Covariance estimation

via tail bounds for eigenvalues of sums of random matrices

Alex Gittens Joel A. Tropp

Department of Computing and Mathematical Sciences
California Institute of Technology
gittens@caltech.edu
jtropp@cms.caltech.edu

SPARS Workshop 2011

* *Research supported by ONR and AFOSR awards and a Sloan Fellowship.*

Problem Statement

Let $\mathbf{x} \in \mathbb{R}^p$ be a zero-mean high-dimensional random vector. Information on the dependence structure of $\mathbf{x}$ is captured by the covariance matrix

$$\mathbf{\Sigma} = \mathbb{E}\mathbf{x}\mathbf{x}^*.$$

The sample covariance matrix is a classical estimator for $\mathbf{\Sigma}$:

$$\widehat{\mathbf{\Sigma}}_n = \frac{1}{n} \sum\nolimits_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^*.$$

How many samples of $\mathbf{x}$ are required so that $\widehat{\mathbf{\Sigma}}_n$ accurately estimates $\mathbf{\Sigma}$?

What is known

Typically accuracy is measured in spectral norm.

How many samples ensure that

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_2 \leq \varepsilon \|\mathbf{\Sigma}\|_2?$$

- for log-concave distributions $\Omega(p)$ samples suffice (Adamczak et al. 2011),
- for distributions with finite fourth moments, $\tilde{\Omega}(p)$ samples suffice (Vershynin 2011a),
- for distributions with finite $2 + \varepsilon$ moments that satisfy a regularity condition, $\Omega(p)$ samples suffice (Vershynin 2011b),
- for distributions with finite second moments, $\Omega(p \log p)$ samples suffice (Rudelson 1999).

What is known

Typically accuracy is measured in spectral norm.

How many samples ensure that

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_2 \leq \varepsilon \|\mathbf{\Sigma}\|_2?$$

- for log-concave distributions $\Omega(p)$ samples suffice (Adamczak et al. 2011),
- for distributions with finite fourth moments, $\tilde{\Omega}(p)$ samples suffice (Vershynin 2011a),
- for distributions with finite $2 + \varepsilon$ moments that satisfy a regularity condition, $\Omega(p)$ samples suffice (Vershynin 2011b),
- for distributions with finite second moments, $\Omega(p \log p)$ samples suffice (Rudelson 1999).

What is known

Typically accuracy is measured in spectral norm.

How many samples ensure that

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_2 \leq \varepsilon\|\mathbf{\Sigma}\|_2?$$

- for log-concave distributions $\Omega(p)$ samples suffice (Adamczak et al. 2011),
- for distributions with finite fourth moments, $\tilde{\Omega}(p)$ samples suffice (Vershynin 2011a),
- for distributions with finite $2 + \varepsilon$ moments that satisfy a regularity condition, $\Omega(p)$ samples suffice (Vershynin 2011b),
- for distributions with finite second moments, $\Omega(p \log p)$ samples suffice (Rudelson 1999).

What is known

Typically accuracy is measured in spectral norm.

How many samples ensure that

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_2 \leq \varepsilon \|\mathbf{\Sigma}\|_2?$$

- for log-concave distributions $\Omega(p)$ samples suffice (Adamczak et al. 2011),
- for distributions with finite fourth moments, $\tilde{\Omega}(p)$ samples suffice (Vershynin 2011a),
- for distributions with finite $2 + \varepsilon$ moments that satisfy a regularity condition, $\Omega(p)$ samples suffice (Vershynin 2011b),
- for distributions with finite second moments, $\Omega(p \log p)$ samples suffice (Rudelson 1999).

What is known

Typically accuracy is measured in spectral norm.

How many samples ensure that

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_2 \leq \varepsilon \|\mathbf{\Sigma}\|_2?$$

- for log-concave distributions $\Omega(p)$ samples suffice (Adamczak et al. 2011),
- for distributions with finite fourth moments, $\tilde{\Omega}(p)$ samples suffice (Vershynin 2011a),
- for distributions with finite $2 + \varepsilon$ moments that satisfy a regularity condition, $\Omega(p)$ samples suffice (Vershynin 2011b),
- for distributions with finite second moments, $\Omega(p \log p)$ samples suffice (Rudelson 1999).

