CHAPTER 5.

BAYES METHODS AND ELEMENTARY DECISION THEORY

1. ELEMENTARY DECISION THEORY

Notation 1.1.

Let

- $\Theta \equiv$ the states of nature.
- $\mathbf{A} \equiv$ the action space
- $\mathbf{X} \equiv$ the sample space of a random variable X with distribution P_θ
- $P : \mathbf{X} \times \Theta \to [0,1]$; $P_\theta(X = x) =$ probability of observing $X = x$ when θ is true.
- $L : \Theta \times \mathbf{A} \to R^+$; a loss function.
- $d : \mathbf{A} \times \mathbf{X} \to [0,1]$; a decision function function. $d(a, x) = d(a \mid x) =$ probability of action a when $X = x$ is observed.
- $\mathbf{D} = \{$all decision functions$\}$.

Definition 1.2. The *risk* for rule $d \in \mathbf{D}$ when $\theta \in \Theta$ is true is

$$
\begin{aligned}
R(\theta, d) &= E_\theta L(\theta, d(\,\cdot \mid X)) \\
&= \int_{\mathbf{X}} \int_{\mathbf{A}} L(\theta, a)\, d(da \mid X = x)\, P_\theta(dx) \\
&= \sum_{i=1}^{k} L(\theta, a_i) \{ \sum_{j=1}^{m} d(a_i, x_j)\, P_\theta(X = x_j) \} \qquad \text{in the discrete case .}
\end{aligned}
$$

We will call $R : \Theta \times \mathbf{D} \to R^+$ the *risk function.*

Definition 1.3. A decision rule d is *inadmissible* if there is a rule d' such that $R(\theta, d') \le R(\theta, d)$ for all θ and $R(\theta, d') < R(\theta, d)$ for some θ. A decision rule d is *admissible* if it is not inadmissible.

Example 1.4. Hypothesis testing. $\mathbf{A} = \{0,1\}$, $\Theta = \Theta_0 \cup \Theta_1$

$$L(\theta, 0) = l_0 1_{\Theta_1}(\theta) , \qquad L(\theta, 1) = l_1 1_{\Theta_0}(\theta) .$$

Since $\mathbf{A}$ contains just two points, $d(1 \mid x) = 1 - d(0 \mid x) \equiv \phi(x)$; and then

$$P_\theta(accept\ \ H_0) = E_\theta d(0 \mid X) = E_\theta(1 - \phi(X)) ,$$

$$P_\theta(reject\ \ H_0) = E_\theta d(1 \mid X) = E_\theta \phi(X) ,$$

and

$$R(\theta, d) = l_1 E_\theta \phi(X) 1_{\Theta_0}(\theta) + l_0 E_\theta(1 - \phi(X)) 1_{\Theta_1}(\theta) .$$

The classical Neyman - Pearson hypothesis testing philosophy bounds $l_1 \alpha \equiv \sup_{\theta \in \Theta_0} R(\theta, d)$ and tries to minimize

$$\mathbf{l}_0(1 - \beta_d(\theta)) \equiv R(\theta, d)$$

for each $\theta \in \Theta_1$.

B. **Estimation.** $\mathbf{A} = \Theta$; typically $\Theta = R^s$ for some s (but sometimes it is useful to consider $\Theta = \mathbf{T}$ where $(T, |\cdot|)$ is a metric space). A typical loss function (often used for mathematical simplicity more than anything else) is

$$L(\theta, a) = K|\theta - a|^2 \qquad \text{for some } K .$$

Then the risk of a rule d is

$$\begin{aligned} R(\theta, d) &= K\, E_\theta|\theta - d(\,\cdot\,|X)|^2 \\ &= K\, E_\theta|\theta - d(X)|^2 \qquad \text{for a non - randomized } d \\ &= K\{Var_\theta d(X) + [bias_\theta(d(X))]^2\} \qquad \text{when } s = 1 . \end{aligned}$$

Definition 1.5. A decision rule d_M is *minimax* if

$$\inf_{d \in \mathbf{D}} \sup_{\theta \in \Theta} R(\theta, d) = \sup_{\theta \in \Theta} R(\theta, d_M) .$$

Definition 1.6. A probability distribution Λ over Θ is called a *prior distribution.*

Definition 1.7. For any given prior Λ and $d \in \mathbf{D}$, the *Bayes risk of* d *with respect to* Λ *is*

$$\begin{aligned} I\!R(\Lambda, d) &= \int_\Theta R(\theta, d)\, d\Lambda(\theta) \\ &= \sum_{i=1}^{l} R(\theta_i, d)\, \lambda_i \qquad \text{if } \Theta = \{\theta_1, \cdots, \theta_l \end{aligned}$$

Definition 1.8. A *Bayes decision rule with respect to* Λ *,* d_Λ *, is any rule satisfying*

$$I\!R(\Lambda, d_\Lambda) = \inf_{d \in \mathbf{D}} I\!R(\Lambda, d) = \text{Bayes risk} .$$

Example 1.9. $\Theta = \{1, 2\}$.

Urn 1: 10 red balls, 20 blue balls, 70 green balls.

Urn 2: 40 red balls, 40 blue balls, 20 green balls.

One ball is drawn from one of the two urns. Problem: decide which urn the ball came from if the losses $L(\theta, a)$ are given by

θ/a	1	2
1	0	10 .
2	6	0

Let $d = (d_R, d_B, d_G)$ with $d_x =$ probability of choosing urn 1 if color $X = x$ is observed. Then,

$$\begin{aligned} R(1,d) &= 10\,P_1(action\ 2) \\ &= 10\{.1(1-d_R) + .2(1-d_B) + .7(1-d_G)\} \end{aligned}$$

and, similarly,

$$R(2,d) = 6\{.4d_R + .4d_B + .2d_G\}.$$

If the prior distribution on the urns is given by $\lambda_1 = \lambda$ and $\lambda_2 = 1-\lambda$, then the Bayes risk is

$$I\!R(\Lambda, d) = 10\lambda + (2.4 - 3.4\lambda)d_R + (2.4 - 4.4\lambda)d_B + (1.2 - 8.2\lambda)d_G.$$

This is minimized by choosing $d_x = 1$ if its coefficient is negative, 0 if its coefficient is positive. For example, if $\lambda = 1/2$, then the Bayes risk equals

$$5 + .7d_R + .2d_B - 2.9d_G,$$

which is minimized by $d_R = d_B = 0$, $d_G = 1$; i.e. the Bayes rule d_Λ with respect to $\lambda = 1/2$ is $d_\Lambda = (0,0,1)$. Note that the Bayes rule is in fact a nonrandomized rule. This give us the Bayes risk for $\lambda = 1/2$ as $I\!R(1/2, d_B) = 2.1$.

I claim that the minimax rule is $d_M = (0, 9/22, 1)$, which is a randomization of the two nonrandom rules $d_2 = (0,0,1)$ and $d_7 = (0,1,1)$. This is easily confirmed by computing $R(1, d_M) = 263/110 = R(2, d_M)$.

Here is a table giving all of the nonrandomized rules:

X	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
R	0	0	0	1	1	1	0	1
B	0	0	1	0	1	0	1	1
G	0	1	0	0	0	1	1	1
$R(1,d)$	10	3	8	9	7	2	1	0
$R(2,d)$	0	1.2	2.4	2.4	4.8	3.6	3.6	6

5.2

2. STRUCTURE OF THE RISK BODY: THE FINITE CASE

Suppose that

$$\Theta = \{\theta_1, \cdots, \theta_l\}, \quad \mathbf{X} = \{x_1, \cdots, x_m\}, \quad \mathbf{A} = \{a_1, \cdots, a_k\}.$$

Definition 2.1. $A \subset R^l$ is *convex* if, for all $\lambda \in [0,1]$ $x, y \in A$, $\lambda x + (1-\lambda)y \in A$.

Lemma 2.2. Let $\lambda = (\lambda_1, \cdots, \lambda_t)$ be a probability distribution, $\lambda_i \geq 0$, $\sum_{i=1}^{t} \lambda_i = 1)$. Then for any $d_1, \cdots, d_t \in \mathbf{D}$, $\sum_{i=1}^{t} \lambda_i d_i \in \mathbf{D}$.

Proof. set $d(a_l, x_j) = \sum_{i=1}^{t} \lambda_i d_i(a_l, x_j)$. Then $d(a_l, x_j) \geq 0$ and

$$\sum_{l=1}^{k} d(a_l, x_j) = \sum_{i=1}^{t} \lambda_i \sum_{l=1}^{k} d_i(a_l, x_j) = 1.$$

□

We call

$$\mathbf{R} = \{(R(\theta_1, d), \cdots, R(\theta_l, d)) \in R^{+l} : d \in \mathbf{D}\}$$

the *risk body.*

Theorem 2.3. The risk body is convex.

Proof. Let $d_1, d_2 \in \mathbf{D}$ denote the rules corresponding to $R_i = (R(\theta_1, d_i), \cdots, R(\theta_l, d_i))$, $i = 1, 2$, and set $d = \lambda d_1 + (1-\lambda) d_2 \in \mathbf{D}$ by the lamma. Then

$$\begin{aligned}
&\lambda R(\theta_l, d_1) + (1-\lambda) R(\theta_l, d_2) \\
&\qquad = \sum_{i=1}^{k} \sum_{j=1}^{m} L(\theta_l, a_i) \{\lambda d_1(a_i, x_j) + (1-\lambda) d_2(a_i, x_j)\} p_{\theta_l}(x_j) \\
&\qquad = R(\theta_l, d),
\end{aligned}$$

so $d \in \mathbf{D}$ has risk point $\lambda R_1 + (1-\lambda) R_2 \in \mathbf{R}$. □

Theorem 2.4. Every $d \in \mathbf{D}$ may be expressed as a convex combination of the nonrandomized rules.

Proof. Any $d \in \mathbf{D}$ may be represented as a $k \times m$ matrix of ≥ 0 real numbers whose columns add to 1. The nonrandomized rules are those whose entries are 0's and 1's. Set $d(a_i, x_j) = d_{ij}$, $d = (d_{ij})$, $1 \leq i \leq k$, $1 \leq j \leq m$.

There are k^m nonrandom decision rules: call them $\delta^{(r)} = (\delta_{ij}^{(r)})$, $1 \le r \le k^m$. Given $d \in \mathbf{D}$, we want to find $\lambda_l \ge 0$, $\sum_{l=1}^{k^m} \lambda_l = 1$, so that

$$\sum_{l=1}^{k^m} \lambda_r \delta^{(r)} = d \ .$$

Claim: $\lambda_r = \prod_{j=1}^{m} \prod_{i=1}^{k} d_{ij}^{\delta_{ij}^{(r)}} = \prod_{j=1}^{m} \{ \sum_{i=1}^{k} d_{ij} \delta_{ij}^{(r)} \}$ work.

Proof. Induction on m. □

3. THE FINITE CASE: RELATIONS BETWEEN BAYES, MINIMAX, ADMISSIBILITY

This section continues our examination of the special, but illuminating, case of of a finite set Θ. In this case we can prove a number of results about Bayes and minimax rules and connections between them which carry over to more general problems with some appropriate reformulation.

Theorem 3.1. A minimax rule always exists.

