BIOS 761 - Advanced Probability and Statistical Inference II

Quefeng Li

1 Elementary Decision Theory

Decision theory, as the name implies, is concerned with the problem of making decisions. Statistical decision theory is concerned with the making of decisions in the presence of statistical knowledge which sheds light on some of the uncertainties involved in the decision problem.

For the most part, these uncertainties can be considered to be unknown numerical quantities (parameters), represented by θ.

Example 1.1

Consider the situation of a drug company deciding whether or not to market a new pain reliever. Two of the many factors affecting its decision are the proportion of people for which the drug will prove effective (θ_1) and the proportion of the market the drug will capture (θ_2). Both θ_1 and θ_2 will be generally unknown, though typically experiments can be conducted to obtain statistical information about them. This example is one of decision theory in that the ultimate purpose is to decide whether or not to market the drug, how much to market, and what price to charge, etc...

Classical statistics is directed toward the use of sample information (the data) in making inferences about θ. The inferences are, for the most part, made without regard to the use to which they are to be put. In decision theory, on the other hand, an attempt is made to combine the sample information with other relevant aspects of the problem in order to make the best decision.

In addition to the data, two other types of information are typically relevant. The first is a knowledge of the possible consequences of the decision. Often this knowledge can be quantified by determining the loss that would be incurred for each possible decision and for the various possible values of θ.

The incorporation of loss functions into statistical analyses was first studied by Wald (1950). Economists and those in business think in terms of utility function = - loss function.

In Example 1.1, suppose we consider estimating θ_1 for an advertising campaign. The loss in underestimating θ_1 arises from making the product appear worse than it really is (adversely affecting sales), while the loss in overestimating θ_1 would be based on the risks of possible penalties for misleading advertising.

The second source of nonsample information is prior information about θ, which will be used in Bayesian decision theory.

Basic Elements

The unknown quantity θ is often called the state of nature. The parameter space $\boldsymbol{\Theta}$ denotes all possible states of nature. For Example 1.1, $\boldsymbol{\Theta} = \{\theta_1, \theta_2\}$.

Decisions are more commonly called actions. A particular action will be denoted by 'a' and the set of all possible actions by $\boldsymbol{\mathcal{A}}$ = the action space. $\boldsymbol{\mathcal{X}}$ = the sample space of a random variable X with distribution P_θ. X can be continuous and have a density with respect to Lebesgue measure, or X can be discrete and have density (probability function) with respect to a counting measure.

$$\frac{dP_\theta}{d\mu} = p_\theta \quad (p_\theta \text{ is a density, } \mu = \text{Lebesgue measure})$$

$$\frac{dP_\theta}{d\mu} = p_\theta \quad (p_\theta \text{ is a probability function, } \mu \text{ is counting measure})$$

$P : \boldsymbol{\mathcal{X}} \times \boldsymbol{\Theta} \to [0,1]$; $P_\theta(B) = \int_B p(x|\theta)\, dx$ when X is continuous,
$P_\theta(B) = \sum_{x \in B} p(x|\theta)$, X is discrete where $B \in \boldsymbol{\mathcal{X}}$.

As noted earlier, the key element in decision theory is the loss function. If a particular action a_1 is taken and θ_1 turns out to be the true state of nature, then a loss $L(\theta_1, a_1)$ will be incurred. The loss function, $L(\theta, a)$, is defined for all $(\theta, a) \in \boldsymbol{\Theta} \times \boldsymbol{\mathcal{A}}$, and thus

$$L : \boldsymbol{\Theta} \times \boldsymbol{\mathcal{A}} \to \boldsymbol{\Re} \quad \text{(loss funtion)}$$

Example 1.2

Decision theory is often used in games and gambling. Suppose an investor must decide whether or not to buy risky bonds. If he buys, they can be redeemed for a net gain of \$500. There may be a default on the bonds in which the original \$1000 would be lost. If he puts money in a "safe" investment, he will be guaranteed a net gain of \$300 over the same period. Here $\boldsymbol{\mathcal{A}} = \{a_1, a_2\}$, a_1 = buying bonds, a_2 = not buying. $\boldsymbol{\Theta} = \{\theta_1, \theta_2\}$, where θ_1 denotes the state of nature "no default occurs" and θ_2 the state of "a default occurs". Recalling that a gain is represented by a negative loss, the loss function $L(\theta, a)$ is given by the following table:

	a_1	a_2
θ_1	-500	-300
θ_2	1000	-300

Note that in this example, there is no data from an associated statistical experiment. Such a problem is called a no-data problem.

The loss table above is often called a loss matrix.

Note: In general we will not consider negative losses (gains), so that for our purposes, we define

$$L : \boldsymbol{\Theta} \times \boldsymbol{\mathcal{A}} \rightarrow \Re^{+} \ .$$

Definition 1.1 (Non-randomized decision function (rule))

A non-randomized decision rule, $d(X)$, is a function from $\boldsymbol{\mathcal{X}}$ into $\boldsymbol{\mathcal{A}}$ ($d : \boldsymbol{\mathcal{X}} \rightarrow \boldsymbol{\mathcal{A}}$). If $X = x$ is the observed value of X, then $d(x)$ is the action that will be taken. When there is no data, a decision rule is simply an action. Two decision rules $d_1(x)$ and $d_2(x)$ are considered equivalent if $P_{\Theta}(d_1(x) = d_2(x)) = 1$ for all θ. $\boldsymbol{\mathcal{D}}$ = the set of all decision functions (rules). In decision theory, we deal with the triplet $(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}}, L)$ and any such triplet defines a decision problem.

The reason $d(x)$ above is called a non-randomized decision rule is because once $X = x$ is observed, we take an action (decision), $d(x)$, with probability 1.

Example 1.1 (continued)

Suppose we wish to estimate θ_2. Since θ_2 is a proportion, it is clear that $\boldsymbol{\Theta} = \{\theta_2 : 0 \leq \theta_2 \leq 1\} = [0, 1]$. Since the goal is to estimate θ_2, the action taken will simply be the choice of a number as an estimate of θ_2. Hence $\boldsymbol{\mathcal{A}} = [0, 1]$, and thus $\boldsymbol{\mathcal{A}} = \boldsymbol{\Theta}$ in estimation problems. The company might determine the loss function to be

$$L(\theta_2, a) = \begin{cases} \theta_2 - a & \text{if } \theta_2 - a \geq 0 \\ 2(a - \theta_2) & \text{if } \theta_2 - a \leq 0 \end{cases}.$$

Thus an overestimate of demand is considered twice as costly as an underestimate of demand. Suppose X = number of people who buy the drug among n people interviewed, so that $X \sim \text{Binomial}(n, \theta_2)$, and

$$p(x|\theta_2) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta_2^x (1-\theta_2)^{n-x} .$$

$d(x) = \frac{X}{n}$ is the standard decision rule for estimating θ_2. In estimation problems, a decision rule is called an estimator. $d(x) = \frac{X}{n}$ does not make use of the loss function. We will see how to make use of the loss function shortly.

Definition 1.2

The risk function of a decision rule $d(x)$ is defined by

$$\begin{aligned} R(\theta, d) &= E_\theta[L(\theta, d(x)] \\ &= \int_{\mathcal{X}} L(\theta, d(x)) P_\theta(dx) \\ &= \int_{\mathcal{X}} L(\theta, d(x)) p(x|\theta)\ dx\ . \end{aligned}$$

$R : \Theta \times \mathcal{D} \to \Re^+$.

When there is no data (no data problem), $R(\theta, d) = L(\theta, d)$. $R(\theta, d)$ is sometimes referred to as the frequentist risk. It is desirable to use a decision rule $d(x)$ which has small $R(\theta, d)$. We often denote a decision rule by $d \equiv d(x)$.

Definition 1.3

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

From Definition 1.3 above, it is clear that an inadmissible rule should not be used since a decision rule with smaller risk can be found. Unfortunately, there is usually a large class of admissible decision rules for a particular problem. These rules often have risk function that cross, i.e., are better in different places.

Example 1.3

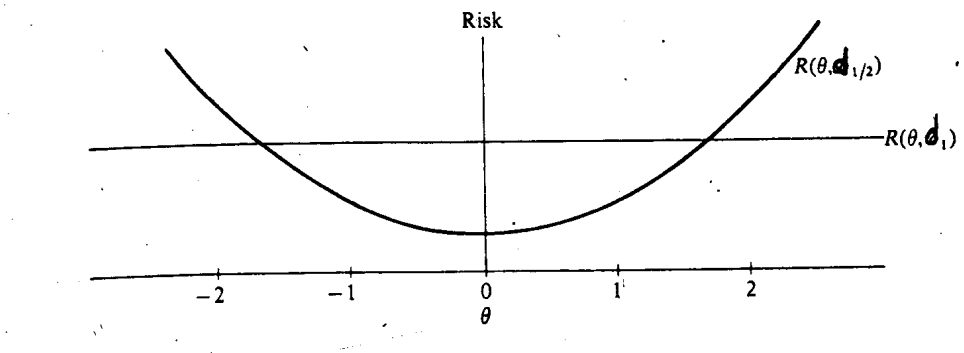
Suppose $X \sim N(\theta, 1)$ and we wish to estimate θ under squared-error loss, $L(\theta, a) = (\theta - a)^2$. Consider the decision rules $d_c(x) = cX$. We have

$$\begin{aligned} R(\theta, d_c) &= E_\theta[L(\theta, d_c(x)] \\ &= E_\theta[(\theta - cX)^2] \\ &= E_\theta[\theta^2 - 2\theta cX + c^2X^2] \\ &= \theta^2 - 2\theta^2 c + c^2(1 + \theta^2) \\ &= c^2 + (1 - c)^2\theta^2 \ . \end{aligned}$$

For $c > 1$,

$$R(\theta, d_1) = 1 < c^2 + (1-c)^2\theta^2 = R(\theta, d_c) \ ,$$

and thus d_1 has less risk than d_c for $c > 1$. Hence d_c is inadmissible for $c > 1$.
On the other hand, for $0 \leq c \leq 1$, the rules are not comparable, since the risk functions cross.


Risk
$R(\theta, d_{1/2})$
$R(\theta, d_1)$
−2
−1
0
1
2
θ

Example 1.4

Consider the following loss matrix of a particular no-data problem.

	a_1	a_2	a_3
θ_1	1	3	4
θ_2	-1	5	5
θ_3	0	-1	-1

The rule (action) a_2 has smaller risk (loss) than a_3 since $L(\theta_i, a_2) \leq L(\theta_i, a_3)$ for all θ_i, with strict inequality for θ_1. Hence a_3 is inadmissible.

The actions a_1 and a_2 are not comparable, in that $L(\theta_i, a_1) < L(\theta_i, a_2)$ for θ_1 and θ_2, while the reverse inequality holds for θ_3. Thus a_1 and a_2 are admissible.

Randomized decision rules

In some situations, it is necessary to take actions in a random manner. Such situations often arise in games and gambling.

Example 1.5 (matching pennies)

You and your opponent are to simultaneously uncover a penny. If the two coins match (i.e., both are heads or both are tails) you win \$1 from your opponent.

If the coins don't match, your opponent wins \$1 from you. The actions are a_1 = choose heads, a_2 = choose tails. The states of nature are θ_1 = the opponent's coin is a head, and θ_2 = the opponent's coin is a tail. The loss matrix is given by

	a_1	a_2
θ_1	-1	1
θ_2	1	-1

Both a_1 and a_2 are admissible. If the game is to be played a large number of times, it would be a very poor idea to decide to use a_1 exclusively or a_2 exclusively. Your opponent would quickly realize your strategy and choose his action to guarantee victory. The only way of preventing ultimate defeat is to choose a_1 and a_2 by some random mechanism. We can choose a_1 with probability p and a_2 with probability $1 - p$.

Definition 1.4

A randomized decision rule $d(a, x)$ is, for each x, a probability distribution on $\boldsymbol{\mathcal{A}}$, that is $d : \boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{X}} \to [0, 1]$, and $d(a, x) \equiv d(a|X) =$ probability of action a when $X = x$ is observed.

Definition 1.5

The loss function $L(\theta, d(\cdot|x))$ of a randomized rule $d(a|x)$ is defined to be

$$\begin{aligned} L(\theta, d(\cdot|x)) &= E_{d(\cdot|x)}[L(\theta, a)] \\ &= \int L(\theta, a)\, d(da|X = x) \\ &= \int L(\theta, a) p(a|x)\, da\ . \end{aligned}$$

In the discrete case

$$L(\theta, d(\cdot|x)) = \sum_{i=1}^{k} L(\theta, a_i) d(a_i, x).$$

Definition 1.6

The risk function for a randomized rule $d \in \mathcal{D}$ when $\theta \in \Theta$ is true is defined as

$$\begin{aligned} R(\theta, d) &= E_\theta[L(\theta, d(\cdot|x))] \\ &= \begin{cases} \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a)\ d(da|X = x)\ P_\theta(dx) \\ \sum_{i=1}^{k} L(\theta, a_i)\{\sum_{j=1}^{m} d(a_i, x_j) P_\theta(X = x_j)\} \quad \text{(discrete case)} \end{cases} \end{aligned}$$

If X has density with respect to Lebesgue measure,

$$R(\theta, d) = \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a) p(a|x) p(x|\theta)\ da\ dx\ .$$

An example of a randomized decision rule is hypothesis testing.

Example 1.6 (Hypothesis testing)

$\mathcal{A} = \{a_0, a_1\} = \{0, 1\}$, $\Theta = \Theta_0 \cup \Theta_1$, $a_0 =$ accept H_0, $(a_0 = 0)$, $a_1 =$ reject H_0 $(1 - a_0 = a_1 = 1)$, $L(\theta, 0) = l_0 1_{\Theta_1}(\theta)$, $L(\theta, 1) = l_1 1_{\Theta_0}(\theta)$, where

$$1_{\Theta_1}(\theta) = \begin{cases} 1, & \theta = \theta_1 \\ 0, & \theta = \theta_0 \end{cases}.$$

	$a_0 = 0$	$a_1 = 1$
θ_0	0	l_1
θ_1	l_0	0

Since $\mathcal{A}$ contains just two points,

$$d(1|x) = 1 - d(0|x) \equiv \phi(x)$$

and

$$\begin{aligned} P_\theta(\text{accept } H_0) &= E_\theta[d(0|x)] = \int d(0|x)p(x|\theta)\, dx \\ &= E_\theta[1-\phi(x)]\ . \end{aligned}$$

$$P_\theta(\text{reject } H_0) = E_\theta[d(1|x)] = E_\theta[\phi(x)] = \beta_d(\theta).$$

and

$$R(\theta,d) = l_1 E_\theta[\phi(x)]1_{\mathbf{\Theta}_0}(\theta) + l_0 E_\theta[1-\phi(x)]1_{\mathbf{\Theta}_1}(\theta)\ .$$

The classic Neyman-Pearson (NP) hypothesis testing philosophy bounds $l_1\alpha_d(\theta) \equiv \sup_{\theta\in\mathbf{\Theta}_0} R(\theta,d)$ and tries to minimize $l_0(1-\beta_d(\theta)) \equiv R(\theta,d)$ for each $\theta \in \mathbf{\Theta}_1$.

Derivation of $R(\theta, d)$ from Example 1.6:

$$\begin{aligned} L(\theta, d(\cdot|x)) &= E_{d(\cdot|x)}[L(\theta, a)] \\ &= L(\theta, a_0)d(a_0|x) + L(\theta, a_1)d(a_1|x) \\ &= L(\theta, 0)d(0|x) + L(\theta, 1)d(1|x) \\ &= l_0 1_{\Theta_1}(\theta)d(0|x) + l_1 1_{\Theta_0}(\theta)d(1|x) \\ &= l_0 1_{\Theta_1}(1 - \phi(x)) + l_1 1_{\Theta_0}\phi(x) \ . \end{aligned}$$

Thus

$$\begin{aligned} R(\theta, d) &= E_\theta[L(\theta, d(\cdot|x))] \\ &= E_\theta[l_0 1_{\Theta_1}(1 - \phi(x)) + l_1 1_{\Theta_0}\phi(x)] \\ &= l_0 1_{\Theta_1}(\theta)E_\theta(1 - \phi(x)) + l_1 1_{\Theta_0}(\theta)E_\theta(\phi(x)). \end{aligned}$$

Example 1.7 (Randomized Decision Rules)

For discrete random variables, randomized decision rules have to be employed to obtain size α tests. Consider $X \sim$ Binomial (n, θ) and we wish to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ where $\theta_1 > \theta_0$. The nonrandomized Uniformly Most Powerful (UMP) test has a rejection region of the form $C = \{x \in \mathcal{X} : x \geq j\}$, $j = 0, 1, 2, \ldots, n$. Since $\mathcal{X}$ is discrete and finite, the size of this test is $\alpha = P_{\theta_0}(C)$ and can attain only a finite number of values. If other values of α are desired, such as $\alpha = .05$, for example, then randomized rules must be used. It suffices to consider the randomized rules given by

$$d_j(a_0|x) = \begin{cases} 1 & \text{if } x < j \\ p & \text{if } x = j \\ 0 & \text{if } x > j \end{cases},$$

where

$d_j(a_0|x) = P(\text{accepting } H_0|x)$ and $d_j(a_1|x) = 1 - d_j(a_0|x)$ and a_i denote accepting H_i, $i = 0, 1$. Thus if $x > j$ is observed, H_0 will be rejected with probability 1. If $x < j$ is observed, then H_0 is accepted (not rejected) with probability 1. If $x = j$ is observed, a randomization will be performed, rejecting H_0 with probability $1 - p$ and accepting H_0 with probability p.

