Stein's Phenomenon

MLE $X_1, \dots X_n \overset{i.i.d.}{\sim} N(\mu, \Sigma)$

$\hat{\mu}_n^{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i \longrightarrow \mu$ Consistency. $\Big\rangle$ $n \to \infty$ fixed p

$tr(\hat{I}^{-1}) = \lim_{n \to \infty} var(\hat{\mu}_n^{MLE}) \leq \lim_{n \to \infty} Var(\hat{\mu}_n)$

Without Loss of Generality. $\Sigma = U \Lambda U^T$ $\Lambda = diag(\lambda_i)$

$Y_i = \Lambda^{-\frac{1}{2}} U^T X_i$ P.C.A.

$Y_i \sim N(\mu, I_p)$

Risk (Mean Square Error / MSE)

Given $\hat{\mu}_n(Y_1, \cdots, Y_n)$

Risk. $R(\hat{\mu}_n, \mu) = E_{Y_1 \cdots Y_n} L(\hat{\mu}_n(Y_1 \cdots Y_n), \mu)$

$\overset{MSE}{=} E \| \hat{\mu}_n - \mu \|^2$ $\hat{\mu}_n, \mu \in \mathbb{R}^p$

Bias - Variance

$$R(\hat{\mu}_n, \mu) = E \| \hat{\mu}_n - E(\hat{\mu}_n) + E(\hat{\mu}_n) - \mu \|^2$$

$$= \underbrace{E \| \hat{\mu}_n - E(\hat{\mu}_n) \|^2}_{\|} + \underbrace{\| E(\hat{\mu}_n) - \mu \|^2}_{\|} + \cancel{2E \langle \hat{\mu}_n - \underset{\overset{\|}{0}}{E}(\hat{\mu}_n), E(\hat{\mu}_n) - \mu \rangle}$$

$$= Var(\hat{\mu}_n) + Bias(\hat{\mu}_n)$$

Examp $Y_i \sim N(\mu, \sigma^2 I_p)$ $\hat{\mu}_n^{MLE} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ $E[Y_i] = \mu$

$Bias(\hat{\mu}_n^{MLE}) = 0$ unbiased!

$Var(\hat{\mu}_n^{MLE}) = \frac{\sigma^2 p}{n}$

In particular, $n = 1$. $Var(\hat{\mu}_1^{MLE}) = \sigma^2 p$.

$R(\hat{\mu}_{n}^{MLE}, \mu) = \sigma^2 p$

Linear Estimator $\hat{\mu}_C(Y) = CY$, $Y \sim N(\mu, \sigma^2 I_p)$ n=1

$C = I \rightarrow$ MLE

$C = \mathrm{diag}(c_i)$ $\min_{\theta} \frac{1}{2}\|Y-\theta\|^2 + \frac{\lambda}{2}\|\theta\|^2$ Ridge Regression

$c_i = \frac{1}{1+\lambda}$ $\hat{\theta} = \frac{1}{1+\lambda} Y$

$$\mathrm{Bias}(\hat{\mu}_C) = \|\mathbb{E}(\hat{\mu}_C) - \mu\|^2 = \|(I-C)\mu\|^2$$ $\mathbb{E}(CY) = C\mu$

$$\mathrm{Var}(\hat{\mu}_C) = \mathrm{tr}\,\mathbb{E}(CY-C\mu)^T(CY-C\mu)$$

$$= \mathbb{E}\,\mathrm{tr}[(Y-\mu)^T C^T C (Y-\mu)] = \mathrm{tr}[C^T C]\,\underbrace{\mathbb{E}(Y-\mu)(Y-\mu)^T}_{\sigma^2 I_p}$$

$$= \sigma^2 \mathrm{tr}(C^T C)$$

$C = \mathrm{diag}(c_i)$

$$R(\hat{\mu}_C, \mu) = \sum_{i=1}^{p} \sigma^2 c_i^2 + \sum_{i=1}^{p} (1-c_i)^2 \mu_i^2 \leq \sum \sigma^2 c_i^2 + \sum (1-c_i)^2 \tau_i^2$$

Statiscal Decision theory : Minimax Risk

$|\mu_i| < \tau_i$ Rectangular class

$$\inf_{c_i} \sup_{|\mu_i|<\tau_i} R(\hat{\mu}_C, \mu) = \sum_{i=1}^{p} \frac{\sigma^2 \tau_i^2}{\tau_i^2 + \sigma^2} \leq \sigma^2 p$$ MLE

sparse family 启示

Problem :

$\hat{\mu}_n$ better estimator ?

