Lecture 7: September 21

Lecturer: Alessandro Rinaldo *Scribes: Jackie Mauro*

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

7.1 Recap: covariance matrix estimation

For A an $m \times n$ matrix, take $A = UDV^T$. U, V orthonormal columns, D diagonal. Then:

$$\begin{aligned}\sigma_{max} &= \max_{x \in \mathbb{S}^{n-1}} ||Ax|| \text{ (largest singular value)} \\ &= \max_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{||Ax||}{||x||} \\ &= \max_{\substack{x \in \mathbb{S}^{n-1} \\ y \in \mathbb{S}^{n-1}}} |y^T Ax|\end{aligned}$$

If A $(n \times n)$ is symmetric, $\sigma_{max}(A) = \max_{x \in \mathbb{S}^{n-1}} |x^T Ax|$.

If A $(n \times n)$ is PSD, $\sigma_{max}(A) = \max_{x \in \mathbb{S}^{n-1}} x^T Ax$, the largest eigenvalue of A.

For a generic A $(m \times n)$, $\sigma_{max}(A)$ also called the "operator norm". $||A||_{op}$ is the L_∞ norm of its singular values.

$||A||_F = \sqrt{\sum_{i,j} A_{ij}^2}$ is the "Frobenius" norm. It is the L_2 norm over the singular values.

Nuclear norm of A: $\sum_i \sigma_i$, the L_1 norm of the singular values.

7.2 Operator Norm

Take A, B to be $m \times n$ matrices.. If $||A - B||_{op} \to 0$ then $|y^T Ax - y^T Bx| \to 0$ uniformly over $x \in \mathbb{S}^{n-1}, y \in \mathbb{S}^{n-1}$. And this implies $max_{ij}|A_{ij} - B_{ij}| \to 0$.

If Σ is the covariance matrix and $\hat{\Sigma}$ an estimator of it (both PSD), then:

$$\begin{aligned}||\Sigma - \hat{\Sigma}||_{op} &\to 0 \\ \Rightarrow \max_{v \in \mathbb{S}^{n-1}} |v^T \Sigma v - v^T \hat{\Sigma} v| &\to 0 \\ \Rightarrow \max_{v \in \mathbb{S}^{n-1}} |\mathbb{V}(v^T X) - \mathbb{V}(v^T \tilde{X})| &\to 0\end{aligned}$$

Where $X \sim (\mu, \Sigma)$ and $\tilde{X} \sim (\tilde{\mu}, \hat{\Sigma})$.

7.3 Weyl Inequality

A, B are $m \times n$ with singular values:

$$\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{min(n,m)}(A)$$
$$\sigma_1(B) \geq \sigma_2(B) \geq \cdots \geq \sigma_{min(n,m)}(B)$$

$\Rightarrow \max_{k=1,\cdots,min(m,n)} |\sigma_k(A) - \sigma_k(B)| \leq ||A-B||_{op} = \sigma_{max}(A-B)$

Recall: a random vector $X \in SG(\sigma^2)$ if:

$$\mathbb{E}(e^{\lambda v^T X}) \leq e^{\frac{\lambda \sigma^2}{2}} \tag{7.1}$$

then $X \in SG_d(\sigma^2)$ if its coordinates are independent $SG(\sigma^2)$ or $X \sim N_d(0, \Sigma)$ with $\sigma^2 = ||\Sigma||_{op}$ because:

$$\mathbb{V}(v^T X) = v^T \Sigma v \Rightarrow \max_{v \in \mathbb{S}^{d-1}} (v^T \Sigma v) = ||\Sigma||_{op} \tag{7.2}$$

Theorem 7.1 *If $X_1, \cdots, X_n \overset{iid}{\sim} (0, \Sigma)$ in $\mathbb{R}^d$ and $\in SG(\sigma^2)$, then, setting*

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \tag{7.3}$$

there exists a constant $c > 0$ such that:

$$\mathbb{P}(||\Sigma - \hat{\Sigma}||_{op} \leq \sigma^2 C \min\{\sqrt{\frac{d + log(2/\delta)}{n}}, \frac{d + log(2/\delta)}{n}\}) \geq 1 - \delta \tag{7.4}$$

for $\delta \in (0,1)$

This implies that if $\Sigma = I, \sigma^2 = 1$ then:

$$||\hat{\Sigma} - I||_{op} \leq \sqrt{\frac{d}{n}} + \frac{d}{n} \tag{7.5}$$

with high probability

Consistency requires $d = o(n)$. Unless you make sparsity assumptions on Σ you must have d grow slowly with n.