By a proper choice of j and p, a most powerful test of the above form can be found for any given size α,

$$\begin{aligned} \alpha &= P_{\theta_0}(X > j) + (1-p)P_{\theta_0}(X = j) \\ &= E_{\theta_0}[1 - d_j(a_0|x)] = E_{\theta_0}[d_j(a_1|x)] \ . \end{aligned}$$

Thus,

$$\begin{aligned}\alpha &= 1-\sum_{k=0}^{j}\binom{n}{k}\theta_0^k(1-\theta_0)^{n-k}+(1-p)\binom{n}{j}\theta_0^j(1-\theta_0)^{n-j}\\ &= 1-\sum_{k=0}^{j-1}\binom{n}{k}\theta_0^k(1-\theta_0)^{n-k}-p\binom{n}{j}\theta_0^j(1-\theta_0)^{n-j}\ .\end{aligned}$$

Remark 1.1

Note that a nonrandomized rule is just a special case of a randomized rule, in that nonrandomized rules consists of randomized rules which, for each x, choose a specific action with probability 1. If $d(x)$ is a nonrandomized rule, then the equivalent randomized rule is

$$I_a(d(x)) = \begin{cases} 1, & \text{if } d(x) \in a \\ 0, & \text{if } d(x) \notin a\ . \end{cases}$$

Example 1.8

Suppose we are in the estimation framework so that $\mathcal{A} = \Theta$, and let $\Theta = \Re^s$ for some s. A typical loss function in estimation problems is

$$L(\boldsymbol{\theta}, \boldsymbol{a}) = k|\boldsymbol{\theta} - \boldsymbol{a}|^2 \quad \text{for some } k, \tag{1.1}$$

where $|\boldsymbol{a}| = \sqrt{\boldsymbol{a}'\boldsymbol{a}}$. This is often called quadratic loss. Other common loss functions are

$$L(\boldsymbol{\theta}, \boldsymbol{a}) = k|\boldsymbol{\theta} - \boldsymbol{a}| \quad \text{(absolute error loss)}$$

$$L(\boldsymbol{\theta}, \boldsymbol{a}) = \begin{cases} 1, & \boldsymbol{\theta} \notin \Theta_0 \\ 0, & \boldsymbol{\theta} \in \Theta_0 \end{cases} \quad (0-1 \text{ loss}) \ ,$$

or more generally

$$L(\boldsymbol{\theta}, \boldsymbol{a}) = \lambda(\boldsymbol{\theta})|\boldsymbol{\theta} - \boldsymbol{a}|^r, \quad r \geq 0.$$

For the quadratic loss function in (1.1), we can derive the risk of a nonrandomized rule $d(x)$ as

$$\begin{aligned} R(\boldsymbol{\theta}, d) &= kE_{\boldsymbol{\theta}}|\boldsymbol{\theta} - d(X)|^2 \\ &= k[\mathrm{Var}_\theta(d(X)) + (E[d(X)] - \theta)^2] \quad (\text{when } s = 1). \end{aligned}$$

For a randomized decision rule

$$\begin{aligned} R(\boldsymbol{\theta}, d) &= kE_\theta|\boldsymbol{\theta} - d(\cdot|X)|^2 \\ &= k \int_{\boldsymbol{\mathcal{X}}} \int_{\boldsymbol{\mathcal{A}}} |\boldsymbol{\theta} - d(\boldsymbol{a}|x)|^2 p(\boldsymbol{a}|x) p(x|\boldsymbol{\theta}) \, d\boldsymbol{a} \, dx \end{aligned}$$

Example 1.7 (cont'd)

Recall Example 1.7 and assume that the loss is zero if a correct decision is made and the loss is 1 if an incorrect decision is made. Thus

$$L(\theta_i, a_k) = \begin{cases} 0 & \text{if } i = k \\ 1 & \text{if } i \neq k \end{cases}$$

$i = 0, 1$, $k = 0, 1$.

The loss of the randomized decision rule d_j when $\theta = \theta_0$ is then given by

$$\begin{aligned} L(\theta_0, d_j(\cdot|x)) &= E_{d_j(\cdot|x)}[L(\theta_0, a)] \\ &= d_j(a_0|x)L(\theta_0, a_0) + d_j(a_1|x)L(\theta_0, a_1) \\ &= d_j(a_1|x) \ . \end{aligned}$$

The risk for the randomized decision rule is given by

$$\begin{aligned}
R(\theta_0, d_j(\cdot|x)) &= E_{\theta_0}[L(\theta_0, d_j(\cdot, x))] \\
&= E_{\theta_0}[d_j(a_1|x)] \\
&= 1 - E_{\theta_0}[d_j(a_0|x)] \\
&= 1 - [pP_{\theta_0}(X = j) + P_{\theta_0}(X < j)] \\
&= 1 - pP_{\theta_0}(X = j) - P_{\theta_0}(X \leq j - 1) \\
&= P_{\theta_0}(\text{reject } H_0) \\
&= \alpha \text{ (probability of type I error).}
\end{aligned}$$

Similarly,

$$\begin{aligned} L(\theta_1, d_j(\cdot|x)) &= E_{d_j(\cdot|x)}[L(\theta_1, a)] \\ &= d_j(a_0|x)L(\theta_1, a_0) + d_j(a_1|x)L(\theta_1, a_1) \\ &= d_j(a_0|x) \ , \end{aligned}$$

so that

$$\begin{aligned} R(\theta_1, d_j(\cdot|x)) &= E_{\theta_1}[d_j(a_0|x)] \\ &= P_{\theta_1}(X < j) + pP_{\theta_1}(X = j) \\ &= P_{\theta_1}(\text{accept } H_0) \\ &= 1 - \beta \quad \text{(probability of type II error)} \end{aligned}$$

Thus, under 0-1 loss in hypothesis testing, the risks are just the probability of type I and type II errors.

Admissibility and Optimality

Recall Definition 1.3 of admissibility. d is inadmissible if there exists a d' such that $R(\theta, d') \leq R(\theta, d)$ for all θ and $R(\theta, d') < R(\theta, d)$ for some θ. d is admissible if it is not inadmissible. The class of admissible rules is usually too large for a given problem, so we need some further restrictions to get 'optimal' rules.

Definition 1.7

A decision rule d is said to optimal if
(i) d is admissible.
(ii) for any other admissible rule d', $d' = d$.
Thus optimal rules consist of unique admissible rules.

Example 1.9

Suppose $X_1, \ldots, X_n$ are iid, with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. Consider the estimation of μ so that $\boldsymbol{\mathcal{A}} = \boldsymbol{\Theta}$. Suppose we consider squared error loss. If we let $\boldsymbol{\mathcal{D}}$ = class of all possible estimators, then there is no optimal rule. Let $\boldsymbol{\mathcal{D}}$ = class of all possible linear estimators, so that $d(x) = \sum_{i=1}^{n} c_i X_i$, where $c_i \in \Re^1$ and known, $i = 1 \ldots, n$. It follows that the risk is

$$R(\theta, d) \equiv \mu^2 \left(\sum_{i=1}^{n} c_i - 1 \right)^2 + \sigma^2 \sum_{i=1}^{n} c_i^2 \ ,$$

where $\theta = (\mu, \sigma^2)$.

Now we show that there does not exist a d^* of the form $d^* = \sum_{i=1}^{n} c_i^* X_i$ such that

$$R(\theta, d^*) \leq R(\theta, d) \quad \text{for all } d \in \mathcal{D}, \theta \in \Theta \ .$$

If there is such a d^*, then $(c_1^*, \ldots, c_n^*)$ is a minimum of the function

$$R(\theta, d) = \mu^2 \left(\sum_{i=1}^{n} c_i - 1 \right)^2 + \sigma^2 \sum_{i=1}^{n} c_i^2 \ . \tag{1.2}$$

If $(c_1^*, \ldots, c_n^*)$ is a minimum of (1.2), then it must be that $c_i^* = \frac{\mu^2}{\sigma^2 + n\mu^2}$, which depends on the parameters, and thus d^* is not a statistic. This shows that there is no optimal estimator in the class of all linear estimators.

However, if we restrict the class to $\mathcal{D}$ = all linear estimators $d(x) = \sum_{i=1}^{n} c_i X_i$, with $\sum_{i=1}^{n} c_i = 1$, then we get an optimal estimator.

With the restriction $\sum_{i=1}^{n} c_i = 1$, (1.2) reduces to

$$R(\theta, d) = \sigma^2 \sum_{i=1}^{n} c_i^2 \ .$$

Now minimizing $\sigma^2 \sum_{i=1}^{n} c_i^2$ subject to $\sum_{i=1}^{n} c_i = 1$ leads to $c_i = \frac{1}{n}$ for all i, and thus $d(x) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$ is an optimal decision rule (admissible and unique) for the class of all linear estimators satisfying $\sum_{i=1}^{n} c_i = 1$.

Minimax Principle

Suppose d is a decision rule (may be randomized) and consider the quantity

$$\sup_{\theta \in \mathbf{\Theta}} R(\theta, d) \ . \tag{1.3}$$

(1.3) represents the worst that can happen if the rule d is used. If it is desired to protect against the worst possible state of nature, one is led to using the minimax principle

The Minimax Principle: A decision rule d_1 is preferred to a rule d_2 if

$$\sup_{\theta \in \mathbf{\Theta}} R(\theta, d_1) < \sup_{\theta \in \mathbf{\Theta}} R(\theta, d_2) \ .$$

Definition 1.8

A decision rule d_M is minimax if

$$\inf_{d \in \mathcal{D}} \{ \sup_{\theta \in \Theta} R(\theta, d) \} = \sup_{\theta \in \Theta} R(\theta, d_M) \ . \tag{1.4}$$

That is, a rule is minimax if it minimizes $\sup_{\theta \in \Theta} R(\theta, d)$ among all (randomized) rules $d \in \mathcal{D}$. The quantity on the right hand side of (1.4) is called the minimax value of the problem.

Example 1.3 (continued)
We have

$$\sup_{\theta} R(\theta, d) = \sup_{\theta} [c^2 + (1-c)^2 \theta^2] = \begin{cases} 1 & \text{if } c = 1 \\ \infty & \text{if } c \neq 1 \ . \end{cases}$$

Hence $d_1 \equiv d_{c=1}$ is best among all rules d_c according to the minimax rule and 1 is the minimax value for the problem.

Example 1.5 (continued) Consider the randomized action in the no-data problem

$$d(a) = \begin{cases} p, & \text{if } a = a_1 \\ 1-p, & \text{if } a = a_2 \end{cases}.$$

The loss (and hence risk) of such a rule is

$$\begin{aligned} R(\theta, d) &\equiv L(\theta, d) \\ &= pL(\theta, a_1) + (1-p)L(\theta, a_2) \\ &= \begin{cases} 1-2p, & \theta = \theta_1 \\ 2p-1, & \theta = \theta_2 \end{cases}. \end{aligned}$$

Recall the loss matrix given by

	a_1	a_2
θ_0	-1	1
θ_1	1	-1

so that $L(\theta_1, a_1) = -1$, $L(\theta_1, a_2) = 1$, $L(\theta_2, a_1) = 1$, and $L(\theta_2, a_2) = -1$.

Hence

$$\sup_{\theta} R(\theta, d) = \max\{1 - 2p, 2p - 1\}, \quad 0 \leq p \leq 1 \ .$$

If we graph the function $1 - 2p$ and $2p - 1$ for $0 \leq p \leq 1$ and noting that the maximum is always the higher of the two lines, it becomes clear that the minimum value of $\max\{1 - 2p, 2p - 1\}$ is 0, occurring at $p = \frac{1}{2}$. Thus $d_{\frac{1}{2}} \equiv d_{p=\frac{1}{2}}$ is the minimax action and 0 is the minimax value for the problem.

A unique minimax estimator is admissible since any estimator better than a minimax estimator is also minimax.
The concept of minimaxity is best explained within a Bayesian framework, since it will turn out that Bayes estimators (with proper priors) are minimax and admissible.

Definition 1.9

A probability distribution Λ over $\boldsymbol{\Theta}$ is called a prior distribution.

Definition 1.10

For any given prior Λ and $d \in \mathcal{D}$, the Bayes risk of d with respect to Λ is

$$\mathcal{R}(\Lambda, d) = \begin{cases} \int_{\boldsymbol{\Theta}} R(\theta, d)\, d\Lambda(\theta) \\ \sum_{i=1}^{l} R(\theta_i, d)\lambda_i, \quad \text{if} \quad \boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_l\} \, . \end{cases}$$

Definition 1.11

A Bayesian decision rule with respect to Λ, d_Λ, is any rule satisfying

$$\mathcal{R}(\Lambda, d_\Lambda) = \inf_{d \in \mathcal{D}} \mathcal{R}(\Lambda, d) = \text{Bayes Risk} .$$

Example 1.10

Suppose $\Theta = \{1, 2\}$.
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 comes from if the losses $L(\theta, a)$ are given by

	a_1	a_2
θ_1	0	10
θ_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

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

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

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

This is minimized by choosing $d_x = 1$ if its coefficient is negative and 0 if its coefficient is positive. For example, if $\lambda = \frac{1}{2}$, 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 $\lambda = \frac{1}{2}$ is $d_\Lambda = (0, 0, 1)$. Note that the Bayes rule is in fact a nonrandomized rule. This gives us the Bayes risk for $\lambda = \frac{1}{2}$ as $\mathcal{R}\left(\frac{1}{2}, d_\Lambda\right) = 2.1$.

The minimax rule is $d_M = (0, \frac{9}{22}, 1)$, 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 $R(1, d_M) = \frac{24}{11} = R(2, d_M)$.

The following table shows 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

Example 1.11

Consider finding the Bayes risk in an estimation problem. In this case $\mathcal{A} = \Theta$. Suppose we assume the loss function $L(\theta, a) = (\theta - a)^2$ and observe the value of the random variable X where $X \sim U(0, \theta)$, and

$$p(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & \text{otherwise} \end{cases}.$$

We want to find the Bayes risk for the rule $d(x) = cX$ with respect to the prior

$$\lambda(\theta) = \begin{cases} \theta e^{-\theta}, & \theta > 0 \\ 0, & \theta < 0 \end{cases}.$$

Now

$$\mathcal{R}(\Lambda, d) = \int_{\Theta} R(\theta, d)\ \lambda(\theta)\ d\theta\ .$$

$$
\begin{aligned}
R(\theta, d) &= \int_0^\theta (\theta - d(x))^2 \left(\frac{1}{\theta}\right) \, dx \\
&= \int_0^\theta (\theta - cx)^2 \left(\frac{1}{\theta}\right) \, dx \\
&= \int_0^\theta (\theta^2 - 2\theta cx + c^2x^2) \left(\frac{1}{\theta}\right) \, dx \\
&= \int_0^\theta \left(\theta - 2cx + \frac{c^2x^2}{\theta}\right) \, dx \\
&= \left(\theta x - cx^2 + \frac{c^2x^3}{3\theta}\right)\Bigg|_0^\theta \\
&= \theta^2 - c\theta^2 + \frac{c^2\theta^3}{3\theta} \\
&= \theta^2(1-c) + \frac{1}{3}c^2\theta^2 \, .
\end{aligned}
$$

Now the Bayes risk is

$$\mathcal{R}(\Lambda, d) = \int_0^\infty \left[\theta^2(1-c) + \frac{1}{3}c^2\theta^2\right]\left(\theta e^{-\theta}\right)\, d\theta$$

which can be evaluated in closed form using integration by parts.
Suppose we wanted to find the Bayes rule d_Λ. d_Λ is the rule that minimizes the Bayes risk,

$$\mathcal{R}(\Lambda, d_\Lambda) = \inf_{d \in \boldsymbol{\mathcal{D}}} \mathcal{R}(\Lambda, d)\ . \tag{1.5}$$

Formula (1.5) is not convenient for finding Bayes rules since we would have to minimize

$$\int_{\boldsymbol{\Theta}} R(\theta, d(x))\lambda(\theta)\, d\theta$$

with respect to $d(x)$, without knowing the form of $d(x)$.

A more convenient way of finding Bayes rules is to first write

$$R(\theta, d) = \int_{\mathcal{X}} L(\theta, d(x)) p(x|\theta)\, dx,$$

and thus the Bayes risk is

$$\begin{aligned}
&\int_{\Theta} \int_{\mathcal{X}} L(\theta, d(x)) p(x|\theta) \lambda(\theta)\, dx\, d\theta \\
&\quad = \int_{\mathcal{X}} \left[\int_{\Theta} L(\theta, d(x)) \left\{ \frac{p(x|\theta)\lambda(\theta)}{\int p(x|\theta)\lambda(\theta)\, d\theta} \right\} d\theta \right] \left(\int p(x|\theta)\lambda(\theta)\, d\theta \right) dx \\
&\quad = \int_{\mathcal{X}} \left\{ \int_{\Theta} L(\theta, d(x)) p(\theta|x)\, d\theta \right\} p(x)\, dx\ ,
\end{aligned}$$

where $p(x) = \int_{\Theta} p(x|\theta)\lambda(\theta)\, d\theta$ is the marginal distribution of X.

And thus by a change in the order of integration, we can write

$$\mathcal{R}(\Lambda, d) = \int_{\boldsymbol{\mathcal{X}}} \left[\int_{\boldsymbol{\Theta}} L(\theta, d(x)) p(\theta|x)\, d\theta \right] p(x)\, dx. \tag{1.6}$$

Using (1.6), it is easy to describe a Bayes decision rule. To find a function $d(x)$ that minimizes the double integral (1.6), we may minimize the inside integral separately for each x, that is, for each x, we minimize

$$\int L(\theta, d(x)) p(\theta|x)\, d\theta\ . \tag{1.7}$$

$p(\theta|x) = \frac{p(x|\theta)\lambda(\theta)}{\int p(x|\theta)\lambda(\theta)\, d\theta}$ is called the posterior distribution of θ. Thus (1.7) is called the posterior expected loss.

Theorem 1.1

The Bayes rule d_Λ, is the rule which minimizes (w.r.t. $d(x)$) the posterior expected loss.

$$\text{Bayes rule} \quad = \arg\min_{d(x)} \left[\int L(\theta, d(x)) p(\theta|x) \, d\theta \right] .$$

The Bayes rule is sometimes referred to as the Bayes action. The Bayes risk is then computed via (1.6).

Again suppose $\mathcal{A} = \Theta$, $L(\theta, a) = c(\theta - a)^2$, $c \in \Re^1$. Let us find the Bayes rule for this loss function. The posterior expected loss is

$$G(a) = \int_\Theta c(\theta - a)^2 p(\theta|x) \, d\theta .$$

Minimizing $G(a)$ with respect to a, we get,

$$\frac{dG(a)}{da} = \int_{\Theta} -2c(\theta - a)p(\theta|x)\, d\theta = 0 \ ,$$

$$\begin{aligned} \Rightarrow \quad a &= \int_{\Theta} \theta p(\theta|x)\, d\theta \\ &= E[\theta|x] \ . \end{aligned}$$

Thus the Bayes rule for quadratic loss is the posterior mean of θ. In estimation problems, we refer to the Bayes rule as the Bayes estimator of θ.

Theorem 1.2

If $L(\theta, a) = c|\theta - a|$, for some $c > 0$, the Bayes estimator of θ is the posterior median.

Theorem 1.3

If

$$L(\theta, a) = \begin{cases} 0, & |\theta - a| \leq c, \quad c > 0 \\ 1, & |\theta - a| > c, \quad c > 0 \end{cases}$$

then the Bayes estimator of θ converges to the posterior mode of θ as $c \to 0$.

Theorem 1.4

If

$$L(\theta, a) = \begin{cases} k_1|\theta - a|, & a \leq \theta \\ k_2|\theta - a|, & a > \theta \end{cases},$$

for some $k_1 > 0$ and $k_2 > 0$, then the Bayes estimator of θ is the posterior *pth* quantile where p depends on k_1 and k_2.

