Math Stats Lecture Notes

1. Example: Sample mean. $X_1, \dots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$
Suppose we have some prior estimation $\hat{\mu}_0$ for μ.

Assume $\hat{\mu} = 0.2\hat{\mu}_0 + 0.8\bar{X}$, then
$\text{Bias}(\hat{\mu}) = 0.2(\hat{\mu}_0 - \mu)$
$\text{Var}(\hat{\mu}) = 0.64\frac{\sigma^2}{n}$
$R(\mu, \hat{\mu}) = 0.04(\hat{\mu}_0 - \mu)^2 + 0.64\frac{\sigma^2}{n}$.
Conclusion: $\hat{\mu}$ is better than $\bar{X}$ if $\hat{\mu}_0$ is close to μ.

2. Ridge Regression
Obs: (x_i, y_i), $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ with model
$y_i = x_i^T\beta^* + \varepsilon$

Ridge regression solves ignore the $\frac{1}{n}$ term in lecture note

$$\min_\beta \underbrace{\sum_{i=1}^n (y_i - x_i^T\beta)^2}_{} + \lambda\|\beta\|_2^2 \quad (1)$$

$\Rightarrow \hat{\beta}_\lambda = (X^TX + \lambda I_p)^{-1}X^Ty \qquad (X = (x_1, \dots x_n)^T \in \mathbb{R}^{n\times p})$
$= (X^TX + \lambda I_p)^{-1}X^T(X\beta^* + \varepsilon)$
$\text{MSE} = \mathbb{E}[\|\hat{\beta}_\lambda - \beta^*\|^2] = \mathbb{E}[\|\hat{\beta}_\lambda - \mathbb{E}\hat{\beta}_\lambda\|_2^2] + \|\mathbb{E}\hat{\beta}_\lambda - \beta^*\|_2^2$
$\text{Bias} = O(\lambda)$, $\sigma^2\text{Tr}((X^TX)^{-1}) - \text{Var} = O(\lambda)$
$\lambda\uparrow \Rightarrow \text{Var}\downarrow \ \text{Bias}\uparrow$

3. General Bias-Variance Tradeoff

Consider model $Y = f(x) + \varepsilon$, $\mathbb{E}\varepsilon = 0$, $\text{Var}(\varepsilon) = \sigma^2$ $X \in \mathbb{R}^p$
kNN estimator: $\hat{f}_k(x_0) = \frac{1}{k}\sum_{k=1}^{k} y_{x(k)}$
where $x_{(1)}, \dots x_{(k)}$ are k nearest points to x_0

$$\mathbb{E}[(Y - \hat{f}_k(x_0))^2 \mid X = x_0] = \sigma^2 + \underbrace{\Big(f(x_0) - \frac{1}{k}\sum_{k=1}^{k} f(x_{(k)})\Big)^2}_{\text{Bias}} + \underbrace{\frac{\sigma^2}{k}}_{\text{Variance}}$$

4. Overparametrization

4. Overparametrization $\hat{f}(x) = x^T\hat{\beta}$

$\text{Var}(\hat{f}(x_0)\mid x_0) = \sigma^2 x_0^T(X^TX)^{-1}x_0$
$\mathbb{E}[\hat{f}(x_0)\mid x_0] = x_0^T\beta \quad \hat{\beta} = (X^TX)^{-1}X^Ty \perp x_0$
$\text{Var}(\hat{f}(x_0)) = \mathbb{E}[\sigma^2 x_0^T(X^TX)^{-1}x_0] + \text{Var}(x_0^T\beta)$
$= \sigma^2\mathbb{E}[\text{tr}((X^TX)^{-1}x_0x_0^T)] + \beta^T\beta$
$= \sigma^2\text{tr}((X^TX)^{-1}\mathbb{E}[x_0x_0^T]) + \beta^T\beta$
$= n\sigma^2\text{tr}((X^TX)^{-1}) + \|\beta\|^2$

$\text{Var}(x^T\hat{\beta}) = \sigma^2 x^T(X^TX)^{-1}x$
If x_0 has zero mean and identity covariance matrix.
$\text{Var}(\hat{f}(x_0)) = \sigma^2\mathbb{E}[x_0^T(X^TX)^{-1}x_0] = \sigma^2\text{tr}(\underbrace{(X^TX)^{-1}}_{\in\mathbb{R}^{p\times p}}) + \|\beta\|^2$
The variance generally increases as p increases
Typically, $X^TX = \sum_{i=1}^n x_ix_i^T \Rightarrow \mathbb{E}[X^TX] = n\Sigma_X \quad \Sigma_X \in \mathbb{R}^{p\times p}$
If the smallest eigenvalue is bounded away from 0.
then the variance is $O(p/n)$

Double Descent and min-norm solution.
When $p > n$: $\min_\beta \|\beta\|_2^2 \quad (2)$ (ridgeless estimator)
subject to $X\beta = y$
Suppose $\text{rank}(XX^T) = n$, then it has a unique solution
$\hat{\beta}_{0+} = \lim_{\lambda\to 0+}\hat{\beta}_\lambda = X^T(XX^T)^{-1}$

5. The implicit regularization of gradient descent

GD: $\beta^{(t+1)} = \beta^{(t)} - \eta\nabla L(\beta)$,
Starting at $\beta^{(0)} = 0$, solving $L(\beta) = \|Y - X\beta\|^2$
Thm. Suppose $p > n$, and $\text{rank}(XX^T) = n$. Then,
if $\eta < [2\lambda_{max}(X^TX)^{-1}]$, then $\beta^{(t)}$ converges, and
$\lim_{t\to\infty}\beta^{(t)} = \hat{\beta}_{0+}$

Rmk. GD prefers the solution. Because
$\nabla L(\beta) = 2X^T(Y - X\beta) \in C(X^T) \Rightarrow \beta^{(t)} \in C(X^T)$
$\Rightarrow \lim_{n\to\infty}\beta^{(t)} \in C(X^T)$
and $\hat{\beta}_{0+}$ is the unique solution in $C(X^T)$

Thm (GD for overparameterized logistic regression)
Consider logistic loss and linear separable data.
From any initializer $\beta^{(0)} \in \mathbb{R}^p$, the gradient iterate
$\beta^{(k+1)} = \beta^{(k)} - \eta\nabla L_n(\beta^{(k)})$ satisfies $\beta^{(k)} = \hat{\beta}\log k + \Delta^{(k)}$
where $\|\Delta^{(k)}\|_2 = O(\log\log k)$ and
$\hat{\beta} = \arg\min\|\beta\|_2^2$ subject to $x_i^T\beta \geq 1$. for $i = 1, \dots n$.

Remark: The DIRECTION of $\beta^{(k)}$ converge to $\hat{\beta}$.
and in logistic model, only direction matters.

Classifying Statistical Models

Parametric Family: $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$

e.g. Family of densities: If $P_\theta \ll \nu$, $\forall \theta \in \Theta$, then we only need to consider $\{\frac{dP_\theta}{d\nu} : \theta \in \Theta\}$, particularly pdf / pmf

- Example: $P \in \mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$. $\theta = (\mu, \sigma^2)^T$
- Example (Linear model) $\mathcal{P} = \{N(\beta^T X, \sigma^2) : \beta \in \mathbb{R}^p, \sigma^2 > 0\}$ $X \in \mathbb{R}^p$

Nonparametric Family: Θ is a infinite dimensional space

Rk. Statistical Methods often use a finite number of estimators to approximate the unknown quantity

- Example (Density estimation): Each distribution is abs cts and the set of pdf is $\{f(x) : |f'(x)| < L\}$ for some $L > 0$

Semiparametric Family: Finite parameters of interest + Infinite Nuisance Parameters

- Example: Linear model with unknown dist of ε

Identifiability

Def. $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is identifiable iff $P_{\theta_1} \neq P_{\theta_2}$ for $\theta_1 \neq \theta_2$.

- Example (Factor Model) Consider $Y \in \mathbb{R}^p$ has the structure
  $$Y = Bf + u, \text{ where } f \sim N(0, I_r),\ u \sim N(0, \Sigma).$$
  Σ: known, $B \in \mathbb{R}^{p \times r}$: unknown.
  Equivalently $Y \sim (0, BB^T + \Sigma)$.
  Since $BR(BR)^T = BB^T$ for any orthogonal matrix $R \in \mathbb{R}^{r \times r}$
  this model is not identifiable.
- Example: Multinomial regression: In classification problem:
  $$p_k = \frac{\exp(\beta_k^T x)}{\sum_{j=1}^{K-1} \exp(\beta_j^T x)} \quad \text{where } k = 1, 2, \dots, K-1$$
  $(\beta_1 \cdots \beta_K)^T$ and $(\beta_1 + c, \dots \beta_K + c)$ stipulate the same model
  - To avoid this: $$p_k = \frac{\exp(\beta_k^T x)}{\sum_{j=1}^{k} \exp(\beta_j^T x) + 1}$$

Statistics

Def. Let $X \in \mathbb{R}^{n \times p}$ be the data. For any msb function T, we call $T(X)$ a statistic
$\sigma(T(X)) \subset \sigma(X)$: $T(X)$ summarizes information from X.

- Example Sample mean $\bar{X}$ and Sample variance S^2
- Order Statistics $X_{(1)} \cdots X_{(n)}$. pdf for $X_{(i)}$ is:
  $$g_i(x) = \frac{n!}{(i-1)!\,(n-i)!} [F(x)]^{i-1} [1 - F(x)]^{n-i} f(x)$$

Exponential Family:

Def. Dist with pdf $f_\theta(\omega) = \exp\{[\eta(\theta)]^T T(\omega) - \xi(\theta)\} h(\omega)$, $\forall \omega \in \Omega$

Canonical Form: $f_\eta(\omega) = \exp\{\eta^T T(\omega) - \zeta(\eta)\} h(\omega)$, $\omega \in \Omega$

e.g.: Binomial, Normal, Dists subsumed in exponential family

Properties. (1) Marginal / Conditional dist remains exponential

(2) For Borel function f: $\int |f| dP_{\eta_0} < \infty$

$\int f(\omega) \exp(\eta^T T(\omega)) h(\omega) d\nu(\omega) \in C^\infty$ — Probability Measure induced by the random Variable.

$\text{Cov. } m(\eta) \in C^\infty$ on Ω

Basic Terminologies in Statistical Decision Theory

- Decision Rule: $T : (\mathbb{R}^k, \mathcal{B}_k) \to (\underbrace{\mathbb{A}}_{\text{Action Space}}, \mathcal{F}_{\mathbb{A}})$
- Loss Function: $L(P, T(x))$
- Risk Function: $R_T(P) = \mathbb{E}[L(P, T(X))]$. For parametric Family, it can be written as $R(\theta)$
  e.g. $MSE_{\hat\theta}(\theta) = \mathbb{E}_\theta[(\hat\theta - \theta)^2] = \int (\hat\theta(x) - \theta)^2 dP_\theta(x)$

Optimal Rule

Def. T_* is an optimal rule if
$R_{T_*}(P) \le R_T(P)$, $\forall P \in \mathcal{P}$, $\forall T \in \{\text{Class of allowable decisions}\}$
Rate optimal if $R_{T_*}(P) \le C R_T(P)$ for some C

Randomized decision rules.

For every x, $\delta(x, \cdot)$ is a probability measure on $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$

Often, $\delta = \sum_{j=1}^{M} 1(Z = j) T_j$ where Z is a discrete random variable

- Risk of randomized rule: $R(\theta, \delta) = \mathbb{E}_{X \sim P_\theta}[\mathbb{E}_{T \sim \delta(X)}[L(\theta, A)]]$

e.g. Stochastic Gradient Descent

Admissibility Let $\mathcal{G}$ be a class of decision rules. $T \in \mathcal{G}$ is $\mathcal{G}$-admissible iff $\forall S \in \mathcal{G}$, $R_S(P) > R_T(P)$, $\forall P \in \mathcal{P}$

Main Approaches for optimality:

- Bayes rule: Given a prior dist Π over $\mathcal{P}$, T_* minimizes
  $$r_T(\Pi) = \int R_T(P)\, d\Pi(P)$$
- Minimax rule: $T_* = \arg\min_T (\sup_P R_T(P))$

Estimation.

For iid $X_1, \cdots X_n$, WTK $\mathbb{E}[X_i]$. Use squared loss $L(P, a) = (\theta - a)^2$, then
$$R_{\bar{X}}(P) = \frac{\sigma^2}{n} \quad \text{where } \sigma^2 = Var(X_i)$$

Bias: $b_T(P) = \mathbb{E}_P[T(X)] - \theta$

Hypothesis Test

$$H_0 : P \in \mathcal{P}_0 \quad vs. \quad H_1 : P \in \mathcal{P}_1$$

Let $T(X) \in \{0, 1\}$, $T(X)$ must have the form $T_C(X) = \{X : T(X) = 1\}$, where C is called the rejection region. Consider the loss:
$$L(P, j) = 0 \text{ if } P \in \mathcal{P}_j, \text{ and } 1 \text{ otherwise}$$
Then
$$R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C), & P \in \mathcal{P}_0 \\ P(T(X) = 0) = P(X \notin C), & P \in \mathcal{P}_1 \end{cases}$$

Sufficient Statistics.

Def. $T(X)$ is sufficient if the conditional dist of $X \mid T(X) = t$ does not depend on P or θ for any t

Factorization Thm: Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by the Lebesgue measure ν. Then $T(X)$ is sufficient iff $\exists$ Borel functions h, g_θ s.t. $\frac{dP_\theta}{d\nu}(x) = g_\theta(T(x)) h(x)$.

Example: exponential family $f_\theta(\omega) = \underbrace{\exp\{[\eta(\theta)]^T T(\omega) - \xi(\theta)\}}_{g_\theta(T(x))} \underbrace{h(\omega)}_{h(x)}$

Connection to information theory:

$T(X)$ is sufficient $\iff I(X; \theta) = I(T(X); \theta)$

where θ is treated as a random variable

***Data Processing Inequality.**

$X \to Y \to Z$ Markov Chain $X \perp\!\!\!\perp Z \mid Y$, then
$$I(X, Z) \le I(X; Y).$$

1. Complete Statistic

32

Def. (Complete Statistics) $T(X)$ is complete for $P \in \mathcal{P}$ iff $\forall$ Borel f, $\mathbb{E}[f(T)] = 0 \Rightarrow f(T) = 0$ a.s. $\mathcal{P}$

Rmk: Requires the statistic has no redundant information.

32

Proposition. Let $\mathcal{P} = \{P_\eta, \eta \in \Xi\}$ be an exponential family of full rank (contains an open set) with pdf

$$f_\eta(x) = \exp\{\eta^T T(x) - \xi(\eta)\} h(x)$$

Then $T(X)$ is sufficient and complete for $\eta \in \Xi$

32

2. UMVUE

32

Def. An unbiased estimator $T(X)$ of ν is called the UMVUE iff $\mathrm{Var}(T(X)) \le \mathrm{Var}(U(X))$ for any $P \in \mathcal{P}$ and any other unbiased estimator $U(X)$ of ν

Rmk UMVUE does not always exist

32

Thm Suppose there exists a sufficient and complete statistic $T(X)$ for P. If ν is estimable then there exists unique UMVUE, which is of form $h(T)$ where h is a Borel function.

exists unbiased estimator

33

3. Construct UMVUE

33

Method 1. Find sufficient and complete statistic $T(X)$, then find $h(T)$ such that $\mathbb{E}[h(T)] = \nu$, $\forall P \in \mathcal{P}$

33

Method 2. Find sufficient, complete $T(X)$ and unbiased $U(X)$, then $\mathbb{E}[U|T]$ is an UMVUE.

33

Method 3. Find UMVUE without knowing complete statistics

33

Thm 2 Let $\mathcal{U} = \{U : \mathbb{E}[U(X)] = 0, \mathrm{Var}(U(X)) < \infty, \forall P \in \mathcal{P}\}$

T is unbiased for ν with $\mathbb{E}[T(X)] < \infty$

(1). $T(X)$ is UMVUE iff

$$\mathbb{E}[T(X)U(X)] = 0, \ \forall U \in \mathcal{U}, \ \forall P \in \mathcal{P}$$

(2) If $\tilde{T}$ is sufficient for P, let

$$\tilde{\mathcal{U}} = \mathcal{U} \cap \{g(\tilde{T}) : g \text{ Borel}\}$$

Then $T = h(\tilde{T})$ is UMVUE iff

$$\mathbb{E}[T(X)U(X)] = 0, \ \forall U \in \tilde{\mathcal{U}}, \ \forall P \in \mathcal{P}$$

38

2. Subgaussian variables and Hoeffding's Inequality

38

Thm 1. Equivalent defs of subgaussian random variables

(1) $P(|X|\ge t) \le 2\exp(-t^2/K_1^2)$, $\forall t \ge 0$ for some $K_1>0$

(2) $\|X\|_p \le K_2\sqrt{p}$ for some $K_2>0$ and all $p\in\mathbb{N}^*$

(3) $\mathbb{E}\exp(\lambda^2X^2) \le \exp(K_3^2\lambda^2)$ for some $K_3>0$ and all λ w. $|\lambda| \le \frac{1}{K_3}$

(4) $\mathbb{E}\exp(X^2/K_4^2) = m_{X^2}(\frac{1}{K_4^2}) \le 2$ for some $K_4>0$

(5) If $\mathbb{E}X=0$, $\mathbb{E}\exp(\lambda X) \le \exp(K_5^2\lambda^2)$, $\forall \lambda\in\mathbb{R}$

Remark. The smallest K_4 is called the subgaussian norm of X, denoted by $\|X\|_{\psi_2}$

39

Examples. · If $X\sim N(0,\sigma^2)$, then $\|X\|_{\psi_2} \le C\sigma$

· If $|X|\le M$ a.s. then $\|X\|_{\psi_2}\le CM$

· $X\sim$ Bernoulli (0.5) $\|X\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}$

41

Maximum of sub-gaussians: Let $X_1, X_2, \cdots$ be a sequence of sub-gaussian random variables, not necessarily indpt. Then
$\mathbb{E}\max_i \frac{|X_i|}{\sqrt{1+\log i}} \le CK$
where $K=\max_i\|X_i\|_{\psi_2}$. For any $N\ge 2$, we have
$\mathbb{E}\max_{i\le N}|X_i| \le CK\sqrt{\log N}$

41

Rmk. The bound is sharp: Let $X_1, X_2, \dots$ iid $N(0,1)$, we have $\mathbb{E}\max_{i\le N} X_i \ge c\sqrt{\log N}$

41

Thm 2. Hoeffding's Inequality Let $X_1, X_2, \dots, X_n$ be independent, mean zero and subgaussian r.v.s. Let $a\in\mathbb{R}^n$, $K:=\max_{1\le i\le n}\|X_i\|_{\psi_2}$. Then for any $t>0$

$P(|\sum_{i=1}^n a_iX_i|>t) \le 2\exp(-\frac{ct^2}{K^2\|a\|_2^2})$

holds for certain absolute constant $c>0$

41

Rmk. More flexible than CLT: Mean → weighted sum

42

Khintchine's inequality Let $X_1, X_2, \dots, X_N$ be independent sub-gaussian random variables with zero-means and unit variance, let $a\in\mathbb{R}^N$. Then for any $p\in[2,+\infty)$:

$\|a\|_2 \le \|\sum_{i=1}^N a_iX_i\|_p \le CK\sqrt{p}\,\|a\|_2$

where $K=\max_i\|X_i\|_{\psi_2}$ and C is an absolute constant

42

3. Subexponential Variables and Bernstein's Inequality

42

Thm 3. Equivalent defs of Subexponential random variables)

(1) $P(|X|>t) \le 2e^{-t/K_1}$, $\forall t\ge 0$ relaxed subgaussian.

(2) $\|X\|_p \le K_2 p$ for all $p\ge 1$

(3) $\mathbb{E}[\exp(\lambda|X|)] \le \exp(K_3\lambda)$, $\forall\lambda\in[0, 1/K_3]$

(4) $\mathbb{E}[\exp(|X|/K_4)] \le 2$

(5) If $\mathbb{E}X=0$, $\mathbb{E}\exp(\lambda X) \le \exp(K_5^2\lambda^2)$, $\forall\lambda\in[-1/K_5, 1/K_5]$

$K_i \le CK_j$, $\forall i\ne j$ for some C independent of the random variable X.

42

Remark

· X is called sub-exponential if any of the condition holds

· The smallest K_2 is called the sub-exponential norm. $\|X\|_{\psi_1}$

· subgaussian => subexponential, $\|X\|_{\psi_1} \le \|X\|_{\psi_2}$

43

Centering For random variables with non-zero mean, conditions (1) and (2) in thm 1 and thm 3 are still equivalent. Moreover, we have bound

$\|X-\mathbb{E}X\|_{\psi_i} \le C\|X\|_{\psi_i}$

43

Cor. Suppose X, Y are subgaussian. Then XY are ~~subgaussian~~ subexponential, with

$\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$

43

Thm 4. Bernstein's Inequality

Let $X_1, \cdots X_n$ be independent r.v. such that $\mathbb{E}X_i=0$ and each X_i is subexponential. Let $a\in\mathbb{R}^n$, $K=\max_{1\le i\le n}\|X_i\|_{\psi_1}$. Then

subgaussian subexponential

$P(|\sum_{i=1}^n a_iX_i|\ge t) \le 2\exp(-c\min(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}))$

for some constant $c>0$ decays no faster than e^{-t}

43

Remark. Subexponential r.v. has "heavier tails" compared with subgaussian r.v.s.

44

Thm 5. Bernstein's Inequality for bounded random variables

Let $X_1 \dots X_n$ be indpt r.v.s, $\mathbb{E}X_i=0$, $\max_i|X_i|\le M$ a.s.

