Principle of Statistics

0 Introduction

- distribution
- p.m.f.
- p.d.f.
- samples
- sample size
- statistical model $\{f(\theta, \cdot)\}$
- law
- parameter space Θ
- correctly specified

Goal.

(i) *Estimation*

(ii) *Testing Hypothesis*

(iii) *Inference*

- estimator
- test
- confidence

1 Likelihood Principle

Setting 1. $\{f(\cdot, \theta) : \theta \in \Theta\}$ *statistical model,* X_i *i.i.d. copy of* X

- likelihood function $L_n(\theta) = \prod f(x_i, \theta)$
- log-likelihood function $l_n(\theta) = \log L_n(\theta)$
- normalized log-likelihood function $\bar{l}_n(\theta) = \frac{1}{n} l_n(\theta)$
- maximum likelihood estimator (MLE) $\hat{\theta} = \hat{\theta}_{MLE}$

– score function $S_n(\theta) = \nabla_\theta l_n(\theta)$

Fact. $S_n(\hat{\theta}) = 0$

Setting 2. *model* $\{f(\cdot, \theta)\}$, $X \sim P$

– $l(\theta) = \mathbb{E}_{\theta_0}(\log(f(X, \theta)))$

Theorem 1.1. $\mathbb{E}|\log(f(X, \theta))| < \infty$, *well specified with* $f(x, \theta_0)$, *then* $l(\theta)$ *maximised at* θ_0

– sample approximation $\bar{l}_n(\theta) = \frac{1}{n}\sum \log(f(x_i, \theta))$

– strict identifiability —— $f(\cdot, \theta) = f(\cdot, \theta') \iff \theta = \theta'$

Fact. *With strict identifiability, maximizer unique hence must be the true value* θ_0

– Kullback-Leibler divergence $KL(P_{\theta_0}, P_\theta) = l(\theta_0) - l(\theta)$

Setting 3. *regular —— integration and differentiation can be interchanged*

Theorem 1.2. *regular, then* $\forall \theta \in int(\Theta)$, $\mathbb{E}[\nabla_\theta \log(f(X, \theta))] = 0$

Fact. $\mathbb{E}_{\theta_0}[\nabla_\theta \log(f(X, \theta))] = 0$

– Fisher information matrix $I(\theta) = \mathbb{E}_\theta[\nabla_\theta \log f(X, \theta) \nabla_\theta \log f(X, \theta)^\top]$

Fact. *1-d case,* $I(\theta) = \mathbb{E}\left[(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X, \theta))^2\right] = Var_\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X, \theta)\right]$

Theorem 1.3. *regularity assumptions,* $\forall \theta \in int(\Theta)$, $I(\theta) = -\mathbb{E}_\theta[\nabla_\theta^2 \log f(X, \theta)]$

Fact. *1-d case, relation between variance of score and curvature of* l

– $I_n(\theta) = \mathbb{E}[\nabla_\theta \log f(X_1, \ldots, X_n, \theta) \nabla \log f(X_1, \ldots, X_n, \theta)^\top]$

Proposition 1.4 (Tensorize)**.** X_i *i.i.d* , $I_n(\theta) = nI(\theta)$

Theorem 1.5 (Cramer-Rao lower bound (1-d))**.** *model* $\{f(\cdot, \theta)\}$, *regular,* $\Theta \subset \mathbb{R}$, *unbiased estimator* $\tilde{\theta}(X_1, \ldots, X_n)$, *then* $\forall \theta \in int(\Theta)$, $Var_\theta(\tilde{\theta}) = \mathbb{E}[(\tilde{\theta} - \theta)^2] \geq \frac{1}{nI(\theta)}$

Corollary 1.6. $Var_\theta(\tilde{\theta}) \geq \frac{(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta(\tilde{\theta}))^2}{nI(\theta)}$

Proposition 1.7. Φ *differentiable functional,* $\tilde{\Phi}$ *unbiased estimator of* $\Phi(\theta)$, *then* $\forall \theta \in int(\Theta)$, $Var_\theta(\tilde{\Phi}) \geq \frac{1}{n}\nabla_\theta \Phi(\theta)^\top I^{-1}(\theta) \nabla_\theta \Phi(\theta)$