Example 1.11 (continued) We can find the Bayes rule by minimizing the posterior expected loss, leading to

$$d(x) = E(\theta|x) \ .$$

The posterior distribution is given by

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta)\lambda(\theta)}{\int p(x|\theta)\lambda(\theta)\, d\theta} \\ &= \frac{\left(\frac{1}{\theta}\right)\theta e^{-\theta} I(0 < x < \theta)}{\int_x^\infty \left(\frac{1}{\theta}\right)(\theta e^{-\theta})\, d\theta} \\ &= e^{-\theta} e^{x} I(\theta > x) \\ &= \begin{cases} e^{x-\theta} & \theta > x \\ 0 & \text{otherwise} \ . \end{cases} \end{aligned}$$

$$E(\theta|x) = \int_x^\infty \theta e^{x-\theta}\, d\theta = d_\Lambda \quad = \quad \text{Bayes rule}$$
$$= \quad x+1.$$

Structure of the Risk Body: The Finite Discrete Case

Suppose that $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_l\}$, $\boldsymbol{\mathcal{X}} = \{x_1, \ldots, x_m\}$, $\boldsymbol{\mathcal{A}} = \{a_1, \ldots, a_k\}$.
$d(a_i, x_j) = P(\text{probability of action } a_i \text{ when } X = x_j)$, $\boldsymbol{\mathcal{D}} = \{\text{all decision functions}\}$.

Definition 1.12

$\boldsymbol{\mathcal{A}} \subset \mathfrak{R}^l$ is convex if , for all $\lambda \in [0,1]$, $x, y \in \boldsymbol{\mathcal{A}}$,

$$\lambda x + (1-\lambda)y \in \boldsymbol{\mathcal{A}}.$$

Lemma 1.1

Let $\lambda = (\lambda_1, \ldots, \lambda_l)$ be a probability distribution, $\lambda_i \geq 0$, $\sum_{i=1}^{l} \lambda_i = 1$. Then for any $d_1, \ldots, d_l \in \mathcal{D}$,

$$\sum_{n=1}^{l} \lambda_n d_n \in \mathcal{D}.$$

Lemma 1.1 says that a convex combination of decision rules belongs to $\mathcal{D}$, the set of all possible decision rules.

Proof:

Set $d(a_i, x_j) = \sum_{n=1}^{l} \lambda_n d_n(a_i, x_j)$. Then $d(a_i, x_j) \geq 0$ and

$$\sum_{i=1}^{k} d(a_i, x_j) = \sum_{n=1}^{l} \lambda_n \sum_{i=1}^{k} d_n(a_i, x_j) = 1.$$

We will call

$$\boldsymbol{R} = \{(R(\theta_1, d), \ldots, R(\theta_l, d)) \in \Re^{+l} : d \in \boldsymbol{\mathcal{D}}\}$$

the risk body (or risk set).

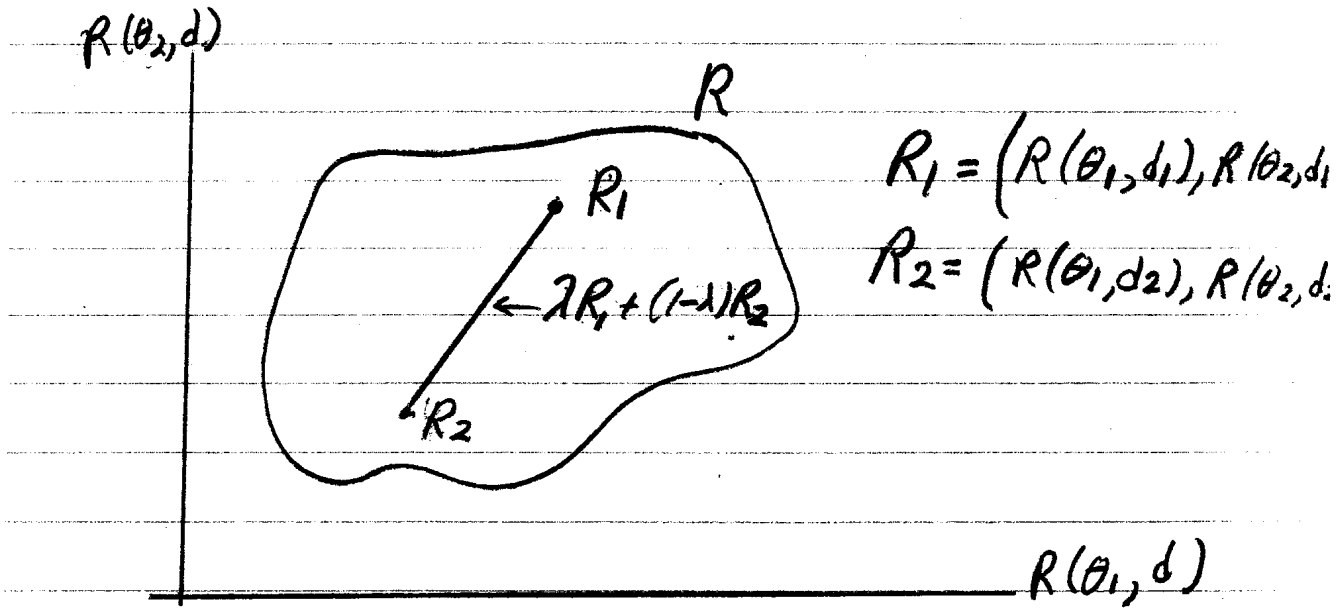
Theorem 1.5

The risk body is convex.

Proof: Let $d_1, d_2 \in \mathcal{D}$ denote the rules corresponding to $\boldsymbol{R}_i = (R(\theta_1, d_i), \ldots, R(\theta_l, d_i))$, $i = 1, 2$. Set $d = \lambda d_1 + (1-\lambda)d_2 \in \mathcal{D}$ by the lemma.
Then

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

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


R(θ2, d)
R
R1
λR1 + (1−λ)R2
R2
R(θ1, d)
R1 = (R(θ1, d1), R(θ2, d1
R2 = (R(θ1, d2), R(θ2, d2

Theorem 1.6

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

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

Then there are k^m nonrandom decision rules, call them $\delta^{(r)} = (\delta_{ij}^{(r)})$, $1 \leq r \leq k^m$, Given $d \in \mathcal{D}$ we want to find $\lambda_l \geq 0$, $\sum_{l=1}^{k^m} \lambda_l = 1$, so that

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

Claim:

$$\begin{aligned}\lambda_r &= \prod_{j=1}^{m}\prod_{i=1}^{k} d_{ij}^{\delta_{ij}^{(r)}} \\ &= \prod_{j=1}^{m}\left\{\sum_{i=1}^{k} d_{ij}\delta_{ij}^{(r)}\right\}\end{aligned}$$

works.

Proof: Induction on m.

The convexity of the risk body and the decision rules imply that the risk body contains all of the information about the decision problem.

Relations Between Bayes, Minimax, and Admissibility for the Finite Discrete Case

We provide a number of results about Bayes and minimax rules and connections between them which carry over to more general problems.

Theorem 1.7

If the risk set contains its boundary points, then a minimax rule always exists.

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

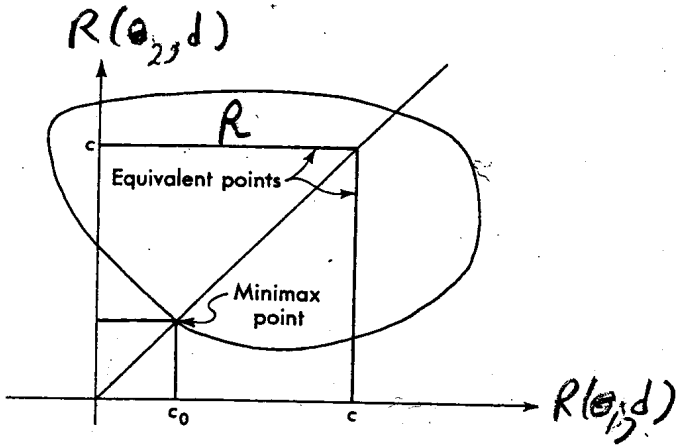


Figure 1.2

Minimax Point:

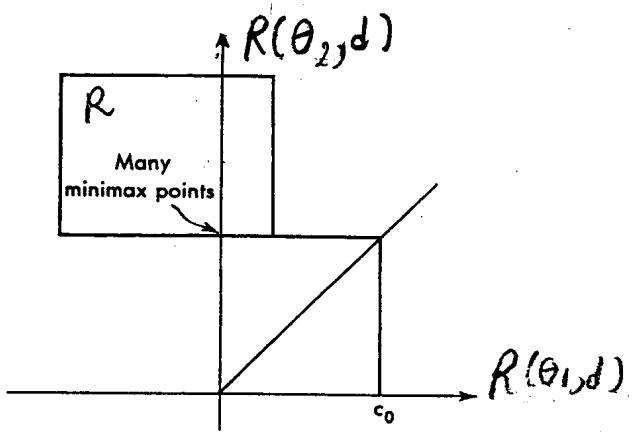
$y_1 = R(\theta_1, d_M) = c_0$

$y_2 = R(\theta_2, d_M) = c_0$

The maximum risk for a fixed rule d is $\max_i R(\theta_i, d) = \max_i y_i$, $y_i = R(\theta_i, d)$, $i = 1, \ldots, l$. Thus any point $\boldsymbol{y}, \boldsymbol{y} \in \boldsymbol{R}$, $\boldsymbol{y} = (y_1, \ldots, y_l)$, that gives rise to the same value of $\max_i y_i$ are equivalent in the ordering given by the minimax principle. Thus, all points on the boundary of the set

$$\boldsymbol{Q}_c = \{(y_1, \ldots, y_l) : y_i \leq c, i = 1, \ldots, l\}$$

for any real number c are equivalent. To find the minimax rules, we find the infinum of those values c, call it c_0, such that the set $\boldsymbol{Q}_c$ intersects $\boldsymbol{R}$ = risk body (risk set). Any decision rule d whose associated risk point is an element of the intersection $\boldsymbol{Q}_{c_0} \cap \boldsymbol{R}$ is a minimax decision rule. Of course, minimax decision rules do not exist when the risk set $\boldsymbol{R}$ does not contain its boundary points.


$R(\theta_2, d)$
R
Many
minimax points
$R(\theta_1, d)$
c_0

We note in the previous figure that there may be many minimax points for a given decision problem and that minimax points do not have to lie on the diagonal line $y_1 = y_2 = \cdots = y_l$.

Theorem 1.8

If $\Lambda = (\lambda_1, \ldots, \lambda_l)$ is a prior, then Bayes decision rules have risk point on the hyperplane $\{\boldsymbol{y} \in \mathfrak{R}^l : \sum_{i=1}^{l} \lambda_i y_i = b_0\}$, where

$$b_0 = \inf\{b \geq 0 : \text{ the plane determined by } \sum_{i=1}^{l} \lambda_i y_i = b \text{ that intersects } \boldsymbol{R} = \text{risk body}\}$$

Proof:
$\mathcal{R}(\Lambda, d) = \sum_{i=1}^{l} \lambda_i R(\theta_i, d)$ is the Bayes risk and we want to minimize it.

Lemma 1.2

An admissible rule is a Bayes rule for some prior Λ.

A prior distribution for Θ, when Θ consists of l points is merely an l-tuple of non-negative numbers $(\lambda_1, \ldots, \lambda_l)$, where $\sum_{i=1}^{l} \lambda_i = 1$, where $\lambda_j =$ probability of $\theta_j, j = 1, \ldots, l$. All points that yield the same expected risk,

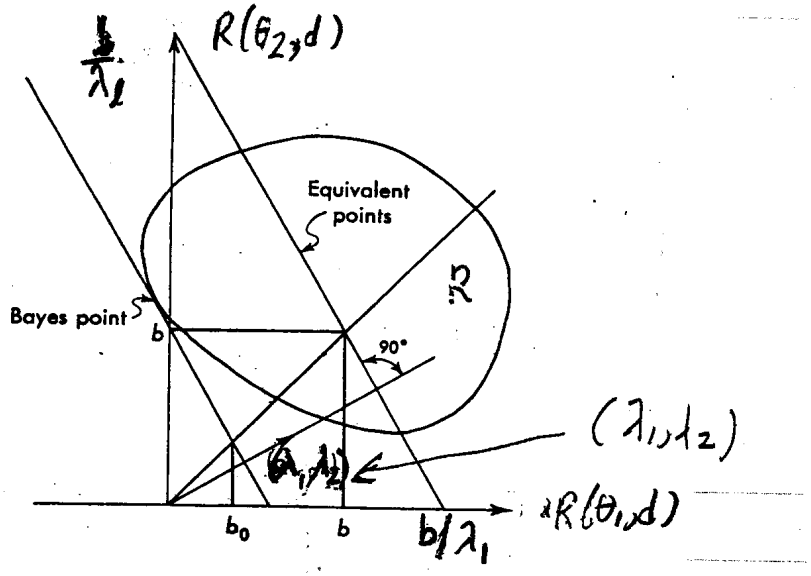
$$\sum_{j=1}^{l} \lambda_j R(\theta_j, d) = \sum_{j=1}^{l} \lambda_j y_j$$

are equivalent in the ordering for the prior $(\lambda_1, \ldots, \lambda_l)$.

Thus all points on the plane $\sum_{j=1}^{l} \lambda_j y_j = b$ for any real number b are equivalent. Every such plane is perpendicular to the vector from the origin to the point $(\lambda_1, \ldots, \lambda_l)$, and because each λ_j is non-negative, the slope of the line of intersection of the plane $\sum_{j=1}^{l} \lambda_j y_j = b$ with the coordinate planes cannot be positive.

The quantity b can be visualized by noting that the point of intersection of the diagonal line $y_1 = y_2 = \cdots = y_l$ with the plane $\sum_{j=1}^{l} \lambda_j y_j = b$ must occur at $(b, \ldots, b)$. To find Bayes rules, we find the infinum of those values of b, call it b_0, for which the plane $\sum_{j=1}^{l} \lambda_j y_j = b$ intersects the risk set $\boldsymbol{R}$.

Decision rules corresponding to points in this intersection are Bayes rules with respect to the prior distribution $(\lambda_1, \ldots, \lambda_l)$. Of course, Bayes rules do not exist when the risk set $\boldsymbol{R}$ does not contain its boundary points.


$R(\theta_2, d)$
$\frac{b}{\lambda_1}$
Equivalent points
Bayes point
b
90°
(λ_1, λ_2)
(λ_1, λ_2)
$R(\theta_1, d)$
b_0
b
b/λ_1

In order to understand these concepts better, recall that the risk set $= \boldsymbol{R} = \{(R(\theta_1, d), \ldots, R(\theta_l, d)) \in \Re^{+l}$, for all $d \in \boldsymbol{\mathcal{D}}\}$. Thus the risk set consist of those points $(y_1, \ldots, y_l)$ for all $d \in \boldsymbol{\mathcal{D}}$. For example, for $d \in \boldsymbol{\mathcal{D}}$, $(y_1, \ldots, y_l)$ is a point in the risk set. For $d^* \in \boldsymbol{\mathcal{D}}$, $(y_1^*, \ldots, y_l^*)$ is a point in the risk set where $y_1^* = R(\theta_1, d^*), \ldots, y_l^* = R(\theta_l, d^*)$. Risk points are determined by different d's. The dimension of the risk points is determined by $\boldsymbol{\Theta}$. The minimax point is the point in the risk set that achieves the minimax value.

$$y_1 = R(\theta_1, d_M) = c_0$$

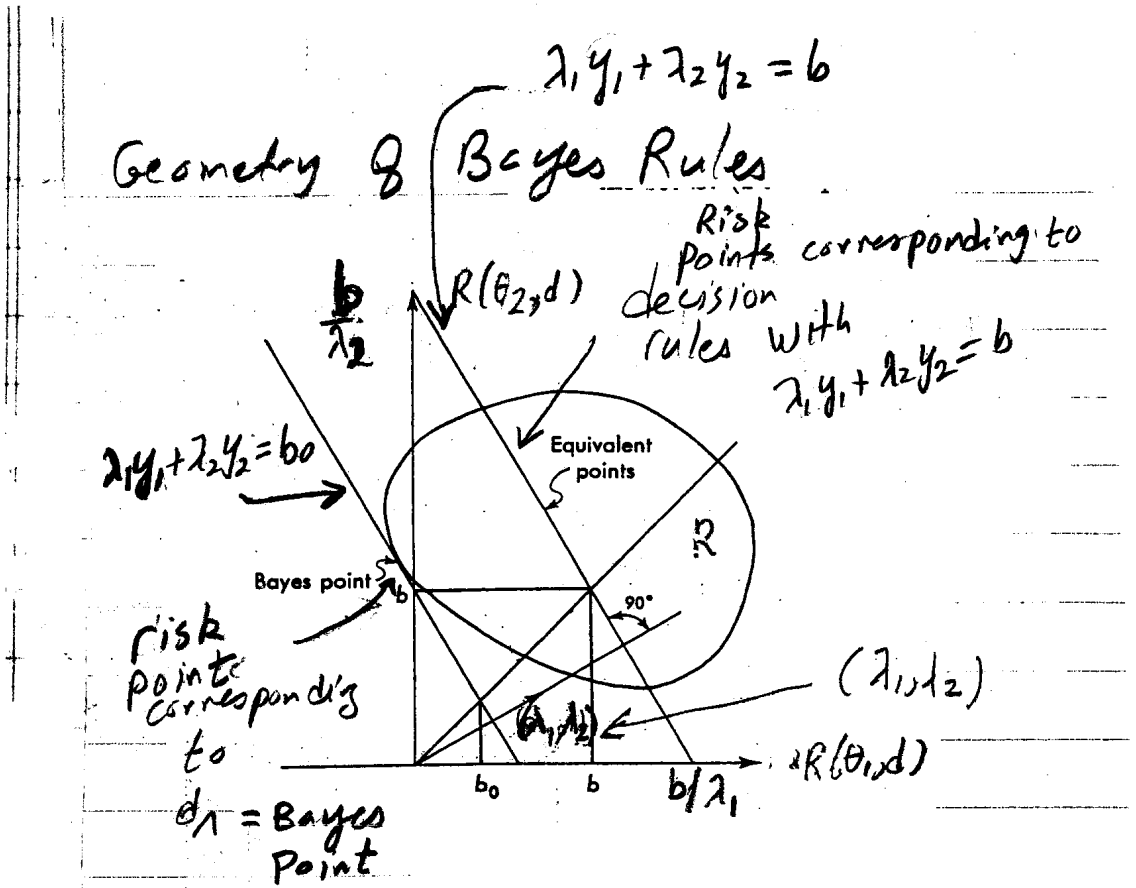
$$y_2 = R(\theta_2, d_M) = c_0$$

The Bayes point is the point in the risk set that achieves the minimum Bayes risk.

