36-755: Advanced Statistical Theory **Spring 2017**

Lecture 7: September 25

Lecturer: Alessandro Rinaldo *Scribes: Alan Mishler*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

7.1 Covariance matrix estimation in the operator norm

Note: We stated and partially proved the following theorem last time. This section restates the theorem and completes the proof.

Theorem 7.1 Let $X_1, \ldots X_n \overset{iid}{\sim} (0, \mathbf{\Sigma})$ *in* $\mathbb{R}^d$ *s.t.* $X_i \in SG(\sigma^2), i = 1, \ldots, n,$ and let $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ be the usual estimator for $\mathbf{\Sigma}$.

Then $\mathbb{P}\left(\frac{\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{op}}{\sigma^2} \leq C * \max\left\{\sqrt{\frac{d+\log(2/\delta)}{n}}, \frac{d+\log(2/\delta)}{n}\right\}\right) \geq 1 - \delta$ for some constant C and for all $\delta \in (0,1)$.

Remarks

- Recall that $X_i \in SG(\sigma^2)$ iff $\mathbf{v}^T X_i \in SG(\sigma^2), \forall v \in S^{d-1}$, where $S^{d-1} = \{y \in \mathbb{R}^d : \|y\| = 1\}$.)
- Note that if $\delta = n^{-C}, C > 0$, then $\log\left(\frac{2}{\delta}\right) \sim \log(n)$.
- If $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{op} \leq \epsilon$, then by Weyl's Theorem, $\max_i |\lambda_i - \hat{\lambda_i}| \leq \epsilon$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ are the eigenvalues of $\mathbf{\Sigma}$ and $\hat{\lambda_1} \geq \hat{\lambda_2} \geq \ldots \geq \hat{\lambda_d}$ are the eigenvalues of $\hat{\mathbf{\Sigma}}$.

  In other words, this estimator of the covariance matrix is a "good" one in some sense.

Proof: The strategy is to use discretization to approximate the max over an infinite set by the max over a finite set; then we use concentration inequalities to bound the individual probabilities.

For a $d \times d$ matrix $\mathbf{A}$, express

$$\|\mathbf{A}\|_{op} = \max_{x \in S^{d-1}} \left|x^T \mathbf{A} x\right| \leq \frac{1}{1-2\epsilon} \max_{z \in \mathcal{N}_\epsilon} \left|z^t \mathbf{A} z\right| \tag{7.1}$$

where $\mathcal{N}_\epsilon$ is an ϵ-covering of S^{d-1}, with $\epsilon < 1/2$. In particular, pick $\epsilon < 1/4 \implies |\mathcal{N}_\epsilon| \leq 9^d$, using volumetric argument bounds.

Let

$$\mathbf{Q} = \hat{\mathbf{\Sigma}} - \mathbf{\Sigma} \implies \|\mathbf{Q}\|_{op} \leq 2 \max_{i=1,\ldots,|\mathcal{N}_\epsilon|} |v_i^T \mathbf{Q} v_i|. \tag{7.2}$$

where $\{v_1, \ldots, v_{\mathcal{N}_\epsilon}\}$ is an ϵ-covering of S^{d-a}. Then

$$\begin{aligned} \mathbb{P}(\|Q\|_{op} \geq t) &\leq \mathbb{P}(max_i |v_i^t \mathbf{Q} v_i| \geq t/2) && \text{(discretization argument)} \quad (7.3)\\ &\leq \sum_{i=1}^{|\mathcal{N}_\epsilon|} (|v_i^T \mathbf{Q} v_i| \geq t/2) && \text{(union bound)} \quad (7.4) \end{aligned}$$

Now we want to bound the individual probabilities $\mathbb{P}(|v_i^T \mathbf{Q} v_i| \geq t/2)$.

