Midterm2

Sufficient Statistic

$p(X_1, \dots, X_n \mid T(X_1, \dots, X_n); \theta)$ does not depend on θ for any t

The factorization theorem (Neyman–Fisher)

$$p(x_1, \dots, x_n; \theta) = h(x_1, \dots, x_n)\, g(T(x_1, \dots, x_n); \theta)$$

Example: $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

$$p(x_1, \dots, x_n; \mu, \sigma^2) = \prod \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \underbrace{\frac{1}{\sqrt{2\pi}^n \sigma^n}}_{h} \underbrace{\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2\right)}_{g}$$

$T(X_1, \dots, X_n) = (\sum^n X_i^2, \sum^n X_i)$ if μ, σ^2 unknown

$T(X_1, \dots, X_n) = (\sum^n X_i)$. if μ unknown, σ known

proof of factorization theorem

$$L(\theta) = g(T(X_1, \dots, X_n); \theta) \underbrace{h(x_1, \dots, x_n)}$$

ignore constants do not depend on θ

$$L(\theta) = g(T(X_1, \dots, X_n); \theta)$$

Minimal sufficiency,

Define ratio: $R = \dfrac{p(y_1, \dots, y_n; \theta)}{p(x_1, \dots, x_n; \theta)}$

T is MSS if R does not depend on θ iff $T(x_1, \dots, x_n) = T(y_1, \dots, y_n)$

Rao–Blackwell theorem:

$$R(\tilde{\theta}, \theta) \le R(\hat{\theta}, \theta). \qquad \tilde{\theta} = \mathbb{E}(\hat{\theta} \mid T)$$

Example: $X_1, \dots, X_n \sim Ber(\theta)$.

$\hat{\theta} = X_1$

$$\tilde{\theta} = E[X_1 \mid \sum^n X_i] = 1 \cdot p(X_1 = 1 \mid \sum^n X_i) = \frac{\cancel{p} \cdot C_{n-1}^{T-1} \cancel{p^{T-1}(1-p)^{(n-1)-(T-1)}}}{C_n^T \cancel{p^T (1-p)^{n-T}}}$$

$$= \frac{T}{n}$$

$R(\hat{\theta}) = \theta(1-\theta)$

$R(\tilde{\theta}) = \frac{\theta(1-\theta)}{n} < \theta(1-\theta)$

Proof of Rao-Blackwell.

$$R(\tilde{\theta}, \theta) = \mathbb{E}\left[(E[\hat{\theta} \mid T] - \theta)^2 \right]$$

$$= \mathbb{E}\left[(E[\hat{\theta} - \theta \mid T])^2 \right]$$

$$\star \quad \le \mathbb{E}\left[E[(\hat{\theta} - \theta)^2 \mid T] \right] \quad \text{by Jensen.}$$

$$= R(\hat{\theta}, \theta)$$

Exponential Family

$$p(x; \theta) = \exp\left[\sum^s \eta_i(\theta) T_i(x) - A(\theta) \right] h(x)$$

$A: \Theta \to \mathbb{R}$

canonical parametrization:

$$p(x; \theta) = \exp\left[\sum^s \theta_i T_i(x) - A(\theta) \right] h(x)$$

θ: natural parameters.

properties of Exponential Families

- Random sampling.

The exponential structure is preserved for an iid sample $\{x_1, \dots, x_n\} \sim p(x_i \dots)$

$$p(x_1, \dots, x_n; \theta) = \prod^n h(x_i) \exp\left[\sum^s \theta_i \sum^n T_i(x_j) - nA(\theta) \right]$$

same natural parameters

$T(x_1, \dots, x_n) = \sum^n T \dots$

$T_i(x_1, \cdots, x_n) = \sum T_i(x_j)$

- $A(\theta)$ log-normalization constant
  log-partition function
  cumulant function

$$A(\theta) = \log \left[\int_x \exp\left[\sum^s \theta_i T_i(x)\right] h(x)\, dx \right]$$

$$\frac{A(\theta)}{\theta_i} = \frac{\int_x \exp\left[\sum^s \theta_i T_i(x)\right] T_i(x)\, h(x)\, dx}{\int_x \exp\left[\sum^s \theta_i T_i(x)\right] h(x)\, dx} = E[T_i(x)]$$

$$\frac{\partial^2 A(\theta)}{\partial\theta_i \partial\theta_j} = \mathrm{cov}(T_i(x), T_j(x))$$

- The log likelihood in an exponential family is concave

$$LL(\theta; x_1, \cdots, x_n) \propto \left[\sum^s \theta_i \sum^n T_i(x_j) - nA(\theta)\right]$$

Hessian is $(-n)\times$ Hessian of A. $\quad$ A is convex

- minimal

  no set of coefficient. such that $\sum a_i T_i(x) = \text{const}$

  Over complete exponential families are not statistically identifiable.

