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

SDS 384 11: Theoretical Statistics

Lecture 18: Covariance estimation

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

Covariance estimation

Theorem

Let $X_1, \ldots X_n \in \mathbb{R}^d$ denote IID draws from a mean-zero sub-gaussian(σ^2) distribution. Let $\hat{\Sigma} = \sum_i X_i X_i^T / n$. Then we have, for $n = \Omega(\log(1/\delta))$,

$$P\left(\|\hat{\Sigma} - \Sigma\| \geq C\sigma^2\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{d}{n}} + \frac{d}{n}\right)\right) \leq \delta$$

Step 1 - discretization

- Let $\mathcal{S}_{d-1} = \{u \in \mathbb{R}^d | \|u\| = 1\}$
- Consider an ϵ cover $\mathcal{N}$ of $\mathcal{S}_{d-1}$
- Let $M = \hat{\Sigma} - \Sigma$
- Recall that $\|M\| = \sup_{u \in \mathcal{S}_{d-1}} |u^T M u|$
- Assume that this is achievable at u^*.
- Write $u^* = u + r$ where $u \in \mathcal{N}$ is such that $\|r\| \leq \epsilon$

$$\|M\| \leq |u^T M u| + 2\|r\|\|M\| + 2\|r\|^2\|M\|$$
$$|u^T M u| \geq \|M\|(1 - 2\epsilon - \epsilon^2)$$

- Set $\epsilon = 1/4$ to get

$$|u^T M u| \geq \|M\|/4$$

Step 2 - fixed unit vector

- How big is $P(|u^T Mu| \geq t)$
- If $Y_i \sim subexp(\nu, b)$ then $\sum_i Y_i$ is subexp$(\nu\sqrt{n}, b)$
- $\sum_i Y_i/n \sim subexp(\nu/\sqrt{n}, b/n)$
- Note that

$$u^T Mu = \sum_i \underbrace{\left((u^T X_i)^2 - E[(u^T X_1)^2]\right)}_{subexp(c_1\sigma^2, c_2\sigma^2)} /n$$
$$\sim subexp(c_1\sigma^2/\sqrt{n}, c_2\sigma^2/n)$$

Step 2 cont.

$$
\begin{aligned}
&P(\|M\| \geq t) \\
&\leq P(|u^T M u| \geq t/4) \\
&\leq \begin{cases} \exp\left(-c_1' n t^2/\sigma^4\right) & \text{If } t \leq c''\sigma^2 \\ \exp\left(-c_2' n t/\sigma^2\right) & \text{o.w.} \end{cases}
\end{aligned}
$$

Putting things together

$$P(\|M\| \geq t) \leq P(|u^T Mu| \geq t/4) \leq 9^d \begin{cases} \exp\left(-c_1' nt^2/\sigma^4\right) & \text{If } t \leq c''\sigma^2 \\ \exp\left(-c_2' nt/\sigma^2\right) & \text{o.w.} \end{cases}$$

- Setting the error probabilities to δ, we have:
- $t = C\sigma^2(\sqrt{\log(1/\delta)/n} + \sqrt{d/n} + d/n)$
- When $d = o(n)$, we have concentration.

PCA

- Find $\hat{v}_1 := \max_{u \in \mathcal{S}_{d-1}} u^T \hat{\Sigma} u$
- Let $v_1 := \max_{u \in \mathcal{S}_{d-1}} u^T \Sigma u$
- Let λ_1, λ_2 be the top two eigenvalues of Σ

Theorem (Davis Kahan - simplified)

$$|\sin(\hat{v}_1, v_1)| \leq c \frac{\|\hat{\Sigma} - \Sigma\|}{\lambda_1 - \lambda_2}$$

Theorem (Weyl's inequality - corollary)

Let $\hat{\lambda}_i, i = 1, \ldots, \min(d, n)$ denote the eigenvalues of $\hat{\Sigma}$.

