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:

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

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

1^{st} 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 !

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

Linear, cyclic property $tr(AB) = tr(BA)$

$tr(ABC) = tr(BCA) = \cdots$

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

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

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

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

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

对称 symmetric, p.s.d.

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

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

$$\log|\Sigma| = +\log|S_n| - \log|S|$$

S (Σ) 变量

S_n sample cov. Σ 无关

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

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

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

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

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

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

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

$\theta \in R^p$ 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

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

MLE 还好估计? Stein's phenomenon.

有限 n, p ≥ 3. ∃ JS, MLE 好!