03/14/18

$y = A(x) + e$ (e and y random)

$P(y|x)$ likelihood function

- frequentist: $\hat{x}(y)$ estimator of x, random
- Bayesian: $P(x|y)$ posterior

$P(x|y) \sim P(y|x)\,P(x)$ ← prior

Def $MSE = \mathbb{E}\,\|\hat{x} - x\|^2$

Prop. $= \underbrace{\mathbb{E}\,\|\hat{x} - \mathbb{E}\hat{x}\|^2}_{\text{Var } \hat{x}} + \underbrace{\|\mathbb{E}\hat{x} - x\|^2}_{\text{Bias}(\hat{x},x)^2}$

Reminder: $\mathbb{E}[f(y)] = \int f(y)\,p(y|x)\,dy$.

Def Fisher score of y: (random)

$$V = \frac{\partial}{\partial x} \ln P(y|x) = \frac{\frac{\partial}{\partial x} P(y|x)}{P(y|x)}$$

Rmk. $\mathbb{E}V = \int \frac{\partial}{\partial x} P(y|x) \cdot \frac{1}{P(y|x)}\, P(y|x)\,dy$

$= \frac{\partial}{\partial x} \int P(y|x)\,dy = \frac{\partial}{\partial x} 1 = 0.$

Def Fisher information of y (number)
("how much information y contains about x").

$$J(x) = \mathbb{E}\left[\left(\frac{\partial}{\partial x} \ln P(y|x)\right)^2\right] \quad (= \mathbb{E}V^2 = \text{Var}\,V)$$

ex. $y = (y_1, \dots, y_n)$ iid, then

$P(y|x) = \prod_i P(y_i|x) \Rightarrow V_y = \sum_{i=1}^{n} V_{y_i}$

$J_y(x) = n\,J_{y_i}(x)$

("additive, and relative to what we want to learn", here x)

Thm (Cramér-Rao lower bound, $x \in \mathbb{R}$)

Let $\hat{x}$ be unbiased: $\mathbb{E}\hat{x} = x$.

Then, $\operatorname{Var} \hat{x} \geq \frac{1}{J(x)}$

Pf Recall $J_y(x) = \mathbb{E}V^2$ $\quad (\mathbb{E}V = 0)$

$\operatorname{Var}\hat{x} = \mathbb{E}(\hat{x}-x)^2$ $\quad (\mathbb{E}\hat{x} = x)$

Cauchy-Schwarz $\quad (\int fg)^2 \leq (\int f^2)(\int g^2)$

$$\underbrace{\left(\mathbb{E}[V(\hat{x}-x)]\right)^2}_{\mathbb{E}V\hat{x}} \leq \underbrace{(\mathbb{E}V^2)}_{J(x)} \underbrace{(\mathbb{E}(\hat{x}-x)^2)}_{\operatorname{Var}\hat{x}}$$

$$\mathbb{E}V\hat{x} = \int \frac{\frac{\partial}{\partial x} p(y|x)}{p(y|x)} \hat{x}(y)\, p(y|x)\, dy$$

$$= \frac{\partial}{\partial x} \int \hat{x}(y)\, p(y|x)\, dy$$

$$= \frac{\partial}{\partial x} \mathbb{E}\hat{x} = \frac{\partial}{\partial x} x = 1$$

□

Thm Biased case:

$$\operatorname{Var}\hat{x} \geq \frac{\left(1 + \frac{\partial}{\partial x}\overbrace{\mathbb{E}\hat{x} - x}^{\operatorname{Bias}(\hat{x},x)}\right)^2}{J(x)}$$

Thm Multivariate, unbiased case:

$$\Sigma_{ij} = \operatorname{Cov}(\hat{x}_i, \hat{x}_j) = \mathbb{E}[\hat{x}_i \hat{x}_j] \qquad (\Sigma_{ii} = \operatorname{Var}\hat{x}_i)$$

$$J_{ij}(x) = \mathbb{E}\left[\frac{\partial}{\partial x_i} \ln p(y|x) \frac{\partial}{\partial x_j} \ln p(y|x)\right]$$

(Fisher information matrix)

then $\Sigma \geq J^{-1}(x)$ meaning $\Sigma - J^{-1}(x) \geq 0$.

in particular $\mathrm{Var}\,\hat{x}_i \geq (J^{-1})_{ii}$

MLE: $\hat{x} = \underset{x}{\operatorname{argmax}}\, p(y|x)$.

