

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 $\forall \theta \in \text{int}(\Theta)$, $\mathbb{E}[\nabla_\theta \log(f(X, \theta))] = 0$

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

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

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

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

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 $\tilde{\theta}(X_1, \dots, X_n)$, then $\forall \theta \in \text{int}(\Theta)$, $\text{Var}_\theta(\tilde{\theta}) = \mathbb{E}[(\tilde{\theta} - \theta)^2] \geq \frac{1}{nI(\theta)}$

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

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

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

Fact. $\text{Cov}_\theta(\tilde{\theta}) \succeq \frac{1}{n} I^{-1}(\theta)$ (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

$$\sqrt{n}(\bar{X}_n - \mathbb{E}(X)) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

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

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

$$\sqrt{n}(\bar{X}_n - \mathbb{E}(X)) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

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

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 i.e. $\tilde{\theta}_{MLE} \xrightarrow{P_\theta} \theta_0$

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 $\sup_\Theta \left| \frac{1}{n} \sum q(X_i, \theta) - \mathbb{E}(q(X, \theta)) \right| \xrightarrow{a.s.} 0$

Assumption 2. In addition to usual regularity assumption,

- (i) true $\theta_0 \in \text{int}(\Theta)$
- (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

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

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 — $nVar_{\theta_0}(\tilde{\theta}_n) \rightarrow I(\theta_0)^{-1}$
- 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, $\hat{i}_n \xrightarrow{P_{\theta_0}} I(\theta_0)$

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

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: $\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta \setminus \Theta_0 \end{cases}$

- decision rule ψ_n
- type-one error (false positive) — $\mathbb{P}_{\theta}(\text{reject } H_0) = \mathbb{E}_{\theta}(\psi_n)$ for $\theta \in \Theta_0$
- type-two error (false negative) — $\mathbb{P}_{\theta}(\text{accept } H_0) = \mathbb{E}_{\theta}(1 - \psi_n)$ 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 $\Theta_0 = \{\theta_0\}$, $\theta_0 \in \text{int}(\Theta)$, then $\Lambda_n(\Theta, \Theta_0) \xrightarrow{d} \chi_p^2$

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

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

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

Fact. Bayesian inference not based on asymptotic distribution, but posterior distribution

- credible set — $\mathcal{C}_n = \left\{ |\nu - \hat{\theta}_n| \leq \frac{R_n}{\sqrt{n}} \right\}$ st $\Pi(\mathcal{C}_n|X_1, \dots, X_n) = 1 - \alpha$
- $\phi_n \sim \mathcal{N}\left(\hat{\theta}_n, \frac{I(\theta_0)^{-1}}{n}\right)$

Theorem 3.1 (Bernstein-von Mises). Under further assumptions, prior with continuous density π at θ_0 , $\pi(\theta_0) > 0$, then $\|\Pi_n - \phi_n\|_{L^1} = \int_{\Theta} |\Pi_n(\theta) - \phi_n(\theta)| d\theta \xrightarrow{a.s.} 0$

Fact. $\Pi_n(A) - \phi_n(A) \rightarrow 0$, so $\phi_n(\mathcal{C}_n) \rightarrow 1 - \alpha$

- $\Phi_0(t) = \mathbb{P}(|Z_0| \leq t)$ — $Z_0 \sim \mathcal{N}(0, I(\theta_0)^{-1})$

Lemma 3.2. Under assumptions, $R_n \xrightarrow{a.s.} \Phi_0^{-1}(1 - \alpha)$

Theorem 3.3. Under assumptions, $\mathbb{P}_{\theta_0}(\theta_0 \in \mathcal{C}_n) \rightarrow 1 - \alpha$

Fact. similar result with posterior mean $\bar{\theta}_n$ instead of $\hat{\theta}_n$

4 Decision Theory

Setting 7. *sample space* \mathcal{X}

- decision problems
- action space \mathcal{A}
- decision rules δ — $\delta : \mathcal{X} \rightarrow \mathcal{A}$
- loss function L — $L : \mathcal{A} \times \Theta \rightarrow [0, \infty)$

Example.

