Loss Functions for Set Estimations

BU-999-MA June, 1992

George Casella, J.T.-Gene Hwang[1], and Christian P. Robert[2]

Cornell University, Ithaca, N.Y. and Paris University (Rouen University?)

Key Words and Phrases: Confidence sets, Decision theory, Bayes estimation.

AMS 1980 Subject Classification: Primary 62C05, Secondardy 62F25, 62A99.

---

[1] Research supported by National Science Foundation Grant No. DMS88-09016.

[2] Research supported by the U.S. Army Research Office through the Mathematical Sciences Institute at Cornell University.

Summary

Approaches to set estimation based on a decision-theoretic formulation have usually used a loss function that is a linear combination of volume and coverage probability. Such loss functions can suffer from paradoxical behavior of the Bayes rules, and thus may not be appropriate. We investigate the behavior of optimal set estimators for different classes of loss functions and study their decision-theoretic properties.

1. Introduction

Set estimators contribute an inherent part of most statistical analyses, providing sets C_x depending on the observation $x \sim f(x|\theta)$ for the parameter of interval θ. Although they are thus bound to be used extensively in decisions, there has been hardly any move towards a decision-theoretic evaluation of these procedures.

Set estimators are usually derive from testing acceptance regions by duality. However, contrary to intuition, these estimators do not necessarily enjoy the optimality properties of their testing dual sets, as shown by Hwang and Casella (1982) for the domination of the usual normal confidence region.

A direct approach to set estimation has to take into account two criteria, size and coverage. The resting alternative can also be perceived this way for one of the two criteria remaining constant, namely the coverage probability of the considered set. Apart from vector loss generalizations, for which a complete ordering is most often impossible, a loss function considered in the literature is the *linear loss function*,

$$L(\theta, C) = a\ \mathrm{vol}(C) + \mathrm{II}(\theta \in C)\ , \tag{1.1}$$

where a is a fixed constant, as in Cohen and Strawderman (1973) and Meeden and Vardeman (1985). This loss function indeed allows for interplay between volume and coverage but it can also suffer from major defects. One of these leads to a paradox pointed to us by James Berger and studied in Casella, Hwang and Robert (1992). We also refer the reader to this paper for additional references on loss estimation.

The so-called paradox exhibits a domination under (1.1) of the Student's t-interval by a truncated set which is empty when s^2, the empirical variance, is *large enough*. However, the Bayes estimator associated with (1.1) and the Lebesgue prior is such that its length is decreasing in $\mathbf{S}$ when $\mathbf{S}$ is large enough. These phenomena incriminate the loss (1.1) and the fact that it does not balance properly the two criteria, giving an undue importance to the volume component.

In this paper, we propose to study the decision-theoretic properties of the losses of the form

(1.2) $$L(\theta, C) = \mathbb{S}[\text{vol}(C)] + \text{II}(\theta \notin C) ,$$

where $\mathbb{S}$ is a *size* function. Casella *et al.* (1992) already showed that some losses in (1.2) can avoid Berger's paradox and came to the following requirements on $\mathbb{S}$: the size function must be increasing and satisfy $\mathbb{S}(0) = 0$, $\lim_{t \to +\infty} \mathbb{S}(t) = 1$ so that volume *and* coverage are weighted equally. These conditions are only necessary since some functions $\mathbb{S}$ satisfy them but still lead to Berger's paradox (see Casella *et al.*, 1992).

Section 2 is dedicated to a classical analysis of losses like (1.2) and the derived notions of admissibility, minimaxity and Bayes data. Section 3 examines the range of Bayes sets for different parameterized families of losses and derives conditions under which conjugate Bayes sets are never empty. Section 4 focuses on two particular classes of losses, one the bounded spaces, the other for unbounded spaces.

2. Decision Theory

In this section, we consider usual decision-theoretic properties derived from the loss function

$$L_S(\theta, C) = S[\mathrm{vol}(C)] + \Pi(\theta \notin C) , \tag{2.1}$$

where S is increasing from 0 to 1. First note that, as mentioned in Joshi (1969), the action space is not $\mathfrak{B}(\Theta)$, set of all measurable v subsets of Θ, but rather $\mathbb{B}(\Theta)$, the quotient space derived from $\mathfrak{B}(\Theta)$ by the equivalence relation $C \sim C'$ iff $\lambda(C \Delta C') = 0$, where λ is the Lebesgue measure on Θ and $C \Delta C'$ denotes $\left(C \mathrm{n} (C')^0\right) \cup \left(C' \mathrm{n} C^0\right)$. In other words, confidence sets can only be defined up to a set of Lebesgue measure zero. Otherwise, no comparison is possible. For this reason, the only acceptable priors are those which are absolutely continuous with the Lebesgue measure.

2.1. Bayes and Generalized Bayes Estimators

For any given loss function $L(\theta, C)$ and prior distribution π on θ, a (decision-theoretic) *Bayes estimator* C^π is a solution of the minimization problem

$$\int_\theta R(\theta, C^\pi)\pi(\theta)d\theta = \min_C \int_\theta R(\theta, C)\pi(\theta)d\theta . \tag{2.2}$$

If $\pi(\theta)$ is a proper prior distribution, for each x the Bayes estimator C_x^π minimizes

$$\int_\theta L(\theta, C)\pi(\theta \mid x)d\theta , \tag{2.3}$$

where $\pi(\theta \mid x)$ is the posterior distribution. If $\pi(\theta)$ is not a proper prior distribution but $\pi(\theta \mid x)$ is proper, then (2.3) is taken as the definition of C_x^π. Thus, in general, C_x^π minimizes the posterior expected loss. For loss functions (2.1), the Bayes estimators are HPD regions.