$$y_1 = R(\theta_1, d_\Lambda)$$

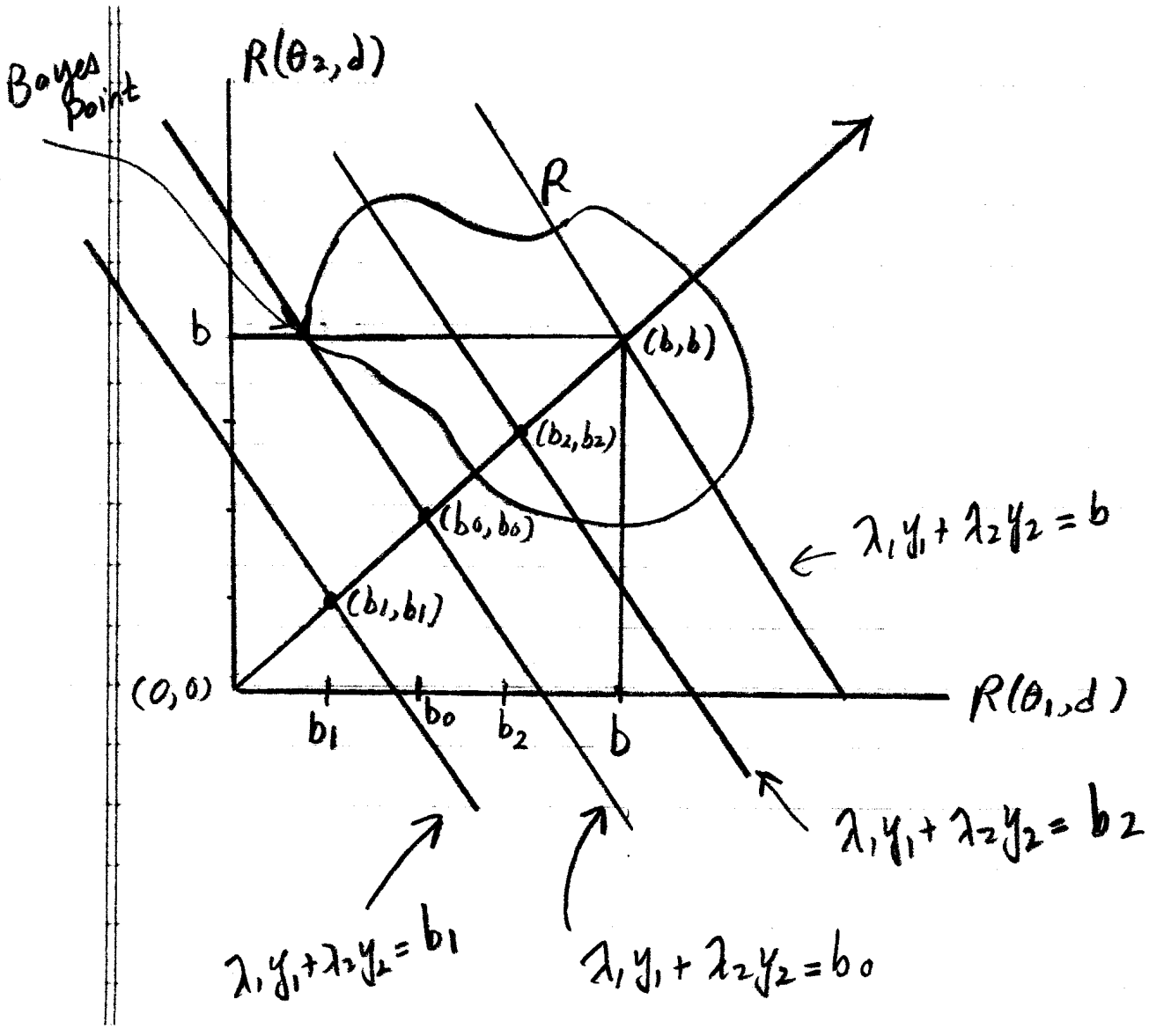
$$y_2 = R(\theta_2, d_\Lambda)$$

$$\mathcal{R}(\Lambda, d_\Lambda) = \lambda_1 y_1 + \lambda_2 y_2 = b_0.$$

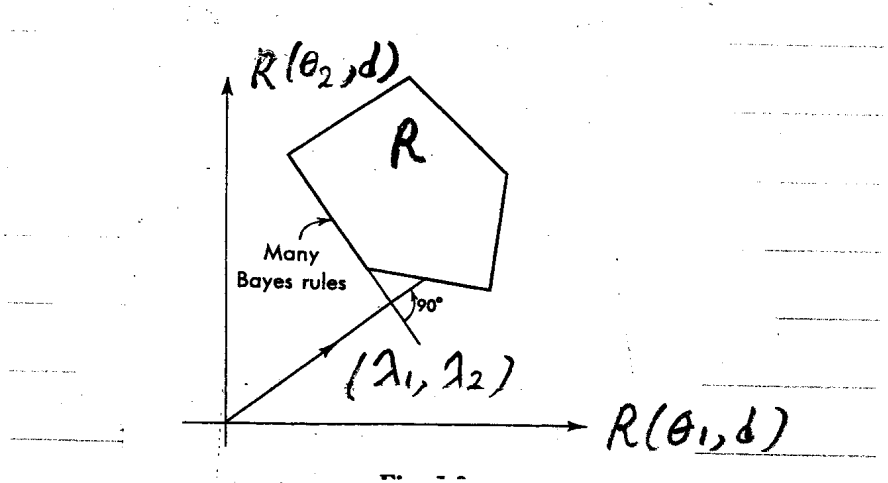

$\lambda_1 y_1 + \lambda_2 y_2 = b$
Geometry of Bayes Rules
Risk Points corresponding to decision rules with $\lambda_1 y_1 + \lambda_2 y_2 = b$
$R(\theta_2, d)$
$\frac{b}{\lambda_2}$
$\lambda_1 y_1 + \lambda_2 y_2 = b_0$
Equivalent points
R
Bayes point
b
90°
risk point corresponding to d_λ = Bayes Point
(λ_1, λ_2)
(θ_1, λ_2)
$R(\theta_1, d)$
b_0
b
b/λ_1

The Bayes point in the previous figure is the point (y_1, y_2) in the risk set $\boldsymbol{R}$ that yields the minimum Bayes risk. That is, the Bayes point is given by $y_1 = R(\theta_1, d_\Lambda)$, $y_2 = R(\theta_2, d_\Lambda)$ and $\mathcal{R}(\Lambda, d_\Lambda) = \lambda_1 y_1 + \lambda_2 y_2 = b_0$. Note that the Bayes risk generates a hyperplane (line) of the form $\lambda_1 y_1 + \lambda_2 y_2 + \cdots \lambda_l y_l$.

Bayes rules are those rules in which the infinum of the values, call it b_0, intersects the risk set $\boldsymbol{R}$. Decision rules corresponding to points in this intersection are Bayes rules with respect to the prior distribution $(\lambda_1, \ldots, \lambda_l)$.


Bayes point
$R(\theta_2, d)$
R
b
(b, b)
(b_2, b_2)
(b_0, b_0)
(b_1, b_1)
$\lambda_1 y_1 + \lambda_2 y_2 = b$
$(0, 0)$
b_1
b_0
b_2
b
$R(\theta_1, d)$
$\lambda_1 y_1 + \lambda_2 y_2 = b_2$
$\lambda_1 y_1 + \lambda_2 y_2 = b_1$
$\lambda_1 y_1 + \lambda_2 y_2 = b_0$

There may be many Bayes points with respect to a given prior, if the prior is $\lambda_1 = 0, \ldots, \lambda_{l-1} = 0, \lambda_l = 1$, for example. The next figure illustrates this point.


$R(\theta_2, d)$
R
Many
Bayes rules
90°
(λ_1, λ_2)
$R(\theta_1, d)$

Now all admissible rules, by definition, are on the boundary of the risk set, and thus all admissible rules are Bayes rules. Note, however that not all Bayes rules are admissible. A Bayes rule, for example with respect to the prior $\lambda_1 = 0, \ldots, \lambda_{l-1} = 0, \lambda_l = 1$ may not be admissible. Thus, the class of admissible rules is a subset of the class of Bayes rules.

Theorem 1.9

Suppose that d_0 is a Bayes rule with respect to $\Lambda = (\lambda_1, \ldots, \lambda_l)$ and $\lambda_i > 0$, $i = 1, \ldots, l$. Then d_0 is admissible.

Proof: Suppose that d_0 is not admissible. Then there is a rule d better than d_0, that is,

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

with strict inequality for some i.

Since $\lambda_i > 0$ for $i = 1, \ldots, l$,

we have that

$$\sum_{i=1}^{l} \lambda_i R(\theta_i, d) < \sum_{i=1}^{l} \lambda_i R(\theta_i, d_0) \ ,$$

and thus $\mathcal{R}(\Lambda, d) < \mathcal{R}(\Lambda, d_0)$. This is a contradiction since d_0 is Bayes w.r.t. λ.

Theorem 1.10

If $d_\Lambda \in \mathcal{D}$ is Bayes for Λ, and it has constant risk, then d_Λ is minimax.

Proof:
Let r_0 be the constant risk. Assume d_Λ is not minimax and let d_M be the minimax rule (which exists). Then

(a)

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

(b)

$$\min_{d\in\mathcal{D}} \mathcal{R}(\Lambda, d) = \mathcal{R}(\Lambda, d_\Lambda) = \sum_{i=1}^{l} \lambda_i R(\theta_i, d_\Lambda) = r_0$$

But (a) yields

$$\mathcal{R}(\Lambda, d_M) = \sum_{i=1}^{l} \lambda_i R(\theta_i, d_M) < \sum_{i=1}^{l} \lambda_i r_0 = r_0$$

which contradicts (b).

Example 1.12

Suppose $X \sim \text{Binomial}(n, \theta)$, $L(\theta, a) = \frac{(\theta-a)^2}{\theta(1-\theta)}$, and $\lambda(\theta) = 1$, $0 \leq \theta < 1$.
(a) Let us find the Bayes rule $d(x)$. The Bayes rule is the rule $d(x)$ which minimizes the posterior expected loss.

$$\begin{aligned} E[L(\theta,a)|x] &= \int_{\Theta} L(\theta,a)p(\theta|x)\,d\theta \\ &= \int_{\Theta} \frac{(\theta-a)^2}{\theta(1-\theta)}p(\theta|x)\,d\theta \end{aligned}$$

$$\begin{aligned} p(\theta|x) = \frac{p(x|\theta)\lambda(\theta)}{\int p(x|\theta)\lambda(\theta)\,d\theta} &= \frac{\theta^x(1-\theta)^{n-x}}{\int \theta^x(1-\theta)^{n-x}\,d\theta} \\ &\propto \theta^x(1-\theta)^{n-x} \\ &= \theta^{x+1-1}(1-\theta)^{n-x+1-1}\,. \end{aligned}$$

Thus $\theta|x \sim \text{Beta}(x+1, n-x+1)$

$$\frac{d}{da}E[L(\theta,a)|x] = 0$$

$$\Rightarrow -2\int_{\Theta}\frac{(\theta-a)}{\theta(1-\theta)}p(\theta|x)=0$$

$$\Rightarrow \int_{\Theta}(1-\theta)^{-1}p(\theta|x)\,d\theta = a\int_{\Theta}\theta^{-1}(1-\theta)^{-1}p(\theta|x)\,d\theta$$

$$\Rightarrow a=\frac{\int_{\Theta}(1-\theta)^{-1}p(\theta|x)\,d\theta}{\int_{\Theta}\theta^{-1}(1-\theta)^{-1}p(\theta|x)\,d\theta}$$

$$\Rightarrow a=\frac{\int_0^1(1-\theta)^{-1}\theta^{x+1-1}(1-\theta)^{n-x+1-1}\,d\theta}{\int_0^1\theta^{-1}(1-\theta)^{-1}\theta^{x+1-1}(1-\theta)^{n-x+1-1}\,d\theta}$$

$$\Rightarrow a=\frac{\int_0^1\theta^{x+1-1}(1-\theta)^{n-x-1}\,d\theta}{\int_0^1\theta^{x-1}(1-\theta)^{n-x-1}\,d\theta}$$

$$\Rightarrow a = \frac{B(x+1, n-x)}{B(x, n-x)} = \frac{x}{n} \equiv d(x) \ .$$

Recall that $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Thus, the Bayes estimator of θ (Bayes rule) is $d_\Lambda = d(x) = \frac{X}{n}$.
Let us calculate the Bayes risk.

$$\mathcal{R}(\Lambda, d_\Lambda) = \int R(\theta, d_\Lambda)\ \lambda(\theta)\ d\theta.$$

$$
\begin{aligned}
R(\theta, d) &= E_{x|\theta}[L(\theta, d(x))] \\
&= E\left\{\frac{(\theta - \frac{x}{n})^2}{\theta(1-\theta)}\right\} \\
&= \theta^{-1}(1-\theta)^{-1} E\left[\theta^2 - \frac{2\theta x}{n} + \frac{x^2}{n^2}\right] \\
&= \theta^{-1}(1-\theta)^{-1}\left[\theta^2 - \frac{2\theta}{n}(n\theta) + \frac{1}{n^2}\left(n\theta(1-\theta) + n^2\theta^2\right)\right] \\
&= \frac{1}{n},
\end{aligned}
$$

and so

$$
\mathcal{R}(\Lambda, d_\Lambda) = \int \frac{1}{n}\lambda(\theta)\, d\theta = \frac{1}{n}\ .
$$

Since the Bayes estimator $d_\Lambda \equiv \frac{X}{n}$ has constant risk $R(\theta, d_\Lambda) = \frac{1}{n}$, $d(x) = \frac{X}{n}$ is minimax.

Calculating Posterior Distributions

Definition 1.13

If $\theta \sim \Lambda$ is a prior over $\boldsymbol{\Theta}$, then the conditional distribution of θ given x, $p(\theta \in A|x)$, $A \subset B(\boldsymbol{\Theta})$, is called the posterior distribution of θ. We write $\Lambda(\theta|x) = p(\theta \leq \theta|X = x)$ for the conditional distribution function of θ given $X = x$.

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

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

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

Definition 1.14

Suppose X has density $p(x|\theta)$, and λ is a prior density, $\lambda \in P_\Lambda$ for θ. If $\lambda(\theta|x)$ has the same form as the prior, i.e. $\lambda(\theta|x) \in P_\Lambda$, then λ is said to be a conjugate prior for θ.

Example 1.13

Suppose $X|\theta \sim \text{Poisson}(\theta)$, $\theta \sim \text{Gamma}(\alpha, \beta)$, then

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

$$\lambda(\theta|x) = \text{Gamma}(x+\alpha, \beta+1), \quad \text{that is}$$

$$\lambda(\theta|x) = \frac{(\beta+1)^{x+\alpha}}{\Gamma(x+\alpha)}\theta^{x+\alpha-1}e^{-\theta(\beta+1)},$$

and thus the gamma prior is the conjugate prior for the Poisson(θ) distribution.

Example 1.14

Suppose $X|\theta \sim N(\theta, \sigma^2)$, $\theta \sim N(\mu, \tau^2)$, σ^2 is known. Then

$$\theta|x \sim N\left(w\mu + (1-w)x, \frac{1}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}}\right)$$

where $w = \frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2}+\frac{1}{\sigma^2}}$.

Note that if we have a random sample $X_1, \ldots, X_n$, then by sufficiency, we can replace X with $\bar{X}$ above and obtain

$$\theta|\bar{x} \sim N\left(w\mu + (1-w)\bar{x}, \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right)$$

where $w = \frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}$.

Thus, the normal prior is the conjugate prior for the $N(\theta, \sigma^2)$ family of distributions.

Example 1.15

If $X|\theta \sim \text{Binomial}(n, \theta)$, $\theta \sim \text{Beta}(\alpha, \beta)$, then $\theta|x \sim \text{Beta}(\alpha + x, \beta + n - x)$, and thus the Beta prior is the conjugate prior for the $\text{Binomial}(n, \theta)$ distribution.

Example 1.16

Suppose that $\boldsymbol{X}|\boldsymbol{\theta} \sim \text{Multinomial}_k(n, \boldsymbol{\theta})$, $\boldsymbol{X} = (X_1, \ldots, X_k)$. $\boldsymbol{\theta}$ is $(k-1) \times 1$, and $\boldsymbol{\theta} \sim \text{Dirichlet}_{k-1}(\alpha_1, \ldots, \alpha_k)$, which means

$$\lambda(\boldsymbol{\theta}) = \left(\frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \right) \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

where $\theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$, $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$, $\alpha_i > 0$.

Then

$$\boldsymbol{\theta}|\boldsymbol{x} \sim \text{Dirichlet}_{k-1}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$$

and the posterior mean of $\boldsymbol{\theta}$ is

$$E(\boldsymbol{\theta}|\boldsymbol{x}) = \left(\frac{\alpha_1 + x_1}{\sum_{i=1}^{k} \alpha_i + n}, \cdots, \frac{\alpha_k + x_k}{\sum_{i=1}^{k} \alpha_i + n} \right) .$$

Finding Bayes Rules

If the loss function is convex in a, then it suffices to consider only non-randomized Bayes rules. Corollary 7.9 of Lehmann and Casella (Theory of Point Estimation), states that: given any randomized estimator, there exists a non-randomized estimator which is uniformly better if the loss function is strictly convex in a, and at least as good when it is convex. A function $f(x)$ is convex if for $x \in S, y \in S$

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) .$$

Remark 1.2

Bayes rules can also be defined with improper priors. An improper prior implies a non finite measure on $\boldsymbol{\Theta}$. An improper prior for $\boldsymbol{\Theta}$, is a prior for which $\int d\Lambda(\theta) = \infty$, or if $\lambda(\theta)$ is the density of θ w.r.t. Lebesgue measure, $\int_{\boldsymbol{\theta}} \lambda(\theta)\, d\lambda = \infty$.

Proper posterior distributions often are still obtained with improper priors. With improper priors, $p(x)$ has infinite mass. Nonetheless we can proceed and seek a function $d(x)$ which minimizes

$$\int L(\theta, d) p(\theta|x)\, d\theta$$

as long as $p(\theta|x)$ is proper, i.e. $\int_{\boldsymbol{\Theta}} p(\theta|x)\, d\theta = 1$. Such a minimizing rule is called a generalized Bayes rule.

Definition 1.15

A rule d_{GB} is said to be a generalized Bayes rule if there exists a measure $\lambda(\theta)$ on $\boldsymbol{\Theta}$ such that,

$$\int L(\theta, d)p(\theta|x)\ d\theta$$

takes on a finite minimum when $d = d_{GB}$.

Example 1.17

Suppose $X|\theta \sim N(\theta, 1)$, $\lambda(\theta) \propto 1$, $L(\theta, a) = (\theta - a)^2$. Consider the decision rule $d(x) \equiv X$. Is $d(x)$ a generalized Bayes rule? We know that the Bayes rule is the posterior mean of θ since $L(\theta, a) = (\theta - a)^2$.
Now, one can easily show that

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)\lambda(\theta) \\ &\propto e^{-\frac{1}{2}(x-\theta)^2} \end{aligned}$$

and thus

$$\theta|x \sim N(x,1)\ ,$$

$$\text{and} \quad E(\theta|x) = x = d(x)\ .$$

Thus, $d(x) = X$ is indeed a generalized Bayes rule with respect to the uniform improper prior on the real line.

Most of our discussion below deals with Bayes rules, that is, rules based on proper priors.