An observation

A relative spectral error bound,

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_2 \leq \varepsilon \|\mathbf{\Sigma}\|_2,$$

ensures recovery of the top eigenpair of $\mathbf{\Sigma}$, ...

but does *not* ensure the recovery of the remaining eigenpairs:

$$|\lambda_k(\mathbf{\Sigma}) - \lambda_k(\widehat{\mathbf{\Sigma}}_n)| < \varepsilon \|\mathbf{\Sigma}\|_2$$

is not meaningful if $\lambda_k \ll \lambda_1$.

Using known relative spectral error bounds, need $\mathrm{O}(\varepsilon^{-2}\kappa(\mathbf{\Sigma}_\ell)^2 p)$ measurements to get relative error recovery of the top ℓ eigenvalues.

An observation

A relative spectral error bound,

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_2 \leq \varepsilon \|\mathbf{\Sigma}\|_2,$$

ensures recovery of the top eigenpair of $\mathbf{\Sigma}$, ...

but does *not* ensure the recovery of the remaining eigenpairs:

$$|\lambda_k(\mathbf{\Sigma}) - \lambda_k(\widehat{\mathbf{\Sigma}}_n)| < \varepsilon \|\mathbf{\Sigma}\|_2$$

is not meaningful if $\lambda_k \ll \lambda_1$.

Using known relative spectral error bounds, need $\mathrm{O}(\varepsilon^{-2}\kappa(\mathbf{\Sigma}_\ell)^2 p)$ measurements to get relative error recovery of the top ℓ eigenvalues.

An observation

A relative spectral error bound,

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_2 \leq \varepsilon \|\mathbf{\Sigma}\|_2,$$

ensures recovery of the top eigenpair of $\mathbf{\Sigma}$, ...

but does *not* ensure the recovery of the remaining eigenpairs:

$$|\lambda_k(\mathbf{\Sigma}) - \lambda_k(\widehat{\mathbf{\Sigma}}_n)| < \varepsilon \|\mathbf{\Sigma}\|_2$$

is not meaningful if $\lambda_k \ll \lambda_1$.

Using known relative spectral error bounds, need $\mathrm{O}(\varepsilon^{-2}\kappa(\mathbf{\Sigma}_\ell)^2 p)$ measurements to get relative error recovery of the top ℓ eigenvalues.

. . . and a question

Maybe $\mathbf{\Sigma}$ has a decaying spectrum. What if we want accurate estimates of a few of its eigenvalues?

How many samples ensure the top $\ell \ll p$ eigenvalues are estimated to relative accuracy,

$$|\lambda_k(\mathbf{\Sigma}) - \lambda_k(\widehat{\mathbf{\Sigma}}_n)| \leq \varepsilon \lambda_k(\mathbf{\Sigma})?$$

Do we really need $\mathrm{O}(p)$ measurements to recover just a few of the top eigenvalues?

A simplified result

Theorem

Let the samples be drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ distribution. Assume λ_k decays sufficiently for $k > \ell$. If $\varepsilon \in (0, 1]$ and

$$n = \Omega(\varepsilon^{-2} \kappa(\mathbf{\Sigma}_\ell)^2 \ell \log p),$$

then with high probability, for each $k = 1, \ldots, \ell$,

$$|\lambda_k(\widehat{\mathbf{\Sigma}}_n) - \lambda_k(\mathbf{\Sigma})| \leq \varepsilon \lambda_k(\mathbf{\Sigma})$$

- Sufficient decay is, (other conditions give other results)

$$\sum_{i > \ell} \lambda_i / \lambda_1 \leq C.$$

  This is satisfied if, e.g., the tail eigenvalues, $k > \ell$, correspond to spread-spectrum noise or decay like $\frac{1}{i^{(1+\iota)}}$ for some $\iota > 0$.
- The approach generalizes to other subgaussian distributions.