Theorem 2.1. *For $X \sim f(x \mid \theta)$ and (possibly improper) prior $\pi(\theta)$, the Bayes estimator against the loss* (2.1) *is given by*

$$C_x^\pi = \left\{\theta : \pi(\theta) \geq k(x)\right\} ,$$

where $k = k(x)$ minimizes the quantity

$$S\left(\mathrm{vol}\{\theta : \pi(\theta \mid x) \geq k\}\right) - \int_{\{\theta : \pi(\theta \mid x) \geq k\}} \pi(\theta \mid x)d\theta ,$$

provided that $\left\{\theta : \pi(\theta \mid x) = \ell\right\}$ has Lebesgue measure zero for every ℓ.

The condition that $\{\theta : \pi(\theta \mid x) = \ell\}$ has zero Lebesgue measure is a technical one, and has been added to eliminate the necessity of considering randomized rules. We chose to include it in the assumptions for ease of understanding. A more general theorem which does not need this assumption is included, for completeness, in the Appendix. In most cases, however, Theorem 2.1 is general enough.

2.2 Admissible Estimators

Given the previous restrictions on the definition of the action space, we will say that an estimator C is *admissible* if, whenever there exists an estimator C′ satisfying $R(\theta,C') \leq R(\theta, C)$ for almost all θ, then $R(\theta,C') = R(\theta, C)$ for almost all θ.

We will consider here the relation between Bayes estimators and admissibility. Things are not as clear as for point estimation and some questions remain unanswered; in particular, we do not yet know if there exist equivalent theorems to the complete class theorems of point estimation.

Theorem 2.2. *If π gives positive measure to any set with positive Lebesgue measure, then the Bayes estimator C^{π} is admissible.*

If π satisfies the conditions of the previous theorem, then all the Bayes sets are admissible, even if C^{π} is not essentially unique. However, unlike the point estimation case with strictly convex loss, it may be possible to exhibit inadmissible proper Bayes estimators. Obviously, generalized Bayes estimators may also be inadmissible (as in the point estimation problem).

Example 2.1. It is easily seen that the usual confidence set

$$C_x^0 = \{\theta : |\theta - x| \leq c\}$$

is generalized Bayes with respect to the Lebesgue measure. However, for any choice of the loss, C_x^0 is inadmissible. Consider the recentered set,

$$C_x^a = \{\theta : |\theta - \delta_a^+(x)| \leq c\},$$

where δ_a^+ is the positive-part James Stein estimator

$$\delta_a^+(x) = \left(1 - \frac{a}{|x|^2}\right)^+ x .$$

Hwang and Casella (1982, 1984) have established that $P_\theta(\theta \in C_x^0) < P_\theta(\theta \in C_x^a)$ for all θ and some values of a (depending on c), and the two sets C_x^0 and C_x^a have the same volume.

It is not true that any $\mathbf{S}(\cdot)$ function considered in (2.1) will result in nonparadoxical behavior. There are, in fact, numerous examples of losses and prior distributions which give ϕ (or Θ) as an admissible constant Bayes set estimator. This problem is considered in Section 4.

2.3. Minimaxity

For a given loss, L, the *minimax risk* is defined by

$$\mathbf{m} = \inf_{C} \sup_{\theta} R(\theta, C) . \tag{2.4}$$

As $R(\theta, C) = 1$ for every θ, it is obvious that $\mathbf{m} \leq 1$. We would like to known when θ is a minimax estimator, i.e., when does $\mathbf{m} = 1$, in order to avoid such losses. In particular, we have the following condition.

Theorem 2.3. *If* $X \sim N_p(\theta,I)$, *a necessary and sufficient condition for* $\mathbf{m} = 1$ *is that*

$$\mathbf{S}\left(\mathfrak{B}c^p\right) \geq P\left(|Z| < c\right)$$

for every $c \geq 0$, *where* $Z \sim N_p(0,I)$ *and* β *is the volume of the unit ball.*

Proof. If there exists c_0 such that $\mathbf{S}\left(\mathfrak{B}c_0^p\right) < P\left(|Z| < c_0\right)$, the estimator

$$C_x^0 = \{\theta : |x-\theta| \leq c_0\}$$

has a constant risk strictly smaller than 0. Hence $\mathbf{m} < 0$, establishing the necessity. For the sufficiency, consider the conjugate priors $N(0, \tau^2 I)$. The corresponding Bayes sets are

$$C_x^{\pi} = \left\{\theta : \left|\theta - \frac{\tau^2}{\tau^2+1}\, x\right| < \ell_\tau^* \sqrt{\frac{\tau^2}{\tau^2+1}}\right\}$$

where ℓ_τ^* attains the minimum

$$\min_{\ell}\left[\mathbf{S}\left(\mathfrak{B}\left(\frac{\tau^2}{\tau^2+1}\right)^{p/2} \ell^p\right) + P\left(|Z| > \ell\right)\right].$$

The minimum value is also the Bayes risk and hence is a lower bound on $\mathbf{m}$. The sufficiency will therefore be established if we can show that the minimum is bounded below by zero as $\tau^2 \to \infty$. Now the minimum equals

$$\min_{\ell}\left[S\left(\mathfrak{B}\ell^p\right) - P\left(|Z| < \left(\frac{\tau^2}{\tau^2+1}\right)^{-1/2}\ell\right)\right] \geq \min_{\ell}\left[S\left(\mathfrak{B}\ell^p\right) - P\left(|Z| < \ell\right)\right]$$

$$+ \min_{\ell}\left[P\left(|Z| < \ell\right) - P\left(|Z| < \left(\frac{\tau^2+1}{\tau^2}\right)^{1/2}\ell\right)\right].$$

The first term of the lower bound is nonnegative by assumption, and the second term goes to zero as $\tau^2 \to \infty$. The proof is now complete. □

An example of loss for which $\mathbf{m} = 0$ is given below. First, we define the effective radius of a p dimensional set C, $\rho(C)$, as $\rho(C) = \left(\text{vol}(C)/\mathfrak{B}\right)^{1/p}$.

