Part II: Principle of Statistics - Revision Notes

Lectures by , notes by Zhimei Liu

1. Likelihood principle

1.1. Basic ideas and concepts

Definition: (*Likelihood function*) Let $\{f(\cdot,\theta):\theta\in\Theta\}$ be a statistical model of pfd/pmf $f(x,\theta)$ for the distribution P of a random variable X and consider iid copies $X_1, X_2, ..., X_n$ of X. Then likelihood function of the model is:

$$L_n(\theta)=\prod_{i=1}^{n} f(x_i,\theta)$$

Log-likelihood function:

$$\ell_n(\theta)=\sum_{i=1}^{n} log(f(x_i,\theta))$$

Normalised log-likelihood function:

$$\bar{\ell}_n(\theta)=\frac{1}{n}\ell_n(\theta)$$

Definition: (*Maximum likelihood estimator*)

$$\hat{\theta}_{MLE}=argmax_{\theta\in\Theta}\ L_n(\theta)$$

Definition: (*Score function*) For $\Theta\subseteq\mathbb{R}^p$,

$$S_n(\theta)=\nabla_\theta\ell_n(\theta)=[\frac{\partial\ell_n(\theta)}{\partial\theta_1},...,\frac{\partial\ell_n(\theta)}{\partial\theta_p}]^T$$

One of the main uses of the score function is to look for the MLE as a solution to $S_n(\hat{\theta})=0$

1.2. Information geometry and likelihood function

Theorem 1.1. Model $\{f(\cdot,\theta):\theta\in\Theta\}$, a variable $X\sim P$ s.t. $\mathbb{E}[|log(f(X,\theta)|]<\infty$, if the model is well specified with $f(x,\theta_0)$ as pdf of P, the function

$$\ell(\theta)=\mathbb{E}_{\theta_0}[log(f(X,\theta))]$$

is maximized at θ_0.

Proof. Compare $\ell(\theta)-\ell(\theta_0)$ with 0 and apply Jensen's inequality with concave function ϕ: $\mathbb{E}[\phi(Z)]\le\phi(\mathbb{E}[Z])$. □

Definition: (*Kullback-Leibler divergence*)

$$\ell(\theta_0)-\ell(\theta)=KL(P_{\theta_0},P_\theta)=\int_{\mathcal{X}} f(x,\theta_0)\log\frac{f(x,\theta_0)}{f(x,\theta)}dx$$

This quantity can be thought of as a 'distance' between distributions.

Theorem 1.2. $\{f(\cdot,\theta):\theta\in\Theta\}$ regular enough that integration and differentiation can be exchanged, then $\forall\ \theta\in int(\Theta)$,

$$\mathbb{E}_\theta[\nabla_\theta logf(X,\theta)]=0$$

Proof. By expanding the expectation and changing integration and differentiation.

Definition: (*Fisher information matrix*)

$$I(\theta)=\mathbb{E}_\theta[\nabla_\theta\log\ f(X,\theta)\ \nabla_\theta\log\ f(X,\theta)^T]$$

$$I_{ij}(\theta)=\mathbb{E}_{ij}[\frac{\partial}{\partial\theta_i}\log f(X,\theta)\frac{\partial}{\partial\theta_j}\log f(X,\theta)]$$

In one dimension,

$$I(\theta)=\mathbb{E}_\theta[(\frac{d}{d\theta}\log f(X,\theta))^2]=\mathrm{Var}_\theta[\frac{d}{d\theta}\log f(X,\theta)]$$

Theorem 1.3.

$$I(\theta)=-\mathbb{E}_\theta[\nabla^2_\theta\log f(X,\theta)]$$

$$I_{ij}(\theta)=-\mathbb{E}_\theta[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X,\theta)]$$

In one dimension,

$$I(\theta)=\mathrm{Var}_\theta[\frac{d}{d\theta}\log f(X,\theta)]=-\mathbb{E}_\theta[\frac{d^2}{d\theta^2}\log f(X,\theta)]$$

