspecified constants.

Motivations and background

Let $X_1, \ldots, X_n$ are centered i.i.d. observations in $\mathbb{R}^p$ with $\Sigma = \mathbb{E}[X_1 X_1^\top]$, and let $\widehat{\Sigma}$ denote the sample covariance matrix.

When $p \asymp n$ or $p \gg n$, there is a large literature on the problem of deriving high-probabilty non-asymptotic bounds on the operator norm error

$$T = \sqrt{n}\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}.$$

However, such bounds are typically only given up to unspecified constants.

In order to solve practical inference problems, such as constructing numerical error bounds for $\widehat{\Sigma}$, or confidence regions for Σ, we need to approximate the distribution of T.

Motivations and background

Let $X_1, \ldots, X_n$ are centered i.i.d. observations in $\mathbb{R}^p$ with $\Sigma = \mathbb{E}[X_1 X_1^\top]$, and let $\widehat{\Sigma}$ denote the sample covariance matrix.

When $p \asymp n$ or $p \gg n$, there is a large literature on the problem of deriving high-probabilty non-asymptotic bounds on the operator norm error

$$T = \sqrt{n}\|\widehat{\Sigma} - \Sigma\|_{\text{op}}.$$

However, such bounds are typically only given up to unspecified constants.

In order to solve practical inference problems, such as constructing numerical error bounds for $\widehat{\Sigma}$, or confidence regions for Σ, we need to approximate the distribution of T.

The recent work of Han, Xu, and Zhou (2018) has explored bootstrap approximations for $\sup_{\|u\|_2 \le 1, \|u\|_0 \le s} \sqrt{n}\left|u^\top(\widehat{\Sigma} - \Sigma)u\right|$ when $s \ll n$, but beyond this, not much is known about bootstrapping T in high dimensions.

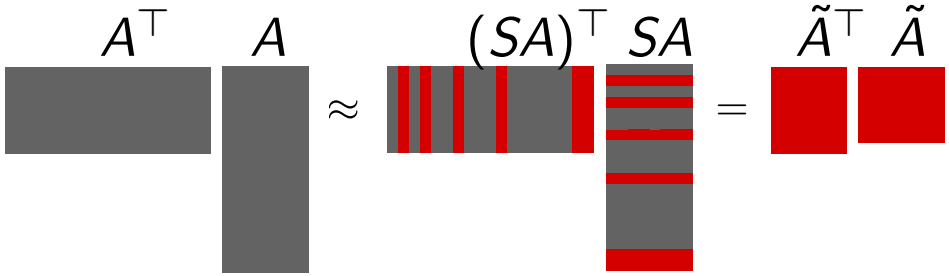
Further motivations (RandNLA and sketching)

Randomized numerical linear algebra (RandNLA) or "matrix sketching" uses randomization to accelerate numerical linear algebra on huge matrices.

Further motivations (RandNLA and sketching)

Randomized numerical linear algebra (RandNLA) or "matrix sketching" uses randomization to accelerate numerical linear algebra on huge matrices.

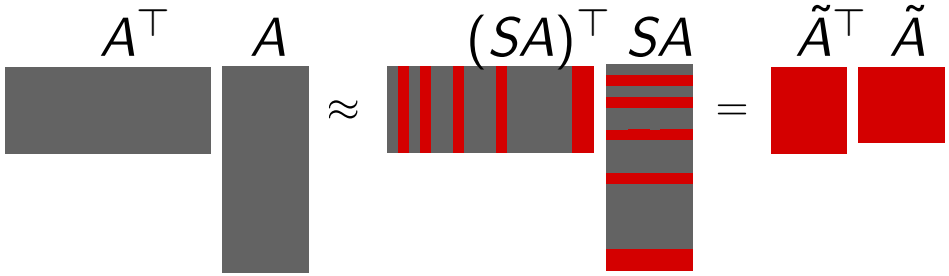
For instance, if A is a very tall (deterministic) matrix, one may try to do computations with a shorter "sketch" $\tilde{A} = SA$, where S is a random short matrix satisfying $\mathbb{E}[S^\top S] = I$.



Further motivations (RandNLA and sketching)

Randomized numerical linear algebra (RandNLA) or "matrix sketching" uses randomization to accelerate numerical linear algebra on huge matrices.

For instance, if A is a very tall (deterministic) matrix, one may try to do computations with a shorter "sketch" $\tilde{A} = SA$, where S is a random short matrix satisfying $\mathbb{E}[S^\top S] = I$.



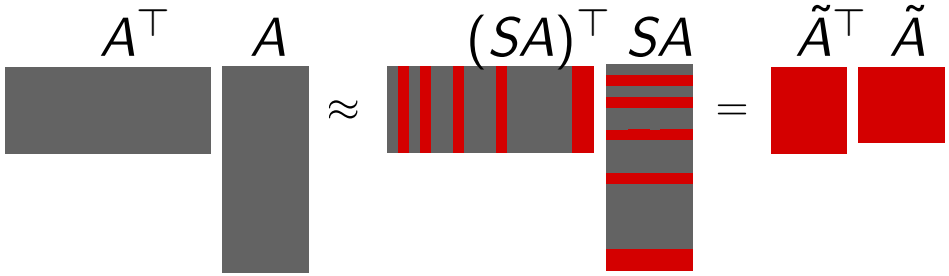
In practice, it is necessary to estimate the algorithmic error

$$\|A^\top S^\top SA - A^\top A\|_{\mathrm{op}}.$$

Further motivations (RandNLA and sketching)

Randomized numerical linear algebra (RandNLA) or "matrix sketching" uses randomization to accelerate numerical linear algebra on huge matrices.

For instance, if A is a very tall (deterministic) matrix, one may try to do computations with a shorter "sketch" $\tilde{A} = SA$, where S is a random short matrix satisfying $\mathbb{E}[S^\top S] = I$.