$P(|\sum_{i=1}^n X_i|\ge t) \le 2\exp(-\frac{t^2/2}{\sigma^2+Mt/3})$

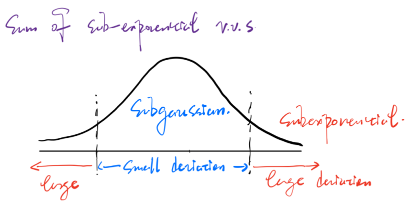
where $\sigma^2=\sum_{i=1}^n\mathbb{E}X_i^2$ is total variance

Variant of Bernstein's Inequality:

$P(|\sum_{i=1}^n X_i| \ge \frac{C(\sigma+M)}{\delta}) \le 2\exp(-\min(\frac{1}{\delta}, \frac{1}{\delta^2}))$, $\forall\delta>0$

44

Sum of sub-exponential r.v.s



1. Fisher Information

34

34

Def. For $X \sim \{P_\theta : \theta \in \Theta\}$:

$$I_X(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^T\right]$$

Rmk. $I(\theta) \succeq 0$. $\frac{\partial}{\partial\theta}\log f_\theta(x)$ is called the score function

34

Proposition

1. If $X \perp\!\!\!\perp Y$, then $I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta)$

2. Suppose f_θ is twice differentiable at θ, under some "regularity condition", we have

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f_\theta(X)\right].$$

2. Cramér-Rao Lower Bound

34

34

Let $\nu = g(\theta)$, $g: \mathbb{R}^d \to \mathbb{R}^k$ differentiable. Suppose X is drawn from $\{P_\theta : \theta \in \Theta\}$, and $T(X)$ is an unbiased estimator of ν.

Thm. Suppose $I(\theta)$ is positive definite, and for any $\theta \in \Theta$,

$$\frac{\partial}{\partial\theta}\int h(x) f_\theta(x)\,d\nu = \int h(x)\frac{\partial}{\partial\theta} f_\theta(x)\,d\nu$$

holds for $h \equiv 1$ or $h(x) = T(x)$. Then,

$$\mathrm{Var}(T(X)) \succeq \left[\frac{\partial}{\partial\theta}g(\theta)\right]^T [I(\theta)]^{-1}\frac{\partial}{\partial\theta}g(\theta)$$

35

Rmk (Reparameterization). If $\theta = \psi(\eta)$, $\psi \in C^1$ is bijective,

then $I_X(\eta) = \frac{\partial}{\partial\eta}\psi(\eta) I_X(\psi(\eta))\frac{\partial}{\partial\eta}\psi(\eta)^T$

is different but

$$\frac{\partial}{\partial\eta}g(\theta) = \frac{\partial}{\partial\theta}g(\theta)\frac{\partial}{\partial\eta}\psi(\eta).$$

Therefore, the C-R lower bound is invariant under reparameterization

35

Rmk. Asymptotic optimality of MLE.

$\mathrm{Var}(\mathrm{MLE}) \to (nI_X(\theta))^{-1}$ matches the C-R lower bound

Rmk. C-R lower bound may not be attained

3. Interpretations of Fisher information.

35

35

3.1. Geometric view.: Larger $I(\theta) \Rightarrow$ Larger local convexity

35

3.2. Information theory. under certain "regularity conditions":

$$D(P_{\theta_0} \| P_{\theta_0+\xi}) = \frac{1}{2}\xi^T I(\theta_0)\xi + o(\|\xi\|^2)$$

"How hard to distinguish two dist in a parametric family under the KL divergence"

5. Examples

36

36

5.1. Exponential Family.

Let $\Theta \subset \mathbb{R}^k$ be an open set. Let $\{f_\theta : \theta \in \Theta\}$ be

$$f_\theta(x) = \exp\{\eta(\theta)^T T(x) - \xi(\theta)\}c(x)$$

36

Proposition

1. The regularity conditions for C-R LB is satisfied for any msb h with $\mathbb{E}[h(X)] < \infty$
2. Consider natural parameter η, $\mathrm{Var}(T(X)) = I(\eta)$
3. $\nu = \mathbb{E}[T(X)]$, then $\mathrm{Var}(T) = (I(\nu))^{-1}$

36

5.2 Linear Models. $X_i = \beta^T Z_i + \varepsilon_i$ $\quad Z \in \mathbb{R}^{n \times p}$ $\quad \varepsilon \sim N(0, \sigma^2 I_n)$

Thm.2 Suppose Z is full column rank

(i) $\ell^T\hat{\beta}$ is UMVUE of $\ell^T\beta$, $\forall \ell \in \mathbb{R}^p$

(ii) $\hat{\sigma}^2 = \frac{1}{n-p}\|X - Z\hat{\beta}\|^2$ is the UMVUE of σ^2

37

Fisher information $\varepsilon \overset{iid}{\sim} N(0, \sigma^2 I_n)$ and the unknown parameters are $\theta = (\beta, \sigma^2)$

$$I(\theta) = \frac{1}{\sigma^2}\begin{pmatrix} Z^T Z & 0 \\ 0 & \frac{n}{2\sigma^2}\end{pmatrix}$$

If Z is not full rank, $I(\theta)$ may not be invertible.

Two distributions P_θ and $P_{\theta'}$ are not distinguishable

Prop. LSE estimate $\hat{\beta}$ attains C-R lower bound.

19

Borel-Cantelli lemma. Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\{A_n\}_{n\geq 1}$ be a sequence of events $A_n \in \mathcal{F}$

1. $\sum_{i=1}^{\infty} P(A_n) < \infty \Rightarrow P(\limsup A_n) = 0$
2. If $\{A_n\}_n$ are independent, then
   $\sum_{i=1}^{\infty} P(A_n) = \infty \Rightarrow P(\limsup A_n) = 1$

$\limsup A_n := \bigcap_{m\geq 1} \bigcup_{n \geq m} A_n$

Op and op Notations.

21

21

Def. Let $\{a_n\}$ be a sequence of real numbers, and $\{b_n\}$ be a sequence of positive numbers. Denote

$$\begin{cases} a_n = O(b_n) & \text{if there exists } C>0 \text{ s.t. } |a_n| \leq Cb_n,\ \forall n \\ a_n = o(b_n) & \text{if } \frac{a_n}{b_n} \to 0 \text{ as } n \to \infty \end{cases}$$

21

Def. Let $X_1, X_2, \dots$ be random vectors and $Y_1, Y_2, \dots$ be random variables defined on a common prob space

(i) $X_n = O_p(Y_n)$ iff. $\forall \varepsilon > 0$, $\exists C > 0$ s.t.

$$\sup_n P(\|X_n\| > C|Y_n|) < \varepsilon$$

(ii) $X_n = o_p(Y_n)$ iff $\|X_n\|/|Y_n| \xrightarrow{P} 0$ as $n \to \infty$

Rmk. $\{X_n\}$ is said to be bounded in probability if $X_n = O_p(1)$

22

Proposition $X_1, \dots X_n$ be random vectors. $Y_1, \dots Y_n$ be random variables defined on a common probability space

(1) If $X_n = o_p(Y_n)$, then $X_n = O_p(Y_n)$

(2) If $X_n = o_p(1)$ and $Y_n = O_p(1)$, then $X_n Y_n = o_p(1)$

3. Estimation Consistency.

22

22

Suppose we have data $X^{(n)}$ for each n drawn from a distribution $P \in \mathcal{P}$, and an estimator $T_n(X) \in \mathbb{R}^p$

Def. Consistency Suppose $T_n(X)$ is an estimator of unknown parameter $\nu \in \mathbb{R}^p$. Let $(a_n)_{n\geq 1}$ be a sequence of positive numbers with $a_n \to \infty$

(1) $T_n(X)$ is consistent for ν iff $T_n(X) \xrightarrow{P} \nu$, $\forall P \in \mathcal{P}$

(2) $T_n(X)$ is a_n-consistent for ν iff $a_n[T_n(X) - \nu] = O_p(1)$ holds for any $P \in \mathcal{P}$

(3) $T_n(X)$ is strongly consistent iff $T_n(X) \xrightarrow{a.s.} \nu$, $\forall P \in \mathcal{P}$

4. Law of Large Numbers

23

23

Thm. (LLN. Simple version) Suppose $X_1, X_2, \dots$ have finite second moment with the same mean $\mu = \mathbb{E}X_i$ and bounded variance $\sigma_i^2 = \mathrm{Var}(X_i) \leq C$ for all i. Further assume that any pair of two random variables are uncorrelated. Then

$$\lim_{n\to\infty} \mathbb{E}[(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu)^2] = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu$$

23

Chebyshev's inequality. If Y is a random variable satisfying $\mathbb{E}[Y^2] < \infty$ and $a > 0$, then

$$P(|Y - \mathbb{E}Y| \geq a) \leq \frac{\mathrm{Var}(Y)}{a^2}$$

Rmk. Finite higher moment implies finite lower moment.

$$(\mathbb{E}\|X\|_p^p)^{1/p} \leq (\mathbb{E}\|X\|_q^q)^{1/q}$$

24

Thm $X_1, X_2 \dots$ be i.i.d. random variables, and $a_n = \mathbb{E}[X_1 1\{|X_1| \leq n\}]$. Suppose $nP(|X_n| > n) \to 0$ as $n \to \infty$. Then:

$$\frac{1}{n}\sum_{i=1}^{n} X_i - a_n \xrightarrow{P} 0$$

Consequence: weak LLN. if $\mathbb{E}[|X_1|] < \infty$, $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}X_1$

5. Strong LLN.

25

25

Tail σ-algebra Let $X_1, X_2 \dots$ be random variables. $\mathcal{F}_n' = \sigma(X_n, X_{n+1}, \dots)$ $\mathcal{T} = \bigcap_{n=1}^{\infty} \mathcal{F}_n'$: tail σ-algebra

Rmk. An event is in tail σ-algebra if it does not depend on any finite number of events

26

SLLN. $\{X_n\}$ iid, $\mathbb{E}[|X_1|] < \infty$ and $c_i \geq 1$ bounded. Then

$$\frac{1}{n}\sum_{i=1}^{n} c_i(X_i - \mathbb{E}[X_i]) \xrightarrow{a.s.} 0$$

7. Consistency of MoM and MLE.

26

26

Consistent of MoM.

Let $\mu \in \mathbb{R}^k$ be the first k-th moments of P_θ and $\mu = h(\theta)$ for some $h: \mathbb{R}^k \to \mathbb{R}^k$ continuous bijection function. Let $\hat{\mu}_n \in \mathbb{R}^k$ be empirical moments based on n iid random variables from P_θ. Then under assumptions on finite moments: $\hat{\mu}_n \xrightarrow{a.s.} \mu \Rightarrow h^{-1}(\hat{\mu}_n) \xrightarrow{a.s.} \theta$ as $n \to \infty$

27

Consistency of MLE.

For iid data, the negative log-likelihood function is the sum of n independent terms for every $\theta \in \Theta$

$$-\ell(\theta; y) = -\frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta)$$

Under some regularity conditions. MLE is $\sqrt{n}$-consistent.

1. Method of Moments

Classical recipe

$X_1, \cdots X_n \overset{iid}{\sim} P_\theta$, $\theta \in \Theta \subset \mathbb{R}^k$, $\mathbb{E}[|X_1|^k] < \infty$

Let $\mu_j = \mathbb{E}[|X_1|^j]$, $\hat{\mu}_j = \frac{1}{n}\sum_i X_i^j$

Suppose we can find certain Borel functions

$h_1, \cdots h_k : \mathbb{R}^k \to \mathbb{R}$ $\quad \mu_j = h_j(\theta)$

Then solve $\hat{\theta}$ by equations $\hat{\mu}_j = h(\hat{\theta})$, $j = 1, \cdots k$

Rmk. $\hat{\theta}$ may not exist, may not unique

Generalized Method of Moments

Suppose $X_1, \cdots X_n \overset{iid}{\sim} P_\theta$, $X_i \in \mathbb{R}^p$, $\theta \in \Theta$, where Θ is compact

For simplicity. Let $g : \mathbb{R}^{p+1} \to \mathbb{R}^m$ be cts s.t. $\mathbb{E}[g(X_i, \theta)] = 0$

Rmk m could be larger than p.

example of g: $g_k(x, \theta) = x^k - \mu_k(\theta)$ (MoM)

$g_k(x, \theta) = \cos(kx) - \mathbb{E}[\cos(kX_1)]$

$g_k(x, \theta) = \mathrm{LeakyReLU}(kx) - \mathbb{E}[\mathrm{LeakyReLU}(kX_1)]$

$$\mathrm{LeakyReLU}(z) = \begin{cases} z, & \text{if } z \geq 0 \\ \alpha z, & \text{if } z < 0 \end{cases}$$

where α is a small positive constant

Let $W \in \mathbb{R}^{m \times m}$, $W \succ 0$. Define GMM estimator as:

$$\hat{\theta} \in \arg\min \left(\frac{1}{n}\sum_{i=1}^n g(X_i, \theta)\right)^T W \left(\frac{1}{n}\sum_{i=1}^n g(X_i, \theta)\right).$$

Rmk. It is advantageous to pick $m > p$.

We hope to pick W s.t. useful moments have large weights.

2. Examples

Second moments and spectral methods

Factor Model Consider Factor Model with $Y_1, \cdots Y_n \in \mathbb{R}^p$ with

$Y_i = Bf_i + u_i$, $B \in \mathbb{R}^{p \times k}$, $f_i \in \mathbb{R}^k$, $u_i \in \mathbb{R}^k$

$\mathbb{E}[f_i] = 0$, $\mathrm{Cov}(f_i) = I_k$, $\mathbb{E}[u_i] = 0$, $\mathrm{Cov}(u_i) = \Sigma$

$\mathrm{Cov}(Y_i) = BB^T + \Sigma$, estimate $L = BB^T$

1. Using covariance matrix (second moment):

$$\hat{L}' = \frac{1}{n}\sum_{i=1}^n Y_i Y_i^T - \Sigma.$$

Rmk: condition $\mathrm{rank}(L) \leq k$ is not used

2. Using spectral decomposition. Let $\hat{L}' = U \Lambda U^T$

$\Lambda = \mathrm{diag}\{\lambda_1, \cdots \lambda_p\}$ (descending eigenvalues)

$U = [u_1, \cdots u_p]$. (corresponding eigenvectors)

put $\hat{L} = \sum_{i=1}^k \lambda_i u_i u_i^T$: new estimator satisfies $\mathrm{rank}(L) \leq k$

Latent Variable Model

Topic Model: k topics in a corpus. h: latent r.v.

Topic is drawn from $P(h = j) = w_j$, $j = 1, \cdots, k$

Word is drawn from a dist over vocabulary of size d: $\mu_j \in \mathbb{R}^d$

$x_t = e_i$ iff the t-th word in the document is i.

Let x_1, x_2, x_3 be word vectors in the same document.

We obtain the population moments:

$$M_2 = \mathbb{E}[x_1 \otimes x_2] = \sum_{k=1}^K w_k \mu_k \otimes \mu_k \in \mathbb{R}^{d \times d}$$

$$M_3 = \mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{k=1}^K w_k \mu_k \otimes \mu_k \otimes \mu_k \in \mathbb{R}^{d \times d \times d}$$

Goal: estimate $\{w_k\}_{k=1}^K$ and $\{\mu_k\}_{k=1}^K$ from M_1 and M_2

Orthogonal Tensor Decomposition: if $k < d$

1. Whitening: Use M_2 to determine a linear transformation $\bar{\mu}_k = B\mu_k \in \mathbb{R}^K$ such that $\bar{\mu}_1, \cdots \bar{\mu}_K$ are orthogonal

Exercise: $M_2 = \sum_{k=1}^K w_k \mu_k \otimes \mu_k = \sum_{k=1}^K w_k \mu_k \mu_k^T$

By spectral theorem. M_2 has eigenvectors $v_1, \cdots v_k, \cdots v_d$ consisting an orthonormal basis of $\mathbb{R}^d$ with descending eigenvalues $\lambda_1, \cdots \lambda_k, 0, \cdots 0$. (Suppose $\mu_1, \cdots \mu_K$ are l.i.)

Then define $\bar{\mu}_k = (v_1, \cdots v_d)^T \mu_k$ $\qquad UU^T = I$

$$\bar{\mu}_i^T \bar{\mu}_j = \mu_i^T (v_1, \cdots v_d)(v_1, \cdots v_d)^T \mu_k = \mu_i^T$$

2. Use B to transform $\bar{M}_3 = \sum_{i=k}^K \lambda_k \bar{\mu}_k \otimes \bar{\mu}_k \otimes \bar{\mu}_k$ for certain $\lambda_1, \cdots \lambda_K$

3. Apply Power Method to derive $\lambda_k, \bar{\mu}_k$ from M_3 numerically

3. Maximum Likelihood Estimation

$Y \sim P_\theta$, $\theta \in \Theta$, pdf: f_θ, $L(\theta, y) = f_\theta(y)$

$$\hat{\theta}_{mle} := \arg\max_{\theta \in \Theta} L(\theta, y) = \arg\max_{\theta \in \Theta} \ell(\theta, y)$$

Properties: Asym consistency, efficiency

Computational Complexity: Community Detection

$Y \in \mathbb{R}^{n \times n}$: symmetric adjacency matrix for n nodes, $Y_{ii} = 0$

$Y_{ij} \sim \mathrm{Bernoulli}(p^*)$ if i and j are in the same community

$Y_{ij} \sim \mathrm{Bernoulli}(q^*)$ if i and j are in different communities.

$\left.\begin{array}{l} z_i^* = 1: \text{ i-th node in comm 1} \\ z_i^* = -1: \text{ i-th node in comm 2} \end{array}\right\}$ equal-sized

$\theta = (p^*, q^*, \{z_i\})$ consider $\{z_i\}$ term:

$$\ell(\theta, Y) \propto \sum_{i<j} \mathbb{1}\{z_i z_j = 1\} Y_{ij}$$

Objective: $\max_{z \in \{\pm 1\}^n : z^T 1_n = 0} \sum_{i<j} \mathbb{1}\{z_i z_j = 1\} Y_{ij}$ (NP Hard)

47

1. Examples of concentration inequalities

47

Random graphs. The E-R graph $G \sim (n, p)$:
with adjacency matrix $A \in \mathbb{R}^{n \times n}$, $A_{ij} \overset{iid}{\sim} \mathrm{Bern}(p)$
· degree $d_i = \sum_{j \neq i} A_{ij}$, $\mathbb{E}[d_i] = (n-1)p$, put $d = (n-1)p$

47

Proposition there's an absolute constant C,
if $d \geq C(\log n)^2$ Then $P(0.9d \leq d_i \leq 1.1d, \forall i) \geq 0.9$

48

Covariance matrix Let $y_1, \cdots y_n \overset{iid}{\sim} Y \in \mathbb{R}^d$. $\mathbb{E}[Y] = 0$,
with subgaussian components. $\max_{1 \leq j \leq d} \|Y_j\|_{\psi_2} \leq K$.
Then the covariance matrix of Y exists
and is denoted by $\Sigma = \mathrm{Cov}(Y) \in \mathbb{R}^{d \times d}$
Empirical variance: $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T$

48

Denote $\|\hat{\Sigma} - \Sigma\|_{\max} := \max_{j,m} |\hat{\Sigma} - \Sigma|_{j,m}$.

$$= \max_{1 \leq j,m \leq d} \left|\frac{1}{n}\sum_{i=1}^{n}(y_{ij} y_{im} - \mathbb{E}[y_{ij} y_{im}])\right|$$

By Bernstein's inequality.

$$P([\hat{\Sigma} - \Sigma]_{j,m} \geq tK) \leq 2\exp(-cn\min(t^2, t))$$

Choose $t = C\sqrt{(\log d)/n}$ with large C s.t.
$cnt^2 = cC^2\log d \geq 4\log d$
Assume n is large enough s.t. $t \leq 1$, then $t^2 \leq t$

$$P(|\hat{\Sigma} - \Sigma|_{jm} \geq CK\sqrt{\frac{\log d}{n}}) \leq 2\exp(-4\log d) = 2d^{-4}$$

Apply union bound we have

$$P(\|\hat{\Sigma} - \Sigma\|_{\max} > CK\sqrt{\frac{\log d}{n}}) \leq 2d^{-2}$$

49

2. Two general concentration inequalities.

49

Thm 1. Gaussian Concentration inequality.
Let $X_1, \cdots X_n \overset{iid}{\sim} N(0,1)$. Let $F: \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz function. i.e. $|F(x) - F(y)| \leq L\|x - y\|$ Then

$$\|F(X_1, \cdots X_n) - \mathbb{E}[F(X_1, \cdots X_n)]\|_{\psi_2} \leq CL$$

where C is a constant

49

Thm 2. Talagrand's inequality Suppose $X_1, \cdots X_n$ are independent random variables with $\max |X_i| \leq K$.
Denote $X = (X_1, \cdots X_n)^T$. Let $F: \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz function. Then

$$P(|F(x) - \mathbb{E}[F(x)]| \geq tL) < C\exp(-\frac{ct^2}{K^2}), \forall t > 0$$

49

Rmk. It says bounded r.v. with indpt coordinates, $F(X) - \mathbb{E}F(X)$ is a subgaussian with norm bounded by $O(KL)$

50

3. Norms of random vectors.

50

Suppose $X = (X_1, \cdots X_n)^T$ with $X_1, \cdots X_n$ indpt and $\mathbb{E}[X_i] = 0$, $\mathrm{Var}(X_i) = 1$. We already have:

1. By LLN and cts mapping thm, $n^{-\frac{1}{2}}\|X\| \xrightarrow{P} 1$ as $n \to \infty$
2. By CLT and delta method, $\|X\| - \sqrt{n} \xrightarrow{d} N(0, 1/2)$
3. By Jensen's inequality, $\mathbb{E}[\|X\|] \leq \sqrt{\mathbb{E}[\|X\|^2]} = \sqrt{n}$
4. If $X_i \sim N(0,1)$ or $\max\|X_i\| \leq C$, then $\|X\| - \mathbb{E}\|X\|$ has subgaussian norm $O(1)$

50

Thm 3. Let $X = (X_1, \cdots X_n)^T$ be a random vector with $X_1, \cdots X_n$ indpt, $\mathbb{E}[X_i^2] = 1$, $\max_i \|X_i\|_{\psi_2} \leq K$. Then,

$$\|\|X\|_2 - \sqrt{n}\|_{\psi_2} \leq CK^2$$

Rmk. High-dim Gaussian is very close to uniform distribution on the sphere of radius $\sqrt{n}$.

51

Thm 4. Uniform spherical distribution is subgaussian.
$X \sim \mathrm{Unif}(\sqrt{n}\,\mathbb{S}^{n-1})$ satisfies

$$\sup_{u \in \mathbb{R}^n, \|u\| = 1} \|\langle X, u \rangle\|_{\psi_2} \leq C$$

Namely, projecting X using any unit vector gives a subgaussian random variable with a bounded norm

51

4. Random Projection.

52

Thm 5. Johnson-Lindenstrauss Lemma Let $\varepsilon \in (0, \frac{1}{2})$.
There exists $C > 0$ s.t. Let $k > (C\log N)/\varepsilon^2$ be an integer. Then $\exists f \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^k)$ s.t.

$$(1-\varepsilon)\|x - y\|^2 \leq \|f(x) - f(y)\|^2 \leq (1+\varepsilon)\|x - y\|^2$$

For all $x, y \in Q$.

