THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 384 11: Theoretical Statistics

Lecture 18: Concentration of vectors and matrices

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

Recall the Markov trick?

- Consider $X_1, \dots X_n \in \mathbb{R}$ mean zero subGaussian random variables.
- $E[\exp(\lambda X_1)] \leq \exp(\lambda^2 \sigma^2/2)$
-

$$
\begin{aligned}
P(\sum_i X_i \geq t) &\leq \exp(-t\lambda) E \exp(\lambda \sum_i X_i) \\
&\leq \exp(-t\lambda) \exp(n\lambda^2 \sigma^2/2) \\
&\leq \inf_{\lambda \geq 0} \exp(-t\lambda) \exp(n\lambda^2 \sigma^2/2)
\end{aligned}
$$

What if X_i were vectors

- What is a subGaussian vector?
- $X \in \mathbb{R}^d$ is subGaussian(σ) if $a^T X$ is subGaussian(σ) for all unit vector $a \in \mathbb{R}^d$
- Prove that - given $X_1, \ldots X_n \sim N(\mu, \Sigma)$,

$$P\left(\|\bar{X}_n - \mu\| \geq \sqrt{\frac{Trace(\Sigma)}{n}} + c\sqrt{\frac{\|\Sigma\| \log(1/\delta)}{n}} \right) \leq \delta$$

How will you prove this?

- $\bar{X}_n - \mu \sim N(0, \Sigma/n)$
- So we can worry about $\|\Sigma_n^{1/2} Z\|$, where $\Sigma_n = \Sigma/n$ and $Z \sim N(0, I_n)$.
- $f(z) = \|Az\|$ - is it Lipschitz?
- Lipschitz: $\|f(z) - f(u)\| \leq \|A(z-u)\| \leq \|A\|_{op}\|z-u\|$
- So yes,lipshitz with $L = \|A\|_{op}$

Recall Gaussian concentration?

Theorem

If $f(Z)$ is a convex $L-$ Lipschitz function of independent Gaussians $Z \sim N(0, I_n)$ then,

$$P(|f(z) - E[f(z)]| \geq t) \leq 2\exp(-Ct^2/L^2)$$

- Apply to $f(Z) = \|\Sigma_n^{1/2} Z\|$ with $L = \|\Sigma_n\|^{1/2}$

$$\begin{aligned} &P(\|\bar{X}_n - \mu\| \geq E\|\bar{X}_n - \mu\| + t) \\ &= P(f(Z) \geq E[f(Z)] + C' L\sqrt{\log(1/\delta)}) \leq \delta \end{aligned}$$

- Now
$E\sqrt{Z^T \Sigma_n Z} \leq \sqrt{E[Z^T \Sigma_n Z]} = \sqrt{E[Trace(Z^T \Sigma_n Z)]} \leq \sqrt{trace(\Sigma_n)}$

Gaussian conc revisited

- $Ef(Z) \leq \sqrt{E[f(Z)]} = \sqrt{E[Trace(z^T \Sigma_n z)]} \leq \sqrt{trace(\Sigma_n)}$
- Putting everything together

$$P(f(Z) \geq \sqrt{trace(\Sigma_n)} + C' L \sqrt{\log(1/\delta)}) \leq \delta$$

$$P(f(Z) \geq \sqrt{trace(\Sigma)/n} + C' \sqrt{\|\Sigma\|_{op}/n} \sqrt{\log(1/\delta)}) \leq \delta$$

How about matrices?

- Let $X_1, \dots X_n$ be $d \times d$ symmetric independent mean zero matrices.
- We are interested in the quantity:

$$P(\| \sum_i X_i \| \geq t)$$

- First lets talk about the matrix $\exp(A)$ where A is symmetric $d \times d$ matrix.
- $\exp(A)$ is also a symmetric matrix with eigenvalues $\exp(\lambda_i(A))$
- What are the main shifts from scalars to matrices?

From scalar to matrices

- A, B commute when $AB = BA$.
- $\exp(A+B) = \exp(A)\exp(B)$ when A,B are scalars.
- For matrices, only true when A and B commute.
- Furthermore, $Tr(\exp(A+B))$ is not equal to $Tr(\exp(A)\exp(B))$ when A, B don't commute.
- However, the celebrated Goldman-Thompson theorem says that:

$$Tr(\exp(A+B)) \leq Tr(\exp(A)\exp(B))$$

- Naive generalization to three matrices does not work.

