STA 3000 F Lecture 2

Admissibility ct'd.

Stein's phenomenon/paradox

(σ^2 known)

$$X \sim \mathcal{N}(\theta, \sigma^2 I_d).$$

(Alternatively, $X_1, X_2, \ldots X_n \overset{iid}{\sim} \mathcal{N}(\theta, I_d)$) equiv to $\sigma^2 = 1/n$.

Natural choice $\hat{\theta} = X$. not admissible when $d \geq 3$.

James - Stein estimator.

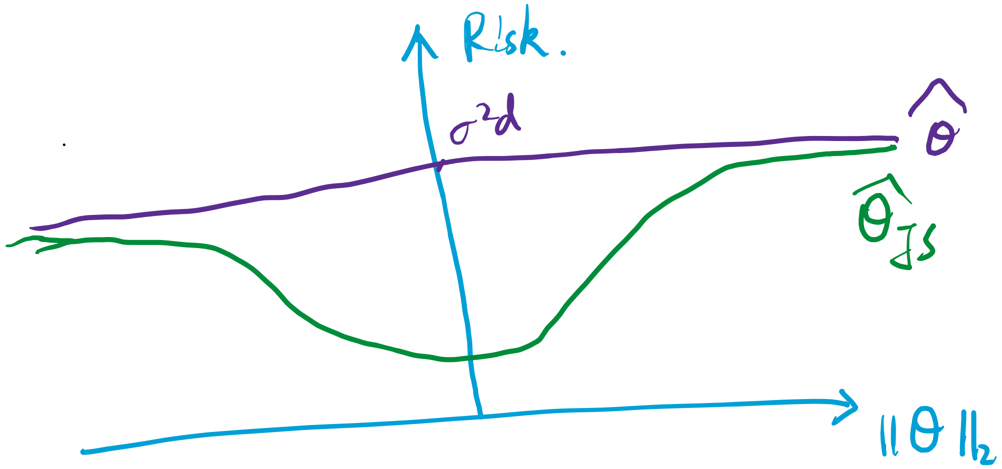
$$\hat{\theta}_{JS} = \left(1 - \frac{(d-2)\sigma^2}{\|X\|_2^2}\right) X.$$

Fact : $\forall \theta \in \mathbb{R}^d$

$$\mathbb{E}_\theta\left[\|\hat{\theta}_{JS} - \theta\|_2^2\right] = \sigma^2 d - \mathbb{E}\left[\left(\frac{(d-2)\sigma^2}{\|X\|_2}\right)^2\right]$$

>0

$$\mathbb{E}_\theta\left[\|\hat{\theta} - \theta\|_2^2\right] = \sigma^2 d.$$



Related to: shrinkage estimation & empirical Bayes.

Other criteria: Bayes and minimax decision rules.

· Bayes risk: π: prior distribution (on Θ).

$$r_\pi(\delta) := \int_\Theta \underbrace{R(\theta;\delta)}_{\text{risk of } \delta} \pi(d\theta)$$

$$\delta_{\text{Bayes},\pi} = \arg\min_\delta \{ r_\pi(\delta) \}$$

Not necessarily following Bayesian framework, $\int \cdots \pi$ can be understood as weighted average.

Calculate the Bayes decision rule.

Suppose $p_\theta(x) = \frac{dP}{d\lambda}(x)$. for some base measure λ.

$$r_\pi(\delta) = \int_\Theta \int_X L(\theta;\delta(x))\, p_\theta(x)\, \lambda(dx) \cdot \pi(d\theta)$$

(Fubini–Tonelli)

$$= \int_{\mathcal{X}} \int_{\Theta} L(\theta; \delta(x))\, p_\theta(x)\, \pi(d\theta)\, \lambda(dx).$$

δ can depend on x arbitrarily.

For each x, minimize inner integral by choosing δ.

Bayes decision rule:

$$\delta_{\text{Bayes},\pi}(x) = \arg\min_{a \in \mathcal{A}} \int_{\Theta} L(\theta; a)\, p_\theta(x) \cdot \pi(d\theta).$$

Posterior distribution.

$$\Pi(d\theta \mid X) = \frac{\pi(d\theta) \cdot p_\theta(X)}{\int_{\Theta} p_{\theta'}(x)\, \pi(d\theta')}.$$

$$\mathbb{E}_{\Pi}[L(\theta; a)] = \frac{\int_{\Theta} L(\theta; a)\, p_\theta(X)\, \pi(d\theta)}{\int_{\Theta} p_\theta(x)\, \pi(d\theta)}$$

Indp of a.

So Bayes rule is

$$\delta_{\text{Bayes},\pi}(x) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{\Pi(\cdot \mid X)}[L(\theta, a)].$$

For estimation problems with MSE risk.

minimize $\int_{\Theta} (g(\theta) - a)^2 \, \Pi(d\theta | X)$.

$\delta_{Bayes, \pi} = \int_{\Theta} g(\theta) \, \Pi(d\theta | X)$ posterior mean.