- The exponential families arise naturally $\begin{cases} 1. \text{ maximize the entropy of distribut} \\ 2. \text{ constraint } \quad \hat{\mu}_i = \mathbb{E}_p[T_i(x)] \end{cases}$

- MLE coincide with MOM

$$LL(\theta; x_1, \cdots, x_n) \propto \left[\sum^s \theta_i \sum^n T_i(x_j) - nA(\theta)\right] \rightarrow \text{concave}.$$

$$\partial LL(\theta; x_1, \cdots, x_n) = \sum^n T_i(x_j) - nA'(\theta)$$

$$= \underbrace{\sum_j^n T_i(x_j) - n E_p[T_i(x)]}$$

Point Estimation

The Method of Moments

The Method of Moments

Maximum Likelihood

The MLE is equivariant

Bayes Estimator

$$p(\theta \mid X_1, \cdots X_n) \propto \mathcal{L}(\theta)\, p(\theta) \qquad \text{Likelihood} \times \text{prior}$$

Example 1: $X_1, \cdots, X_n \sim \mathrm{Ber}(\theta)$. $\theta \sim \mathrm{Beta}(\alpha, \beta)$ prior

$$p(\theta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$$p(\theta \mid X) \propto \mathcal{L}(\theta)\, p(\theta) = \theta^{\sum X_i} (1-\theta)^{n-\sum X_i}\, \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$$p(\theta \mid X) = \mathrm{Beta}(\alpha + \textstyle\sum X_i,\ n - \sum X_i + \beta)$$

$$\hat{\theta}_{Bayes} = \frac{\alpha + \sum X_i}{n + \alpha + \beta} = (1-\lambda)\hat{\theta}_{MLE} + \lambda \bar{\theta}$$

Example 2: $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, $\mu \sim N(m, \tau^2)$

$$p(\mu \mid X) \propto \cancel{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n} \exp\left(-\frac{\sum_{i}^{n} (X_i - \mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\tau} \exp\left(\frac{(\mu - m)^2}{-2\tau^2}\right)$$

$$\frac{1}{\sqrt{2\pi}\tau} \exp\Big(- \frac{\tau^2 \sum^n (X_i - \mu)^2 + (\mu - m)^2 \sigma^2}{2\sigma^2\tau^2}$$

$$\tau^2 \left(n\mu^2 - 2\sum^n X_i \mu + \sum^n X_i^2\right) +$$

$$\sigma^2 (\mu^2 - 2m\mu + m^2)$$

$$(n\tau^2 + \sigma^2)\mu^2 - \left(2\sum^n X_i \tau^2 + 2m\right)\mu + \cdots$$

$$\frac{\left(\mu - \dfrac{\sum^n X_i \tau^2 + m\sigma^2}{n\tau^2 + \sigma^2}\right)^2}{\dfrac{2\sigma^2\tau^2}{n\tau^2 + \sigma^2}}$$

$$\hat{\mu}_{Bayes} = \frac{\frac{1}{n}\sum X_i \tau^2 + \frac{1}{n} m\sigma^2}{\tau^2 + \frac{\sigma^2}{n}}$$

$$\mathrm{Var}(\mu \mid X) = \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}$$

Evaluating Point Estimators.

Mean Squared Error

$$\mathbb{E}_\theta (\hat{\theta} - \theta)^2 = \left(E_\theta(\hat{\theta} - \theta)\right)^2 + Var_\theta(\hat{\theta})$$

$$= B^2 + V$$

Log-likelihood: $LL(\theta) = \sum_i^n \log p(x_i; \theta)$

Score: $S(\theta) = \frac{\partial LL(\theta)}{\partial \theta}$

Fish Information: $I(\theta) = \mathbb{E}[\, s(\theta) s(\theta)^T \,]$

Score function has mean 0

proof:

$$E_\theta(S(\theta)) = \sum^n \int_x \nabla_\theta \log p(x_i;\theta)\, p(x_i;\theta)\, dx_i$$

$$= n \int \nabla_\theta \log p(x;\theta)\, p(x;\theta)\, dx$$

$$= \int \frac{\nabla_\theta p(x;\theta)}{p(x;\theta)} p(x;\theta)\, dx \quad \text{dominated convergence theorem}$$

$$= \nabla_\theta \int p(x;\theta) dx = \nabla_\theta 1 = 0$$

$$\boxed{I_1(\theta) = \mathbb{E}[-\nabla_\theta^2 \log p(x;\theta)]}$$

Proof:

$$\nabla_\theta^2 \log p(x;\theta) = \nabla_\theta \frac{\nabla_\theta p(x;\theta)}{p(x;\theta)}$$

$$= \frac{\nabla_\theta p(x;\theta)}{p(x;\theta)} - \frac{\nabla_\theta p(x;\theta) \nabla_\theta p(x;\theta)^T}{p(x;\theta)^2}$$

$$= \frac{\nabla_\theta p(x;\theta)}{p(x;\theta)} - s(\theta) s(\theta)^T$$

$$\mathbb{E}\nabla_\theta^2 \log p(x;\theta) = \mathbb{E}\frac{\nabla_\theta^2 p(x;\theta)}{p(x;\theta)} - \mathbb{E}\, s(\theta)s(\theta)^T$$
$$= -\mathbb{E}\, s(\theta)s(\theta)^T$$
$$= -I(\theta)$$

$$\boxed{\begin{array}{l}\text{Cramér-Rao Bound} \\ \quad \hat{\theta} \text{ unbiased estimater} \\ \quad\quad Var(\hat{\theta}) \geq \dfrac{1}{nI_1(\theta)}\end{array}}$$

Multivariate Generalization

$$Var(\hat{\theta}) \succeq I(\theta)^{-1}$$

KL loss: $KL(p(x;\theta), p(x;\alpha)) = E_{x\sim p(\theta)} \log\left(\frac{p(x;\theta)}{p(x;\alpha)}\right)$

Risk: $R(\theta, \hat{\theta}(X)) = E_{x\sim\theta}\, L(\theta, \hat{\theta}(x))$

Bayes Risk: $B_\pi(\hat{\theta}) = \int R(\theta,\hat{\theta})\pi(\theta)\,d\theta.$

$\alpha = \beta = \sqrt{\frac{n}{4}}$. $R(\hat{\theta}_B, \theta) = \frac{n}{4(n+\sqrt{n})^2}$

posterior risk

$\boxed{\hat{\theta}_{Bayes} \text{ minimize } r(\hat{\theta}|x^n)}$ $\quad r(\hat{\theta}|x^n) = \int L(\theta, \hat{\theta}(x^n))\pi(\theta|x^n)\,d\theta.$

Proof:
$$B_\pi(\hat{\theta}) = \int R(\theta,\hat{\theta})\pi(\theta)d\theta = \int\left(\int L(\theta,\hat{\theta})\, p(x|\theta)\,dx^n\right)\pi(\theta)\,d\theta$$
$$= \iint L(\theta,\hat{\theta}(x_n))\, \underline{\underline{p(x,\theta)}}\,dx^n\,d\theta$$
$$= \iint L(\theta, \hat{\theta}(x_n))\,\pi(\theta|x^n)\,m(x^n)\,dx^n\,d\theta$$
$$= \iint L(\theta, \hat{\theta}(x_n))\,\pi(\theta|x^n)\,d\theta\; m(x^n)\,dx^n$$
$$= \iint r(\hat{\theta}|x^n)\,m(x^n)\,dx^n$$

$\hat{\theta}_{Bayes}$ minimize $r(\hat{\theta}|x^n)$

If $L(\theta,\hat\theta)=|\theta-\hat\theta|^2$, then $\hat\theta_{Bayes}=\mathbb{E}(\hat\theta|X^n)$

$$\int |\theta-\hat\theta|^2 p(\theta|x)\,d\theta$$

$$2\int(\theta-\hat\theta)\,p(\theta|x)\,d\theta \Rightarrow \hat\theta=\int \theta p(\theta|x)d\theta = E_\theta(\theta|X^n)$$

$$\boxed{\text{The risk: } R(\theta,\hat\theta)=\mathbb{E}_{x\sim p(\theta)}L(\theta,\hat\theta)=\int_X L(\theta,\hat\theta(X_n))\,p(X_n;\theta)\,dX_n}$$

when $L(\theta,\hat\theta)$ is squared loss. MSE is the risk,

Example: comparing risk functions

$X_1,\cdots,X_n\sim Ber(p)$

$$\hat p_1=\bar X,\quad \hat p_2=\frac{n\bar X+\alpha}{n+\alpha+\beta}$$

$$r(p,\hat p_1)=Var(\bar X)=\frac{p(1-p)}{n}$$

$$r(p,\hat p_2)=\left(\frac{\alpha(1-p)-\beta p}{n+\alpha+\beta}\right)^2+\frac{np(1-p)}{(n+\alpha+\beta)^2}$$