Proof. If $0 \notin \mathbf{R}$, then for some $c > 0$ the cube with all sides of length c in the positive quadrant does not intersect $\mathbf{R}$. Let c increase until the square intersects $\mathbf{R}$; call this number c_0. Any decision rule with a risk point which intersects the c_0 − square is minimax. (Note the the equation for the boundary of a cube is $\max_i R(\theta_i, d) = constant$. □

Theorem 3.2. If $\lambda = (\lambda_1, \cdots, \lambda_l)$ is a prior, then Bayes decision rules have risk points on the hyperplane $\{x \in R^l : \sum_i \lambda_i x_i = c^*\}$ where

$$c^* = \inf\{ c \geq 0 : \text{the plane determined by } \sum_i \lambda_i x_i = c \quad \text{intersects } \mathbf{R}\} .$$

Proof. $I\!R(\Lambda, d) = \sum \lambda_i R(\theta_i, d)$ is the Bayes risk and we want to minimize it. □

Lemma 3.3. An admissible rule is a Bayes rule for some prior λ .

Proof. Picture! □

Note that *not every* Bayes rule is admissible.

Theorem 3.4. Suppose that d_0 is Bayes w.r.t. $\lambda = (\lambda_1, \cdots, \lambda_l)$ and $\lambda_i > 0$, $i = 1, \cdots, l$. Then d_0 is admissible.

Proof. Suppose that d_0 is *not* admissible; then there is a rule d better than d_0 : i.e.

$$R(\theta_i, d) \leq R(\theta_i, d_0) \qquad \text{for all } i$$

with $<$ for some i . Since $\lambda_i > 0$ for $i = 1, \cdots, l$, contradicting the fact that d_0 is Bayes with respect to λ . □

Theorem 3.5. If $d_B \in \mathbf{D}$ is Bayes for λ and it has *constant risk*, then d_B is minimax.

Proof. Let r_0 be the constant risk. Assume d_B is *not* minimax and let d_m be the minimax rule (which exists). Then

(a) $$R(\theta_i, d_M) \leq \max_{1 \leq j \leq l} R(\theta_j, d_M) < \max_{1 \leq j \leq l} R(\theta_j, d_B) = r_0$$

and

(b) $$\min_{d \in \mathbf{D}} \mathbf{R}(\lambda, d) = \mathbf{R}(\lambda, d_B) = \sum \lambda_i R(\theta_i, d_B) = r_0 .$$

But (a) yields

(c) $$\mathbf{R}(\lambda, d_M) = \sum \lambda_i R(\theta_i, d_M) < \sum \lambda_i r_0 = r_0$$

which contradicts (b). □

4. POSTERIOR DISTRIBUTIONS

Now we will examine Bayes rules more generally.

Definition 4.1. If $\boldsymbol{\theta} \sim \Lambda$, a prior distribution over Θ, then the conditional distribution of $\boldsymbol{\theta}$ given X, $P(\boldsymbol{\theta} \in A \mid X)$, $A \in \mathbf{B}(\Theta)$, is called the *posterior* distribution of $\boldsymbol{\theta}$. Write $\Lambda(\theta \mid x) = P(\boldsymbol{\theta} \leq \theta \mid X = x)$ for the conditional distribution function if Θ is Euclidean.

If Λ has density λ with respect to ν and P_θ has density p_θ with respect to μ, then

$$\lambda(\theta \mid x) = \frac{p(x \mid \theta)\lambda(\theta)}{\int_\Theta p(x \mid \theta)\, \lambda(\theta)\, d\nu(\theta)} = \frac{p(x \mid \theta)\lambda(\theta)}{p(x)}$$

is the *posterior density* of $\boldsymbol{\theta}$ given $X = x$.

Definition 4.2. Suppose X has density $p(x \mid \theta)$, and $\lambda = \lambda(\,\cdot\,)$ is a prior density, $\lambda \in \mathbf{P}_\Lambda$. If the posterior density $\lambda(\,\cdot \mid x)$ has the same form (i.e. $\lambda(\,\cdot \mid x) \in \mathbf{P}_\Lambda$ for a.e. x), then λ is said to be a *conjugate prior* for $p(\,\cdot \mid \theta)$.

Examples 4.3.

Example 4.3.1. $(X \mid \boldsymbol{\theta} = \theta) \sim \mathit{Poisson}(\theta)$, $\boldsymbol{\theta} \sim \Gamma(\alpha, \beta)$. Then

$$\begin{aligned} \lambda(\theta \mid x) &= \frac{p(x \mid \theta)\, \lambda(\theta)}{p(x)} \propto p(x \mid \theta)\, \lambda(\theta) \\ &= e^{-\theta} \frac{\theta^x}{x!} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha - 1} e^{-\beta\theta} \\ &= \frac{\beta^\alpha}{x!\, \Gamma(\alpha)} \theta^{\alpha + x - 1} e^{-(\beta + 1)\theta}, \end{aligned}$$

so $(\boldsymbol{\theta} \mid X = x) \sim \Gamma(\alpha + x, \beta + 1)$, and gamma is a conjugate prior for Poisson.

Example 4.3.2. (Example 1.3, Lehmann, TPE, page 243). $(X \mid \boldsymbol{\theta} = \theta) \sim N(\theta, \sigma^2)$, $\boldsymbol{\theta} \sim N(\mu, \tau^2)$. Then

$$(\boldsymbol{\theta} \mid X = x) \sim N\Big(\frac{\mu/\tau^2 + x/\sigma^2}{1/\tau^2 + 1/\sigma^2}, \frac{1}{1/\tau^2 + 1/\sigma^2}\Big).$$

Note that if X is replaced by $X_1, \cdots, X_n$ iid $N(\theta, \sigma^2)$, then by sufficiency,

$$\begin{aligned} (\boldsymbol{\theta} \mid \underline{X} = \underline{x}) &= (\boldsymbol{\theta} \mid \overline{X} = \overline{x}) \\ &\sim N\Big(\frac{n/\sigma^2}{1/\tau^2 + n/\sigma^2} \overline{x} + \frac{1/\tau^2}{1/\tau^2 + n/\sigma^2} \mu, \frac{1}{1/\tau^2 + n/\sigma^2}\Big). \end{aligned}$$

Example 4.3.3. If $(X \mid \boldsymbol{\theta} = \theta) \sim \mathit{Binomial}(n, \theta)$, $\boldsymbol{\theta} \sim \mathrm{B}(\alpha, \beta)$, then

$$\begin{aligned}\lambda(\theta \mid x) &= \frac{p(x \mid \theta)\,\lambda(\theta)}{p(x)} \propto p(x \mid \theta)\,\lambda(\theta) \\ &= \binom{n}{x}\theta^2(1-\theta)^{n-x}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}\,B_{\alpha,\beta} \\ &= \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\,B_{\alpha,\beta}\binom{n}{x}\,,\end{aligned}$$

so $(\boldsymbol{\theta} \mid X = x) \sim \mathrm{B}(\alpha + x\,, \beta + n - x)$.

Example 4.3.4. Multinomial / Dirichlet. Suppose that $(X \mid \boldsymbol{\theta} = \theta) \sim \mathit{Multinomial}_k(n\,, \underline{\theta})$ and $\boldsymbol{\theta} \sim \mathit{Dirichlet}(\underline{\alpha})$; i.e.

$$\lambda(\theta) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\prod_{i=1}^k \Gamma(\alpha_i)}\,\theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

for $\theta_i \geq 0\,,\ \sum \theta_i = 1\,,\ \alpha_i > 0$. Then $(\underline{\boldsymbol{\theta}} \mid \underline{X} = \underline{x}) \sim \mathit{Dirichlet}(\underline{\alpha} + \underline{x})$,

$$E(\boldsymbol{\theta}) = \frac{\underline{\alpha}}{\sum \alpha_i}\,,$$

and

$$d_\Lambda(\underline{X}) = \frac{\underline{\alpha} + \underline{X}}{\sum \alpha_i + n} = \frac{\sum \alpha_i}{\sum \alpha_i + n}\,\frac{\underline{\alpha}}{\sum \alpha_i} + \frac{n}{\sum \alpha_i + n}\,\frac{\underline{X}}{n}\,.$$

5. FINDING BAYES RULES

For convex loss function we can restrict attention to non-randomized rules; see corollary 1.6.2, TPE, page 51. The following theorem gives a useful recipe for finding Bayes rules.

Theorem 5.1. Suppose that $\boldsymbol{\theta} \sim \Lambda$, $(X \,|\, \boldsymbol{\theta} = \theta) \sim P_\theta$, and $L(\theta, a) \geq 0$ for all $\theta \in \Theta$, $a \in \mathbf{A}$. If

(i) There exists a rule d_0 with finite risk, and

(ii) For a.e. x there exists $d_\Lambda(x)$ minimizing

$$E\{L(\boldsymbol{\theta}, d(\,\cdot\,|\,x)) \,|\, X = x\} = \int_\Theta L(\theta, d(\,\cdot\,|\,x))\, \lambda(\theta \,|\, x)\, d\nu(\theta) ,$$

then $d_\Lambda(\,\cdot\,|\,x)$ is a Bayes rule.

Proof. For any rule d with finite risk

(a) $$E\{L(\boldsymbol{\theta}, d(X)) \,|\, X\} \geq E\{L(\boldsymbol{\theta}, d_\Lambda(X)) \,|\, X\} \quad \text{a.s.}$$

by (ii). Hence

(b) $$\begin{aligned} I\!R(\Lambda, d) &= EE\{L(\boldsymbol{\theta}, d(X)) \,|\, X\} = E\, L(\boldsymbol{\theta}, d(X)) \\ &\geq EE\{L(\boldsymbol{\theta}, d_\Lambda(X)) \,|\, X\} = E\, L(\boldsymbol{\theta}, d_\Lambda(X)) \\ &\geq I\!R(\Lambda, d_\Lambda) . \end{aligned}$$

Corollary 5.2. (Estimation with weighted squared - error loss). If $\Theta = \mathbf{A}$ and $L(\theta, a) = K(\theta)|\theta - a|^2$, then

$$d_\Lambda(x) = \frac{E\{K(\boldsymbol{\theta})\boldsymbol{\theta} \,|\, X = x\}}{E\{K(\boldsymbol{\theta}) \,|\, X = x\}} = \frac{\int \theta\, K(\theta)\, d\Lambda(\theta \,|\, x)}{\int K(\theta)\, d\Lambda(\theta \,|\, x)} .$$

When $K(\theta) \equiv 1$, then

$$d_\Lambda(x) = \int \theta\, d\Lambda(\theta \,|\, x) = E\{\boldsymbol{\theta} \,|\, X = x\} \equiv \text{the posterior mean}$$

Proof. For an arbitrary (nonrandomized) rule $d \in \mathbf{D}$

$$\begin{aligned} \int K(\theta)\,|\theta - d(x)|^2\, d\Lambda(\theta \,|\, x) &= \int K(\theta)\,|\theta - d_\Lambda(x) + d_\Lambda(x) - d(x)|^2\, d\Lambda(\theta \,|\, x) \\ &= \int K(\theta)\,|\theta - d_\Lambda(x)|^2\, d\Lambda(\theta \,|\, x) \\ &\quad + 2(d_\Lambda(x) - d(x)) \int K(\theta)\{\theta - d_\Lambda(x)\}\, d\Lambda(\theta \,|\, x) \end{aligned}$$

$$+ \ (d_\Lambda(x) - d(x))^2 \int K(\theta)\, d\Lambda(\theta \,|\, x)$$

$$\geq \int K(\theta)\, |\theta - d_\Lambda(x)|^2 \, d\Lambda(\theta \,|\, x)$$

with equality if $d(x) = d_\Lambda(x)$. □

Corollary 5.3. (Estimation with $L_1 -$ loss). If $\Theta = \mathbf{A}$ and $L(\theta, a) = |\theta - a|$, then

$$d_\Lambda(x) = \text{any median of } \Lambda(\theta \,|\, x) .$$

Corollary 5.4. (Testing with $0 - 1$ loss). If $\mathbf{A} = \{0, 1\}$, $\Theta = \Theta_0 + \Theta_1$, $L(\theta, a_i) = l_i 1_{\Theta_i^c}(\theta)$, $i = 0, 1$, then any rule of the form

$$d_\Lambda(x) = \begin{cases} 1 & P(\boldsymbol{\theta} \in \Theta_1 \,|\, X = x) > (l_1/l_0) P(\boldsymbol{\theta} \in \Theta_0 \,|\, X = x) \\ \gamma(x) \ \text{ if } & P(\boldsymbol{\theta} \in \Theta_1 \,|\, X = x) = (l_1/l_0) P(\boldsymbol{\theta} \in \Theta_0 \,|\, X = x) \\ 0 & P(\boldsymbol{\theta} \in \Theta_1 \,|\, X = x) < (l_1/l_0) P(\boldsymbol{\theta} \in \Theta_0 \,|\, X = x) \end{cases}$$

is Bayes with respect to Λ. Note that this reduces to a test of the Neyman - Pearson form when $\Theta_i = \{\theta_i\}$, $i = 0, 1$.