- *hypothesis testing* — $\mathcal{A} = \{0, 1\}$, $\delta(X)$ *test*
- *estimation problem* — $\mathcal{A} = \Theta$, $\delta(X) = \hat{\theta}(X)$
- *inference problem* — $\mathcal{A} = \mathcal{B}(\Theta)$, $\delta(X) = \mathcal{C}(X)$
- misclassification error — $L(a, \theta) = \mathbb{1}_{\{a \neq \theta\}}$
- absolute error — $L(a, \theta) = |a - \theta|$
- squared error — $L(a, \theta) = |a - \theta|^2$
- risk / average loss $R(\delta, \theta) = \mathbb{E}_\theta(L(\delta(X), \theta)) = \int_{\mathcal{X}} L(\delta(x), \theta) f(x, \theta) dx$
- quadratic risk / mean squared error (MSE) $\mathbb{E}_\theta[(\delta(X) - \theta)^2]$
- π -Bayes risk $R_\pi(\delta) = \mathbb{E}_\pi[R(\delta, \theta)] = \int_{\Theta} R(\delta, \theta) \pi(\theta) d\theta$ — prior π
- π -Bayes decision rule δ_π — minimizer of $R_\pi(\delta)$
- posterior risk R_Π — $R_\Pi(\delta) = \mathbb{E}_\Pi[L(\delta(x), \theta) | x]$, expectation over θ
- δ_Π minimise R_Π — $\mathbb{E}_\Pi[L(\delta_\Pi(x), \theta)] \leq \mathbb{E}_\Pi[L(\delta(x), \theta)]$ for all x

Proposition 4.1. δ minimizes $R_\Pi \Rightarrow$ minimizes R_π

Fact. For quadratic risk, $\delta_\Pi(X) = \mathbb{E}_\Pi[\theta | X]$

- unbiased decision rule — $\mathbb{E}_\theta[\delta(X)] = \theta$
- $Q(x, \theta) = f(x, \theta) \pi(\theta)$

Proposition 4.2. δ unbiased, π -Bayes rule under quadratic risk, then $\mathbb{E}_Q[(\delta(X) - \theta)^2] = 0$

Fact. unbiased estimator typically disjoint from Bayes estimators

- prior λ least favorable — $R_\lambda(\delta_\lambda) \geq R_{\lambda'}(\delta_{\lambda'})$ for all prior λ'
- maximal risk $R_m(\delta, \Theta) = \sup_{\Theta} R(\delta, \theta)$
- minimax risk $\inf_{\delta} R_m(\delta, \Theta)$
- minimax — δ attain minimax risk

Proposition 4.3. any prior λ , δ then $R_\lambda(\delta) \leq R_m(\delta, \Theta)$

Proposition 4.4. λ prior, δ_λ Bayes rule, $R_\lambda(\delta_\lambda) = R_m(\delta_\lambda, \Theta)$, then

- (i) δ_λ minimax
- (ii) if δ_λ unique Bayes rule, then unique minimax
- (iii) prior λ least favorable

Corollary 4.5. Bayes rule δ_λ constant risk in θ , then minimax

- δ inadmissible — $\exists \delta'$ st $R(\delta', \theta) \leq R(\delta, \theta)$ for all θ , strict inequality for some θ

Proposition 4.6.

- (i) unique Bayes rule admissible
- (ii) δ admissible, constant risk, then minimax

Proposition 4.7. $X_i \sim \mathcal{N}(\theta, \sigma^2)$ i.i.d., known σ^2 , then $\theta_{MLE} = \bar{X}_n$ admissible, minimax in quadratic risk

Fact. all minimax rules are limits of Bayes rule (dimension $p = 1, 2$, false for $p \geq 3$)

- James-Stein estimator $\delta^{JS}(X) = \left(1 - \frac{p-2}{\|X\|^2}\right) X$

Setting 8. $X \sim \mathcal{N}(\theta, I_p)$

Fact. $R(\hat{\theta}_{MLE}, \theta) = p$

Lemma 4.8 (Stein's lemma). $X \sim \mathcal{N}(\theta, 1)$, g bounded, differentiable, $\mathbb{E}|g'(X)| < \infty$, then $\mathbb{E}[(X - \theta)g(X)] = \mathbb{E}[g'(X)]$