Example 2.2. Let $X \sim \mathcal{N}\left(\theta, I_p\right)$, $Z \sim \mathcal{N}\left(0, I_p\right)$ and the loss is

$$L(\theta, C) = P\left(|Z| < 2\rho(C)\right) + \mathbb{I}(\theta \notin C),$$

under the prior distribution $N(0, \tau^2)$. Then the posterior distribution is $N\left(\frac{\tau^2}{\tau^2+1}x, \frac{\tau^2}{\tau^2+1}I\right)$ and the Bayes set is of the form

$$C_x^{\pi} = \left\{\theta : \left|\theta - \frac{\tau^2}{\tau^2+1}x\right| \leq \ell^* \sqrt{\frac{\tau^2}{\tau^2+1}}\right\},$$

where ℓ^* minimizes

$$P\left(|Z| \leq 2\ell\sqrt{\frac{\tau^2}{\tau^2+1}}\right) - P\left(|Z| \leq \ell\right). \tag{2.5}$$

If $[\tau^2/(\tau^2+1)] > 1/4$, a solution of (2.5) is $\ell^* = 0$ or ∞; thus $C_x^{\pi} = \emptyset$ or Θ. Since π satisfies the conditions of Theorem 2.2, $\emptyset$ is admissible and $\mathbf{m} = 0$.

The above example illustrates a technique used to establish that $\mathbf{m} = 1$, namely to show that some Bayes estimators are ϕ. The nonexistence of trivial Bayes sets is actually a necessary condition for $\mathbf{m} < 1$ (see Section 2). Another way to show that $\mathbf{m} < 1$ is to exhibit an estimator with maximum risk (in θ) less than 1, as illustrated in the following example, for a bounded parameter space [0, 1]. In such a case, size function is the length (volume).

Example 2.3. Let $X \sim$ Binomial(n, θ). Then, because $\theta \in [0, 1]$, the simple loss

$$L(\theta, C) = \ell(C) + \mathbb{I}(\theta \notin C), \tag{2.6}$$

where ℓ is the length of C, fits the requirements of (2.1). If $\pi(\theta)$ is the uniform distribution on [0, 1],

then $\pi(\theta \mid x)$ is Beta$(x+1,\ n-x+1)$. We will see in Section 5 that the Bayes rule is necessarily the set $\{\pi(\theta \mid x) \geq 1\}$, which is an interval $\left[\theta_x^{\ell},\ \theta_x^{u}\right]$. Numerical solutions show that, for all the observed values of n, the frequentist risk is strictly negative for every $\theta \in [0,\ 1]$. Therefore, under (2.6), $\mathbf{m} < 1$. Note that the Bayes sets $\left[\theta_x^{\ell},\ \theta_x^{u}\right]$ are not the UMPU intervals (Blyth and Hutchinson, 1961). For selected values of n, Table 2.1 gives the minimum coverage probability and maximum length of the Bayes rule.

Table 2.1. Minimum coverage probability and maximal length of the Bayes set for binomial θ using a uniform prior.

n	10	20	25	30	35	40	45
Length	.41	.34	.32	.30	.29	.28	.27
Posterior coverage	.79	.87	.88	.90	.91	.91	.92

However, the Bayes rule is not minimax as the associated Bayes risk is always strictly lower than the frequentist risk.

3. Ranges of Bayes Sets

As previously seen, one problem with a loss function approach to set estimation is the range of resulting Bayes sets. In particular, it is possible to derive proper Bayes sets that are trivial, that is, Bayes sets that may be either empty or equal to the entire parameter space. In Section 3.1, we derive a sufficient condition for a Bayes set to be nontrivial. We say that the Bayes set is *nontrivial* if it is neither $\emptyset$ nor Θ with positive posterior probability. In Section 3.2 we consider particular classes of losses and study the ranges of the Bayes sets. We give explicit bounds on this range in the normal case.

3.1. Conditions for Existence of Nontrivial Bayes Sets

We can derive a useful sufficient condition by differentiating

$$\$\Big(\mathrm{vol}\Big(C_x^{\pi}(k)\Big)\Big) - \int_{C_x^{\pi}(k)} \pi(\theta \mid x)\,d\theta\,, \tag{3.1}$$

where $C_x^{\pi}(k) = \{\theta : \pi(\theta \mid x) > k\}$, with respect to k.

