36-710: Advanced Statistical Theory **Fall 2018**

Lecture 1: August 27

Lecturer: Alessandro Rinaldo *Scribes: Shamindra Shrotriya*

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 Recap of Parametric Statistics

A (parametric) statistical model is specified by $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$. Where:

- Θ is the parameter space, usually an open subset of $\mathbb{R}^d$
- $\forall \quad \theta \in \Theta$, P_θ is a probability measure on $(\mathbb{R}^s, \mathcal{B}^s)$
- $P_\theta \ll \mu \quad \forall \theta \in \Theta$ where μ is a $\sigma-$finite measure
- $\dim(\Theta) = d$ is the dimension of the parameter space

Example 1.1.1 (Multivariate Gaussian)**.** $\Theta = \{(\mu, \Sigma) \mid \mu \in \mathbb{R}^s, \Sigma \in S_k^+\}$, where S_k^+ is the cone of positive semidefinite matrices and we have $\mathcal{P} = \{\mathcal{N}(\mu, \Sigma) \mid (\mu, \Sigma) \in \Theta\}$ with $\dim(\Theta) = \frac{k(k+1)}{2} + k$ is the dimension of the parameter space

Example 1.1.2 (Regression)**.** $(X, Y) \in \mathbb{R}^d \times \mathbb{R} \sim P_{(\theta, \sigma^2)}$ where $\theta \in \mathbb{R}^d, \sigma \in \mathbb{R}_+$. Moreover we have $Y = X^T\theta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The additional assumptions are:

- $X \sim P$ on $(\mathbb{R}^d, \mathcal{B}^d)$ and $X \perp\!\!\!\perp \epsilon$ and $\mathrm{Var}[X] = I_d$
- $\Theta = \{(\mu, \sigma^2) \mid \theta \in \mathbb{R}^d, \sigma > 0\}$
- $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$
- $\dim(\Theta) = d + 1$ since we need to additionally specify the σ parameter

1.2 Key Tools in Parametric Low-Dimensional Inference

Consider $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \overset{\text{i.i.d.}}{\sim} P_{\theta_0}$ where $\theta_0 \in \Theta$ and X has distribution $P_{\theta_0}^n$ (product measure for $X_1, X_2 \ldots X_n$).

You have learned in `STAT-36705` that if $\mathcal{P}$ is sufficiently well behaved, inference in the unknown parameter can be done optimally [See also the lectures in `STAT-36-752` on April 24-26 2018[1]]. We then have $\tilde{\theta} \xrightarrow{p} \theta_0$ and $\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}_d(0, I_d^{-1}(\theta_0))$.

We need the following tools

- Weak Law of Large Numbers (WLLN)
- Central Limit Theorem (CLT)
- Continuous Mapping Theorem (CMT)

Key point: These tools require d, θ_0 to be **fixed** which may not meet more complicated real-life modeling requirements.

1.2.1 Shortcomings of Low Dimensional Parametric Models

In such described low dimensional parametric models we require:

- **Asymptotic results** i.e. $n \to \infty$
- Require d to be **fixed** which limits the use of such models where we gather data of higher complexity

1.3 High Dimensional Parametric Statistical Models

Such models are defined as a sequence $\{\mathcal{P}_n\}_n^\infty$ of parametric statistical models indexed by sample size n. More specifically we have:

1. $\mathcal{P}_n$ is a parametric model with parameter space $\Theta_n \subset \mathbb{R}^{d_n}$, sample space $(\mathbb{R}^{S_n}, \mathcal{B}^{S_n})$ where both d_n and S_n are **non-decreasing** in n.
2. Observe that $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \overset{\text{i.i.d.}}{\sim} P_{\theta_0}$ where $\theta_{0,n} \in \Theta_n$

Remark. 1. We have 2 high-dimensional regimes $d \leq n$ and $d = o(n)$ see [portnoy1988]

2. $d_n \gg n$ where we need structural assumptions

Example 1.3.1. In regression we have $(X_i, Y_i)_{i=1}^n$ with

$$Y_i = X_i^T \theta_0 + \epsilon_i, \theta_0 \in \mathbb{R}^d$$

However if $d \gg n$ then we need the additional structural assumption of **sparsity**. That is

$$\|\theta\|_0 = \#\{i \mid \theta_{0,i} \neq 0\} \ll n$$

Remark. A high-dimensional statistical model is not probablistically consistent.