Proof: The proof uses a discretization argument. Operator norm is the max over an infinite set, so we need to discretize. Take $X \in SG(\sigma^2), X - E(X^2) \in SE(\nu^2, \alpha), \nu = \alpha = 16\sigma^2$.

We also need the discretization lemma:

Lemma 7.2 *Let A ($n \times n$) symmetric (will eventually be $\Sigma - \hat{\Sigma}$) and $\mathcal{N}_\epsilon$ be an ϵ-net of $\mathbb{S}^{n-1}$. Then*

$$||A||_{op} = \max_{x \in \mathbb{S}^{n-1}} |x^T A x| \leq (1-2\epsilon)^{-1} \max_{y \in \mathcal{N}_\epsilon} (y^T A y) \tag{7.6}$$

Proof: Let $x^* \in \mathbb{S}^{n-1}$ st $||A||_{op} = |x^{*T} A x^*|$. Let $y \in \mathcal{N}_\epsilon$ st $||x^* - y|| \leq \epsilon$. Then:

$$\begin{aligned}
|x^T A x^* - y^T A y| &= |x^T A(x^* - y) + y^T A(x^* - y)| \text{ by symmetry} \\
&\leq |x^T A(x^* - y)| + |y^T A(x^* - y)| \\
&\leq ||x^*|| |A(x^* - y)| + ||y|| ||A(x^* - y)| \\
&\leq 2||A||_{op} ||x^* - y|| \\
&\leq 2\epsilon ||A||_{op}
\end{aligned}$$

Where the second to last inequality follows from $||Az|| \leq ||A||_{op}||z||$.

This gives:

$$\begin{aligned} |y^T Ay| &\geq |x^{*T}Ax^*| - 2\epsilon||A||_{op} &(7.7)\\ \Rightarrow ||A||_{op} &\leq \frac{1}{1-2\epsilon}|y^T Ay| &(7.8)\\ &\leq \frac{1}{1-2\epsilon}\max_{y\in\mathbb{S}^{n-1}}|y^T Ay| &(7.9)\end{aligned}$$

■

Now set $A = \hat{\Sigma} - \Sigma$ ($d \times d$ and symmetric) and consider $\mathcal{N}_{\frac{1}{4}}$ a $1/4$ - net of $\mathbb{S}^{d-1}$, then:

$$||\hat{\Sigma} - \Sigma||_{op} = ||A||_{op} \leq 2\max_i |v_i^T A v_i| \tag{7.10}$$

where $\{v_i, \cdots, v_n\} = \mathcal{N}_{\frac{1}{4}}$. Note that $|\mathcal{N}_{\frac{1}{4}}| \leq 9^d$ because it is a volume calculation. So, $\forall t > 0$:

$$\begin{aligned} \mathbb{P}(||\hat{\Sigma} - \Sigma||_{op} \geq t) &\leq \mathbb{P}(\max_i |v_i^T(\hat{\Sigma} - \Sigma)v_i| \geq t/2)\\ &\leq \sum_{i\leq 9^d}\mathbb{P}(|v_i^T(\hat{\Sigma} - \Sigma)v_i| \geq t/2)\end{aligned}$$