Theorem 1.11

Suppose $\theta \sim \Lambda$, $(x|\theta) \sim P_{\boldsymbol{\theta}}$ and $L(\theta, a) \geq 0$ for all $\theta \in \boldsymbol{\Theta}, a \in \boldsymbol{\mathcal{A}}$.

If

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

(ii) there exists a d_Λ minimizing

$$E\left[L(\theta, d(x))|X = x)\right] = \int_{\Theta} L(\theta, d(x))\lambda(\theta|x)\, d\nu(\theta)$$

for each x, then $d_\Lambda(x)$ is a Bayes rule.

Proof: For any rule d with finite risk

$$E[L(\theta, d(x))|x] \geq E[L(\theta, d_\Lambda(x))|x]$$

almost surely by **(ii)**. Hence

$$\begin{aligned}
\mathcal{R}(\Lambda, d) &= E_x E_{\theta|x}[L(\theta, d(x))|x] = E_{(x,\theta)}[L(\theta, d(x))] \\
&\geq E_x E_{\theta|x}\left[L(\theta, d_\Lambda(x))|x\right] = E_{(x,\theta)}\left[L(\theta, d_\Lambda(x))\right] \\
&= \mathcal{R}(\Lambda, d_\Lambda).
\end{aligned}$$

Corollary 1.1

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

$$\begin{aligned} d_\Lambda(x) &= \frac{E[k(\theta)\theta|X = x]}{E[k(\theta)|X = x]} \\ &= \frac{\int \theta\, k(\theta)\, d\Lambda(\theta|x)}{\int k(\theta)\, d\Lambda(\theta|x)} \\ &= \frac{\int \theta\, k(\theta)\lambda(\theta|x)\, d\theta}{\int k(\theta)\lambda(\theta|x)\, d\theta} \quad \text{(Lebesgue Measure).} \end{aligned}$$

When $k(\theta) = 1$, then

$$\begin{aligned} d_\Lambda(x) &= \int \theta\, d\Lambda(\theta|x) = E[\theta|X = x] \\ &= \text{posterior mean .} \end{aligned}$$

Proof: For an arbitrary (nonrandomized) rule $d \in \mathcal{D}$,

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

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

Note: when $\boldsymbol{\theta} \in \Re^s$, $|\boldsymbol{\theta} - \boldsymbol{a}|^2 = (\boldsymbol{\theta} - \boldsymbol{a})'(\boldsymbol{\theta} - \boldsymbol{a})$.

Corollary 1.2

If $\Theta = \mathcal{A}$ and $L(\boldsymbol{\theta}, \boldsymbol{a}) = |\boldsymbol{\theta} - \boldsymbol{a}|$, then $d_\Lambda(x) =$ any median of $\Lambda(\boldsymbol{\theta}|x)$.

Corollary 1.3

If $A = \{0, 1\}$, $\Theta = \Theta_0 \cup \Theta_1$, $L(\theta, a_i) = l_i 1_{\Theta_i^c}(\theta)$, $i = 0, 1$, then any rule of the form

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

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

$$
\begin{aligned}
& E[L(\theta, \phi(x))|X = x] \\
&= \int_{\boldsymbol{\Theta}} L(\theta, \phi(x))\, d\Lambda(\theta|x) \\
&= \int_{\boldsymbol{\Theta}} \{l_1 \phi(x) 1_{\boldsymbol{\Theta}_0}(\theta) + l_0(1 - \phi(x)) 1_{\boldsymbol{\Theta}_1}(\theta)\}\, d\Lambda(\theta|x) \\
&= l_1 \phi(x) p(\theta \in \boldsymbol{\Theta}_0 | X = x) + l_0(1 - \phi(x)) p(\theta \in \boldsymbol{\Theta}_1 | X = x) \\
&= l_0 p(\theta \in \boldsymbol{\Theta}_1 | X = x) + \phi(x)\{l_1 p(\theta \in \boldsymbol{\Theta}_0 | X = x) - l_0 p(\theta \in \boldsymbol{\Theta}_1 | X = x)\},
\end{aligned}
$$

which is minimized by any rule of the form d_Λ.

Corollary 1.4

If $\boldsymbol{\Theta} = \Re$, $\boldsymbol{\mathcal{A}} = \{0, 1\}$, $\boldsymbol{\Theta}_0 = (-\infty, \theta_0]$, $\boldsymbol{\Theta}_1 = (\theta_0, \infty)$ and $L(\theta, 0) = (\theta - \theta_0) 1_{(\theta_0, \infty)}(\theta)$, $L(\theta, 1) = (\theta_0 - \theta) 1_{(-\infty, \theta_0]}(\theta)$;

then

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

is a Bayes rule with respect to Λ.

Proof:
Again, it suffices to minimize

$$\begin{aligned} & E[L(\theta, \phi(x)|X=x)] \\ &= \int_{\Theta} \{\phi(x)(\theta_0 - \theta)1_{(-\infty,\theta_0]}(\theta) + (1-\phi(x))(\theta - \theta_0)1_{(\theta_0,\infty)}(\theta)\}\, d\Lambda(\theta|x) \\ &= \int_{\Theta} (\theta - \theta_0)1_{(-\infty,\theta_0]}(\theta) d\Lambda(\theta|x) + (1-\phi(x))[E(\theta|X=x) - \theta_0]\ , \end{aligned}$$

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

Example 1.18

Suppose $X \sim \text{Binomial}(n, \theta)$, $\theta \sim \text{Beta}(\alpha, \beta)$, and thus $\theta | X = x \sim \text{Beta}(\alpha + x, \beta + n - x)$. Let $L(\theta, a) = (\theta - a)^2$, so that

$$\begin{aligned} d_\Lambda(x) = E[\theta | X = x] &= \frac{\alpha + x}{\alpha + \beta + n} \\ &= w_n \left(\frac{\alpha}{\alpha + \beta} \right) + (1 - w_n) \frac{x}{n} \end{aligned}$$

where $w_n = \frac{\alpha+\beta}{\alpha+\beta+n}$, and $E(\theta) = \frac{\alpha}{\alpha+\beta}$. Note that $w_n \to 0$ as $n \to \infty$. If the loss is $L(\theta, a) = \frac{(\theta-a)^2}{\theta(1-\theta)}$, then $k(\theta) = \theta^{-1}(1-\theta)^{-1}$, and

$$E[\theta k(\theta) | X = x] = \frac{B(\alpha + x, \beta + n - x - 1)}{B(\alpha + x, \beta + n - x)},$$

$$E[k(\theta) | X = x] = \frac{B(\alpha + x - 1, \beta + n - x - 1)}{B(\alpha + x, \beta + n - x)},$$

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

$$\begin{aligned} d_\Lambda(x) &= \frac{B(\alpha+x, \beta+n-x-1)}{B(\alpha+x-1, \beta+n-x-1)} = \frac{\alpha+x-1}{\alpha+\beta+n-2} \\ &= w_n \left(\frac{\alpha-1}{\alpha+\beta-2} \right) + (1-w_n) \left(\frac{x}{n} \right) \end{aligned}$$

where $w_n = \frac{\alpha+\beta-2}{\alpha+\beta+n-2}$, $w_n \to 0$ as $n \to \infty$. Note that when $\alpha = \beta = 1$, the Bayes estimator for this loss function becomes the MLE, $\hat{\theta} = \frac{X}{n}$.

Example 1.19

If $\boldsymbol{X}|\boldsymbol{\theta} \sim \text{Multinomial}_k(n, \boldsymbol{\theta})$, $\boldsymbol{\theta} \sim \text{Dirichlet}_{k-1}(\boldsymbol{\alpha})$, then

$$E(\boldsymbol{\theta}) = \frac{\boldsymbol{\alpha}}{\sum_{i=1}^{k} \alpha_i} = \left(\frac{\alpha_1}{\sum_{i=1}^{k} \alpha_i}, \cdots, \frac{\alpha_k}{\sum_{i=1}^{k} \alpha_i} \right) .$$

With $L(\boldsymbol{\theta}, \boldsymbol{a}) = |\boldsymbol{\theta} - \boldsymbol{a}|^2 \equiv (\boldsymbol{\theta} - \boldsymbol{a})'(\boldsymbol{\theta} - \boldsymbol{a})$, we have

$$d_\Lambda(\boldsymbol{x}) = E(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\boldsymbol{\alpha} + \boldsymbol{x}}{n + \sum_{i=1}^{k} \alpha_i} ,$$

where $\boldsymbol{x} = (x_1, \ldots, x_k)$.

Example 1.20

If $X|\theta \sim N(\theta, \sigma^2)$, $\theta \sim N(\mu, \tau^2)$, σ^2 is known, then $\theta|x \sim N\left(w\mu + (1-w)x, \frac{1}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}}\right)$, $w = \frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}}$, and with $L(\theta, a) = (\theta - a)^2$,

$$d_\Lambda(x) = E(\theta|x) = w\mu + (1-w)x .$$

This result remains true if $L(\theta, a) = \rho(\theta - a)$ where $\rho(\cdot)$ is a convex and even function.

Finding Minimax Rules

Definition 1.16

A prior Λ_0 for which $\mathcal{R}(\Lambda, d_\Lambda)$ is maximized is called a least favorable prior:

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

Theorem 1.12

Suppose that Λ is a prior distribution on $\boldsymbol{\Theta}$ on such that

$$\begin{aligned} \mathcal{R}(\Lambda, d_\Lambda) &= \int_{\boldsymbol{\Theta}} R(\theta, d_\Lambda)\, d\Lambda(\theta) \\ &= \sup_{\theta} R(\theta, d_\Lambda). \end{aligned} \tag{1.8}$$

Then

(i) d_Λ is minimax.

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

(iii) Λ is least favorable.

Proof:

(i) let d be another rule. Then

$$\begin{aligned}
\sup_{\theta\in\Theta} R(\theta,d) &\geq \int_{\Theta} R(\theta,d)\, d\Lambda(\theta) \\
&\geq \int_{\Theta} R(\theta,d_\Lambda)\, d\Lambda(\theta) \quad \text{(since } d_\Lambda \text{ is Bayes w.r.t. } \Lambda\text{)} \\
&= \sup_{\theta\in\Theta} R(\theta,d_\Lambda) \quad \text{by (1.8).}
\end{aligned}$$

Hence d_Λ is minimax.

Note here that

$$\begin{aligned}\int_{\Theta} R(\theta,d)\, d\Lambda(\theta) &\leq \int_{\Theta} \sup_{\theta\in\Theta} R(\theta,d)\, d\Lambda(\theta)\\ &= \sup_{\theta\in\Theta} R(\theta,d) \int_{\Theta} d\Lambda(\theta)\\ &= \sup_{\theta\in\Theta} R(\theta,d).\end{aligned}$$

(ii) If d_Λ is unique Bayes, then strict inequality holds in **(i)**, so d_Λ is unique minimax.

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

$$\begin{aligned}
\mathcal{R}(\Lambda^*, d_{\Lambda^*}) &= \int_{\Theta} R(\theta, d_{\Lambda^*})\, d\Lambda^*(\theta) \\
&\leq \int_{\Theta} R(\theta, d_{\Lambda})\, d\Lambda^*(\theta) \quad \text{(since } d^*_{\Lambda} \text{ is Bayes w.r.t. } \Lambda^*\text{)} \\
&\leq \sup_{\theta} R(\theta, d_{\Lambda}) \\
&= \mathcal{R}(\Lambda, d_{\Lambda}) \equiv r_{\Lambda} \quad \text{by (1.8)} .
\end{aligned}$$

Corollary 1.5

If d_Λ is Bayes w.r.t. Λ and has constant risk, $R(\theta, d_\Lambda) = \text{constant}$, then d_Λ is minimax.

Proof: If d_Λ has constant risk, then (1.8) holds.

Corollary 1.6

Let $\boldsymbol{\Theta}_\Lambda = \{\theta \in \boldsymbol{\Theta} : R(\theta, d_\Lambda) = \sup_{\theta'} R(\theta', d_\Lambda)\}$ be the set of θ's where the risk of d_Λ attains its maximum. Then d_Λ is minimax if $\Lambda(\boldsymbol{\Theta}_\Lambda) = 1$.

Equivalently, d_Λ is minimax if there is a set $\boldsymbol{\Theta}_\Lambda$ with $\Lambda(\boldsymbol{\Theta}_\Lambda) = 1$ and $R(\theta, d_\Lambda) = \sup_{\theta'} R(\theta', d_\Lambda)$ for all $\theta \in \boldsymbol{\Theta}_\Lambda$.

Example 1.21

Suppose $X \sim \text{Binomial}(n, \theta)$, $\theta \sim \text{Beta}(\alpha, \beta)$, $L(\theta, a) = (\theta - a)^2$.
$\theta | X = x \sim \text{Beta}(\alpha + x, \beta + n - x)$, and $d_\Lambda(x) = w_n \left(\frac{\alpha}{\alpha+\beta}\right) + (1 - w_n)\left(\frac{X}{n}\right)$,
$w_n = \frac{\alpha+\beta}{\alpha+\beta+n}$.
Consequently,

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

if $2\alpha(\alpha+\beta) = n$ and $(\alpha+\beta)^2 = n$.

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

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

is Bayes with respect to $\Lambda = \text{Beta}(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2})$ and has risk $R(\theta, d_M) = \frac{1}{4(1+\sqrt{n})^2}$. Hence the Beta$(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2})$ distribution is least favorable and d_M is minimax. Note that d_M is a consistent estimator of θ, that is,

$$d_M(x) \xrightarrow{p} \theta \quad \text{as } n \to \infty \,,$$

but its bias is of the order $n^{-1/2}$. The bias equals $\frac{(\frac{1}{2}-\theta)}{1+\sqrt{n}}$.

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 proper prior distributions, let d_k denote the Bayes estimator corresponding to Λ_k, and set $\mathcal{R}_k \equiv \int_{\boldsymbol{\Theta}} R(\theta, d_k)\, d\Lambda_k(\theta)$.
Suppose that $\mathcal{R}_k \to r < \infty$ as $k \to \infty$.

Definition 1.17

The sequence of prior distributions $\{\Lambda_k\}$ with Bayes risks $\{\mathcal{R}_k\}$ is said to be least favorable if $\mathcal{R}_\Lambda \equiv \mathcal{R}(\Lambda, d_\Lambda) \leq r$ for any prior distribution Λ.

Theorem 1.13

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

(i) $\mathcal{R}_k \to r$

(ii) d is an estimator for which $\sup_{\theta \in \Theta} R(\theta, d) = r$.

Then

A. d is minimax, and

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

Proof:

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

$$\sup_{\theta} R(\theta, d^*) \geq \int R(\theta, d^*)\, d\Lambda_k(\theta) \geq \mathcal{R}_k$$

for all $k \geq 1$. Hence

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

so d is minimax.

B. If Λ is any prior distribution, then

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

by (ii).

Example 1.22

Suppose $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$ (given θ), and suppose $\theta \sim N(\mu, \tau^2)$, σ^2 is known. Then under squared error loss,

$$d_\Lambda(\boldsymbol{x}) = w_n \mu + (1 - w_n)\bar{X}\ ,$$

$\boldsymbol{x} = (x_1, \ldots, x_n)$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $w_n = \frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}$, and

$$\begin{aligned}
\mathcal{R}(\Lambda, d_\Lambda) \equiv r_\Lambda &= E(\theta - d_\Lambda)^2 \\
&= E_x E_{\Lambda|x}[(\theta - d_\Lambda(x))^2 | X = x] \\
&= E_x[\mathrm{Var}(\theta|x)] \\
&= \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \\
&\to \frac{\sigma^2}{n} \quad \text{as } \tau^2 \to \infty \\
&= R(\theta, \bar{X}) \quad \text{for all } \theta \in \boldsymbol{\Theta}.
\end{aligned}$$

Hence by Theorem 1.13 , $\bar{X}$ is a minimax estimator of θ.

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

Lemma 1.3

Suppose that $x \sim P \in \boldsymbol{M} \equiv \{$ all P's on $\boldsymbol{\mathcal{X}}\}$, and that $\nu : P \subset \boldsymbol{M} \to \mathfrak{R}^1$ is a functional (e.g., $\nu(P) = E_P(X) = \int x\, dP(x)$, $\nu(P) = \mathrm{Var}_P(X)$, $\nu(P) = P_f$ for some fixed function f, or if $\boldsymbol{\mathcal{X}} = \mathfrak{R}^1$, $\nu(P) = F_P^{-1}(t)$ for fixed $t \in (0,1)$, where $F_P(x) = P(X \leq x)$.) Suppose that d is a minimax estimator of $\nu(P)$ for $P \in P_0 \subset P_1$. If

$$\sup_{P \in P_0} R(P,d) = \sup_{P \in P_1} R(P,d)\ ,$$

then d is minimax for estimating $\nu(P)$, $P \in P_1$.

Proof:

Suppose that d is not minimax in P_1 (but minimax in P_0). Then there exists a rule d^* with smaller maximum risk:

$$\begin{aligned} \sup_{P\in P_0} R(P,d^*) &\leq \sup_{P\in P_1} R(P,d^*) \\ &< \sup_{P\in P_1} R(P,d) = \sup_{P\in P_0} R(P,d) \ , \end{aligned}$$

which contradicts the hypothesis that d is minimax for P_0.

Example 1.23

Suppose $X_1,\ldots,X_n$ are iid $N(\theta,\sigma^2)$ and both (θ,σ^2) are unknown. Thus $\boldsymbol{\Theta} = \{(\theta,\sigma^2) : \theta \in \mathfrak{R}, 0 < \sigma^2 < \infty\}$. So to get a reasonable minimax estimator, we need to restrict σ^2. Let

$$P_0 = \{N(\theta,M) : \theta \in \mathfrak{R}\}$$

$$P_1 = \{N(\theta,\sigma^2) : \theta \in \mathfrak{R},\ 0 \leq \sigma^2 \leq M\} \ .$$

Then $\bar{X}$ is minimax for P_0 by Example 1.22 , $P_0 \subset P_1$ and

$$\sup_{P_1} R(P, \bar{X}) = \sup_{P_0} R(P, \bar{X}) = \frac{M}{n}.$$

Thus $\bar{X}$ is a minimax estimator of θ for P_1.