[1] See the lecture scribed lectures here `http://www.stat.cmu.edu/~arinaldo/Teaching/36752/S18/schedule.html`

1.4 Recap: High Dimensional Statistics from Final Lectures in STAT-36752

1.4.1 Different Topology in High Dimensions

When $d \to \infty$, norms in $\mathbb{R}^d$ have widely different geometry/ topology compared to low dimensions.

If d is fixed, all L_p norms with $p \geq 1$ are topologically equivalent. In other words, for fixed d and any 2 norms, say $\|\cdot\|_p$ and $\|\cdot\|_q$ we have that $c \leq \frac{\|x\|_p}{\|x\|_q} \leq C$, Such norms include $\|x\|_p = (\sum_i^n |x_i|^p)^{\frac{1}{p}}$ and $\|x\|_\infty = \max_i |x_i|$.

Recall that the volume of the unit Euclidean Ball in $\mathbb{R}^d$ is

$$\begin{aligned} V_d &= \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} \\ &\sim \left(\frac{2\pi e}{d}\right)^{\frac{d}{2}} (d\pi)^{-\frac{1}{2}} \\ &\to 0 \quad \text{very fast as } d \to \infty \end{aligned}$$

The volume of the unit L_∞ ball in $\mathbb{R}^d$ is

$$vol([0,1]^d) = 1$$

You need approximately atleast $\left(\frac{d}{2\pi\epsilon}\right)^{-\frac{d}{2}} \sqrt{(d\pi)}$ unit euclidean balls to cover $[0,1]^d$ as $d \to \infty$. The value of the multiplicative constants of the 2 balls depends on the dimensions of the balls.

1.4.2 Concentration of Measure Phenomenon

1.4.2.1 Uniform measure on the Euclidean unit ball in $\mathbb{R}^d$

$\forall \epsilon \in (0,1)$, arbitrarily small and fixed $\mathbb{P}(X \in B_d(0, 1-\epsilon))$. Then $X \sim$ Uniform on $B_d(0,1)$, where $B_d(r) = \{x \in \mathbb{R}^d \mid \|x\| \leq r\}$. We find that most of the mass for the uniform ball on $B_d(0,1)$ for high dimensions is near the boundary of the ball.

If d is large and $X \sim \mathcal{N}_d(0, I_d)$ with high probability we have $\|X\| \sim \sqrt{d}$

Example 1.4.1 (Covariance Estimation)**.** Here we consider $X_1, X_2, \ldots X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$ in $\mathbb{R}^d$.

We can then estimate Σ with $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$. We note that $\mathbb{E}(\hat{\Sigma}) = \Sigma$ (unbiased).

For a $d \times d$ matrix A let $\|A\|_\infty = \max_{i,j} |A_{i,j}|$. Let's look at $\|\hat{\Sigma} - \Sigma\|_\infty$. Then for d fixed we have for any i, j that $\hat{\Sigma}_{i,j} = \frac{1}{n}{}^n{}_{i=1} Z_k^{(i,j)}$ where $Z_k^{(i,j)} = \frac{1}{n} = X_{k,i}X_{k,j} \quad \forall i,j \in [d]$. Then we have by WLLN that $\hat{\Sigma}_{i,j} \xrightarrow{p} \Sigma_{i,j}$. Furthermore:

$$\begin{aligned} \|\hat{\Sigma} - \Sigma\|_\infty &\leq \sum_{i,j} |\underbrace{\hat{\Sigma}_{i,j} - \Sigma_{i,j}}_{o_p(1)}| \\ &= \frac{d(d+1)}{2} o_p(1) \\ &= o_p(1) \quad \text{since } d \text{ is fixed} \end{aligned}$$

Under an extra moment assumption $\|\hat{\Sigma} - \Sigma\|_\infty = O\left(\frac{1}{\sqrt{n}}\right)$ using CLT. If $d \nearrow$ with n, under some tail assumptions we have have

$$\|\hat{\Sigma} - \Sigma\|_\infty = O\left(\sqrt{\frac{\log d_n}{n}}\right)$$

with probability 1.

References

[portnoy1988] S. Portnoy, "Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity," *The Annals of Statistics*, 1988, pp. 356–366.