Inadmissible : $\hat{\mu}_n$ is inadmissible

$\exists\, \mu_n^*$ s.t. $\mathbb{E}\|\mu_n^* - \mu\|^2 \leq \mathbb{E}\|\hat{\mu}_n - \mu\|^2$ for all μ $(\in \mathbb{R}^p)$ $(\in \Theta)$

$\exists\, \mu_0$ $R(\mu_n^*, \mu_0) < R(\hat{\mu}_n, \mu_0)$

$\hat{\mu}_n^{MLE}$ inadmissible ? Stein '1956 , James-Stein '1961

Yes

$$\hat{\mu}_n^{JS} = \left(1 - \frac{\sigma^2(p-2)}{\|\hat{\mu}_n^{MLE}\|}\right)\hat{\mu}_n^{MLE}, \quad Y \sim N(\mu, \sigma^2 I_p).$$

Thm

$$R(\hat{\mu}_n^{JS}, \mu) < R(\hat{\mu}_n^{MLE}, \mu), \quad \exists \mu \in R^p, \; p \geq 3$$

几乎所有的 μ

where $\hat{\mu}_{n=1}^{MLE} = Y$, $\hat{\mu}_{n=1}^{JS} = \left(1 - \frac{\sigma^2(p-2)}{\|Y\|^2}\right) Y.$

Stein's Unbiased Risk Estimates (SURE) : $Y \sim N(\mu, I_p).$

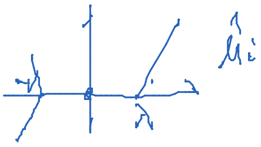
$$\hat{\mu}(Y) = Y + g(Y), \quad g \text{ nonlinear}$$

例子: ✓ 1) Linear Est. $\hat{\mu} = CY$, $g(Y) = (C - I)Y$

✓ 2) Soft Thresholding $g_{ST}^i(Y) = \begin{cases} -\lambda & Y_i > \lambda \\ -Y_i & |Y_i| \leq \lambda \\ \lambda & Y_i < -\lambda \end{cases}$

$\min_{\hat{\mu}} \frac{1}{2}\|Y - \hat{\mu}\|^2 + \lambda\|\hat{\mu}\|_1 \Rightarrow \hat{\mu}_{ST}$: $Y - \hat{\mu}^{ST} + \lambda \partial \|\hat{\mu}\|_1 = 0$

$\partial \|\hat{\mu}\|_1 = \operatorname{sign}(\hat{\mu}_i)$

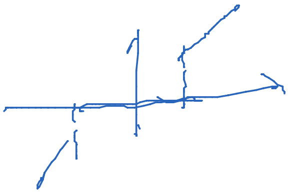


$\Leftarrow \begin{cases} \hat{\mu}_i \neq 0 : & Y_i - \hat{\mu}_i + \lambda \operatorname{sign}(\hat{\mu}_i) \Rightarrow 0 \\ \hat{\mu}_j = 0 & |Y_j - \hat{\mu}_j| < \lambda \end{cases}$

✗ 3) Hard Thresholding $g_{HT}^i(Y) = \begin{cases} 0 & |Y_i| > \lambda \\ -Y_i & |Y_i| \leq \lambda \end{cases}$

– not weakly differentiable

$\frac{1}{2}\|Y - \hat{\mu}\|^2 + \lambda\|\hat{\mu}\|_0$



✓ 4) JS $g_{JS}(Y) = -\frac{\sigma^2(p-2)}{\|Y\|^2}$ $\sigma^2 = 1$

$g(x)$ 1) weakly differentiable. $g(x_i, x_{-i})$ ~ $\forall i$. x_{-i}

absolutely continuous w.r.t. x_i

2) $\sum_{i=1}^{p} \int |\partial_i g_i(x)| \, dx_i < \infty$

$\hat{\mu}(Y) = Y + g(Y)$ suitable g. above

Lemma (Stein '61)

$$R(\hat{\mu}, \mu) = \mathbb{E}\left[\, p + 2\nabla^T g(Y) + \|g(Y)\|^2 \right]$$

$$\underline{\nabla^T g(Y) := \sum_{i=1}^{p} \frac{\partial}{\partial Y} g_i(Y)}$$