Simple trace inequalities

- Von Neumann's trace identity:
- Let A, B denote $n \times n$ matrices with singular values $\sigma_i(A), \sigma_i(B), i \in [n]$.

$$Tr(AB) \leq \sum_i \sigma_i(A)\sigma_i(B) \leq \|B\| \sum_i \sigma_i(A)$$

- If A, B are symmetric PSD,

$$Tr(AB) \leq \|B\| Trace(A)$$

How about matrices?

- Let $X_1, \dots X_n$ be $d \times d$ symmetric independent mean zero matrices.
- We are interested in the quantity:

$$P(\| \sum_i X_i \| \geq t)$$

- First lets talk about the matrix $\exp(A)$ where A is symmetric $d \times d$ matrix.
- $\exp(A)$ is also a symmetric matrix with eigenvalues $\exp(\lambda_i(A))$
- Let $S_n = \sum_i X_i$ and so

$$\begin{aligned} P(\exists i \in [d], \lambda_i(S_n) \geq t) &\leq P(\mathit{Tr}(\exp(\lambda S_n)) \geq \exp(\lambda t)) \\ &\leq E \exp(-\lambda t)) \exp(\mathit{Tr}(\exp(\lambda(S_n))) \end{aligned}$$

How about matrices?

- Let $S_n = \sum_i X_i$ and so

$$
\begin{aligned}
P(\exists i \in [d], \lambda_i(S_n) \geq t) &\leq P(Tr(\exp(\lambda S_n)) \geq \exp(\lambda t)) \\
&\leq \exp(-\lambda t) E Tr(\exp(\lambda S_{n-1} + \lambda X_n)) \\
&\leq \exp(-\lambda t) E\left[Tr\left(\exp(\lambda S_{n-1}) \exp(\lambda X_n)\right)\right] \\
&= \exp(-\lambda t)\, Tr\left(E\left[\exp(\lambda S_{n-1}) \exp(\lambda X_n)\right]\right) \\
&= \exp(-\lambda t)\, Tr\left(E\left[\exp(\lambda S_{n-1})\right] E\left[\exp(\lambda X_n)\right]\right) \\
&\leq \exp(-\lambda t) \|E[\exp(\lambda X_n)]\|\, Tr(E[\exp(\lambda S_{n-1})]) \\
&\leq d \exp(-\lambda t) \prod_{i=1}^{n} \|E[\exp(\lambda X_i)]\|
\end{aligned}
$$

How about matrices?

- Let $S_n = \sum_i X_i$ and so

$$P(\|S_n\| \geq t) \leq P(\exists i \in [d], \lambda_i(S_n) \geq t) + P(\exists i \in [d], \lambda_i(S_n) \leq -t)$$
$$\leq d \exp(-\lambda t) \left(\prod_{i=1}^{n} \|E[\exp(\lambda X_i)]\| + \prod_{i=1}^{n} \|E[\exp(-\lambda X_i)]\| \right)$$

- Assume $\|X_i\| \leq 1$ and $\|E[X_i^2]\| \leq B$.
- Recall that for $x \in [-1, 1]$, $\exp(x) \leq 1 + x + x^2$
- So, for $|\lambda| \leq 1$, $\exp(\lambda X_i) \preceq 1 + \lambda X_i + \lambda^2 X_i^2$
- $\|E(\exp(\lambda X_i))\| \leq 1 + \lambda \|EX_i\| + \lambda^2 \|EX_i^2\|$

How about matrices?