54

2. Operator Norm of Subgaussian Random Matrix.

54

Thm 1. Let $A \in \mathbb{R}^{m \times n}$ have independent subgaussian entries (A_{ij} jointly indpt. subgaussian with $\mathbb{E} A_{ij} = 0$)

Then $\|A\| \le CK(\sqrt{m} + \sqrt{n} + t)$ with probability at least $1 - 2e^{-t^2}$ for certain constant C.

and $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$

54

Lemma 2. Let $X_1, \cdots X_n$ be independent subgaussian random variables with $\mathbb{E} X_i = 0$, then $\sum_{i=1}^{n} X_i$ is also subgaussian with

$$\left\|\sum_{i=1}^{n} X_i\right\|_{\psi_2}^2 \le C \sum_{i=1}^{n} \|X_i\|_{\psi_2}^2$$

54

Rmk $\|A\|$ is roughly of order $\sqrt{m} + \sqrt{n}$ with high probability. Assume $n \ge m$, $\sigma^2 = \mathbb{E} A_{ij}^2$

$$\|A\|_F^2 = \sum_{i,j} A_{ij}^2 \xrightarrow{P} nm\sigma^2$$

Since $\|A\|_F^2 = \sum_{j=1}^{m} [\sigma_j(A)]^2$, we have $\frac{1}{m} \sum_{k=1}^{m} \sigma_k^2(A) \xrightarrow{P} n\sigma^2$

The averaged singular value is of order $\sqrt{n}\,\sigma$.

with high prob $\Rightarrow$ so does the largest singular value.

55

3. Covariance Estimation

55

Suppose $X_1, \cdots X_n \in \mathbb{R}^p$ iid random vectors with $\mathbb{E} X_0 = 0$ and $\mathrm{Cov}(X_0) = \Sigma$ exists. By LLN, $\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \xrightarrow{P} \Sigma$

Thm 2. Assume $X_0 \in \mathbb{R}^p$ is a subgaussian random vector, ($\sup_{u \in S^{p-1}} \|\langle X_0, u \rangle\|_{\psi_2} \le K$.) Suppose $X_1, \cdots X_n$ are iid copies of X_0, $X = (X_1, \cdots X_n)^T \in \mathbb{R}^{n \times p}$

Then for every $t \ge 0$, with prob at least $1 - 2e^{-C_K t^2}$

$\left\|\frac{1}{n} X^T X - \Sigma\right\| \le \max\{\delta, \delta^2\} \|\Sigma\|$ where $\delta = C_K \sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}$

55

Thm 2'. Covariance Estimation (HDP).

Let $X_0 \in \mathbb{R}^p$ subgaussian ($\|\langle X, x \rangle\|_{\psi_2} \le K \|\langle X, x \rangle\|_{L^2}$ for some $K \ge 1$, any $x \in \mathbb{R}^p$). Then

$$\mathbb{E}\left[\left\|\frac{1}{n} X^T X - \Sigma\right\|\right] \le CK^2 \left(\sqrt{\frac{p}{n}} + \frac{p}{n}\right) \|\Sigma\|.$$

1. Principle Component Analysis.

56

Suppose $X_1, \cdots X_n \overset{iid}{\sim} X_0$, $\mathbb{E}[X_0] = 0$. $\Sigma = Cov(X_0)$ exists.
Let $\mu_1 \cdots \mu_p$ be orthonormal eigenvectors of Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
Then consider the empirical covariance $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$ with orthonormal eigenvectors $\hat{\mu}_1, \cdots \hat{\mu}_p$ and eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$.

1.1. Fixed Dimension

56

Thm 1 Consistency of PCA in fixed dimension.
Suppose $\lambda_1 > \lambda_2$. Then as $n \to \infty$, we have

$$\max_{k \leq p} |\hat{\lambda}_k - \lambda_k| \xrightarrow{P} 0 \qquad \min_{S \in \{\pm 1\}} \|S\hat{u}_1 - u_1\|_2 \xrightarrow{P} 0$$

Rmk. $\lambda_1 > \lambda_2$ ensures λ_1 has multiplicity of 1.

58

Davis-Kahan Theorem (HDP) Let S and T be symmetric matrices with the same dimensions. Assume

$$\min_{j \neq i} |\lambda_i(S) - \lambda_j(S)| = \delta > 0$$

Then the angle $\sin(v_i(S), v_i(T)) \leq \frac{2\|S - T\|}{\delta}$
where $v_i(\ )$ gives the i-th largest eigenvector.

1.2. Growing Dimension.

58

p_n, Σ_n depend on n. $X_0 \sim N(0, \Sigma_n)$
Spiked model: $\Sigma_n = (\lambda_1 - \lambda_2) u_1 u_1^T + \lambda_2 I_n$
eigenvalues are λ_1 and λ_2 with multiplicity $p-1$
Note that $\frac{1}{2}\min_{S \in \{\pm 1\}} \|S\hat{u}_1 - u_1\|_2^2 = 1 - |\langle \tilde{u}_1, u_1 \rangle|$

58

Inconsistency of PCA in high dimensions:
Assume $\|\Sigma_n\|_{op} \leq C$ and $\lambda_1(\Sigma_n) - \lambda_2(\Sigma_n) \geq \kappa > 0$ holds for certain constant C and κ

Rmk. population and sample principle direction are almost orthogonal in high dims.

2. Linear Regressions in High Dims.

59

Consider (x_i, y_i), $x_i \in \mathbb{R}^p$ and $i = 1, 2, \ldots, n$.

$$y_i = x_i^T\beta + \varepsilon_i, \quad i = 1, 2, \ldots n.$$

$X = (x_1, \ldots x_n)^T \in \mathbb{R}^{n \times p}$ $\quad y = (y_1, \cdots y_n)^T \in \mathbb{R}^n$
$\varepsilon = (\varepsilon_1, \cdots \varepsilon_n)^T \in \mathbb{R}^n$, $\mathbb{E}[\varepsilon_i] = 0$, $Var(\varepsilon_i) = \sigma^2$
- LSE: $\hat{\beta} = (X^TX)^{-1}X^Ty$
- Prediction: $\hat{f}(x_0) = x_0^T\hat{\beta}$

Under Parameterized Regime

59

Proposition 1. Suppose $x_0, x_1, \cdots x_n \overset{iid}{\sim} N(0, I_p)$.
Assume $p = p_n$, $\frac{p_n}{n} = o(1)$ as $n \to \infty$, we have

$$\mathbb{E}[\|\hat{\beta} - \beta\|^2 | X] = (1 + o(1))\frac{\sigma^2 p_n}{n}$$

$$\mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta)^2 | X] = (1 + o_p(1))\frac{\sigma^2 p_n}{n} \quad \text{as } n \to \infty$$

Rmk. the MSE increases approximately linear in p_n

Overparameterized Regime.

60

Minimum-Norm Interpolator:

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|^2$$
$$\text{subject to } y = X\beta$$

Assume $rank(X) = n$. The solution is $\underbrace{\hat{\beta} = X^T(XX^T)^{-1}y}_{\text{unbiased}}$
$\forall \beta^*$ s.t. $X\beta^* = y$, $\exists\, v \in N(X)$ s.t. $\beta^* = \hat{\beta} + v$.
Note that $\hat{\beta} \in C(X^T)$. and $C(X^T) \perp N(X)$
we have $\|\beta^*\|^2 = \|\hat{\beta}\|^2 + \|v\|^2 \geq \|\hat{\beta}\|^2$

60

Proposition 2. Suppose $x_1, \cdots x_n \sim N(0, I_p)$ and $p = p_n$ depends on n and $n/p_n = o(1)$ as $n \to \infty$. We have $rank(X) = n$ with probability $1 - o(1)$ and

$$\mathbb{E}[\|\hat{\beta} - \beta\|^2 | X] = (1 + o_p(1)) \cdot \left[(1 - \frac{n}{p_n})\|\beta\|^2 + \frac{\sigma^2 n}{p_n}\right]$$

$$\mathbb{E}[(x_0^T\hat{\beta} - x_0\beta)^2 | X] = (1 + o_p(1)) \cdot \left[\underbrace{(1 - \frac{n}{p_n})\|\beta\|^2}_{\text{Bias}} + \underbrace{\frac{\sigma^2 n}{p_n}}_{\text{Variance: Smaller when } p_n \gg n}\right]$$

63

1. Motivation

63

1.1. High-D Perspective

Let $X = \mu + Z$, $Z \sim N_p(0, I_p)$

When p is large:

$\|X\|^2 = \|\mu\|^2 + \|Z\|^2 + 2\langle \mu, Z\rangle \approx \|\mu\|^2 + p$

$\|X\| \gg \|\mu\|$ with high probability

Hoeffding:

$P(|\langle \mu, Z\rangle| > t) \leq 2\exp\left(-c\frac{t^2}{\|\mu\|^2}\right)$

if $\|\mu\| = O(1) \Rightarrow \langle \mu, Z\rangle = O_p(1)$

63

1.2. Bias-Variance trade-off perspective

$R_T(\mu) = \mathbb{E}\|T(X) - \mu\|^2 \quad X \sim N(\mu, I_p)$

Consider $\hat{\mu} = (1-\varepsilon)X$

$$\begin{aligned}\mathbb{E}\|\hat{\mu} - \mu\|^2 &= \varepsilon^2\|\mu\|^2 + (1-\varepsilon)^2 p \\ &= (1-2\varepsilon)p + \varepsilon^2(\|\mu\|^2 + p) \\ &= (1-2\varepsilon)p + O(\varepsilon^2)\end{aligned}$$

63

Goal: $\min_{\varepsilon \in [0,1]} \{\varepsilon\|\mu\|^2 + (1-\varepsilon)^2 p\}$.

$\|\mu\|$ is unknown, but we can use $\|X\|^2 \approx \|\mu\|^2 + p$ when p is large to estimate $\|\mu\|^2$

64

2. Shrinkage Estimators

64

James-Stein Estimator: noise $p \downarrow \Rightarrow \gamma \downarrow$

$\hat{\mu}_{JS} = \left(1 - \frac{p-2}{\|X\|^2}\right)X \qquad \hat{\mu}_{JS+} = \left(1 - \frac{p-2}{\|X\|^2}\right)_+ X$

64

Rmk. Shrinkage to 0 is not the only direction that works. For any $c \in \mathbb{R}^p$, we can make

$\hat{\mu}_{JS} = \left(1 - \frac{\gamma(p-2)}{\|X-c\|^2}\right)(X-c) + c$

$\hat{\mu}_{JS+} = \left(1 - \frac{\gamma(p-2)}{\|X-c\|^2}\right)_+(X-c) + c$

64

Theorem 1. If $\hat{\mu}$ is any of the two shrinkage estimators then for $\gamma \in (0,2)$, when $p \geq 3$

$$R_{\hat{\mu}}(\mu) < R_X(\mu) \text{ for every } \mu \in \mathbb{R}^p$$

Moreover, for $\hat{\mu} = \hat{\mu}_{JS}$, we have

$$R_{\hat{\mu}}(\mu) = p - (2\gamma - \gamma^2)(p-2)^2\,\mathbb{E}[\|X-c\|^{-2}]$$

64

Thm 2. Stein's Lemma: If $g \in C^1(\mathbb{R}^p)$ with $\mathbb{E}[\|\nabla g(Z)\|] < \infty$ where $Z \sim N(0, I_p)$, then

$$\mathbb{E}[\nabla g(Z)] = \mathbb{E}[Z g(Z)]$$

66

3.2. Shrinkage under Superharmonic Function

66

Def. For $f \in C^2(\mathbb{R}^p)$, denote the Laplace operator by $\Delta f(x) = \sum_{j=1}^p \partial_j^2 f(x)$. We say f is superharmonic iff $\Delta f(x) \leq 0$, $\forall x \in \mathbb{R}^p$

66

Thm 3. Let $f \in C^2(\mathbb{R}^p)$, $f > 0$, $\sqrt{f}$ is superharmonic. If $\mathbb{E}\left[\frac{1}{f(X)}\sum_{j=1}^p |\partial_j^2 f(X)|\right] < \infty$, $\mathbb{E}[\|\nabla \log f(X)\|^2] < \infty$ we have $\mathbb{E}[\|X + \nabla \log f(X) - \mu\|^2] = p + 4\mathbb{E}\left[\frac{\Delta\sqrt{f(X)}}{\sqrt{f(X)}}\right] \leq p$

66

Rmk. X is a minimax estimator $\Rightarrow X + \nabla \log f(X)$ is also a minimax estimator

67

4. Diffusion Model:

67

Forward Process: $X_t = \sqrt{1-\sigma_t^2}\,X_{t-1} + \sigma_t Z_t$

$Z_t \overset{iid}{\sim} N(0, I)$.

Score Function p_t: pdf of X_t

$\mathbb{E}[X_{t-1} | X_t = x_t] = \frac{1}{\sqrt{1-\sigma_t^2}}x_t + \frac{\sigma_t^2}{\sqrt{1-\sigma_t^2}}\underbrace{\nabla \log p_t(x_t)}_{\text{Score Function}}$

$X_{t-1} | X_t = x_t \sim N\left(\frac{1}{\sqrt{1-\sigma_t^2}}x_t + \frac{\sigma_t^2}{\sqrt{1-\sigma_t^2}}\nabla \log p_t(x_t),\ \sigma_t^2 I\right)$

67

Reconstruction: $X_{t-1} = \frac{1}{\sqrt{1-\sigma_t^2}}X_t + \frac{\sigma_t^2}{\sqrt{1-\sigma_t^2}}S(\theta, X_t) + \sigma_t Z_t$

$S(\theta, X_t)$ is trained to approximate $\nabla \log p_t(X_t)$:

$$\min_\theta \mathbb{E}_{t \sim \mathrm{Unif}(0,1),\, X_0 \sim p_0,\, X_t \sim p(X_t|X_0)}\left[\sigma_t\|\nabla \log p_t(X_t) - S_\theta(X_t, t)\|^2\right]$$

Lecture 6. Statistical Decision theory and Sufficiency

Basic Terminology

$$\text{Data} \sim (\Omega, \mathcal{F}, P)$$

$X_1, \dots X_n$: random variables $X = (X_1, \dots X_n)^T \in \mathbb{R}^{n\times p}$

Example. Graph Data E-R model $\mathcal{G}(n,p)$:

Graph G, n vertices, A: adjacency matrix.

$A_{ij} \sim \text{Bernoulli}(p)$, then $\mathbb{E}[A] = p\mathbb{1}_n\mathbb{1}_n^T$

$A \simeq \mathbb{E}A$ under some conditions. $T(A) \simeq T(\mathbb{E}A)$

Classifying Statistical Models

Parametric family. $P \in \{P_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$

e.g. family of densities: if $P_\theta \ll \nu$, $\forall \theta \in \Theta$, then we only need to consider $\{\frac{dP_\theta}{d\nu} : \theta \in \Theta\}$, particularly pdf / pmf

Example. $P \in \mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$. $\theta = (\mu, \sigma^2)^T$

Example (linear model) $\mathcal{P} = \{N(\beta^T X, \sigma^2) : \beta \in \mathbb{R}^p, \sigma^2 > 0\}$ $X \in \mathbb{R}^p$

Nonparametric family. Θ is a infinite dimensional space

Rk. Statistical Methods often use a finite number of estimates to approximate the unknown quantity

Example (Density estimation): Each distribution is abs cts and the set of pdf is $\{f(x) : |f'(x)| < L\}$ for some $L > 0$

Semiparametric Family: Finite parameters of interest + Infinite Nuisance Parameters

Example: Linear model with unknown dist of ε

Identifiability:

Def. $\mathbb{P} = \{P_\theta : \theta \in \Theta\}$ is identifiable iff $P_{\theta_1} \neq P_{\theta_2}$ for $\theta_1 \neq \theta_2$.

Example (Factor Model) Consider $Y \in \mathbb{R}^p$ has the structure

$$Y = Bf + u, \text{ where } f \sim N_r(0, I_r), \ u \sim N(0, \Sigma).$$

Σ: known, $B \in \mathbb{R}^{p\times r}$: unknown.

Equivalently $Y \sim (0, BB^T + \Sigma)$.

Since $BR(BR)^T = BB^T$ for any orthogonal matrix $R \in \mathbb{R}^{r\times r}$

this model is not identifiable.

Example: Multinomial regression: In classification problem:

$$p_k = \frac{\exp(\beta_k^T X)}{\sum_{j=1}^{K-1} \exp(\beta_j^T X)} \quad \text{where } k = 1, 2, \dots, K-1$$

$(\beta_1, \dots, \beta_K)^T$ and $(\beta_1 + c, \dots, \beta_K + c)$ stipulate the same model

To avoid this: $p_k = \dfrac{\exp(\beta_k^T x)}{\sum_{j=1}^{k}\exp(\beta_j^T x)+1}$

Statistics

Def. Let $X \in \mathbb{R}^{n\times p}$ be the data. For any msb function T, we call $T(X)$ a statistic

$\sigma(T(X)) \subset \sigma(X)$: $T(X)$ summarizes information from X.

Example Sample mean $\bar{X}$ and sample variance S^2

Order Statistics $X_{(1)} \cdots X_{(n)}$. pdf for $X_{(i)}$ is:

$$g_i(x) = \frac{n!}{(i-1)!\,(n-i)!}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{n-i} f(x)$$

Exponential Family:

Def. Dist with pdf $f_\theta(\omega) = \exp\{[\eta(\theta)]^T T(\omega) - \xi(\theta)\} h(\omega)$, $\forall \omega \in \Omega$

Canonical Form: $f_\eta(\omega) = \exp\{\eta^T T(\omega) - \xi(\eta)\} h(\omega)$, $\omega \in \Omega$

e.g.: Binomial, Normal, Dists subsumed in exponential family

Properties: (1) Marginal / Conditional dist remains exponential

(2) For Borel function f: $\int |f| \, dP_{\eta_0} < \infty$ → Probability Measure induced by the random variable.

$$\int f(\omega) \exp(\eta^T T(\omega)) h(\omega) \, d\nu(\omega) \in C^\infty$$

Cor. $m(t) \in C^\infty$ on Ω.

Basic Terminologies in Statistical Decision Theory.

- Decision Rule: $T: (\mathbb{R}^k, \mathcal{B}_k) \to (\underbrace{A}_{\text{Action Space}}, \mathcal{F}_A)$
- Loss Function: $L(P, T(x))$.
- Risk Function: $R_T(P) = \mathbb{E}[L(P, T(x))]$. For parametric family, it can be written as $R(\theta)$

  e.g. $MSE_{\hat\theta}(\theta) = \mathbb{E}_\theta[(\hat\theta - \theta)^2] = \int (\hat\theta(x) - \theta)^2 \, dP_\theta(x)$

Optimal Rule

Def. T_* is an optimal rule if

$R_{T_*}(P) \le R_T(P)$, $\forall P \in \mathcal{P}$, $\forall T \in \{\text{Class of allowable decisions}\}$

rate optimal if $R_{T^*}(P) < C R_T(P)$ for some C

Randomized decision rules.

For every x, $\delta(x, \cdot)$ is a probability measure on $(A, \mathcal{F}_A)$

Often, $\delta = \sum_{j=1}^{N} 1(Z = j) T_j$ where Z is a discrete random variable

- Risk of randomized rule: $R(\theta, \delta) = \mathbb{E}_{X \sim P_\theta}[\mathbb{E}_{T \sim \delta(x)}[L(\theta, A)]]$

e.g. Stochastic Gradient Descent

Admissibility

Let $\mathcal{G}$ be a class of decision rules. $T \in \mathcal{G}$ is $\mathcal{G}$-admissible iff $\forall S \in \mathcal{G}$, $R_S(P) > R_T(P)$, $\forall P \in \mathcal{P}$

Main Approaches for optimality:

- Bayes rule: Given a prior dist Π over $\mathcal{P}$, T_* minimizes

$$r_T(\Pi) = \int R_T(P)\, d\Pi(P)$$

- Minimax rule: $T_* = \arg\min_T (\sup_P R_T(P))$

Estimation.

For iid $X_1, \cdots X_n$, WTK $\mathbb{E}[X_i]$. Use squared loss $L(P, a) = (\theta - a)^2$, then

$$R_{\bar{X}}(P) = \frac{\sigma^2}{n} \quad \text{where } \sigma^2 = Var(X_i)$$

Bias: $b_T(P) = \mathbb{E}_P[T(X)] - \theta$

Hypothesis Test

$$H_0: P \in \mathcal{P}_0 \quad v.s. \quad H_1: P \in \mathcal{P}_1$$

Let $T(X) \in \{0, 1\}$, $T(X)$ must have the form $T_C(X) = \{X: T(X) = 1\}$. where C is called the rejection region. Consider the loss:

$$L(P, j) = 0 \text{ if } P \in \mathcal{P}_j, \text{ and } 1 \text{ otherwise}$$

Then
$$R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C), & P \in \mathcal{P}_0 \\ P(T(X) = 0) = P(X \notin C), & P \in \mathcal{P}_1 \end{cases}$$

Sufficient Statistics

Def. $T(x)$ is sufficient if the conditional dist of $X \mid T(X)=t$ does not depend on P or θ for any t

Factorization Thm: Suppose $\mathcal{P}=\{P_\theta : \theta \in \Theta\}$ is dominated by the Lebesgue measure ν. Then $T(x)$ is sufficient iff $\exists$ Borel functions h, g_θ s.t. $\frac{dP_\theta}{d\nu}(x) = g_\theta(T(x))\,h(x)$.

Example: exponential family $f_\theta(\omega) = \underbrace{\exp\{[\eta(\theta)]^T T(\omega) - \xi(\theta)\}}_{g_\theta(T(x))}\,\underbrace{h(\omega)}_{h(x)}$

Connection to information theory:

$T(x)$ is sufficient $\Leftrightarrow$ $I(X;\theta) = I(T(x);\theta)$

where θ is treated as a random variable

*** Data Processing Inequality.**

$X \to Y \to Z$ Markov Chain $X \perp\!\!\!\perp Z \mid Y$, then

$$I(X, Z) \le I(X; Y).$$

Lecture 7. Bias - Variance Trade - off.

