MATHEMATICS OF DATA SCIENCE

December 12, 2023

Today: Last lecture (and last concentration inequalities)

RANDOM MATRICES AND COVARIANCE ESTIMATION

(1) Random matrices

→ Random matrices can be generated from random variables, random vectors or directly from a random matrix distribution

Example of matrix distribution: Wishart distribution

$A \in \mathbb{R}^{m \times m}$ follows a Wishart distribution $W_n(\Sigma, n)$ where $m \geq 1$, $n \geq 1$ and $\Sigma = \Sigma^T > 0$ if its probability density is

$$p(A = B) = \frac{1}{2^{\frac{mn}{2}} \det(\Sigma)^{\frac{n}{2}} \Gamma_n\left(\frac{n}{2}\right)} \det(B)^{\frac{n-m-1}{2}} e^{-\frac{1}{2}\operatorname{trace}(\Sigma^{-1}B)}$$

Γ_n: Gamma function

→ Generalization of the χ^2 distribution ("chi-squared")

↳ Concentration inequalities for random matrices are obtained thanks to assumptions on the coefficients of those matrices or on the matrices themselves.

Theorem Let $A \in \mathbb{R}^{m \times n}$ be a rectangular random matrix with independent, mean zero, subgaussian entries.

(Every A_{ij} is mean zero and subgaussian $\forall i = 1..m$, $j = 1..n$, and the A_{ij}'s are independent)

Then $\exists C>0$ such that $\forall t \geq 0$,

$$\mathbb{P}\left(\|A\| \leq CK(\sqrt{m}+\sqrt{n}+t) \right) \geq 1-2\exp(-t^2)$$

where $\|A\| = \max_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|}$ and $K = \max_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \|A_{ij}\|_{\psi_2}$ (subgaussian norm)

Corollary Let $A \in \mathbb{R}^{m \times m}$ is a random symmetric matrix such that $\{A_{ij}\}_{1 \leq i \leq j \leq m}$ (because of symmetry) are independent, mean zero, subgaussian random variables.

Then, $\forall t \geq 0$,

$$\mathbb{P}(\|A\| \leq CK(\sqrt{m}+t)) \geq 1-4\exp(-t^2)$$

where $C>0$, $K = \max_{1 \leq i \leq j \leq m} \|A_{ij}\|_{\psi_2}$.

Proof: Split A into its upper and lower triangular part

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{1m} & \cdots & A_{mm} \end{bmatrix} = \underbrace{\begin{bmatrix} A_{11} & & A_{1m} \\ & \ddots & \\ 0 & & A_{mm} \end{bmatrix}}_{A^u} + \underbrace{\begin{bmatrix} 0 & & 0 \\ A_{12} & \ddots & \\ A_{1m} & A_{m-1,m} & 0 \end{bmatrix}}_{A^l}$$

$$\|A\| \leq \|A^u\| + \|A^l\|$$

$$\mathbb{P}(\|A\| \leq CK(\sqrt{m}+t)) \geq \mathbb{P}(\|A^u\| \leq CK(\sqrt{m}+t) \text{ and } \|A^l\| \leq CK(\sqrt{n}+t))$$

$$= 1 - \mathbb{P}\left(\|A^{(1)}\| > CK(\sqrt{n}+t) \text{ or } \|A^{(2)}\| > CK(\sqrt{m}+t) \right)$$

$$\geq 1 - \mathbb{P}\left(\|A^{(1)}\| > CK(\sqrt{n}+t) \right) - \mathbb{P}\left(\|A^{(2)}\| > CK(\sqrt{m}+t) \right)$$

$$= 1 - \left(1 - \mathbb{P}\left(\|A^{(1)}\| \leq CK(\sqrt{n}+t) \right)\right) - \left(1 - \mathbb{P}\left(\|A^{(2)}\| \leq CK(\sqrt{m}+t) \right)\right)$$

$$= \mathbb{P}\left(\|A^{(1)}\| \leq CK(\sqrt{n}+t) \right) + \mathbb{P}\left(\|A^{(2)}\| \leq CK(\sqrt{m}+t) \right) - 1$$

↑ Independent entries

↑ Independent entries

$\mathbb{P}(E_1 \cup E_2) \leq \mathbb{P}(E_1) + \mathbb{P}(E_2)$

↑ "Union bound"

Matrix Bernstein's inequality (2010)

Let $A_1, \dots, A_N \in \mathbb{R}^{m \times m}$ independent, zero mean, random matrices in $\mathbb{R}^{m \times m}$ such that A_i is symmetric and $\|A_i\| \leq K$ almost surely for every $i = 1 \dots N$

almost surely ↓ $\mathbb{P}_{A_i}(\|A_i\| \leq K) = 1$

Then, for any $t \geq 0$,

$$\mathbb{P}\left(\left\| \sum_{i=1}^N A_i \right\| \geq t \right) \leq \cdot 2m \exp\left(\frac{-t^2/2}{\sigma^2 + \frac{Kt}{3}} \right)$$

where $\sigma^2 = \left\| \sum_{i=1}^N \mathbb{E}(A_i^2) \right\|$ "norm of the matrix variance"