for a fixed $v \in \mathbb{S}^{d-1}$,

$$\begin{aligned} v^T(\hat{\Sigma} - \Sigma)v &= \frac{1}{n}\sum_{j=1}^{n}(v^T X_i)^2 - v^T\Sigma v\\ \text{Note: } v^T\hat{\Sigma}v = v^T(\frac{1}{n}\sum_{j=1}^{n}X_iX_i^T)v &= \frac{1}{n}\sum_{j=1}^{n}v^T X_i X_i^T v = \frac{1}{n}\sum_{j=1}^{n}(v^T X_i)^2\\ &= \frac{1}{n}\sum_{j=1}^{n}[Z_i^2 - \mathbb{E}(Z_i^2)]\end{aligned}$$

where $Z_i = v^T X_i, \Sigma = \mathbb{E}(XX^T)$ We know $Z_i^2 - \mathbb{E}(Z_i^2) \in SE(\nu^2, \alpha), Z_i iid$. For each $v_i \in \mathcal{N}_{\frac{1}{4}}$, by Bernstein:

$$\mathbb{P}(|v_i(\hat{\Sigma} - \Sigma)v_i| \geq t/2) \leq 2exp\{-\frac{n}{2}min\{(\frac{t}{32\sigma^2})^2, \frac{t}{32\sigma^2}\}\}(*) \tag{7.11}$$

Then :

$$\begin{aligned} \mathbb{P}(\frac{||\hat{\Sigma} - \Sigma||}{\sigma^2} \geq t) &\leq 2\cdot 9^d \cdot (*)\\ \text{set: } &\leq \delta\\ \Rightarrow \frac{t}{32} &\geq \sigma\min\{\frac{2}{n}dlog(9) + \frac{2}{n}log(2/\delta), \sqrt{\frac{2}{n}dlog(9) + \frac{2}{n}log(2/\delta)}\}\end{aligned}$$

■

7.4 (Sparse) Linear Models

Setup: $Y = X\beta^* + \epsilon$, with $\epsilon_1, \cdots, \epsilon_n$ independent $SG(\sigma^2)$

Generally, X is considered fixed [Buja15].

There are 2 settings we are interested in where d grows with n (or even $d > n$).

- Prediction
- Estimation

7.4.1 Prediction

Prediction or mean estimation. Suppose we observe a new batch of data $\tilde{Y}$ and want to estimate β^* with $\hat{\beta}$ and we would like to predict Y as follows:

$$\text{minimize: } \frac{1}{n}\mathbb{E}[||\tilde{Y} - X\hat{\beta}||^2] \tag{7.12}$$

this is the same as minimizing $\frac{1}{n}\mathbb{E}[||X(\beta^* - \hat{\beta})||^2] + \mathbb{E}[||\epsilon||^2]$. So we minimize $\frac{1}{n}\mathbb{E}[MSE(X\hat{\beta})]$:

$$MSE(X\hat{\beta}) = ||X\beta^* - X\hat{\beta}||^2, \hat{\beta} = f(Y) \tag{7.13}$$

7.4.2 Parameter estimation

minimize $\mathbb{E}[||\beta^* - \hat{\beta}||^2]$

Prediction is simpler, because parameter estimation requires the *true model*

7.5 Least Squares in High Dimensions

Usual: $\hat{\beta}^{LS} = (X^TX)^{-1}X^TY$ if $(X^TX)^{-1}$ exists. But $\hat{\beta}^{LS}$ is not defined if $d > n$ or X is rank deficient (linearly dependent).

Still, you can find a solution to:

$$\min_{\beta \in \mathbb{R}^d} ||Y - X\beta||^2 \tag{7.14}$$

The function $\beta \to ||Y - X\beta||^2$ is convex.

To find its minimum, we set the gradient to zero:

$$\Rightarrow X^TX\beta = X^TY \tag{7.15}$$

Any β satisfying this will be a minimum.

If $(X^TX)^{-1}$ does not exist, we have infinitely many solutionns. But if all we want is $X\beta$ (rather than β) we can take any such solution.

References

[Buja15] A. Buja, R. Berk,L. Brown, E. George,E. Pitkin,M. Traskin, L. Zhao and K. Zhang "Models as Approximations–A Conspiracy of Random Regressors and Model Deviation Against Classical Inference in Regression," *Submitted to Statistical Science*, 2015.