Fact. $Var_\theta(\alpha^\top \tilde{\theta}) \geq \frac{1}{n}\alpha^\top I^{-1}(\theta)\alpha$

Fact. $Cov_\theta(\tilde{\theta}) \succeq \frac{1}{n}I^{-1}(\theta)$ *(positive semi-definite)*

2 Asymptotic Theory for MLE

- convergence almost surely
- convergence in probability
- convergence in distribution

Proposition 2.1. *convergence a.s. $\Rightarrow$ in prob $\Rightarrow$ in distribution*

Proposition 2.2 (Continuous mapping theorem)**.** *g continuous, then $X_n \xrightarrow{a.s./P/d} X \Rightarrow g(X_n) \xrightarrow{a.s./P/d} g(X)$*

Proposition 2.3 (Slutsky's lemma)**.** *$X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} c$ deterministic, then*

(i) $Y_n \xrightarrow{P} c$

(ii) $X_n + Y_n \xrightarrow{d} X + c$

(iii) $X_n Y_n \xrightarrow{d} cX$

(iv) $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$ *if* $c \neq 0$

Random matrices $(A_n)_{ij} \xrightarrow{P} A_{ij}$ deterministic, then

(i) $A_n X_n \xrightarrow{d} AX$

- bounded in probability $O_P(1)$ —— $\forall \epsilon > 0, \exists M(\epsilon), \sup_n \mathbb{P}(\|X_n\| > M(\epsilon)) < \epsilon$

Proposition 2.4. *$X_n \xrightarrow{d} X$, then (X_n) bounded in probability*

Proposition 2.5 (Weak law of large numbers)**.** *X_i i.i.d. , $Var(X) < \infty$ (unnecessary), then $\bar{X}_n = \frac{1}{n}\sum X_i \xrightarrow{P} \mathbb{E}(X)$*

Theorem 2.6 (Strong law of large numbers)**.** *X_i i.i.d. ,$\mathbb{E}|X| < \infty$, then $\bar{X}_n \xrightarrow{a.s.} \mathbb{E}(X)$*

Theorem 2.7 (Central limit theorem(1-d))**.** *X_i i.i.d. , $Var(X) = \sigma^2 < \infty$, then $\sqrt{n}(\bar{X}_n - \mathbb{E}(X)) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$*

- $\mathcal{N}(\mu, \Sigma)$ —— p.d.f. $\frac{1}{(2\pi)^{k/2}|\det(\Sigma)|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$

Fact. *$X \sim \mathcal{N}(\mu, \Sigma)$, then $\alpha^\top X \sim \mathcal{N}(\alpha^\top \mu, \alpha^\top \Sigma \alpha)$*

Proposition 2.8. *$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top)$*

Proposition 2.9. *Σ diagonal, $X_{(j)}$ independent*

Theorem 2.10 (Central limit theorem(n-d))**.** *X_i i.i.d. , $Cov(X) = \Sigma$ positive definite, then $\sqrt{n}\left(\bar{X}_n - \mathbb{E}(X)\right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$*

- asymptotic efficiency —— $nVar_{\theta_0}(\tilde{\theta}_0) \to I^{-1}(\theta_0)$

Fact. *Under suitable assumptions,* $\theta_{MLE} \approx \mathcal{N}(\theta, I^{-1}(\theta_0)/n)$

Example (Confidence interval)**.**

- *confidence region* $\mathcal{C}_n = \left\{|\mu - \bar{X}| \leq \frac{\sigma z_\alpha}{\sqrt{n}}\right\}$
- *asymptotic level* $1-\alpha$ *confidence set*

Setting 4. X_i *i.i.d. , arising from* $\{P_\theta\}$

- consistency —— $\tilde{\theta}_n \xrightarrow{P_\theta} \theta_0$

Assumption 1 (Usual regularity assumptions)**.** $\{f(\cdot,\theta)\}$ *statistical model of p.d.f. or p.m.f. st*

