36-788: Topics in High Dimensional Statistics I **Fall 2015**

Lecture 1: September 1

Lecturer: Alessandro Rinaldo *Scribes: Jisu KIM*

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

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

1.1 Introduction

homepage : http://www.stat.cmu.edu/˜arinaldo/36788

This course covers Concentration inequalities and how they work.

Concentration inequalities were initially studied in Functional Analysis, in particular in the Geometry of Banach spaces. Concentration inequalities also have a lot of applications in Computer Science and Discrete Math. In particular, they are important in the study Randomized Algorithms and Combinatorial Optimization. They are also important in Statistics and Machine Learning.

1.2 Parametric Statistics

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ with d fixed. P_θ takes values on $(\mathbb{R}^n, \mathcal{B}^n)$.

Observe $X_1, \cdots, X_n \overset{iid}{\sim} P_{\theta^n}$ with $\theta^n \in \Theta$. Parametric estimator is as $\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n)$.

1.2.1 Tools

1) WLLN

$$\forall \epsilon > 0, \lim_{n \to \infty} P\left(d(\hat{\theta}_n, \theta^n) > \epsilon\right) = 0$$

2) CLT

$$\underbrace{A_n}_{\text{seq of scaling matrices}} \left(\bar{X}_n - \mu_n\right) \rightsquigarrow N_d(0, I)$$

More generallly,

$$A_n\left(\hat{\theta}_n - \mu_n\right) \rightsquigarrow N_d(0, I).$$

Especially, if $X_1, \cdots, X_n \overset{iid}{\sim} (\mu, \Sigma)$,

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \rightsquigarrow N_d(0, \Sigma).$$

3) Aside: Berry Esseen Bound

If $d = 1$: if $X_1, \cdots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$ and such that $E|X_1 - \mu|^3 < \infty$, then

$$\sup_{x \in \mathbb{R}} \left| \underbrace{P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right)}_{\text{cdf of } \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}} - \underbrace{\Phi(x)}_{\text{cdf of } N(0,1)} \right| \leq \frac{C(\mu, \sigma^2)}{\sqrt{n}}.$$

This is refinement of CLT: it is finite sample result.

If $d > 1$, C will depend on d. The optimal const is still not known. Especially requires $d < n$ to hold. [1]

1.2.2 Issues

1) Asymptotic Results

Asymptotic results may not necessarily hold under finite n. We want Sample Complexity guarantees, which is to fix $\epsilon > 0$, and find smallest $n(\epsilon)$ s.t. $\forall n > n(\epsilon)$, $\mathbb{E}\left[d(\hat{\theta}_n - \theta^n)\right] \leq \epsilon$.

2) They require d fixed.

d does not change with n. Issue 2 in particular is problematic in high dimensional statistical models!

1.3 High dimensional statistics

Definition 1.1 *High dimensional statistical model is a sequence $\{\mathcal{P}_k\}_k$ of parametric models, where $\mathcal{P}_k = \{P_\theta, \theta \in \Theta \subset \mathbb{R}^{d_k}\}$ and d_k increases as k increases.*

We observe iid samples $X_1, \cdots, X_n \sim P_{\theta^k}$, where $P_{\theta^k} \in \mathcal{P}_k$. We take $k = n$ (or $k = k(n)$ increasing in n will be enough).

Example. (X_1, Y_1), ..., $(X_n, Y_n) \sim P_\theta$ with $\theta \in \mathbb{R}^{d_k}$. The model is parametrized as

$$Y_i = < X_i, \theta > + \epsilon_i.$$

1.3.1 Regimes of Complexity

1) $d_n \to \infty$ as $n \to \infty$ but $d_n = o(n)$. This is more classic and we try not to make assumptions.

2) $d_n \gg n$: require structural assumption(sparsity).

Concentration inequalities can handle both cases. They provide finite sample bounds (n remains finite) in which the dimension is often an explicit parameter.

[1] Theorem 14 and Theorem 16 in [WKR2014], p.1202-1205, Theorem 1.1 in [B2003], p.385-387, [P1986], p.571-572

1.3.2 General form of concentration inequalities

Let $X = (X_1, \cdots, X_n) \overset{iid}{\sim} P$ and $X \in \mathscr{X}$. Let $f : \mathscr{X} \to \mathbb{R}$ that is "smooth" (or 1 Lipschitz[2]). Then, the bounds we are after are of the form

$$\forall \epsilon > 0,\ P\left(|f(x) - E[f(x)]| > \epsilon\right) \leq g(\epsilon, n)$$

where $g(\epsilon, n) \to 0$ as $n \to \infty$ "fast", typically $g(\epsilon, n) = C_1 \exp\left\{-n\epsilon^2 C_2\right\}$.

For any 1-Liptschitz f, typically, $f(X_1, \cdots, X_n) = \sum\limits_{i=1}^{n} X_i$

Example. Covariance estimation in L_∞.