For fixed $v \in S^{d-1}$,

$$\begin{aligned} v^T \mathbf{Q} v &= v^T (\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}) v && (7.5)\\ &= \frac{1}{n} \sum_{i=1}^{n} \left((v^T X_i)^2 - v^T \mathbf{\Sigma} v \right) && (7.6)\\ &= \frac{1}{n} \sum_{i=1}^{n} \left(v^T X_i X_i v - v^T \mathbf{\Sigma} v \right) && (7.7)\\ &= \frac{1}{n} \sum_{i=1}^{n} \left(Z_i^2 - \mathbb{E}[Z_i^2] \right), \text{ where } z_i = v^t X_i && (7.8) \end{aligned}$$

Recall that if $Z_i \in SG(\sigma^2)$, then $Z_i^2 \in SE((16\sigma^2)^2, 16\sigma^2)$, so we can use the probability bounds for sub-exponential variables:

$$\begin{aligned} &\mathbb{P}(|v_i^t \mathbf{Q} v_i| \geq t/2) \leq 2 \exp \left\{ -\frac{n}{2} \min \left\{ \left(\frac{t}{32\sigma^2} \right)^2, \frac{t}{32\sigma^2} \right\} \right\}, \forall i && (7.9)\\ &\mathbb{P}(\|Q\|_{op} \geq t\sigma^2 \leq |\mathcal{N}_\epsilon| && (7.10) \end{aligned}$$

The RHS is $\leq \delta$ if $t \geq 32 \max\{\epsilon_n, \sqrt{\epsilon_n}\}$, where $\epsilon_n = \frac{2d}{n} \log(9) + \frac{2}{n} \log \frac{2}{\delta}$. ■

Remark: Here's another way of framing this result. Assume $X_i = \mathbf{\Sigma}^{1/2}, Y_i \in SG_d(1)$. Then $X_i \in SG(\|\mathbf{\Sigma}\|_{op})$, so this produces a bound on $\|\mathbf{\Sigma}\|_{op} / \|\mathbf{\Sigma}\|_{op}$, instead of $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{op} / \sigma^2$.

7.2 From bound on probability to bound on expectation

Theorem 7.2 *Assume the same conditions as in Theorem 7.1. Then*

$$\mathbb{E}[\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{op}] \leq C_1 \|\mathbf{\Sigma}\|_{op} \max \left\{ \frac{d}{n}, \sqrt{dn} \right\} \tag{7.11}$$

for some $C_1 > 0$.

Remarks

- You need that $d = o(n)$ in order to get that $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{op} \to 0$ in both probability and expectation.
- This rate is minimax optimal.
- If $d = o(n)$, then the bound is of order $\|\mathbf{\Sigma}\|_{op} \sqrt{d} \frac{1}{\sqrt{n}}$. This bound can actually be improved to $\|\sqrt{d_{\text{int}}}\|_{op}$, where $d_{\text{int}} := \text{tr}(\mathbf{\Sigma}) / \|\Sigma\|_{op}$ is the "intrinsic" (aka "effective") dimension of $\mathbf{\Sigma}$. Note that $d_{\text{int}} \leq d$, because $\text{tr}(\mathbf{\Sigma}) \leq d \|\mathbf{\Sigma}\|_{op}$.[1]

[1] See Bunea & Xiao, 2015.

7.3 Matrix Bernstein inequality

References
Joel Tropp (2011): User-friendly tail bounds for sums of random matrices
Joel Tropp (2014): Introduction to matrix concentration inequalities.

7.3.1 Matrix calculus preliminaries

Below, assume all matrices are symmetric.

A $d \times d$ matrix $\mathbf{A}$ is positive semi-definite (PSD) iff $x^T \mathbf{A} x \geq 0, \forall x \in \mathbb{R}^d \iff \lambda_i(\mathbf{A}) \geq 0, \forall i$, where $\lambda_i, \ldots, \lambda_d$ are the eigenvalues of $\mathbf{A}$.

A symmetric matrix has a spectral decomposition $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where

$\Lambda = \text{diag}(\lambda_i, \ldots, \lambda_d)$, where λ_i is the i^{th} eigenvalue of $\mathbf{A}$, including multiplicities.
$\mathbf{U} = [u_1 \ldots u_d]$, a $d \times d$ matrix in which u_i is the i^{th} eigenvector of $\mathbf{A}$.