Recall. Bernstein's inequality for independent, zero mean and subexponential variables $y_1, \dots, y_N$

$$\forall t \geq 0, \quad \mathbb{P}\left(\left| \sum_{i=1}^N y_i \right| \geq t \right) \leq 2 \exp\left(-c \min\left\{ \frac{t^2}{N \max_{1 \leq i \leq N} \|y_i\|_{\psi_1}}, \frac{t}{\max_{1 \leq i \leq N} \|y_i\|_{\psi_1}} \right\} \right)$$

Remark: Matrix Bernstein's inequality does not require the coefficients of $A_1, \dots, A_N$ to be independent

(2) Covariance matrix estimation

↳ Consider $y_1, \dots, y_N$ iid random vectors in $\mathbb{R}^m$

Let $\mu = \mathbb{E}[y_i] \in \mathbb{R}^m$ (mean vector)

and $\Sigma = \mathbb{E}\left[(y_i - \mathbb{E}[y_i])(y_i - \mathbb{E}[y_i])^T \right] \in \mathbb{R}^{m \times m}$

(covariance matrix)

→ Estimating μ : "easy" with $\frac{1}{N} \sum_{i=1}^N y_i$ ⇒ this concentrates around μ

→ Estimating Σ :

Use $\Sigma_N = \frac{1}{N} \sum_{i=1}^N \left(y_i - \frac{1}{N} \sum_{j=1}^N y_j \right) \left(y_i - \frac{1}{N} \sum_{j=1}^N y_j \right)^T$

(if mean known, use $\Sigma_N = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)(y_i - \mu)^T$)

Theorem Suppose that the y_i s satisfy

$$\|y_i\|^2 \leq C \, \mathbb{E}[\|y_i\|^2] \quad \text{almost surely}$$

$\|y_i - \mu\|^2 \lesssim \|y_i - \mu\|^2$ for some $C > 0$

Then,

$$\mathbb{E}_{y_1, \dots, y_N} \left[\|\Sigma_N - \Sigma\| \right] \leq \hat{C} \|\Sigma\| \left(\sqrt{\frac{m \ln(m)}{N}} + \frac{m \ln(m)}{N} \right)$$

Corollary $\mathbb{E}[\|\Sigma_N - \Sigma\|] \leq \varepsilon \|\Sigma\|$ (for small $\varepsilon > 0$) holds

when $N \geq \bar{C}\, \varepsilon^{-2} \ln(m)\, m$ for $\bar{C} > 0$

Proof idea: Apply matrix Bernstein to the random matrices

$$\left\{ \underbrace{(y_i - \mu)}_{m\times 1} \underbrace{(y_i - \mu)^T}_{1\times m} - \underbrace{\Sigma}_{m\times m} \right\}_{i=1..N} \in \mathbb{R}^{m\times m}.$$

→ independent
→ Symmetric
→ zero mean

$$\mathbb{E}\left[(y_i-\mu)(y_i-\mu)^T - \Sigma\right]$$

$$= \mathbb{E}\left[(y_i-\mu)(y_i-\mu)^T\right] - \Sigma = \Sigma - \Sigma = 0$$

To apply matrix Bernstein, we show that $\|(y_i-\mu)(y_i-\mu)^T - \Sigma\| \leq K$ almost surely

$$\|(y_i-\mu)(y_i-\mu)^T - \Sigma\| \leq \|(y_i-\mu)(y_i-\mu)^T\| + \|\Sigma\|$$

$$= \|y_i - \mathbb{E}[y_i]\|^2 + \|\Sigma\|$$

$$\leq \frac{1}{2}\|y_i\|^2 + \frac{1}{2}\|\mathbb{E}[y_i]\|^2 + \|\Sigma\|$$

$$\leq \frac{1}{2} C\, \mathbb{E}[\|y_i\|^2] + \frac{1}{2}\|\mu\|^2 + \|\Sigma\|$$

$$\leq \frac{1}{2}\max(C,1)\Big[\underbrace{\mathbb{E}[\|y_i\|^2] + \|\mu\|^2}_{= \operatorname{trace}(\Sigma)}\Big] + \|\Sigma\|$$

Concluding remarks

→ Concentration inequalities: apply to random variables, random vectors and random matrices

⇒ Guarantee $\mathbb{P}(Y \geq t) \longrightarrow 0$

Exponentially fast as $t \to \infty$

→ The point is NOT to remember all the inequalities!

→ Two cases of study:

- Derive concentration inequalities using convex optimization
- Apply them to provide probabilistic guarantees