$$\max_i |\hat{\lambda}_i - \lambda_i| \leq \|\hat{\Sigma} - \Sigma\|$$

High dimensional setting

- $d/n \to \gamma$

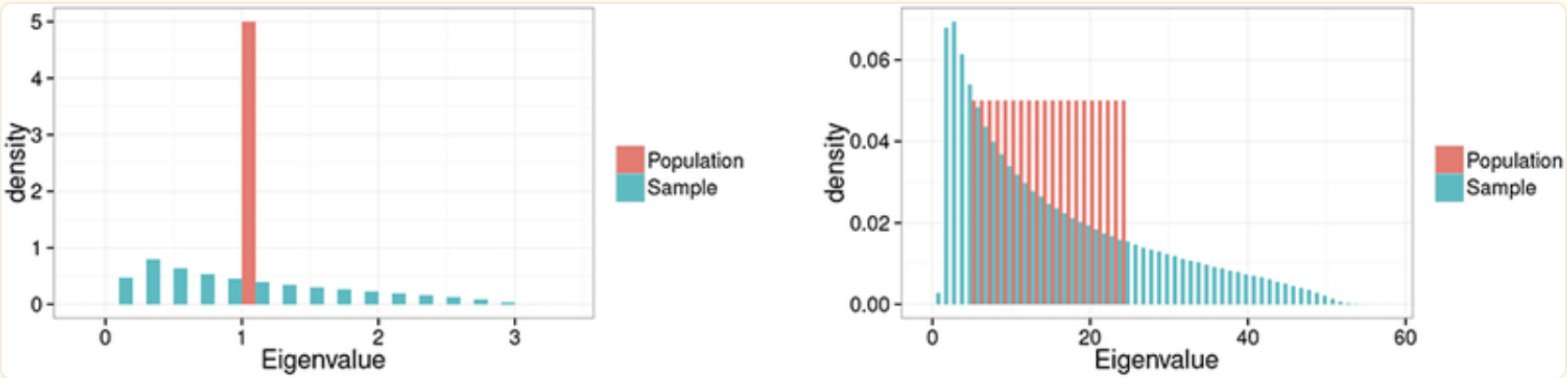


Fig. 2.

Eigenvalue spreading, for $p = 100$, $n = 200$. Population eigenvalues shown as histograms in red: left all at 1, $\sum = I$, right equally spaced on [5, 25]: $\sum = \text{diag}(25,..., 5)$. Corresponding histograms of sample eigenvalues shown in blue. Figure credit: Brett Naul.

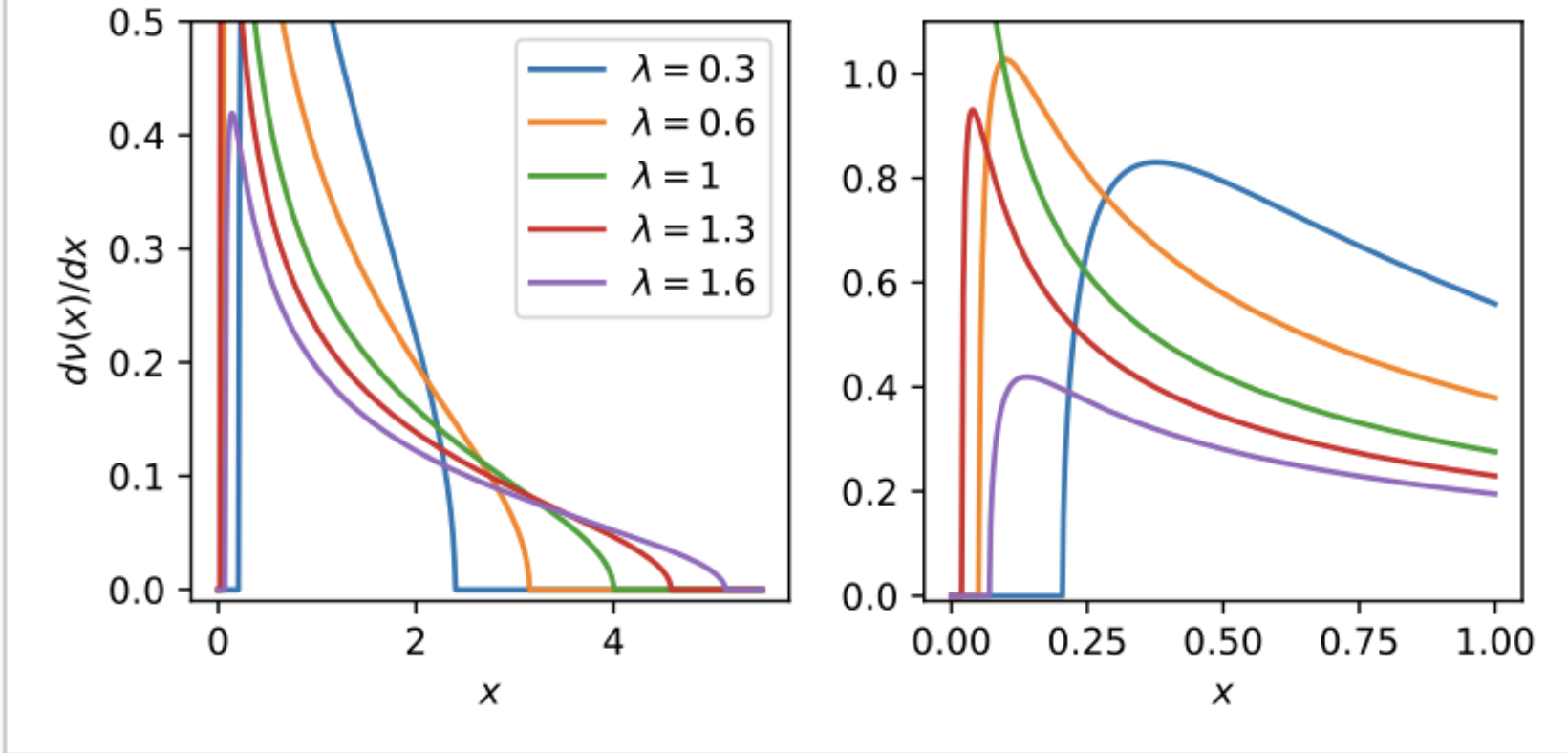
Figure 1: Courtesy: PCA in High Dimensions: An orientation, Johnstone and Paul