Example 1.24

Let $X_1, \ldots, X_n$ be iid $P \in P_\mu$, where
$P_\mu = \{\text{all probability measures } P : E_P|X| < \infty\}$ and consider estimation of $\nu(P) = E_P(X)$ with squared error loss for the families

(i) $P_{b\sigma^2} \equiv \{P \in P_\mu : \text{Var}_p(X) \leq M < \infty\}$

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

Then

A) $\bar{X}$ is minimax for $P_{b\sigma^2} = P_1$ by the lemma since it is minimax for $P_0 = \{N(\theta, \sigma^2) : \theta \in \Re, 0 < \sigma^2 \leq M\}$ and

$$\sup_{P \in P_0} R(P, \bar{X}) = \sup_{P \in P_1} R(P, \bar{X}) \ .$$

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

$$P_1 = \{P : P([0,1]) = 1\}$$

$$P_0 = \{P \in P_1 : P(X = 1) = \theta, \ P(X = 0) = 1 - \theta \quad \text{for some } 0 < \theta < 1\}.$$

For P_0, we know that the minimax estimator is

$$d_M(\mathbf{x}) = \left(\frac{1}{1 + \sqrt{n}}\right)\left(\frac{1}{2}\right) + \left(\frac{\sqrt{n}}{1 + \sqrt{n}}\right)\bar{X} \ .$$

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

$$\begin{aligned} R(P, d_M) &= \frac{1}{(1+\sqrt{n})^2}\left\{\mathrm{Var}_P(X) + \left(\frac{1}{2} - \theta\right)^2\right\} \\ &\leq \frac{1}{(1+\sqrt{n})^2}\{\theta - \theta^2 + \frac{1}{4} - \theta + \theta^2\} \end{aligned}$$

(since $0 \leq X \leq 1$ implies $E_P(X^2) \leq E_P(X) \equiv \theta$).
Thus

$$\sup_{P \in P_1} R(P, d_M) = \sup_{P \in P_0} R(P, d_M),$$

and by the lemma, d_M is minimax.

Theorem 1.14

Any unique Bayes estimator with finite Bayes risk 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

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

which contradicts uniqueness of d_Λ. Thus d_Λ is admissible.

Example 1.25

Consider the Bayes estimator of a normal mean $d_\Lambda = w_n\mu + (1-w_n)\bar{X}$, $w_n = \frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}$. The Bayes risk is finite. Hence d_Λ is unique Bayes and admissible. (The Bayes risk is $\mathcal{R}(\Lambda, d_\Lambda) = w_n^2\tau^2 + (1-w_n)^2\frac{\sigma^2}{n}$.)

Example 1.25 (continued)

For $d_\Lambda = w_n\mu + (1-w_n)\bar{X}$, the Bayes risk is

$$\mathcal{R}(\Lambda, d_\Lambda) = \int R(\theta, d)\, d\Lambda(\theta).$$

$$
\begin{aligned}
R(\theta, d) &= E_\theta[L(\theta, a)] \\
&= E_\theta[w_n\mu + (1-w_n)\bar{X} - \theta]^2 \\
&= E_\theta[(w_n\mu + (1-w_n)\bar{X})^2 - 2\theta(w_n\mu + (1-w_n)\bar{X}) + \theta^2] \\
&= E_\theta[w_n^2\mu^2 + 2w_n(1-w_n)\mu\bar{X} + (1-w_n)^2\bar{X}^2 - 2\theta w_n\mu \\
&\quad -2\theta(1-w_n)\bar{X} + \theta^2] \\
&= w_n^2\mu^2 + 2w_n(1-w_n)\mu\theta + (1-w_n)^2\left(\frac{\sigma^2}{n} + \theta^2\right) - 2\theta w_n\mu \\
&\quad -2\theta^2(1-w_n) + \theta^2 \\
&= w_n^2\mu^2 + 2w_n\mu\theta - 2w_n^2\mu\theta + (w_n^2 - 2w_n + 1)\theta^2 \\
&\quad +(w_n^2 - 2w_n^2 + 1)\frac{\sigma}{n} - 2\theta w_n\mu - 2\theta^2(1-w_n) + \theta^2 \\
&= w_n^2\mu^2 - 2w_n^2\mu\theta + \theta^2(w_n^2 - 2w_n + 1 - 2 + 2w_n + 1) \\
&\quad +(w_n^2 - 2w_n + 1)\left(\frac{\sigma^2}{n}\right) \\
&= w_n^2\mu^2 - 2w_n^2\mu\theta + \theta^2 w_n^2 + (1-w_n)^2\left(\frac{\sigma^2}{n}\right) \\
&= w_n^2(\theta-\mu)^2 + (1-w_n)^2\left(\frac{\sigma^2}{n}\right).
\end{aligned}
$$

Hence

$$\begin{aligned}
\mathcal{R}(\Lambda, d_\Lambda) &= E_\Lambda \left[w_n^2 (\theta - \mu)^2 + (1 - w_n)^2 \left(\frac{\sigma^2}{n} \right) \right] \\
&= w_n^2 E_\Lambda (\theta - \mu)^2 + (1 - w_n)^2 \left(\frac{\sigma^2}{n} \right) \\
&= w_n^2 \mathrm{Var}(\theta) + (1 - w_n)^2 \left(\frac{\sigma^2}{n} \right) \\
&= w_n^2 \tau^2 + (1 - w_n)^2 \left(\frac{\sigma^2}{n} \right) = \text{Bayes risk.}
\end{aligned}$$

(recall that $\theta \sim N(\mu, \tau^2)$).

Theorem 1.15

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

$$\begin{aligned} R(\theta, aX + b) &= a^2\sigma^2 + \{(a-1)\theta + b\}^2 \\ &= \rho(a, b). \end{aligned}$$

(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

$$\begin{aligned}
\rho(a,b) &\geq \{(a-1)\theta + b\}^2 \\
&= (a-1)^2\left\{\theta + \frac{b}{a-1}\right\}^2 \\
&> \left\{\theta + \frac{b}{a-1}\right\}^2 = \rho\left(0, \frac{-b}{a-1}\right).
\end{aligned}$$

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

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

so that $aX + b$ is dominated by X.

If we have a random sample $X_1, \ldots, X_n$, $a\bar{X} + b$ is inadmissible (under squared error loss) 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 1.16

If $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$, $\theta \in \mathbf{\Theta}$ with σ^2 known, then under squared error loss, $\bar{X}$ is an admissible estimator of θ.

Proof: (Limiting Bayes Method)

Suppose that $\bar{X}$ is inadmissible (and $\sigma^2 = 1$). Then there is an estimator d^* such that $R(\theta, d^*) \leq \frac{1}{n} = R(\theta, \bar{X})$ for all θ with risk strictly less than for some θ. Now $R(\theta, d) = E[\theta - d(X)]^2$ is continuous in θ for every d, and hence there exists an $\epsilon > 0$ and $\theta_0 < \theta_1$, so that $R(\theta, d^*) < \frac{1}{n} - \epsilon$ 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

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

so $r_\tau \leq r_{\tau^*}$. Thus

$$\begin{aligned}
\frac{\frac{1}{n} - r_{\tau^*}}{\frac{1}{n} - r_\tau} &= \frac{\frac{1}{\sqrt{2\pi}\tau}\int_{-\infty}^{\infty}\{\frac{1}{n} - R(\theta, d^*)\}\exp\{\frac{-\theta^2}{2\tau^2}\}\, d\theta}{(\frac{1}{n} - \frac{\tau^2}{1+n\tau^2})} \\
&\geq \frac{\frac{1}{\sqrt{2\pi}\tau}\epsilon\int_{\theta_0}^{\theta_1}\exp\{\frac{-\theta^2}{2\tau^2}\}\, d\theta}{\frac{1}{n(1+n\tau^2)}} \\
&= \frac{n(1+n\tau^2)}{\sqrt{2\pi}\tau}\epsilon\int_{\theta_0}^{\theta_1}\exp\left\{\frac{-\theta^2}{2\tau^2}\right\}\, d\theta \\
&\to \infty(\theta_1 - \theta_0) = \infty \quad \text{as } \tau \to \infty.
\end{aligned}$$

Hence

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

for $\tau >$ some τ_0, which contradicts d_Λ that is Bayes w.r.t. Λ (with Bayes risk r_τ). Hence $\bar{X}$ is admissible.

Now let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be iid $N_k(\boldsymbol{\theta}, \boldsymbol{I})$, and consider estimation of $\boldsymbol{\theta} \in \Re^k$ with loss function $L(\boldsymbol{\theta}, \boldsymbol{a}) = |\boldsymbol{\theta} - \boldsymbol{a}|^2 = (\boldsymbol{\theta} - \boldsymbol{a})'(\boldsymbol{\theta} - \boldsymbol{a})$. Let $\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i$.

Note that

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

and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})$.

Theorem 1.17

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

Proof:

Let $g : \Re^k \to \Re^k$ have $E|\frac{\partial}{\partial x_i} g_i(x)| < \infty$ and consider estimators of the form $\hat{\boldsymbol{\theta}}_n = \bar{\boldsymbol{X}} + n^{-1} g(\bar{\boldsymbol{X}})$.

Now

$$\begin{aligned} & E_{\boldsymbol{\theta}}|\bar{\boldsymbol{X}} - \boldsymbol{\theta}|^2 - E_{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}|^2 \\ &= E_{\boldsymbol{\theta}}|\bar{\boldsymbol{X}} - \boldsymbol{\theta}|^2 - E_{\boldsymbol{\theta}}|\bar{\boldsymbol{X}} - \boldsymbol{\theta} + n^{-1}g(\bar{\boldsymbol{X}})|^2 \\ &= -2n^{-1}E_{\boldsymbol{\theta}}\{(\bar{\boldsymbol{X}} - \boldsymbol{\theta})'g(\bar{\boldsymbol{X}})\} - n^{-2}E_{\boldsymbol{\theta}}|g(\bar{\boldsymbol{X}})|^2 \ . \end{aligned} \tag{1.9}$$

To proceed further, we need an identity due to Stein.

Lemma 1.4

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 $\boldsymbol{X} \sim N_k(\boldsymbol{\theta}, \sigma^2\mathbf{I})$, and $g : \mathfrak{R}^k \to \mathfrak{R}^k$ satisfies

$$E\left|\frac{\partial}{\partial x_i}g_i(\boldsymbol{x})\right| < \infty,$$

then

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

Proof:

With out loss of generality, suppose that $\theta = 0$ and $\sigma^2 = 1$. Integration by parts gives

$$\begin{aligned} E[g'(X)] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g'(x) \exp\left(\frac{-x^2}{2}\right) \, dx \\ &= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x) \exp\left(\frac{-x^2}{2}\right) \left(\frac{-2x}{2}\right) \, dx \\ &= E\left[Xg(X)\right] \, . \end{aligned}$$

In applying integration by parts here, we have ignored the term $uv|_{-\infty}^{\infty} = g(x)\phi(x)|_{-\infty}^{\infty}$. To prove that this term does in fact vanish, it is easiest to apply Fubini's theorem (twice) in a slightly modified 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

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

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

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

$$\begin{aligned}
\sigma^2 E\left[\frac{\partial}{\partial X_i} g_i(\boldsymbol{X})\right] &= \sigma^2 E\left[E\left\{\frac{\partial}{\partial X_i} g_i(\boldsymbol{X}) | \boldsymbol{X}^{(i)}\right\}\right] \\
&= E[E\{(X_i - \theta_i) g_i(\boldsymbol{X}) | \boldsymbol{X}^{(i)}\}] \quad \text{by (i)} \\
&= E[(X_i - \theta_i) g_i(\boldsymbol{X})] \,.
\end{aligned}$$

Now we return to the proof of the theorem. Using (ii) of Lemma 1.4 in (1.9) shows that the right side of (1.9) equals

$$-2n^{-2} E_\theta \left[\sum_{i=1}^{k} \frac{\partial g_i}{\partial X_i}(\bar{\boldsymbol{X}})\right] - n^{-2} E_\theta |g(\bar{\boldsymbol{X}})|^2. \tag{1.10}$$

Now let $\psi : \mathfrak{R}^k \to \mathfrak{R}$ be twice differentiable and set

$$\begin{aligned} g(\boldsymbol{x}) &= \nabla\{\log \psi(\boldsymbol{x})\} \\ &= \frac{1}{\psi(\boldsymbol{x})}\left(\frac{\partial \psi(\boldsymbol{x})}{\partial x_1}, \cdots, \frac{\delta \psi(\boldsymbol{x})}{\delta x_k}\right) . \end{aligned}$$

Thus

$$\begin{aligned} \frac{\partial g_i(\boldsymbol{x})}{\partial x_i} &= \frac{1}{\psi(\boldsymbol{x})}\frac{\partial^2}{\partial x_i^2}\psi(\boldsymbol{x}) - \frac{1}{\psi^2(\boldsymbol{x})}\left(\frac{\partial \psi(\boldsymbol{x})}{\partial x_i}\right)^2 \\ &= \frac{1}{\psi(\boldsymbol{x})}\frac{\partial^2 \psi(\boldsymbol{x})}{\partial x_i^2} - (g_i(\boldsymbol{x}))^2 \end{aligned}$$

and

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

where $\nabla^2 \psi(\boldsymbol{x}) = \sum_{i=1}^{k} \frac{\partial^2 \psi(\boldsymbol{x})}{\partial x_i^2}$. Hence the right side of (1.10) is

$$n^{-2} E_\theta |g(\bar{\boldsymbol{X}})|^2 - 2n^{-2} E_\theta \left\{ \frac{1}{\psi(\boldsymbol{X})} \nabla^2 \psi(\bar{\boldsymbol{X}}) \right\},$$

which is greater than 0 if $\psi(\boldsymbol{x}) \geq 0$ and $\nabla^2 \psi(\boldsymbol{x}) \leq 0$, $g \neq 0$ (i.e., ψ is superharmonic). An example of a superharmonic function is

$$\psi(\boldsymbol{x}) = |\boldsymbol{x}|^{-(k-2)} = \{x_1^2 + \cdots + x_k^2\}^{-\frac{(k-2)}{2}} .$$

Then

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

and $\nabla^2\psi(\boldsymbol{x}) = 0$, and so $\psi(\boldsymbol{x})$ is harmonic. Thus

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

and

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

since $|\boldsymbol{X}|^2 \sim \chi^2_k$ with $E\left(\frac{1}{\chi^2_k}\right) = \frac{1}{k-2}$. Hence

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

For general $\boldsymbol{\theta}$,

$$\begin{aligned} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n) &= \frac{k}{n} - \frac{(k-2)^2}{n} E_0\left(\frac{1}{|\boldsymbol{X} + \sqrt{n}\boldsymbol{\theta}|^2}\right) \\ &= \frac{k}{n}\left(1 - \left(\frac{k-2}{k}\right) E_0\left(\frac{k-2}{\chi^2_k(\delta)}\right)\right) \end{aligned}$$

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

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

Remark 1.3

Another interesting function ψ is

$$\psi(\boldsymbol{x}) = \begin{cases} |\boldsymbol{x}|^{-(k-2)}, & |\boldsymbol{x}| \geq \sqrt{k-2} \\ (k-2)^{-\frac{(k-2)}{2}} \exp\left\{\frac{1}{2}[(k-2) - |\boldsymbol{x}|^2]\right\}, & |\boldsymbol{x}| < \sqrt{k-2} \end{cases}$$

For this ψ,

$$g(\boldsymbol{x}) = \nabla \log \psi(\boldsymbol{x}) = \begin{cases} -\frac{(k-2)}{|\boldsymbol{x}|^2}, & |\boldsymbol{x}| \leq \sqrt{k-2} \\ -\boldsymbol{x}, & |\boldsymbol{x}| \geq \sqrt{k-2} \end{cases},$$

Further Results and Examples in Decision Theory

Example 1.3 (continued) Suppose $X \sim N(\theta, 1)$, $L(\theta, a) = (\theta - a)^2$, and we consider the class of estimators $d_c(x) = cX$. Suppose $\theta \sim N(0, \tau^2)$, τ^2 known. Then the Bayes risk for $d_c(x)$ is

$$\begin{aligned}
\mathcal{R}(\Lambda, d) &= \int R(\theta, d)\, d\Lambda(\theta) \\
&= \int [c^2 + (1-c)^2\theta^2]\, d\Lambda(\theta) \\
&= c^2 + (1-c)^2 \int \theta^2\, d\Lambda(\theta) \\
&= c^2 + (1-c)^2 E[\theta^2] \\
&= c^2 + (1-c)^2\tau^2\ . \qquad (1.11)
\end{aligned}$$

Now let us examine values of c where $\mathcal{R}(\Lambda, d)$ is minimized. Differentiating (1.14) with respect to c and setting equal to zero shows that $c_0 = \frac{\tau^2}{1+\tau^2}$ is the unique minimizer and thus $d_{c_0}(x) = \left(\frac{\tau^2}{1+\tau^2}\right) X$ is the unique Bayes rule. Since $d_{c_0}(x)$ is the unique Bayes rule, it is admissible by Theorem (1.14). This example also shows that all decision rules of the form $d_c(x) = cX$, $0 < c < 1$ are admissible, since τ is arbitrary in $d_{c_0}(x)$. $d_1(x) = X$ is admissible by Theorem 1.16. The Bayes risk for $d_{c_0}(x)$ obtained by substituting $c_0 = \frac{\tau^2}{1+\tau^2}$ into (1.16), yielding

$$\begin{aligned} \mathcal{R}(\Lambda, d_{c_0}) &= \left(\frac{\tau^2}{1+\tau^2}\right)^2 + \left(1 - \frac{\tau^2}{1+\tau^2}\right)^2 \tau^2 \\ &= \frac{\tau^2}{1+\tau^2}. \end{aligned}$$

Now we ask the question: Is d_{c_0} minimax? Theorem 1.12 says that d_{c_0} is minimax if

$$\mathcal{R}(\Lambda, d_\Lambda) = \sup_\theta R(\theta, d_\Lambda) \ .$$

Now

$$\begin{aligned} R(\theta, d_\Lambda) &= R(\theta, d_{c_0}) = c_0 + (1-c_0)^2\theta^2 \\ &= \left(\frac{\tau^2}{1+\tau^2}\right)^2 + \left(\frac{1}{1+\tau^2}\right)^2 \theta^2 \ . \end{aligned}$$

clearly $\sup_\theta R(\theta, d_{c_0}) = \infty \neq \frac{\tau^2}{1+\tau^2} = \mathcal{R}(\Lambda, d_{c_0})$ so that Theorem 1.12 cannot be used here. We try and use Theorem 1.13.

To do this, notice that $\{\Lambda_k\} = N(0,k)$, $k = \tau^2$. We have $r_k = \mathcal{R}(\Lambda, d_{\Lambda_k}) = \frac{k}{k+1}$, $\lim_{k\to\infty} r_k = 1 \equiv r$, and note that $d_1(x) = X$ has risk $R(\theta, d_1(x)) = 1$, and so $\sup_\theta R(\theta, d_1(x)) = 1 = r$ and thus $d_1(x) = X$ is minimax, and $\{\Lambda_k\} = N(0,k)$ is least favorable.

An alternative form of Theorem 1.13 is

Theorem 1.13* Assume that $\{\Lambda_k\}$ is a sequence of proper priors and d is a decision rule such that

$$R(\theta, d) \leq \lim_{k\to\infty} \mathcal{R}(\Lambda_k, d_{\Lambda_k}) < \infty, \quad \text{for all } \theta$$

where d_{Λ_k} is the Bayes rule corresponding to $\{\Lambda_k\}$. Then d is a minimax rule.