Proposition 4.9. $X \sim \mathcal{N}(\theta, I_p)$, $p \geq 3$, then $R(\delta^{JS}, \theta) < p$ for all θ

Fact. δ^{JS} , $\hat{\theta}_{MLE}$ same maximal risk

Fact. δ^{JS} dominated by $\delta^{JS+}(X) = \left(1 - \frac{p-2}{\|X\|^2}\right)^{+X}$

Fact. admissible must be smooth

- $\begin{cases} X|Y=0 \sim f_0(x) \\ X|Y=1 \sim f_1(x) \end{cases}$
- classification rule $\delta_{\mathcal{R}}(X) = \begin{cases} 1 & \text{if } x \in \mathcal{R} \\ 0 & \text{if } x \in \mathcal{R}^c \end{cases}$
- $\mathbb{P}_1(X \in \mathcal{R}^c) = \mathbb{P}(X \in \mathcal{R}^c | Y = 1)$
- $\mathbb{P}_0(X \in \mathcal{R}) = \mathbb{P}(X \in \mathcal{R} | Y = 0)$
- risk function $R_\pi(\delta_{\mathcal{R}}) = \pi_0 \mathbb{P}_0(X \in \mathcal{R}) + \pi_1 \mathbb{P}_1(X \in \mathcal{R}^c)$
- marginal distribution P_X

- $\eta(x) = \mathbb{P}(1|X = x)$
- $Q(x, y) = f(x, y)\pi(x)$

Proposition 4.10. $R_\pi(\delta) = \mathbb{P}_Q(\delta(X) \neq Y) = \mathbb{E}_Q[\mathbb{1}\{\delta(X) \neq Y\}] = \int_{\mathcal{X}} \Pi(\delta^c(x)|x) dP_X(x)$

Setting 9. prior $\pi = (\pi_0, \pi_1)$

- Bayes classifier $\delta_\pi = \delta_{\mathcal{R}} = \begin{cases} 1 & \text{if } x \in \mathcal{R} \\ 0 & \text{if } x \in \mathcal{R}^c \end{cases}$
- $\mathcal{R} = \{\eta(x) \geq 1 - \eta(x)\}$

Proposition 4.11.

- (i) δ_π minimizes Bayes classification risk
- (ii) If $\mathbb{P}(\eta(x) = 1 - \eta(x)) = 0$, then Bayes rule unique
- discriminant function $D(X) = X^\top \Sigma(\mu_1 - \mu_0)$

Example (Linear discriminant analysis). $\left\{ \begin{array}{l} f_0 = \mathcal{N}(\mu_0, \Sigma) \\ f_1 = \mathcal{N}(\mu_1, \Sigma) \end{array} \right.$ can show Bayes rule only depends on $D(X)$

5 Multivariate analysis & PCA

- covariance $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$
- correlation $\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$
- sample correlation coefficient $\hat{\rho}_{X,Y} = \frac{\frac{1}{n} \sum (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\frac{1}{n} \sum (X_i - \bar{X}_n)^2} \sqrt{\frac{1}{n} \sum (Y_i - \bar{Y}_n)^2}}$

Fact. $Var(X), Var(Y)$ positive, finite, then sample version of $Var(X), Var(Y), Cov(X, Y)$ and $\hat{\rho}_{X,Y}$ consistent

Fact. $[\rho]_{ij} = \rho_{X^{(i)}, X^{(j)}}$

- (i) coefficients in $[-1, 1]$
- (ii) diagonal coefficients equals to 1
- (iii) positive semidefinite

Proposition 5.1.