Properties: $(x \in \mathbb{R})$

- Consistency: $\hat{x} \xrightarrow{p} x$ as $n \to \infty$ $(y \in \mathbb{R}^n)$ (conv. in probability)
- asymptotic efficiency:

$$\frac{(\mathrm{Var}\,\hat{x})\, J(x)}{(1 + \mathrm{Bias}(\hat{x}, x))^2} \to 1 \quad \text{as } n \to \infty.$$

(with or without denom) so reaches Cramer-Rao as $n \to \infty$

- asymptotic normality:

$$\sqrt{n}\,(\hat{x} - x) \xrightarrow{d} N(0, J^{-1}(x))$$

(d: convergence in distribution; in the vector case as well)

suggests $\mathrm{Bias}(\hat{x}, x) \sim 1/\sqrt{n}$

although:

- asymptotic unbiasedness:

$$\mathrm{Bias}(\hat{x}, x) \sim 1/n$$

Rmk: $J_{ij} = -E\left[\frac{\partial^2}{\partial x_i \partial x_j} \ln p(y|x)\right]$.

Ex. $y = \underset{\in \mathbb{R}^N}{Ax} + \underset{\in \mathbb{R}^N}{e}$, $e \sim N(0, \Sigma)$

$\mathbb{E} e_i e_j = \Sigma_{ij}$

$y|x \sim N(Ax, \Sigma)$

$p(y|x) = C \exp\left[-\frac{1}{2}(y-Ax)^T \Sigma^{-1} (y-Ax)\right]$

MLE: $\max_x p(y|x) \iff \min_x -\ln p(y|x)$

$\iff \min_x \frac{1}{2}(y-Ax)^T \Sigma^{-1}(y-Ax)$

$\|y-Ax\|^2_{\Sigma^{-1}}$

$\nabla_x = A^T\Sigma^{-1}(Ax-y) = 0$

$\Rightarrow A^T\Sigma^{-1}Ax = A^T\Sigma^{-1}y$

$\hat{x} = (A^T\Sigma^{-1}A)^{-1}A^T\Sigma^{-1}y$

(provided invertible (overdet).

$\hat{x}|x$ is Normal with expectation

$(A^T\Sigma^{-1}A)^{-1}A^T\Sigma^{-1}Ax = x \rightarrow \mathbb{E}\hat{x} = x$, unbiased.

[Lemma: $y \sim N(0, \Sigma)$
$\Rightarrow My \sim N(0, M\Sigma M^T)$
here $M = (A^T\Sigma^{-1}A)^{-1}A^T\Sigma^{-1}$.]

and $\mathrm{Cov}(\hat{x}) = (A^T\Sigma^{-1}A)^{-1}A^T\Sigma^{-1}\Sigma\Sigma^{-1}A(A^T\Sigma^{-1}A)^{-1}$
$= (A^T\Sigma^{-1}A)^{-1}$

suitably decreases as rows are added to A.

$$J_{ij}(x) = E\left[\frac{\partial}{\partial x_i} \ln p(y|x) \frac{\partial}{\partial x_j} \ln p(y|x)\right]$$

$$\frac{\partial}{\partial x_i} \ln p(y|x) = -\left[A^T\Sigma^{-1}(Ax-y)\right]_i$$

$$J(x) = E\left[A^T\Sigma^{-1}(Ax-y)(Ax-y)^T\Sigma^{-1}A\right]$$

$$= A^T\Sigma^{-1} E(yy^T) \Sigma^{-1}A$$

$$= A^T\Sigma^{-1}\Sigma\Sigma^{-1}A$$

$$= \boxed{A^T\Sigma^{-1}A}$$

$$J^{-1}(x) = (A^T\Sigma^{-1}A)^{-1} = Cov(\hat{x})$$

→ efficient (rare occurrence!).

Rmk: $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \Rightarrow J(x) = A_1^T\Sigma^{-1}A_1 + A_2^T\Sigma^{-1}A_2$

additive, also known as "precision matrix" $= Cov(\hat{x})^{-1}$

→ build up precision by adding data.

Rmk: $f(x) = \frac{1}{2}(y-Ax)^T\Sigma^{-1}(y-Ax)$

$\nabla\nabla f(x) = A^T\Sigma^{-1}A$ Hessian

→ another interpretation of $J(x)$.

ex. $y = x + e$ $\quad x \in \mathbb{R}$

$A = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ "ones$(n,1)$"

$e \sim N(0, \sigma^2 I)$

$A^T\Sigma^{-1}A = n\sigma^{-2}$

$\hat{x} = (A^T\Sigma^{-1}A)^{-1}A^T\Sigma^{-1}y = \frac{\sigma^2}{n}\sigma^{-2}[1 \cdots 1]y = \frac{1}{n}\sum y_i$

$E\hat{x} = x$

$Var\,\hat{x} = (A^T\Sigma^{-1}A)^{-1} = \frac{\sigma^2}{n}$