STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Huan Cheng; Ip Man Fai

Lecture 8

1 Bayes Estimators and Average Risk Optimality

We need to introduce a measure Λ over the parameter space Ω. This measure Λ can be viewed as an assignment if weights to each of the parameters values $\theta \in \Theta$ a priori. [i.e. before any data is observed]

Remark 1 *The parameter of interest θ is not fixed and unknown constant.*

Given a measure Λ, our objective is to find an estimator δ_Λ which minimizes the average risk, which is given by

$$r(\Lambda, \delta) = \int R(\theta, \delta) d\Lambda(\theta) = E_\theta(R(\theta, \delta)). \tag{1}$$

If Λ is a probability distribution on Ω, we call Λ the prior distribution. Correspondingly, the estimator δ_Λ, if exists, is called the Bayes estimator with respect to Λ, and the minimized average risk is called the **Bayes risk**.

$$r(\Lambda, \delta) = E_{(X,\Theta)}(L(\Theta, \delta(X))) = E_\Theta(E_X(L(\Theta, \delta(X) \mid \Theta))) = E_\Theta(R(\Theta, \delta)). \tag{2}$$

We shall pay attention to $E(L(\Theta, \delta(X)) \mid X = x)$, the conditional risk at (almost) every value of X. Notice that the expectation have is taken with respect to the conditional distribution of Θ given X, i.e. $(\Theta \mid X = x)$.

Theorem 1 *Suppose $\Theta \sim \Lambda$ and $X \mid \Theta = \theta \sim P_\theta$. If*

(a) There exists δ_0, an estimator of $g(\theta)$ with finite risk for all θ, and

(b) There exists a value $\delta_\Lambda(X)$ that minimizes $E(L(\Theta, \delta_\Lambda(X)) \mid X = x)$ for almost every X,

then δ_Λ is a Bayes estimator with respect to Λ.

Note that the almost sure statement is defined with respect to the marginal distribution of X, which is given by

$$P(X \in A) = \int P_\theta(X \in A) d\Lambda(\theta) \tag{3}$$

Proof 1 *Under the assumptions of theorem (a) and (b), for any other estimator δ', say, and for almost surely X, $E(L(\Theta, \delta_\Lambda(X)) \mid X = x) \le E(L(\Theta, \delta'_\Lambda(X)) \mid X = x)$. After taking expectation over X, we obtain $E(L(\Theta, \delta_\Lambda(X))) \le E(L(\Theta, \delta'_\Lambda(X)))$ for all δ'.*

Example 1 (Bayes estimator of L^2 loss) *If we consider the squared loss function $L(\theta, d) = (\theta - d)^2$, to find the Bayes estimator. We need to minimize $E((g(\Theta) - \delta(X))^2 \mid X = x)$ and in this case, the Bayes estimator is $\delta_\Lambda(X) = E(g(\Theta) \mid X)$, the posterior mean of $g(\Theta)$ given $X = x$*

Consider the Risk function, $E(L(\Theta, \delta(X)) \mid X = x)$, we can observe that

$$
\begin{aligned}
&E(\{g(\Theta) - E(g(\Theta) \mid X) + E(g(\Theta) \mid X) - \delta(X)\}^2 \mid X = x) \\
&= E(\{g(\Theta) - E(g(\Theta) \mid X)\}^2 \mid X = x) + E(\{E(g(\Theta) \mid X) - \delta(X)\}^2 \mid X = x)
\end{aligned}
$$

which shows the risk function could be minimized by posterior mean if it is the Bayes estimator.

Remark 2 *To calculate the posterior mean $E(g(\Theta) \mid X)$, we should find out the posterior distribution first. Since posterior = joint / marginal = priori $\times$ likelihood / marginal , which is equivalent to $p(\theta \mid X) = p(\theta, X)/\int p(\theta', X)d\theta' = p(X \mid \theta) \times \pi(\theta)/\int p(\theta', X)d\theta'$ by Bayes's Theorem , posterior distribution could be derived as posterior $\propto$ prior $\times$ likelihood.*

Example 2 (Binomial-Beta) *Suppose $X \sim Binomial(n, \theta)$ given $\Theta = \theta$ and that Θ has a prior distribution $Beta(\alpha, \beta)$, with hyperparameters α and β. The prior density is given by*

$$
\pi(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathbf{1}_{\{0<\theta<1\}}. \tag{4}
$$

Obviously, the model density is $f(X;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, in which case the posterior distribution of Θ given X is

$$
\begin{aligned}
\pi(\theta \mid X) &\propto \underbrace{\binom{n}{x}\theta^x(1-\theta)^{n-x}}_{\text{Likelihood}} \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{Prior}} \\
&\propto \underbrace{\theta^{(x+\alpha)-1}(1-\theta)^{(n-x+\beta)-1}}_{\text{Kernel part}} \\
&\sim Beta(x+\alpha, n-x+\beta),
\end{aligned}
$$

where $\int p(\theta', X)d\theta'$ (the denominator part of posterior) is normalising constant, meaning that the posterior of $\Theta \mid X = (x+\alpha)/(n+\alpha+\beta)$.