Proof. Let $\phi(x) = d(1 \,|\, x)$. Then

$$E\{L(\boldsymbol{\theta}, \phi(x)) \,|\, X = x\} = \int_\Theta L(\theta, \phi(x))\, d\Lambda(\theta \,|\, x)$$

$$= \int_\Theta \{l_1 \phi(x) 1_{\Theta_0}(\theta) + l_0 (1 - \phi(x)) 1_{\Theta_1}(\theta)\}\, d\Lambda(\theta \,|\, x)$$

$$= l_1 \phi(x)\, P(\boldsymbol{\theta} \in \Theta_0 \,|\, X = x) + l_0 (1 - \phi(x))\, P(\boldsymbol{\theta} \in \Theta_1 \,|\, X = x)$$

$$= l_0\, P(\boldsymbol{\theta} \in \Theta_1 \,|\, X = x) + \phi(x) \{l_1 P(\boldsymbol{\theta} \in \Theta_0 \,|\, X = x) - l_0 P(\boldsymbol{\theta} \in \Theta_1 \,|\, X = x)\}$$

which is minimized by any rule of the form d_Λ. □

Corollary 5.5. (Testing with linear loss). If $\Theta = R$, $\mathbf{A} = \{0, 1\}$, $\Theta_0 = (-\infty, \theta_0]$, $\Theta_1 = (\theta_0, \infty)$ and

$$L(\theta, 0) = (\theta - \theta_0) 1_{(\theta_0, \infty)}(\theta) ; \quad L(\theta, 1) = (\theta_0 - \theta) 1_{(-\infty, \theta_0]}(\theta) ;$$

then

$$d_\Lambda(x) = \begin{cases} 1 & E(\boldsymbol{\theta} \,|\, X = x) > \theta_0 \\ \gamma(x) \ \text{ if } & E(\boldsymbol{\theta} \,|\, X = x) = \theta_0 \\ 0 & E(\boldsymbol{\theta} \,|\, X = x) < \theta_0 \end{cases}$$

is Bayes with respect to Λ.

Proof. Again, by theorem 5.1 it suffices to minimize

$$E\{L(\boldsymbol{\theta}, \phi(x)) \,|\, X = x\}$$

$$= \int_{\Theta} \{\phi(x)(\theta_0 - \theta)1_{(-\infty,\theta_0]}(\theta) + (1-\phi(x))(\theta-\theta_0)1_{(\theta_0,\infty)}(\theta)\} d\Lambda(\theta \mid x)$$

$$= \int_{\Theta} (\theta - \theta_0)1_{(-\infty,\theta_0]}(\theta)\, d\Lambda(\theta \mid x) + (1 - \phi(x))\{E(\boldsymbol{\theta} \mid X = x) - \theta_0\}$$

which is minimized for each fixed x by any rule of the form d_Λ. □

More examples 5.6: Estimation with squared error and weighted squared error loss.

Example 5.6.1. Binomial(n, θ) with $\boldsymbol{\theta} \sim Beta(\alpha, \beta)$. Thus $(\boldsymbol{\theta} \mid X = x) \sim \mathrm{B}(\alpha + x, \beta + n - x)$, so that

$$E(\boldsymbol{\theta}) = \frac{\alpha}{\alpha+\beta}$$

and

$$d_\Lambda(X) = \frac{\alpha + X}{\alpha+\beta+n}$$

$$= \frac{\alpha+\beta}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\frac{X}{n}.$$

If the loss is $L(\theta, a) = (\theta - a)^2/\{\theta(1-\theta)\}$ instead of squared - error loss, then

$$E\{\boldsymbol{\theta}K(\boldsymbol{\theta}) \mid X = x\} = \frac{B(\alpha+x, \beta+n-x-1)}{B(\alpha+x, \beta+n-x)},$$

$$E\{K(\boldsymbol{\theta}) \mid X = x\} = \frac{B(\alpha+x-1, \beta+n-x-1)}{B(\alpha+x, \beta+n-x)},$$

and hence the Bayes rule with respect to Λ for this loss function is

$$d_\Lambda(X) = \frac{B(\alpha+X, \beta+n-X-1)}{B(\alpha+X-1, \beta+n-X-1)} = \frac{\alpha+X-1}{\alpha+\beta+n-2}$$

$$= \lambda_n(\alpha,\beta)\frac{\alpha-1}{\alpha+\beta-2} + (1-\lambda_n)(\alpha,\beta)\frac{X}{n}$$

where $\lambda_n(\alpha, \beta) = (\alpha+\beta-2)/(\alpha+\beta+n-2)$. Note that when $\alpha = 1 = \beta$, the Bayes estimator for this loss function becomes the familiar maximum likelihood estimator $\hat{p} = X/n$.

Example 5.6.2. Multinomial/Dirichlet. If $(\underline{X} \mid \underline{\theta}) \sim Multinomial_k(n, \underline{\theta})$, $\underline{\boldsymbol{\theta}} \sim Dirichlet(\underline{\alpha})$, then

$$E(\underline{\boldsymbol{\theta}}) = \frac{\underline{\alpha}}{\sum \alpha_i},$$

and

$$d_\Lambda(\underline{X}) = E(\underline{\boldsymbol{\theta}} \mid \underline{X}) = \frac{\underline{\alpha} + \underline{X}}{\sum \alpha_i + n}.$$

Example 5.6.3. Normal with normal prior. If $(X \mid \boldsymbol{\theta} = \theta) \sim N(\theta, \sigma^2)$, $\boldsymbol{\theta} \sim N(\mu, \tau^2)$, then

$$(\boldsymbol{\theta} \mid X) \sim N\left(\frac{1/\tau^2}{1/\tau^2 + 1/\sigma^2}\,\mu + \frac{1/\sigma^2}{1/\tau^2 + 1/\sigma^2}\,X\,,\ \frac{1}{1/\tau^2 + 1/\sigma^2}\right).$$

Consequently

$$E(\boldsymbol{\theta}) = \mu\,,$$

and

$$d_\Lambda(X) = E(\boldsymbol{\theta} \mid X) = \frac{1/\tau^2}{1/\tau^2 + 1/\sigma^2}\,\mu + \frac{1/\sigma^2}{1/\tau^2 + 1/\sigma^2}\,X \qquad .$$

This remains true if $L(\theta, a) = \rho(\theta - a)$ where ρ is convex and even (see e.g. Lehmann, TPE, page 244).

5.6

6. FINDING MINIMAX RULES

Definition 6.1. A prior Λ_0 for which $I\!R(\Lambda, d_\Lambda)$ is maximized is called a *least favorable prior*:

$$I\!R(\Lambda_0, d_{\Lambda_0}) = \sup_\Lambda I\!R(\Lambda, d_\Lambda) .$$

Theorem 6.2. Suppose that Λ is a prior distribution on Θ such that

$$(1) \qquad I\!R(\Lambda, d_\Lambda) = \int_\Theta R(\theta, d_\Lambda)\, d\Lambda(\theta) = \sup_\theta R(\theta, d_\Lambda) .$$

Then:

(i) d_Λ is minimax.

(ii) If d_Λ is unique Bayes with respect to Λ, d_Λ is unique minimax.

(iii) Λ is least favorable.

Proof. (i) Let d be another rule. Then

$$(a) \qquad \sup_\theta R(\theta, d) \geq \int_\Theta R(\theta, d) d\Lambda(\theta)$$

$$(b) \qquad \geq \int_\Theta R(\theta, d_\Lambda)\, d\Lambda(\theta)$$

since d_Λ is Bayes wrt Λ

$$= \sup_\theta R(\theta, d_\Lambda) \qquad \text{by (1)} .$$

Hence d_Λ is minimax.

(ii). If d_Λ is unique Bayes, then $>$ holds in (b), so d_Λ is unique minimax.

(iii). Let Λ^* be some other prior distribution. Then

$$r_{\Lambda^*} \equiv \int_\Theta R(\theta, d_{\Lambda^*})\, d\Lambda^*(\theta) \leq \int_\Theta R(\theta, d_\Lambda)\, d\Lambda^*(\theta)$$

since d_{Λ^*} is Bayes wrt Λ^*

$$\leq \sup_\theta R(\theta, d_\Lambda)$$

$$= I\!R(\Lambda, d_\Lambda) \equiv r_\Lambda \qquad \text{by (1).}$$

□

Corollary 6.3. If d_Λ is Bayes with respect to Λ and has constant risk, $R(\theta, d_\Lambda) =$ constant, then d_Λ is minimax.

Proof. If d_Λ has constant risk, then (1) holds. □

Corollary 6.4. Let

$$\Theta_\Lambda \equiv \{\theta \in \Theta : R(\theta, d_\Lambda) = \sup_{\theta'} R(\theta', d_\Lambda)\}$$

be the set of θ's where the risk of d_Λ assumes its maximum. Then d_Λ is minimax if $\Lambda(\Theta_\Lambda) = 1$. Equivalently, d_Λ is minimax if there is a set Θ_Λ with $\Lambda(\Theta_\Lambda) = 1$ and $R(\theta, d_\Lambda) = \sup_{\theta'} R(\theta', d_\Lambda)$ for all $\theta \in \Theta_\Lambda$.