(i) $f(x,\theta) > 0$

(ii) $\int_{X f(x,\theta)} dx = 1$

(iii) $f(x,\cdot)$ *continuous*

(iv) Θ *compact*

(v) $f(\cdot,\theta) = f(\cdot,\theta') \Rightarrow \theta = \theta'$

(vi) $\mathbb{E}_\theta \sup_\theta |\log f(X,\theta)| < \infty$

Theorem 2.11 (Consistency of the MLE)**.** *Usual regularity assumptions,* X_i *i.i.d. , then*

(i) *MLE exists*

(ii) *MLE consistent*

Fact. *proof can be simplified when* l_n *differentiable, in this case* Θ *compact not needed*

Theorem 2.12 (Uniform law of large numbers)**.** Θ *compact,* $q(x,\cdot)$ *continuous,* $\mathbb{E}\sup_\Theta |q(X,\theta)| < \infty$*, then* $\sup_\Theta |\frac{1}{n}\sum q(X_i,\theta) - \mathbb{E}(q(X,\theta))| \xrightarrow{a.s.} 0$

Assumption 2. *In addition to usual regularity assumption,*

(i) *true* $\theta_0 \in int(\Theta)$

(ii) $\exists U$ *open nbhd of* θ_0 *st* $f(x,\cdot) \in C^2$

(iii) $I(\theta_0)$ *non-singular,* $\mathbb{E}_{\theta_0}\|\nabla_\theta \log f(X,\theta_0)\| < \infty$

(iv) $\exists K \subset U$ *compact, non-empty interior containing* θ_0 *st*

$$\mathbb{E}_{\theta_0} \sup_K \left\|\nabla_\theta^2 \log f(X,\theta)\right\| < \infty$$
$$\int_X \sup_K \|\nabla_\theta \log f(X,\theta)\| dx < \infty$$
$$\int_X \sup_K \left\|\nabla_\theta^2 \log f(X,\theta)\right\| dx < \infty$$

Theorem 2.13. *Further usual assumption,* $\hat{\theta_n}$ *MLE of i.i.d.* $X_i \sim P_{\theta_0}$*, then* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$

- asymptotic efficiency —— $nVar_{\theta_0}(\tilde{\theta}_n) \to I(\theta_0)^{-1}$
- Hodge estimator —— $\tilde{\theta}_n = \begin{cases} \hat{\theta}_n & \text{if } |\hat{\theta}_n| > n^{-1/4} \\ 0 & \text{otherwise} \end{cases}$
- profile likelihood $L^{(p)}(\theta_1) = \sup_{\Theta_2} L((\theta_1, \theta_2))$
- plug-in MLE $\Phi(\hat{\theta}_{MLE})$

Fact. *under new parametrization* $\{f(\cdot, \phi) : \phi = \Phi(\theta)\}$, $\hat{\phi}_{MLE} = \Phi(\hat{\theta}_{MLE})$

Theorem 2.14 (Delta method)**.** $\Phi \in C^1$ *at* θ_0, $\nabla_\theta \Phi(\theta_0) \neq 0$, *let* $(\hat{\theta}_n)$ *st* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z$, *then* $\sqrt{n}(\Phi(\hat{\theta}_n) - \Phi(\theta_0)) \xrightarrow{d} \nabla_\theta \Phi(\theta_0)^\top Z$

Fact. *if* $\hat{\theta}_n$ *MLE with asymptotic normality, then* $\sqrt{n}(\Phi(\hat{\theta}_n) - \Phi(\theta_0)) \xrightarrow{d} \mathcal{N}(0, \nabla_\theta \Phi(\theta_0)^\top I^{-1}(\theta_0) \nabla_\theta \Phi(\theta_0))$

Fact. *plug in MLE asymptotically efficient*

- observed Fisher information $i_n(\theta) = \frac{1}{n} \sum \nabla_\theta \log f(X_i, \theta) \nabla_\theta \log f(X_i, \theta)^\top$
- $\hat{i}_n = i_n(\hat{\theta}_{MLE})$