Remark 3 *The posterior mean can be rewritten as:*

$$
\overbrace{\frac{X+\alpha}{n+\alpha+\beta}}^{\text{Shrink the estimate from prior mean}} = \underbrace{\frac{n}{n+\alpha+\beta}}_{\omega}(\frac{X}{n}) + \overbrace{\underbrace{\frac{\alpha+\beta}{n+\alpha+\beta}}_{1-\omega}(\frac{\alpha}{1+\beta})}^{\text{Contribution from prior}}
$$

ω and $1-\omega$ can be treated as the weight average of the sample mean $\bar{X}_n$ and the prior mean $\alpha/(\alpha+\beta), correspondingly$. As $n \to \infty$ (by empirical evidence and observations), $E(\Theta \mid X) \to \bar{X}_n$. (Let the data "speak for themselves.")

Example 3 (Normal Mean Estimation) *Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\Theta, \sigma^2)$, with σ^2 known. Let $\Theta \sim N(\mu, b^2)$*

where μ and b^2 are two fixed prior hyperparameters. Then the posterior distribution of $\Theta \mid X$ is

$$\begin{aligned}
\pi(\theta \mid X) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(X_i-\theta)^2\right\} \times \frac{1}{\sqrt{2\pi b^2}} \exp\left\{-\frac{1}{2b^2}(\theta-\mu)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2 - \frac{1}{2b^2}(\theta-\mu)^2\right\} \\
&\propto \cdots \\
&\propto \exp\left\{-\frac{1}{2}\Big(\frac{n}{\sigma^2}+\frac{1}{b^2}\Big)\theta^2 + \Big(\frac{n\bar{X}}{\sigma^2}+\frac{\mu}{b^2}\Big)\theta\right\} \\
&\propto \exp\left\{-\frac{1}{2\tilde{\sigma}^2}(\theta-\tilde{\mu})^2\right\}.
\end{aligned}$$

The posterior distribution of Θ given X is $N(\tilde{\mu}, \tilde{\sigma}^2)$ where

$$\tilde{\mu} = \frac{n\bar{X}/\sigma^2 + \mu/b^2}{n/\sigma^2 + 1/b^2} \qquad \textit{and} \qquad \tilde{\sigma}^2 = \tfrac{1}{n/\sigma^2+1/b^2}$$

Hence, the posterior mean of $\Theta \mid X$ is $\dfrac{n\bar{X}/\sigma^2 + \mu/b^2}{n/\sigma^2 + 1/b^2}$ and similarly we can rewrite as

$$\underbrace{\frac{n/\sigma^2}{n/\sigma^2 + 1/b^2}}_{1\ as\ n\to\infty} \bar{X} + \underbrace{\frac{1/b^2}{n/\sigma^2 + 1/b^2}}_{0\ as\ n\to\infty} \mu$$

Thus, Bayes estimator δ_Λ is $\tilde{\mu}$ if we adopt the squared loss function.

Example 4 (Bayes estimator of weighted L^2 loss) *Assume that we consider $L(\theta, d) = \omega(\theta)\{d - g(\theta)\}^2$, where $\omega(\theta) \geqslant 0$, which can be interpreted as a weight function. Our goal is to find the corresponding Bayes estimator, which minimizes $E(\omega(\Theta)\{g(\Theta) - d\}^2 \mid X = x)$ $(*)$with respect to d.*
$(*)$ can be rewritten as

$$d^2 E(\omega(\Theta) \mid X = x) - 2dE(\omega(\Theta)g(\Theta) \mid X = x) + E(\omega(\Theta)g(\Theta)^2 \mid X = x). \tag{$\dagger$}$$

Tanking derivative of ($\dagger$) with respect to d, we obtain

$$2d^* E(\omega(\Theta) \mid X = x) - 2E(\omega(\Theta)g(\Theta) \mid X = x) = 0.$$

Thus

$$\delta_\Lambda(x) = d^* = \frac{E(\omega(\Theta)g(\Theta) \mid X = x)}{E(\omega(\Theta) \mid X = x)}. \tag{5}$$

In particular, if $\omega(\cdot) \equiv 1$, $\delta_\Lambda(x)$ (with $\omega(\cdot) \equiv 1$)=$E(g(\Theta) \mid X = x)$.