Lemma 3.1. *For the set* $C_x^{\pi}(k) = \{\theta : \pi(\theta \mid x) \geq k\}$, *is* $\$$ *is differentiable, then*

$$\frac{\partial}{\partial k}\left[\$\Big(\mathrm{vol}\Big(C_x^{\pi}(k)\Big)\Big) - \int_{C_x^{\pi}(k))} \pi(\theta \mid x)\,d\theta\right] = \left[k - \$'\Big(\mathrm{vol}\Big(C_x^{\pi}(k)\Big)\Big)\right] \int_{\{\theta:\pi(\theta|x)=k\}} \frac{ds}{|\nabla\pi(\theta|x)|}\,. \tag{3.2}$$

Here, ds *represents the infinitesimal surface area of the set* $\{\theta : \pi(\theta \mid x) = k\}$ *and* $\nabla\pi(\theta \mid x)$ *is the gradient of* $\pi(\theta \mid x)$ *for fixed* x.

Proof. Straightforward differentiation yields

$$\frac{\partial}{\partial k}\left[\$\Big(\mathrm{vol}\Big(C_x^{\pi}(k)\Big)\Big) - \int_{C_x^{\pi}(k)} \pi(\theta \mid x)\,d\theta\right] = \$'\Big(\mathrm{vol}\Big(C_x^{\pi}(k)\Big)\Big)\frac{\partial}{\partial k}\,\mathrm{vol}\Big(C_x^{\pi}(k)\Big) - \frac{\partial}{\partial k}\int_{C_x^{\pi}(k)} \pi(\theta \mid x)\,d\theta\,.$$

From the definition of $C_x^{\pi}(k)$,

$$\frac{\partial}{\partial k}\,\mathrm{vol}\Big(C_x^{\pi}(k)\Big) = \frac{\partial}{\partial k}\int_{\{\theta:\pi(\theta|x)\,\geq\,k\}} d\theta = \lim_{\Delta k \to 0}\frac{1}{\Delta k}\int_{\{\theta:k\,+\,\Delta k\,\geq\,\pi(\theta|x)\,\geq\,k\}} d\theta\,. \tag{3.2}$$

Let $h_x(\theta)$ be the perpendicular infinitesimal distance between the sets $\{\theta : \pi(\theta \mid x) = k\}$ and $\{\theta : \pi(\theta \mid x) = k + \Delta k\}$. Clearly $\Delta k = h_x(\theta)\,|\nabla\pi(\theta \mid x)|$ and hence from (3.2) we obtain

$$\frac{\partial}{\partial k}\,\mathrm{vol}\left(C_x^{\pi}(k)\right) = \lim_{\Delta k \to 0} \frac{1}{\Delta k} \int h_x(\theta)\,ds = -\int \frac{ds}{|\nabla \pi(\theta \mid x)|}\,. \tag{3.4}$$

Similarly, we have

$$\frac{\partial}{\partial k} \int_{\{\theta:\pi(\theta|x) \geq k\}} \pi(\theta|x)\,d\theta = -k \int \frac{ds}{|\nabla \pi(\theta \mid x)|} \tag{3.5}$$

which, together with (3.4), establishes the lemma. □

One nice feature of equation (3.2) is that to evaluate the sign of this derivative, only the term $\left[k - S'\left(\mathrm{vol}\left(C_x^{\pi}(k)\right)\right)\right]$ is important, as the sign of the integral is always nonnegative. Therefore we only need this function to determine whether an interior point is a minimum or maximum and the difficult task of calculating this surface integral can be avoided.

A useful sufficient condition can now be easily established. Due to the fact that the posterior loss (3.1) is equal to zero for $k = 0$ and $k = \infty$, straightforward application of Lemma 2.1 yields the following sufficient conditions.

Theorem 3.2. *If there exists* $k_0 > 0$ *such that*

$$\left[k - S'\left(\mathrm{vol}\left(C_x^{\pi}(k)\right)\right)\right] < 0 \quad \textit{for } k \textit{ satisfying } 0 < k < k_0\,,$$

or

$$\left[k - S'\left(\mathrm{vol}\left(C_x^{\pi}(k)\right)\right)\right] > 0 \quad \textit{for all } k > k_0\,,$$

the Bayes set C_x^{π} *is nontrivial.*

3.2. The Normal Case

We now investigate estimation of θ in the normal setting with

$$X \mid \theta \sim N_p(\theta, \sigma^2 I), \qquad \theta \sim N_p(\mu, \tau^2 I), \qquad 0 < \tau^2 \leq \infty\,.$$

The posterior distribution of $\theta \mid X = x$ is $N_p\left(m(x), \sigma_\tau^2 I\right)$, with

$$m(x) = \frac{\sigma_\tau^2}{\tau^2}\,\mu + \left(1 - \frac{\sigma_\tau^2}{\tau^2}\right)x, \quad \sigma_\tau^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\,.$$

Instead of using Theorem 3.2, we take a slightly different approach in this setting. We feel that this variation, which is only appropriate in the normal setting, will provide more insight into the problem.

Recall the definition of the effective radius $\rho(C)$, of a set C. We consider three classes of size functions expressed in terms of effective radius $\rho = \rho(C)$:

$$\text{i)}\quad \bar{S}_a(\rho) = \rho^p \Big/ \left(a + \rho^p\right), \qquad a > 0$$

$$\text{ii)}\quad \bar{S}_a(\rho) = P\left(|Z| \le a\rho\right), \qquad a > 0,\ Z \sim N_p(0, 1)$$

$$\text{iii)}\quad \bar{S}_a(\rho) = 1 - e^{-a\rho^2/2}, \qquad a > 0\ .$$

It appears that the class of losses defined using (3.6i) is a reasonable class for most situations (see Section 4). Examination of the losses given by (3.6ii) and (3.6iii) will further reveal the bounds of the theory considered here.