Proposition 2.15. *Under further assumption,* $\hat{i}_n \xrightarrow{P_{\theta_0}} I(\theta_0)$

- $j_n(\theta) = -\frac{1}{n} \sum \nabla^2_\theta \log f(X_i, \theta)$
- $\hat{j}_n = j_n(\hat{\theta}_{MLE})$
- Wald statistic $W_n(\theta) = n(\hat{\theta}_{MLE} - \theta)^\top \hat{i}_n (\hat{\theta}_{MLE} - \theta)$
- ξ_α —— $\mathbb{P}(\chi^2_p \leq \xi_\alpha) = 1 - \alpha$

Proposition 2.16 (Confidence ellipsoids)**.** *Under further assumption, define* $\mathcal{C}_n = \{\theta : W_n(\theta) \leq \xi_\alpha\}$, *then* $\mathcal{C}_n$ α*-level asymptotic confidence region*

Setting 5. *hypothesis testing:* $\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta \backslash \Theta_0 \end{cases}$

- decision rule ψ_n
- type-one error (false positive) —— $\mathbb{P}_\theta(reject\ H_0) = \mathbb{E}_\theta(\psi_n)$ for $\theta \in \Theta_0$
- type-two error (false negative) —— $\mathbb{P}_\theta(accept\ H_0) = \mathbb{E}_\theta(1 - \psi_n)$ for $\theta \in \Theta_1$
- likelihood ratio test $\Lambda_n(\Theta, \Theta_0) = 2 \log \frac{\sup_\Theta \prod f(X_i, \theta)}{\sup_{\Theta_0} \prod f(X_i, \theta)} = 2 \log \frac{\prod f(X_i, \hat{\theta}_{MLE})}{\prod f(X_i, \hat{\theta}_{MLE,0})}$

Theorem 2.17 (Wilks theorem)**.** *Under further assumption, hypothesis test with* $\Theta_0 = \{\theta_0\}$, $\theta_0 \in int(\Theta)$, *then* $\Lambda_n(\Theta, \Theta_0) \xrightarrow{d} \chi^2_p$

Fact. *test* $\psi_n = \mathbb{1}\{\Lambda_n(\Theta, \Theta_0) \geq \xi_\alpha\}$ *controls type-one error at symptotic level* $1 - \alpha$

Fact. Θ_0 *dimension* $p_0 < p$, *then* $\Lambda_n(\Theta, \Theta_0) \xrightarrow{d} \chi^2_{p - p_0}$

3 Bayesian Inference

Setting 6. $\mathcal{X}$ *sample space, probability measure* $Q(x, \theta) = f(x, \theta)\pi(\theta)$

- prior distribution π
- posterior distribution $\Pi(\theta|X)$
- conjugate prior —— $\pi(\theta)$ and $\Pi(\theta|X)$ same family of distributions

Example.

(i) *normal prior, normal sampling, normal posterior*

(ii) *Beta prior, binomial sampling, Bata posterior*

(iii) *Gamma prior, Poisson sampling, Gamma posterior*

- improper prior —— infinite integral over Θ
- Jeffreys prior —— $\pi(\theta)$ proportional to $\sqrt{\det I(\theta)}$

Goal.

(i) *Estimation*

(ii) *Uncertainty Quantification*

(iii) *Hypothesis Testing*

- posterior mean $\bar{\theta}$ —— $\bar{\theta}(X_1, \ldots, X_n) = \mathbb{E}_\Pi(\theta|X_1, \ldots, X_n)$
- credible set $\mathcal{C}_n$ —— $\Pi(\mathcal{C}_n|X_1, \ldots, X_n) = 1 - \alpha$
- Bayes factor —— $\frac{\mathbb{P}(X_1,\ldots,X_n|\Theta_0)}{\mathbb{P}(X_1,\ldots,X_n|\theta_1)} = \frac{\Pi(\Theta_0|X_1,\ldots,X_n)}{\Pi(\Theta_1|X_1,\ldots,X_n)}$