A simplified result

Theorem

Let the samples be drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ distribution. Assume λ_k decays sufficiently for $k > \ell$. If $\varepsilon \in (0, 1]$ and

$$n = \Omega(\varepsilon^{-2} \kappa(\mathbf{\Sigma}_\ell)^2 \ell \log p),$$

then with high probability, for each $k = 1, \ldots, \ell$,

$$|\lambda_k(\widehat{\mathbf{\Sigma}}_n) - \lambda_k(\mathbf{\Sigma})| \leq \varepsilon \lambda_k(\mathbf{\Sigma})$$

- Sufficient decay is, (other conditions give other results)

$$\sum_{i>\ell} \lambda_i / \lambda_1 \leq C.$$

  This is satisfied if, e.g., the tail eigenvalues, $k > \ell$, correspond to spread-spectrum noise or decay like $\frac{1}{i^{(1+\iota)}}$ for some $\iota > 0$.

- The approach generalizes to other subgaussian distributions.

A simplified result

Theorem

Let the samples be drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ *distribution. Assume* λ_k *decays sufficiently for* $k > \ell$. *If* $\varepsilon \in (0, 1]$ *and*

$$n = \Omega(\varepsilon^{-2} \kappa(\mathbf{\Sigma}_\ell)^2 \ell \log p),$$

then with high probability, for each $k = 1, \ldots, \ell$,

$$|\lambda_k(\widehat{\mathbf{\Sigma}}_n) - \lambda_k(\mathbf{\Sigma})| \leq \varepsilon \lambda_k(\mathbf{\Sigma})$$

- Sufficient decay is, (other conditions give other results)

$$\sum_{i > \ell} \lambda_i / \lambda_1 \leq C.$$

  This is satisfied if, e.g., the tail eigenvalues, $k > \ell$, correspond to spread-spectrum noise or decay like $\frac{1}{i^{(1+\iota)}}$ for some $\iota > 0$.
- The approach generalizes to other subgaussian distributions.

More generally

Restrict, for each k, probability that $\hat{\lambda}_k$ under/overestimates λ_k.

- an upper bound on λ_k

$$n = \frac{8}{3\varepsilon^2} \boxed{\kappa(\mathbf{\Sigma}_k)\frac{\operatorname{tr}\mathbf{\Sigma}_k}{\lambda_k}(\log k + \beta \log p)} \Rightarrow \mathbb{P}\left\{\frac{\hat{\lambda}_k}{1-\epsilon} > \lambda_k\right\} > 1 - p^{-\beta}$$

- a lower bound on λ_k

$$n = \frac{1}{32\varepsilon^2} \boxed{\frac{\left(\sum_{i \geq k} \lambda_i\right)}{\lambda_k}(\log(p-k+1) + \beta \log p)}$$

$$\Rightarrow \mathbb{P}\left\{\frac{\hat{\lambda}_k}{1+\varepsilon} < \lambda_k\right\} > 1 - p^{-\beta}.$$

- Assuming decay

	upper bound	lower bound
λ_1	$\mathrm{O}(\log p)$	$\mathrm{O}(\ell \log p)$
λ_ℓ	$\mathrm{O}(\kappa^2(\mathbf{\Sigma}_\ell)\ell \log p)$	$\mathrm{O}(\kappa(\mathbf{\Sigma}_\ell) \log p)$

More generally

Restrict, for each k, probability that $\hat{\lambda}_k$ under/overestimates λ_k.

