Maximum Likelihood Estimate (极大似然估计)

Fisher

Statistical Model: $f(x|\theta)$, $x \in \mathbb{R}^d$ etc. $\theta \in \mathbb{R}^p$

probability model

Data: i.i.d. $x_1 \cdots x_n \sim f(x|\theta_0)$ $\theta_0 \in \mathbb{R}^p$

目标: $\hat{\theta} = G(x_1, \ldots, x_n) \xrightarrow[n\to\infty]{} \theta_0$?

MLE $\hat{\theta}^{MLE} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} f(x_i|\theta)$ (M-estimate)

$= \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ln f(x_i|\theta)$

例子: $f(x|\theta) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$, $\theta = (\mu, \Sigma)$

$x \in \mathbb{R}^p$

$x_1, \ldots, x_n$ i.i.d.

MLE $\Rightarrow$?

Log Likelihood:

$\log f(x|\theta) = -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) - \frac{1}{2}\ln|\Sigma| + \text{Const}$

$I_n = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta) = -\frac{1}{2n}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) - \frac{\cancel{n}}{2\cancel{n}}\ln|\Sigma| + C$

1st order condition

$0 = \frac{\partial I_n}{\partial \mu} = -\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}(x_i-\mu) \Rightarrow \hat{\mu}^{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample mean !

$$\operatorname{trace}[I_n(\Sigma)] = -\frac{1}{2n}\sum_{i=1}^{n}\operatorname{tr}\left[(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right] - \frac{1}{2}\operatorname{tr}\left[\log|\Sigma|\right] + C$$

Linear, cyclic property $\operatorname{tr}(AB) = \operatorname{tr}(BA)$

$\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \cdots$

$$\frac{1}{2n}\sum_{i=1}^{n}\operatorname{trace}\left[(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right] = \frac{1}{2n}\sum_{i=1}^{n}\operatorname{tr}\left(\Sigma^{-1}(X_i-\mu)(X_i-\mu)^T\right)$$

$$= \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\underline{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^T\right)}\right)$$

$$\hat{S}_n = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})^T, \quad S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^T$$

$$= \frac{1}{2}\operatorname{tr}(\Sigma^{-1}S_n) \qquad S_n^{\frac{1}{2}}S_n^{\frac{1}{2}}$$

$$= \frac{1}{2}\operatorname{tr}\left(\underline{S_n^{\frac{1}{2}}\Sigma^{-1}S_n^{\frac{1}{2}}}\right)$$

对称. Symmetric, p.s.d.

$$S = S_n^{\frac{1}{2}}\Sigma^{-1}S_n^{\frac{1}{2}} = U\Lambda U^T \qquad \Lambda = \operatorname{diag}(\lambda_i)_{i=1,p} \quad \lambda_i \geq 0$$

$$\Sigma = S_n^{+\frac{1}{2}}S^{-1}S_n^{+\frac{1}{2}} \qquad \det(AB) = |AB| = |A|\cdot|B|$$

$$\log|\Sigma| = +\log|S_n| - \log|S| \qquad S\ (\Sigma)\ \text{变量}$$

S_n sample cov. Σ无关

$$\operatorname{argmax} I_n(\Sigma) = -\frac{1}{2}\operatorname{trace}(S) + \frac{1}{2}\log|S| + C(S_n, 1)$$

$$= -\frac{1}{2}\sum_{i=1}^{p}\lambda_i + \frac{1}{2}\sum_{i=1}^{p}\log\lambda_i + C$$

$$\frac{\partial I_n}{\partial\lambda_i} = -\frac{1}{2} + \frac{1}{2\lambda_i} \Rightarrow \lambda_i = 1$$

$$S = I_p = S_n^{\frac{1}{2}}\Sigma^{-1}S_n^{\frac{1}{2}} \Rightarrow \Sigma^{MLE} = S_n.$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^T$$

Note $\hat{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^T$

总结：

1) $\hat{\mu}^{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$ sample mean

2) $\hat{S}_n^{MLE} = \frac{1}{n}\sum_{i=1}^{n} (X_i-\hat{\mu}^{MLE})(X_i-\hat{\mu}^{MLE})^T$ sample covariance

为什么 MLE.