1. Example: Sample mean. $X_1, \dots X_n \overset{iid}{\sim} (\mu, \sigma^2)$

Suppose we have some prior estimation $\hat{\mu}_0$ for μ.

Construct $\hat{\mu} = 0.2\hat{\mu}_0 + 0.8\bar{X}$, then

$$\text{Bias}(\hat{\mu}) = 0.2(\hat{\mu}_0 - \mu)$$

$$\text{Var}(\hat{\mu}) = 0.64\frac{\sigma^2}{n}$$

$$R(\mu, \hat{\mu}) = 0.04(\hat{\mu}_0 - \mu)^2 + 0.64\frac{\sigma^2}{n}.$$

Conclusion: $\hat{\mu}$ is better than $\bar{X}$ if $\hat{\mu}_0$ is close to μ.

2. Ridge Regression.

Obs: (x_i, y_i), $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ with model

$$y_i = x_i^T\beta^* + \varepsilon$$

ridge regression solves   ignore the $\frac{1}{n}$ term in lecture note

$$\min_{\beta} \underbrace{\sum_{i=1}^{n}(y_i - x_i^T\beta)^2}_{} + \lambda\|\beta\|_2^2 \quad (1)$$

$$\Rightarrow \hat{\beta}_\lambda = (X^TX + \lambda I_p)^{-1}X^Ty \qquad (X = (x_1, \dots x_n)^T \in \mathbb{R}^{n\times p})$$

$$= (X^TX + \lambda I_p)^{-1}X^T(X\beta^* + \varepsilon)$$

$$\text{MSE} = \mathbb{E}[\|\hat{\beta}_\lambda - \beta^*\|^2] = \mathbb{E}[\|\hat{\beta}_\lambda - \mathbb{E}\hat{\beta}_\lambda\|_2^2] + \|\mathbb{E}\hat{\beta}_\lambda - \beta^*\|_2^2$$

$$\text{Bias} = O(\lambda), \quad \sigma^2\text{Tr}((X^TX)^{-1}) - \text{Var} = O(\lambda)$$

$$\lambda\uparrow \Rightarrow \text{Var}\downarrow \quad \text{Bias}\uparrow$$

Exercise.

Bias:

$$\mathbb{E}[\hat\beta_\lambda] = (X^TX + \lambda I_p)^{-1} X^T X\beta^*$$
$$= (X^TX + \lambda I_p)^{-1}(X^TX + \lambda I_p - \lambda I_p)\beta^*$$
$$= \beta^* + \lambda (X^TX + \lambda I_p)^{-1}\beta^*$$

Let $M = (X^TX + \lambda I_p) \in \mathbb{R}^{p\times p}$

$$\|M^{-1}\| = \max_{x\neq 0} \frac{\|M^{-1}x\|}{\|x\|}$$
$$= \max_{x\neq 0} \frac{\|x\|}{\|Mx\|}$$
$$= \left(\min_{x\neq 0} \frac{\|Mx\|}{\|x\|}\right)^{-1}$$
$$= (\mu_{min}^*)^{-1} \leq (\mu_{min})^{-1}$$

where μ_{min}^* is the minimum eigenvalue of M.

μ_{min} is the smallest eigenvalue of X^TX

$$\text{Bias} := \|\lambda (X^TX + \lambda I_p)^{-1}\beta^*\|$$
$$\leq \lambda \|(X^TX + \lambda I_p)^{-1}\|\beta^*$$
$$\leq \lambda (\mu_{min})^{-1}\beta^* \qquad \Rightarrow \text{Bias} = O(\lambda)$$

Variance

$$\hat\beta_\lambda = (X^TX + \lambda I_p)^{-1}X^TX\beta^* + (X^TX + \lambda I_p)^{-1}X^T\varepsilon$$

$$\hat\beta_\lambda - \mathbb{E}[\hat\beta_\lambda] = (X^TX + \lambda I_p)^{-1}X^T\varepsilon$$

$$\mathbb{E}[\|\hat\beta_\lambda - \mathbb{E}\hat\beta\|^2] = \mathbb{E}[\varepsilon^T X (X^TX + \lambda I_p)^{-2}X^T\varepsilon]$$
$$= \sigma^2 \operatorname{tr}(X(X^TX + \lambda I_p)^{-2}X^T)$$

Spectral Decomposition: $X^TX = U\Lambda U^T$. $\quad X^TX + \lambda I_p = U(\Lambda + \lambda I_p)U^T$

$$X(X^TX+\lambda I_p)^{-2}X^T$$
$$= U\Lambda U^T U(\Lambda+\lambda I_p)^{-2}U^T U\Lambda U^T$$
$$= U\Lambda(\Lambda+\lambda I_p)\Lambda U^T$$
$$= U \,\mathrm{diag}\left\{\frac{\mu_i}{(\mu_i+\lambda)^2}\right\} U^T$$

where μ_i are eigenvalues of X^TX

$$\mathrm{Var}(\hat{\beta}_\lambda) = \sigma^2 \sum_{i=1}^{n} \frac{\mu_i}{(\mu_i+\lambda)^2} \qquad \frac{\mu_i}{(\mu_i+\lambda)^2} =$$

WTS: $\sigma^2 \mathrm{Tr}((X^TX)^{-1}) - \mathrm{Var} = O(\lambda)$

3. General Bias-Variance Tradeoff

Consider model $Y = f(x) + \varepsilon$, $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$, $x \in \mathbb{R}^p$

kNN estimator: $\hat{f}_k(x_0) = \frac{1}{k}\sum_{k=1}^{k} y_{x_{(k)}}$

where $x_{(1)}, \dots, x_{(k)}$ are k nearest points to x_0

$$\mathbb{E}[(Y - \hat{f}_k(x_0))^2 \mid X = x_0] = \sigma^2 + \underbrace{\left(f(x_0) - \frac{1}{k}\sum_{k=1}^{k} f(x_{(k)})\right)^2}_{\text{Bias}} + \underbrace{\frac{\sigma^2}{k}}_{\text{Variance}}$$

4. Overparametrization

$$\hat{f}(x) = x^T\hat{\beta}$$

$$\mathrm{Var}(\hat{f}(x_0) \mid x_0) = \sigma^2 x_0^T (X^TX)^{-1} x_0$$

$$\mathbb{E}[\hat{f}(x_0) \mid x_0] = x_0^T\beta \qquad \hat{\beta} = (X^TX)^{-1}X^Ty \perp\!\!\!\perp x_0$$

$$\mathrm{Var}(\hat{f}(x_0)) = \mathbb{E}[\sigma^2 x_0^T (X^TX)^{-1} x_0] + \mathrm{Var}(x_0^T\beta)$$

$$= \sigma^2 \mathbb{E}[\mathrm{tr}((X^TX)^{-1} x_0 x_0^T)] + \beta^T\beta$$

$$= \sigma^2 \mathrm{tr}((X^TX)^{-1}\mathbb{E}[x_0 x_0^T]) + \beta^T\beta$$

$$= n\sigma^2 \mathrm{tr}((X^TX)^{-1}) + \|\beta\|^2$$

$$\mathrm{Var}(x^T\hat{\beta}) = \sigma^2 x^T (X^TX)^{-1} x$$

If x_0 has zero mean and identity covariance matrix.

$$\mathrm{Var}(\hat{f}(x_0)) = \sigma^2 \mathbb{E}[x_0^T (X^TX)^{-1} x_0] = \sigma^2 \mathrm{tr}(\underbrace{(X^TX)^{-1}}_{\hookrightarrow \mathbb{R}^{p\times p}}) + \|\beta\|^2$$

The variance generally increases as p increases

Typically. $X^TX = \sum_{i=1}^{n} x_i^T x_i \Rightarrow \mathbb{E}[X^TX] = n\Sigma_x \quad \Sigma_x \in \mathbb{R}^{p\times p}$

If the smallest eigenvalue is bounded away from 0.

then the variance is $O(p/n)$

Double Descent and min-norm solution.

When $p > n$: $\min_\beta \|\beta\|_2^2$ (2) (ridgeless estimator)

subject to $X\beta = y$

Suppose $\mathrm{rank}(XX^T) = n$, then it has a unique solution

$$\hat{\beta}_{0+} = \lim_{\lambda \to 0+} \hat{\beta}_\lambda = X^T(XX^T)^{-1}$$

5. The implicit regularization of gradient descent

GD: $\beta^{(t+1)} = \beta^{(t)} - \eta \nabla L(\beta)$,

Starting at $\beta^{(0)} = 0$, solving $L(\beta) = \|Y - X\beta\|^2$

Thm. Suppose $p > n$, and $\text{rank}(XX^T) = n$. Then, if $\eta < [2\lambda_{max}(X^TX)^{-1}]$, then $\beta^{(t)}$ converges, and

$$\lim_{t\to\infty} \beta^{(t)} = \hat{\beta}_{ot}$$

Rmk. GD prefers the solution. Because

$$\nabla L(\beta) = 2X^T(Y - X\beta) \in C(X^T) \Rightarrow \beta^{(t)} \in C(X^T)$$

$$\Rightarrow \lim_{n\to\infty} \beta^{(t)} \in C(X^T)$$

and $\hat{\beta}_{ot}$ is the unique solution in $C(X^T)$

Thm (GD for overparametrized logistic regression)

Consider logistic loss and linear separable data.

From any initializer $\beta^{(0)} \in \mathbb{R}^p$, the gradient iterate

$\beta^{(k+1)} = \beta^{(k)} - \eta \nabla L_n(\beta^{(k)})$ satisfies $\beta^{(k)} = \hat{\beta} \log k + \Delta^{(k)}$

where $\|\Delta^{(k)}\|_2 = O(\log\log k)$ and

$\hat{\beta} = \arg\min \|\beta\|_2^2$ subject to $x_i^T\beta \geq 1$. for $i = 1, \cdots n$.

Remark: The DIRECTION of $\beta^{(k)}$ converge to $\hat{\beta}$.

and in logistic model, only direction matters.

Lecture 7 Exercises

Exercise 1: Solve ridge regression

$$\min_{\beta} \mathcal{L}(\beta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_2^2 \quad (1)$$

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta} = \frac{1}{2n} \sum_{i=1}^{n} 2(-x_i)(y_i - x_i^T \beta) + 2\lambda\beta$$

$$= \frac{-1}{n} \sum_{i=1}^{n} x_i (y_i - x_i^T \beta) + 2\lambda\beta$$

$$= \frac{-1}{n} \left(\sum_i x_i y_i - \sum_i x_i^T \beta x_i \right) + 2\lambda\beta$$

$$= \frac{-1}{n} (X^T y - X^T X \beta) + 2\lambda\beta$$

$$= \frac{-1}{n} X^T y + \left(\frac{1}{n} X^T X + 2\lambda\right)\beta$$

$$\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta^2} = \frac{1}{n} X^T X + 2\lambda I \succeq 0$$

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta} = 0 \Rightarrow \hat{\beta} = \left(\frac{1}{n} X^T X + 2\lambda I\right)^{-1} \frac{1}{n} X^T y$$

Exercise 2. Prove the ridge less solution

$$\hat{\beta}^{?} = \lim_{\lambda \to 0^+} \hat{\beta}_\lambda = X^T (X X^T)^{-1} y$$

First we prove two optimization problem are the same when $\lambda \to 0$

$\min_\beta L(\beta) = \|Y - X\beta\|^2 + \lambda \|\beta\|^2$ Let $B = \{\beta : Y - X\beta = 0\}$

For any $\beta \in B^c$ and $\tilde{\beta} \in B$

$L(\beta) - L(\tilde{\beta}) = \|Y - X\beta\|^2 + \lambda(\|\beta\| - \|\tilde{\beta}\|^2) > 0$ as $\lambda \to 0$

Then the minimizer must satisfy $Y - X\beta = 0$

Then we solve: $\min_\beta \|\beta\|_2^2$ (2)

subject to $X\beta = y$

Let $h(\beta) = X\beta - y$. Let $L(\beta, \lambda)$ be the Lagrangian

$L(\beta, \lambda) = \|\beta\|_2^2 - \lambda^T (X\beta - y)$ for some $\lambda \in \mathbb{R}^n$

$$\nabla L(\beta, \lambda) = \begin{pmatrix} 2\beta - X^T \lambda \\ X\beta - y \end{pmatrix} = 0$$

$$\Rightarrow \begin{array}{l} \beta = \frac{1}{2} X^T \lambda \\ X\beta = \frac{1}{2} X X^T \lambda - y = 0 \end{array}$$

$\Rightarrow \lambda = (\frac{1}{2} X X^T)^{-1} y$ since $X X^T$ invertible.

Then we have $\beta = X^T (X X^T)^{-1} y$

Lecture 8. Basic Estimation Methods

1. Method of Moments

Classical recipe

$X_1, \dots X_n \overset{iid}{\sim} P_\theta$, $\theta \in \Theta \subset \mathbb{R}^k$, $\mathbb{E}[|X_1|^k] < \infty$

Let $\mu_j = \mathbb{E}[|X_1|^j]$, $\hat{\mu}_j = \frac{1}{n}\sum_i X_i^j$

Suppose we can find certain Borel functions

$h_1, \dots h_k : \mathbb{R}^k \to \mathbb{R}$ $\quad \mu_j = h_j(\theta)$

Then solve $\hat{\theta}$ by equations $\hat{\mu}_j = h(\hat{\theta})$, $j = 1, \dots k$

Rmk. $\hat{\theta}$ may not exist, may not unique

Generalized Method of Moments

Suppose $X_1, \dots X_n \overset{iid}{\sim} P_\theta$, $X_i \in \mathbb{R}^p$, $\theta \in \Theta$, where Θ is compact

For simplicity. Let $g : \mathbb{R}^{p+1} \to \mathbb{R}^m$ be cts s.t. $\mathbb{E}[g(X_i, \theta)] = 0$

Rmk m could be larger than p.

example of g: $g_k(x, \theta) = x^k - \mu_k(\theta)$ (MoM)

$g_k(x, \theta) = \cos(kx) - \mathbb{E}[\cos(kX_1)]$

$g_k(x, \theta) = \mathrm{LeakyReLU}(kx) - \mathbb{E}[\mathrm{LeakyReLU}(kX_1)]$

$$\mathrm{LeakyReLU}(z) = \begin{cases} z, & \text{if } z \geq 0 \\ \lambda z, & \text{if } z < 0 \end{cases}$$

where λ is a small positive constant

Let $W \in \mathbb{R}^{m\times m}$, $W \succ 0$. Define GMM estimator as:

$$\hat{\theta} \in \operatorname{argmin} \left(\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta)\right)^T W \left(\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta)\right).$$

Rmk. It is advantageous to pick $m > p$.

We hope to pick W s.t. useful moments have large weights.

2. Examples

Second moments and spectral methods

Factor Model Consider Factor Model with $Y_1, \cdots Y_n \in \mathbb{R}^p$ with

$Y_i = Bf_i + u_i$, $B \in \mathbb{R}^{p\times k}$. $f_i \in \mathbb{R}^k$, $u_i \in \mathbb{R}^k$

$\mathbb{E}[f_i] = 0$, $\operatorname{Cov}(f_i) = I_k$, $\mathbb{E}[u_i] = 0$, $\operatorname{Cov}(u_i) = \Sigma$

$\operatorname{Cov}(Y_i) = BB^T + \Sigma$, estimate $L = BB^T$

1. Using covariance matrix (second moment):

$$\hat{L}' = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^T - \Sigma.$$

Rmk: condition $\operatorname{rank}(L) \le k$ is not used

2. Using spectral decomposition. Let $\hat{L}' = U \Lambda U^T$

$\Lambda = \operatorname{diag}\{\lambda_1, \cdots \lambda_p\}$ (descending eigenvalues)

$U = [u_1, \cdots u_p]$. (corresponding eigenvectors)

put $\hat{L} = \sum_{i=1}^{k} \lambda_i u_i u_i^T$: new estimator satisfies $\operatorname{rank}(L) \le k$

Latent Variable Model

Topic Model: k topics in a corpus. h: latent r.v.

Topic is drawn from $P(h=j)=w_j$, $j=1,\cdots,k$

Word is drawn from a dist over vocabulary of size d: $\mu_j \in \mathbb{R}^d$

$x_t = e_i$ iff the t-th word in the document is i.

Let x_1, x_2, x_3 be word vectors in the same document.

We obtain the population moments:

$$M_2 = \mathbb{E}[x_1 \otimes x_2] = \sum_{k=1}^{k} w_k \mu_k \otimes \mu_k \in \mathbb{R}^{d\times d}$$

$$M_3 = \mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{k=1}^{k} w_k \mu_k \otimes \mu_k \otimes \mu_k \in \mathbb{R}^{d\times d\times d}$$

Goal: estimate $\{w_k\}_{k=1}^{k}$ and $\{\mu_k\}_{k=1}^{k}$ from M_1 and M_2

Orthogonal Tensor Decomposition: if $k<d$

1. **Whitening:** Use M_2 to determin a linear transformation $\bar{\mu}_k = B\mu_k \in \mathbb{R}^k$ such that $\bar{\mu}_1, \cdots \bar{\mu}_k$ are orthogonal

Exercise: $M_2 = \sum_{k=1}^{k} w_k \mu_k \otimes \mu_k = \sum_{k=1}^{k} w_k \mu_k \mu_k^T$

By spectral theorem. M_2 has eigenvectors $v_1, \cdots v_k, \cdots v_d$ consisting an orthonormal basis of $\mathbb{R}^d$ with descending eigenvalues $\lambda_1, \cdots \lambda_k, 0, \cdots 0$. (Suppose $\mu_1 \cdots \mu_k$ are l.i.)

Then define $\bar{\mu}_k = (v_1 \cdots v_d)^T \mu_k$ $\qquad UU^T = I$

$$\bar{\mu}_i^T \bar{\mu}_j = \mu_i^T (v_1 \cdots v_d)(v_1 \cdots v_d)^T \mu_k = \mu_i^T$$

2. Use B to transform $\overline{M}_3 = \sum_{i=k}^{k} \lambda_k \overline{\mu}_k \otimes \overline{\mu}_k \otimes \overline{\mu}_k$ for certain $\lambda_1, \dots, \lambda_K$

3. Apply Power Method to derive $\lambda_k, \overline{\mu}_k$ from M_3 numerically

3. Maximum Likelihood Estimation

$Y \sim P_\theta$. $\theta \in \Theta$, pdf: f_θ, $L(\theta, y) = f_\theta(y)$

$$\hat{\theta}_{mle} := \arg\max_{\theta \in \Theta} L(\theta, y) = \arg\max_{\theta \in \Theta} \ell(\theta, y)$$

Properties: Asym consistency, efficiency

Computational Complexity: Community Detection

$Y \in \mathbb{R}^{n \times n}$: symmetric adjacency matrix for n nodes, $Y_{ii} = 0$

$Y_{ij} \sim \text{Bernoulli}(p^*)$ if i and j are in the same community

$Y_{ij} \sim \text{Bernoulli}(q^*)$ if i and j are in different communities

$$\left.\begin{array}{l} z_i^* = 1: \ i\text{th node in Comm } 1 \\ z_i^* = -1: \ i\text{th node in Comm } 2 \end{array}\right\} \text{equal-sized}$$

$\theta = (p^*, q^*, \{z_i\})$ consider $\{z_i\}$ term:

$$\ell(\theta, Y) \propto \sum_{i<j} \mathbb{1}\{z_i z_j = 1\} Y_{ij}$$

Objective: $\max_{z \in \{\pm 1\}^n : z^T 1_n = 0} \sum_{i<j} \mathbb{1}\{z_i z_j = 1\} Y_{ij}$ (NP Hard)

Note that $\mathbb{1}\{z_i z_j = 1\} = \dfrac{1 + z_i z_j}{2}$

$$\max_{z \in \{\pm 1\}^n : z^T 1_n = 0} \sum_{i<j} \mathbb{1}\{z_i z_j = 1\} Y_{ij}$$

$$\Leftrightarrow \max_{z \in \{\pm 1\}^n : z^T 1_n = 0} \sum_{i<j} z_i Y_{ij} z_j$$

$$\Leftrightarrow \max_{z \in \{\pm 1\}^n : z^T 1_n = 0} z^T Y z$$

$$\xrightarrow{\text{relax}} \max_{\|z\| = \sqrt{n}} z^T Y z$$

$\tilde{z}$: largest eigenvector of Y with $\|z\| = \sqrt{n}$

Discretization: $\hat{z}_i = 1$ if $\tilde{z}_i > 0$, otherwise, $\hat{z}_i = -1$

Lecture 9. Law of Large Number and Estimation Consistency

Ex: $X_n \xrightarrow{a.s} X \Rightarrow X_n \xrightarrow{P} X$.

$$P(\omega: \lim_{n\to\infty} X_n(\omega) = X(\omega)) = 1$$

$$P(\bigcup_{N=1}^{\infty} \bigcap_{n\geq N} \{\|X_n(\omega) - X(\omega)\| < \varepsilon\} = 1 \quad \text{for any } \varepsilon > 0.$$

put $A_{n,\varepsilon} = \{\omega : \|X_n(\omega) - X(\omega)\| < \varepsilon\}$

$$B_{N,\varepsilon} = \bigcap_{n\geq N} A_{n,\varepsilon} \qquad B_{1,\varepsilon} \subseteq B_{2,\varepsilon} \subseteq \cdots$$

$$\lim_{n\to\infty} P(B_{n,\varepsilon}) = P(\bigcup_{N=1}^{\infty} B_{N,\varepsilon}) = 1$$

$$B_{N,\varepsilon} \subseteq A_{n,\varepsilon} \Rightarrow \lim_{n\to\infty} P(A_{n,\varepsilon}) = 1$$

$$\Leftrightarrow \lim_{n\to\infty} P(\|X_n(\omega) - X(\omega)\|_2 > \varepsilon) = 0$$

Borel-Cantelli Lemma. Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\{A_n\}_{n\geq 1}$ be a sequence of events $A_n \subset \mathcal{F}$

1. $\sum_{i=1}^{\infty} P(A_n) < \infty \Rightarrow P(\limsup A_n) = 0$
2. If $\{A_n\}_n$ are independent, then
   $\sum_{i=1}^{\infty} P(A_n) = \infty \Rightarrow P(\limsup A_n) = 1$

$\limsup A_n := \bigcap_{m\geq 1} \bigcup_{n\geq m} A_n$

Pf. 1. Let $N(\omega) = \sum_{n=1}^{\infty} 1_{A_n}(\omega)$

$$E[N(\omega)] \overset{MCT}{=} \sum_{n=1}^{\infty} E[1_{A_n}] = \sum_{n=1}^{\infty} P(A_n) < \infty$$

$$\Rightarrow P(N = \infty) = 0$$

Alternatively:

$$P(\bigcap_{m\geq 1}\bigcup_{n\geq m} A_n) = \lim_{m\to\infty} P(\bigcup_{n\geq m} A_n) \leq \lim_{m\to\infty} \sum_{n=m}^{\infty} P(A_n) = 0$$

2. For $M < N$, $P(\bigcap_{n=M}^{N} A_n^c) = \prod_{n=M}^{N} (1 - P(A_n))$

$$\leq \prod_{n=M}^{N} e^{-P(A_n)}$$

$$= \exp(-\sum_{n=M}^{N} P(A_n))$$

$$\to 0 \quad \text{as } N \to \infty$$

$$P((\limsup_n A_n)^c) = \Pr((\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n^c))$$

$$= \lim_{m\to\infty} P(\bigcap_{n=m}^{\infty} A_n^c) = 0$$

$P(\bigcap_{n=M}^{\infty} A_n^c) = 0 \Rightarrow P(\bigcup_{n=M}^{\infty} A_n) = 1$ for any M

$$\Rightarrow P(\bigcap_{M=1}^{\infty}\bigcup_{n=M}^{\infty} A_n) = 1$$

Exercise. $\lim_{N\to\infty} \sum_{n=N}^{\infty} P(\|X_n - X\|_2 > \varepsilon) = 0$, $\forall \varepsilon > 0$

$$\Rightarrow \lim_{N\to\infty} P(\bigcup_{n=N}^{\infty} \{\|X_n - X\|_2 > \varepsilon\}) = 0$$

$\Leftrightarrow P(\bigcap_{N=1}^{\infty}\bigcup_{n=N}^{\infty} \{\|X_n - X\|_2 < \varepsilon\}) = 1$ i.e. $X_n \xrightarrow{a.s.} X$

Pf. $\lim_{N\to\infty} \sum_{n=N}^{\infty} P(\|X_n - X\|_2 > \varepsilon) = 0$

$\Rightarrow \sum_{n=1}^{\infty} P(\|X_n - X\|_2 > \varepsilon) < \infty$ i.e. $X_n \xrightarrow{a.s.} X$.

$$\Uparrow$$

$\Rightarrow P(\bigcap_{N=1}^{\infty}\bigcup_{n=N}^{\infty} \{\|X_n - X\|_2 > \varepsilon\}) = 0 \Rightarrow P(\bigcup_{N=1}^{\infty}\bigcap_{n=N}^{\infty} \{\|X_n - X\| \leq \varepsilon\}) = 1$

O_p and o_p Notations.