In practice, it is necessary to estimate the algorithmic error

$$\|A^\top S^\top SA - A^\top A\|_{\text{op}}.$$

However, most theoretical work has focused on bounds that hold up to constants, and only a handful of papers have addressed error estimation in this context:

(e.g. Liberty et al., (2007), Woolfe et al., (2008), Halko, Martinsson and Tropp (2011), Lopes, Wang and Mahoney, (2017) (2018)).

A model with spectrum decay

Suppose $X \in \mathbb{R}^{n \times p}$ is a data matrix with rows generated as

$$X_i = \Sigma^{1/2} Z_i$$

where the vectors $Z_1, \ldots, Z_n \in \mathbb{R}^p$ are i.i.d. and have i.i.d. entries with $\mathbb{E}[Z_{11}] = 0$, $\mathbb{E}[Z_{11}^2] = 1$, $\mathbb{E}[Z_{11}^4] > 1$, and $\|Z_{11}\|_{\psi_2} \leq c_0$ for some constant $c_0 > 0$ not depending on n.

A model with spectrum decay

Suppose $X \in \mathbb{R}^{n \times p}$ is a data matrix with rows generated as

$$X_i = \Sigma^{1/2} Z_i$$

where the vectors $Z_1, \ldots, Z_n \in \mathbb{R}^p$ are i.i.d. and have i.i.d. entries with $\mathbb{E}[Z_{11}] = 0$, $\mathbb{E}[Z_{11}^2] = 1$, $\mathbb{E}[Z_{11}^4] > 1$, and $\|Z_{11}\|_{\psi_2} \leq c_0$ for some constant $c_0 > 0$ not depending on n.

There are constants $\beta > 1/2$ and $c_1, c_2 > 0$, not depending on n, such that for each $j \in \{1, \ldots, p\}$,

$$c_1 j^{-2\beta} \leq \lambda_j(\Sigma) \leq c_2 j^{-2\beta}.$$

Main result (rate of bootstrap approximation)

Recall that we aim to approximate the law of $T = \sqrt{n}\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$.

Let $(X_1^*, \ldots, X_n^*)$ be drawn with replacement from $(X_1, \ldots, X_n)$, and let

$$\widehat{\Sigma}^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*(X_i^*)^\top.$$

Also, define the bootstrapped statistic $T^* = \sqrt{n}\|\widehat{\Sigma}^* - \widehat{\Sigma}\|_{\mathrm{op}}$.

Main result (rate of bootstrap approximation)

Recall that we aim to approximate the law of $T = \sqrt{n}\|\widehat{\Sigma} - \Sigma\|_{\text{op}}$.

Let $(X_1^*, \ldots, X_n^*)$ be drawn with replacement from $(X_1, \ldots, X_n)$, and let

$$\widehat{\Sigma}^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*(X_i^*)^\top.$$

Also, define the bootstrapped statistic $T^* = \sqrt{n}\|\widehat{\Sigma}^* - \widehat{\Sigma}\|_{\text{op}}$.

Theorem 2 (LEM 2019)

Let d_{K} denote the Kolmogorov metric. Then, under the stated model, there is a constant $c > 0$ not depending on n such that the event

$$d_{\text{K}}\big(\mathcal{L}(T)\,,\, \mathcal{L}(T^*|X)\big) \;\leq\; c\, n^{-\frac{\beta - 1/2}{6\beta + 4}}\, \log(n)^c$$

occurs with probability at least $1 - \frac{c}{n}$.

Connections to other works

In recent years, there have been several influential works by Chernozhukov, Chetverikov, and Kato (CCK) (2013), (2014), (2017) on bootstrapping maxima of empirical processes

$$\sup_{f \in \mathscr{F}} \mathbb{G}_n(f)$$

where $\mathbb{G}_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(Z_i) - \mathbb{E}[f(Z_i)]$.

Connections to other works

In recent years, there have been several influential works by Chernozhukov, Chetverikov, and Kato (CCK) (2013), (2014), (2017) on bootstrapping maxima of empirical processes

$$\sup_{f \in \mathscr{F}} \mathbb{G}_n(f)$$

where $\mathbb{G}_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z_i)]$.

The statistic $T = \sqrt{n}\|\widehat{\Sigma} - \Sigma\|_{\text{op}}$ can be represented in this form by taking

$$f(Z_i) = \pm\langle v, Z_i\rangle^2,$$

with $v = \Sigma^{1/2}u$ for some unit vector u (so $\mathscr{F}$ corresponds to a signed ellipsoid).

Connections to other works

In recent years, there have been several influential works by Chernozhukov, Chetverikov, and Kato (CCK) (2013), (2014), (2017) on bootstrapping maxima of empirical processes

$$\sup_{f \in \mathscr{F}} \mathbb{G}_n(f)$$

where $\mathbb{G}_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z_i)]$.

The statistic $T = \sqrt{n}\|\widehat{\Sigma} - \Sigma\|_{\text{op}}$ can be represented in this form by taking

$$f(Z_i) = \pm\langle v, Z_i\rangle^2,$$

with $v = \Sigma^{1/2}u$ for some unit vector u (so $\mathscr{F}$ corresponds to a signed ellipsoid).

However, the results of CCK are not directly applicable to T, because such results typically involve a "minimum variance condition", such as

$$\inf_{f \in \mathscr{F}} \text{var}(\mathbb{G}_n(f)) \geq c$$

for some $c > 0$ not depending on n, which fails for the ellipsoidal index set.

Why $\beta - 1/2$?

Recall the rate of bootstrap approximation

$$n^{-\frac{\beta-1/2}{6\beta+4}} \log(n)^c.$$

The role of $\beta - 1/2$ can be understood in terms of the error that comes from discretizing $\mathscr{F}$,

$$\Delta_n(\epsilon) := \sup_{\text{dist}(f,\tilde{f})\leq\epsilon} |\mathbb{G}_n(f) - \mathbb{G}_n(\tilde{f})|.$$

Why $\beta - 1/2$?