- Let $S_n = \sum_i X_i$ and so

$$P(\|S_n\| \geq t) \leq d \exp(-\lambda t) \prod_{i=1}^{n} (\|E[\exp(\lambda X_i)]\| + \|E[\exp(-\lambda X_i)]\|)$$
$$\leq 2d \exp(-\lambda t + nB\lambda^2)$$

- Optimize over $\lambda \in [-1, 1]$, to get:

$$t = Bn\epsilon \qquad \lambda = \frac{\epsilon}{2}$$

$$P(\|S_n\| \geq t) \leq 2d \exp(-n\epsilon^2 B/4)$$

Covariance estimation

Theorem

Let $\epsilon := \sqrt{\frac{4\log(2d/\delta)}{nB}}$. *Assume* $\|X_i\| \leq 1$ *and* $\|E[X_i^2]\| \leq B$. *We have:*

$$P\left(\|S_n\| \geq 2\sqrt{Bn\log(2d/\delta)}\right) \leq \delta$$

- Consider $Y_1, \ldots, Y_n \in \mathbb{R}^d$. Let $\|Y_i\| \leq R$.
- Let $EY_1 = 0$ and $E[Y_1 Y_1^T] = \Sigma$, $\lambda_1 = \|\Sigma\|$
- Let $X_1 = \frac{Y_1 Y_1^T - \Sigma}{R^2 + \lambda_1}$ and

$$\|E[X_1^2]\| = \frac{\|E[(Y_1 Y_1^T)^2] - \Sigma^2\|}{(R^2+\lambda_1)^2} \leq \frac{R^2\lambda_1}{(R^2+\lambda_1)^2} \leq \frac{\lambda_1}{R^2+\lambda_1}$$

Covariance estimation

- Let $S_n = \sum_i X_i$ and $\hat{\Sigma}_n = \sum_i Y_i Y_i / n$.
- Apply the last theorem. $P\left(\|S_n\| \geq 2\sqrt{\frac{n\lambda_1 \log(2d/\delta)}{\lambda_1 + R^2}} \right) \leq \delta.$
- Now note that $\hat{\Sigma}_n - \Sigma = (R^2 + \lambda_1) S_n$.

$$P\left(\|\hat{\Sigma}_n - \Sigma\| \geq 2\sqrt{\frac{\lambda_1(\lambda_1 + R^2)\log(2d/\delta)}{n}} \right) \leq \delta$$

Matrix Bernstein

Theorem

Let $Y_1, \ldots, Y_n$ be mean-zero IID random vectors. Assume $\|Y_i Y_i^T - E[Y_i Y_i^T]\| \leq C$ and $\|E[(Y_i Y_i^T)^2]\| = \sigma^2$. We have:

$$P\left(\|\sum_i (Y_i Y_i^T - \Sigma)\| \geq t\right) \leq 2d \exp\left(\frac{-t^2/2}{n\sigma^2 + Ct/3}\right)$$

- With probability $1-\delta$, $\|\hat{\Sigma}_n - \Sigma\| = 2 \cdot \max\left\{\sqrt{\frac{\sigma^2}{n} \log\frac{d}{\delta}}\,,\ \frac{C}{n}\log\frac{d}{\delta}\right\}$

Covariance estimation for Gaussians

- $X_1, \ldots, X_n \sim N(0, I)$
- Let $\mathcal{G} = \{\max_i \|X_i\| \leq \sqrt{d} + c\sqrt{n/\delta}\}$
- How do I use the previous result?

Covariance estimation for Gaussians

- Define the event $\mathcal{E} := \left\{X : \|X\|^2 \leq \tau\right\}$, $\tau := d + 3\sqrt{d \log\left(\frac{1}{\delta}\right)}$
- Recall the chi-squared tail bound (Lemma 1, Laurent and Massart, 2000) ,

$$\mathbb{P}(\|X\|^2 \geq d + \sqrt{2dt} + 2t) \leq e^{-t}$$

- Therefore, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. Furthermore, conditioned on the event $\mathcal{E}$,

$$\left\|XX^\top - I\right\| \leq \tau + 1 \leq 2\tau$$

- Next, we need to bound $\left\|\mathbb{E}[XX^\top|\mathcal{E}] - I\right\|$ and $\left\|\mathbb{E}[(XX^\top)^2|\mathcal{E}]\right\|$