Def. Let $\{a_n\}$ be a sequence of real numbers, and $\{b_n\}$ be a sequence of positive numbers. Denote

$$\begin{cases} a_n = O(b_n) & \text{if there exists } C>0 \text{ s.t. } |a_n| < C b_n,\ \forall n \\ a_n = o(b_n) & \text{if } \frac{a_n}{b_n} \to 0 \text{ as } n \to \infty \end{cases}$$

Def. Let $X_1, X_2, \dots$ be random vectors and $Y_1, Y_2, \dots$ be random variables defined on a common prob space

(i) $X_n = O_p(Y_n)$ iff. $\forall \varepsilon > 0$, $\exists C > 0$ s.t.

$$\sup_n P(\|X_n\| > C|Y_n|) < \varepsilon$$

(ii) $X_n = o_p(Y_n)$ iff $\|X_n\|/|Y_n| \xrightarrow{P} 0$ as $n \to \infty$

Rmk. $\{X_n\}$ is said to be bounded in probability if $X_n = O_p(1)$

Exercise. $X_n \sim N(0, \sigma_n^2)$, $n = 1, 2, \dots$, then

$$\lim_{n\to\infty} \sigma_n^2 = \sigma^2 < \infty \iff X_n = O_p(1)$$

Pf. $P(|X_n| \geq C) \leq \dfrac{Var(X_n)}{C^2}$ by Markov's inequality

$$\limsup_n P(|X_n| \geq C) \leq \frac{\sigma^2}{C^2}$$

we could choose C^2 large enough s.t. $\frac{\sigma^2}{C^2} < \varepsilon$

"$\Leftarrow$". Suppose $\sigma_n^2 \to \infty$ as $n \to \infty$.

Note that $\frac{X_n}{\sigma_n} \sim N(0,1)$

$$P(|X_n| \geq C) \geq P(X_n \geq C) = (1 - \Phi(\frac{C}{\sigma_n}))$$

For any $C > 0$ we find σ_n large enough

Such that $\Phi(\frac{C}{\sigma_n}) \leq 0.8$.

$P(|X_n| \geq C) \geq 0.2$ for large enough n.

$\{|X_n|\}$ is not bounded

Proposition $X_1 \cdots X_n$ be random vectors. $Y_1, \cdots Y_n$ be random variables defined on a common probability space

(1) If $X_n = o_p(Y_n)$, then $X_n = O_p(Y_n)$

(2) If $X_n = o_p(1)$ and $Y_n = O_p(1)$, then $X_n Y_n = o_p(1)$

3. Estimation Consistency.

Suppose we have data $X^{(n)}$ for each n drawn from a distribution $P \in \mathcal{P}$, and an estimator $T_n(X) \in \mathbb{R}^p$

Def. Consistency Suppose $T_n(X)$ is an estimator of unknown parameter $\nu \in \mathbb{R}^p$. Let $(a_n)_{n \geq 1}$ be a sequence of positive numbers with $a_n \to \infty$

(1) $T_n(X)$ is consistent for ν iff $T_n(X) \xrightarrow{P} \nu$, $\forall P \in \mathcal{P}$

(2) $T_n(X)$ is a_n-consistent for ν iff $a_n[T_n(X) - \nu] = O_p(1)$ holds for any $P \in \mathcal{P}$

(3) $T_n(X)$ is strongly consistent iff $T_n(X) \xrightarrow{a.s.} \nu$, $\forall P \in \mathcal{P}$

4. Law of Large Numbers

Thm. (LLN. Simple Version) Suppose $X_1, X_2, \dots$ have finite second moment with the same mean $\mu = \mathbb{E}X_i$ and bounded variance $\sigma_i^2 = \mathrm{Var}(X_i) \le C$ for all i. Further assume that any pair of two random variables are uncorrelated. Then

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right)^2\right] = 0. \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mu.$$

Chebyshev's inequality. If Y is a random variable satisfying $\mathbb{E}[Y^2] < \infty$ and $a > 0$, then

$$P(|Y - \mathbb{E}Y| \ge a) \le \frac{\mathrm{Var}(Y)}{a^2}$$

Rmk. Finite higher moment implies finite lower moment.

$$(\mathbb{E}\|X\|_p^p)^{1/p} \le (\mathbb{E}\|X\|_q^q)^{1/q}$$

Exercise: Converse is false.

Let X has density function $f(x) = \frac{\alpha C^\alpha}{x^{\alpha+1}}$ for $x \ge C$, $1 < \alpha < 2$

$$\int f(x)\,dx = \alpha C^\alpha \int_c^\infty x^{-(\alpha+1)}\,dx = \alpha C^\alpha \left[-\frac{1}{\alpha} x^{-\alpha}\right]_c^\infty = 1$$

$$\mathbb{E}[X] = \alpha C^\alpha \int_c^\infty x^{-\alpha}\,dx = \frac{\alpha C}{\alpha - 1}$$

$$\mathbb{E}[X^2] = \alpha C^\alpha \int_c^\infty x^{-\alpha+1} \quad \text{diverges}$$

Thm $X_1, X_2 \dots$ be i.i.d. random variables, and $a_n = \mathbb{E}[X_1 \mathbb{1}\{|X_1| \le n\}]$

Suppose $nP(|X_n| > n) \to 0$ as $n \to \infty$. Then:

$$\frac{1}{n}\sum_{i=1}^{n} X_i - a_n \xrightarrow{P} 0$$

Consequence: weak LLN. if $\mathbb{E}[|X_1|] < \infty$, $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}X_1$

Pf (Exercise).

First we prove the consequence $a_n \to \mathbb{E}[X_1]$ by DCT.

$$nP(|X_n| > n) \le n \frac{\mathbb{E}|X_n|}{n} = \mathbb{E}|X_1|$$

Suppose for contradiction that $nP(|X_n| > n)$ does not converge to 0. Then $\exists$ subsequence $\{n_k\}$ s.t.

$$n_k P(|X_{n_k}| > n_k) \ge c \text{ for all } k \text{ and some } c > 0$$

$$P(|X_{n_k}| > n_k) \ge \frac{c}{n_k}$$

Then
$$\mathbb{E}[|X_1|] = \int_0^\infty P(|X_{n_k}| > t)\,dt$$

$$\ge \int_{n_k}^\infty P(|X_{n_k}| > t)\,dt$$

$$\ge \int_{n_k}^\infty \frac{c}{t}\,dt = c\,[\ln t]_{n_k}^\infty = \infty$$

Contradicts to $\mathbb{E}[|X_1|] < \infty$

Therefore, $nP(|X_n| > n) \to 0$

For the major part. Let $Y_{nj} = X_j \mathbb{1}\{|X_j| \le n\}$. Then

$$P\left(\frac{1}{n}\sum_{j=1}^{n} X_j \ne \frac{1}{n}\sum_{j=1}^{n} Y_{nj}\right) \le \sum_{j=1}^{n} P(X_j \ne Y_{nj}) = nP(|X_j| > n) \to 0$$

Thus, we only need to work on $\{Y_n\}$. Let $T_n = \frac{1}{n}\sum_{j\le n} Y_{nj}$

$$P(|T_n - \mathbb{E}T_n| \ge \varepsilon) \le \frac{\mathrm{Var}(Y_{nj})}{n\varepsilon^2} \le \frac{\mathbb{E}[Y_{nj}^2]}{n\varepsilon^2}$$

$$\le \frac{1}{n\varepsilon^2}\mathbb{E}[\min\{|X_j|, n\}^2]$$

$$= \frac{1}{n\varepsilon^2}\mathbb{E}\left[\int_0^n 2t\,\mathbb{1}(|X_j| \ge t)\,dt\right]$$

$$(\text{Fubini}) = \frac{2}{n\varepsilon^2}\int_0^n tP(|X_j| \ge t)\,dt \to 0$$

if $|X_j| \le t$

$\int_0^{|X_j|} 2t\,dt = |X_j|^2$

if $|X_j| \ge t$

$\int_0^n 2t\,dt = t^2$

$\frac{1}{n}\int_0^n tP(|X_j| \ge t)\,dt$: average of $tP(|X_j| \ge t)$

and $tP(|X_j| \ge t) \to 0$

5. Strong LLN.

Tail σ-algebra Let $X_1, X_2, \dots$ be random variables.

$\mathcal{F}_n' = \sigma(X_n, X_{n+1}, \dots)$ $\quad \mathcal{T} = \bigcap_{n=1}^{\infty} \mathcal{F}_n'$: tail σ-algebra

Rmk. An event is in tail σ-algebra if it does not depend on any finite number of events

Kolmogorov's 0-1 law $A \in \mathcal{T}$. $P(A) \in \{0, 1\}$

? Exercise $A_1 = \left\{\limsup_n \frac{X_1 + \dots + X_n}{n} \ge \mathbb{E}X_1 + \varepsilon\right\}$

$A_2 = \left\{\limsup_n \frac{X_1 + \dots + X_n}{n} \le \mathbb{E}X_1 - \varepsilon\right\}$

$P(A_1) = 1$

$\boxed{\text{SLLN}}$. $\{X_n\}$ iid, $\mathbb{E}[|X_1|] < \infty$ and $c_i \geq 1$ bounded. Then

$$\frac{1}{n}\sum_{i=1}^{n} c_i(X_i - \mathbb{E}[X_i]) \xrightarrow{a.s.} 0$$

7. Consistency of MoM and MLE.

Exercise. $X_1, \dots X_n \overset{iid}{\sim} P \in \mathcal{P}$. Assume $\mathbb{E}|X_1| = \mu < \infty$

$Var(X_1) = \sigma^2 < \infty$, Then $S_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

and $\frac{n-1}{n} S_n$ are strongly consistent for σ^2

Pf. $S_n = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i \bar{X} + \sum_{i=1}^{n} \bar{X}^2\right)$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right)$$

$$= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right) \xrightarrow{a.s.} \mathbb{E}[X^2] - \mathbb{E}[X]^2 = Var(X)$$

$$\frac{n-1}{n} S_n \xrightarrow{a.s.} Var(X)$$

Consistence of MoM.

Let $\mu \in \mathbb{R}^k$ be the first k-th moments of P_θ and $\mu = h(\theta)$ for some $h: \mathbb{R}^k \to \mathbb{R}^k$ continuous bijection function. Let $\hat{\mu}_n \in \mathbb{R}^k$ be empirical moments based on n iid random variables from P_θ. Then under assumptions on finite moments: $\hat{\mu}_n \xrightarrow{a.s.} \mu \Rightarrow h^{-1}(\hat{\mu}_n) \xrightarrow{a.s.} \theta$ as $n \to \infty$

Consistency of MLE.

For iid data, the negative log-likelihood function is the sum of n independent terms for every $\theta \in \Theta$

$$-\ell(\theta; y) = -\frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta)$$

Under some regularity conditions. MLE is $\sqrt{n}$ - consistant.

Lecture 10. Weak Convergence and CLT

1. Weak Convergence

Equivalence definition for $X_n \xrightarrow{d} X$:

- cdf $F_n(x) \to F(x)$ for each continuity point x of F.
- $\lim_{n\to\infty} \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$ for all bdd cts function h on $\mathbb{R}^d$
- Characteristic function $\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$ for all $t \in \mathbb{R}^d$
- $X \xrightarrow{d} X$ iff $c^T X_n \xrightarrow{d} c^T X$, $\forall c \in \mathbb{R}^d$

Thm. Let $\{X_i\}_{i\in\mathbb{Z}^+}$ be random vectors in $\mathbb{R}^d$

(i) $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{d} X$

(ii) $X_n \xrightarrow{d} c$, $c \in \mathbb{R}^k$ then $X_n \xrightarrow{P} c$

(iii). $X_n \xrightarrow{d} X$, then $X_n = O_p(1)$

2. Central Limit Theorem

Thm Let $X_1, X_2, \dots$ be iid random vectors in $\mathbb{R}^d$ with finite second moment. Let $\Sigma = \mathrm{Cov}(X_1)$, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \xrightarrow{d} N(0, \Sigma)$$

Lecture 11. Asymptotic normality and Delta Method

1. Convergence of Transformations

Continuous Mapping Theorem Let $X_1, X_2, \ldots$ be random vectors in $\mathbb{R}^d$ defined on a common probability space and g be a measurable function from $(\mathbb{R}^d, \mathcal{B}^d)$ to $(\mathbb{R}^k, \mathcal{B}^k)$. Suppose g is cts a.s. w.r.t. μ_X, then

(i) $X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$

(ii) $X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X)$

(iii) $X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$

Pf. (i). $A_0 = \{\omega : \lim_{n\to\infty} X_n = X\}$ $D = \{x \in \mathbb{R}^k . g \text{ cts at } x\}$

$\forall \omega \in A = A_0 \cap X^{-1}(D)$ $\lim g(X_n) = g(X)$

$A^c = A_0^c \cup X^{-1}(D)$ $P(A^c) \leq P(A_0^c) + P(X^{-1}(D)) = 0 \Rightarrow P(A) = 1$

(ii). $\forall \varepsilon > 0$. $\exists \delta > 0$ $\|g(x) - g(y)\| \leq \varepsilon$ if $\|x - y\| < \delta$

Since $\lim P(|X_n - X| > \delta) = 0$

$P(\|g(X_n) - g(X)\| > \varepsilon) \leq P(\|X_n - X\| \geq \delta) \to 0$ as $n \to \infty$

(iii) $X_n \xrightarrow{d} X \Rightarrow \exists Y_1, Y_2, \ldots,$ $Y_i \stackrel{d}{=} X_i$. and $Y_n \xrightarrow{a.s.} Y$

Slutsky's Theorem Let $X_1, X_2, \dots, Y_1, Y_2, \dots$ be random variables on a probability space. Suppose $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} c$. Then

(i) $X_n + Y_n \xrightarrow{d} X + c$

(ii) $X_n Y_n \xrightarrow{d} cX$

(iii) $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$

2. Delta Method

Thm (Delta method) Let $X_1, X_2, \dots$ and Y be random vectors in $\mathbb{R}^k$ satisfying $a_n(X_n - c) \xrightarrow{d} Y$. For $c \in \mathbb{R}^k$ and $\{a_n\}$ being a sequence of positive numbers with $\lim_{n\to\infty} a_n = \infty$. Let $g: \mathbb{R}^k \to \mathbb{R}$. If g is differentiable at c:

$$a_n[g(X_n) - g(c)] \xrightarrow{d} [\nabla g(c)]^T Y$$

3. Prevalence of asymptotic normality

Asymptotic Variance $\{a_n\}$. Positive sequence and either $a_n \to \infty$ or $a_n \to a > 0$. Assume

$$a_n(T_n(X) - \theta) \xrightarrow{d} Y \quad \text{with } 0 < \mathbb{E}Y^2 < \infty$$

(1) The asymptotic variance of T_n is defined as $\dfrac{\mathrm{Var}(Y)}{a_n^2}$

(2) Let $T_n'(X)$ be another estimator. Asymptotic relative efficiency is defined to be the ratio between the

the asymptotic variance

as $\mathbb{E}[Y^2] \leq \liminf_{n \to \infty} \mathbb{E}[a_n^2 (T_n - \theta)^2]$

Lecture 12. Unbiased Estimation and UMVUE.

1. Complete Statistic

Def. (Complete Statistics) $T(X)$ is complete for $P \in \mathcal{P}$ iff $\forall$ Borel f, $E[f(T)] = 0 \Rightarrow f(T) = 0$ a.s. $\mathcal{P}$

Rmk: Requires the statistic has no redundant information.

Proposition. Let $\mathcal{P} = \{P_\eta, \eta \in \Xi\}$ be an exponential family of full rank (contains an open set) with pdf

$$f_\eta(x) = \exp\{\eta^T T(x) - \xi(\eta)\} h(x)$$

Then $T(X)$ is sufficient and complete for $\eta \in \Xi$

2. UMVUE

Def. An unbiased estimator $T(X)$ of ν is called the UMVUE iff $Var(T(X)) \le Var(U(X))$ for any $P \in \mathcal{P}$ and any other unbiased estimator $U(X)$ of ν

Rmk UMVUE does not always exist

Thm. Suppose there exists a sufficient and complete statistic $T(X)$ for $\mathcal{P}$. If ν is estimable then there exists unique UMVUE, which is of form $h(T)$ where h is a Borel function.

exists unbiased estimator

3. Construct UMVUE

Method 1. Find sufficient and complete statistic $T(X)$, then find $h(T)$ such that $E[h(T)] = \nu$, $\forall P \in \mathcal{P}$

Method 2. Find sufficient, complete $T(X)$ and unbiased $U(X)$, then $E[U|T]$ is an UMVUE.

Method 3. Find UMVUE without knowing complete statistics

Thm 2 Let $\mathcal{U} = \{U : E[U(X)] = 0,\ Var(U(X)) < \infty,\ \forall P \in \mathcal{P}\}$

T is unbiased for ν with $E[T(X)] < \infty$

(1). $T(X)$ is UMVUE iff

$$E[T(X)U(X)] = 0,\ \forall U \in \mathcal{U},\ \forall P \in \mathcal{P}$$

(2) If $\tilde{T}$ is sufficient for P, let

$$\tilde{\mathcal{U}} = \mathcal{U} \cap \{g(\tilde{T}) : g \text{ Borel}\}$$

Then $T = h(\tilde{T})$ is UMVUE iff

$$E[T(X)U(X)] = 0,\ \forall U \in \tilde{\mathcal{U}},\ \forall P \in \mathcal{P}$$

Method 4*: Variational Calculus

Lecture 13: Fisher information and C-R Lower Bound.

1. Fisher Information

Def. For $X \sim \{P_\theta : \theta \in \Theta\}$:

$$I_X(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^T\right]$$

Rmk. $I(\theta) \succeq 0$. $\frac{\partial}{\partial\theta}\log f_\theta(x)$ is called the score function

Proposition

1. If $X \perp\!\!\!\perp Y$, then $I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta)$

2. Suppose f_θ is twice differentiable at θ, under some "regularity condition", we have

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f_\theta(X)\right].$$

2. Cramér-Rao Lower Bound

Let $v = g(\theta)$, $g: \mathbb{R}^d \to \mathbb{R}^k$ differentiable. Suppose X is drawn from $\{P_\theta : \theta \in \Theta\}$. and $T(X)$ is an unbiased estimator of v.

Thm. Suppose $I(\theta)$ is positive definite, and for any $\theta \in \Theta$,

$$\frac{\partial}{\partial\theta}\int h(x) f_\theta(x)\,d\nu = \int h(x)\frac{\partial}{\partial\theta}f_\theta(x)\,d\nu$$

holds for $h \equiv 1$ or $h(x) = T(x)$. Then,

$$\mathrm{Var}(T(X)) \succeq \left[\frac{\partial}{\partial\theta}g(\theta)\right]^T \left[I(\theta)\right]^{-1}\frac{\partial}{\partial\theta}g(\theta)$$

Rmk (Reparameterization). If $\theta = \psi(\eta)$, $\psi \in C^1$ is bijective,

then $I_X(\eta) = \frac{\partial}{\partial \eta}\psi(\eta)\, I_X(\psi(\eta))\, \frac{\partial}{\partial \eta}\psi(\eta)^T$

is different but

$$\frac{\partial}{\partial \eta} g(\theta) = \frac{\partial}{\partial \theta} g(\theta) \frac{\partial}{\partial \eta}\psi(\eta).$$

Therefore, the C-R lower bound is **invariant** under reparameterization

Rmk. Asymptotic optimality of MLE.