Recall the rate of bootstrap approximation

$$n^{-\frac{\beta-1/2}{6\beta+4}} \log(n)^c.$$

The role of $\beta - 1/2$ can be understood in terms of the error that comes from discretizing $\mathscr{F}$,

$$\Delta_n(\epsilon) := \sup_{\text{dist}(f,\tilde{f})\leq\epsilon} |\mathbb{G}_n(f) - \mathbb{G}_n(\tilde{f})|.$$

In order for discrete approximation to work, we should have $\mathbb{E}[\Delta_n(\epsilon)] \to 0$ as $\epsilon \to 0$.

This imposes an implicit constraint on the complexity of $\mathscr{F}$ (i.e. the complexity Σ).

Why $\beta - 1/2$?

Recall the rate of bootstrap approximation

$$n^{-\frac{\beta-1/2}{6\beta+4}} \log(n)^c.$$

The role of $\beta - 1/2$ can be understood in terms of the error that comes from discretizing $\mathscr{F}$,

$$\Delta_n(\epsilon) := \sup_{\mathrm{dist}(f,\tilde{f})\leq\epsilon} |\mathbb{G}_n(f) - \mathbb{G}_n(\tilde{f})|.$$

In order for discrete approximation to work, we should have $\mathbb{E}[\Delta_n(\epsilon)] \to 0$ as $\epsilon \to 0$.

This imposes an implicit constraint on the complexity of $\mathscr{F}$ (i.e. the complexity Σ).

If we consider a simpler situation where $\mathbb{G}_n(f)$ is replaced by a linear Gaussian process indexed by the same $\mathscr{F}$, say

$$\mathbb{G}'_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \langle v, \zeta_i \rangle,$$

with $\zeta_1, \ldots, \zeta_n$ i.i.d. $N(0, I)$, then it follows from classical results that the associated discretization error satisfies the lower bound

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

Hence, the condition $\beta - 1/2 > 0$ is needed even in the linear Gaussian case.

A new general-purpose error bound

To analyze the bootstrap, it was necessary to use dimension-free high-probability upper bounds on $\|\widehat{\Sigma} - \Sigma\|_{\text{op}}$.

A new general-purpose error bound

To analyze the bootstrap, it was necessary to use dimension-free high-probability upper bounds on $\|\widehat{\Sigma} - \Sigma\|_{\text{op}}$.

Existing dimension-free bounds generally require either $\|X_i\|_2 \leq c$ almost surely, or $\|\langle u, X_i \rangle\|_{\psi_2} \asymp \|\langle u, X_i \rangle\|_{L_2}$ for all $\|u\|_2 = 1$.

(e.g. Rudelson and Vershynin, (2007), Oliveira, (2010), Hsu, Kakade and Zhang, (2012), Koltchinskii and Lounici, (2017), Minsker, (2017))

A new general-purpose error bound

To analyze the bootstrap, it was necessary to use dimension-free high-probability upper bounds on $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$.

Existing dimension-free bounds generally require either $\|X_i\|_2 \leq c$ almost surely, or $\|\langle u, X_i \rangle\|_{\psi_2} \asymp \|\langle u, X_i \rangle\|_{L_2}$ for all $\|u\|_2 = 1$.

(e.g. Rudelson and Vershynin, (2007), Oliveira, (2010), Hsu, Kakade and Zhang, (2012), Koltchinskii and Lounici, (2017), Minsker, (2017))

However, the ℓ_2-boundedness condition is often restrictive, while the ψ_2-L_2 equivalence condition is not well-suited to the discrete distributions that arise from resampling.

A new general-purpose error bound

To analyze the bootstrap, it was necessary to use dimension-free high-probability upper bounds on $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$.

Existing dimension-free bounds generally require either $\|X_i\|_2 \leq c$ almost surely, or $\|\langle u, X_i \rangle\|_{\psi_2} \asymp \|\langle u, X_i \rangle\|_{L_2}$ for all $\|u\|_2 = 1$.

(e.g. Rudelson and Vershynin, (2007), Oliveira, (2010), Hsu, Kakade and Zhang, (2012), Koltchinskii and Lounici, (2017), Minsker, (2017))

However, the ℓ_2-boundedness condition is often restrictive, while the ψ_2-L_2 equivalence condition is not well-suited to the discrete distributions that arise from resampling.

As a way to streamline our analysis of both $(X_1, \ldots, X_n)$ and the bootstrap samples $(X_1^*, \ldots, X_n^*)$, it is of interest to develop a dimension-free bound that can be applied in a more general-purpose way.

A new general-purpose error bound

Proposition 1 (LEM 2019)

Let $\xi_1, \ldots, \xi_n \in \mathbb{R}^p$ be i.i.d. random vectors, and for any $q \geq 1$, define the quantity

$$\mathtt{r}(q) = q \, \frac{\left(\mathbb{E}[\|\xi_1\|_2^{2q}]\right)^{\frac{1}{q}}}{\left\|\mathbb{E}[\xi_1 \xi_1^\top]\right\|_{\mathrm{op}}}. \tag{4}$$

A new general-purpose error bound

Proposition 1 (LEM 2019)

Let $\xi_1, \ldots, \xi_n \in \mathbb{R}^p$ be i.i.d. random vectors, and for any $q \geq 1$, define the quantity

$$\mathtt{r}(q) = q \, \frac{\left(\mathbb{E}[\|\xi_1\|_2^{2q}]\right)^{\frac{1}{q}}}{\left\|\mathbb{E}[\xi_1 \xi_1^\top]\right\|_{\mathrm{op}}}. \tag{4}$$

Then, there is an absolute constant $c > 0$, such that for any $q \geq 3 \vee \log(n)$, the event

$$\left\| \frac{1}{n} \sum_{i=1}^n \xi_i \xi_i^\top - \mathbb{E}[\xi_i \xi_i^\top] \right\|_{\mathrm{op}} \leq c \cdot \left\|\mathbb{E}[\xi_1 \xi_1^\top]\right\|_{\mathrm{op}} \cdot \left(\sqrt{\frac{\mathtt{r}(q)}{n}} \vee \frac{\mathtt{r}(q)}{n} \right)$$

holds with probability at least $1 - \frac{1}{n}$.