The Marcenko-Pastur Law

- Consider the simple case of $\Sigma = I$
- Consider the empirical distribution $F_d(x) := \sum_i 1(\hat{\lambda}_i \leq x)/d$
- When $d/n \to \gamma$, $F_d(x) \to F(x)$ a.s. where

$$f(x) = \frac{1}{2\pi\gamma x}\sqrt{(b_+ - x)(x - b_-)}1(x \in [b_-, b_+])$$

where $b_+ = (1 + \sqrt{\gamma})^2$ and $b_- = (1 - \sqrt{\gamma})^2$.



Plot of the Marchenko-Pastur distribution for various values of lambda

Figure 2: Courtesy: Wikipedia - $\lambda = \gamma$

Covariance matrices with structure

- Σ is sparse
- Goal: estimate Σ
- Given some threshold $\lambda > 0$, define the thresholding operator

$$T_\lambda(x) = x1(|x| > \lambda)$$

- $T_\lambda(M)$ for matrix M is another matrix with the same dimensions that applies the above operator entrywise.

Thresholding based covariance estimation

Theorem

Consider $X_1, \ldots, X_n$ IID mean zero random vectors with common covariance Σ, and $X_{ij} \sim$ subgaussian(σ^2). Let $n \geq \log d$. Then for $\delta > 0$, the thresholding operator with $t_n/\sigma^2 = 8\sqrt{\log d/n} + \delta$ satisfies:

$$P\left(\|T_{t_n}(\hat{\Sigma}) - \Sigma\| \geq 2\|A\| t_n\right) \leq 8\exp(-n\min(\delta, \delta^2)/16)$$

, where $A_{ij} = 1(\Sigma_{ij} \neq 0)$.

- If A is sparse (i.e. Σ is sparse) such that it has at most s nonzero entries per row, then $\|A\| \leq s$.
- Setting $\delta = \sqrt{\log d/n}$, gives:

$$P\left(\|T_{t_n}(\hat{\Sigma}) - \Sigma\| \geq c_1 s\sqrt{\log d/n}\right) \leq 8\exp(-n\min(\delta, \delta^2)/16)$$

Proof - step 1

- Let $c = \max_{ij} |\hat{\Sigma}_{ij} - \Sigma_{ij}|$.
- Let $E_{ij} = |T_c(\hat{\Sigma})(i,j) - \Sigma_{ij}|$
- Note that

$$E_{ij} \begin{cases} \leq 2c & \text{If } A_{ij} = 1 \\ = 0 & \text{If } A_{ij} = 0 \end{cases}$$

- Thus, elementwise, we have:

$$E_{ij} \leq 2cA_{ij}$$

Proof - Step 2

Theorem

Consider two symmetric non-negative square matrices $A, B \in \mathbb{R}^n$ such that $B_{ij} \leq A_{ij}$. We have $\|B\| \leq \|A\|$.

Proof.

