Lecture 6. Statistical Decision Theory and Efficiency

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 $G(n,p)$:

Graph G, n vertices, A: adjacency matrix.

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

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

Classifying Statistical Models

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

eg. 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

Rmk. Statistical Methods often use a finite number of estimates to approximate the unknow 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:

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

$(\beta_1, \ldots \beta_K)^T$ and $(\beta_1 + C, \ldots \beta_K + C)$ septulate 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 nice function T, we call $T(X)$ a statistic

$\sigma(T(X)) \subset \sigma(X)$: $T(X)$ summarizes information for 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 $f_\theta(\omega) = \exp\{[\eta(\theta)]^T T(\omega) - \zeta(\theta)\} h(\omega)$, $\forall \omega \in \Omega$

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

eg: Binomial, Normal, Dises 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.

$\int f(\omega) \exp(\eta^T T(\omega)) h(\omega) d\nu(\omega) \in C^\infty$ → on Ω.

Cor. $m(T) \in C^\infty$

Basic Terminologies in Statistical Decision Theory.

- Decision Rule: $T: (\mathbb{R}^k, \mathcal{B}_k) \to (\underline{\Delta}, \mathcal{F}_\Delta)$ — Action Space
- Loss Function: $L(P, T(X))$.
- Risk Function: $R_T(P) = \mathbb{E}_T L(P, T(X))$. 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

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

Rule 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 $(\Delta, \mathcal{F}_\Delta)$

Often, $\delta = \sum_{j=1}^{M} \mathbb{I}(Z = j) T_j$ 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 G be a class of decision rules. $T \in G$ is G-admissible iff $\forall S \in G$, $R_S(P) > R_T(P)$, $\forall P \in \mathcal{P}$

Main Approaches for optimality:

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

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

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

Estimation:

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

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

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

Hypothesis Test

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

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 $f_\theta(w) = \underbrace{\exp\{[\eta(\theta)]^T T(w) - \xi(\theta)\}}_{g_\theta(T(x))} \underbrace{h(w)}_{h(x)}$

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

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

Lecture 7. Bias - Variance Trade-off.

1. Example: Sample mean. $X_1, \cdots X_n$ iid (μ, σ^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)

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

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

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

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

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

$\lambda \uparrow \Rightarrow$ Var $\downarrow$ Bias $\uparrow$

Exercise.

Bias:

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

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

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

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

$$\text{Bias} := \|\lambda(X^TX + \lambda I_p)^{-1}\beta^*\|$$

$$\leq \lambda\|(X^TX + \lambda I_p)^{-1}\|\beta^*$$

$$\leq \lambda(\mu_{min})^{-1}\beta^* \quad \Rightarrow \text{Bias} = O(\lambda)$$

Variance

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

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

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

$$= \sigma^2 \operatorname{tr}(X(X^TX + \lambda I_p)^{-2}X^T)$$

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

$$X(X^TX+\lambda\tilde{I}_p)^{-2}X^T$$

$$= U\Lambda U^T U(\Lambda+\lambda\tilde{I}_p)^{-2}U^T U\Lambda U^T$$

$$= U\Lambda(\Lambda+\lambda\tilde{I}_p)\Lambda U^T$$

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

where μ_i are eigenvalues of X^TX

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

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

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

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

4. Overparametrization

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

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

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

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

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

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

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

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

If x_0 has zero mean and identity covariance matrix.

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

$(X^TX) \hookrightarrow \mathbb{R}^{p\times p}$

The variance generally increases as p increases

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

If the smallest eigenvalue is bounded away from 0.

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

Double Decent 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 overparameterized logistic regression)

Consider logistic loss and linear separable data.

From any initialize $\beta^{(0)} \in \mathbb{R}^p$, 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.

Lecture 7 Exercises

Exercise 1: Solve ridge regression

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

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

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

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

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

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

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

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

Exercise 2: Prove the ridge less solution

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

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

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

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

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

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

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

Subject to $X\beta = y$

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

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

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

$$\Rightarrow \beta = \frac{1}{2}X^T\lambda$$

$$X\beta = \frac{1}{2}XX^T\lambda - y = 0$$

$$\Rightarrow \lambda = (\tfrac{1}{2}XX^T)^{-1}y \quad \text{Since } XX^T \text{ invertible.}$$

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

Lecture 8. Basic Estimation Methods

1. Method of Moments

Classical recipe

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

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

Remark. $\hat{\theta}$ may not exist may not unique

Generalized Method of Moments

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

For simplicity. Let $g: \mathbb{R}^{p+1} \to \mathbb{R}^m$ be cts s.t. $\mathbb{E}[g(X_1, \theta)] = 0$

Rmk m could be larger than p.