Example 6.5. Binomial(n, θ) with $\boldsymbol{\theta} \sim Beta(\alpha, \beta)$. Thus $(\boldsymbol{\theta} \mid X = x) \sim \mathrm{B}(\alpha + x, \beta + n - x)$, and as we computed in section 5,

$$d_\Lambda(X) = \frac{\alpha + X}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n}\frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n}\frac{X}{n}.$$

Consequently,

$$\begin{aligned} R(\theta, d_\Lambda) &= (\frac{n}{\alpha + \beta + n})^2 \frac{\theta(1-\theta)}{n} + (\frac{\alpha + n\theta}{\alpha + \beta + n} - \theta)^2 \\ &= \frac{1}{(\alpha + \beta + n)^2}\{n\theta(1-\theta) + (\alpha - \alpha\theta - \beta\theta)^2\} \\ &= \frac{1}{(\alpha + \beta + n)^2}\{\alpha^2 + (n - 2\alpha(\alpha + \beta))\theta + ((\alpha + \beta)^2 - n)\theta^2\} \\ &= \frac{1}{(\alpha + \beta + n)^2}\{\alpha^2 + (n - 2\alpha(\alpha + \beta))\theta + ((\alpha + \beta)^2 - n)\theta^2\} \\ &= \frac{1}{(\alpha + \beta + n)^2}\alpha^2 \end{aligned}$$

if $2\alpha(\alpha + \beta) = n$ and $(\alpha + \beta)^2 = n$. But solving these two equations yields $\alpha = \beta = \sqrt{n}/2$. Thus for these choices of α and β, the risk of the Bayes rule is constant in θ, and hence

$$d_M(X) = \frac{1}{1 + \sqrt{n}}\frac{1}{2} + \frac{\sqrt{n}}{1 + \sqrt{n}}\frac{X}{n}$$

is Bayes with respect to $\Lambda = \mathrm{Beta}(\sqrt{n}/2, \sqrt{n}/2)$ and has risk

$$R(\theta, d_M) = \frac{1}{4(1 + \sqrt{n})^2}.$$

Hence $\mathrm{Beta}(\sqrt{n}/2, \sqrt{n}/2)$ is least favorable and d_M is minimax. Note that d_M is a consistent estimator of θ:

$$d_M(X) \to_p \theta$$

as $n \to \infty$, but its bias is of the order $n^{-1/2}$ (the bias equals $(1/2 - \theta)/(1 + \sqrt{n})$. Furthermore d_M is a (locally) regular estimator of θ, but the limit distribution is *not* centered at zero if $\theta \neq 1/2$: under P_{θ_n}

$$\sqrt{n}(d_M(X) - \theta_n) \to_d N(1/2 - \theta, \theta(1-\theta)).$$

If a least - favorable prior does not exist, then we can still consider improper priors or limits of proper priors:

Let $\{\Lambda_k\}$ be a sequence of prior distributions, let d_k denote the Bayes estimator corresponding to Λ_k, and set

$$r_k \equiv \int_\Theta R(\theta, d_k)\, d\Lambda_k(\theta) .$$

Suppose that

(2) $$r_k \to r < \infty \quad \text{as} \quad k \to \infty .$$

Definition 6.6. The sequence of prior distributions $\{\Lambda_k\}$ with Bayes risks $\{r_k\}$ is said to be *least favorable* if $r_\Lambda \equiv I\!R(\Lambda, d_\Lambda) \le r$ for any prior distribution Λ.

Theorem 6.7. Suppose that $\{\Lambda_k\}$ is a sequence of prior distributions with Bayes risks satisfying

(i) $r_k \to r$,

and d is an estimator for which

(ii) $\sup_\theta R(\theta, d) = r$.

Then:

A. d is minimax, and

B. $\{\Lambda_k\}$ is least - favorable.

Proof. A. Suppose d^* is any other estimator. Then

$$\sup_\theta R(\theta, d^*) \ge \int R(\theta, d^*)\, d\Lambda_k(\theta) \ge r_k$$

for all $k \ge 1$. Hence

$$\sup_\theta R(\theta, d^*) \ge r = \sup_\theta R(\theta, d) \quad \text{by (ii)},$$

so d is minimax.

B. If Λ is any prior distribution, then

$$\begin{aligned} r_\Lambda = \int R(\theta, d_\Lambda)\, d\Lambda(\theta) &\le \int R(\theta, d)\, d\Lambda(\theta) \\ &\le \sup_\theta R(\theta, d) = r \end{aligned}$$

by (ii). □

Example 6.8. Let $X_1, \cdots, X_n$ be iid $N(\theta, \sigma^2)$ given $\boldsymbol{\theta} = \theta$, and suppose that $\boldsymbol{\theta} \sim N(\mu, \tau^2)$ and that σ^2 is known. Then it follows from 5.6 that

$$d_\Lambda(\underline{X}) = E(\boldsymbol{\theta} \mid \underline{X}) = \frac{1/\tau^2}{1/\tau^2 + n/\sigma^2}\,\mu + \frac{n/\sigma^2}{1/\tau^2 + n/\sigma^2}\,\bar{X}$$

and

$$\begin{aligned}
I\!R(\Lambda, d_\Lambda) \equiv r_\Lambda &= E\{\boldsymbol{\theta} - d_\Lambda(\underline{X})\}^2 \\
&= EE\{[\boldsymbol{\theta} - d_\Lambda(\underline{X})]^2 \mid \underline{X}\} \\
&= E\,Var\{\boldsymbol{\theta} \mid \underline{X}\} \\
&= \frac{1}{1/\tau^2 + n/\sigma^2} \\
&\to \frac{\sigma^2}{n} \quad \text{as} \quad \tau^2 \to \infty \\
&= R(\theta, \overline{X})
\end{aligned}$$

for all $\theta \in \Theta$. Hence, by theorem 6.7, $\overline{X}$ is a minimax estimator of θ. □

The remainder of this section is aimed at extending minimaxity of estimators from smaller models to larger ones.

Lemma 6.9. Suppse that $X \sim P \in \mathbf{M} \equiv \{\text{all } P's \ \ on\, \mathbf{X}\}$, and that $\nu : \mathbf{P} \subset \mathbf{M} \to R$ is a functional (e.g. $\nu(P) = E_P X = \int x\, dP(x)$, $\nu(P) = Var_P(X)$, $\nu(P) = Pf$ for some fixed function f; or, in the case of $\mathbf{X} = R$, $\nu(P) = F_P^{-1}(t)$ for fixed $t \in (0,1)$ where $F_P(x) = P(X \le x)$. Suppose that d is a minimax estimator of $\nu(P)$ for $P \in \mathbf{P}_0 \subset \mathbf{P}_1$. If

$$\sup_{P \in \mathbf{P}_0} R(P, d) = \sup_{P \in \mathbf{P}_1} R(P, d),$$

then d is also minimax for estimating $\nu(P)$, $P \in \mathbf{P}_1$.

Proof. Suppose that d is not minimax. Then there exists a rule d^* with smaller maximum risk:

$$\sup_{P \in \mathbf{P}_0} R(P, d^*) \le \sup_{P \in \mathbf{P}_1} R(P, d^*) < \sup_{P \in \mathbf{P}_1} R(P, d) \le \sup_{P \in \mathbf{P}_0} R(P, d)$$

which contradict the hypothesis that d is minimax for $\mathbf{P}_0$. □

Example 6.10. This continues the normal mean example 6.8. Suppose that σ^2 is unknown. Thus, with $\Theta = \{(\theta, \sigma^2) : \theta \in R,\ 0 < \sigma^2 < \infty\}$,
So, to get something reasonable, we need to restrict σ^2. Let

$$\begin{aligned}
\mathbf{P}_0 &= \{N(\theta, M) : \theta \in R\}, \\
\mathbf{P}_1 &= \{N(\theta, \sigma^2) : \theta \in R,\ 0 \le \sigma^2 \le M\}.
\end{aligned}$$

Then $\overline{X}$ is minimax for $\mathbf{P}_0$ by our earlier calculation, $\mathbf{P}_0 \subset \mathbf{P}_1\}$, and

$$\sup_{\mathbf{P}_1} R(P, \overline{X}) = \sup_{\mathbf{P}_0} R(P, \overline{X}) = \frac{M}{n}.$$

Thus $\bar{X}$ is a minimax estimator of θ for $\mathbf{P}_1$. $\square$

Example 6.11. Let $X_1, \cdots, X_n$ be iid $P \in \mathbf{P}_\mu$ where

$$\mathbf{P}_\mu = \{\text{all probability measures } P : E_P|X| < \infty\}$$

and consider estimation of $\nu(P) = E_P(X)$ with squared error loss for the families

(i) $\mathbf{P}_{b\sigma^2} \equiv \{P \in \mathbf{P}_\mu : Var_P(X) \le M < \infty\}$.

(ii) $\mathbf{P}_{br} \equiv \{P \in \mathbf{P}_\mu : P(a \le X \le b) = 1\}$ for some fixed $a, b \in R$.

Then:

A. $\bar{X}$ is minimax for $\mathbf{F}_{b\sigma^2} = \mathbf{P}_1$ by 6.9 since it is minimax for $\mathbf{P}_0 \equiv \{N(\theta, \sigma^2) : \theta \in R, 0 < \sigma^2 \le M\}$ and

$$\sup_{P \in \mathbf{P}_0} R(P, \bar{X}) = \sup_{P \in \mathbf{P}_1} R(P, \bar{X}).$$

B. Without loss of generality suppose $a = 0$ and $b = 1$. Let

$$\mathbf{P}_1 = \{P : P([0,1]) = 1\},$$

$$\mathbf{P}_0 = \{P \in \mathbf{P}_1 : P(X=1) = p, \; P(X=0) = 1-p \text{ for some } 0 < p < 1\}.$$

For $\mathbf{P}_0$ we know that the minimax estimator is

$$d_M(\underline{X}) = \frac{1}{1+\sqrt{n}}\frac{1}{2} + \frac{\sqrt{n}}{1+\sqrt{n}}\bar{X}.$$

Now, with $E_P X \equiv \theta$,

$$\begin{aligned} R(P, d_M) &= \frac{1}{(1+\sqrt{n})^2}\{Var_P(X) + (\frac{1}{2} - \theta)^2\} \\ &\le \frac{1}{(1+\sqrt{n})^2}\{\theta - \theta^2 + \frac{1}{4} - \theta + \theta^2 \\ &\qquad \text{since } 0 \le X \le 1 \quad \text{implies } E_P X^2 \le E_P X \equiv \theta \\ &= \frac{1/4}{(1+\sqrt{n})^2}. \end{aligned}$$

Thus

$$\sup_{P \in \mathbf{P}_1} R(P, d_M) = \sup_{P \in \mathbf{P}_0} R(P, d_M),$$

and by 6.9 d_M is minimax.

Notes. 1. The restriction $\sigma^2 \le M$ in 6.10 (i) is crucial; note that $\nu(P) = E_P(X)$ is discontinuous at every P in the sense that we can easily have $P_m \to_d P$, but $\nu(P_m)$ fails to converge to $\nu(P)$.

2. If $\Theta = [-M, M] \subset R$, then $\bar{X}$ is no longer minimax in the normal case;

see Bickel (1981), *Ann. Statist.* **9,** 1301 - 1309. (The approximately least favorable densities for $M \to \infty$ are $\lambda_M(\theta) = M^{-1}\cos^2(\frac{\pi}{2}\frac{\theta}{M})\, 1_{[-M,M]}(\theta)$.)

5.7

7. ADMISSIBILITY AND INADMISSIBILITY

Our goal in this section is to establish some basic and simple results about admissibility / inadmissibility of some classical estimators. In particular, we will show that $\bar{X}$ is admissible for the Gaussian location model for $k = 1$, but inadmissible for the Gaussian location model for $k \geq 3$. The fundamental work in this area is that of Stein (1956) and James and Stein (1961). For an interesting expository paper, see Efron and Morris (1977); for current work motivated by admissibility issues see e.g. Perlman and Marden (1990) or Strawderman and Brandwein (1991).

Theorem 7.1. Any unique Bayes estimator is admissible.