$\mathrm{Var}(\mathrm{MLE}) \to (n I_X(\theta))^{-1}$ matches the C-R lower bound

Rmk. C-R lower bound may not be attained

3. Interpretations of Fisher information.

3.1. Geometric view: Larger $I(\theta)$ => larger local convexity

3.2. Information theory: under certain "regularity conditions":

$$D(P_{\theta_0} \| P_{\theta_0+\xi}) = \frac{1}{2}\xi^T I(\theta_0)\xi + o(\|\xi\|^2)$$

"How hard to distinguish two dist in a parametric family under the KL divergence"

5. Examples

5.1. Exponential Family.

Let $\Theta \subset \mathbb{R}^k$ be an open set. Let $\{f_\theta : \theta \in \Theta\}$ be

$$f_\theta(x) = \exp\{\eta(\theta)^T T(x) - \xi(\theta)\} c(x)$$

Proposition

1. The regularity conditions for C-R lb is satisfied for any msb h with $\mathbb{E}[h(X)] < \infty$
2. Consider natural parameter η, $Var(T(X)) = I(\eta)$
3. $\nu = \mathbb{E}[T(X)]$, then $Var(T) = (I(\nu))^{-1}$

5.2 Linear Models. $X_i = \beta^T Z_i + \varepsilon_i$ $\quad Z \in \mathbb{R}^{n \times p}$ $\quad \varepsilon \sim N(0, \sigma^2 I_n)$

Thm.2 Suppose Z is full column rank

(i) $\ell^T\hat{\beta}$ is UMVUE of $\ell^T\beta$, $\forall \ell \in \mathbb{R}^p$

(ii) $\hat{\sigma}^2 = \frac{1}{n-p}\|X - Z\hat{\beta}\|^2$ is the UMVUE of σ^2

Pf (Sketch) $(Z\hat{\beta}, \|X - Z\hat{\beta}\|^2)$ is complete and sufficient for $\theta = (\beta, \sigma^2)$. Further. verify unbiasedness

Thm 3 Under assumptions in thm 2, $\ell^T\hat{\beta}$ is independent of $\hat{\sigma}^2$, moreover,

$$\ell^T\hat{\beta} \sim N(\ell^T\beta, \sigma^2 \ell^T(Z^TZ)^{-1}\ell), \quad \hat{\sigma}^2 \frac{n-p}{\sigma^2} \sim \chi^2_{n-p}$$

Fisher information $\varepsilon \overset{iid}{\sim} N(0, \sigma^2 I_n)$ and the unknown parameters are $\theta = (\beta, \sigma^2)$

$$I(\theta) = \frac{1}{\sigma^2}\begin{pmatrix} Z^T Z & 0 \\ 0 & \frac{n}{2\sigma^2} \end{pmatrix}$$

If Z is not full rank, $I(\theta)$ may not be invertible.

Two distributions P_θ and $P_{\theta'}$ are not distinguishable

Prop. LSE estimate $\hat{\beta}$ attains C-R lower bound.

Lecture 14. Concentration Inequality

1. Motivation Non-asymptotic bounds (finite sample).

Gaussian Tail Inequality Let $G \sim N(0,1)$. then

$$\left(\frac{1}{t}-\frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \le P(G \ge t) \le \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \quad \forall t>0$$

Let F and P be the Gaussian CDF and pdf. then

$$1-F(t) = (1+o(1))\frac{1}{t}P(t)$$

Berry-Essen Inequality. Suppose $X_1, \cdots X_n$ iid with $\mathbb{E}[X_1]=0$ and $\mathbb{E}[X_1^3] < \infty$, then

$$\left| P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \ge t\right) - P(G \ge t) \right| \le \frac{C\,\mathbb{E}[X^3]}{\sqrt{n}}$$

Rmk. The general bound is too loss.

2. Subgaussian variables and Hoeffding's Inequality

Thm 1. Equivalent defs of Subgaussian random variables

(1) $P(|X| \ge t) \le 2\exp(-t^2/k_1^2)$, $\forall t>0$ for some $k_1>0$

(2) $\|X\|_p \le k_2\sqrt{p}$ for some $k_2>0$ and all $p \in \mathbb{N}^+$

(3) $\mathbb{E}\exp(\lambda^2 X^2) \le \exp(k_3^2\lambda^2)$ for some $k_3>0$ and all λ w. $|\lambda| \le \frac{1}{|k_3|}$

(4) $\mathbb{E}\exp(X^2/k_4^2) = m_{X^2}\left(\frac{1}{k_4^2}\right) \le 2$ for some $k_4>0$

(5) If $\mathbb{E}X = 0$, $\mathbb{E}\exp(\lambda X) \le \exp(K_5^2 \lambda^2)$, $\forall \lambda \in \mathbb{R}$

Remark. The smallest K_2 is called the subgaussian norm of X, denoted by $\|X\|_{\psi_2}$

Examples
- If $X \sim N(0, \sigma^2)$, then $\|X\|_{\psi_2} \le C\sigma$
- If $|X| \le M$ a.s. then $\|X\|_{\psi_2} \le CM$
- $X \sim$ Bernoulli (0.5) $\|X\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}$

Pf. (1) => (2). Assume $K_1 = 1$:

$$\mathbb{E}|X|^p = \int_0^\infty P(|X|^p \ge u)\, du$$
$$= \int_0^\infty P(|X| \ge t)\, dt^p$$
$$\le \int_0^\infty 2e^{-t^2} p t^{p-1}\, dt$$
$$= 2p \int_0^\infty e^{-v} v^{\frac{1}{2}(p-1)}\, dv^{\frac{1}{2}}$$
$$= p \int_0^\infty e^{-v} v^{\frac{1}{2}p-1}\, dv$$
$$= p\,\Gamma(\tfrac{1}{2}p)$$
$$\le 3p(p/2)^{p/2} \Rightarrow \|X\|_p \le \frac{3}{\sqrt{2}} p^{1/p} p^{1/2} \le \frac{3}{\sqrt{2}}\sqrt{p}$$

we have (2) with $K_2 \le 3$.

Gamma Function:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$$

Stirling's Approximation

$$n! / \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \to 1$$
$$n! \le \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$$
$$\Gamma(x) \sim x^{x-\frac{1}{2}} e^{-x} \sqrt{2\pi}$$
$$\Gamma(x) \le 3x^x$$

Stirling's approximation also yields $p! \ge (p/e)^p$

(2) => (1) Assume $K_3 = 1$.

$$\mathbb{E}\exp(\lambda^2 X^2) = \mathbb{E}\left[1 + \sum_{p\ge 1}^\infty \frac{(\lambda^2 X^2)^p}{p!}\right] = 1 + \sum_{p\ge 1}^\infty \frac{\lambda^{2p}\mathbb{E}[X^{2p}]}{p!}$$

$$\leq 1+\sum_{p\geq 1}^{\infty}\frac{\lambda^{2p}(2p)^p}{(p/e)^p}=\sum_{p\geq 0}^{\infty}(2e\lambda^2)^p=\frac{1}{1-2e\lambda^2}\leq \exp(4e\lambda^2)$$

Suppose $2e\lambda^2<1$ $\quad$ $2e\lambda^2\in[0,\frac{1}{2}]$

For $|\lambda|\leq\frac{1}{2\sqrt{e}}$. and $K_3=2\sqrt{e}$

(3) => (4): Trivial.

(4) => (1). Assume $K_4=1$.

$$P(|X|\geq t)=P(e^{X^2}\geq e^{t^2})$$
$$\leq e^{-t^2}\mathbb{E}e^{X^2}\leq 2e^{-t^2}$$

$$P(X\geq a)\leq\frac{\mathbb{E}[X]}{a}$$

When we have $\mathbb{E}X=0$:

(3) => (5): Assume $K_3=1$.

$$\mathbb{E}[e^{\lambda X}]\leq\mathbb{E}[\lambda X+e^{\lambda^2X^2}]$$
$$=\mathbb{E}[e^{\lambda^2X^2}]\leq e^{\lambda^2}\text{ for }|\lambda|\leq 1.$$

When $|\lambda|\geq 1$: $\quad 2\lambda X\leq\lambda^2+X^2$

$$\mathbb{E}e^{\lambda X}\leq e^{\lambda^2/2}\mathbb{E}e^{X^2/2}\leq e^{\lambda^2/2}e^{1/2}\leq e^{\lambda^2}$$

(5) => (1): For any $\lambda>0$:

$$P(X\geq t)=P(e^{\lambda X}\geq e^{\lambda t})$$
$$\leq e^{-\lambda t}\mathbb{E}e^{\lambda X}$$
$$\leq e^{-\lambda t}e^{\lambda^2}$$
$$=e^{-\lambda t+\lambda^2}$$
$$\leq e^{-t^2/2}$$

Maximum of sub-gaussians. Let $X_1, X_2, \dots$ be a sequence of sub-gaussian random variables, not necessarily indpe.
Then $\mathbb{E}\max_i \frac{|X_i|}{\sqrt{1+\log i}} \le CK$
where $K = \max_i \|X_i\|_{\psi_2}$. For any $N \ge 2$, we have
$$\mathbb{E}\max_{i \le N}|X_i| \le CK\sqrt{\log N}$$

Rmk. the bound is sharp: Let $X_1, X_2, \dots \overset{iid}{\sim} N(0,1)$, we have $\mathbb{E}\max_{i \le N} X_i \ge c\sqrt{\log N}$.

Thm 2. Hoeffding's Inequality Let $X_1, X_2, \dots X_n$ be independent, mean zero and subgaussian r.v.s.
Let $a \in \mathbb{R}^n$, $K := \max_{1 \le i \le n}\|X_i\|_{\psi_2}$. Then for any $t > 0$
$$P\left(\left|\sum_{i=1}^n a_i X_i\right| > t\right) \le 2\exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right)$$
holds for certain absolute constant $c > 0$

Rmk. More flexible than CLT: Mean $\to$ weighted sum

Pf. $\mathbb{E}[\exp(\lambda \sum_{i=1}^n a_i X_i)] = \prod_{i=1}^n \mathbb{E}[\exp(\lambda a_i X_i)]$
$$\le \prod_{i=1}^n \exp(C\lambda^2 a_i^2 \|X_i\|_{\psi_2}^2)$$
$$\le \exp(C\lambda^2 \sum_{i=1}^n a_i^2 K^2)$$
$$= \exp(C\lambda^2\|a\|^2 K^2)$$ Then use Thm 1 (1)

Khintchine's inequality Let $X_1, X_2, \dots X_N$ be independent sub-gaussian random variables with zero-means and unit variance, let $a \in \mathbb{R}^N$. Then for any $p \in [2, +\infty)$,

$$\|a\|_2 \le \|\sum_{i=1}^{N} a_i X_i\|_p \le CK\sqrt{p}\,\|a\|_2$$

where $K = \max_i \|X_i\|_{\psi_2}$ and C is an absolute constant

3. Subexponential Variables and Bernstein's Inequality

Thm 3. (Equivalent defs of Subexponential random variables)

(1) $P(|X| > t) \le 2e^{-t/K_1}$, $\forall t \ge 0$ relaxed subgaussian.

(2) $\|X\|_p \le K_2 p$ for all $p \ge 1$

(3) $\mathbb{E}[\exp(\lambda|X|)] \le \exp(K_3\lambda)$, $\forall \lambda \in [0, 1/K_3]$

(4) $\mathbb{E}[\exp(|X|/K_4)] \le 2$

(5) If $\mathbb{E}X = 0$, $\mathbb{E}\exp(\lambda X) \le \exp(K_5^2\lambda^2)$, $\forall \lambda \in [-1/K_5, 1/K_5]$

$K_i \le CK_j$. $\forall i \ne j$ for some C independent of the random variable X.

Remark

- X is called sub-exponential if any of the condition holds
- The smallest K_2 is called the sub-exponential norm $\|X\|_{\psi_1}$
- subgaussian $\Rightarrow$ subexponential, $\|X\|_{\psi_1} \le \|X\|_{\psi_2}$

Centering For random variables with non-zero mean, conditions (1) and (2) in thm 1 and thm 3 are still equivalence. Moreover, we have bound

$$\|X-\mathbb{E}X\|_{\psi_2} \le C\|X\|_{\psi_2}$$

Cor. Suppose X, Y are subgaussian. Then XY are ~~subgaussian~~ subexponential, with

$$\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$$

Pf.

$$\left[\mathbb{E}(|X|^p|Y|^p)\right]^{1/p} \le \left[\mathbb{E}|X|^{2p}\,\mathbb{E}|Y|^{2p}\right]^{1/2p}$$

$$\le \|X\|_{\psi_2}\sqrt{p}\,\|Y\|_{\psi_2}\sqrt{p}$$

$$\frac{1}{p}\left[\mathbb{E}(|X|^p|Y|^p)\right]^{1/p} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$$

Thm 4. Bernstein's Inequality

Let $X_1, \cdots X_n$ be independent r.v. such that $\mathbb{E}X_i = 0$ and each X_i is subexponential. Let $a \in \mathbb{R}^n$, $K = \max_{1\le i\le n}\|X_i\|_{\psi_1}$. Then

$$P\left(\left|\sum_{i=1}^n a_iX_i\right| \ge t\right) \le 2\exp\Big(-c\underbrace{\min\Big(\overset{\text{subgaussian}}{\frac{t^2}{K^2\|a\|^2}}, \overset{\text{subexponential}}{\frac{t}{K\|a\|_\infty}}\Big)}_{\text{decays no faster than } e^{-t}}\Big)$$

for some constant $c > 0$

Remark. Subexponential r.v. has "heavier tails" compared with subgaussian r.v.s.

Thm 5. Bernstein's Inequality for bounded random variables

Let $X_1 \cdots X_n$ be indpt r.v.s, $\mathbb{E}X_i = 0$, $\max_i |X_i| \le M$ a.s.

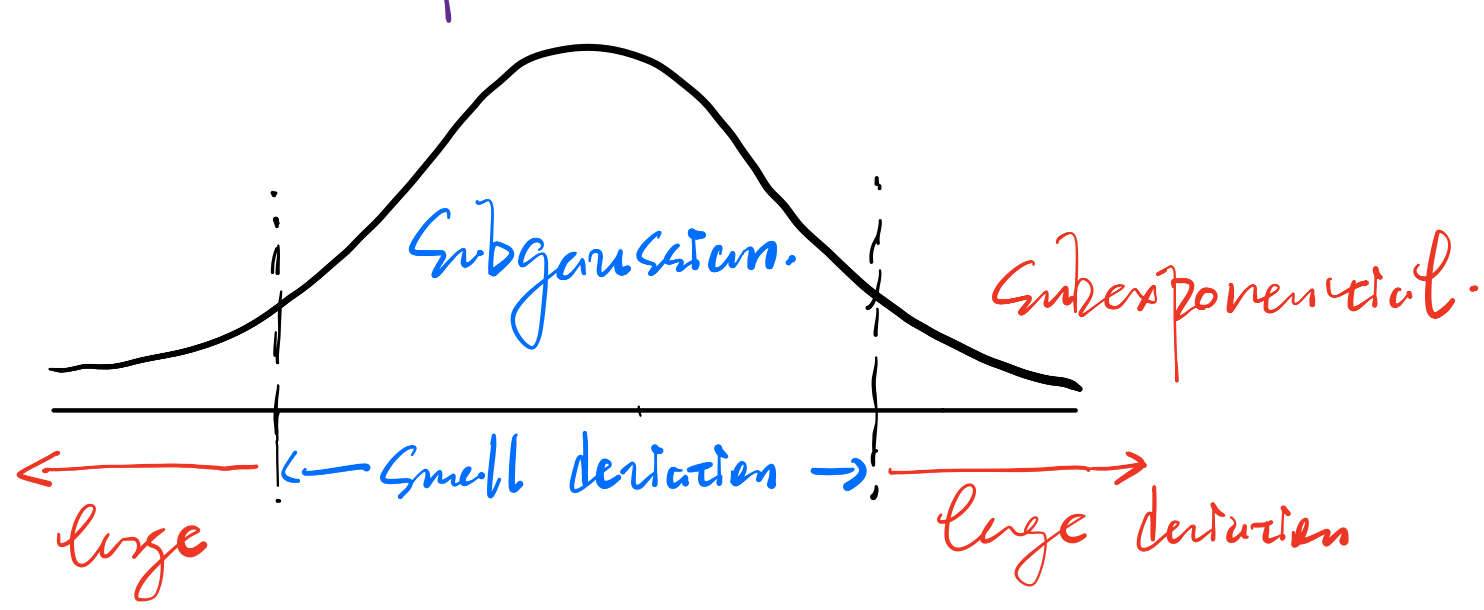
$$P(|\sum_{i=1}^n X_i| \ge t) \le 2\exp\left(-\frac{t^2/2}{\sigma^2 + Mt/3}\right)$$

where $\sigma^2 = \sum_{i=1}^n \mathbb{E}X_i^2$ is total variance

Variant of Berstein's Inequality:

$$P\left(|\sum_{i=1}^n X_i| \ge \frac{C(\sigma+M)}{\delta}\right) \le 2\exp\left(-\min\left(\frac{1}{\delta}, \frac{1}{\delta^2}\right)\right), \ \forall \delta > 0$$

Sum of sub-exponential r.v.s



HDP 2.3. Chernoff's inequality

Thm (Chernoff's inequality) Let X_i be independent Bernoulli random variables with parameters p_i. Consider $S_N = \sum_{i=1}^{N} X_i$, $\mu := \mathbb{E} S_N$. Then for any $t > \mu$:

$$\mathbb{P}(S_N > t) \le e^{-\mu}\left(\frac{e\mu}{t}\right)^t$$

For any $t < \mu$:

$$\mathbb{P}(S_N \le t) \le e^{-\mu}\left(\frac{e\mu}{t}\right)^t$$

Cor. Poisson Tails: Let $X \sim \text{Poisson}(\lambda)$ for any $t > \lambda$:

$$\mathbb{P}(X \geq t) \le e^{-\lambda}\left(\frac{e\lambda}{t}\right)^t$$

Rmk. using Stirling's formula we have:

$$\mathbb{P}(X = k) \simeq \frac{1}{\sqrt{2\pi k}} e^{-\lambda}\left(\frac{e\lambda}{k}\right)^k$$

So our bound on the entire tail of X has the same form as the probability of hitting one value k in the tail.

Cor. (Small deviations). For some absolute constant $c > 0$

$$\mathbb{P}(|S_N - \mu| > \delta\mu) \le 2e^{-c\mu\delta^2}$$

Cor. (Poisson distribution near the mean). Let $X \sim \text{Poisson}(\lambda)$. For $t \in (0, \lambda]$, we have

$$P(|X-\lambda| \geq t) \leq 2\exp\left(\frac{-ct^2}{\lambda}\right)$$

Rmk.

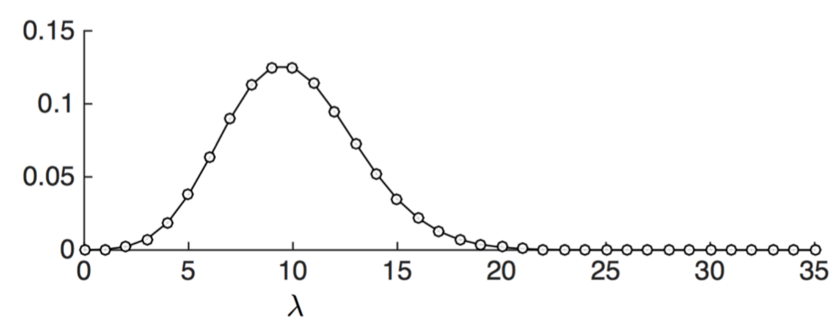


Figure 2.1 The probability mass function of the Poisson distribution $\text{Pois}(\lambda)$ with $\lambda = 10$. The distribution is approximately normal near the mean λ, but to the right from the mean the tails are heavier.

Normal Approximation to Poisson: $X \sim \text{Poisson}(\lambda)$

as $\lambda \to \infty$, we have

$$\frac{X-\lambda}{\sqrt{\lambda}} \xrightarrow{d} N(0,1)$$

Lecture 15. Random vectors in high dims.

1. Examples of concentration inequalities

Random graphs. The E-R graph $G\sim(n,p)$:
with adjacency matrix $A\in\mathbb{R}^{n\times n}$, $A_{ij}\overset{iid}{\sim}\mathrm{Bern}(p)$
. degree $d_i=\sum_{j\neq i}A_{ij}$. , $\mathbb{E}[d_i]=(n-1)p$, put $d=(n-1)p$

Proposition there's an absolute constant C,
if $d\geq C(\log n)^2$ Then $P(0.9d\leq d_i\leq 1.1d,\ \forall i)\geq 0.9$

Pf. Apply Bernstein's inequality. For any i, any $\delta>0$

$$P\left(|d-d_i|\geq\frac{C_1(\sigma+M)}{\delta}\right)\leq 2\exp\left(-\min\left(\frac{1}{\delta},\frac{1}{\delta^2}\right)\right)$$

where $\sigma=\sqrt{\mathrm{Var}(d_i)}$, $M\leq 2$ being a simple bound.

$\mathrm{Var}(d_i)=(n-1)p(1-p)\leq d$. Pick $\delta=(a\log n)^{-1}$ for see $a>1$.

$$P\left(|d-d_i|\geq\frac{C_1(\sqrt{d}+2)}{\delta}\right)\leq 2\exp\left(-\min\left(\frac{1}{\delta},\frac{1}{\delta^2}\right)\right)=2n^{-a}$$

Note that $C_1(\sqrt{d}+2)\,a\log n\leq 0.1d$ for large C

Then, by union bound, we have

$$P(\exists i,\ |d_i-d|\geq 0.1d)\leq\sum_{i=1}^{n}P\left(|d_i-d|\geq\frac{C_1(\sigma+M)}{\delta}\right)\leq 2n^{1-a}$$

Covariance matrix Let $y_1, \cdots, y_n$ iid $Y \in \mathbb{R}^d$. $\mathbb{E}[Y]=0$, with subgaussian components. $\max_{1\le j\le d} \|Y_j\|_{\psi_2} \le K$.

Then the covariance matrix of Y exists and is denoted by $\Sigma = \mathrm{Cov}(Y) \in \mathbb{R}^{d\times d}$

Empirical variance: $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T$

Denote $\|\hat{\Sigma} - \Sigma\|_{max} := \max_{j,m} |\hat{\Sigma} - \Sigma|_{j,m}$.

$$= \max_{1\le j,m\le d} \left|\frac{1}{n}\sum_{i=1}^{n}(y_{ij}y_{im} - \mathbb{E}[y_{ij}y_{im}])\right|$$

By Bernstein's inequality.

$$P([\hat{\Sigma}-\Sigma]_{j,m} \ge tK) \le 2\exp(-cn\min(t^2,t))$$

Choose $t = C\sqrt{(\log d)/n}$ with large C s.t.