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

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

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

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

where α is a small positive constant

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

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

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 $Y_1, \dots Y_n \in \mathbb{R}^p$ with

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

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

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

1. Using covariance matrix (second moment):

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

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

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

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

$U = [u_1, \dots u_p]$ (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: recover $\{w_k\}_{k=1}^{K}$ 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 $\bar{\mu}_k = (v_1 \cdots v_d)^T \mu_k$ $\quad UU^T = I$

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

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 λ_k, $\overline{\mu_k}$ from M_3 numerically

3. Maximum Likelihood Estimation

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

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

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.

$z_i^* = 1$: ith node in comm 1

$z_i^* = -1$: ith node in comm 2 } 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: $\max_{z \in \{\pm 1\}^n : z^T \mathbf{1}_n = 0} \sum_{i<j} \mathbb{1}\{z_i z_j = 1\} Y_{ij}$ (NP Hard)

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

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

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

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

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

$\tilde{z}$: Lingese 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: $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X$.

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

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

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

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

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\Rightarrow \sum_{n=1}^{\infty} P(\|X_n - X\|_2 > \varepsilon) < \infty \qquad \text{i.e. } X_n \xrightarrow{L^2} X.$$

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

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

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

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. $P(|X_n| \ge C) \le \frac{Var(X_n)}{C^2}$ by Markov's inequality

$$\limsup_n P(|X_n| \ge C) \le \frac{\sigma^2}{C^2}$$

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 $X_n = O_p(Y_n)$

as 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 $v \in \mathbb{R}^p$. Let $(a_n)_{n \geq 1}$ be a sequence of positive numbers with $a_n \to \infty$

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

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

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

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.

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

Interchange: Converse is false.

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

$$\int f(x)\,dx = 2C^2 \int_C^\infty x^{-(2+1)}\,dx = 2C^2\left[\frac{-1}{2}x^{-2}\right]_C^\infty = 1$$

$$\mathbb{E}[X] = 2C^2 \int_C^\infty x^{-2}\,dx = \frac{2C}{2-1}$$

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

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 $n\mathbb{P}(|X_n| > n) \to 0$ 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.

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

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

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

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

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

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

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

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

Therefore, $n\mathbb{P}(|X_n| > n) \to 0$

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

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

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

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

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

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

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

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

$P(A_1) = 1$

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 $\mathbb{E}|X_1| = \mu < \infty$

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

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

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

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

Consistency 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.

Equivalent definitions 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 $X_1, X_2, \ldots$ 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 $X_1, X_2, \cdots$ 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). $A_0 = \{\omega : \lim_{n\to\infty} X_n = X\}$ $D = \{x \in \mathbb{R}^k : g \text{ cts at } x\}$.

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

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

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

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

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

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

Slutsky's Theorem Let $X_1, X_2 \ldots, Y_1, Y_2, \ldots$ 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 $X_1, X_2, \ldots$ 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

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

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

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

the asymptotic variance

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

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, $E_P[f(T)] = 0 \Rightarrow f(T) = 0$ a.s. 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 $\mathrm{Var}(T(X)) \le \mathrm{Var}(U(X))$ for any $P \in \mathcal{P}$ and any other unbiased estimator $U(X)$ of ν

Rmk UMVUE does not always exist

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

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

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 $\mathcal{P}$, 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. $I(\theta) \geq 0$. $\frac{\partial}{\partial\theta}\log f_\theta(x)$ 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 on 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 $\nu = g(\theta)$, $g: \mathbb{R}^d \to \mathbb{R}^k$ 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 $\theta \in \Theta$,

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

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

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

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

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

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.

$\mathrm{Var}(\text{MLE}) \to (n I_X(\theta))^{-1}$ 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 $I(\theta) \Rightarrow$ larger local convexity

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

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

"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, $\mathrm{Var}(T(X)) = I(\eta)$
3. $v = \mathbb{E}[T(X)]$, then $\mathrm{Var}(T) = (I(v))^{-1}$

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

Thm.2 Suppose Z is full column rank

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

(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, $c^T\hat\beta$ is independent of $\hat\sigma^2$, moreover,

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

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 $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^3] < \infty$, then

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

WTS: Tail prob of $\bar{X}$ is $O(e^{-nt^2/2})$

2. Subgaussian variables and Hoeffding's Inequality

Thm 1. Subgaussian random variables

Let X be an r.v. with $\mathbb{E}X = 0$. Then the conditions

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

(2) $\sup_{p \ge 1} \frac{1}{\sqrt{p}} (\mathbb{E}|X|^p)^{1/p} \le k_2$ for some $k_2 > 0$

(3) $\mathbb{E}[e^{\lambda X}] \le e^{k_3^2 \lambda^2}$, $\forall\, \lambda \in \mathbb{R}$ Constants independent of ↑ random variable

are equivalent in the sense that $k_i < C_{ij} k_j$ $\forall\, i \ne j$.