Proof. Suppose that d_Λ is unique Bayes with respect to Λ and is inadmissible. Then there exists an estimator d such that $R(\theta, d) \leq R(\theta, d_\Lambda)$ with strict inequality for some θ , and hence

$$\int_\Theta R(\theta, d)\, \Lambda(\theta) \;\leq\; \int_\Theta R(\theta, d_\Lambda)\, d\Lambda(\theta)$$

which contradicts uniqueness of d_Λ . Thus d_Λ is admissible. □

Lemma 7.2. If the loss function $L(\theta, a)$ is squared error (or is convex in a) and, with Q defined by $Q(A) = \int P_\theta(X \in A)\, d\Lambda(\theta)$, a.e. Q implies a.e. $\mathbf{P}$, then a Bayes rule with finite Bayes risk is unique a.e. $\mathbf{P}$.
[a.e. $\mathbf{P}$ means $P(N) = 0$ for all $P \in \mathbf{P}$.]

Proof. See Lehmann TPE Corollary 4.1.2, page 240. □

Example 7.3. Consider the Bayes estimator of a normal mean, $d_\Lambda = p_n\mu + (1 - p_n)\bar{X}$, $p_n = (1/\tau^2)/(1/\tau^2 + n/\sigma^2)$. The Bayes risk is finite and a.e. Q implies a.e. $\mathbf{P}$. Hence d_Λ is unique Bayes and admissible.

Theorem 7.4. If X is a random variable with mean θ and variance σ^2 , then $aX + b$ is *inadmissible* as an estimator of θ for squared error loss if:

(i) $a > 1$

(ii) $a < 0$

(iii) $a = 1 ,\ b \neq 0$.

Proof. For any a , b the risk of the rule $aX + b$ is

$$R(\theta\,, aX + b) \;=\; a^2\sigma^2 \;+\; \{(a-1)\theta + b\}^2 \;\equiv\; \rho(a, b)\,.$$

(i) If $a > 1$

$$\rho(a, b) \;\geq\; a^2\sigma^2 \;>\; \sigma^2 \;=\; \rho(1, 0)\,,$$

so $aX + b$ is dominated by X .
(ii) If $a < 0$, then $(a-1)^2 > 1$ and

$$\rho(a, b) \;\geq\; \{(a-1)\theta + b\}^2 \;=\; (a-1)^2\{\theta \;+\; \frac{b}{a-1}\}^2$$

$$> \{\theta \;+\; \frac{b}{a-1}\}^2 \;=\; \rho(0\,, -\,\frac{b}{a-1})\,.$$

(iii). If $a = 1$, $b \neq 0$,

$$\rho(1, b) \;=\; \sigma^2 \;+\; b^2 \;>\; \sigma^2 \;=\; \rho(1, 0)\,,$$

so $X + b$ is dominated by X . □

Example 7.5, continued. Thus $a\bar{X} + b$ is inadmissible for $a < 0$ or $a > 1$. When $a = 0$, $d = b$ is admissible since it is the only estimator with zero risk at $\theta = b$. When $a = 1$, $b \neq 0$, d is inadmissible. What is left is $\bar{X}$.

Theorem 7.6. If $X_1, \cdots, X_n$ are iid $N(\theta\,, \sigma^2)$, $\theta \in \Theta = R$ with σ^2 known, then $\bar{X}$ is an admissible estimator of θ .

Proof A. (Limiting Bayes method). Suppose that $\bar{X}$ is inadmissible (and $\sigma = 1$). Then there is an estimator d^* such that $R(\theta\,, d^*) \leq \dfrac{1}{n} = R(\theta\,, \bar{X})$ for all θ with risk $<$ for some θ . Now

$$R(\theta\,, d) \;=\; E(\theta - d(X))^2$$

is continuous in θ for every d , and hence there exists $\epsilon > 0$ and $\theta_0 < \theta_1$ so that

$$R(\theta\,, d^*) \;<\; \frac{1}{n} - \epsilon \qquad \text{for all} \;\; \theta_0 < \theta < \theta_1\,.$$

Let

$$r^*_\tau \;\equiv\; \int_\Theta R(\theta\,, d^*)\, d\Lambda(\theta)$$

where $\Lambda = N(0\,, \tau^2)$. Thus

$$r_\tau \;\equiv\; \int_\Theta R(\theta\,, d_\tau)\, d\Lambda(\theta) \;=\; \frac{1}{1/\tau^2 + n} \;=\; \frac{\tau^2}{1 + n\tau^2}$$

so $r_\tau \leq r^*_\tau$. Thus

$$\frac{1/n - r_\tau^*}{1/n - r_\tau} = \frac{\frac{1}{\sqrt{2\pi}\tau}\int_{-\infty}^{\infty}\{\frac{1}{n} - R(\theta, d^*)\}\exp(-\theta^2/2\tau^2)\,d\theta}{1/n - \tau^2/(1+n\tau^2)}$$

$$\geq \frac{\frac{1}{\sqrt{2\pi}\tau}\,\epsilon\int_{\theta_0}^{\theta_1}\exp(-\theta^2/2\tau^2)\,d\theta}{1/n(1+n\tau^2)}$$

$$= \frac{n(1+n\tau^2)}{\sqrt{2\pi}\tau}\,\epsilon\int_{\theta_0}^{\theta_1}\exp(-\theta^2/2\tau^2)\,d\theta$$

$$\to \infty\,(\theta_1 - \theta_0) = \infty \qquad \text{as} \quad \tau \to \infty\,.$$

Hence

$$\frac{1}{n} - r_\tau^* > \frac{1}{n} - r_\tau$$

for $\tau >$ some τ_0 , or,

$$r_\tau^* < r_\tau$$

for $\tau >$ some τ_0 , which contradicts d_τ Bayes wrt Λ_τ with Bayes risk r_τ . Hence $\overline{X}$ is admissible. □

Proof B. (Information inequality method). The risk of any rule d is

$$R(\theta, d) = E_\theta(d - \theta)^2 = Var_\theta[d(X)] + b^2(\theta)$$

$$\geq \frac{[1+b'(\theta)]^2}{n} + b^2(\theta)$$

since $I(\theta) = 1$. Suppose that d is any estimator satisfying

$$R(\theta, d) \leq 1/n \qquad \text{for all} \quad \theta\,.$$

Then

(a) $$\frac{1}{n} \geq b^2(\theta) + \frac{[1+b'(\theta)]^2}{n}\,.$$

But (a) implies that:

(i) $|b(\theta)| \leq 1/\sqrt{n}$ for all θ ; i.e. b is bounded.

(ii) $b'(\theta) \leq 0$ since $1 + 2b'(\theta) + [b'(\theta)]^2 \leq 1$.

(iii) There exists $\theta_i \to \infty$ such that $b'(\theta_i) \to 0$. [If $b'(\theta) \leq -\epsilon$ for all $\theta > \theta_0$, then $b(\theta)$ is not bounded.]

(iv)

$$b^2(\theta) \leq \frac{1}{n} - \frac{[1+b'(\theta)]^2}{n}$$

$$= -\frac{2b'(\theta) + [b'(\theta)]^2}{n}$$

Hence $b(\theta) = 0$, and $R(\theta, d) \equiv 1/n$. □

Now let $X_1, \cdots, X_n$ be iid $N_k(\theta, I)$, and consider estimation of $\theta \in R^k$ with loss function $L(\theta, a) = |\theta - a|^2$. Let $\bar{X} = \sum_{i=1}^{n} X_i/n$. Note that

$$R(\theta, \bar{X}) = E_\theta|\bar{X} - \theta|^2 = \frac{k}{n} \quad \text{for all } \theta .$$

Theorem 7.7. If $k \geq 3$, then $\bar{X}$ is inadmissible.

Proof. Let $g: R^k \to R^k$ have $E|\frac{\partial}{\partial x_i} g_i(X)| < \infty$, and consider estimators of the form

$$\hat{\theta}_n = \bar{X} + n^{-1} g(\bar{X}) .$$

Now

$$\begin{aligned} & E_\theta|\bar{X} - \theta|^2 - E_\theta|\hat{\theta}_n - \theta|^2 \\ &\quad = E_\theta|\bar{X} - \theta|^2 - E_\theta|\bar{X} - \theta + n^{-1} g(\bar{X})|^2 \\ \text{(a)} \qquad &\quad = -2n^{-1} E_\theta < \bar{X} - \theta, g(\bar{X}) > - n^{-2} E_\theta |g(\bar{X})|^2 . \end{aligned}$$

To proceed further we need an identity due to Stein.

Lemma 7.8. If $X \sim N(\theta, \sigma^2)$, and g is a function with $E|g'(X)| < \infty$, then

(i) $$\sigma^2 E\, g'(X) = E(X - \theta) g(X) .$$

If $X \sim N_k(\theta, \sigma^2 I)$, and $g: R^k \to R^k$ has

$$E|\frac{\partial}{\partial x_i} g_i(X)| < \infty ,$$

then:

(ii) $$\sigma^2 E \frac{\partial}{\partial x_i} g_i(X) = E(X_i - \theta_i) g_i(X) , \qquad i = 1, \cdots, k .$$

Proof. Without loss of generality, suppose that $\theta = 0$ and $\sigma^2 = 1$. Integration by parts gives

$$\begin{aligned} Eg'(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g'(x)\, exp(-\frac{x^2}{2})\, dx \\ &= \frac{-1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x)\, exp(-\frac{x^2}{2}) \{ -\frac{2x}{2} \}\, dx \\ &= EXg(X) . \end{aligned}$$

[In applying integration by parts here, we have ignored the term

$uv|_{-\infty}^{\infty} = g(x)\phi(x)|_{-\infty}^{\infty}$. To prove that this does in fact vanish, it is easiest to apply Fubini's theorem (twice!) in a slightly devious way as follows. Let $\phi(t) = (2\pi)^{-1/2}\exp(-t^2/2)$ be the standard normal density. Since $\phi'(t) = -t\phi(t)$, we have both

$$\phi(x) = -\int_x^\infty \phi'(t)\,dt = \int_x^\infty t\phi(t)\,dt$$

and

$$\phi(x) = \int_{-\infty}^x \phi'(t)\,dt = -\int_{-\infty}^x t\phi(t)\,dt .$$

Therefore we can write

$$\begin{aligned}
Eg'(X) &= \int_{-\infty}^\infty g'(x)\phi(x)\,dx \\
&= \int_0^\infty g'(x)\int_x^\infty t\phi(t)dt\,dx - \int_{-\infty}^0 g'(x)\int_{-\infty}^x t\phi(t)dt\,dx \\
&= \int_0^\infty t\phi(t)\left\{\int_0^t g'(x)dx\right\}dt - \int_{-\infty}^0 t\phi(t)\left\{\int_t^0 g'(x)dx\right\}dt \\
&= \left(\int_0^\infty + \int_{-\infty}^0\right)\{t\phi(t)[g(t)-g(0)]\}\,dt \\
&= \int_{-\infty}^\infty tg(t)\phi(t)\,dt = EXg(X) .
\end{aligned}$$

Here the third equality is justified by the hypothesis $E|g'(X)| < \infty$ and Fubini's theorem.]