A new general-purpose error bound

Proposition 1 (LEM 2019)

Let $\xi_1, \ldots, \xi_n \in \mathbb{R}^p$ be i.i.d. random vectors, and for any $q \geq 1$, define the quantity

$$\mathtt{r}(q) = q \, \frac{\left(\mathbb{E}[\|\xi_1\|_2^{2q}]\right)^{\frac{1}{q}}}{\left\|\mathbb{E}[\xi_1 \xi_1^\top]\right\|_{\mathrm{op}}}. \tag{4}$$

Then, there is an absolute constant $c > 0$, such that for any $q \geq 3 \vee \log(n)$, the event

$$\left\| \frac{1}{n} \sum_{i=1}^n \xi_i \xi_i^\top - \mathbb{E}[\xi_i \xi_i^\top] \right\|_{\mathrm{op}} \leq c \cdot \left\|\mathbb{E}[\xi_1 \xi_1^\top]\right\|_{\mathrm{op}} \cdot \left(\sqrt{\frac{\mathtt{r}(q)}{n}} \vee \frac{\mathtt{r}(q)}{n} \right)$$

holds with probability at least $1 - \frac{1}{n}$.

The proof extends an argument from Rudelson and Vershynin (2007) to the case of unbounded $\xi_1, \ldots, \xi_n$. The essential step is based on the non-commutative Khinchine inequality of Lust-Piquard (1986).

Coverage probabilities (error bounds or confidence regions)

Simulation settings:

- $n \in \{300, 500, 700\}$ and $p = 1{,}000$
- Repeated leading eigenvalues:

  $\lambda_1(\Sigma) = \cdots = \lambda_5(\Sigma) = 1 \qquad \text{and} \qquad \lambda_j(\Sigma) = j^{-2\beta} \quad \text{for } j \in \{6, \ldots, p\}$
- True eigenvectors were drawn from the Haar (uniform) distribution.
- Entries Z_{ij} drawn from $N(0, 1)$ or standardized t_{20}.

Coverage probabilities (error bounds or confidence regions)

Simulation settings:

- $n \in \{300, 500, 700\}$ and $p = 1{,}000$
- Repeated leading eigenvalues:

  $\lambda_1(\Sigma) = \cdots = \lambda_5(\Sigma) = 1$ and $\lambda_j(\Sigma) = j^{-2\beta}$ for $j \in \{6, \ldots, p\}$
- True eigenvectors were drawn from the Haar (uniform) distribution.
- Entries Z_{ij} drawn from $N(0, 1)$ or standardized t_{20}.

---

decay param. β	sample size n 300	500	700
0.75	92.83%	92.26%	91.96%
1.00	92.66%	91.70%	91.23%
1.25	92.43%	91.53%	91.16%

(a) $N(0, 1)$ distribution

decay param. β	sample size n 300	500	700
0.75	92.90%	91.96%	91.93%
1.00	92.53%	91.76%	91.63%
1.25	92.50%	91.70%	91.46%

(b) standardized t_{20} distribution

Simultaneous confidence intervals for true eigenvalues

It is of interest to construct confidence intervals $\mathcal{I}_1, \ldots, \mathcal{I}_p$ that satisfy

$$\mathbb{P}\left(\bigcap_{j=1}^{p} \{\lambda_j(\Sigma) \in \mathcal{I}_j\} \right) \geq 1 - \alpha. \qquad (*)$$

Simultaneous confidence intervals for true eigenvalues

It is of interest to construct confidence intervals $\mathcal{I}_1, \ldots, \mathcal{I}_p$ that satisfy

$$\mathbb{P}\left(\bigcap_{j=1}^{p} \{\lambda_j(\Sigma) \in \mathcal{I}_j\} \right) \geq 1 - \alpha. \qquad (*)$$

To do this, it is helpful to consider the (deterministic) Weyl inequality,

$$\max_{1 \leq j \leq p} |\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)| \leq \|\widehat{\Sigma} - \Sigma\|_{\text{op}}.$$

Simultaneous confidence intervals for true eigenvalues

It is of interest to construct confidence intervals $\mathcal{I}_1, \ldots, \mathcal{I}_p$ that satisfy

$$\mathbb{P}\left(\bigcap_{j=1}^{p} \{\lambda_j(\Sigma) \in \mathcal{I}_j\} \right) \geq 1 - \alpha. \qquad (*)$$

To do this, it is helpful to consider the (deterministic) Weyl inequality,

$$\max_{1 \leq j \leq p} |\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)| \leq \|\widehat{\Sigma} - \Sigma\|_{\text{op}}.$$

If $q_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of T, then Weyl's inequality implies that (*) must hold for $\mathcal{I}_j := [\lambda_j(\widehat{\Sigma}) \pm q_{1-\alpha}/\sqrt{n}]$. Hence, we may construct approximate intervals by replacing $q_{1-\alpha}$ with the bootstrap estimate $\widehat{q}_{1-\alpha}$.

Simultaneous confidence intervals for true eigenvalues

It is of interest to construct confidence intervals $\mathcal{I}_1, \ldots, \mathcal{I}_p$ that satisfy

$$\mathbb{P}\left(\bigcap_{j=1}^{p} \{\lambda_j(\Sigma) \in \mathcal{I}_j\} \right) \geq 1 - \alpha. \qquad (*)$$

To do this, it is helpful to consider the (deterministic) Weyl inequality,

$$\max_{1 \leq j \leq p} |\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)| \leq \|\widehat{\Sigma} - \Sigma\|_{\text{op}}.$$

If $q_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of T, then Weyl's inequality implies that (*) must hold for $\mathcal{I}_j := [\lambda_j(\widehat{\Sigma}) \pm q_{1-\alpha}/\sqrt{n}]$. Hence, we may construct approximate intervals by replacing $q_{1-\alpha}$ with the bootstrap estimate $\widehat{q}_{1-\alpha}$.

decay param. β	sample size n = 300	500	700
0.75	94.46%	94.26%	93.26%
1.00	93.13%	92.06%	91.53%
1.25	92.63%	91.73%	91.23%