X is said to be a subgaussian r.v if they holds.

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 $\|X\|_{\psi_2} \le CM$

Proof of Thm 1:

(1) $\Rightarrow$ (2) $\quad \mathbb{E}|X|^p = \int_0^\infty P(|X| > t)\, p\, t^{p-1}\, dt$

w.l.o.g. $K_1 = 1$ (we can always rescale X)

$$\le \int_0^\infty 2p\, t^{p-1} e^{-t^2}$$

$$= p \int_0^\infty u^{(p-2)/2} e^{-u}\, du$$

$$= p\,\Gamma(p/2) \le 3p \left(\frac{p}{2}\right)^{p/2}$$

$$\|X\|_p \le (3p)^{1/p} \left(\frac{p}{2}\right)^{\frac{1}{2}} (p)^{-\frac{1}{2}} = \frac{\sqrt{2}}{2} (3p)^{1/p}$$

Let $g(p) = \log (3p)^{1/p} = \frac{1}{p} \log(3p)$

$$g'(p) = -\frac{1}{p^2} \log(3p) + \frac{1}{3p^2} = \frac{1 - 3\log(3p)}{3p^2}$$

$$g(p) \le g\left(\tfrac{1}{3} e^{\frac{1}{3}}\right) \Rightarrow \|X_p\| \le \frac{\sqrt{2}}{2} \exp\left(g\left(\tfrac{1}{3} e^{\frac{1}{3}}\right)\right)$$

(3) $\Rightarrow$ (1). For $\lambda > 0$, $t > 0$

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

Choose $\lambda = t/2$ to minimize $-\lambda t + \lambda^2$

$$\leq e^{-\lambda t + \lambda^2} \leq e^{-t^2/4}$$

Similarly, $\mathbb{P}(-X \geq t) \leq e^{-t^2/4}$

Then, $\mathbb{P}(|X| > t) \leq 2e^{-t^2/4}$

(2) $\Rightarrow$ (3): W.L.O.G. $K_1 = 1$. Let $\tilde{X} \overset{d}{=} X$ be an independent copy of X. Then $X - \tilde{X} \overset{d}{=} \tilde{X} - X$ because

$$\phi_{X-\tilde{X}}(t) = \phi_X(t)\,\phi_{\tilde{X}}(-t) = \phi_{\tilde{X}}(t)\,\phi_X(-t) = \phi_{\tilde{X}-X}(t)$$

By Jensen's inequality we have $\mathbb{E}e^{-\lambda\tilde{X}} \geq e^{-\lambda\mathbb{E}\tilde{X}} = 1$

Therefore, $\mathbb{E}e^{\lambda X} \leq \mathbb{E}e^{\lambda X}\mathbb{E}e^{-\lambda\tilde{X}} = \mathbb{E}e^{\lambda(X-\tilde{X})}$

$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(\lambda(X-\tilde{X}))^k}{k!}\right]$$

$$= 1 + \sum_{k:\text{even}}^{\infty} \frac{\lambda^k \mathbb{E}[(X-\tilde{X})^k]}{k!}$$

$$\leq 1 + \sum_{k:\text{even}}^{\infty} C_1 \frac{(2\lambda)^k k^{k/2}}{\sqrt{2\pi k}(k/e)^k} \quad \text{(by Stirling's)} \; ?$$

$$\leq 1 + \sum_{k:\text{even}}^{\infty} C_2 \frac{(2\sqrt{e}\lambda)^k}{\sqrt{2\pi k/2}(k/(2e))^{k/2}}$$

$$\leq 1 + \sum_{m=0}^{\infty} \frac{(4C_2 e\lambda^2)^m}{m!} \quad (m := k/2)$$

$$= e^{K_2\lambda^2} \quad (\text{Choose } K_2 = 4C_2 e)$$

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

$$P\left(\left|\sum_{i=1}^n a_i X_i\right| > t\right) \le 2\exp\left(-\frac{Ct^2}{K\|a\|^2}\right)$$

holds for certain absolute constant $C > 0$

3. Subexponential Variables and Bernstein's Inequality

Thm 3. (Subexponential random variables) Let X be a random variable with $\mathbb{E}X = 0$. Then

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

(2) $\sup_{p \ge 1} \frac{1}{p} \mathbb{E}[|X|^p]^{1/p} \le K_2$

(3) $\mathbb{E}[e^{\lambda X}] \le e^{K_3 \lambda^2}$, $\forall |\lambda| \le 1/K_3$

are equivalent in the sense that $K_i \le C_{ij} K_j$

Constants 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 equivalent. Moreover, we have bound

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

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

$\langle X, Y\rangle_p \le \|X\|_{2p}\|Y\|_{2p}$?

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

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

for some constant $c > 0$

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