To prove (ii), write $X^{(i)} \equiv (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_k)$. Then

$$\begin{aligned}
\sigma^2 E\frac{\partial}{\partial x_i} g_i(X) &= \sigma^2 EE\{\frac{\partial}{\partial x_i} g_i(X) \mid X^{(i)}\} \\
&= EE\{(X_i - \theta_i)g_i(X) \mid X^{(i)}\} \qquad \text{by (i)} \\
&= E(X_i - \theta_i)g_i(X) .
\end{aligned}$$

□

Now we return to the proof of the theorem. Using (ii) in (a) shows that the right side of (a) equals

$$\text{(b)} \qquad -2n^{-2}E_\theta\sum_{i=1}^k \frac{\partial g_i}{\partial x_i}(\bar{X}) - n^{-2}E_\theta|g(X)|^2 .$$

Now let $\psi : R^k \to R$ be twice differentiable, and set

$$g(x) = \nabla\{\log\psi(x)\} = \frac{1}{\psi(x)}\left(\frac{\partial\psi(x)}{\partial x_1}, \cdots, \frac{\partial\psi(x)}{\partial x_k}\right) .$$

Thus

$$\frac{\partial g_i}{\partial x_i}(x) = \frac{1}{\psi(x)} \frac{\partial^2}{\partial x_i^2} \psi(x) - \frac{1}{\psi^2(x)} \left(\frac{\partial \psi(x)}{\partial x_i}\right)^2$$
$$= \frac{1}{\psi(x)} \frac{\partial^2}{\partial x_i^2} \psi(x) - g_i(x)^2$$

and

$$\sum_{i=1}^{k} \frac{\partial g_i}{\partial x_i}(x) = \frac{1}{\psi(x)} \nabla^2 \psi(x) - |g(x)|^2 .$$

Hence the right side of (b) is

$$\text{(c)} \qquad = n^{-2} E_\theta |g(\bar{X})|^2 - 2\, n^{-2} E_\theta \{ \frac{1}{\psi(\bar{X})} \nabla^2 \psi(\bar{X}) \}$$
$$> 0$$

if $\psi(x) \geq 0$ and $\nabla^2 \psi \leq 0$, $g \neq 0$ (i.e. ψ superharmonic). Here is one example of such a function:

A. Suppose that $\psi(x) = |x|^{-(k-2)} = \{x_1^2 + \cdots + x_k^2\}^{-(k-2)/2}$. then

$$g(x) = \nabla \log \psi(x) = -\frac{k-2}{|x|^2} x$$

and

$$\nabla^2 \psi(x) = 0 ,$$

so ψ is *harmonic*. Thus

$$\hat{\theta}_n = (1 - \frac{k-2}{n|\bar{X}|^2})\bar{X}$$

and

$$E|\bar{X} - \theta|^2 - E|\hat{\theta} - \theta|^2 = n^{-2} E_\theta |g(\bar{X})|^2$$
$$= \left(\frac{k-2}{n}\right)^2 E_\theta |\bar{X}|^{-2}$$
$$= \left(\frac{k-2}{\sqrt{n}}\right)^2 E_\theta |\sqrt{n}(\bar{X} - \theta) + \sqrt{n}\theta|^{-2}$$
$$= \left(\frac{k-2}{\sqrt{n}}\right)^2 E_0 |X + \sqrt{n}\theta|^{-2}$$
$$= \begin{cases} \left(\frac{k-2}{n}\right)^2 E_0 |n^{-1/2} X + \theta|^{-2} = O(n^{-2}) , & \theta \neq 0 \\ \frac{(k-2)^2}{n} \frac{1}{k-2} = \frac{k-2}{n} , & \theta = 0 \end{cases}$$

since $|X|^2 \sim \chi_k^2$ with $E(1/\chi_k^2) = 1/(k-2)$. Hence

$$\frac{E_0|\hat{\theta} - \theta|^2}{E_0|\bar{X} - \theta|^2} = \frac{2}{k} < 1 .$$

For general θ ,

$$\begin{aligned} R(\theta, \hat{\theta}) &= \frac{k}{n} - \frac{(k-2)^2}{n} E_0\Big(\frac{1}{|X + \sqrt{n}\theta|^2}\Big) \\ &= \frac{k}{n}\Big(1 - \frac{k-2}{k} E_0\Big(\frac{k-2}{\chi_k^2(\delta^2)}\Big)\Big) \end{aligned}$$

since $|X + \sqrt{n}\theta|^2 \sim \chi_k^2(\delta^2)$ with $\delta^2 = n|\theta|^2/2$. Thus

$$\begin{aligned} \frac{R(\theta, \hat{\theta})}{R(\theta, \bar{X})} &= \Big(1 - \frac{k-2}{k} E_0\Big(\frac{k-2}{\chi_k^2(\delta^2)}\Big)\Big) \\ &= 2/k \quad \text{when} \quad \theta = 0 , \\ &\to 1 \quad \text{as} \quad n \to \infty \quad \text{for fixed} \quad \theta \neq 0 , \\ &\to 1 \quad \text{as} \quad |\theta| \to \infty \quad \text{for fixed} \quad n . \end{aligned}$$

$\square$

Remark 7.9. Note that the James - Stein estimator

$$\hat{\theta}_n = \Big(1 - \frac{k-2}{n|\bar{X}|^2}\Big)\bar{X}$$

is not regular at $\theta = 0$: if $\theta_n = tn^{-1/2}$, then $\sqrt{n}(\bar{X} - \theta_n) \stackrel{d}{=} Z \sim N_k(0, \sigma^2 I)$ under P_{θ_n} , so that

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta) &= \sqrt{n}(\bar{X} - \theta) - \frac{k-2}{|\sqrt{n}(\bar{X} - \theta_n) + t|^2}\{\sqrt{n}(\bar{X} - \theta_n) + t\} \\ &= Z - \frac{k-2}{|Z + t|^2}(Z + t) \end{aligned}$$

which has a distribution dependent on t .

Remark 7.10. It turns out that the James - Stein estimator $\hat{\theta}_n$ is itself inadmissible; see Lehmann TPE pages 295, 299, 302-307.

Remark 7.11. Another interesting function ψ is

$$\psi(x) = \begin{cases} |x|^{-(k-2)} & |x| \geq \sqrt{k-2} \\ (k-2)^{-(k-2)/2} \exp\big(\frac{1}{2}[(k-2) - |x|^2]\big) & |x| < \sqrt{k-2} \end{cases} .$$

For this ψ ,

$$g(x) = \nabla \log \psi(x) = \begin{cases} -(k-2)/|x|^2 & \text{if} \quad |x| \leq \sqrt{k-2} \\ -x & \phantom{\text{if}} \quad |x| \geq \sqrt{k-2} \end{cases}$$

8. Asymptotic Theory of Bayes Estimators

We begin with a basic result for Bayesian estimation of a Bernoulli probability θ.

Suppose that $(Y_1, Y_2, \cdots) \mid \boldsymbol{\theta} = \theta$ are iid Bernoulli(θ). Suppose that the prior distribution is an arbitrary distribution on $(0,1)$. The Bayes estimator of θ wrt squared error loss is the posterior mean $d(\underline{Y}) = d(Y_1, \cdots, Y_n) = E(\boldsymbol{\theta} \mid \underline{Y})$. Now this sequence is a (uniformly integrable) martingale; this type of martingale is sometimes known as *Doob's martingale.* Hence, by the martingale convergence theorem,

$$d(\underline{Y}_n) = E(\boldsymbol{\theta} \mid \underline{Y}_n) \to_{a.s.} E(\boldsymbol{\theta} \mid Y_1, Y_2, \cdots).$$

To prove consistency, it remains only to show that $E(\boldsymbol{\theta} \mid Y_1, Y_2, \cdots) = \boldsymbol{\theta}$.

To see this, we compute conditionally on $\boldsymbol{\theta}$: now

$$E(\sum_{i=1}^{n} Y_i \mid \boldsymbol{\theta}) = n\boldsymbol{\theta}, \qquad Var(\sum_{i=1}^{n} Y_i \mid \boldsymbol{\theta}) = n\boldsymbol{\theta}(1-\boldsymbol{\theta}).$$

Hence, with $\hat{\theta}_n \equiv n^{-1} \sum_{i=1}^{n} Y_i$,

$$E(\hat{\theta}_n - \boldsymbol{\theta})^2 \le n^{-1},$$

and by Chebychev's inequality, $\hat{\theta}_n$ converges in probability to $\boldsymbol{\theta}$. Therefore $\hat{\theta}_{n_k} \to_p \boldsymbol{\theta}$ almost surely for some subsequence n_k. Hence $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ a.s. where $\tilde{\boldsymbol{\theta}} \equiv \lim_k \hat{\theta}_{n_k}$ is measurable with respect to $Y_1, Y_2, \cdots$. Thus we have

$$\begin{aligned} E(\boldsymbol{\theta} \mid Y_1, Y_2, \cdots) &= E(\tilde{\boldsymbol{\theta}} \mid Y_1, Y_2, \cdots) \\ &= \tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} \qquad \text{a.s. } P_\Lambda \end{aligned}$$

where

$$P_\Lambda(\underline{Y} \in A, \boldsymbol{\theta} \in B) = \int_B P_\theta(\underline{Y} \in A)\, d\Lambda(\theta).$$

Therefore

$$P_\Lambda(d(\underline{Y}_n) \to \boldsymbol{\theta}) = \int_{[0,1]} P_\theta(d(\underline{Y}_n) \to \theta)\, d\Lambda(\theta) = 1,$$

and this implies that

$$P_\theta(d(\underline{Y}_n) \to \theta) = 1 \qquad \text{a.e. } \Lambda.$$

Thus the Bayes estimator $d(\underline{Y}_n) \equiv E(\boldsymbol{\theta} \mid \underline{Y}_n)$ is consistent for Λ a.e. θ.

A general result of this type was established by Doob (1948), and has been generalized by Breiman, Le Cam, and Schwartz (1964), Freedman (1963), and Schwartz (1965). The latter authors prove that for smooth, finite - dimensional families (θ, Λ) is consistent if and only if θ is in the support of Λ. However the assumption of finite - dimensionality is important: Freedman (1963) gives an example of an inconsistent Bayes rule when the sample space is $\mathbf{X} = \mathbb{Z}^+$ and Θ is the collection of all probabilities on $\mathbf{X}$. The recent work of Diaconis and Freedman (1986) on the consistency of Bayes estimates is an outgrowth of the work of Freedman (1963).

Asymptotic Efficiency of Bayes Estimators

Now we consider a more general situation, but still with a real-valued parameter θ. Let $X_1, \cdots, X_n$ be iid $p_\theta(x)$ with respect to μ where $\theta \in \Theta \subset R$, and $\theta_0 \in \Theta$ (an open interval) is true; write P for P_{θ_0}.

Assumptions:

B1: (i) - (vi) of theorem 1.1 (TPE, page 406) hold. Thus it follows that

$$l(\theta) - l(\theta_0) = (\theta - \theta_0)\dot{\mathbf{l}}_\theta(\underline{X}\,;\theta_0) - \frac{1}{2}(\theta - \theta_0)^2\{nI(\theta_0) + R_n(\theta)\} \tag{1}$$

where

$$\frac{1}{n} R_n(\theta) \to_p 0\,;$$

recall problems 2.22 and 7.2, TPE. We strengthen this to:

B2: For any $\epsilon > 0$ there exists a $\delta > 0$ such that

$$P\{\sup_{|\theta - \theta_0| \le \delta} |\frac{1}{n} R_n(\theta)| \ge \epsilon\} \to 0 \quad \text{as} \quad n \to \infty\,.$$

B3: For any $\delta > 0$ there exists an $\epsilon > 0$ such that

$$P\{\sup_{|\theta - \theta_0| \ge \delta} \frac{1}{n}(l(\theta) - l(\theta_0)) \le -\epsilon\} \to 1\,.$$

B4: The prior density λ of θ is continuous and > 0 for all $\theta \in \Theta$.

B5: $E_\lambda|\theta| < \infty$.