- an upper bound on λ_k

$$n = \frac{8}{3\varepsilon^2} \boxed{\kappa(\mathbf{\Sigma}_k)\frac{\operatorname{tr}\mathbf{\Sigma}_k}{\lambda_k}(\log k + \beta \log p)} \Rightarrow \mathbb{P}\left\{\frac{\hat{\lambda}_k}{1-\epsilon} > \lambda_k\right\} > 1 - p^{-\beta}$$

- a lower bound on λ_k

$$n = \frac{1}{32\varepsilon^2} \boxed{\frac{\left(\sum_{i \geq k} \lambda_i\right)}{\lambda_k}(\log(p-k+1) + \beta \log p)}$$

$$\Rightarrow \mathbb{P}\left\{\frac{\hat{\lambda}_k}{1+\varepsilon} < \lambda_k\right\} > 1 - p^{-\beta}.$$

- Assuming decay

	upper bound	lower bound
λ_1	$\mathrm{O}(\log p)$	$\mathrm{O}(\ell \log p)$
λ_ℓ	$\mathrm{O}(\kappa^2(\mathbf{\Sigma}_\ell)\ell \log p)$	$\mathrm{O}(\kappa(\mathbf{\Sigma}_\ell) \log p)$

More generally

Restrict, for each k, probability that $\hat{\lambda}_k$ under/overestimates λ_k.

- an upper bound on λ_k

$$n = \frac{8}{3\varepsilon^2} \boxed{\kappa(\mathbf{\Sigma}_k)\frac{\operatorname{tr} \mathbf{\Sigma}_k}{\lambda_k}(\log k + \beta \log p)} \Rightarrow \mathbb{P}\left\{\frac{\hat{\lambda}_k}{1-\epsilon} > \lambda_k\right\} > 1 - p^{-\beta}$$

- a lower bound on λ_k

$$n = \frac{1}{32\varepsilon^2} \boxed{\frac{\left(\sum_{i \geq k} \lambda_i\right)}{\lambda_k}(\log(p-k+1) + \beta \log p)}$$

$$\Rightarrow \mathbb{P}\left\{\frac{\hat{\lambda}_k}{1+\varepsilon} < \lambda_k\right\} > 1 - p^{-\beta}.$$

- Assuming decay

	upper bound	lower bound
λ_1	$\mathrm{O}(\log p)$	$\mathrm{O}(\ell \log p)$
λ_ℓ	$\mathrm{O}(\kappa^2(\mathbf{\Sigma}_\ell)\ell \log p)$	$\mathrm{O}(\kappa(\mathbf{\Sigma}_\ell) \log p)$

More generally

Restrict, for each k, probability that $\hat{\lambda}_k$ under/overestimates λ_k.

- an upper bound on λ_k

$$n = \frac{8}{3\varepsilon^2} \kappa(\mathbf{\Sigma}_k) \frac{\operatorname{tr} \mathbf{\Sigma}_k}{\lambda_k} (\log k + \beta \log p) \Rightarrow \mathbb{P}\left\{ \frac{\hat{\lambda}_k}{1-\epsilon} > \lambda_k \right\} > 1 - p^{-\beta}$$

- a lower bound on λ_k

$$n = \frac{1}{32\varepsilon^2} \frac{\left(\sum_{i \geq k} \lambda_i\right)}{\lambda_k} (\log(p-k+1) + \beta \log p)$$

$$\Rightarrow \mathbb{P}\left\{ \frac{\hat{\lambda}_k}{1+\varepsilon} < \lambda_k \right\} > 1 - p^{-\beta}.$$

- Assuming decay

	upper bound	lower bound
λ_1	$\mathrm{O}(\log p)$	$\mathrm{O}(\ell \log p)$
λ_ℓ	$\mathrm{O}(\kappa^2(\mathbf{\Sigma}_\ell)\ell \log p)$	$\mathrm{O}(\kappa(\mathbf{\Sigma}_\ell) \log p)$

Proof sketch

It suffices to show

$$\mathbb{P}\left\{\hat{\lambda}_k \geq (1+\varepsilon)\lambda_k\right\} \quad \text{and} \quad \mathbb{P}\left\{\hat{\lambda}_k \leq (1-\varepsilon)\lambda_k\right\}$$

decay like $\mathrm{C}\exp(-\mathrm{c}n\epsilon^2)$ when ϵ is sufficiently small.

1. Reduce the probability of each case occuring to the probability that the norm of an appropriate matrix is large.
2. Use matrix Bernstein bounds to establish the correct decay of these norms.
3. Take a union bound over the indices k.