We can write HPD regions for θ in the form

$$C_x^{\pi}(\ell) = \left\{\theta : |\theta - m(x)| \le \ell\sigma_\tau\right\}$$

for some ℓ which does not depend on x. In fact, the Bayes rule chooses ℓ to minimize the posterior expected loss

$$L\left(C_x^{\pi}(\ell)\,|\,x\right) = \bar{S}\left(\ell\sigma_\tau\right) + P\left(|Z| > \ell\right),$$

where $Z \sim N_p(0, 1)$.

If we now apply Theorem 3.2, we get a sufficient condition: the Bayes sets are nontrivial if there exists ℓ_0 such that either

$$\sigma_\tau \bar{S}'\left(\ell\sigma_\tau\right) - f_Z(\ell) < 0 \qquad \text{for } 0 < \ell < \ell_0 \tag{3.7}$$

or

$$\sigma_\tau \bar{S}'\left(\ell\sigma_\tau\right) - f_Z(\ell) > 0 \qquad \text{for } \ell > \ell_0\ , \tag{3.8}$$

where $f_Z(\cdot)$ is the density of $|Z|$ and $f_Z(t) = t^{p-1}e^{-t^2/2} \Big/ \left[2^{(p/2)-1}\Gamma(p/2)\right]$.

We can now apply these formulas to the size functions in (3.6) to obtain a number of interesting results.

Theorem 3.3. *Let* $X\,|\,\theta \sim N_p\left(\theta, \sigma^2 I\right)$, $\theta \sim N_p\left(\mu, \tau^2 I\right)$

a. For the rational size function (3.6i), the Bayes sets are nontrivial for any $a > 0$.

b. For the probability size function (3.6ii), the Bayes sets are nontrivial for all Z if and only if $a < \sigma^{-1}$.

For the exponential size function (3.6iii), the Bayes sets are nontrivial if $p = 1$. *For* $p \geq 2$, *the Bayes sets are nontrivial for all* Z *if and only if* $a < \sigma^{-1}$.

Proof a. Condition (3.8) is obviously satisfied, as the rational size function has a heavier tail than $f_Z(\cdot)$, which decreases exponentially fast.

b. For the probability size function we have

$$L\left(C_x^{\pi}(\ell)\middle|x\right) = P\left(|Z| \leq a\ell\sigma_{\tau}\right) + P\left(|Z| > \ell\right),$$

which is clearly negative if $a\sigma_{\tau} < 1$. Thus, since $\sigma_{\tau}{-}1 > \sigma{-}2$, part b is proved.

c. From conditions (3.7)–(3.8), we will have nontrivial Bayes sets if

$$\frac{\partial}{\partial\ell} L\left(C_x^{\pi}(\ell)\middle|x\right) = \sigma_{\tau}^2 a\ell e^{-a(\ell\sigma_{\tau})^2/2} - \ell^{p-1}e^{-\ell^2/2} \Big/ \left[2^{(p/2)-1}\Gamma(p/2)\right]$$

is positive for all sufficiently large ℓ or negative for ℓ sufficiently close to zero. If $p = 1$, it is straightforward to check that $(\partial/\partial\ell)\, L\left(C_x^{\pi}(\ell)\middle|x\right) < 0$ for ℓ near 0, so the Bayes sets are nontrivial. If $p \geq 2$ and $a\sigma_{\tau} \geq 1$, then it is also straightforward to check that $L\left(C_x^{\pi}(\ell)\middle|x\right) \geq 0$ for all ℓ, so the Bayes sets are trivial by Theorem 4.1. If $a\sigma_{\tau} < 1$, then it can be shown that (3.9) is positive for all sufficiently large ℓ and hence, the Bayes sets are nontrivial. The condition follows since $\sigma_{\tau}{-}1 > \sigma{-}2$. □

The three size functions of (3.6) have been chosen because they span a range of possibilities. The relationship between the size function and the density function is of utmost importance in determining the behavior of the Bayes sets. With the size functions in (3.6), we have tails that can be heavier than normal (rational size), lighter than normal (probability size) and equal to normal (exponential size). It follows from the previous developments that the size function must have a tail heavier than the posterior density in order for the Bayes sets to be nontrivial. In less technical words, if the penalty for large size increases rapidly enough (to its upper bound), or increases slowly enough (from its lower bound), the Bayes sets will be trivial.

3.3. Bounds on Bayes Sets

In this section, we still concentrate on the normal case using the size functions defined in (3.6). For these cases, we find that as the value of a changes, there is a smallest, nonempty Bayes set.

However, there is no largest bounded Bayes set (that is, the maximum volume of a Bayes set is infinity.)

To be explicit, for a specified value of a we write the Bayes set

$$C_x^{\pi}[\ell(a)] = \left\{\theta : |\theta - m(x)| \le \ell(a)\sigma_\tau\right\}.$$

For each loss given in (3.6), there is a unique value of $\ell(a)$ that describes the Bayes set. We now find the range of this function as a varies. Consider first the rational size function.

Theorem 3.6. *If* $X \mid \theta \sim N_p(\theta, \sigma^2 I)$ *and* $\theta \sim N_p(\mu, \tau^2 I)$ *the Bayes set under rational size is such that* $\ell(a)$ *satisfies*

i) $\max_a \ell(a) = \infty$

ii) $\min_a \ell(a) = \ell^* > 0$.