Theorem 8.1. Suppose that $\lambda^*(t|\underline{X})$ is the posterior density of $\sqrt{n}(\theta - T_n)$ where

$$T_n \equiv \theta_0 + \frac{1}{I(\theta_0)}\{n^{-1}\dot{\mathbf{l}}_\theta(\underline{X}\,;\theta_0)\}\,.$$

(i) If B1 - B4 hold, then

$$d_{TV}(\Lambda^*(\cdot\,|\,\underline{X}), N(0, \frac{1}{I(\theta_0)})) = \int |\lambda^*(t\,|\,\underline{X}) - \sqrt{I(\theta_0)}\,\phi(t\sqrt{I(\theta_0)})\,|\,dt \to_p 0\,. \tag{2}$$

(ii) If B1 - B5 hold, then

(3) $$\int (1 + |t|)\,|\lambda^*(t \mid \underline{X}) \;-\; \sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)})\,|\,dt \;\;\rightarrow_p\;\; 0 \;.$$

Theorem 8.2. If B1 - B5 hold, and $\tilde{\theta}_n$ is the Bayes estimator wrt λ for squared error loss, then

(4) $$\sqrt{n}(\tilde{\theta}_n \;-\; \theta_0) \;\rightarrow_d\;\; N(0\,,\frac{1}{I(\theta_0)})\,,$$

so that $\tilde{\theta}_n$ is $\sqrt{n}-$consistent and asymptotically efficient.

Example 8.3. Suppose that $X_1\,,\cdots,X_n$ are iid $N(\theta\,,\sigma^2)$ and that the prior on θ is $N(\mu\,,\tau^2)$. Then the posterior distribution is given by $(\theta \mid \overline{X}) \;\sim\; N(p_n\overline{X} + (1-p_n)\mu\,,\sigma_n^2/n)$ where

$$p_n \;\equiv\; \frac{n/\sigma^2}{n/\sigma^2 \;+\; 1/\tau^2}\,, \qquad\qquad \sigma_n^2 \;=\; \frac{n}{n/\sigma^2 + 1/\tau^2}$$

so that $p_n \rightarrow 1$, $\sqrt{n}(1-p_n) \;\rightarrow\; 0$, and $\sigma_n^2 \;\rightarrow\; \sigma^2$. Note that

$$T_n \;=\; \theta_0 \;+\; \frac{1}{1/\sigma^2}\,\frac{(\overline{X} \;-\; \theta_0)}{\sigma^2} \;=\; \overline{X}\,,$$

and the Bayes estimator with squared error loss is

$$E(\theta \mid \overline{X}) \;=\; p_n\overline{X} \;+\; (1-p_n)\mu\;.$$

Thus

$$\begin{aligned}(\sqrt{n}(\theta \;-\; \overline{X}) \mid \overline{X}) \;&\sim\; N(\sqrt{n}(1-p_n)(\mu - \overline{X})\,,\sigma_n^2)\\ &\rightarrow_d\; N(0\,,\sigma^2) \;=\; N(0\,,\frac{1}{I(\theta_0)}) \qquad \text{a.s.}\end{aligned}$$

in agreement with theorem 1, while

$$\begin{aligned}\sqrt{n}(E(\theta \mid \overline{X}) \;-\; \theta_0) \;&=\; \sqrt{n}(p_n\overline{X} + (1-p_n)\mu \;-\; \theta_0)\\ &=\; p_n\sqrt{n}(\overline{X} \;-\; \theta_0) \;+\; \sqrt{n}\,(1-p_n)\,(\mu - \theta_0)\\ &=_d\; p_n\,N(0\,,\sigma^2) \;+\; \sqrt{n}\,(1-p_n)\,(\mu - \theta_0)\\ &\rightarrow_d\; N(0\,,\sigma^2)\end{aligned}$$

in agreement with theorem 2. □

Proof of theorem 8.2. Suppose that theorem 8.1 is proved, and write

(a) $$\sqrt{n}(\tilde{\theta}_n \;-\; \theta_0) \;=\; \sqrt{n}(\tilde{\theta}_n \;-\; T_n) \;+\; \sqrt{n}(T_n \;-\; \theta_0)\,.$$

Since

(b) $$\sqrt{n}(T_n - \theta_0) \to_d N(0, \frac{1}{I(\theta_0)}),$$

it suffices to show that

(c) $$\sqrt{n}(\tilde{\theta}_n - T_n) \to_p 0.$$

To prove (c), note that by definition of $\tilde{\theta}_n$

$$\tilde{\theta}_n = \int \theta\, \lambda(\theta \mid \underline{X})\, d\theta = \int \{ \frac{t}{\sqrt{n}} + T_n \} \lambda^*(t \mid \underline{X})\, dt,$$

and hence

$$\sqrt{n}(\tilde{\theta}_n - T_n) = \int t\, \lambda^*(t \mid \underline{X})\, dt - \int t \sqrt{I(\theta_0)} \phi(t \sqrt{I(\theta_0)})\, dt.$$

It follows that

$$\sqrt{n}|\tilde{\theta}_n - T_n| \le \int |t|\, |\lambda^*(t \mid \underline{X}) - \int \sqrt{I(\theta_0)}\, \phi(t\sqrt{I(\theta_0)})\, |\, dt \to_p 0 \quad \text{as } n \to \infty$$

by theorem 1(ii). Thus (c) holds, and the proof is complete; it remains only to prove theorem 8.1. □

Proof of theorem 8.1: (i) By definition of T_n

$$\lambda^*(t \mid \underline{X}) = \frac{\lambda(T_n + n^{-1/2}t) \exp(l(T_n + n^{-1/2}t))}{\int \lambda(T_n + n^{-1/2}u) \exp(l(T_n + n^{-1/2}u))\, du} \equiv e^{\omega(t)} \lambda(T_n + n^{-1/2}t)/C_n$$

where

(a) $$\omega(t) \equiv l(T_n + n^{-1/2}t) - l(\theta_0) - \frac{1}{2nI(\theta_0)} \{\dot{\mathbf{l}}_\theta(\underline{X}\,; \theta_0)\}^2$$

and

(b) $$C_n \equiv \int e^{\omega(u)} \lambda(T_n + n^{-1/2}u)\, du.$$

Now if we can show that

$$J_{1n} \equiv \int |e^{\omega(t)} \lambda(T_n + n^{-1/2}t) - e^{-t^2 I(\theta_0)/2} \lambda(\theta_0)|\, dt \to_p 0,$$

then

(d) $$C_n \to_p \int e^{-t^2 I(\theta_0)/2} \lambda(\theta_0)\, dt = \lambda(\theta_0) \sqrt{2\pi / I(\theta_0)},$$

and the left side of (2) is J_n / C_n where

(e) $$J_n \equiv \int |e^{\omega(t)}\lambda(T_n + n^{-1/2}t) - C_n \sqrt{I(\theta_0)}\,\phi(t\sqrt{I(\theta_0)})\,|\,dt\ .$$

Further $J_n \le J_{1n} + J_{2n}$ where J_{1n} is defined in (c) and

(f) $$J_{2n} \equiv \int |\,C_n \sqrt{I(\theta_0)}\,\phi(t\sqrt{I(\theta_0)}) - e^{-t^2 I(\theta_0)/2}\,\lambda(\theta_0)\,|\,dt$$

$$= |\,\frac{C_n\sqrt{I(\theta_0)}}{\sqrt{2\pi}} - \lambda(\theta_0)| \int e^{-t^2 I(\theta_0)/2}\,dt \ \to_p\ 0$$

by (d). Hence it remains only to prove that (c) holds.

(ii) The same proof works with J'_{1n} replacing J_{1n} where J'_{1n} has an additional factor of $(1 + |t|)$.

Lemma 8.4. The following identity holds:

(5) $$\omega(t) \equiv l(T_n + n^{-1/2}t) - l(\theta_0) - \frac{1}{2nI(\theta_0)}\{\dot{\mathbf{l}}_\theta(\underline{X}\,;\theta_0)\}^2$$

$$= -I(\theta_0)\frac{t^2}{2} - \frac{1}{2n}R_n(T_n + \frac{t}{\sqrt{n}})\{t + \frac{1}{I(\theta_0)}\frac{\dot{\mathbf{l}}_\theta(\underline{X}\,;\theta_0)}{\sqrt{n}}\}^2\ .$$

Proof. This follows immediately from the definition of T_n, (1), and algebra. □

Proof that $J_{1n} \to_p 0$**:** Recall the definition of J_{1n} given in (c) of the proof of theorem 1. Divide the region of integration into:

(i) $|t| < M$;
(ii) $M \le t \le \delta\sqrt{n}$; and
(iii) $\delta\sqrt{n} < t < \infty$.

For the region (i), the integral is bounded by $2M$ times the supremum of the integrand over the set $|t| \le M$, and by application of lemma 1 it is seen that this supremum converges to zero in probability if

(a) $$\sup_{|t| \le M} |\,\frac{1}{n}R_n(T_n + n^{-1/2}t)(t + \frac{1}{I(\theta_0)}\frac{\dot{\mathbf{l}}_\theta(\underline{X}\,;\theta_0)}{\sqrt{n}})^2| \ \to_p\ 0$$

and

(b) $$\sup_{|t| \le M} |\lambda(T_n + n^{-1/2}t - \lambda(\theta_0)| \ \to_p\ 0\ .$$

Now λ is continuous by B4 and $T_n \to_p \theta_0$ by B1, so λ is uniformly continuous in a neighborhood of θ_0 and

(c) $$\theta_0 - \delta \le T_n - \frac{M}{\sqrt{n}} \le T_n - \frac{t}{\sqrt{n}} \le T_n + \frac{M}{\sqrt{n}} \le \theta_0 + \delta$$

for $|t| \le M$, so (b) holds. Now $n^{-1/2}\dot{\mathbf{l}}_\theta(\underline{X}\,;\theta_0)$ is bounded in probability (by B1 and the central limit theorem), so it follows from (c) and B2 that (a) holds.

(ii) For the region $M \le |t| \le \delta\sqrt{n}$ it suffices to show that the integrand is bounded by an integrable function with probability close to one, since then the integral can be made small by choosing M suficiently large. Since the second term of the integrand in J_{1n} is integrable, it suffices to find such an integrable bound for the first term. In particular, it can be shown that for every $\epsilon > 0$ there exists a $C = C(\epsilon, \delta)$ and an integer $N = N(\epsilon, \delta)$ so that, for $n \ge N$,

$$P(e^{\omega(t)}\lambda(T_n + \frac{t}{\sqrt{n}}) \le C\exp(-t^2 I(\theta_0)/4) \text{ for all } |t| \le \delta\sqrt{n}) \ge 1-\epsilon\,; \tag{d}$$

this follows from the assumption B2; see TPE, pages 463 - 464.

(iii) The integral over the region $|t| \ge \delta\sqrt{n}$ converges to zero by argument similar to that for the region (ii), but now the assumption B3 comes into play; see TPE pages 464-465.

The proof for J'_{1n} requires only trivial changes, using the assumption B5. □

Notes:

1. This proof is from Lehmann, TPE, pages 454 - 465; this type of results apparently dates back to Laplace (1820), and was later rederived by Bernstein (1917), and von Mises (1931). More general versions have been established by Le Cam (1958), Bickel and Yahav (1969), and Ibragimov and Has'minskii (1972). See the discussion on page 482 of TPE. For a proof with weaker regularity conditions, see Hartigan (1983), chapter 11, section 3.