Proof sketch

It suffices to show

$$\mathbb{P}\left\{\hat{\lambda}_k \geq (1+\varepsilon)\lambda_k\right\} \quad \text{and} \quad \mathbb{P}\left\{\hat{\lambda}_k \leq (1-\varepsilon)\lambda_k\right\}$$

decay like $\mathrm{C}\exp(-\mathrm{c}n\epsilon^2)$ when ϵ is sufficiently small.

1. Reduce the probability of each case occuring to the probability that the norm of an appropriate matrix is large.
2. Use matrix Bernstein bounds to establish the correct decay of these norms.
3. Take a union bound over the indices k.

Proof sketch

It suffices to show

$$\mathbb{P}\left\{\hat{\lambda}_k \geq (1+\varepsilon)\lambda_k\right\} \quad \text{and} \quad \mathbb{P}\left\{\hat{\lambda}_k \leq (1-\varepsilon)\lambda_k\right\}$$

decay like $\mathrm{C}\exp(-\mathrm{c}n\epsilon^2)$ when ϵ is sufficiently small.

1. Reduce the probability of each case occuring to the probability that the norm of an appropriate matrix is large.
2. Use matrix Bernstein bounds to establish the correct decay of these norms.
3. Take a union bound over the indices k.

Proof sketch

It suffices to show

$$\mathbb{P}\left\{\hat{\lambda}_k \geq (1+\varepsilon)\lambda_k\right\} \quad \text{and} \quad \mathbb{P}\left\{\hat{\lambda}_k \leq (1-\varepsilon)\lambda_k\right\}$$

decay like $\mathrm{C}\exp(-\mathrm{c}n\epsilon^2)$ when ϵ is sufficiently small.

1. Reduce the probability of each case occuring to the probability that the norm of an appropriate matrix is large.
2. Use matrix Bernstein bounds to establish the correct decay of these norms.
3. Take a union bound over the indices k.

Reduction for $\hat{\lambda}_k \geq \lambda_k + t$

Let $\mathbf{B}$ have orthonormal columns and span the bottom $(p-k+1)$-dimensional invariant subspace of $\mathbf{\Sigma}$.

Claim

$$\mathbb{P}\left\{ \hat{\lambda}_k \ \geq \ \lambda_k \ + t \right\} \leq \mathbb{P}\left\{ \lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}) \ \geq \ \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B}) \ + t \right\}.$$

Proof.

By Courant–Fischer,

$$\lambda_k(\mathbf{\Sigma}) = \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B})$$

and

$$\lambda_k(\widehat{\mathbf{\Sigma}}_n) = \min_{\substack{\mathbf{V}\in\mathbb{C}^{p\times(p-k+1)} \\ \mathbf{V}^*\mathbf{V}=\mathbf{I}}} \lambda_1(\mathbf{V}^*\widehat{\mathbf{\Sigma}}_n\mathbf{V}) \leq \lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}).$$

$\square$

Reduction for $\hat{\lambda}_k \geq \lambda_k + t$

Let $\mathbf{B}$ have orthonormal columns and span the bottom $(p-k+1)$-dimensional invariant subspace of $\mathbf{\Sigma}$.

Claim

$$\mathbb{P}\left\{ \hat{\lambda}_k \ \geq \ \lambda_k \ + t \right\} \leq \mathbb{P}\left\{ \lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}) \ \geq \ \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B}) \ + t \right\}.$$

Proof.
By Courant–Fischer,

$$\lambda_k(\mathbf{\Sigma}) = \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B})$$

and

$$\lambda_k(\widehat{\mathbf{\Sigma}}_n) = \min_{\substack{\mathbf{V}\in\mathbb{C}^{p\times(p-k+1)}\\ \mathbf{V}^*\mathbf{V}=\mathbf{I}}} \lambda_1(\mathbf{V}^*\widehat{\mathbf{\Sigma}}_n\mathbf{V}) \leq \lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}).$$

$\square$

Reduction for $\hat{\lambda}_k \geq \lambda_k + t$

Let $\mathbf{B}$ have orthonormal columns and span the bottom $(p-k+1)$-dimensional invariant subspace of $\mathbf{\Sigma}$.