Proof. We can write the posterior loss of $C_x^{\pi}[\ell(a)]$ as

$$L\left(C_x^{\pi}[\ell(a)] \,\middle|\, x\right) = \frac{(\ell(a)\sigma_\tau)^p}{a + (\ell(a)\sigma_\tau)^p} + P\left(\chi_p^2 > \ell^2(a)\right).$$

As $a \to 0$, $L\left(C_x^{\pi}[\ell(a)] \,\middle|\, x\right)$ can remain negative (as it must, by Theorem 3.3) only if $\ell \to \infty$. This establishes that $(\partial/\partial a)\, L\left(C_x^{\pi}[\ell(a)] \,\middle|\, x\right) < 0$. This implies that the minimum value of $\ell(a)$ is attained for $0 < a < \infty$, establishing part ii). □

It is straightforward to verify that the minimum value of $\ell(a)$ is independent of σ_τ. Although analytic determination of this minimum is difficult, a numerical algorithm can be used. The following

table gives some selected values of $\ell^* = \min_a \ell(a)$:

Table 3.1. For dimension p = 1,2,···,7, values of the minimum length ℓ^* and corresponding coverage of conjugate Bayes sets against the rational loss.

p	1	2	3	4	5	6	7
ℓ^*	1.90	2.07	2.21	2.29	1.40	1.40	1.29
Posterior coverage	.94	.88	.82	.74	.31	.08	.02

Of course, the posterior coverage of the Bayes set $C_X^{\pi}(\ell^*)$ is the minimum probability, and corresponds to the usual frequentist coverage when the prior is taken to be Lebesgue measure. Table 3.1 shows that for small values of p, the rational size function dictates a rather large coverage probability, and hence a restricted range of Bayes sets. As p increases, however, the range of the coverage probabilities increases.

We next turn to the probability size function.

Theorem 3.5. *For* $X \mid \theta \sim N_p(\theta, \sigma^2 I)$ *and* $\theta \sim N_p(\mu, \tau^2 I)$, *under probability size function, with* $a < \sigma^{-1}$, *the Bayes set is associated with*

$$\ell(a) = \left[\frac{-2p\log(a\sigma_\tau)^2}{1-(a\sigma_\tau)^2}\right]^{\frac{1}{2}}$$

and $\min_{a < \sigma_\tau^{-1}} \ell(a) = p^{1/2}$ *and* $\max_{a < \sigma_\tau^{-1}} \ell(a) = \infty$.

Thus the probability size function also results in an unbounded range of Bayes sets. As before, the coverage probability of the smallest Bayes set, $C_X^{\pi}(p^{1/2})$, is independent of σ_τ. However, for this size function the range of the Bayes sets is wider for small p, as evidenced in Table 3.2. Moreover, it can be established analytically that for this loss

$$\min_a P\left[\left(\theta \in C_X^{\pi}[\ell(a)]\right) \middle| x\right) \geq \frac{1}{2} \quad \text{for any } p\,.$$

Table 3.2. For selected dimensions p, values of the coverage of the minimum length Bayes set against the probability size loss.

p	1	3	5	10	20	40	100
Posterior coverage	.69	.61	.59	.56	.54	.50	.50

For the exponential size function, the results are similar to those for the probability size function, and will only be summarized.

Theorem 3.6. *For* $X \mid \theta \sim N_p(\theta, \sigma^2 I)$, $\theta \sim N_p(\mu, \tau^2 I)$ *and the expo size function, when* $a < \sigma^{-1}$ *and* $p > 2$, *the Bayes set is associated with*

$$(p-2)^{1/2} \leq \ell(a) < \infty .$$

Again, as with the probability size function, it can be established that, as lone as $p > 2$ and $a < \sigma_\tau^{-1}$, the minimum coverage probability of the Bayes set is at least 1/2. In fact, an analytic bound on the length is

$$\min_a \ell(a) \geq \min_a \sqrt{\frac{p-2}{1-a^2}} = \sqrt{p-2} .$$

The coverage probabilities of the Bayes sets of minimum volume are given in Table 3.3. They indicate that the minimum length is presumably much larger than $\sqrt{p-2}$.

Table 3.3. For selected dimensions p, values of the minimum length ℓ^* and coverage of the Bayes set (4.15) against the loss (4.18).

p	3	4	5	6	7	8	9
ℓ^*	2.24	2.66	3.00	3.27	3.51	3.74	3.96
Posterior coverage	.83	.87	.89	.90	.91	.92	.93

p	10	11	12	13	14	15	20
ℓ^*	4.15	4.32	4.49	4.65	4.80	4.96	5.65
Posterior coverage	.93	.93	.93	.94	.94	.94	.96

4. Relations Between the Loss and the Distribution

When we consider a given loss,

$$L(\theta, C) = \$\,\mathrm{vol}(C) + \mathbb{I}(\theta \notin C) , \tag{4.1}$$

one important problem is to determine how much this loss depends on the distribution. For instance, we can look at the conditions for admissibility or minimaxity of different sets, and see how these conditions depend on the distribution. Another subject of interest is the value of the minimal coverage probability over a class of losses, as treated in the previous section. We can ask if this lower bound depends on the distribution.

In general, there will be a relationship between the loss and the distribution. We can then try to restrict ourselves to the consideration of "robust" or "universal" losses. Although this notion is not easy to define, it is certain that we cannot define a loss that will be valid for any kind of problem. This problem is not unique to set estimation, but also is the case for point estimation problems. In point estimation problems, it is well known that different combinations of loss functions can give different "optimal" answers for the same distribution.