Theorem 2 *If δ is unbiased for $g(\theta)$ with $r(\Lambda, \delta) < \infty$ and $E(g(\Theta)^2) < \infty$, then δ is not Bayes under the squared loss function unless its average risk is zero, which is*

$$E_{(X,\Theta)}(\{\delta(X) - g(\Theta)\}^2) = 0. \tag{6}$$

Proof 2 *Let δ be an unbiased estimator under the squared loss function. Then we know that δ is the posterior mean, which is*

$$\delta(X) = E(g(\Theta) \mid X),$$

almost surely. Thus, we have

$$\begin{aligned} E(\delta(X)g(\Theta)) &= E(E(\delta(X)g(\Theta) \mid X)) \\ &= E(\delta(X)E(g(\Theta) \mid X)) \\ &= E(\delta^2(X)). \end{aligned} \tag{7}$$

Also,

$$\begin{aligned} E(\delta(X)g(\Theta)) &= E(E(\delta(X)g(\Theta) \mid \Theta)) \\ &= E(g(\Theta)E(\delta(X) \mid \Theta)) \\ &= E(g^2(\Theta)). \end{aligned} \tag{8}$$

Observe that

$$\begin{aligned} E(\{\delta(X) - g(\Theta)\}^2) &= E(\delta^2(X)) - 2E(\delta(X)g(\Theta)) + E(g^2(\Theta)) \\ &= E(\delta^2(X)) - E(\delta(X)g(\Theta)) + E(g^2(\Theta)) - E(\delta(X)g(\Theta)) \\ &= E(\delta^2(X)) - E(\delta^2(X)) + E(g^2(\Theta)) - E(g^2(\Theta)) \ (due\ to\ (7)\ and\ (8)) \\ &= 0. \end{aligned}$$

Thus we have that $E(\{\delta(X) - g(\Theta)\}^2) = 0$, which means the average risk is zero. The claim is thus proved.

Example 5 (Application of Theorem 2) *Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\Theta, \sigma^2)$, with σ^2 known. Is $\bar{X}$ Bayes under the squared loss function for some choice of the prior distribution?*
Observe that $E(\bar{X} \mid \theta) = \theta$, hence $\bar{X}$ is unbiased for θ. The corresponding average risk under the squared loss function is given by

$$E_{(X,\Theta)}(\{\bar{X} - \Theta\}^2) = \frac{\sigma^2}{n} \neq 0.$$

So $\bar{X}$ is not Bayes estimator under any prior distribution.

Theorem 3 (Admissibility) *A unique Bayes estimator (almost surely for all P_θ) is admissible.*
An estimator is admissible if it is not uniformly dominated by some other estimator. δ is said to be inadmissible if and only if there exists δ' such that

$$\begin{cases} R(\theta, \delta') \leq R(\theta, \delta), \ for\ any\ \theta \in \Omega \\ R(\theta, \delta') < R(\theta, \delta), \ for\ some\ \theta \in \Omega \end{cases}$$

Proof 3 *Suppose δ_Λ is Bayes for Λ, and for some δ', $R(\theta, \delta') \leq R(\theta, \delta_\Lambda)$ for all $\theta \in \Omega$. If we take expectation with respect to Θ, the inequality above is preserved and we can write*

$$\int_{\theta \in \Omega} R(\theta, \delta') d\Lambda(\theta) \leq \int_{\theta \in \Omega} R(\theta, \delta_\Lambda) d\Lambda(\theta)$$

This implies that δ' is also Bayes because δ' has less (or equal) risk than δ_Λ which minimizes the average risk. Hence $\delta' = \delta_\Lambda$ with probability one for all P_θ.

$\boxed{\text{Question: When is a Bayes estimator unique?}}$

Theorem 4 (Uniqueness) *Let Q be the marginal distribution of X, that is*

$$Q(E) = \int P(X \in E \mid \theta) d\Lambda(\theta)$$

Then, under a strictly convex loss function, δ_Λ is unique (almost surely for all P_θ) if
(a) $r(\Lambda, \delta_\Lambda)$ is finite and
(b) $P_\theta \ll Q$ (absolute continuity)

$$\text{Benefits of Bayes} \begin{cases} (i) \text{ Admissible} \\ (ii) \text{ Incorporate } \underbrace{\text{prior information}}_{\text{domain knowledge}} \longrightarrow \text{frequentist} \\ (iii) \dots \end{cases}$$

2 Next Lecture

1. Minimax Estimator

   Considering

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

2. Worst-case Scenario/Optimality

3. Testing of Statistical Hypothesis (UMP, UMPU...)