Claim

$$\mathbb{P}\left\{ \hat{\lambda}_k \geq \lambda_k + t \right\} \leq \mathbb{P}\left\{ \lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}) \geq \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B}) + t \right\}.$$

Proof.
By Courant–Fischer,

$$\lambda_k(\mathbf{\Sigma}) = \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B})$$

and

$$\lambda_k(\widehat{\mathbf{\Sigma}}_n) = \min_{\substack{\mathbf{V}\in\mathbb{C}^{p\times(p-k+1)} \\ \mathbf{V}^*\mathbf{V}=\mathbf{I}}} \lambda_1(\mathbf{V}^*\widehat{\mathbf{\Sigma}}_n\mathbf{V}) \leq \lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}).$$

$\square$

Using the reduction

Need to control RHS of

$$\mathbb{P}\left\{\hat{\lambda}_k \geq \lambda_k + t\right\} \leq \mathbb{P}\left\{\lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}) \geq \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B}) + t\right\}$$

Note:

- $\lambda_1(\mathbf{B}^*\hat{\mathbf{\Sigma}}_n\mathbf{B}) \to \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B})$, and
- $\mathbf{B}^*\hat{\mathbf{\Sigma}}_n\mathbf{B} = \sum_i \mathbf{B}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{B}$ is a sum of independent random matrices.

Use estimates of the matrix moments of the summands to quantify the convergence.

- If $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, then for $m \geq 2$,

$$\mathbb{E}(\mathbf{g}\mathbf{g}^*)^m \preceq 2^m m! \, (\operatorname{tr} \mathbf{C})^{m-1} \cdot \mathbf{C}.$$

- Other subgaussian distributions satisfy similar relations. Can also substitute bounds on matrix moment generating functions,

$$\mathbb{E} \exp\left(\theta \mathbf{y}\mathbf{y}^*\right) \preceq \mathbf{U}(\theta).$$

Using the reduction

Need to control RHS of

$$\mathbb{P}\left\{\hat{\lambda}_k \geq \lambda_k + t\right\} \leq \mathbb{P}\left\{\lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}) \geq \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B}) + t\right\}$$

Note:

- $\lambda_1(\mathbf{B}^*\hat{\mathbf{\Sigma}}_n\mathbf{B}) \to \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B})$, and
- $\mathbf{B}^*\hat{\mathbf{\Sigma}}_n\mathbf{B} = \sum_i \mathbf{B}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{B}$ is a sum of independent random matrices.

Use estimates of the matrix moments of the summands to quantify the convergence.

- If $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, then for $m \geq 2$,

$$\mathbb{E}(\mathbf{g}\mathbf{g}^*)^m \preceq 2^m m!\,(\operatorname{tr}\mathbf{C})^{m-1} \cdot \mathbf{C}.$$

- Other subgaussian distributions satisfy similar relations. Can also substitute bounds on matrix moment generating functions,

$$\mathbb{E}\exp\left(\theta\mathbf{y}\mathbf{y}^*\right) \preceq \mathbf{U}(\theta).$$

Using the reduction

Need to control RHS of

$$\mathbb{P}\left\{\hat{\lambda}_k \geq \lambda_k + t\right\} \leq \mathbb{P}\left\{\lambda_1(\mathbf{B}^*\widehat{\boldsymbol{\Sigma}}_n\mathbf{B}) \geq \lambda_1(\mathbf{B}^*\boldsymbol{\Sigma}\mathbf{B}) + t\right\}$$

Note:

- $\lambda_1(\mathbf{B}^*\hat{\boldsymbol{\Sigma}}_n\mathbf{B}) \to \lambda_1(\mathbf{B}^*\boldsymbol{\Sigma}\mathbf{B})$, and
- $\mathbf{B}^*\hat{\boldsymbol{\Sigma}}_n\mathbf{B} = \sum_i \mathbf{B}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{B}$ is a sum of independent random matrices.

Use estimates of the matrix moments of the summands to quantify the convergence.