2. See Ibragimov and Has'minskii (1981) for more on the consistency and efficiency of Bayes estimators; also see Strasser (1981) as referenced by Lehmann.

3. Read the examples and discussion on pages 458 - 461 of TPE.

4. Analogues of these theorems for nonparametric and semiparametric problems are still in a relatively undeveloped state; Diaconis and Freedman (1986) is a good starting point. Cox (1993) treats Bayesian nonparametric regression.

Now we will continue the development for convex likelihoods started in section 4.7. The notation will be as in section 4.7, and we will make use of lemmas 4.7.1 and 4.7.2. Here, as in section 4.7, we will *not* assume that the distribution P governing the data is a member of the parametric model $\mathbf{P} = \{P_\theta : \theta \in \Theta \subset R^d\}$.

Here are the hypotheses which will be used.

Suppose that $\lambda(\theta)$ is the prior density (with respect to Lebesgue measure ν on R^d). We will assume that:

C1. $\lambda(\theta) \le C_1 \exp(C_2|\theta|)$ for all θ for some positive constants C_1, C_2.

C2. $\lambda(\theta)$ is continuous at θ_0.

C3. $\log p(x, \theta_0 + t) - \log p(x, \theta_0) = \psi(x)^T t + R(x; t)$ is concave in t.

C4. $E_P \psi(X_1)\psi(X_1)^T \equiv K$, $E_P \psi(X_1) = 0$.

C5. $E_P R(X_1; t) = -t^T J t/2 + o(|t|^2)$ and $Var_P(R(X_1; t)) = o(|t|^2)$ where J is symmetric, positive definite.

C6. $X_1, \cdots, X_n$ are i.i.d. P.

Theorem 8.5. Suppose that C1 - C6 hold. Then the maximum likelihood estimator $\hat{\theta}_n$ is $\sqrt{n}$ – consistent for $\theta_0 = \theta_0(P)$ and satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N_d(0, J^{-1}K(J^{-1})^T).$$

Moreover the Bayes estimator $\theta_n^* = E\{\boldsymbol{\theta} \,|\, \underline{X}_n\}$ satisfies

$$\sqrt{n}(\theta_n^* - \hat{\theta}_n) \to_p 0,$$

and hence also

$$\sqrt{n}(\theta_n^* - \theta_0) \to_d N_d(0, J^{-1}K(J^{-1})^T).$$

Proof. Let $L_n(\theta) \equiv \prod_{i=1}^n p(X_i, \theta)$ denote the likelihood function. Define random convex functions A_n by

$$\exp(-A_n(t)) = L_n(\hat{\theta}_n + t/\sqrt{n})/L_n(\hat{\theta}_n).$$

By definition of the MLE, A_n achieves its minimum value of zero at $t = 0$. Then

$$\theta_n^* = \frac{\int \theta\, L_n(\theta)\lambda(\theta)\, d\theta}{\int L_n(\theta)\lambda(\theta)\, d\theta}$$

$$= \hat{\theta}_n + \frac{1}{\sqrt{n}} \frac{\int t \exp(-A_n(t))\lambda(\hat{\theta}_n + t/\sqrt{n})\exp(-C_2|\hat{\theta}_n|)\, dt}{\int \exp(-A_n(t))\lambda(\hat{\theta}_n + t/\sqrt{n})\exp(-C_2|\hat{\theta}_n|)\, dt}$$

by the change of variable $\theta = \hat{\theta}_n + t/\sqrt{n}$.

Claim: The random functions A_n converge in probability uniformly on compact sets to $t^T J t/2$.

Proof. Let

$$A_n^0(t) = n\, I\!P_n\{\log p(x; \theta_0 + t/\sqrt{n}) - \log p(x; \theta_0)\}$$

From the proof of theorem 4.7.6 (with $h(x;\theta) \equiv -\log p(x;\theta)$ and multiplication by -1)

$$(*) \qquad A_n^0(t) \;=\; U_n^T t \;+\; \frac{1}{2}\, t^T J t \;+\; o_p(1)\;,$$

and this holds uniformly in t in compact sets by lemma 4.7.1. Note that

$$A_n(t) \;=\; A_n^0(t+\sqrt{n}(\hat{\theta}_n-\theta_0)) \;-\; A_n^0(\sqrt{n}(\hat{\theta}_n-\theta_0))\;.$$

Since $\sqrt{n}(\hat{\theta}_n-\theta_0) \;=\; O_p(1)$ so that for every $\epsilon>0$ there exists a compact set $K \equiv K_\epsilon$ such that $P(\sqrt{n}(\hat{\theta}_n-\theta_0) \in K) \geq 1-\epsilon$. Hence by the uniformity in t in the convergence in (*),

$$\begin{aligned}
A_n(t) \;&=\; U_n^T(t+\sqrt{n}(\hat{\theta}_n-\theta_0)) \;-\; U_n^T(\sqrt{n}(\hat{\theta}_n-\theta_0)) \\
&\qquad +\; \frac{1}{2}\,\{(t+\sqrt{n}(\hat{\theta}_n-\theta_0))^T J (t+\sqrt{n}(\hat{\theta}_n-\theta_0)) \\
&\qquad\quad -\; \sqrt{n}(\hat{\theta}_n-\theta_0)^T J \sqrt{n}(\hat{\theta}_n-\theta_0)\} \;+\; o_p(1) \\
&=\; U_n^T t \;+\; \frac{1}{2}\, t^T J t \;+\; \sqrt{n}(\hat{\theta}_n-\theta_0)^T J t \;+\; o_p(1) \\
&=\; \frac{1}{2}\, t^T J t \;+\; o_p(1)
\end{aligned}$$

where the last line follows because $\sqrt{n}(\hat{\theta}_n-\theta_0) \;=\; -\,J^{-1}U_n \;+\; o_p(1)$. Since A_n is convex and converges pointwise in probability to $t^T J t/2$, it converges uniformly in probability on compact subsets by lemma 4.7.1. □

Now we return to the main proof. Define $\gamma_n \;\equiv\; \inf_{|u|\,=\,1} A_n(u)$. It converges in probability to $\gamma_0 \;=\; \inf_{|u|\,=\,1} u^T J u/2 \;>\; 0$. By the same argument used in lemma 4.7.2 it follows that

$$A_n(t) \;\geq\; \gamma_n |t| \qquad \text{for} \quad |t|>1\;.$$

Now the integrand in the numerator above converges in probability for each fixed t to

$$t\exp(\,-\,t^T J t/2)\,\lambda(\theta_0)\exp(\,-\,C_2|\theta_0|)\;,$$

and similarly the integrand of the denominator converges for each fixed t to

$$\exp(\,-\,t^T J t/2)\,\lambda(\theta_0)\exp(\,-\,C_2|\theta_0|)\;.$$

Furthermore the domination hypotheses of the following convergence lemma hold for both the numerator and denominator with dominating function

$$D(t) \;\equiv\; \begin{cases} 2C_1 & \text{if} \quad |t|\leq 1 \\ C_1|t|\exp(\,-\,\gamma_0|t|/2) & \text{if} \quad |t|>1 \end{cases}.$$

Hence the ratio of integrals converges in probability to

$$\frac{\displaystyle\int t\exp(\,-\,t^T J t/2)\,\lambda(\theta_0)\exp(\,-\,C_2|\theta_0|)\,dt}{\displaystyle\int \exp(\,-\,t^T J t/2)\,\lambda(\theta_0)\exp(\,-\,C_2|\theta_0|)\,dt} \;=\; 0\;.$$

□

Lemma 8.6. Suppose that $X_n(t)$, $Y_n(t)$ are jointly measure random functions, $X_n(t,\omega)$, $Y_n(t,\omega)$ for $(t,\omega) \in K \times \Omega$ where $K \subset R^d$ is compact, that λ is a measure on R^d, and:

(i) $Y_n(t) \to_p Y(t)$ and $X_n(t) \to_p X(t)$ for λ almost all $t \in K$.

(ii) $\int_K Y_n(t)\,d\lambda(t) \to_p \int_K Y(t)\,d\lambda(t)$ with $|\int_K Y(t)\,d\lambda(t)| < \infty$ almost surely.

Then $\int_K X_n(t)\,d\lambda(t) \to_p \int_K X(t)\,d\lambda(t)$.

Proof. It suffices to show almost sure convergence for some further subsequence of any given subsequence. By convergence in probability for fixed t, and then dominated convergence

$$\begin{aligned} H_n(\epsilon) &\equiv (P \otimes \lambda)(\{(\omega,t) : |X_n(t,\omega) - X(t,\omega)| > \epsilon\}) \\ &= \int_K P(\{\omega : |X_n(t,\omega) - X(t,\omega)| > \epsilon\})\,dt \to 0 \end{aligned}$$

by the dominated convergence theorem, and similarly for $\{Y_n\}$. By replacing ϵ by $\epsilon_n \downarrow 0$ and extraction of subsequences we can find a subsequence $\{n'\}$ so that the integrals $\int_K X_n\,d\lambda$ converge and for some set $N \subset K \times \Omega$ with $P \otimes \lambda(N) = 0$ we get convergence for all $(\omega,t) \in N^c$ of both $X_{n'}$ and $Y_{n'}$. Then we have $\lambda(\{t : (\omega,t) \in N\}) = 0$ for almost all ω, Hence by Fatou's lemma applied to $Y_{n'} \pm X_{n'}$ we deduce that

$$\int_K X_{n'}(t)\,d\lambda(t) \to_{a.s.} \int_K X(t)\,d\lambda(t).$$

□

Corollary 8.7. Suppose that:

(i) $G_n(t) \to_p G(t)$ for each fixed $t \in R^d$.

(ii) $P(|G_n(t)| \le D(t)$ for all $t \in R^d) \to 1$.

(iii) $\int D(t)\,dt < \infty$.

Then $\int G_n(t)\,dt \to_p \int G(t)\,dt$.

Proof. Let $\epsilon > 0$; choose a compact set $K = K_\epsilon$ so that $\int_K D(t)\,dt < \epsilon$. This is possible since $\int D(t)\,dt < \infty$. Then, using $|G(t)| \le D(t)$,

$$|\int G_n(t)\,dt - \int G(t)\,dt|$$

$$\leq \ |\int_K G_n(t)\,dt \ - \ \int_K G(t)\,dt|$$
$$+ \int_{K^c} |G_n(t)|\,dt \ + \ \int_{K^c} |G(t)|\,dt|$$
$$\leq \ |\int_K G_n(t)\,1_{[|G_n(t) \leq D(t)]}\,dt \ - \ \int_K G(t)\,dt|$$
$$+ \int_{K \cap [|G_n(t) > D(t)]} |G_n(t)|\,dt \ + \ \int_{K^c \cap [|G_n(t) > D(t)]} |G_n(t)|\,dt$$
$$+ \ 2\int_{K^c} D(t)\,dt$$

Now apply lemma 8.6 with $Y_n(t) \ \equiv \ D(t)$ and

$$X_n(t) \ \equiv \ G_n(t) 1_{[|G_n(t)| \leq D(t)]} \ .$$

Then $X_n(t) \ \to_p \ G(t)$ and the second hypothesis of the lemma holds easily. Hence $\int X_n(t)\,dt \ \to_p \ \int G(t)\,dt$ so that the first term above $\to_p \ 0$. The second and third terms converge in probability to 0 because with the set over which the integral is taken is empty with arbitrarily high probability; and the third term is bounded by 2ϵ by choice of K. □

5.9

9. Nonparametric Bayes Estimation