- For any integer $m \geq 1$, $B^m(i,j) \leq A^m(i,j)$
- Thus $\|B^m\|_F \leq \|A^m\|_F$ for all $m \geq 1$
- Let the eigenvalues values of A be $|\sigma_1| \geq \cdots \geq |\sigma_n|$
- Similarly, the eigenvalues of B are $|\nu_1| \geq \cdots \geq |\nu_n|$
- Also, $\|A\|_F^2 = \sum_i \sigma_i^2$, so $\|A^m\|_F = (\sum_i \sigma_i^{2m})^{1/2}$
- Now take $m \to \infty$ and use $\|v\|_\infty \leq \|v\|_m \leq n^{1/m}\|v\|_\infty$

□

Step 3

- Let $M = \hat{\Sigma} - \Sigma$
- We have for all $k = \ell$, $X_{ik}^2 \sim \textit{subexponential}(c_1\sigma^2, c_2\sigma^2)$
- We have for all k, $M_{kk} \sim \textit{subexponential}(c_1\sigma^2/n, c_2\sigma^2/n)$
- $$P(|M_{ii}| \geq t) \leq \begin{cases} 2\exp(-cnt^2/2\sigma^4) & \text{If } t \leq c\sigma^2 \\ 2\exp(-c'nt/\sigma^2) & \text{o.w.} \end{cases}$$
- Another way to say this is:

$$P(|M_{ii}|/\sigma^2 \geq t) \leq 2\exp(-cn\min(t, t^2))$$

Off diagonal terms

- We have for all $k \neq \ell$, $\hat{\Sigma}_{k\ell} = \frac{1}{n}\sum_{i=1}^{n} X_{ik}X_{i\ell}$
- $2X_{ik}X_{i\ell} = (X_{ik} + X_{i\ell})^2 - X_{ik}^2 - X_{i\ell}^2$
- $X_{ik} + X_{i\ell} \sim \textit{subgaussian}(2\sigma)$
- Can we remember how to do this?
- Warmup with subGaussian RVs.

Lemma

$X_1, \ldots X_2$ are subGaussian(σ_1), subgaussian(σ_2) RVs. We have:
$X_1 + X_2 \sim \textit{subgaussian}(\sigma_1 + \sigma_2)$

- $E(\exp(t\sum_i X_i) \leq \sqrt{E\exp(2tX_1)E\exp(2tX_2)} \leq \exp(t^2(\sigma_1^2 + \sigma_2^2))$
- So this is subexponential$(\sqrt{2(\sigma_1^2 + \sigma_2^2)})$

Off diagonal terms

- Instead of Cauchy-Schwartz, use Holder with $1/p + 1/q = 1$:

$$E[|XY|] \leq \left(E[|X|^p]\right)^{1/p} \left(E[|Y|^q]\right)^{1/q}$$

$$\begin{aligned} E(\exp(t(X_1 + X_2)) &\leq \left(E\exp(ptX_1)\right)^{1/p} \left(E\exp(qtX_2)\right)^{1/q} \\ &\leq \exp((pt^2\sigma_1^2 + qt^2\sigma_2^2)/2) \end{aligned}$$

- Optimize over p to get $p = (\sigma_1 + \sigma_2)/\sigma_1$

$$E(\exp(t(X_1 + X_2)) \leq \exp(t^2(\sigma_1 + \sigma_2)^2/2) \sim \textit{subgaussian}((\sigma_1 + \sigma_2))$$

- We have for all $k \neq \ell$, $\hat{\Sigma}_{k\ell} = \frac{1}{n}\sum_{i=1}^{n} X_{ik}X_{i\ell}$
- $2X_{ik}X_{i\ell} = (X_{ik} + X_{i\ell})^2 - X_{ik}^2 - X_{i\ell}^2$
- $X_{ik} + X_{i\ell} \sim \mathit{subgaussian}(2\sigma)$
- $(X_{ik} + X_{i\ell})^2 \sim \mathit{subexponential}(c_1'\sigma^2, c_2'\sigma^2)$
- So $\hat{\Sigma}_{k\ell} = \frac{1}{n}\sum_{i=1}^{n} X_{ik}X_{i\ell} = \frac{1}{n}\sum_{i} \frac{(X_{ik} + X_{i\ell})^2 - X_{ik}^2 - X_{jk}^2}{2}$
- So
  $2M_{k\ell} = \frac{1}{n}\sum_{i=1}^{n}\Big((X_{ik} + X_{i\ell})^2 - (\Sigma_{kk} + \Sigma_{\ell\ell} + 2\Sigma_{k\ell})\Big) - M_{kk} - M_{k\ell} \sim$
  $\mathit{subexponential}(d_1\sigma^2/n, d_2\sigma^2/n)$

Putting everything together

- For simplicity take $\sigma = 1$.
- $P(\max_{k,\ell} |M_{k\ell}| \geq t) \leq d^2 \exp(-cn \min(t^2, t))$
- For $t_n = c_1 \sqrt{\log d/n} + \delta$, since $n \geq \log d$, the exponent is of the form $-c' n \min(\log d/n, \sqrt{\log d/n}) - nc'' \min(\delta, \delta^2)$.
- Just pick the constant c_1 to be large enough to cancel out the $2 \log d$ coming from the exponent.