36-755 – Advanced Statistical Theory I **Fall 2017**

Lecture 1: August 28

Lecturer: Alessandro Rinaldo *Scribes: Minshi Peng*

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.*

1.1 Low-dim and high-dim model

1.1.1 Low-dim model

$X = (X_1, \cdots, X_n) \overset{i.i.d}{\sim} P_\theta$. Parametric model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}, \Theta \subseteq \mathbb{R}^d$. We have

- WLLN: $\tilde{\theta}_n \xrightarrow{P} \theta_0$
- CLT: $\sqrt{n} A_n(\tilde{\theta}_n - b_n) \Rightarrow N(0, I_d)$

1.1.2 High-dim model

$\{\mathcal{P}_n\}, n = 1, 2 \cdots$, sequence of parametric models, where $d = d(n) \nearrow \infty$ as $n \to \infty$. WLLN and CLT require fixed d.

1.1.3 How is HD difference from LD

Geometry of HD spaces is different! Concentration of measure phenomenon.[Ball97]

Example 1 Let $B_d(r) = \{x \in \mathbb{R}^d, \|x\|_2 \leq r\}$. The volume of the ball is

$$\text{Vol}\big(B_d(r)\big) = \frac{\pi^{d/2} r^d}{\Gamma(d/2+1)} \sim \Big(\frac{2\pi e r^2}{d}\Big)^{d/2} (d\pi)^{-1/2}$$

where $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, x > 0$. Thus $\text{Vol}(B_d(r)) \to 0$, as $d \to \infty$. Consider now the unit ball for norm $\|x\|_\infty = \max_i |x_i|$, the volume is $\text{Vol}([0,1]^d) = 1$.

Example 2 Let $C_d(\epsilon r) = \{x \in B_d(r), \|x\| > \epsilon r\}, \epsilon \in (0,1)$, $\epsilon = 0.99$ for example.

$$\frac{\text{Vol}\big(C_d(\epsilon r)\big)}{\text{Vol}(B_d(r))} = 1 - \epsilon^d \to 1 \text{ fast}$$

Example 3 $X \sim N(0, I_d)$, with high probability $\|X\|$ tightly concentrates around $\sqrt{d}$.

1.1.4 Statistical examples

1) Covariance matrix estimation. $X_1, \ldots, X_n \sim (0, \Sigma)$ in $\mathbb{R}^d$, $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$. For $A = (A_{ij}), i, j = 1, \cdots, d$, $\|A\|_{\max} = \max |A_{i,j}|$. We want to know $\|\Sigma - \hat{\Sigma}\|_{\max}$.

In low-dim(fixed d), $\hat{\Sigma}_{ij} = \frac{1}{n}\sum_{k=1}^n Z_k^{(i,j)}$, $Z_k^{(i,j)} = X_{k,i}X_{k,j}$, where $X_{k,i}$ is the i-th element of X_k. $Z_k^{(i,j)}$ are i.i.d. By WLLN $\hat{\Sigma}_{ij} \xrightarrow{P} \Sigma_{ij}$. So

$$\|\hat{\Sigma} - \Sigma\|_{\max} \le \sum_{i,j} |\hat{\Sigma}_{ij} - \Sigma_{ij}| = \frac{d(d+1)}{2} o_P(1) = o_P(1)$$

In high-dim, we will see that, under some mild assumptions,

$$\|\hat{\Sigma} - \Sigma\|_{\max} \le C\sqrt{\frac{\log d + \log n}{n}}$$

with high probability, where C is a universal constant. For different norm, we will get different dependence in d.

1.2 Concentration inequalities

References:

- Chapter 2
- Boucheron , Lugosi & Massart: Concentration Inequalities: A Nonasymptotic Theory of Independence
- Concentration of measure for the analysis of randomized algorithm

1.2.1 Motivation

$X_1, \ldots, X_n \overset{i.i.d}{\sim} (\mu, \sigma^2)$, $\bar{X}_n := \frac{1}{n}\sum_i X_i \xrightarrow{P} \mu$, we want to know $\mathbb{P}(|\bar{X}_n - \mu| \ge t) \le ?$ when $t > 0$. By CLT

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \Rightarrow N(0, 1)$$

So $\bar{X}_n$ is a $\sqrt{n}$-consistent estimator of μ $\bar{X}_n = \mu + O_P(\frac{\sigma}{\sqrt{n}})$

$$\mathbb{P}(\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \ge t) \to \mathbb{P}(Z \ge t) \le \frac{1}{2}e^{-t^2/2}$$

where $Z \sim N(0, 1)$. So that

$$\mathbb{P}(\bar{X}_n - \mu \ge t) \le e^{-nt^2/2\sigma^2} \text{approximately}$$

Our goal is to establish such result

- a) For all n (finite sample)
- b) Far all distribution in a large class "Distribution Free".
- c) Dependence on d is explicit.

1.2.2 Markov Inequality

If $X \geq 0$ and $\mathbb{E}[X] < \infty$

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \forall t > 0$$
$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2},\ \sigma^2 = \mathbb{V}[X]$$

If we want to upper bound $\mathbb{P}(|X - \mu| \geq t)$, we could observe

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\mathbb{E}\big[|X - \mu|^k\big]}{t^k},\quad k = 1, 2, \dots$$
$$\implies \mathbb{P}(|X - \mu| \geq t) \leq \min_{k=1,2,\cdots} \frac{\mathbb{E}\big[|X - \mu|^k\big]}{t^k}$$

This is a good bound but we need to know all moments of X which requires strong and unrealistic assumptions on X.