Theorem 1.13 (and Theorem 1.13*) suggests that a good way to come up with a minimax rule is to guess a least favorable prior distribution, whose limit is an improper prior, i.e. $\lim_{k\to\infty} \Lambda_k = \Lambda_0$ where Λ_0 is improper, and then derive the generalized Bayes rule for Λ_0. Such a generalized Bayes rule will often be minimax via Theorem 1.13. In our example, we have

$$\{\Lambda_k\} = N(0,k) \quad \text{(sequence of proper priors)}$$

$$\lim_{k\to\infty} \Lambda_k = \Lambda_0 \ ,$$

where Λ_0 is the (improper) uniform prior on $\Re$. The generalized Bayes rule under quadratic loss is $d_{GB} = X$, which is the minimax rule by Theorem 1.13.

Example 1.26

Suppose $X \sim N(\theta, \sigma^2)$, σ^2 is known and that it is desired to estimate θ under squared error loss. Suppose $\theta \sim N(\mu, \tau^2)$, and we focus on linear estimators, of the form $d(x) = aX + b\mu$.

The Bayes risk of the rule $d(x)$ is

$$\begin{aligned}
\mathcal{R}(\Lambda, d) &= E_\Lambda E_{X|\theta}(aX + b\mu - \theta)^2 \\
&= E_\Lambda E_{X|\theta}[(a(X-\theta) + (a-1)(\theta-\mu) + (b+a-1)\mu)^2] \\
&= E_\Lambda[a^2\sigma^2 + \{(a-1)(\theta-\mu) + (b+a-1)\mu\}^2] \\
&= a^2\sigma^2 + (a-1)^2\tau^2 + (b+a-1)^2\mu^2 \ .
\end{aligned} \tag{1.12}$$

Now to obtain the Bayes rule for this class of estimators, we minimize $\mathcal{R}$ with respect to (a, b). Taking derivatives of (1.12) with respect to (a, b) and setting equal to 0, we get

$$b = 1 - a \ , \ a = \frac{\tau^2}{\sigma^2 + \tau^2} \ .$$

and thus, the unique Bayes rule (estimator) is

$$\begin{aligned} d_\Lambda &= \left(\frac{\tau^2}{\sigma^2+\tau^2}\right)X + \left(\frac{\sigma^2}{\sigma^2+\tau^2}\right)\mu \\ &= \left(\frac{1/\sigma^2}{1/\tau^2+1/\sigma^2}\right)X + \left(\frac{1/\tau^2}{1/\tau^2+1/\sigma^2}\right)\mu \\ &= (1-w)X + w\mu, \quad w = \frac{1/\tau^2}{1/\tau^2+1/\sigma^2} \end{aligned}$$

and the Bayes risk of d_Λ is thus given by

$$\begin{aligned} \mathcal{R}(\Lambda, d_\Lambda) &= \left(\frac{\tau^2}{\sigma^2+\tau^2}\right)^2 \sigma^2 + \left(\frac{\tau^2}{\sigma^2+\tau^2} - 1\right)^2 \tau^2 \\ &= \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}. \end{aligned}$$

More on the admissibility of Bayes rules

Theorem 1.18

Assume that the risk functions $R(\theta, d)$ are continuous in θ for all decision rules d. Assume also that the prior Λ gives positive probability to any open subset of $\boldsymbol{\Theta}$. Then if $\mathcal{R}(\Lambda, d_\Lambda) < \infty$, the Bayes rule, d_Λ, with respect to Λ is admissible.

Example 1.27 (Bayes rules may inadmissible)

Suppose that $X \sim N(\theta, 1)$, $\theta \sim N(0, 1)$, and

$$L(\theta, a) = \exp\left\{\frac{3\theta^2}{4}\right\}(\theta - a)^2.$$

It can easily be shown that $\Lambda(\theta|x) = N(\frac{x}{2}, \frac{1}{2})$. Using Corollary 1.1, we obtain the Bayes rule $d_\Lambda = 2X$.

Then it follows that

$$\begin{aligned} R(\theta, d_\Lambda) &= \exp\left\{\frac{3\theta^2}{4}\right\}(4+\theta^2) \\ &> \exp\left\{\frac{3\theta^2}{4}\right\}(1) = R(\theta, d_1) \end{aligned}$$

where $d_1 = X$. Thus d_Λ is inadmissible.

Note that

$$\begin{aligned} \mathcal{R}(\Lambda, d_\Lambda) &= E_\Lambda[R(\theta, d_\Lambda)] \\ &= \int_{-\infty}^{\infty} \exp\left(\frac{3\theta^2}{4}\right)(4+\theta)^2 \lambda(\theta)\, d\theta \\ &= \infty, \end{aligned}$$

where $\lambda(\theta) = (2\pi)^{-1/2} e^{-\frac{\theta^2}{2}}$.

In fact it also follows that $\mathcal{R}(\Lambda, d(\cdot|x)) = \infty$ for all randomized decision rules $d(\cdot|x)$. This example thus shows that Bayes rules may not be admissible if their Bayes risks are infinite.

Admissibility of Generalized Bayes Rules

As with Bayes rules, generalized Bayes rules need not be admissible. One situation in which a generalized Bayes rule can be shown to be admissible is when

$$\mathcal{R}(\Lambda, d_{GB}) = \int_{\Theta} R(\theta, d_{GB})\, d\Lambda(\theta) < \infty\,, \tag{1.13}$$

where $\Lambda(\theta)$ is an improper prior. If $\Lambda(\theta)$ was a proper prior distribution, then (1.13) would be the Bayes risk. For improper Λ, the meaning of $\mathcal{R}(\Lambda, d_{GB})$ is unclear. Nonetheless, the generalized Bayes estimator is still obtained by minimizing the posterior expected loss.

It is unfortunately rather rare to have $\mathcal{R}(\Lambda, d_{GB}) < \infty$ for improper Λ. When $\mathcal{R}(\Lambda, d_{GB}) = \infty$, generalized Bayes estimators can be inadmissible. The following example illustrates this point.

Example 1.28

Suppose $X \sim \text{Gamma}(\alpha, \frac{1}{\theta})$, $\alpha > 1$, α known, where $p(x|\theta) = \frac{\theta^{-\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\theta}}$. We wish to estimate θ assuming $L(\theta, a) = (\theta - a)^2$. We observe X. Since θ is a scale parameter, we consider the improper prior $\lambda(\theta) \propto \theta^{-1}$. The posterior density of θ is given by

$$\begin{aligned} \lambda(\theta|x) &\propto p(x|\theta)\lambda(\theta) \\ &\propto \theta^{-\alpha} e^{-x/\theta} \theta^{-1} \\ &= \theta^{-\alpha-1} e^{-x/\theta}, \end{aligned}$$

which can be recognized as the kernel of an $\text{IG}(\alpha, x)$ distribution.

Thus

$$\lambda(\theta|x) = \text{Inverse-Gamma}\,(\alpha, x)\ .$$

Since the loss is squared error loss, the Bayes rule (estimator) is the posterior mean of θ, which is $E(\theta|x) = d_{GB}(x) = \frac{X}{\alpha-1}$. Now consider the risk of the estimator $d_c(x) = cX$. [We note that $E(X|\theta) = \alpha\theta$ and $\text{Var}(X|\theta) = \alpha\theta^2$, and $Z \sim \text{IG}(a,b)$ if

$$p(z|a,b) = \frac{b^a}{\Gamma(a)} z^{-a-1} e^{-b/z}\ ,$$

$X = \frac{1}{Z} \sim \text{Gamma}(a,b)$].

The risk of $d_c(x)$ is given by

$$\begin{aligned} R(\theta, d_c) &= E_{X|\theta}[cX - \theta]^2 \\ &= E_{X|\theta}[c(X - \alpha\theta) + (c\alpha - 1)\theta]^2 \\ &= c^2\alpha\theta^2 + (c\alpha - 1)^2\theta^2 \\ &= \theta^2[c^2\alpha + (c\alpha - 1)^2] \,. \end{aligned} \tag{1.14}$$

Differentiating with respect to c and setting equal to zero shows that the value of c minimizing (1.14) is unique and is given by $c_0 = (\alpha + 1)^{-1}$. Thus $d_{c_0} = \frac{X}{\alpha+1}$ is admissible.

It follows that if $c \neq c_0$, then $R(\theta, d_{c_0}) < R(\theta, d_c)$ for all θ, showing in particular that $d_{GB} = \frac{X}{\alpha-1}$ is inadmissible. The ratio of risks of d_{GB} and d_{c_0} is

$$\frac{R(\theta, d_{GB})}{R(\theta, d_{c_0})} = \frac{\alpha(\alpha-1)^{-2} + \left(\frac{\alpha}{\alpha-1} - 1\right)^2}{\alpha(\alpha+1)^{-2} + \left(\frac{\alpha}{\alpha+1} - 1\right)^2} = \frac{(\alpha-1)^{-2}}{(\alpha+1)^{-2}} .$$

For small α, d_{GB} has significantly worse risk than $d_{c_0} \equiv \frac{X}{\alpha+1}$.

Example 1.29

Suppose $\boldsymbol{X} = (X_1, \ldots, X_k)$, $\boldsymbol{X} \sim N_k(\boldsymbol{\theta}, \boldsymbol{I})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. It is desired to estimate $\boldsymbol{\theta}$ under squared error loss, $L(\boldsymbol{\theta}, \boldsymbol{a}) = (\boldsymbol{\theta} - \boldsymbol{a})'(\boldsymbol{\theta} - \boldsymbol{a})$, or $L(\boldsymbol{\theta}, \boldsymbol{a}) = (\boldsymbol{\theta} - \boldsymbol{a})'\boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{a})$, $\boldsymbol{B}$ is positive definite $(k \times k)$. In either loss function, the Bayes estimator of $\boldsymbol{\theta}$ is the posterior mean of $\boldsymbol{\theta}$.

Suppose we have one observation $\boldsymbol{X} = (X_1, \ldots, X_k)$. We consider a uniform improper prior for $\boldsymbol{\theta}$ in $\Re^k$, and thus $\lambda(\boldsymbol{\theta}) \propto 1$. Under this prior, it is easily shown that $\lambda(\boldsymbol{\theta}|\boldsymbol{x}) = N_k(\boldsymbol{x}, \boldsymbol{I})$. The generalized Bayes estimator of $\boldsymbol{\theta}$ is thus $d_{GB}(\boldsymbol{x}) = \boldsymbol{X}$. Note that if a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ was taken, the generalized Bayes estimator would just be the vector of sample means, that is, $d_{GB}(\boldsymbol{x}) = (\bar{\boldsymbol{X}}_1, \ldots, \bar{\boldsymbol{X}}_k)$, where $\bar{\boldsymbol{X}}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$.

This most standard of estimators is admissible for $k = 1$ or 2, but surprisingly is inadmissible for $k \geq 3$, as shown in Theorem 1.17. James and Stein (1960) showed that

$$d_{JS}(\boldsymbol{x}) = \left(1 - \frac{(k-2)}{|\boldsymbol{X}|^2}\right)\boldsymbol{X}$$

has $R(\boldsymbol{\theta}, d_{JS}) < R(\boldsymbol{\theta}, d_{GB})$ for all $\boldsymbol{\theta}$ if $k \geq 3$. It should be noted that inadmissibility in this example is in some sense, less serious than that of Example 1.28. In fact, the ratio of $R(\boldsymbol{\theta}, d_{JS})$ to $R(\boldsymbol{\theta}, d_{GB})$ is very close to 1 over most of the parameter space. Only in a small region near 0 (several standard deviations wide) will the ratio of risks be significantly smaller than 1. This is in contrast to the situation of Example 1.28, in which the ratio of risks can be uniformly bad. The estimator d_{JS} can be modified so as to adjust the region of significant improvement to coincide with prior knowledge concerning $\boldsymbol{\theta}$, that is, a proper prior for $\boldsymbol{\theta}$.

An example of such a modified estimator when $\boldsymbol{\theta} \sim N_k(\boldsymbol{\mu}, \tau^2 \boldsymbol{I})$ is

$$d_S(\boldsymbol{x}) = \boldsymbol{X} - \min\left\{ \frac{\sigma^2}{\sigma^2 + \tau^2} \, , \frac{2(k-2)\sigma^2}{|\boldsymbol{X} - \boldsymbol{\mu}|^2} \right\} (\boldsymbol{X} - \boldsymbol{\mu}) \, .$$

It can be shown that $R(\boldsymbol{\theta}, d_S) < k\sigma^2$, while $R(\boldsymbol{\theta}, d_{GB}) \equiv k\sigma^2$. Hence d_S is better than $d_{GB}(\boldsymbol{x}) = \boldsymbol{X}$. If an improper prior for $\boldsymbol{\theta}$ is used, then use of $d_{JS}(\boldsymbol{x})$ (or some modification of it) will not be significantly beneficial and $d_{GB}(\boldsymbol{x}) = \boldsymbol{X}$ might as well be used.

Example 1.30

Suppose that $\boldsymbol{X} \sim N_k(\boldsymbol{\theta}, \boldsymbol{I})$ and that it is desired to estimate $\eta = \boldsymbol{\theta}'\boldsymbol{\theta} = \sum_{i=1}^{k} \theta_i^2$. Suppose we specify the improper prior $\lambda(\boldsymbol{\theta}) \propto 1$. Under squared error loss, the generalized Bayes estimator of η is the posterior mean of η,

and thus

$$\begin{aligned} d_{GB}(\boldsymbol{x}) = E(\eta|\boldsymbol{x}) &= \sum_{i=1}^{k} E(\theta_i^2|\boldsymbol{x}) \\ &= \sum_{i=1}^{k} E[\{(\theta_i - x_i) + x_i\}^2|\boldsymbol{x}] \\ &= \sum_{i=1}^{k}(1 + x_i^2) = k + \sum_{i=1}^{k} x_i^2 \ . \end{aligned}$$

This result seems somewhat counter-intuitive. For instance, as $k \to \infty$ and under mild conditions, the law of large numbers shows that

$$\frac{1}{k}\sum_{i=1}^{k} X_i^2 \to \frac{1}{k}\sum_{i=1}^{k} \theta_i^2 + 1 \ ,$$

so that

$$\begin{aligned} \frac{1}{k} d_{GB}(\boldsymbol{x}) &= 1 + \frac{1}{k} \sum_{i=1}^{k} X_i^2 \\ &\to 1 + \frac{1}{k} \sum_{i=1}^{k} \theta_i^2 + 1 = 2 + \frac{\eta}{k} \, . \end{aligned}$$

This seems to indicate that $d_{GB}(\boldsymbol{x})$ substantially overestimates η. Moreover, $d_{GB}(\boldsymbol{x})$ can be shown to be inadmissible. Specifically, it can be shown that the estimator $d_c(\boldsymbol{x}) = |\boldsymbol{X}|^2 - c$ has $R(\boldsymbol{\theta}, d_c) = 2k + (k-c)^2 + 4|\boldsymbol{\theta}|^2$, so that $R(\boldsymbol{\theta}, d_{GB}) - R(\boldsymbol{\theta}, d_k) = 4k^2$, a substantial uniform difference (although again $\frac{R(\boldsymbol{\theta}, d_{GB})}{R(\boldsymbol{\theta}, d_k)} \to 1$ as $|\boldsymbol{\theta}|^2 \to \infty$.) Here $d_k = d_k(\boldsymbol{x}) = |\boldsymbol{X}|^2 - k$, i.e. $d_c(\boldsymbol{x})|_{c=k}$. The reason that $\lambda(\boldsymbol{\theta}) \propto 1$ seems to give bad results in this example can be seen by transforming $\lambda(\boldsymbol{\theta})$ to a density λ^* on η.

The result is that $\lambda^*(\eta) \propto \eta^{\frac{k-2}{2}}$. Thus, by using $\lambda(\boldsymbol{\theta}) \propto 1$, one is implicitly assuming that very large η are more likely than small η.
In particular, since the likelihood function is concentrated in the area where η is near $|\boldsymbol{x}|^2$, it is the part of $\lambda^*(\eta)$ near $|\boldsymbol{x}|^2$ that is operationally relevant, and the prior bias towards large η effects a pronounced shift upwards in the estimate.
This example demonstrates that "standard" choices of noninformative priors may not always be accurate reflections of true vague prior beliefs. It could be the case that $\lambda(\boldsymbol{\theta}) \propto 1$ but it may be more reasonable to assume that $\lambda(\eta) \propto 1$.

Example 1.31

Suppose that $X \sim \text{Exponential}(\frac{1}{\theta})$ is observed, $p(x|\theta) = \theta^{-1}e^{-x/\theta}$. It is desired to test $H_0 : \ 0 < \theta \leq 1$ versus $H_1 : \ 2 \leq \theta < \infty$. Here the parameter space is $\boldsymbol{\Theta} = (0, 1] \cup [2, \infty)$. Let a_i denote accepting $H_i, i = 0, 1$.

The loss function is given by

$$L(\theta, a_0) = \begin{cases} 0, & 0 < \theta \leq 1 \\ 4, & \theta \geq 2 \end{cases},$$

$$L(\theta, a_1) = \begin{cases} 0, & \theta \geq 2 \\ 5 - \theta, & 0 < \theta \leq 1 \end{cases}.$$

It seems plausible that the least favorable prior is one which makes H_0 and H_1 as hard to distinguish as possible, namely one which gives positive probability on to the points $\theta = 1$ and $\theta = 2$. Assume that $\lambda(\theta)$ is such a prior, and let $\lambda_1 = \lambda(1) = 1 - \lambda(2)$. We still must find a least favorable choice λ_1.

To find the Bayes rule with respect to λ, note that the posterior expected losses of a_0 and a_1 are

$$E_{\theta|x}[L(\theta, a_0)] = \frac{4(1-\lambda_1)(\frac{1}{2}e^{-\frac{x}{2}})}{p(x)},$$

$$E_{\theta|x}[L(\theta, a_1)] = \frac{4\lambda_1 e^{-x}}{p(x)},$$

where $p(x) = \int_{\Theta} p(x|\theta)\, d\Lambda(\theta)$ is the marginal distribution of X. Thus the Bayes rule is to choose a_0 if

$$\frac{(1-\lambda_1)(e^{-x/2})}{p(x)} < \frac{2\lambda_1 e^{-x}}{p(x)},$$

which can be rewritten as $x < 2\log[\frac{2\lambda_1}{1-\lambda_1}] = c$.

We will, therefore, consider the class of rules defined by the test functions $\phi_c(x) = I(c < x < \infty)$. To find a least favorable λ_1, or alternatively a least favorable c, note that

$$\begin{aligned} R(\theta, \phi_c) &= \begin{cases} \int_c^\infty (5-\theta)\theta^{-1} e^{-x/\theta} dx, & 0 < \theta \leq 1 \\ \int_0^c 4\theta^{-1} e^{-x/\theta} dx, & \theta \geq 2 \end{cases} \\ &= \begin{cases} (5-\theta)e^{-c/\theta}, & 0 < \theta \leq 1 \\ 4(1-e^{-c/\theta}), & \theta \geq 2 \end{cases} . \end{aligned}$$

Define

$$h(c) = \sup_{0 \leq \theta \leq 1} (5-\theta)e^{-c/\theta}$$

and

$$g(c) = \sup_{2 \le \theta \le \infty} 4(1 - e^{-c/\theta}) \ .$$

Clearly $\sup_{\theta \in \mathbf{\Theta}} R(\theta, \phi_c) = \max\{h(c), g(c)\}$. Note that $h(c)$ is strictly decreasing in c, with $h(0) = 5$ and $\lim_{c \to \infty} h(c) = 0$. Also, $g(c)$ is strictly increasing in c, with $g(0) = 0$ and $\lim_{c \to \infty} g(c) = 4$. Hence $h(c)$ and $g(c)$ are equal for just one value of c, call it c_0, and

$$\sup_{\theta \in \mathbf{\Theta}} R(\theta, \phi_{c_0}) = \inf_c \sup_{\theta \in \mathbf{\Theta}} R(\theta, \phi_c) \ .$$

The test ϕ_{c_0} is thus a good candidate for a minimax test.