7.3.1.1 Matrix functions

Let S_d^+ be the cone of the PSD $d \times d$ matrices and S_d^{++} be the cone of positive definite (PD) matrices. Given a symmetric matrix $\mathbf{A}$ as above, let $f^* : \mathbb{R} \to \mathbb{R}$ be a real-valued function. Then let

$$f(\mathbf{A}) := \mathbf{U} f(\boldsymbol{\Lambda}) \mathbf{U}^T \tag{7.12}$$

$$:= \mathbf{U}\, \text{diag}(f(\lambda_i), f(\ldots), f(\lambda_d)) \mathbf{U}^T \tag{7.13}$$

$$= \sum_{i=1}^{d} f(\lambda_i) \mathbf{U}_i \mathbf{U}_i^T \tag{7.14}$$

7.3.1.2 “Transfer rule”

Define a partial ordering on S_d^+ by $\mathbf{A} \preccurlyeq \mathbf{B}$ if $\mathbf{B} - \mathbf{A} \in S_d^+$. Note that $S_d^+ \succcurlyeq \mathbf{0}$ by the same ordering rule.

Transfer rule: Let $f^*, g^* : I \subseteq \mathbb{R} \to \mathbb{R}$ be functions on an interval I s.t. spectrum$(\mathbf{A}) \subset I$ and $f^*(x) \leq g^*(x), \forall x \in I$. Then $f(\mathbf{A}) \preccurlyeq g(\mathbf{A})$, where f and g are defined as above.

7.3.1.3 Examples

- $\exp(\mathbf{A}) = \mathbf{I} + \sum_{i=1}^{\infty} \frac{\mathbf{A}^i}{i!}$ and $\mathbf{A}^i = \mathbf{U}\Lambda^i\mathbf{U}^T \implies \mathbf{I} + \mathbf{A} \preccurlyeq \exp(\mathbf{A})$
- The log matrix function is defined as the function that satisfies $\log \exp(\mathbf{A}) = \mathbf{A}$. If $\mathbf{0} \preccurlyeq \mathbf{A} \preccurlyeq \mathbf{B}$, then $\log(\mathbf{A}) \preccurlyeq \log(\mathbf{B})$.
- **tr-exp inequality:** Recall that $\text{tr}(\mathbf{A}) = \sum_i A_{ii} = \sum_i \lambda_i$. If $\mathbf{A} \preccurlyeq \mathbf{B}$, then $\text{tr} \exp(\mathbf{A}) \leq \text{tr} \exp(\mathbf{B})$.
- **Golden-Thomspson inequality:** $\text{tr} \exp(\mathbf{A} + \mathbf{B}) \leq \text{tr} \exp(\mathbf{A})$. Note that in general, $\exp(\mathbf{A} + \mathbf{B}) \neq \exp(\mathbf{A}) \exp(\mathbf{B})$, because $\mathbf{AB} \neq \mathbf{BA}$ generally.

Theorem 7.3 (Matrix Bernstein inequality) *Let $X_1, \ldots, X_n$ be 0-mean $d \times d$ symmetric matrices s.t. $\|X_i\|_{op} \leq C$ a.e., $\forall i$.* (This is like assuming that we have independent 0-mean random variables.) *Then*

$$\mathbb{P}(\|\sum_{i=1}^{n} X_i\|_{op} \geq t) \leq 2d \exp\left\{-\frac{t^2}{2(\sigma^2 + \frac{tC}{3})}\right\} \tag{7.15}$$

where $\sigma^2 = \| \sum_{i=1}^{n} \mathbb{E}[X_i^2]\|_{op}$. With $d = 1$, this yields the ordinary Bernstein inequality.

Proof:
Step 1. We use the Chernoff procedure to bound the mgf of a random matrix.

First, note that $\|A\|_{op} = \max\{\lambda_{\max}(\mathbf{A}), \lambda_{\max}(-\mathbf{A})\}$, where $\lambda_{\max}(\mathbf{A})$ is the largest eigenvalue $\lambda_i(\mathbf{A})$. Hence, it suffices to bound $\mathbb{P}(\lambda_{\max}(\sum_{i=1}^{n} X_i \geq t))$.