1.2.3 Chernoff Bound

Let, for $\lambda \in \mathbb{R}$, $\psi_X(\lambda) = \log \mathbb{E}[e^{\lambda(X-\mu)}]$ and assume it exists $\forall |\lambda| < b \leq \infty$.

$$\begin{aligned}\mathbb{P}(X - \mu \geq t) =& \mathbb{P}(e^{X-\mu} \geq e^t)\\ =& \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda t}),\ \lambda > 0\\ \leq& \mathbb{E}[e^{\lambda(X-\mu)}]e^{-\lambda t} \text{ by Markov iequality}\\ =& \exp\big\{\psi_X(\lambda) - \lambda t\big\}\end{aligned}$$

Which implies $\mathbb{P}(X_\mu) \geq t \leq \exp(-\psi_\lambda^*(t))$ where $\psi_\lambda^*(t) = \sup_{\lambda\in(0,b)}\{\lambda t - \psi_X(\lambda)\}$.

Example Let $X \sim N(\mu, \sigma^2)$ We know $\mathbb{E}[e^{\lambda X}] = \exp(\mu\lambda + \sigma^2\lambda^2/2), \forall \lambda \in \mathbb{R}$. So

$$\sup_{\lambda>0} \big\{\lambda t - \log \mathbb{E}\big[e^{\lambda(X-\mu)}\big]\big\} = \sup_{\lambda>0} \big\{\lambda t - \frac{\sigma^2\lambda^2}{2}\big\} = \frac{t^2}{2\sigma^2}$$

By using Chernoff, $t > 0$,

$$\mathbb{P}(X - \mu \geq t) \leq \exp\Big\{-\frac{t^2}{2\sigma^2}\Big\},\ t \geq 0$$

By symmetry

$$\mathbb{P}(|X - \mu| \geq t) \leq 2\exp\Big\{-\frac{t^2}{2\sigma^2}\Big\},\ \forall t \geq 0$$

This is not a bad bound, since

$$\sup_{t\geq 0} \mathbb{P}(Z \geq t)\exp\big\{t^2/2\big\} = \frac{1}{2}$$

We want bounds of the form

$$\mathbb{P}(|X - \mu| \geq t) \leq C_1 \exp\Big\{-C_2 t^2\Big\},\quad C_1, C_2 > 0$$

1.2.4 Sub-Gaussian Random Variable

Definition 1.1 *A random variable X with finite $\mu = \mathbb{E}[X]$ is said to be sub-gaussian with parameter σ^2, $X \in SG(\sigma^2)$, if*

$$\mathbb{E}\big[e^{\lambda(X-\mu)}\big] \le \exp\big\{\frac{\lambda^2\sigma^2}{2}\big\},\ \forall \lambda \in \mathbb{R}$$

Remark Always center X! If $X \in SG(\sigma^2)$

$$\mathbb{P}(X - \mu \ge t) \le \exp\big\{-\frac{t^2}{2\sigma^2}\big\},\ \ \forall t > 0$$

Since $X \in SG(\sigma^2)$, if $-X \in SG(\sigma^2)$, we get

$$\mathbb{P}(|X - \mu| \ge t) \le 2\exp\big\{-\frac{t^2}{2\sigma^2}\big\},\ \ \forall t > 0$$

Properties. Assume $X \in SG(\sigma^2)$

- p1) $\mathbb{V}[X] \le \sigma^2$
  Proof: By Taylor expansion

$$1 + \lambda\mathbb{E}[X-\mu] + \lambda^2\mathbb{E}[(X-\mu)^2] + o(\lambda^2) \le 1 + \frac{\lambda^2\sigma^2}{2} + o(\lambda^2)$$

  Divide by λ^2, let $\lambda \to 0$ and we get $\mathbb{E}[(X-\mu)^2] \le \sigma^2$.

- p2) If $-\infty < a \le X - \mu \le b < \infty$ *a.s.*, then $X \in SG\big((\frac{b-a}{2})^2\big)$
  Proof: Notice that $\mathbb{V}[X] \le (\frac{b-a}{2})^2$, because $|X - \frac{b+a}{2}| \le \frac{b-a}{2}$. For any λ, let Z_λ be a random variable whose distribution P_{Z_λ} is s.t. $\frac{dP_{Z_\lambda}}{dP_X}(z) = e^{\lambda z}e^{-\psi_x(\lambda)}$. Then $a \le Z_\lambda \le b$ *a.e.* and $\mathbb{V}[Z_\lambda] = \psi''_X(\lambda)$. So $\psi''_X(\lambda) \le (\frac{b-a}{2})^2$. Since $\psi_X(0) = \log 1 = 0$ and $\psi'_X(0) = \mathbb{E}[X] = 0$,

$$\psi_X(\lambda) = \int_0^\lambda \psi'_X(\lambda')d\lambda' = \int_0^\lambda \int_0^{\lambda'} \psi''_X(\lambda'')d\lambda''d\lambda'$$
$$\le \frac{\lambda^2}{2}\frac{(b-a)^2}{4} = \frac{\lambda^2(b-a)^2}{8}$$

References

[Ball97] Ball, Keith, "An elementary introduction to modern convex geometry," *Flavors of geometry* 31 (1997): 1-58.