To find c_0, note first that $4(1 - e^{-c/\theta})$ is decreasing in θ, so that $g(c) = 4(1 - e^{-c/2})$. Next observe that

$$\begin{aligned} \frac{d}{d\theta}[(5-\theta)e^{-c/\theta}] &= -e^{-c/\theta} + (5-\theta)c\theta^{-2}e^{-c/\theta} \\ &= \theta^{-2}e^{-c/\theta}(-\theta^2 - \theta c + 5c) \ . \end{aligned}$$

The denominator is positive for $0 \leq \theta \leq 1$, provided that $\theta^2 + \theta c - 5c < 0$. The roots of the equation $\theta^2 + \theta c - 5c = 0$ are $\frac{1}{2}(-c \pm (c^2 + 20c)^{1/2})$, one of which is negative, while the other is larger than 1 for $c > 1/4$. Hence, if $c > 1/4$ and $0 < \theta \leq 1$, it follows that $\theta^2 + \theta c - 5c < 0$, and $(5-\theta)e^{-c/\theta}$ is maximized at $\theta = 1$. Thus $h(c) = 4e^{-c}$ for $c > 1/4$. Let's assume the solution to $h(c) = g(c)$ is some $c > 1/4$. Then we want to solve the equation

$$4e^{-c} = 4(1 - e^{-c/2}) \ .$$

Letting $z = e^{-c/2}$, this is equivalent to solving $4z^2 = 4(1-z)$. The positive solution of this latter equation is $z \cong 0.618$, which corresponds to $c \cong 0.96$ in the original equation. Since $c > 1/4$, this is indeed the unique solution. Note that $c = 0.96$ corresponds to $\lambda_1 \cong .45$. Observe finally that for $\lambda_1 \cong .45$,

$$\begin{aligned} \mathcal{R}(\Lambda, \phi_{.96}) &= \lambda_1 R(1, \phi_{.96}) + (1-\lambda_1) R(2, \phi_{.96}) \\ &= \lambda_1 h(.96) + (1-\lambda_1) g(.96) \\ &= h(.96) \end{aligned}$$

since $h(.96) = g(.96)$. Also, as shown earlier, $R(\theta, \phi_{.96}) \leq h(.96)$ for all $\theta \in \mathbf{\Theta}$. Hence Theorem 1.12 can be used to conclude that $\phi_{.96}$ is a minimax rule (test), and λ_1 is least favorable.

We have now seen several ways of finding minimax rules

(1) Direct method: find d_M such that

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

(2) Use Theorem 1.12 to find a d_Λ such that

$$\mathcal{R}(\Lambda, d_\Lambda) = \sup_{\theta\in\Theta} R(\theta, d_\Lambda) \ .$$

(3) Use Theorem 1.13 and Theorem 1.13*. Find a sequence of proper priors $\{\Lambda_k\}$ such that the Bayes risks $r_k \to r$, and find a d such that

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

That is , we find a least favorable prior distribution which yields a limiting Bayes risk r, and then find a rule d with $\sup_{\theta\in\Theta} R(\theta, d) = r$.

(4) If d_Λ is Bayes with constant risk, then d_Λ is minimax (Corollary 1.5).

Rules with constant risk are sometimes referred to as equalizer rules.

Example 1.32

Suppose $X \sim \text{Binomial}(n, \theta)$ is observed, and that it is desired to estimate θ under squared error loss. We will try and find an equalizer rule of the form $d(x) = aX + b$. Clearly

$$\begin{aligned} R(\theta, d) &= E_\theta[aX + b - \theta]^2 \\ &= E_\theta[a(X - n\theta) + \{b + (an - 1)\theta\}]^2 \\ &= a^2 n\theta(1-\theta) + \{b + (an-1)\theta\}^2 \\ &= \theta^2[-a^2 n + (an-1)^2] + \theta[a^2 n + 2b(an-1)] + b^2. \end{aligned}$$

For the risk to be constant in θ, we must have $-a^2 n + (an - 1) = 0$ and $a^2 n + 2b(an - 1) = 0$. Solving these equations for a and b gives $a = (n + \sqrt{n})^{-1}$ and $b = \frac{\sqrt{n}}{2(n+\sqrt{n})}$.

Thus, $d_0(x) = \frac{x+\sqrt{n}/2}{n+\sqrt{n}}$ is an equalizer rule.

To complete the argument, we must show that $d_0(x)$ is Bayes. With a Beta(α, β) prior for θ, $\theta|x \sim$ Beta$(\alpha + x, \beta + n - x)$, and the Bayes rule is the posterior mean of θ, given by $d_\Lambda = \frac{\alpha+X}{\alpha+\beta+n}$. The equalizer rule is clearly of this form with $\alpha = \beta = \frac{\sqrt{n}}{2}$. Hence $d_0(x)$ is Bayes (with $\alpha = \beta = \frac{\sqrt{n}}{2}$), and by Corollary 1.5, is minimax. Thus equalizer rules that are Bayes are minimax.

Example 1.33

Suppose it is desired to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, based on X. Let a_i denote accepting H_i $(i = 0, 1)$, and assume 0-l_i loss, that is,

	a_0	a_1
θ_0	0	l_1
θ_1	l_0	0

As usual, our rule is represented by the test function $\phi(x)$, where $\phi(x) = d(1|x)$.
Define

$$\alpha = E_{\theta_0}[\phi(X)] = \text{probability of Type I error},$$

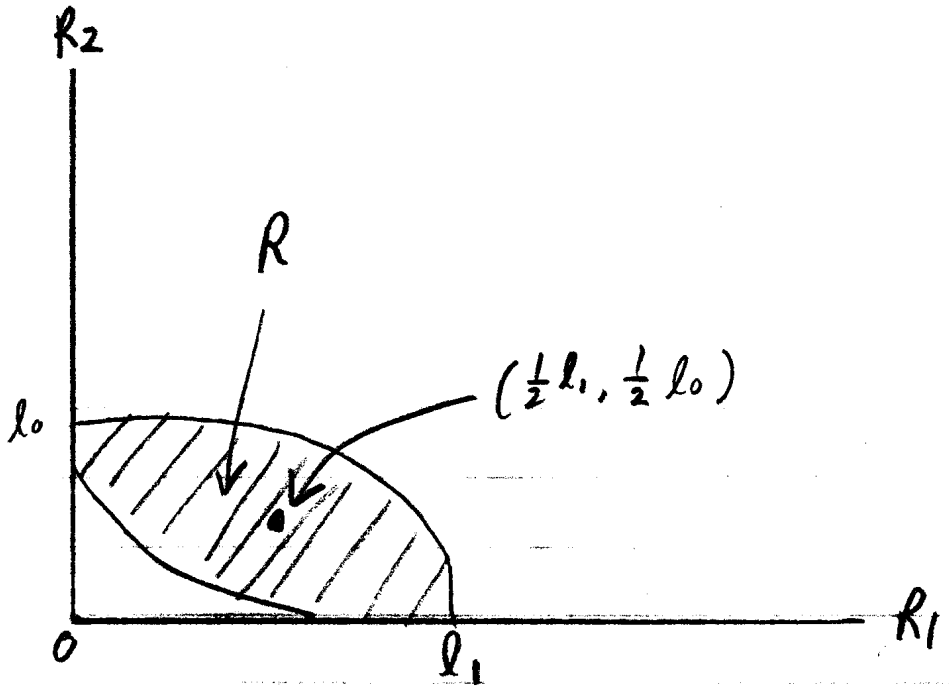
$$\beta^* = E_{\theta_1}[1 - \phi(X)] = \text{probability of Type II error},$$

$\beta^* = 1 - \beta$, $\beta =$ power.

It is easy to check that $R(\theta_0, \phi) = l_1\alpha$ and $R(\theta_1, \phi) = l_0\beta^*$. The risk set is given by

$$\boldsymbol{R} = \{(l_1\alpha, l_0\beta^*) : \phi \in D\} .$$

The risk set is closed and bounded . Thus a minimax rule exists, and the boundary of the risk set is a subset of the Bayes risk points.


R_2
R
$(\frac{1}{2}l_1, \frac{1}{2}l_0)$
l_0
0
l_1
R_1

The form in the figure above is clear. The form of the risk set $\boldsymbol{R}$ above follows from noting that $(i)\ \ 0 \leq \alpha \leq 1,\ \ 0 \leq \beta^* \leq 1$, (ii) $(0, l_0)$ and $(l_1, 0)$ are in $\boldsymbol{R}$, arising form $\phi(x) \equiv 0$ and $\phi(x) \equiv 1$, respectively, and (iii) $\boldsymbol{R}$ is symmetric about $\left(\frac{l_1}{2}, \frac{l_0}{2}\right)$, because if (R_1, R_2) arises from ϕ, then $(l_1 - R_1, l_0 - R_2)$ arises from $1 - \phi$. From the previous figure, it is clear that the minimax risk point is the point in $\boldsymbol{B}(\boldsymbol{R})$ with equal coordinates, where $\boldsymbol{B}(\boldsymbol{R})$ denotes the boundary of the risk set $\boldsymbol{R}$. Furthermore, this minimax point is the unique minimax point (corner point on the boundary). Since any point in $\boldsymbol{B}(\boldsymbol{R})$ is a Bayes risk point, it follows that the minimax rule is the Bayes rule which has constant risk.

Bayes rules in this situation are very simple.

They are of the form

$$\phi(x) = \begin{cases} 1 & \text{if} \quad l_0\lambda(\theta_1|x) > l_1\lambda(\theta_0|x) \\ \gamma(x) & \text{if} \quad l_0\lambda(\theta_1|x) = l_1\lambda(\theta_0|x) \\ 0 & \text{if} \quad l_0\lambda(\theta_1|x) < l_1\lambda(\theta_0|x) \end{cases},$$

where $0 \leq \gamma(x) \leq 1$ is arbitrary and $\lambda(\theta_1|x)$ is the posterior probability of θ_i given x. From the definition of posterior probability, it follows that ϕ can be written as

$$\phi(x) = \begin{cases} 1 & \text{if} \quad l_0\lambda_1 p(x|\theta_1) > l_1\lambda_0 p(x|\theta_0) \\ \gamma(x) & \text{if} \quad l_0\lambda_1 p(x|\theta_1) = l_1\lambda_0 p(x|\theta_0) \\ 0 & \text{if} \quad l_0\lambda_1 p(x|\theta_1) < l_1\lambda_0 p(x|\theta_0) \end{cases},$$

where λ_i is the prior probability of θ_i, $i = 0, 1$. These tests are simply the usual most powerful tests of H_0 versus H_1. The test of this form for which $l_1\alpha = l_0\beta^*$ will be minimax.

Example 1.34

Assume $X \sim \text{Binomial}(n, \theta)$ is observed, and that it is desired to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where $\theta_1 > \theta_0$. Note that $p(x|\theta)$ has a monotone likelihood ratio, i.e., if $\theta < \theta'$, the likelihood ratio $\frac{p(x|\theta')}{p(x|\theta)}$ is a nondecreasing function of x. Hence the Bayes test can be written as

$$\phi(x) = \begin{cases} 1 & \text{if} \quad x > j \\ \gamma & \text{if} \quad x = j \\ 0 & \text{if} \quad x < j \end{cases},$$

where j is some integer (depending on l_i and λ_i). For a test of this form,

$$\alpha = P_{\theta_0}(X > j) + \gamma P_{\theta_0}(X = j)$$

and

$$\beta^* = P_{\theta_1}(X < j) + (1-\gamma)P_{\theta_1}(X = j),$$

$\beta^* = 1 - \beta$, and $\beta =$ Power. We seek j and γ for which $l_1\alpha = l_0\beta^*$. As an explicit example, assume that $n = 15$, $\theta_0 = 1/4$, $\theta_1 = 1/2$, $l_1 = 1$, and $l_0 = 2$. A table of binomial probabilities shows that only $j = 5$ can possibly work. For this value of j,

$$l_1\alpha = .1484 + \gamma(.1651)$$

and

$$l_0\beta^* = 2[.0592 + (1-\gamma)(.0916)] .$$

Setting these expressions equal to each other and solving for γ gives $\gamma \cong .44$.

The minimax test is thus

$$\phi(x) = \begin{cases} 1 & \text{if } x > 5 \\ .44 & \text{if } x = 5 \\ 0 & \text{if } x < 5 \end{cases} .$$

Example 1.35 (Classification Problems)

Suppose $X \sim$ Exponential $\left(\frac{1}{\theta}\right)$, and $\Theta = \{\theta_1, \ldots, \theta_l\}$, $p(x|\theta) = \theta^{-1} e^{-x/\theta}$, where $\theta_1 < \theta_2 < \cdots < \theta_l$. It is desired to classify X as arising from the distribution indexed by $\theta_1, \theta_2, \ldots,$ or θ_l, with loss function given by

$$L(\theta_i, a_j) = \begin{cases} 1 & \text{if } \ i \neq j \\ 0 & \text{if } \ i = j \end{cases} .$$

Let λ_i be the prior probability for θ_i, $\sum_{i=1}^{l} \lambda_i = 1$. To find the minimax rule, we seek a constant risk Bayes rule. In determining the form of a Bayes rule, note that for $i < j$,

$$\begin{aligned} \lambda_i p(x|\theta_i) &= \lambda_i \theta_i^{-1} e^{-x/\theta_i} \\ &> \lambda_j \theta_j^{-1} e^{-x/\theta_j} = \lambda_j p(x|\theta_j) \end{aligned} \tag{1.15}$$

if and only if

$$x < c_{ij} = \frac{\theta_i \theta_j}{\theta_j - \theta_i} \left(\log\left(\frac{\lambda_i}{\lambda_j}\right) + \log\left(\frac{\theta_j}{\theta_i}\right) \right) . \tag{1.16}$$

On the other hand, for $j < i$,

$$\lambda_i p(x|\theta_i) > \lambda_j p(x|\theta_j) \tag{1.17}$$

if and only if $x > c_{ji}$.

We will show that these relationships imply the existence of constants $0 = \alpha_0 \leq \alpha_1 \leq \cdots \leq \alpha_l = \infty$, such that

$$\phi_i(x) = 1 \quad \text{for} \quad \alpha_{i-1} < x < \alpha_i, \ i = 1, \ldots, l \, . \tag{1.18}$$

Note first that $\lambda_i p(x|\theta_i) = \lambda_j p(x|\theta_j)$ only if $x = c_{ij}$. Since $\{x : x = c_{ij}, \ 1 \leq i \leq l, \ 1 \leq j \leq l\}$ is finite, it follows that, with probability 1, $\phi_i(x)$ is 0 or 1. But (1.15), (1.16), and (1.17) imply that $\phi_i(x) = 1$ if and only if

$$\max_{1 \leq j < i} \{c_{ji}\} < x < \min_{i < j \leq l} \{c_{ij}\} \, . \tag{1.19}$$

This shows that $\phi_i(x) = 1$ for x in some interval. The fact that the intervals must be ordered according to i follows directly from (1.15) and (1.16), noting that if $i < j$, $\phi_i(x_i) = 1$, and $\phi_j(x_j) = 1$ (so that $\lambda_i p(x_i|\theta_i) > \lambda_j p(x_i|\theta_j)$ and $\lambda_i p(x_j|\theta_i) < \lambda_j p(x_j|\theta_j)$), then $x_i < c_{ij} < x_j$.

For a decision rule of the form (1.18), it is clear that

$$\begin{aligned} R(\theta_i, \boldsymbol{\phi}) &= \sum_{j=1}^{l} L(\theta_i, a_j) E_{\theta_i}[\phi_j(X)] \\ &= 1 - E_{\theta_i}[\phi_i(X)] \ , \end{aligned}$$

and thus

$$\begin{aligned} R(\theta_i, \phi) &= 1 - \int_{\alpha_{i-1}}^{\alpha_i} \theta_i^{-1} \exp\left\{\frac{-x}{\theta_i}\right\} \, dx \\ &= 1 - \exp\left\{\frac{-\alpha_{i-1}}{\theta_i}\right\} + \exp\left\{\frac{-\alpha_i}{\theta_i}\right\} \ , \end{aligned}$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_l)$.

To find a constant risk rule of this form, simply set the $R(\theta_i, \phi)$ equal to each other and solve for the α_i.

As an explicit example, assume that $l = 3$, $\theta_1 = 1$, $\theta_2 = 2$, and $\theta_3 = 4$. Setting that risks equal to each other, results in the equations

$$e^{-\alpha_1} = 1 - e^{-\alpha_1/2} + e^{-\alpha_2/2} = 1 - e^{-\alpha_2/4} .$$

Letting $z = e^{-\alpha_2/4}$ and $y = e^{-\alpha_1/2}$, these equations become $y^2 = 1 - y + z^2 = 1 - z$. Clearly $z = 1 - y^2$, so that the equation $y^2 = 1 - y + z^2$ can be rewritten $y^2 = 1 - y + (1 - y^2)^2$, or $y^4 - 3y^2 - y - 2 = 0$.

The appropriate solution to this is $y \approx .74$, from which it can be calculated that $z \approx 1 - (.74)^2 = .45$, $\alpha_1 \approx .60$ and $\alpha_2 \approx 1.60$. These can be checked to correspond to the prior with $\lambda_1 = .23$, $\lambda_2 = .33$, and $\lambda_3 = .44$. The minimax rule is thus to decide a_1 (i.e., classify x as arising from $p(x|\theta_1)$) if $0 < x < .6$, decide a_2 if $.6 < x < 1.6$, and decide a_3 if $x > 1.6$.

Example 1.36

Suppose $\boldsymbol{X} \sim N_k(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ is known, and that $L(\boldsymbol{\theta}, \boldsymbol{a}) = (\boldsymbol{\theta} - \boldsymbol{a})' \boldsymbol{Q} (\boldsymbol{\theta} - \boldsymbol{a})$, where $\boldsymbol{Q}$ is a known $k \times k$ positive definite matrix. The "standard" minimax estimator is $d(\boldsymbol{x}) = \boldsymbol{X}$, which is the least squares estimator in standard regression settings. This estimator is standard in the sense that it is a minimax equalizer rule (exercise). For $k = 1$ or 2, we know that this estimator is admissible, but it is inadmissible for $k \geq 3$ (Stein phenomenon), so other minimax estimators exist.