- (i) Σ positive semidefinite, then $\exists X$ st $\mathbb{E}X = 0, \Sigma = Cov(X)$
- (ii) ρ positive semidefinite, 1 on the diagonal, then $\exists X$ st $\mathbb{E}X = 0, \Sigma = [\rho]$

Proposition 5.2. (????) Under $X, Y \sim \mathcal{N}(0, I_p)$ independent, density of $\hat{\rho}_{X,Y}$

$$f_{\hat{\rho}}(r) = \frac{\Gamma(\frac{1}{2}(n-1))}{\Gamma(\frac{1}{2}(n-2))} (1 - r^2)^{\frac{1}{2}(n-4)}$$

Setting 10. $X \sim \mathcal{N}(\mu, \Sigma)$, $X = (X^{(1)}, X^{(2)})$

- covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$
- covariance of $X^{(1)}|X^{(2)}$ — $\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$
- partial correlation of $X^{(1)(i)}, X^{(1)(j)}$ — $\rho_{i,j|2} = \frac{(\Sigma_{11|2})_{ij}}{\sqrt{(\Sigma_{11|2})_{ii}(\Sigma_{11|2})_{jj}}}$

Example (PCA).

Method: projecting the observations on directions with maximum variance

Idea: recover information about leading eigenspaces of the covariance matrix Σ

Setting 11. $E[X] = 0, \Sigma = \mathbb{E}[XX^\top]$

- $\Sigma = V\Lambda V^\top$ — $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$
- $U = V^\top X$

Fact. $\mathbb{E}[UU^\top] = \Lambda$

Proposition 5.3. $X \sim \mathcal{N}(0, \Sigma)$, then

- (i) $U \sim \mathcal{N}(0, \Lambda)$
- (ii) U_i independent

Fact. $X = V(V^\top X) = \sum v_i \sqrt{\lambda_i} Z_i$ where $Z \sim \mathcal{N}(0, I_p)$

- $T_{(-i)}$
- jackknife bias estimate $\hat{B}_n = (n-1) \left(\frac{1}{n} \sum T_{(-i)} - T_n \right)$
- jackknife bias corrected estimate $\tilde{T}_{JACK} = T_n - \hat{B}_n$

Proposition 5.4. regular bias $B(\theta)$, then $|\mathbb{E}(\tilde{T}_{JACK}) - \theta| = O\left(\frac{1}{n^2}\right)$

Example (Bootstrap).

Setting 12. discrete probability distribution $\mathbb{P}_n(\cdot|X_1, \dots, X_n)$ generate $(X_{n,i}^b : 1 \leq i \leq n)$

- $\mathbb{P}_n(X_n^b = X_i) = \frac{1}{n}$

Proposition 5.5. $\mathbb{E}(X_n^b) = \bar{X}_n$

- bootstrap sample mean \bar{X}_n^b

Idea: $\bar{X}_n^b - \bar{X}_n$ approximates $\bar{X}_n - \mu$

- bootstrap confidence set $\mathcal{C}_n^b = \left\{ |\nu - \bar{X}_n| \leq \frac{R_n^b}{\sqrt{n}} \right\}$ — $\mathbb{P}_n(|\bar{X}_n^b - \bar{X}_n| \leq \frac{R_n^b}{\sqrt{n}} | X_1, \dots, X_n) = 1 - \alpha$
- Φ — c.d.f. of $\mathcal{N}(0, \sigma^2)$

Theorem 5.6. X_i i.i.d. , mean μ , finite variance σ^2 , then
 $\sup_{\mathbb{R}} |\mathbb{P}_n(\sqrt{n}(\bar{X}_n^b - \bar{X}_n) \leq t | X_1, \dots, X_n) - \Phi(t)| \xrightarrow{a.s.} 0$

Fact. $\mathbb{P}(\mu \in \mathcal{C}_n^b) \rightarrow 1 - \alpha$

Lemma 5.7.

(i) $A_n \sim f_n$ with c.d.f. F_n

(ii) $A \sim f$ with continuous c.d.f. F

Then, $f_n \xrightarrow{d} f \Rightarrow \sup |F_n(t) - F(t)| \rightarrow 0$

– triangular array $(Z_{n,i} : i \leq n)$ — $Z_{n,1}, \dots, Z_{n,n}$ i.i.d. from distribution Q_n

Assumption 3.