Generally: $X_i \sim f(x_i|\theta_0) \leftarrow$ unknown.

$$\hat{\theta}^{MLE} = \arg\max_{\theta} L(x_{1,n}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$\theta \in \mathbb{R}^p$ p fixed. $n \to \infty$ Limiting properties

1) Consistency $\hat{\theta}^{MLE} \xrightarrow{n\to\infty} \theta_0$ (prob. / almost sure)

2) Asymptotic Normality $\sqrt{n}(\hat{\theta}^{MLE} - \theta_0) \xrightarrow{d} N(0, I^{-1})$

I: Fisher Information matrix

$$I_{ij} = -E_X\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln f(X|\theta_0)\right]_{p\times p} \geq 0$$

3) Asymptotic Efficiency: (second order)

$$tr(I^{-1}) = \lim_{n\to\infty} Var(\hat{\theta}^{MLE}) \leq \lim_{n\to\infty} Var(\hat{\theta}) \;;\; \hat{\theta} \text{ unbiased}$$

参数空间 有限维 $\lim_{n\to\infty} \frac{p}{n} = 0$ 传统统计学

Big Data $n \to \infty$ $p_n \to \infty$ 高维统计学

$\lim_{n\to\infty} \frac{p_n}{n} \to c \neq 0$, ∞

MLE 还好估计？ Stein's phenomenon.

有限 n, p ≥ 3. ∃ JS, MLE 好!

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. $n \to \infty$ fixed p

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

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

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

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

Bias-Variance

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

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

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

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

$Bias(\hat{\mu}_n^{MLE}) = 0$ 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 $\hat{\mu}_C(Y) = CY, \quad Y \sim N(\mu, \sigma^2 I_p) \quad n=1$

$C = I \rightarrow$ MLE

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

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

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

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

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

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

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

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

Statisical Decision theory : Minimax Risk

$|\mu_i| < \tau_i$ Rectangular class

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

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

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

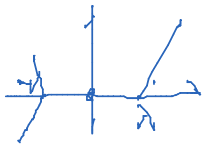
Stein's Unbiased Risk Estimates (SURE): $Y \sim N(\mu, I_p)$.

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

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

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

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

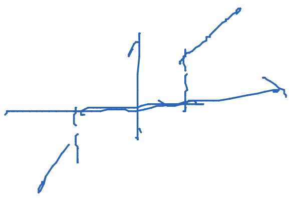


$$\Leftarrow \begin{cases} \hat{\mu}_i \neq 0: & Y_i - \hat{\mu}_i + \lambda\, \mathrm{sign}(\hat{\mu}_i) \Rightarrow \\ \hat{\mu}_j = 0 & |Y_j - \hat{\mu}_j| < \lambda \end{cases}$$

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

— not weakly differentiable

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



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

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

absolutely continuous w.r.t. x_i

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

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

Lemma (Stein '61)

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

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

Proof (Integration by Parts)

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

$$= \underbrace{\mathbb{E}\|Y-\mu\|^2}_{= p} + 2\mathbb{E}(Y-\mu)^T g(Y) + \mathbb{E}\|g(Y)\|^2$$

$Y \sim N(\mu, \sigma^2 I_p)$, $\sigma^2 = 1$

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

$y \sim N(\mu, 1)$ $\qquad \phi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2}}$

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

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

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

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

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

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

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 $p \ge 3$

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) $0 \le \text{eigval}(C) \le 1$

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

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

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

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

MLE. 有限 n. 有限 p. (≥ 3)

不是 inadmissible

MSE (JS) < MSE (MLE)

MSE (ST) < 〃
(Lin)

"Shrinkage." better than MLE,