Let $S = \sum_{i=1}^{n} X_i$. For $\lambda \geq 0$ and $t \in \mathbb{R}$:

$$\begin{aligned}
\mathbb{P}(\lambda_{\max}(S) \geq t) &\leq e^{\lambda t}\mathbb{E}\left[e^{\lambda \cdot \lambda_{\max}}(S)\right] && \text{(Chernoff)} && (7.16)\\
&= e^{\lambda t}\mathbb{E}\left[e^{\lambda_{\max}(\lambda S)}\right] && && (7.17)\\
&= e^{\lambda t}\mathbb{E}\left[\lambda_{\max} e^{(\lambda S)}\right] && && (7.18)\\
&\leq e^{\lambda t}\mathbb{E}\left[\operatorname{tr} e^{(\lambda S)}\right] && \text{(because } e^{\mathbf{A}} \succcurlyeq \mathbf{0}\text{)} && (7.19)
\end{aligned}$$

Step 2. We apply *Lieb's inequality.*

Let $\mathbf{B}$ be a $d \times d$ symmetric matrix. Then the function

$$\mathbf{A} \in S_d^{+} \to \operatorname{tr}\exp(\mathbf{B} + \log \mathbf{A}) \tag{7.20}$$

is concave on S_d^{+}. So if $\mathbf{X}$ is a random $d \times d$ symmetric matrix, then, by Jensen's inequality for matrices:

$$\mathbb{E}[\operatorname{tr}\exp(\mathbf{B} + \mathbf{X})] \leq \operatorname{tr}(\exp \mathbf{B} + \log(\mathbb{E}[e^{\mathbf{X}}])) \tag{7.21}$$

We can apply this inequality to $\mathbf{B}$ and $\mathbf{A} = e^{\mathbf{X}}$. We need to bound $\mathbb{E}[\operatorname{tr}\exp(\lambda \sum_{i=1}^{n}[bX_i)]$:

$$\begin{aligned}
&\mathbb{E}[\operatorname{tr}\exp(\lambda \sum_{i=1}^{n}[bX_i)] && (7.22)\\
=&\mathbb{E}[\operatorname{tr}\exp(\lambda \sum_{i=1}^{n-1} \mathbf{X}_i + \lambda \mathbf{X}_n)] && (7.23)\\
=&\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_{n-1}}\left[\mathbb{E}_{\mathbf{X}_n|\mathbf{X}_1,\ldots,\mathbf{X}_{n-1}}\left[\operatorname{tr}\exp(\lambda \sum_{i=1}^{n-1} \mathbf{X}_i + \lambda \mathbf{X}_n)\right] \mid \mathbf{X}_1, \ldots, \mathbf{X}_{n-1}\right] && (7.24)\\
=&\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_{n-1}}\left[\mathbb{E}_{\mathbf{X}_n}\left[\operatorname{tr}\exp(\lambda \sum_{i=1}^{n-1} \mathbf{X}_i + \lambda \mathbf{X}_n)\right] \mid \mathbf{X}_1, \ldots, \mathbf{X}_{n-1}\right] && (7.25)\\
\leq&\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_{n-1}}\left[\operatorname{tr}\exp(\lambda \sum_{i=1}^{n-1} \mathbf{X}_i + \log \mathbb{E}_{e\mathbf{X}_n} \lambda \mathbf{X}_n)\right] && (7.26)\\
&\vdots && (7.27)\\
\leq&\operatorname{tr}\exp\left(\sum_{i=1}^{n} \log \mathbb{E}\left[e^{\lambda X_i}\right]\right) && (7.28)
\end{aligned}$$

where the third line uses the law of total expectation, the fourth line uses the independence of the $\mathbf{X}_i$'s, and the fifth line uses Lieb's theorem and Jensen's inequality for matrices as defined above. Essentially, we used the law of total expectation to derive a non-random "**B**" matrix so that we could apply the combination of Lieb's theorem and Jensen's inequality. We repeat this process to produce the final line.

This leads to the "master tail bound theorem", from which various inequalities can be derived:

$$\mathbb{P}(\lambda_{\max}(\sum_{i=1}^{n} X_i) \geq t) \leq \inf_{\lambda>0} \left\{ e^{\lambda t} \operatorname{tr} \exp(\sum_{i=1}^{n} \log \mathbb{E}[e^{\lambda X_i}]) \right\}, \forall t \in \mathbb{R} \tag{7.29}$$

All that remains is to bound $\mathbb{E}[e^{\lambda X_i}]$ (to be continued in the next lecture...) ■