$cnt^2 = cC^2\log d \ge 4\log d$

Assume n is large enough s.t. $t \le 1$, then $t^2 \le t$

$$P(|\hat{\Sigma}-\Sigma|_{jm} \ge CK\sqrt{\frac{\log d}{n}}) \le 2\exp(-4\log d) = 2d^{-4}$$

Apply union bound we have

$$P(\|\hat{\Sigma}-\Sigma\|_{max} > CK\sqrt{\frac{\log d}{n}}) \le 2d^{-2}$$

2. Two general concentration inequalities.

Thm 1. Gaussian concentration inequality.

Let $X_1, \cdots X_n \overset{iid}{\sim} N(0,1)$. Let $F: \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz function. i.e. $|F(x) - F(y)| \le L\|x-y\|$ Then

$$\| F(X_1, \cdots X_n) - \mathbb{E}[F(X_1, \cdots X_n)] \|_{\psi_2} \le CL$$

where C is a constant

Example 1. Suppose Y is a matrix of i.i.d standard normal entries. Let $\|\cdot\|$ be either Frobenius norm or matrix operator norm. Then

$$P(|\|Y\| - \mathbb{E}\|Y\|| > t) \le 2e^{-ct^2}$$

Pf. Note that $|\|X\| - \|Y\|| \le \|X - Y\|$, then apply thm 1.

Thm 2. Talagrand's inequality Suppose $X_1, \cdots X_n$ are independent random variables with $\max_i |X_i| \le K$. Denote $X = (X_1, \cdots X_n)^T$. Let $F: \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz function. Then

$$P(|F(X) - \mathbb{E}[F(X)]| \ge tL) < C\exp(-\frac{ct^2}{K^2}), \ \forall t > 0$$

Rmk. It says bounded r.v. with indpt coordinates, $F(X) - \mathbb{E}F(X)$ is a subgaussian with norm bounded by $O(KL)$

3. Norms of random vectors.

Suppose $X = (X_1, \dots X_n)^T$ with $X_1 \dots X_n$ indpt and $\mathbb{E}[X_i] = 0$, $Var(X_i) = 1$. We already have:

1. By LLN and cts mapping thm, $n^{-\frac{1}{2}}\|X\| \xrightarrow{P} 1$ as $n \to \infty$
2. By CLT and delta method, $\|X\| - \sqrt{n} \xrightarrow{d} N(0, 1/2)$
3. By Jensen's inequality, $\mathbb{E}[\|X\|] \leq \sqrt{\mathbb{E}[\|X\|^2]} = \sqrt{n}$
4. If $X_i \sim N(0,1)$ or $\max \|X_i\| \leq C$, then $\|X\| - \mathbb{E}\|X\|$ has subgaussian norm $O(1)$

Thm 3. Let $X = (X_1, \dots X_n)^T$ be a random vector with $X_1, \dots X_n$ indpt, $\mathbb{E}[X_i^2] = 1$, $\max_i \|X_i\|_{\psi_2} \leq K$. Then,

$$\big\| \|X\|_2 - \sqrt{n} \big\|_{\psi_2} \leq CK^2$$

Pf. Assume $K \geq 1$. Let $Y_i = X_i^2 - 1$, Y_i is subexponential with $\|Y_i\|_{\psi_1} \leq CK^2$. Applying Bernstein's inequality,

$$P\left(\left|\frac{1}{n}\|X\|^2 - 1\right| \geq u\right) \leq 2\exp\left(-\frac{cn}{K^4}\min(u, u^2)\right), \text{ for all } u > 0$$

Note that $|z-1| \geq \delta \Rightarrow |z^2 - 1| \geq \max(\delta, \delta^2)$, $\forall \delta > 0$.

Then we obtain $P\left(\left|\frac{1}{\sqrt{n}}\|X\| - 1\right| \geq \delta\right)$

$$\leq P\left(\left|\frac{1}{n}\|X\|^2 - 1\right| \geq \max(\delta, \delta^2)\right)$$

$$\leq 2\exp\left(-\frac{cn}{K^4}\delta^2\right)$$

Changing variables to $t = \delta\sqrt{n}$, we therefore obtain

$$P\left(\left|\|X\| - \sqrt{n}\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{K^4}\right), \ \forall t > 0$$

Rmk. High-dim Gaussian is very close to uniform distribution on the sphere of radius $\sqrt{n}$.

Uniform spherical random variable

rS^{n-1}: sphere in $\mathbb{R}^d$ with radius r.

Thm3: $N(0, I_n)$ is similar to $\mathrm{Unif}(\sqrt{n}\, S^{n-1})$ in high dimensions.

If $G \sim N(0, I_n)$, then

$$G = UR \quad \text{where } U = G/\|G\|,\ R = \|G\|.$$

satisfies $U \sim \mathrm{Uniform}(S^{n-1})$, U and R are independent.

Thm4. Uniform spherical distribution is subgaussian.

$X \sim \mathrm{Unif}(\sqrt{n}\, S^{n-1})$ satisfies

$$\sup_{u \in \mathbb{R}^n, \|u\|=1} \| \langle X, u \rangle \|_{\psi_2} \le C$$

Namely, projecting X using any unit vector gives a subgaussian random variable with a bounded norm

4. Random Projection.

A large collection high-dim points $Q := \{x_i\}_{i \in N} \subset \mathbb{R}^d$.

Goal: Dimensionality Reduction.

Thm 5. Johnson–Lindenstrauss Lemma Let $\varepsilon \in (0, \frac{1}{2})$. There exists $C > 0$ s.t. Let $k \geq (C \log N)/\varepsilon^2$ be an integer. Then $\exists f \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^k)$ s.t.

$$(1-\varepsilon)\|x-y\|^2 \leq \|f(x) - f(y)\|^2 \leq (1+\varepsilon)\|x-y\|^2$$

for all $x, y \in Q$.

Pf.

Lecture 16. Norms of Random Matrices

1. Linear Algebra Prep

Def (ε-net). Let (T, d) be a metric space, then consider a subset $K \subset T$ and some $\varepsilon > 0$, then a subset $\mathcal{N} \subset K$ is called a ε-net of K if $\forall x \in K$, $\exists x_0 \in \mathcal{N}$ s.t. $d(x_0, x) < \varepsilon$.

Def. Covering Number the number of points in the smallest ε-net of K, denoted by $\mathcal{N}(K, d, \varepsilon)$

Prop 1. Consider unit ball $B_2^n = \{x \in \mathbb{R}^n \mid \|x\|_2 \le 1\} \subset \mathbb{R}^n$

$$\left(\frac{1}{\varepsilon}\right)^n \le \mathcal{N}(B_2^n, d_{L_2}, \varepsilon) \le \left(1 + \frac{2}{\varepsilon}\right)^n$$

Lemma 1. $\forall \varepsilon \in (0,1)$, $\mathcal{N}$ is an ε-net of S^{n-1}, then

$$\sup_{x \in \mathcal{N}} \|Ax\|_2 \le \|A\|_{op} \le \frac{1}{1-\varepsilon} \sup_{x \in \mathcal{N}} \|Ax\|_2$$

If in addition there's a ε-net $\mathcal{M}$ of S^{m-1}, then

$$\sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle \le \|A\|_{op} \le \frac{1}{1-2\varepsilon} \sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle$$

Pf. $\forall x \in S^{n-1}$, $\exists x_0 \in \mathcal{N}$ s.t. $\|x - x_0\| < \varepsilon$. Let x be the top eigenvector of A^TA, then

$$\|Ax - Ax_0\| \le \|A\| \|x - x_0\| \le \varepsilon \|A\|.$$

$$\|Ax_0\| \geq \|Ax\| - \|Ax - Ax_0\| \geq \|A\| - \varepsilon\|A\|$$

$$\|A\| \leq \frac{1}{1-\varepsilon}\|Ax_0\|$$

Rmk $\|A\|$ is approximately the same on a discretized space.

2 Operator Norm of Subgaussian Random Matrix.

Thm 1. Let $A \in \mathbb{R}^{m\times n}$ have independent subgaussian entries (A_{ij} jointly indpt. subgaussian with $\mathbb{E}A_{ij} = 0$)

Then $\|A\| \leq CK(\sqrt{m} + \sqrt{n} + t)$ with probability at least $1-2e^{-t^2}$ for certain constant C.

and $K = \max_{ij} \|A_{ij}\|_{\psi_2}$

Lemma 2. Let $X_1, \cdots X_n$ be independent subgaussian random variables with $\mathbb{E}X_i = 0$, then $\sum_{i=1}^n X_i$ is also subgaussian with

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_2}^2 \leq C\sum_{i=1}^n \|X_i\|_{\psi_2}^2$$

Rmk $\|A\|$ is roughly of order $\sqrt{m} + \sqrt{n}$ with high probability. Assume $n \geq m$, $\sigma^2 = \mathbb{E}A_{ij}^2$

$$\|A\|_F^2 = \sum_{ij} A_{ij}^2 \xrightarrow{P} nm\sigma^2$$

Since $\|A\|_F^2 = \sum_{j=1}^m [\sigma_j(A)]^2$, we have $\frac{1}{m}\sum_{k=1}^m \sigma_k^2(A) \xrightarrow{P} n\sigma^2$

The averaged singular value is of order $\sqrt{n}\,\sigma$.

with high prob $\Rightarrow$ so does the largest singular value.

3. Covariance Estimation

Suppose $X_1 \cdots X_n \in \mathbb{R}^p$ iid random vectors with $\mathbb{E}X_0 = 0$ and $\mathrm{Cov}(X_0) = \Sigma$ exists. By LLN, $\frac{1}{n}\sum_{i=1}^n X_i X_i^T \xrightarrow{P} \Sigma$

Thm 2. Assume $X_0 \in \mathbb{R}^p$ is a subgaussian random vector, ($\sup_{u \in S^{p-1}} \|\langle X_0, u\rangle\|_{\psi_2} \le K$.) Suppose $X_1, \cdots X_n$ are iid copies of X_0, $X = (X_1, \cdots X_n)^T \in \mathbb{R}^{n\times p}$

Then for every $t \ge 0$, with prob at least $1 - 2e^{-C_K t^2}$

$\|\frac{1}{n}X^TX - \Sigma\| \le \max\{\delta, \delta^2\}\|\Sigma\|$ where $\delta = C_K\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}$

Thm 2'. Covariance Estimation (HDP).

Let $X_0 \in \mathbb{R}^p$ subgaussian ($\|\langle X, x\rangle\|_{\psi_2} \le K\|\langle X, x\rangle\|_{L^2}$ for some $K \ge 1$, any $x \in \mathbb{R}^p$). Then

$$\mathbb{E}[\|\frac{1}{n}X^TX - \Sigma\|] \le CK^2(\sqrt{\frac{p}{n}} + \frac{p}{n})\|\Sigma\|.$$

Example: Subgaussian Random Vectors.

- $X \sim N(0, I_n)$
- $X \sim \mathrm{Unif}(\{\pm1\}^n)$.
- $X \sim \mathrm{Unif}(\sqrt{n}\, S^{n-1})$

Lecture 17. High-dim Statistical Phenomena

1. Principle Component Analysis.

Suppose $X_1, \dots X_n \overset{iid}{\sim} X_0$, $\mathbb{E}[X_0]=0$. $\Sigma = \mathrm{Cov}(X_0)$ exists.
Let $\mu_1 \cdots \mu_p$ be orthonormal eigenvectors of Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
Then consider the empirical covariance $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$ with orthonormal eigenvectors $\hat{\mu}_1, \cdots \hat{\mu}_p$ and eigenvalues $\hat\lambda_1 \geq \hat\lambda_2 \geq \cdots \geq \hat\lambda_p \geq 0$.

1.1. Fixed Dimension

Thm 1 Consistency of PCA in fixed dimension.

Suppose $\lambda_1 > \lambda_2$. Then as $n \to \infty$, we have

$$\max_{k\leq p} |\hat\lambda_k - \lambda_k| \xrightarrow{P} 0 \qquad \min_{s\in\{\pm1\}} \|s\hat{u}_1 - u_1\|_2 \xrightarrow{P} 0$$

Rmk: $\lambda_1 > \lambda_2$ ensures λ_1 has multiplicity of 1.

Pf By LLN we have $\hat\Sigma \xrightarrow{P} \Sigma$

By Weyl's inequality

$$\max_{k\leq p} |\hat\lambda_k - \lambda_k| \leq \|\hat\Sigma - \Sigma\|_{op}$$

Since $\|\hat\Sigma - \Sigma\|_{op} \to 0$, $\max_{k\leq p} |\hat\lambda_k - \lambda_k| \to 0$

Thm 2 Davis-Kahan Sin-Θ Theorem (Simplified)

Let $u_1, \tilde{u}_1$ be the first normalized eigenvectors of 2 positive-definite matrices $\Sigma, \tilde{\Sigma} \in \mathbb{R}^{p\times p}$.

Assume $\lambda_1 > \lambda_2$, then

$$\min_{s\in\{\pm1\}} \| s\tilde{u}_1 - u_1\|_2 \le \frac{2\sqrt{2}\,\|\tilde{\Sigma}-\Sigma\|_{op}}{\underbrace{\lambda_1-\lambda_2}_{\text{eigen gap}}}$$

Pf. Suppose $\lambda_1 - \lambda_2 < 2\|\tilde{\Sigma}-\Sigma\|$, then we have

$$LHS \le \|2u_1\|_2 = \sqrt{2} \le \sqrt{2}\cdot\frac{2\|\tilde{\Sigma}-\Sigma\|_{op}}{\lambda_1-\lambda_2} = RHS$$

If $\lambda_1 - \lambda_2 > 2\|\tilde{\Sigma}-\Sigma\|_{op}$.

Pick $s_0 = 1$ if $\langle \tilde{u}_1, u_1\rangle > 0$, $s_0 = -1$ if $\langle \tilde{u}_1, u_1\rangle < 0$

$$\min_{s\in\{\pm1\}} \|s\tilde{u}_1 - u_1\|_2 \le \|s_0\tilde{u}_1 - u_1\|_2 = \sqrt{2-2\langle s_0\tilde{u}_1, u_1\rangle}$$

$$\le \sqrt{2-2\langle s_0\tilde{u}_1, u_1\rangle}\sqrt{1+\langle s_0\tilde{u}_1, u_1\rangle}$$

$$= \sqrt{2}\sqrt{1-\langle s_0\tilde{u}_1, u_1\rangle^2}$$

$$= \sqrt{2}\sqrt{1-\langle \hat{u}_1, u_1\rangle^2}$$

Since $u_1, \cdots u_p$ is an orthonormal basis

$$\langle u_1, \tilde{u}_1\rangle^2 + \sum_{k\ge2}^{p}\langle u_k, \tilde{u}_1\rangle^2 = 1.$$

It suffice to show $\sum_{k\ge2}^{p}\langle u_k, \tilde{u}_1\rangle^2 \le \frac{4\|\tilde{\Sigma}-\Sigma\|_{op}^2}{(\lambda_1-\lambda_2)^2}$

$$\langle u_k, \tilde{\Sigma}\tilde{u}_1\rangle = \tilde{\lambda}_1\langle u_k, \tilde{u}_1\rangle$$

$$\langle \tilde{u}_1, \Sigma u_k\rangle = \lambda_k\langle u_k, \tilde{u}_1\rangle$$

$$\Rightarrow \langle u_k\ (\tilde{\Sigma}-\Sigma)\,\tilde{u}_1\rangle = (\tilde{\lambda}_1 - \lambda_k)\langle u_k, \tilde{u}_1\rangle$$

$$\tilde{\lambda}_1 - \lambda_k \geq \tilde{\lambda}_1 - \lambda_2 \geq \lambda_1 - \lambda_2 - \|\hat{\Sigma} - \Sigma\|_{op} \geq \frac{1}{2}(\lambda_1 - \lambda_2)$$

where $\lambda_1 - \lambda_2 \geq 2\|\hat{\Sigma} - \Sigma\|_{op}$. Thus we have

$$\sum_{k>1} \langle u_k, \tilde{u}_k \rangle^2 \leq \sum_{k>0} \frac{\langle u_k, (\hat{\Sigma} - \Sigma)\hat{u}_1 \rangle^2}{(\tilde{\lambda}_1 - \lambda_k)^2} \leq \frac{4\|(\hat{\Sigma} - \Sigma)\tilde{u}_1\|_2^2}{(\lambda_1 - \lambda_2)^2}$$

$$\leq \frac{4\|\hat{\Sigma} - \Sigma\|_{op}^2}{(\lambda_1 - \lambda_2)^2}$$

Davis-Kahan Theorem (HDP) Let S and T be symmetric matrices with the same dimensions. Assume

$$\min_{j \neq i} |\lambda_i(S) - \lambda_j(S)| = \delta > 0$$

Then the angle $\sin(v_i(S), v_i(T)) \leq \frac{2\|S - T\|}{\delta}$

where $v_i(\ )$ gives the i-th largest eigenvector.

1.2. Growing Dimension.

p_n, Σ_n depend on n. $X_0 \sim N(0, \Sigma_n)$

Spiked model: $\Sigma_n = (\lambda_1 - \lambda_2) u_1 u_1^T + \lambda_2 I_n$

eigenvalues are λ_1 and λ_2 with multiplicity $p-1$

Note that $\frac{1}{2} \min_{s \in \{\pm 1\}} \|s\hat{u}_1 - u_1\|_2^2 = 1 - |\langle \tilde{u}_1, u_1 \rangle|$

Inconsistency of PCA in high dimensions:

Assume $\|\Sigma_n\|_{op} \leq C$ and $\lambda_1(\Sigma_n) - \lambda_2(\Sigma_n) \geq k > 0$

holds for certain constant C and k

Suppose $\frac{p_n}{n} \to \gamma \in [0, +\infty)$

- If $\gamma = 0$, then there exists a constant $C_0 > 0$ such that $\min_{s \in \{\pm 1\}} \|s\hat{u}_1 - u_1\|_2 \xrightarrow{P} 0$ as $n \to \infty$

- If $\gamma > 0$, under spiked model assumption we have $|\langle \hat{u}_1, u_1 \rangle| \xrightarrow{P} \eta_\gamma$, where $\eta_\gamma < 1$.
  Moreover, $\exists \gamma^*$, if $\gamma > \gamma^*$, $|\langle \hat{u}_1, u_1 \rangle| \xrightarrow{P} 0$

Rmk. population and sample principle direction are almost orthogonal in high dims.

2. Linear Regressions in High Dims.

Consider (x_i, y_i), $x_i \in \mathbb{R}^p$ and $i = 1, 2, \ldots, n$.

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, 2, \ldots n.$$

$X = (x_1, \ldots x_n)^T \in \mathbb{R}^{n \times p}$ $\quad y = (y_1, \ldots y_n)^T \in \mathbb{R}^n$

$\varepsilon = (\varepsilon_1, \ldots \varepsilon_n)^T \in \mathbb{R}^n$, $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$

- LSE: $\hat{\beta} = (X^T X)^{-1} X^T y$ $\quad$ • Prediction: $\hat{f}(x_0) = x_0^T \hat{\beta}$

Under Parametrized Regime

Proposition 1. Suppose $x_0, x_1, \cdots x_n \overset{iid}{\sim} N(0, I_p)$. Assume $p = p_n$, $\frac{p_n}{n} = o(1)$ as $n \to \infty$, we have.

$$\mathbb{E}[\|\hat\beta-\beta\|^2|X] = (1+o(1))\frac{\sigma^2 p_n}{n}$$

$$\mathbb{E}[(x_0^T\hat\beta - x_0^T\beta)^2|X] = (1+o_p(1))\frac{\sigma^2 p_n}{n} \quad \text{as } n\to\infty$$

Rk. the MSE increases approximately linear in p_n

Overparameterized Regime:

Minimum-Norm Interpolator:

$$\min_{\beta\in\mathbb{R}^p} \|\beta\|^2$$

$$\text{subject to } y = X\beta$$

Assume rank$(X)=n$. The solution is $\underbrace{\hat\beta = X^T(XX^T)^{-1}y}_{}$ unbiased

$\forall \beta^*$ st $X\beta^*=y$, $\exists\, v\in N(X)$ s.t. $\beta^* = \hat\beta + v$.

Note that $\hat\beta\in C(X^T)$ and $C(X^T)\perp N(X)$

we have $\|\beta^*\|^2 = \|\hat\beta\|^2 + \|v\|^2 \geq \|\hat\beta\|^2$

Proposition 2. Suppose $x_1,\dots,x_n \sim N(0, I_p)$ and $p=p_n$ depends on n and $n/p_n = o(1)$ as $n\to\infty$. We have rank$(X)=n$ with probability $1-o(1)$ and

$$\mathbb{E}[\|\hat\beta-\beta\|^2|X] = (1+o_p(1))\cdot\left[(1-\frac{n}{p_n})\|\beta\|^2 + \frac{\sigma^2 n}{p_n}\right]$$

$$\mathbb{E}[(x_0^T\hat\beta - x_0^T\beta)^2|X] = (1+o_p(1))\cdot\Big[\underbrace{(1-\frac{n}{p_n})\|\beta\|^2}_{\text{Bias}} + \underbrace{\frac{\sigma^2 n}{p_n}}_{\text{Variance: Smaller when } p_n\gg n}\Big]$$

3. Community Detection

$z^* \in \{\pm 1\}^n$: unknown membership

Stochastic Block Model. Let $A \in \mathbb{R}^{n\times n}$ be the adjacency matrix, $P(A_{ij}=1) = \begin{cases} p & \text{if } z_i^* = z_j^* \\ q & \text{if } z_i^* \neq z_j^* \end{cases}$

with $p > q$. A_{ij} jointly indpt

Assume each community has equal size: $1^T z^* = 0$

Spectral Method.

Eigen Structure of A:

$$A^* = \mathbb{E}[A] = \frac{p-q}{2} z^* z^{*T} + \frac{p+q}{2} 1_n 1_n^T$$

It has only 2 positive eigenvalues:

$\lambda_1^* = \frac{(p+q)n}{2}$ and $\lambda_2^* = \frac{(p-q)n}{2}$ with corresponding eigenvectors $u_1^* = 1_n/\sqrt{n}$ and $u_2^* = z^*/\sqrt{n}$

Spectral Clustering.