Let $\alpha=\beta=\sqrt{\frac{n}{4}}$, $\quad R(p,\hat p_2)=\frac{n}{4(n+\sqrt n)^2}$

maximum risk: $\bar R(\hat\theta)=\max_\theta R(\theta,\hat\theta)$

Bayes risk: $B_\pi(\hat\theta)=\int R(\theta,\hat\theta)\,\pi(\theta)\,d\theta$

$$p(\theta|x)=\frac{p(x|\theta)\,\pi(\theta)}{p(x)},\quad \pi(\theta|X^n)=\frac{p(X^n|\theta)\,\pi(\theta)}{m(X^n)}$$

$m(X^n)$ → marginal distribution of X^n.

$$\boxed{\text{Posterior risk: } r(\hat\theta|x^n)=\int_\theta L(\theta,\hat\theta(x^n))\,\pi(\theta|x^n)\,d\theta}$$

※ The difference between $R(\theta,\hat\theta)$ and $r(\hat\theta|x^n)$

Bayes risk: $B_\pi(\hat\theta)=\int_\theta R(\theta,\hat\theta)\,\pi(\theta)\,d\theta$

relationship with posterior risk

$$=\int_\theta\int_x L(\theta,\hat\theta)\,p(x|\theta)\,dx\,\pi(\theta)\,d\theta$$

$$=\int_\theta\int_x L(\theta,\hat\theta)\,p(x)\,dx\,d\theta$$

$$= \int_\Theta \int_X L(\theta, \tilde{\theta})\, p(\theta|x)\, m(x)\, dx\, d\theta$$

$$= \int_X \int_\Theta L(\theta, \tilde{\theta})\, p(\theta|x)\, d\theta\; m(x)\, dx$$

$$= \int_X r(\theta|x^n)\, m(x^n)\, dx^n \quad ☆$$

Bayes estimator :	minimize Bayes risk (definition)
	minimize $r(\theta|x^n)$ posterior risk

$\hat{\theta}_{Bayes}$ $\quad L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, $\quad r(\hat{\theta}|x^n) = \int (\theta - \hat{\theta})^2 \pi(\theta|x^n)\, d\theta$

$$\nabla r(\hat{\theta}|x^n) = \int -2(\theta - \hat{\theta})\pi(\theta|x^n)\, d\theta = 0$$

$$\hat{\theta} = \mathbb{E}(\theta|x^n)$$

$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ $\quad r(\hat{\theta}|x^n) = \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta)\pi(\theta|x^n)\, d\theta$

$$+ \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta})\pi(\theta|x^n)\, d\theta$$

$$\nabla r(\hat{\theta}|x^n) = 0 \Rightarrow \int_{-\infty}^{\hat{\theta}} \pi(\theta|x^n)\, d\theta = \int_{\hat{\theta}}^{\infty} \pi(\theta|x^n)\, d\theta = \frac{1}{2}$$

$$\hat{\theta} = \text{median}(\theta|x^n)$$

$L(\theta, \hat{\theta}) = \mathbb{1}(\theta \neq \hat{\theta})$ $\quad r(\hat{\theta}|x^n) = \int \mathbb{1}(\theta \neq \hat{\theta})\pi(\theta|x^n)\, d\theta$

$$= 1 - \pi(\hat{\theta}|x^n)\, d\theta$$

$$\hat{\theta} = \text{mode}(\theta|x^n)$$

Minimax Estimator through Bayes Estimator

→ Bounding the Minimax Risk.

$$B_\pi(\hat{\theta}_{Bayes}) \le B_\pi(\hat{\theta}_{minimax}) \le \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \tilde{\theta}) \le \sup_{\theta \in \Theta} R(\theta, \hat{\theta}_{Bayes})$$

Example : $X_1, \dots, X_n \sim N(\theta, I_d)$

$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$ show that $\hat{\theta}$ is minimax estimator.

$$\hat{\theta} \sim N(\theta, \frac{I_d}{n})$$

$$R(\theta,\hat{\theta}) = \mathbb{E}\left[\sum^{d}(\hat{\theta}_i - \theta_i)^2\right] = \mathbb{E}\left[\sum^{d} Z_i^2\right].$$

$$Z_i \sim N(0, \frac{1}{n})$$

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} R(\theta,\hat{\theta}) \le R(\theta,\hat{\theta}) = \frac{d}{n}$$

take prior $\pi = N(0, c^2 I_d)$?