(i) $\forall \delta > 0, n\mathbb{P}_n(|Z_{n,1}| > \delta\sqrt{n}) \rightarrow 0$

(ii) $\text{Var}(Z_{n,1} \mathbb{1}_{\{|Z_{n,1}| \leq \sqrt{n}\}}) \rightarrow \sigma^2$

(iii) $\sqrt{n}\mathbb{E}(Z_{n,1} \mathbb{1}_{\{|Z_{n,1}| > \sqrt{n}\}}) \rightarrow 0$

Proposition 5.8 (CLT for triangular arrays). Under assumption above, $(Z_{n,i})$ triangular array, finite variance, $\text{Var}(Z_{n,i}) = \sigma_n^2 \rightarrow \sigma^2$, then $\sqrt{n}(\frac{1}{n} \sum Z_{n,i} - \mathbb{E}_n(Z_{n,i})) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$

– nonparametric bootstrap

– parametric bootstrap

– pseudo-random uniform sample — $\mathbb{P}(U_1^* \leq u_1, \dots, U_N^* \leq u_N) = u_1 \dots u_N$

Proposition 5.9. $U_i \sim U[0, 1]$ i.i.d. , $X_i = \sum x_j \mathbb{1}_{\left\{U_i \in \left(\frac{j-1}{n}, \frac{j}{n}\right]\right\}}$, then X_i i.i.d uniform over $\{x_1, \dots, x_n\}$

– generalized inverse F^- — $F^-(u) = \inf \{x : u \leq F(x)\}$

Proposition 5.10. distribution P with c.d.f. F , U uniform over $[0, 1]$, then $X = F^-(U) \sim P$

Example (Importance sampling). density f, h , $\text{supp} f \subset \text{supp} h$, then $\mathbb{E}_h\left(\frac{g(X)}{h(X)}f(X)\right) = \mathbb{E}_f(g(X))$

Fact. (X_1^*, \dots, X_N^*) with density h , then $\frac{1}{N} \sum \frac{g(X_i^*)}{h(X_i^*)} f(X_i^*) \xrightarrow{a.s.} \mathbb{E}_f(g(X))$

Example (Accept/reject algorithm). $f \leq Mh$

Step 1 generate $X \sim h$, $U \sim U[0, 1]$

Step 2 if $U \leq \frac{f(X)}{Mh(X)}$, $Y = X$

Example (Gibbs sampler). bivariate (X, Y) , $X_0 = x$, $\begin{cases} Y_t \sim f_{Y|X}(\cdot | X = X_{t-1}) \\ X_t \sim f_{X|Y}(\cdot | Y = Y_t) \end{cases}$

By ergodic theorem, $\frac{1}{N} \sum g(X_t, Y_t) \xrightarrow{a.s.} \mathbb{E}_{X,Y}(g(X, Y))$

- empirical distribution function F_n — $F_n(t) = \frac{1}{n} \sum \mathbb{1}_{(-\infty, t]}(X_i) = \frac{\#\{i: X_i \leq t\}}{n}$

Theorem 5.11 (Glivenko-Cantelli). $F, \sup |F_n(t) - F(t)| \rightarrow 0$

- Brownian motion — $\begin{cases} W_0 = 0 \\ W_t - W_s \sim \mathcal{N}(0, t - s) \end{cases}$
- Brownian bridge — (B_t) Brownian motion conditioned $B_1 = 0$
 $\begin{cases} B_0 = B_1 = 0 \\ B_t \sim \mathcal{N}(0, t(1 - t)), \text{Cov}(B_s, B_t) = s(1 - t) \end{cases}$

Fact. *can be constructed by taking $B_t = W_t - tW_1$*

Theorem 5.12 (Donsker-Kolmogorov-Doob). \mathcal{G}_F Gaussian process, $\mathcal{G}_F(t) = B_{F(t)}, \text{Cov}(\mathcal{G}_F(s), \mathcal{G}_F(t)) = F(s)(1 - F(t)), \sqrt{n}(F_n - F) \xrightarrow{d} \mathcal{G}_F$

Theorem 5.13 (Kolmogorov-Smirnov). $\sqrt{n}\|F_n - F\|_\infty \xrightarrow{d} \|\mathcal{G}_F\|_\infty = \sup_{[0,1]} |B_t|$