(In practice : use MCMC/VB to compute $\int \cdots \Pi(\cdot | X)$).

Computable case:

e.g. $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. $\pi = N(0, v^2)$.

$$\Pi(\theta | X_1^n) \propto \pi(\theta) \cdot p_\theta(X_1) p_\theta(X_2) \cdots p_\theta(X_n)$$

$$= N\left(\frac{v^2 n \cdot \bar{X}_n}{v^2 n + 1}, \frac{v^2}{v^2 n + 1} \right).$$

v larger $\Longleftrightarrow$ Less informative prior $\Longleftrightarrow$ less shrinkage.

Conjugate families.

(e.g. Beta-Binomial, Dirichlet - Categorical)

Minimax decision rule.

$$\inf_{\delta} \sup_{\theta} R(\theta; \delta).$$

(Usually we can only find "minimax up to constant factors").

Game-theoretic interpretation: (both parties can make randomized choices).

- Statistician first choose δ
- Nature (by observing δ) chooses θ (π). adversarially to maximize risk

Weak duality.

$$\underbrace{\boxed{\sup_{\pi} \inf_{\delta} \{ r_{\pi}(\delta) \}}}_{\text{worst-case Bayes risk.}} \le \inf_{\delta} \sup_{\pi} \{ r_{\pi}(\delta) \}$$

Strong duality. ($\underbrace{\text{finite } \Theta, \mathbb{X}, \mathbb{A}}_{\text{(can be relaxed)}}$).

$$\inf_{\delta} \sup_{\pi} \{ r_{\pi}(\delta) \} = \sup_{\pi} \inf_{\delta} \{ r_{\pi}(\delta) \}$$

(von Neumann minimax theorem).

Minimax rule is the Bayes rule for worst-case prior.

Simple cases.

Fact. Constant-risk Bayes rule δ is minimax.

Proof: Consider decision rule δ'

$$\sup_{\theta\in\Theta} R(\theta,\delta') \geq \int_{\Theta} R(\theta,\delta')\,\pi(d\theta)$$

$$\geq \int_{\Theta} R(\theta,\delta)\,\pi(d\theta)$$

(by optimality of δ)

$$= \sup_{\theta\in\Theta} R(\theta,\delta).$$

e.g. $X \sim \text{Binom}(n,p)$ Goal: estimate p under MSE.

Natural choice: $\hat{p} = \dfrac{X}{n}$

$$R(p,\hat{p}) = \frac{p(1-p)}{n}. \qquad \sup_{p} R(p,\hat{p}) = \frac{1}{4n}.$$

By choosing prior $\pi = \text{Beta}(\alpha,\beta)$.

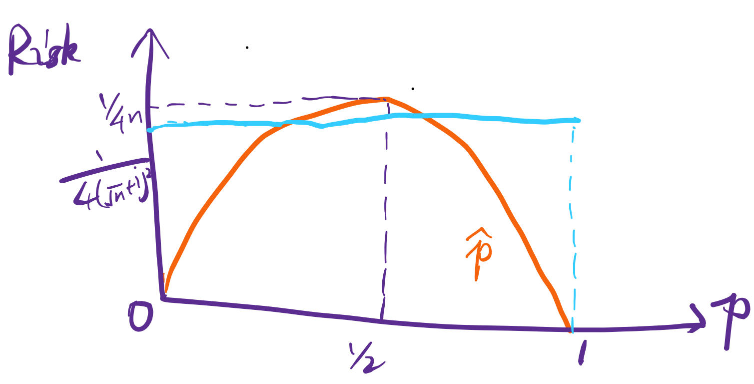
Posterior mean $\delta_{\text{Bayes},\pi}(X) = \dfrac{\alpha+X}{\alpha+\beta+n}$.

$$R(p,\delta_{\text{Bayes},\pi}) = \frac{n(1-p)p + (\alpha(1-p)-\beta p)^2}{(\alpha+\beta+n)^2}$$

$$\alpha=\beta=\sqrt{n}/2. \qquad R(\dots) = \frac{1}{4(1+\sqrt{n})^2}$$

$\shortparallel$

R_{minimax}.



In practice: $\hat{p}$ may be preferred due to adaptivity.
(local minimax, to be discussed).

Extension: Suppose that $\exists \left(\pi_j, r_j\right)_{j=1}^{+\infty}$ r_j = Bayes risk under π_j.

s.t. $\liminf_{j \to +\infty} r_j \geq \sup_\theta R(\theta, \delta)$ for some δ.

then δ is minimax.

Proof. For another decision rule δ'

$$\sup_\theta R(\theta, \delta') \geq \liminf_{j \to +\infty} \int_\Theta R(\theta, \delta') \, \pi_j(d\theta)$$

$$\geq \liminf_{j \to +\infty} \int_\Theta R(\theta, \delta_{Bayes, \pi_j}) \, \pi_j(d\theta)$$

$$= \liminf_{j \to +\infty} r_j \geq \sup_\theta R(\theta, \delta).$$

e.g. $X_1, \cdots X_n \overset{iid}{\sim} N(\theta, I_d)$. $\pi_k = N(0, k I_d)$

$$r_{\pi_k} = \frac{kd}{kn+1} \longrightarrow \frac{d}{n}. \quad (k \to +\infty).$$

For $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$. risk $= \frac{d}{n}$ $(\forall \theta \in \Theta)$.

$$= \liminf_{k \to +\infty} r_{\pi_k}.$$

So $\bar{X}_n$ is minimax.