→ Least Favorable Prior

$R(\theta, \hat{\theta}) \le B_\pi(\hat{\theta})$ for all θ

π is least favorable prior, $\hat{\theta}$ is minimax

MLE Asymptotics

Consistency: $\hat{\theta}_{MLE} \xrightarrow{P} \theta$

Asymptotic distribution $\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} N(0, \frac{1}{I(\theta)})$

mle. maximize $\log L(\theta)$

equals minimize $\frac{1}{n}\sum^{n} \frac{\log p(X;\theta)}{\log p(X;\tilde{\theta})}$ → emperical risk

population risk ← $\underbrace{E_\theta \frac{\log p(X;\theta)}{\log p(X;\tilde{\theta})}}$ KL divergence.

Conditions for consistency:

1. Identifiability: if $\theta_1 \ne \theta_2$, then $p(X;\theta_1) \ne p(X;\theta_2)$

2. Strong identifiability: $\forall \epsilon > 0$, $\inf_{\tilde{\theta}: |\tilde{\theta}-\theta| \ge \epsilon} KL(p(X;\theta) \| p(X;\tilde{\theta})) > 0$

3. Uniform LLN: $\sup_{\tilde{\theta}} |R_n(\theta,\tilde{\theta}) - R(\theta,\tilde{\theta})| \xrightarrow{P} 0$

Asymptotic: $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0, \frac{1}{I(\theta)})$

proof:

$$0 = LL'(\hat{\theta}) = LL'(\theta) + (\hat{\theta}-\theta)LL''(\tilde{\theta})$$

$$\hat{\theta}-\theta = -\frac{LL'(\theta)}{LL''(\tilde{\theta})}$$

$$\sqrt{n}(\hat{\theta}-\theta) = -\frac{\frac{LL'(\theta)}{\sqrt{n}}}{\frac{LL''(\tilde{\theta})}{n}}$$

numerator: $\frac{LL'(\theta)}{\sqrt{n}} = \sqrt{n} \times \frac{1}{n}\sum \nabla_\theta \log p(x_i;\theta)$

$$= \sqrt{n} \times \left(\frac{1}{n}\sum^n \nabla_\theta \log p(x_i;\theta) - \mathbb{E}\nabla_\theta \log p(x_i;\theta)\right)$$

by CLT, and $\mathbb{E}(S(\theta))\Big|_\theta = 0$ $\xrightarrow{d} N(0, Var(S(\theta))) \xrightarrow{d} N(0, I(\theta))$

denominator: $-\frac{LL''(\tilde{\theta})}{n} = -\frac{1}{n}\sum^n \nabla_\theta^2 \log p(x_i;\tilde{\theta}) \xrightarrow{d} I(\theta)$

$\hat{\theta} \xrightarrow{P} \theta$

by Slustsky's

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} \frac{1}{I(\theta)} N(0, I(\theta)) \xrightarrow{d} N(0, \frac{1}{I(\theta)})$$

Example: $x_1, \dots, x_n \sim Exp(\theta)$

$$p(x) = \theta e^{-\theta x}$$

$$LL(\theta) = n\log\theta - \theta\sum^n X_i$$

$$S(\theta) = \frac{n}{\theta} - \sum^n X_i$$

$$I(\theta) = E\left[-\,-\frac{n}{\theta^2}\right] = \frac{n}{\theta^2}$$

$\hat{\theta}_{MLE} = \frac{\sum^n X_i}{n}$, $\sqrt{n}(\hat{\theta}-\theta) \to N(0, \frac{\theta^2}{n})$

Counterexample:

Unifor distribution.

Definitions: minimal representation: no set of $a \in \mathbb{R}^s$

$$\sum_{i=1}^{s} a_i T_i(x) = \text{const. for all } x \in \mathcal{X}$$

Non-minimal exponential families are over-complete.

not identifiable.

full-rank exponential family: space of θ_i is s-dim.

↳ sufficient statistics are minimal

Minimal Sufficiency:

$T(x_1, \cdots, x_n)$ is sufficient, and for any other sufficient statistic $S(x_1, \cdots, x_n)$. we can write $T(x_1, \cdots, x_n) = g(S(x_1, \cdots x_n))$

Condition:

$$R(x_1, \cdots, x_n, y_1, \cdots y_n; \theta) = \frac{p(x_1, \cdots x_n; \theta)}{p(y_1, \cdots y_n; \theta)}$$

does not depend on θ

iff $T(x_1, \cdots, x_n) = T(y_1, \cdots, y_n)$

Inconsistency of the MLE:

- not identifiable
- the parameter space is two large. fail of Uniform LLN

MLE under misspecification

when q does not belong to our model Θ

$KL(q \| p_{\hat{\theta}_{MLE}}) \leq KL(q \| p_\theta)$ for all $\theta \in \Theta$

MLE is estimating the KL projection of q onto our model.