1. Calculate the second eigenvector $u = u_2$ of A

   If p, q are known, use $A - \frac{p+q}{2} 1_n 1_n^T$

2. Estimate $z_i = 1$ if $u \geq 0$, $z_i = -1$ if $u < 0$

3.1. Theoretical guarantee of spectral method.

Thm 5. For any $\varepsilon > 0$, we have

$$P\left(\min_{s \in \{\pm 1\}} \frac{1}{\sqrt{n}} \|sz - z^*\|_2 > \varepsilon\right) = o(1)$$

Rmk It says the mismatch ratio is arbitrarily small if $n \to \infty$

Lecture 18. Stein's phenomenon and shrinkage estimation.

1. Motivation.

1.1. High-D Perspective

Let $X = \mu + Z$, $Z \sim N_p(0, I_p)$

When p is large:

$$\|X\|^2 = \|\mu\|^2 + \|Z\|^2 + 2\langle \mu, Z\rangle \approx \|\mu\|^2 + p$$

$\|X\| \gg \|\mu\|$ with high probability

Hoeffding:

$$P(|\langle \mu, Z\rangle| > t) \le 2\exp\left(-c\frac{t^2}{\|\mu\|^2}\right)$$

if $\|\mu\| = O(1) \Rightarrow \langle \mu, Z\rangle = O_p(1)$

1.2. Bias-Variance trade-off perspective

$$R_T(\mu) := \mathbb{E}\|T(X) - \mu\|^2 \quad X \sim N(\mu, I_p).$$

Consider $\hat{\mu} = (1-\varepsilon)X$

$$\begin{aligned}\mathbb{E}\|\hat{\mu} - \mu\|^2 &= \varepsilon^2\|\mu\|^2 + (1-\varepsilon)^2 p \\ &= (1-2\varepsilon)p + \varepsilon^2(\|\mu\|^2 + p) \\ &= (1-2\varepsilon)p + O(\varepsilon^2)\end{aligned}$$

Goal: $\min\limits_{\varepsilon \in [0,1]} \{\varepsilon\|\mu\|^2 + (1-\varepsilon)^2 p\}$.

$\|\mu\|$ is unknown, but we can use $\|X\|^2 \approx \|\mu\|^2 + p$ when p is large to estimate $\|\mu\|^2$

2. Shrinkage Estimators

signal $\mu \uparrow$
noise $p \downarrow$ $\Rightarrow \varepsilon \downarrow$

James-Stein Estimator:

$$\hat{\mu}_{JS} = \left(1 - \frac{p-2}{\|X\|^2}\right) X \qquad \hat{\mu}_{JS+} = \left(1 - \frac{p-2}{\|X\|^2}\right)_+ X$$

Rmk. Shrinkage to 0 is not the only direction that works. For any $c \in \mathbb{R}^p$, we can make

$$\hat{\mu}_{JS} = \left(1 - \frac{\gamma(p-2)}{\|X-c\|^2}\right)(X-c) + c$$

$$\hat{\mu}_{JS+} = \left(1 - \frac{\gamma(p-2)}{\|X-c\|^2}\right)_+ (X-c) + c$$

Theorem 1. If $\hat{\mu}$ is any of the two shrinkage estimators then for $\gamma \in (0,2)$, when $p \geq 3$

$$R_{\hat{\mu}}(\mu) < R_X(\mu) \text{ for every } \mu \in \mathbb{R}^p$$

Moreover, for $\hat{\mu} = \hat{\mu}_{JS}$, we have

$$R_{\hat{\mu}}(\mu) = p - (2\gamma - \gamma^2)(p-2)^2 \mathbb{E}[\|X-c\|^{-2}]$$

Thm 2. Stein's Lemma. If $g \in C^1(\mathbb{R}^p)$ with $\mathbb{E}[\|\nabla g(Z)\|] < \infty$ where $Z \sim N(0, I_p)$, then

$$\mathbb{E}[\nabla g(Z)] = \mathbb{E}[Z g(Z)]$$

Pf. $\mathbb{E}[\partial_i g(Z)] = \int \partial_i g(z) \frac{1}{\sqrt{2\pi}} e^{-\|z\|^2/2} dz$

(integration by parts) $= -\int g(z) \frac{1}{\sqrt{2\pi}} \partial_i e^{-\|z\|^2/2}$

$$= \int g(z) \frac{1}{\sqrt{2\pi}} z_i e^{-\|z\|^2/2}$$

$$= \mathbb{E}[Z g(Z)]$$

Rmk General case: $g \in C^1(\mathbb{R}^p, \mathbb{R}^p)$, then

$$\mathbb{E}[\nabla \cdot g(Z)] = \mathbb{E}[Z^T g(Z)]$$

3. Tweedie's Formula.

3.1. Bayesian Perspective.

Suppose μ has a prior distribution $\mu \sim \pi$

$X|\mu \sim N(\mu, I_p)$. The marginal density of X is:

$$f_x(x) = \int_{\mathbb{R}^p} f_{X,\mu}(x, \mu) d\mu = \frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \exp(-\|x-\mu\|^2/2) \pi(\mu) d\mu.$$

Consider $\phi: \mathbb{R}^p \to \mathbb{R}^p$

$$MSE = \mathbb{E}\|\mu - \phi(X)\|^2 = \mathbb{E}_X[\mathbb{E}_{\mu|X}[\|\mu - \phi(X)\|^2]]$$

$$\mathbb{E}_{\mu|X=x}[\|\mu - \phi(x)\|^2] \geq \mathbb{E}_{\mu|X=x}[\|\mu - \mathbb{E}_{\mu|X=x}[\mu]\|^2], \ \forall x$$

We only need to choose $\phi(x) = \mathbb{E}[\mu | X=x]$

$$\mathbb{E}[\mu|X=x] = \frac{\int \mu f_{X|\mu}(x|\mu) \pi(\mu) d\mu}{\int f_{X|\mu}(x|\mu) \pi(\mu) d\mu}$$

$$\mathbb{E}[\mu - x | X=x] = \frac{\int (\mu - x) e^{-\|x-\mu\|^2/2} \pi(\mu) d\mu}{\int e^{-\|x-\mu\|^2/2} \pi(\mu) d\mu}$$

$$= \frac{\nabla f_X(x)}{f_X(x)} = \nabla \log f_X(x)$$

$$\psi_\pi(X) = \mathbb{E}[\mu \mid X] = X + \nabla \log f_X(X)$$

Rmk. From this perspective, shrinkage is a prior on the unknown π.

3.2. Shrinkage under superharmonic function

Def. For $f \in C^2(\mathbb{R}^p)$, denote the Laplace operator by $\Delta f(x) = \sum_{j=1}^p \partial_j^2 f(x)$. We say f is superharmonic iff $\Delta f(x) \leq 0$, $\forall x \in \mathbb{R}^p$

Thm 3. Let $f \in C^2(\mathbb{R}^p)$, $f > 0$. If f is superharmonic iff $\mathbb{E}[\frac{1}{f(X)} \sum_{j=1}^p |\partial_j^2 f(X)|] < \infty$, $\mathbb{E}[\|\nabla \log f(X)\|^2] < \infty$ we have $\mathbb{E}[\|X + \nabla \log f(X) - \mu\|^2] = p + 4\mathbb{E}[\frac{\Delta\sqrt{f(X)}}{\sqrt{f(X)}}] \leq p$

Rmk. X is a minimax estimator

$\Rightarrow X + \nabla \log f(X)$ is also a minimax estimator

4. Diffusion Model:

Forward Process: $X_t = \sqrt{1-\sigma_t^2}\, X_{t-1} + \sigma_t Z_t$

$Z_t \overset{iid}{\sim} N(0, I)$.

Score Function p_t: pdf of X_t

$$\mathbb{E}[X_{t-1} \mid X_t = x_t] = \frac{1}{\sqrt{1-\sigma_t^2}} x_t + \frac{\sigma_t^2}{\sqrt{1-\sigma_t^2}} \underbrace{\nabla \log p_t(x_t)}_{\text{Score function}}$$

$$X_{t-1} \mid X_t = x_t \sim N\left(\frac{1}{\sqrt{1-\sigma_t^2}} x_t + \frac{\sigma_t^2}{\sqrt{1-\sigma_t^2}} \nabla \log p_t(x_t),\ \sigma_t^2 I\right)$$

Reconstruction: $X_{t-1} = \frac{1}{\sqrt{1-\sigma_t^2}} X_t + \frac{\sigma_t^2}{\sqrt{1-\sigma_t^2}} S(\theta, X_t) + \sigma_t Z_t$

$S(\theta, x_t)$ is trained to approximate $\nabla \log p_t(x_t)$:

$$\min_\theta \mathbb{E}_{t \sim \mathrm{Unif}(0,1),\, X_0 \sim p_0,\, X_t \sim p(x_t \mid x_0)} \left[\sigma_t \|\nabla \log p_t(X_t) - S_\theta(X_t, t)\|^2\right]$$

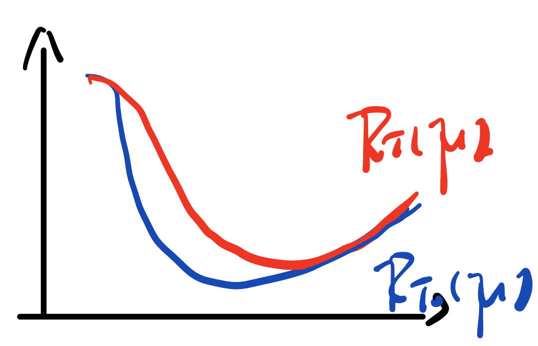
Recall:

Admissibility:

$T(x)$ is admissible if there's no T_0 s.t.

① $R_{T_0}(\mu) \le R_T(\mu)$ for all μ

② $R_{T_0}(\mu) < R_T(\mu)$ for some μ



Minimaxity: $T(x)$ is minimax if for any $T_0(x)$

$$\sup_\mu R_T(\mu) \le \sup_\mu R_{T_0}(\mu)$$

Setting $X \sim N(\mu, I_p)$, estimate μ

① Is X minimax? Yes.

The supremum of risk is controled even if the risk is point-wise smaller than J-S estimator

② Is X admissible? $\begin{cases} \text{Yes when } p = 1, 2 \\ \text{NO when } p \geq 3 \end{cases}$

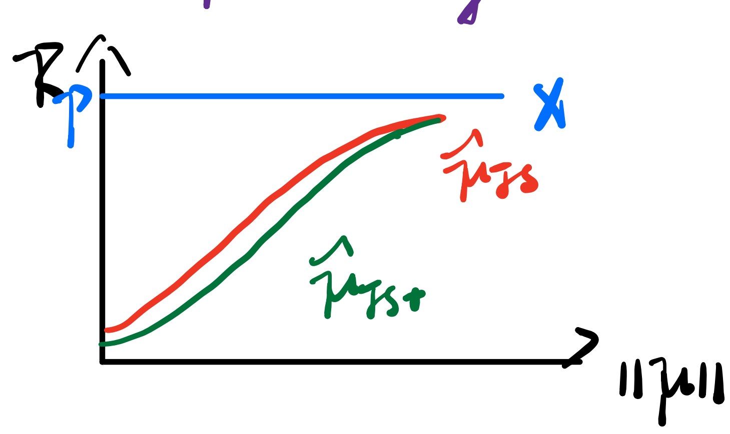
Consider $r = 1$. For any $p \geq 3$:

$R_{\hat{\mu}_{JS}}(\mu) = p - (p-2)^2 \underbrace{\mathbb{E}[\|X\|^{-2}]} < R_X(\mu) = p$

Two extreme case: $\|X\|^2 \sim \chi_p$ $\quad \mathbb{E}[\|X\|^{-2}] = \frac{1}{p-2}$.

$R_{\hat{\mu}_{JS}}(0) = p - (p-2)^2 \frac{1}{p-2}$

Reason why we need $p \geq 3$



$R_{\hat{\mu}_{JS}}(\infty) = p$

$\cdot R_{\hat{\mu}_{JS+}} > R_{\hat{\mu}_{JS}}$

$\Rightarrow \hat{\mu}_{JS}$ is inadmissible

Remark. $\hat{\mu}_{JS+}$ is also inadmissible

Empirical Bayes Perspective:

$X \sim N(\mu, I_p) \quad \mu \sim N(0, \tau^2 I_p)$ $\;\tau^2 \to$ hyperpara

Bayes estimator. $\delta(X) = \underset{\delta(X)}{\arg\min}\, \mathbb{E}[R_\tau(\mu)] = (1 - \frac{1}{1+\tau^2}) X$

$p(x|\tau) = N(0, (1+\tau^2) I_p)$

Empirical Bayes: estimate τ from data using $p(x|\tau)$

- UMVUE of $\frac{1}{1+\tau^2}$: $\frac{p-2}{\|x\|^2}$

  plug in: $\hat{\delta}(x) = (1 - \frac{p-2}{\|x\|^2}) X = \hat{\mu}_{JS}$

- MLE of $1+\tau^2$ is $\frac{\|x\|^2}{p}$

  plug in: $\hat{\delta}(x) = (1 - \frac{p}{\|x\|^2})_+ X \approx \hat{\mu}_{JS+}$

Lecture 19. Denoising via Shrinkage

1. Sparsity in data Analysis

1.2 Quantifying sparsity.

Def. we say $x \in \mathbb{R}^n$ is s-sparse if

$$\|x\|_0 := \sum_{j=1}^{n} \mathbb{1}(x_j \neq 0) \leq s$$

We can also relax the definition using $\|x\|_q$, $q \in (0, 2]$

Rmk. if $q < 1$. $\|x\|_q$ is only a "quasinorm". (Not a norm)

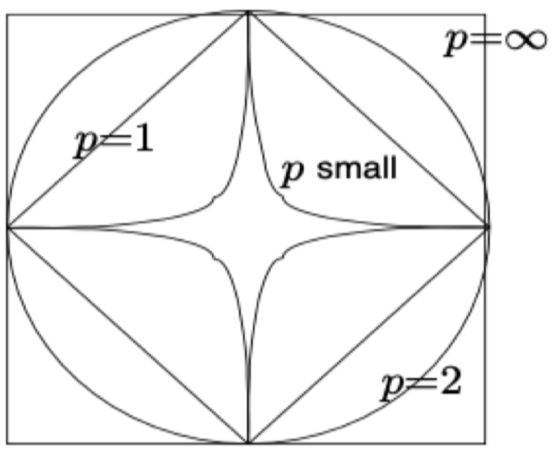


Figure 1.5 Contours of ℓ_p balls

3. Thresholding.

3.1. Gaussian sequence model.

Assume our observations follows $Y_k = \theta_k + \sigma Z_k$, $Z_k \overset{iid}{\sim} N(0,1)$

Raw measurements $X \in \mathbb{R}^n$

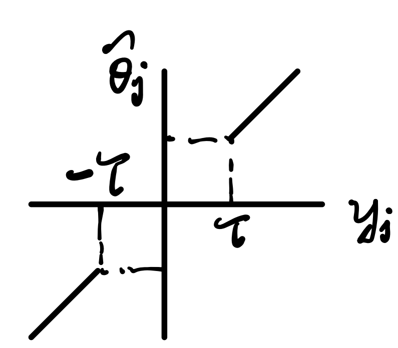
$$X_k = f(t_k) + \sigma W_k, \quad W_k \overset{iid}{\sim} N(0,1)$$

Apply orthogonal transform $Y = WX$ where W is an orthogonal matrix. $\theta = W\bar{f}(t)$, $Y \sim N(\theta_k, \sigma^2 I_n)$

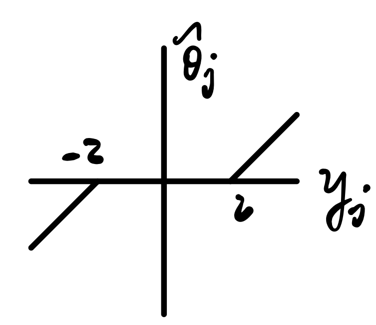
Sparsity Assumption: $\|\theta\|_0 = \sum_{j=1}^{n} I(\theta_j \neq 0) \le S$

unbiased estimator: Y. J-S estimator: $\left(1 - \frac{(p-2)\sigma^2}{\|Y\|^2}\right) Y$.

3.2. Thresholding Functions.



Hard Thresholding: $\hat{\theta}_j = \begin{cases} Y_j, & \text{if } |Y_j| > \tau \\ 0 & \text{if } |Y_j| < \tau \end{cases}$



Soft Thresholding: $T_\tau(u) = \mathrm{sign}(u)(|u| - \tau)_+$

Determining τ: $\|Z\|_\infty \approx \sqrt{2\sigma^2 \log n}$

we may choose $\tau = \sqrt{2\sigma^2 \log n}$

Further, by Gaussian Concentration inequality. $\max Z_i$ is close to $\mathbb{E}[\max_i Z_i]$, and

$$\sqrt{2\log n} - 1 - \sqrt{2\log 2} \le \frac{1}{\sigma}\mathbb{E}\max_i Z_i \le \sqrt{2\log n}$$

So choosing $\tau = (1+0.01)\sqrt{2\sigma^2 \log n}$ is guaranteed to remove all noise with high probability.

4. Denoising theory:

Oracle-aided ideal risk. $R(\theta, \sigma) = \sum_{i=1}^{n} \min\{\theta_i^2, \sigma^2\}$

If θ is S-sparse: $R(\theta, \sigma^2) = S\sigma^2$

Thm 1. Suppose $Y \sim N(\theta, \sigma^2 I_n)$. For soft-thresholding estimator $\hat{\theta}_{\tau_n} = T_{\tau_n}(Y)$ with the level $\tau_n = \sqrt{2\sigma \log n}$,

$$\mathbb{E}[\|\hat{\theta}_{\tau_n} - \theta\|_2^2] \leq (2\log n + 1)[\sigma^2 + R(\theta, \sigma)]$$

Rmk. We don't know the indices of the large signals. The risk is increased by a log factor compared to the ideal risk

4.2 Oracle Inequalities for S-sparse Signals

Theorem 2. *Suppose $S = S_n$ satisfies*

$$\frac{S_n}{n} \to 0, \quad n \to \infty.$$

Then, as $n \to \infty$,

(1) The soft-thresholding estimator $\widehat{\theta}_{\tau_n}$ with $\tau_n = \sqrt{2\sigma^2 \log(n/S_n)}$,

$$\sup_{\theta \in \Theta_{n,s}} \mathbb{E}\big[\|\widehat{\theta}_{\tau_n} - \theta\|^2\big] \leq (1 + o(1)) \cdot 2S_n\sigma^2 \log\left(\frac{n}{S_n}\right).$$

(2) The soft-thresholding estimator matches the minimax lower bound:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta_{n,s}} \mathbb{E}\big[\|\widehat{\theta}_{\tau_n} - \theta\|^2\big] \geq (1 + o(1)) \cdot 2S_n\sigma^2 \log\left(\frac{n}{S_n}\right).$$

Final: 3 problems with subproblems, 16 subproblems in total

Most: No more than 3 lines of proofs

Lecture 20: Basis pursuit, compressed sensing

1. Basis Pursuit

identifiable?

Find sparse solution θ^* to $Y = X\theta^*$, $X \in \mathbb{R}^{n\times p}$, $n<p$

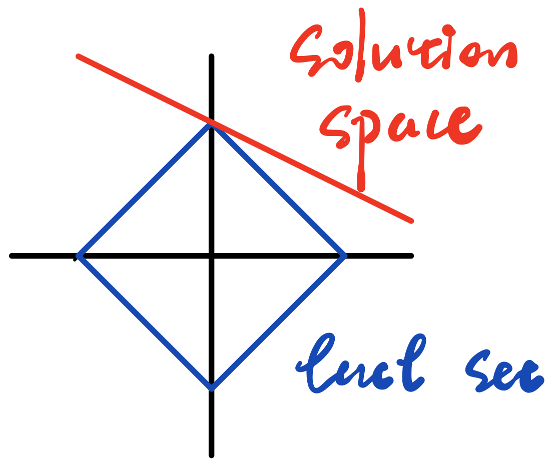
$$\min_{\theta\in\mathbb{R}^p} \|\theta\|_1$$

affine linear subspace

subject to $y = X\theta \iff \theta \in \theta^* + \ker(X)$

1.1. Restricted null space property.

Consider the level set $S_r := \{\theta : \|\theta\|_1 = r\}$.



Rmk High-dim ℓ_1 balls are sharp.

In $\mathbb{R}^p$, the probability of failing to eliminate some parameters is $\frac{1}{2^{p-1}}$

RNS: Formalizing geometric intuitions.

Let $S = \{j \le p : \theta_j^* \neq 0\}$.

$$C(S) = \{\Delta \in \mathbb{R}^p : \|\Delta_{S^c}\|_1 \le \|\Delta_S\|_1\}.$$

where $\Delta_S = (\Delta_j \mathbb{1}(j\in S))_{j\le p}$

We say X satisfies restricted null space (RNS) if $\ker(X) \cap C(S) = \{0\}$.

Proposition 1. X satisfies RNS, then BP recovers θ^*

1.2. Pairwise incoherence

Def Pairwise incoherence. Let $X = (X_1, \dots X_p) \in \mathbb{R}^{n \times p}$

The pairwise incoherence is:

$$\delta_{pw}(X) = \| \frac{1}{n} X^T X - I_p \|_{max} = \max_{k,j} | \frac{1}{n} \langle X_k, X_j \rangle - 1(k=j) |$$

Proposition. Suppose $X = (X_{kj})_{k \le n, j \le p}$, where $X_{kj} \overset{iid}{\sim} N(0,1)$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\delta_{pw}(X) \le C \frac{\sqrt{\log p} + \sqrt{\log(1/\delta)}}{\sqrt{n}}$$

Rmk. Given exponentially many random normal vectors, all pairs of vectors are nearly orthogonal

Thm 1. If X has pair-wise incoherence, $\delta_{pw}(X) < \frac{1}{2k}$, and $|S| \le k$, then X satisfies RNS property.

Cor. Assume $X_{kj} \overset{iid}{\sim} N(0,1)$ and θ^* is k-sparse. Then BP recovers θ^* with high probability if

$$n \ge C k^2 \log p$$

where C is an absolute constant