Connection between Bayes and admissible rules.

- Fact: unique Bayes rules δ are admissible.

Proof: Suppose δ' dominates δ, i.e. $R(\theta,\delta') \le R(\theta,\delta)$ $(\forall \theta \in \Theta)$.

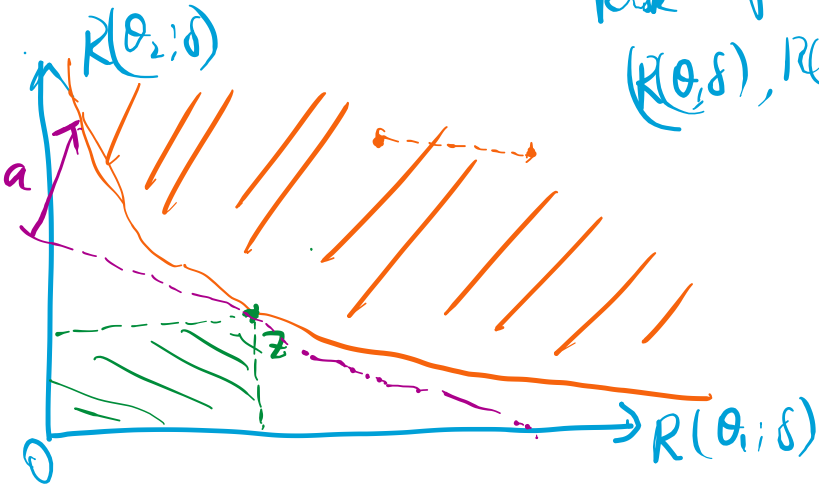
$$r_\pi(\delta') = \int R(\theta,\delta')\,\pi(d\theta) \le \int R(\theta,\delta)\,\pi(d\theta) = r_\pi(\delta).$$

So $\delta' = \delta$ by uniqueness.

- Fact: When $|\Theta| < +\infty$, admissible rules are Bayes.

Proof:

Risk of δ: $(R(\theta_1,\delta), R(\theta_2,\delta), \ldots, R(\theta_K,\delta))$.



$$C = \{ (R(\theta_j,\delta))_{j=1}^{K} : \text{for some decision rule } \delta \}$$

C is convex: given δ_1, δ_2, $\lambda \in [0,1]$.

take randomized decision rule $\begin{cases} \delta_1 & \text{w.p. } \lambda \\ \delta_2 & \text{w.p. } 1-\lambda \end{cases}$

Admissibility of risk vector z

$$C \cap \{x : x \le z\} = \{z\}.$$

Separating hyperplane:

(guaranteed to exist by separating hyperplane thm).

$$\exists\ a \in \mathbb{R}^k,\ b \in \mathbb{R} \quad \text{s.t.} \quad a^T x \begin{cases} \geq b & \text{when } x \in C \\ \leq b & \text{when } x \in Z. \end{cases}$$

$a_i \geq 0 \quad \forall i$ since it does not cross the rectangular region.

Then we take $\pi = \dfrac{a}{\|a\|_1}$

Z is the risk of Bayes estimator under π.

"Sufficient statistics".

Motivation: eg. $X_1, X_2, \dots X_n \overset{iid}{\sim} Ber(p)$

$T(X_1^n) = X_1 + \dots + X_n \sim Binom(n, p)$.

If we condition on $\{T = t\}$

$\{X_i\}_{i=1}^n$ is indicator of uniform random subset of $\{1, 2, \dots n\}$ with size t.

The conditional distribution does not depend on p.

Def. $X \sim \mathbb{P}_\theta \in \mathcal{P}_\Theta$, we call $T = T(X)$ sufficient if $\forall t, \theta$, conditional distribution of X given $T = t$ is independent of θ. under $\mathbb{P}_\theta$

Thm. T is sufficient, for any decision rule δ

$\exists\ \hat{\delta}(T(X))$ (depends on X only through T)

s.t. $\delta(X) \stackrel{d}{=} \hat{\delta}(T(X))$ under any $\mathbb{P}_\theta$.

Proof: $\underbrace{\text{sample } X \mid T(X)}_{\text{indp of } P_\theta}$ and apply δ

Thm (Rao-Blackwellization)

If $L(\theta;\cdot)$ is convex (e.g. MSE).

Given δ, $\eta(T(X)) := \underbrace{\mathbb{E}[\delta(X) \mid T(X)]}_{\text{computable, indp of } \theta}$.

then $R(\theta; \eta) \le R(\theta; \delta)$.

Proof: Jensen's ineq.