We show in Section 4.1 that a natural loss exists for bounded parameter spaces. This loss can be called "universal" in the sense that all Bayes sets are nontrivial, independent of the distribution of the observations. In the more general case of unbounded parameter spaces, some of the losses we have considered will not behave properly for more general models. For instance, the size function $\$^*(C) = P(T_\nu \in C)$ gives $\emptyset$ as a minimax set if the observations have a t distribution with $n < \nu$ degrees of freedom. The rational loss thus appears as the most appealing candidate.

4.1. Bounded Parameter Space

We now consider the case where X has a distribution depending on a parameter $\theta \in \Theta$ with $\mathrm{vol}(\Theta) = \Omega < +\infty$ (the volume being determined by Lebesgue measure). Example 2.3 is a particular case of this situation, and we still see that some of the results in Example 2.3 for the binomial distribution can be generalized to distributions with bounded parameter spaces. In fact, a natural loss in this setup is

$$L(\theta, c) = \frac{\mathrm{vol}(C)}{\Omega} + \mathbb{1}(\theta \notin C) , \tag{4.2}$$

as the volume function is increasing, bounded by one and satisfies $L(\theta, \emptyset) = L(\theta, \Theta) = 0$. Furthermore, the following theorem shows that the loss is independent of the distribution of the observations.

Theorem 4.1. *For any distribution* $f(x \mid \theta)$ *and any prior distribution* π *on* Θ*, the Bayes set is*

$$C_x^\pi = \left\{\theta : \pi(\theta \mid x) \geq \frac{1}{\Omega}\right\} .$$

Proof. In the case of the loss function (4.2), we have $S'(t) = 1/\Omega$. Thus, for a set of the form $C_x^\pi(k) = \{\theta : \pi(\theta \mid x) \geq k\}$ it follows from Lemma 2.1 that

$$\frac{\partial}{\partial k}\Big(L\Big(\theta, C_x^\pi(k)\Big)\Big) = \Big[k - \frac{1}{\Omega}\Big] \int_{\{\pi(\theta \mid x) = k\}} \frac{ds}{|\nabla \pi(\theta \mid x)|} . \tag{4.3}$$

The integral in (4.3) is always nonnegative. Furthermore, this integral has to be positive for some values of k, otherwise π would not be a density on Θ. Therefore, as k increases from 0, $\frac{\partial}{\partial k}\Big(L\Big(\theta, C_x^\pi(k)\Big)\Big)$ changes sign only once from negative to positive. This implies that $k = 1/\Omega$ gives a minimum of $L\Big(\theta, C_x^\pi(k) \mid x\Big)$, for any distribution $f(x \mid \theta)$. □

Theorem 4.2. *For any prior distribution,* $\pi(\theta)$*, the Bayes set associated with the loss function (4.2),* C_x^π*, is nontrivial.*

Proof. We first prove that the posterior distribution cannot be uniform. If $f(x \mid \theta)$ is the density of X with respect to a measure ν, we have $\pi(\theta \mid x) = f(x \mid \theta)\pi(\theta) \Big/ \int_\Theta f(x \mid \theta)\pi(\theta)d\theta$. If $\pi(\theta \mid x) = 1/\Omega$ for almost every x (with respect to ν), then $f(x \mid \theta)\pi(\theta) = (1/\Omega) \int_\Theta f(x \mid \theta)\pi(\theta)d\theta$. Integrating with respect to x gives

$$\begin{aligned} \pi(\theta) = \pi(\theta) \int_{\mathfrak{X}} f(x \mid \theta)d\nu(x) &= \frac{1}{\Omega} \int_{\mathfrak{X}} \int_\Theta f(x \mid \theta)\pi(\theta)d\theta d\nu(x) \\ &= \frac{1}{\Omega} \int_\Theta \int_{\mathfrak{X}} f(x \mid \theta)d\nu(x)\pi(\theta)d\theta = \frac{1}{\Omega} . \end{aligned}$$

But if $f(x \mid \theta)$ actually depends on θ, the posterior distribution associated with $\pi(\theta) = 1/\Omega$ is not uniform. Thus we have a contradiction.

As $\int_\Theta \pi(\theta \mid x)d\theta = 1$ and $\pi(\theta \mid x) \neq 1/\Omega$, it follows that the sets $\{\theta : \pi(\theta \mid x) > (1/\Omega)\}$ and $\{\theta : \pi(\theta \mid x) < (1/\Omega)\}$ are not equivalent to the empty set or the entire space. □

We can deduce from Theorem 4.2 that ϕ is not Bayes for (4.2) and, therefore, less likely to be admissible or minimax.

4.2. The Rational Loss

For the general case of unbounded spaces, we can no longer use the "natural" loss of Section 4.1. Closer examination of the bounded parameter space case helps us understand why the loss seems "universal." First, the linearity of the loss dictates that the tail behavior of the linear loss is appropriate. In fact, since the derivative of linear loss is constant, it follows that, as $x \to \infty$, every density function $f(x|\theta)$ will eventually be below this constant. This is the fact that leads to the universality of linear loss for bounded parameter spaces.

Secondly, only in bounded parameter spaces will the linear loss be coherent in the sense that $\mathbb{S}(0) = 0$, $\mathbb{S}(\infty) = 1$, and $0 < \mathbb{S}(v) < 1$ for $0 < v < \infty$. Thus, to try to carry these conditions over to the unbounded parameter space problem, we must consider a loss function whose derivative is not too small. A reasonable candidate is the family of 'rational' losses, since they avoid trivial Bayes sets for conjugate priors and for $p \geq 5$, the minimum coverage probability phenomenon is not of consequence in the normal case. An additional sufficient condition for nontriviality is given by the following result.