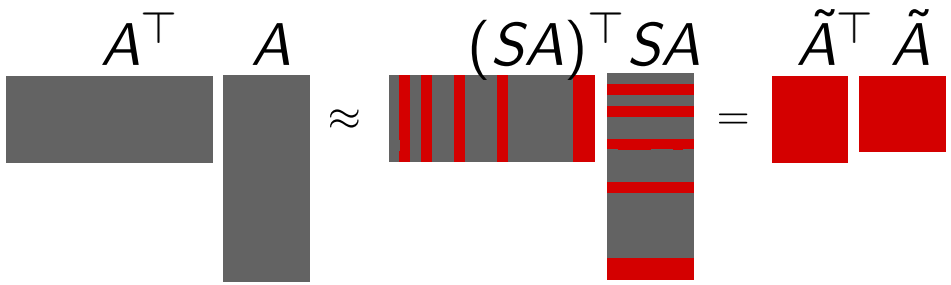
(a) $N(0,1)$ distribution

decay param. β	sample size n = 300	500	700
0.75	94.03%	93.87%	93.76%
1.00	92.90%	91.66%	91.46%
1.25	92.56%	91.40%	91.38%

(b) standardized t_{20} distribution

Application to randomized numerical linear algebra

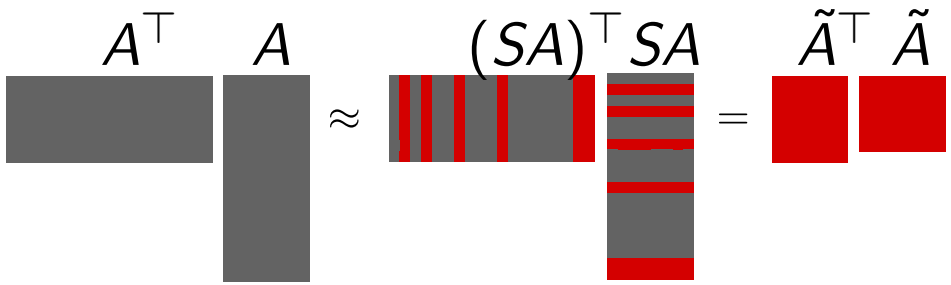
Recall the schematic for randomized matrix multiplication:



where $A \in \mathbb{R}^{d \times p}$ is deterministic with $d \geq p$, and the sketching matrix $S \in \mathbb{R}^{n \times d}$ is random with $n \ll d$.

Application to randomized numerical linear algebra

Recall the schematic for randomized matrix multiplication:

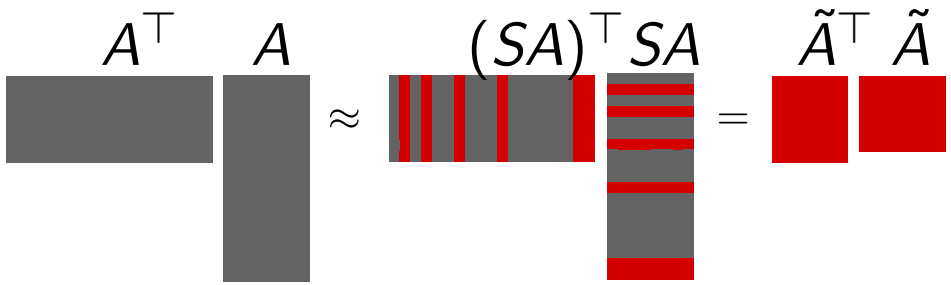


where $A \in \mathbb{R}^{d \times p}$ is deterministic with $d \geq p$, and the sketching matrix $S \in \mathbb{R}^{n \times d}$ is random with $n \ll d$.

Also, the rows of S are (usually) generated to be i.i.d. with $\mathbb{E}[S^\top S] = I$.

Application to randomized numerical linear algebra

Recall the schematic for randomized matrix multiplication:



where $A \in \mathbb{R}^{d \times p}$ is deterministic with $d \geq p$, and the sketching matrix $S \in \mathbb{R}^{n \times d}$ is random with $n \ll d$.

Also, the rows of S are (usually) generated to be i.i.d. with $\mathbb{E}[S^\top S] = I$.

To bootstrap the algorithmic error $\|\tilde{A}^\top \tilde{A} - A^\top A\|_{\text{op}}$, we may regard $\tilde{A}$ as a "data matrix" and sample its rows with replacement.

Note that the user generates the matrix S, which leaves no question about model assumptions!

Computational cost of bootstrapping

Historically, the bootstrap has been labeled as computationally intensive.

Hence, it seems counterintuitive to apply it in the service of comptuation.

Computational cost of bootstrapping

Historically, the bootstrap has been labeled as computationally intensive.

Hence, it seems counterintuitive to apply it in the service of comptuation.

Key considerations.

- In many large-scale computations, communication is the bottleneck, and the user may only be able to access A once or twice.

Computational cost of bootstrapping

Historically, the bootstrap has been labeled as computationally intensive.

Hence, it seems counterintuitive to apply it in the service of comptuation.

Key considerations.

- In many large-scale computations, communication is the bottleneck, and the user may only be able to access A once or twice.
- Whereas the computation of $\tilde{A}$ requires access to A, the bootstrap computations do not.

Computational cost of bootstrapping

Historically, the bootstrap has been labeled as computationally intensive.

Hence, it seems counterintuitive to apply it in the service of comptuation.

Key considerations.

- In many large-scale computations, communication is the bottleneck, and the user may only be able to access A once or twice.
- Whereas the computation of $\tilde{A}$ requires access to A, the bootstrap computations do not.
- Furthermore, the bootstrap computations only involve the small matrix $\tilde{A}$.

Computational cost of bootstrapping

Historically, the bootstrap has been labeled as computationally intensive.

Hence, it seems counterintuitive to apply it in the service of comptuation.

Key considerations.

- In many large-scale computations, communication is the bottleneck, and the user may only be able to access A once or twice.
- Whereas the computation of $\tilde{A}$ requires access to A, the bootstrap computations do not.
- Furthermore, the bootstrap computations only involve the small matrix $\tilde{A}$.
- Lastly, the bootstrap computations can be accelerated via extrapolation.