$X_1, \cdots, X_n \overset{iid}{\sim} P$ in $\mathbb{R}^d$, satisfying sub-Gaussianity condition[3](a critical regularity condition). Let Σ be $Cov[X_1]$, $\hat{\Sigma}_n = \frac{1}{n} \sum\limits_{i=1}^{n} (X_i - \bar{X}_n)(X_i - \bar{X}_n)^T$ be the empirical covariance matrix with $\bar{X}_n = \frac{1}{n} \sum X_i$. Then $\exists C$ s.t.

$$\forall t > 0,\ \|\hat{\Sigma}_n - \Sigma\|_\infty \leq c\sqrt{\frac{t + \log d}{n}} \text{ with prob } \geq 1 - e^{-t},$$

where $\|A_\infty\| = \max\limits_{i,j} |A_{i,j}|$. Especially, take $t = \log n$,

$$\|\hat{\Sigma}_n - \Sigma\|_\infty = O\left(\max\left\{\sqrt{\frac{\log n}{n}}, \sqrt{\frac{\log d}{n}}\right\}\right) \to 0$$

even if $d \gg n$, with prob$\geq 1 - \frac{1}{n}$.

Example.

- Compressed sensing [4]

- Performance of Lasso [5]

Example. Let v_n be the volume of unit ball in $\mathbb{R}^n$. Then volume of a ball of radius $(1 - \epsilon)$, $\epsilon > 0$ small, is $(1 - \epsilon)^n v_n$. By taking ratio,

$$\frac{(1 - \epsilon)^n v_n}{v_n} \to 0$$

fast, for any $\epsilon > 0$. So as $n \to \infty$, most of mass is on the boundary, for uniform distribution.

Example. If $X = X_1, \cdots, X_n \overset{iid}{\sim} N(0, I)$, then the norm $\|X\|$ will be highly concentrated on $\approx \sqrt{n}$. [6]

Example. Let X be uniformly distributed on $\{0, 1\}^n$(hypercube), and $d_H(x, y) = \frac{1}{n} |\{i :\ x_i \neq y_i\}|$ be the distance onthe hypercube. Then, for any f 1$-$Liptshitz ,

$$P\left(f(X) \geq E[f(X)] + r\right) \leq e^{-\frac{nr^2}{2}}.$$

Take a subset $A \subset \{0, 1\}^n$ with $P(A) \geq \frac{1}{2}$, and enlarge A a little by $A_r = \{x \in \mathscr{X} :\ d(x, A) < r\}$, as in Figure 1.1. Then

$$P(A_r^c) \leq e^{-nr^2 c}.$$

$r \sim \frac{1}{\sqrt{n}}$ is enough for A_r to capture almost all mass. [7]

---

[2] f is1-Lipschitz if $|f(x) - f(y)| \leq d(x, y)$. Refer to 2^{nd} lecture.
[3] X is sub-Gaussian if $\exists \nu$ s.t. $\forall \lambda, \log\left(E\left[e^{\lambda X}\right]\right) \leq \frac{\lambda^2 \nu^2}{2}$. Refer to lecture 3.
[4] Theorem 5.2 in [BDDW20007]
[5] Theorem 1 in [W2009]
[6] Proposition 2.2, Corollary 2.3 in [B2005], p.5-8
[7] Theorem 2.11 in [L2005], p.3, p.31, Corollary 4.4 in [B2005], p.17

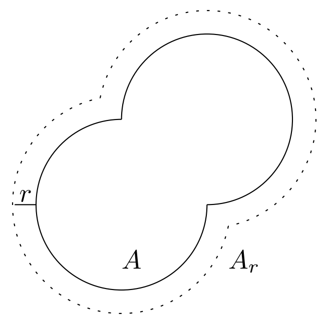

Figure 1.1: A and $A_r = \{x \in \mathscr{X} : d(x, A) < r\}$

Reference

[B2005] Measure Concentration, Lecture Notes for Math 710, by Alexander Barvinok, 2005

[B2003] Bentkus, V. (2003). On the dependence of the BerryEsseen bound on dimension, Journal of Statistical Planning and Inference, 113, 385-402.

[BDDW2007] A simple proof of the restricted isometry property for random matrices, R Baraniuk, M Davenport, R DeVore, M Wakin, Constructive Approximation 28 (3), 253-263

[L2005] The Concentration of Measure Phenomenon, by M. Ledoux, 2005, AMS.

[P1986] Portnoy, S. (1986). On the central limit theorem in R p when $p \rightarrow \infty$, Probability Theory and Related Fields, 73(4), 571-583

[W2009] M. J. Wainwright (2009), Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso). IEEE Transactions on Information Theory, 55:2183–2202.

[WKR2014] Wasserman, L, Kolar, M. and Rinaldo, A. (2014). Berry-Esseen bounds for estimating undirected graphs, Electronic Journal of Statistics, 8(1), 1188-1224