Covariance estimation for Gaussians

Note that

$$\left\|\mathbb{E}[XX^\top|\mathcal{E}] - I\right\| = \left\|\mathbb{E}[(XX^\top - I)|\mathcal{E}]\right\| = \frac{1}{\mathbb{P}(\mathcal{E})}\left\|\mathbb{E}[(XX^\top - I)\mathbb{1}(\mathcal{E})]\right\|$$

$$\begin{aligned}
\left\|\mathbb{E}[(XX^\top - I)\mathbb{1}(\mathcal{E})]\right\| &= \left\|\mathbb{E}[(XX^\top - I)\mathbb{1}(\mathcal{E}^\complement)]\right\| \\
&= \sup_{v\in\mathcal{R}^d,\|v\|=1} \mathbb{E}[((v^\top X)^2 + 1)\mathbb{1}(\mathcal{E}^\complement)] \\
&\leq 2 \sup_{v\in\mathcal{R}^d,\|v\|=1} \sqrt{\mathbb{E}[(v^\top X)^4 + 1]\mathbb{P}(\mathcal{E}^\complement)} \\
&\leq C\sqrt{\delta} \sup_{v\in\mathcal{R}^d,\|v\|=1} \mathbb{E}[(v^\top X)^2] = C\sqrt{\delta}(\left\|\mathbb{E}[XX^\top]\right\| + 1) \\
&\leq C\sqrt{\delta}
\end{aligned}$$

The last line follows because for $v \in \mathcal{R}^d, \|v\| = 1$, $(v^\top X)$ is a standard normal random variable and it satisfies $E[(v^T X)^4] \leq C^2 E[(v^T X)^2]^2$ for some $C \geq 1$.

Covariance estimation for Gaussians

- Similarly, $\left\|\mathbb{E}[(XX^\top)^2|\mathcal{E}]\right\| = \left\|\mathbb{E}[XX^\top \|X\|^2 |\mathcal{E}]\right\| \leq \tau \left\|\mathbb{E}[XX^\top|\mathcal{E}]\right\|$
- From previous result, $\left\|\mathbb{E}[XX^\top|\mathcal{E}]\right\| \leq 1 + \frac{\sqrt{\delta}}{1-\delta}$

Putting everything together,

$$\left\|\mathbb{E}[XX^\top|\mathcal{E}] - I\right\| \leq \frac{\sqrt{3\delta}}{1-\delta}, \quad \left\|\mathbb{E}[(XX^\top)^2|\mathcal{E}]\right\| \leq \tau\left(1 + \frac{\sqrt{3\delta}}{1-\delta}\right), \text{ and}$$

$$\left\|XX^\top - I\right\| \leq 2\tau, \text{ conditioned on } \mathcal{E}$$

Now we are ready to apply Matrix-Bernstein.

Covariance estimation for Gaussians

Let $X_1, X_2, \cdots X_n \sim \mathcal{N}(0, I)$ for $n \geq 2$. Define $\tau_n := d + 3\sqrt{d \log\left(\frac{n^2}{\delta}\right)}$ and events $\mathcal{E}_i := \left\{\|X_i\|_2^2 \leq \tau_n\right\}$. Then, for each X_i, $\mathbb{P}(\mathcal{E}_i) \geq 1 - \frac{\delta}{n^2}$.

Let $\mathcal{A} := \cup_{i \in [n]} \mathcal{E}_i$. Then, by a union-bound, $\mathbb{P}(\mathcal{A}) \geq 1 - \delta/n$.
Furthermore, conditioned on $\mathcal{A}$, via Matrix-Bernstein and the calculation on the previous slide,

$$\left\|\hat{\Sigma}_n - \mathbb{E}\left[XX^\top | \mathcal{A}\right]\right\| \leq 2 \cdot \max\left\{\sqrt{\frac{\mathcal{V}^2}{n} \log\frac{d}{\delta}}\,,\ \frac{\mathcal{M}}{n} \log\frac{d}{\delta}\right\}$$

where, since $\delta \in (0, 1)$, $\mathcal{V}^2, \mathcal{M} \lesssim \tau_n$.

Finally, we note that $\left\|\mathbb{E}\left[XX^\top | \mathcal{A}\right] - I\right\| \leq 4\sqrt{\frac{\delta}{n^2}}$.

Frame Title

Putting everything together, with probability atleast $1 - \delta$,

$$\left\|\hat{\Sigma}_n - I\right\| \leq \max\left\{\sqrt{\frac{\tau_n}{n}\log\frac{d}{\delta}}\,,\ \frac{\tau_n}{n}\log\frac{d}{\delta}\right\} + \frac{4}{n}$$

where $\tau_n := d + 3\sqrt{d\log\left(\frac{n^2}{\delta}\right)}$.