A simple and effective extrapolation rule

Recall that the sketched matrix $\tilde{A}$ is of size $n \times p$, and let $\mathrm{q}_{1-\alpha} = \mathrm{q}_{1-\alpha}(n)$ denote the $1-\alpha$ quantile of the error $\|\tilde{A}^\top \tilde{A} - A^\top A\|_{\mathrm{op}}$.

Since the error typically has fluctuations of order $1/\sqrt{n}$, we may expect the following relationship between a small "initial" sketch size n_0, and a larger "final" sketch size n_1,

$$\mathrm{q}_{1-\alpha}(n_1) \;\approx\; \sqrt{\tfrac{n_0}{n_1}}\; \mathrm{q}_{1-\alpha}(n_0). \tag{6}$$

A simple and effective extrapolation rule

Recall that the sketched matrix $\tilde{A}$ is of size $n \times p$, and let $\mathsf{q}_{1-\alpha} = \mathsf{q}_{1-\alpha}(n)$ denote the $1-\alpha$ quantile of the error $\|\tilde{A}^\top \tilde{A} - A^\top A\|_{\mathsf{op}}$.

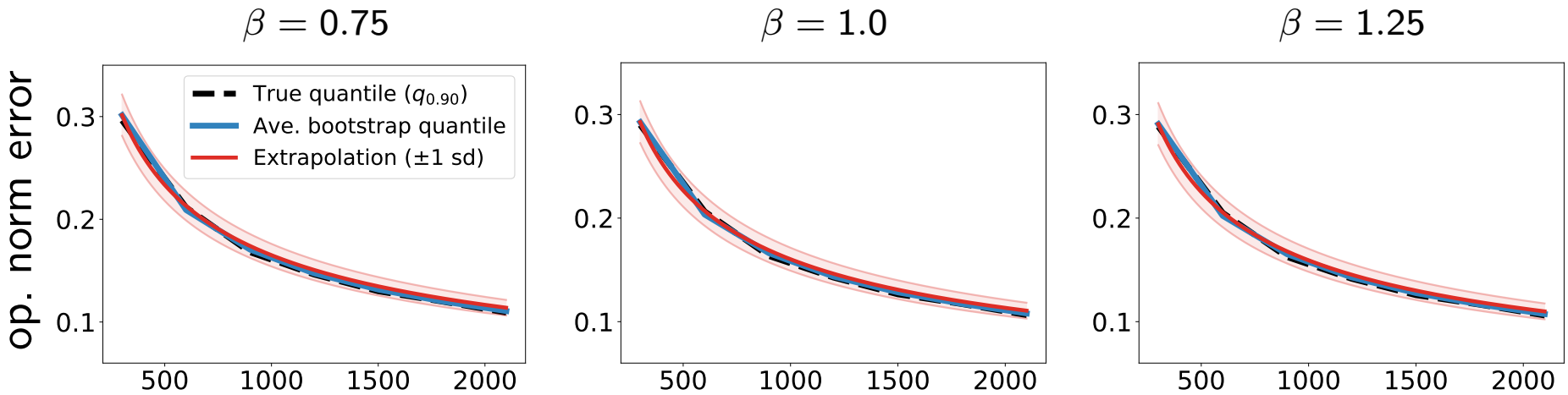
Since the error typically has fluctuations of order $1/\sqrt{n}$, we may expect the following relationship between a small "initial" sketch size n_0, and a larger "final" sketch size n_1,

$$\mathsf{q}_{1-\alpha}(n_1) \;\approx\; \sqrt{\tfrac{n_0}{n_1}}\; \mathsf{q}_{1-\alpha}(n_0). \tag{6}$$

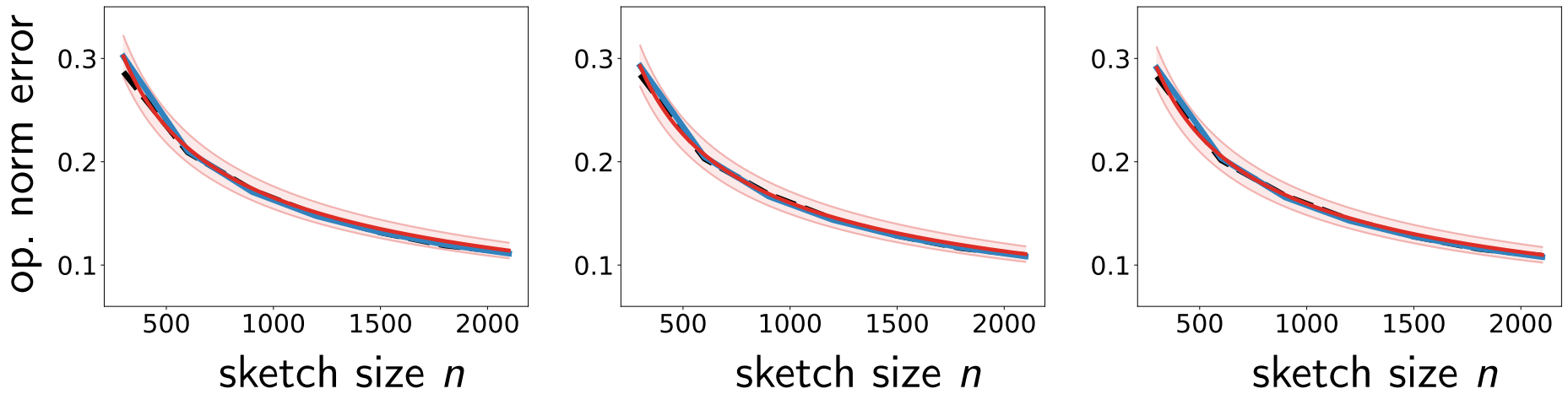
This leads to the extrapolation rule

$$\widehat{\mathsf{q}}_{1-\alpha}(n_1) := \sqrt{\tfrac{n_0}{n_1}}\, \widehat{\mathsf{q}}_{1-\alpha}(n_0) \qquad \text{for any } n_1 \geq n_0.$$

Important: $\widehat{\mathsf{q}}_{1-\alpha}(n_0)$ is much cheaper to compute than $\widehat{\mathsf{q}}_{1-\alpha}(n_1)$.