Proof (Integration by Parts)

$$\mathbb{E}\|\hat{\mu} - \mu\|^2 = \mathbb{E}\|Y + g(Y) - \mu\|^2 = \mathbb{E}\|(Y-\mu) + g(Y)\|^2$$

$$= \underset{\overset{\|}{p.}}{\underline{\mathbb{E}\|Y-\mu\|^2}} + 2\mathbb{E}(Y-\mu)^T g(Y) + \mathbb{E}\|g(Y)\|^2$$

$Y \sim N(\mu, \overset{\overset{1}{\|}}{\sigma^2} I_p)$

$$\mathbb{E}\left[(Y-\mu)^T g(Y)\right] = \sum_{i=1}^{p} \int_{-\infty}^{+\infty} \frac{\partial g_i(Y)}{\partial Y_i} \phi(Y-\mu)\, dY = \underline{\mathbb{E}[\nabla^T g(Y)]}$$

$y \sim N(\mu, 1)$ $\quad \phi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2}}$ $\quad \int_{-\infty}^{+\infty} \underline{(y-\mu)}\, g(y)\, \underline{\phi(y-\mu)}\, dy$

$\frac{\partial}{\partial y}\phi = -(y-\mu)\phi(y)$ $\quad = -\int_{-\infty}^{+\infty} g(y) \frac{\partial}{\partial y}\phi(y-\mu)\, dy$

$$= \underset{\overset{\|}{0}}{-g(y)\phi(y-\mu)\Big|_{-\infty}^{+\infty}} + \int_{-\infty}^{+\infty} \phi(y-\mu) \frac{dg(y)}{dy}$$

$$\therefore \text{ above} = \underline{\mathbb{E}\left[\, p + \nabla^T g(Y) + \|g(Y)\|^2 \right]}$$

$$U(Y) = p + \nabla^T g(Y) + \|g(Y)\|^2$$

$$U(Y) = p + \nabla^T g(Y) + \|g(Y)\|^2$$

1) JS.

$$g(Y) = -\frac{p-2}{\|Y\|^2} Y$$

$$U(Y) = p + 2\left[-\sum_{i=1}^{p}\frac{\partial}{\partial Y_i}\left(\frac{p-2}{\|Y\|^2} Y_i\right)\right] + \frac{(p-2)^2}{\|Y\|^2}$$

$$= p - 2\frac{(p-2)^2}{\|Y\|^2} + \frac{(p-2)^2}{\|Y\|^2}$$

$$\underline{R(\hat{\mu}^{JS}, \mu) = \mathbb{E}\, U(Y) = p - \mathbb{E}\frac{(p-2)^2}{\|Y\|^2}} < p = R(\hat{\mu}^{MLE}, \mu)$$

$n=1,\ \sigma=1$

$\underline{\sigma^2/n}$

$\hat{\mu}^{MLE}$ inadmissible $p \geq 3$

Note $Y \sim N(\mu, I_p)$ $\forall\, \hat{\mu} = CY$ (Lin. est)

$\hat{\mu}$ is admissible iff

1) C sym. $C = C^T$

2) $0 \leq \text{eigval}(C) \leq 1$

3) $\text{eigval}_i(C) = 1$. for at most two i

Lemma 2.8. Johnstone (GE)

例: $\hat{\mu}^{JS+} = \left(1 - \frac{p-2}{\|Y\|^2}\right)_+ Y$ better than MLE JS.

ST. $R(\hat{\mu}^{ST}, \mu) \asymp 1 + (2\log p + 1)\sum_{i=1}^{p}(\mu_i^2 \wedge 1)$

注: $R(\hat{\mu}^{JS}, \mu) = 2 + c\left(\left(\sum_{i=1}^{p}\mu_i^2\right) \wedge p\right)$ $c \in (\frac{1}{2}, 1)$

ST < JS. sparse. $\mu = (*, 0, \cdots, 0)$ $2\log p + 1 < O(p)$

dense. $\mu = (1, \cdots, 1)$ $R^{ST} < R^{JS}$

MLE. 有限 n. 有限 p. (≥ 3)

又是 inadmissible

$\text{MSE}(\text{JS}) < \text{MSE}(\text{MLE})$

$\text{MSE}(\text{ST}) <$ 〃
(Lin)

"Shrinkage." better than M.LE,