Definition: For a random vector $X=(X_1,...,X_n)\in\mathbb{R}^n$, the fisher information matrix is

$$I_\theta(\theta)=\mathbb{E}_\theta[\nabla_\theta\log f(X_1,...,X_n,\theta)\nabla_\theta\log f(X_1,...,X_n,\theta)^T]$$

Proposition 1.1. For a random vector $X=(X_1,...,X_n)$, fisher information tensorizes,

$$I_n(\theta)=nI(\theta)$$

where $I(\theta)$ is the Fisher information for one copy X_i.

Theorem 1.4. (Cramer-Rao lower bound) $\tilde{\theta}=\tilde{\theta}(X_1,...,X_n)$ is an unbiased estimator.

$$\mathrm{Var}_\theta(\tilde{\theta})=\mathbb{E}_\theta[(\tilde{\theta}-\theta)^2]\ge\frac{1}{nI(\theta)}$$

Proof. Apply Cauchy-Schwartz inequality:

$$\mathrm{Cov}(Y,Z)^2\le\mathrm{Var}(Y)\mathrm{Var}(Z)$$

and taking $Y=\tilde{\theta}$ and $Z=\frac{d}{d\theta}\log f(X,\theta)$. Then true for $n=1$.

Proposition 1.2. Let $\tilde{\Phi}$ be an unbiased estimator of $\Phi(\theta)$. Then

$$\mathrm{Var}_\theta(\tilde{\Phi}) \geq \frac{1}{n} \nabla_\theta \Phi(\theta)^T I^{-1}(\theta) \nabla_\theta \Phi(\theta)$$

2. Asymptotic theory for MLE

2.1. Stochastic convergence concepts

Let $(X_n)_{n\geq 0}$, X be random vectors in $\mathbb{R}^k$.

Definition: X_n converges to X **almost surely**, or $X_n \to^{a.s.} X$ as $n \to \infty$, if

$$\mathbb{P}(||X_n - X|| \to 0 \text{ as } n \to \infty) = 1$$

Definition: X_n converges to X **in probability**, or $X_n \to^P X$ as $n \to \infty$, if $\forall \epsilon > 0$

$$\mathbb{P}(||X_n - X|| \geq \epsilon) \to 0$$

Definition: X_n converges to X **in distribution**, or $X_n \to^d X$ as $n \to \infty$, if $\forall\ t$

$$\mathbb{P}(X_n \preccurlyeq t) \to \mathbb{P}(X \preccurlyeq t)$$

Proposition 2.1. almost surely $\implies$ in probability $\implies$ in distribution

Proposition 2.2. (Continuous mapping theorem) For $g : \mathcal{X} \to \mathbb{R}$ continuous, have

$$X_n \to^{a.s./P/d} X \quad \implies g(X_n) \to^{a.s./P/d} g(X)$$

Proposition 2.3. (Slusky's lemma)

Proposition 2.5. (Weak law of large numbers) Let $X_1, ..., X_n$ be iid copies of X with $\mathrm{Var}(X) < \infty$. Then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \to^P \mathbb{E}[X]$$

Proof. Use Chebyshev's inequality to $Z_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X])$.

Proposition 2.5. (Strong law of large numbers) Let $X_1, ..., X_n$ be iid copies of $X \sim P$ with $\mathbb{E}[||X||] < \infty$. Then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \to^{a.s.} \mathbb{E}[X]$$

Proof. Use Chebyshev's inequality to $Z_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X])$.

2.2. Law of large numbers and CLT

Theorem 2.2. (Central limit theorem) Let $X_1, ..., X_n$ be iid copies of $X \sim P$ on $\mathbb{R}$ and assume $\mathrm{Var}(X) = \sigma^2 < \infty$. As $n \to \infty$,

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X]) \to^d \mathcal{N}(0, \sigma^2)$$