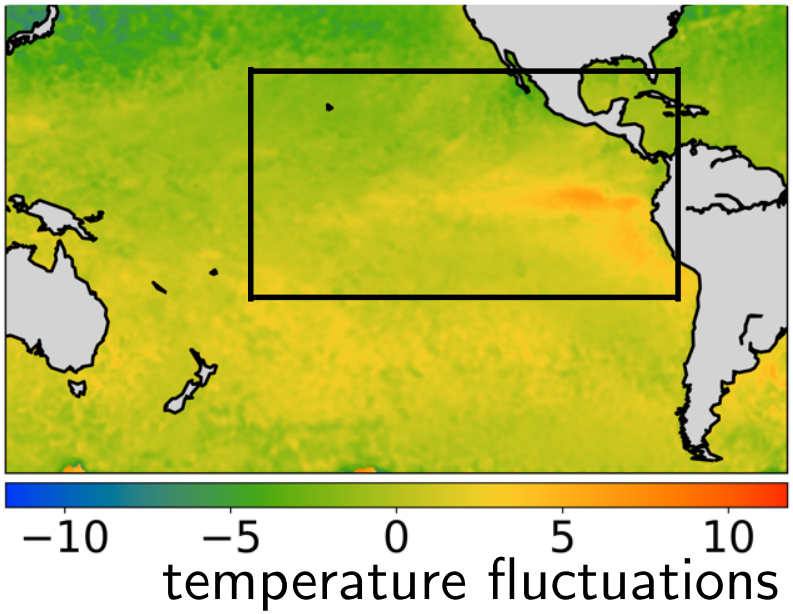
(a) Sketching with Gaussian random projections.



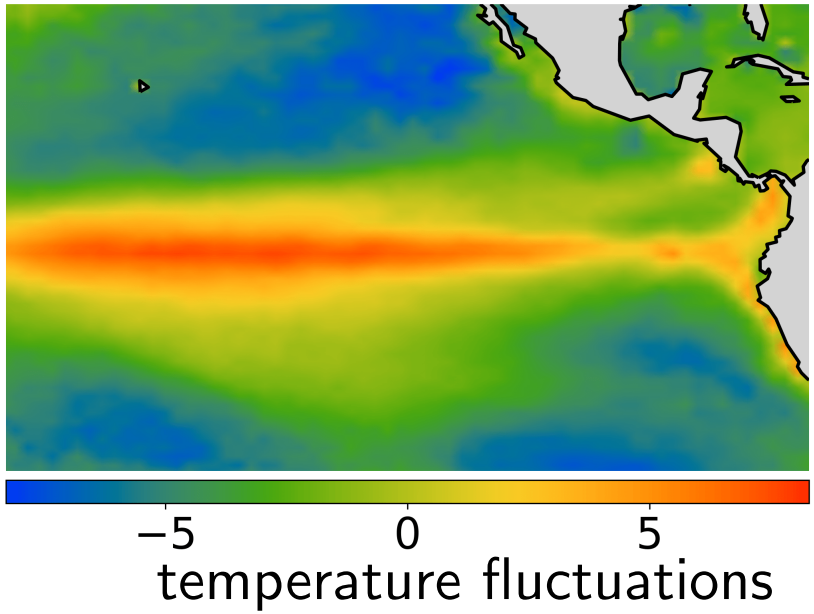
(b) Sketching with uniform row sampling.

Figure: Bootstrap estimates for the 90% quantile of the error $\|\tilde{A}^\top \tilde{A} - A^\top A\|_{\text{op}}$, where A is of size $10,000 \times 1,000$. Initial sketch size is only 300.