- If $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, then for $m \geq 2$,

$$\mathbb{E}(\mathbf{g}\mathbf{g}^*)^m \preceq 2^m m!\,(\operatorname{tr}\mathbf{C})^{m-1} \cdot \mathbf{C}.$$

- Other subgaussian distributions satisfy similar relations. Can also substitute bounds on matrix moment generating functions,

$$\mathbb{E}\exp(\theta\mathbf{y}\mathbf{y}^*) \preceq \mathbf{U}(\theta).$$

Using the reduction

Need to control RHS of

$$\mathbb{P}\left\{\hat{\lambda}_k \geq \lambda_k + t\right\} \leq \mathbb{P}\left\{\lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}) \geq \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B}) + t\right\}$$

Note:

- $\lambda_1(\mathbf{B}^*\hat{\mathbf{\Sigma}}_n\mathbf{B}) \to \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B})$, and
- $\mathbf{B}^*\hat{\mathbf{\Sigma}}_n\mathbf{B} = \sum_i \mathbf{B}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{B}$ is a sum of independent random matrices.

Use estimates of the matrix moments of the summands to quantify the convergence.

- If $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, then for $m \geq 2$,

$$\mathbb{E}(\mathbf{g}\mathbf{g}^*)^m \preceq 2^m m!\,(\operatorname{tr}\mathbf{C})^{m-1} \cdot \mathbf{C}.$$

- Other subgaussian distributions satisfy similar relations. Can also substitute bounds on matrix moment generating functions,

$$\mathbb{E}\exp\left(\theta\mathbf{y}\mathbf{y}^*\right) \preceq \mathbf{U}(\theta).$$

Matrix Bernstein inequality

We use a moment-based matrix analog of Bernstein's inequality.

Theorem (Matrix Moment-Bernstein Inequality)

Suppose self-adjoint matrices $\{\mathbf{G}_i\}$ *have dimension* d *and*

$$\mathbb{E}(\mathbf{G}_i^m) \preceq \frac{m!}{2} A^{m-2} \cdot \mathbf{C}_i^2 \quad \text{for } m = 2, 3, 4, \ldots.$$

Set

$$\mu = \lambda_1\Big(\sum\nolimits_i \mathbb{E}\mathbf{G}_i\Big) \quad \text{and} \quad \sigma^2 = \lambda_1\Big(\sum\nolimits_i \mathbf{C}_i^2\Big).$$

Then, for any $t \geq 0$,

$$\mathbb{P}\Big\{\lambda_1\Big(\sum\nolimits_i \mathbf{G}_i\Big) \geq \mu + t\Big\} \leq d \cdot \exp\Big(-\frac{t^2/2}{\sigma^2 + At}\Big).$$

Finishing the argument

After computing A and $\mathbf{C}_i^2$ for the summands $\mathbf{B}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{B}$, this gives

$$\mathbb{P}\left\{\hat{\lambda}_k \geq \lambda_k + t\right\} \leq (p-k+1)\cdot\exp\left(\frac{-nt^2}{32\lambda_k\sum_{i\geq k}\lambda_i}\right) \quad \text{for } t \leq 4n\lambda_k.$$

Finally, take $t = \varepsilon\lambda_k$ to see

$$\mathbb{P}\left\{\hat{\lambda}_k \geq (1+\varepsilon)\lambda_k\right\} \leq (p-k+1)\cdot\exp\left(\frac{-n\varepsilon^2}{32\sum_{i\geq k}\frac{\lambda_i}{\lambda_k}}\right) \quad \text{for } \varepsilon \leq 4n.$$

The proof for the case $\hat{\lambda}_k \leq \lambda_k - t$ is similar. $\square$

Details

"*Tail Bounds for All Eigenvalues of A Sum of Random Matrices*", Gittens and Tropp, 2011. Preprint, arXiv:1104.4513.

- Elaboration on the relative error estimation results.
- Similar arguments to find tail bounds for all eigenvalues of a sum of *arbitrary* random matrices.
- An application to column subsampling.