An example: Sea surface temperature data



(a) ENSO region

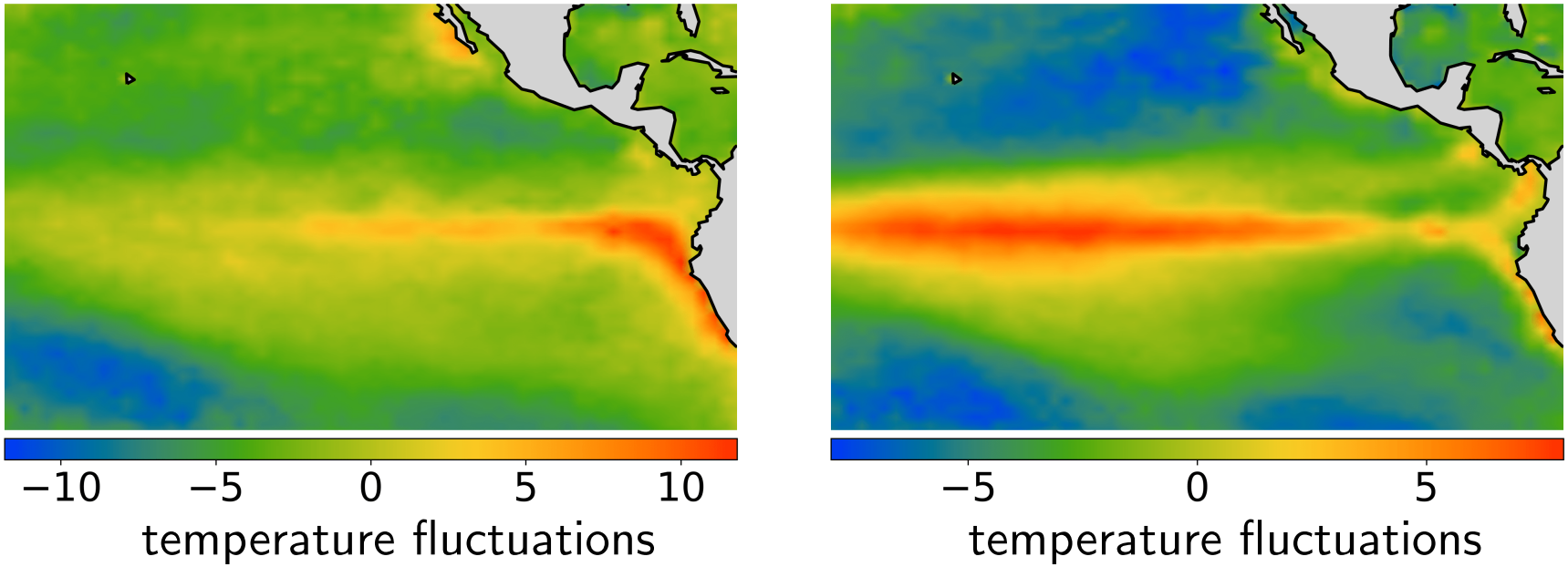


(b) exact ENSO mode

Figure: The rows of $A \in \mathbb{R}^{13{,}271 \times 3{,}944}$ are 13,271 snapshots of the ENSO region. (cf. NOAA SST dataset and Reynolds et al., 2002)

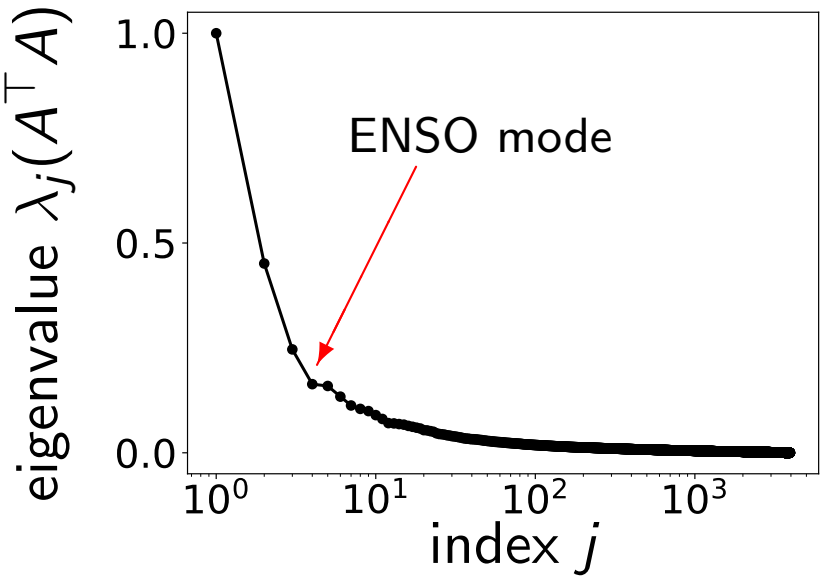
Panel (a): The ENSO region, marked with a rectangle.

Panel (b): The true ENSO mode, obtained by exact computation with the full product $A^\top A$.

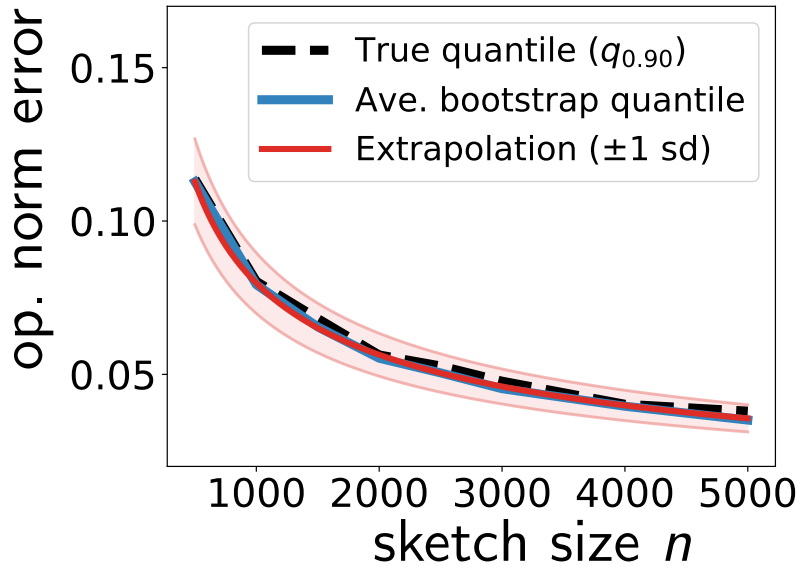


(a) approx. ENSO mode, $n = 500$ (b) approx. ENSO mode, $n = 3{,}000$

Figure: The left and right panels show approximations to the ENSO mode based on the approximate product $\tilde{A}^\top \tilde{A}$, obtained from Gaussian random projections with sketch sizes $n = 500$ and $n = 3{,}000$. A comparison with the exact ENSO mode shows that an insufficient sketch size can lead to a substantial distortion.



(a) spectrum of $A^\top A$



(b) error estimation

Figure: Panel (a): decaying eigenvalues of $A^\top A$.

Panel (b): The extrapolated and non-extrapolated bootstrap methods accurately estimate the 90% quantile of the sketching error $\|\tilde{A}^\top \tilde{A} - A^\top A\|_{\text{op}}$ 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 $n_0 = 500$.

Part II summary: Bootstrap for operator norm error

- For estimating the error of $\widehat{\Sigma}$ or constructing confidence regions for Σ, we need distributional approximation for $T = \sqrt{n}\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$.
- Under the spectrum decay condition $\lambda_j(\Sigma) \asymp j^{-2\beta}$ with $\beta > 1/2$, the ordinary non-parametric bootstrap works, with the rate of approximation being dimension-free.
- The bootstrap approximation guarantee for T is robust against the effect of repeated (or closely spaced) population eigenvalues.
- The bootstrap has a largely untapped potential for estimating the errors of randomized algorithms. (This is a relatively new area with many opportunities at the intersection of computer science, high-dimensional statistics, and random matrix theory.)

Thanks to

- the organizers for hospitality
- you for your attention
- NSF grants DMS-1613218 and DMS-1915786