Theorem 4.3. *If* $\{\theta : \pi(\theta|x) > (1/a)\}$ *has positive Lebesgue measure, the Bayes rule* $C_x^\pi = \{\theta : \pi(\theta|x) > k\}$ *against the rational loss function is nontrivial, where* $k = k(x) < 1/a$.

Proof. See Casella, Hwang and Robert (1992).

That $\{\pi(\theta|x) > (1/a)\} = \phi$ does not necessarily imply that the Bayes rule is ϕ, as illustrated by the following example.

Example 4.1. Suppose $X \sim N_p(\theta, I)$ and $\theta \sim N_p(0, \tau^2 I)$. Then $\pi(\theta|x)$ is $N_p(\eta x, \eta I)$ and the Bayes set is $C_x^\pi = \{\theta : |\theta - \eta x|^2 \leq c\}$ where $\eta = \tau^2/(\tau^2+1)$. If a is such that $(1/a) > (2\pi\eta)^{-p/2}$, it follows that $\{\pi(\theta|x) > (1/a)\}$ is empty. However, as seen in Section 3.3, there is still a nonempty Bayes set.

5. Conclusion

This paper shows how difficult it is to treat the set estimation problem in a decision-theoretic way, since it points out that many losses, even the nonlinear losses like in (1.2), suffer from defects reproched to the original linear loss. It also provides some guidelines in the selection of the size function $\mathbf{S}$, in particular through the "coherency conditions" $\mathbf{S}(0) = 0$, $\mathbf{S}(\infty) = s$. Unfortunately, it appears that the distribution of the observations, $f(x|\theta)$, also plays an important role in the apperition of paradoxes and undesirable features, like the existence of trivial Bayes sets or the minimaxity of ϕ. This implies further robustness studies for specific families of losses and distributions.

Despite these limitations, we deeply recommend a decision-theoretic approach to confidence set estimation and this for three reasons. First, an independent approach separates the testing imperatives from the confidence imperative and allows us to evaluate the confidence procedures for themselves. Second, in a related way, it compels the decision-maker to evaluate the consequences of his/her actions related to the confidence procedure, thus repositions confidence procedures at the core of the inferential process instead of presenting them as secondary tools. Last, by allowing the use of classical decision-theoretic notions like admissibility, minimaxity, etc., this approach brings the decision-maker on a sounder ground since he/she can compare estimators and look for an optimal solution, as is the case in point estimation, instead of borrowing from resting perspectives or devising *ad hoc* optimality notions.

A deeper purpose of this paper is also to call for a reflection on the actual goals of confidence set estimation. Indeed, although the previous sections show that a decision-theoretic approach is possible for a careful choice of the loss function (1.3), the true purpose of decision theory is still to provide the decision-maker with a tool to help him to make rational/coherent decisions. Therefore, following De Groot (1970), Berger (1985) or Lindley (1965), this implies that the loss function he/she uses in this process is obtained from a corresponding utility function. It is yet unclear how this utility function can be constructed in practice, given the available choice of acceptable loss functions, but it is clear that the loss function should not be chosen for its mathematical properties or even for its compliance

with an "intuitive" behavior.

Borrowing from Berger (1985, Chap. 2), we propose the following approximation for the determination of the size function, assuming it belongs to a parametrized family $\mathbf{S}_a$, like the rational size $\mathbf{S}_a(v) = v/a + \sigma$. If the decision-maker is in fact able to compare the consequences of a large set with those of a smaller set, he can produce a sequence of volumes $v_1,\cdots,v_m$ and a corresponding sequence of weights $\rho_2,\cdots,\rho_m$ such that $\mathbf{S}_a(v_2) = \rho_2\mathbf{S}_a(v_1),\cdots,\mathbf{S}_a(v_m) = \rho_m\mathbf{S}_a(v_2)$. An approximate value of a can then be estimated based on this information.

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. Springer-Verlag.

Blyth, C. R. and Hutchinson, D. W. (1961). Tables of Neyman-Shortest Confidence Intervals for the Binomial Parameter. *Biometrika* **47**, 381-391.

Casella, G., Hwang, J. T-G., and Robert, C. P. (1992). A paradox in decision-theoretic interval estimation. *Statist. Sinica*, to appear.

Cohen, A. and Strawderman, W. E. (1973). Admissibility implications for different criteria in confidence estimation. *Ann. Statist.* **1**, 363-366.

De Groot, M. (1970). *Optimal Statistical Decisions.* John Wiley & Sons, New York.

Hwang, J. T. and Casella, G. (1982). Minimax confidence sets for the mean of a multivariate normal distribution. *Ann. Statist.* **10**, 868-881.

Hwang, J. T. and Casella, G. (1984). Improved set estimators for a multivariate normal mean. *Statistics and Decisions, Supplement Issue* **1**, 3-16.

Joshi, V. M. (1969). Admissibility of the usual confidence set for the mean of a univariate or bivariate normal population. *Ann. Math. Statist.* **40**, 1042-1067.

Lindley, D. (1985. *Making Decisions.* John Wiley & Sons, New York.

Meeden, G. and Vardeman, S. (1985). Bayes and admissible set estimation. *J. Amer. Statist. Assoc.* **80**, 465-471.

Appendix

Theorem (Generalization of Theorem 3.1). *For* $X \sim f(x \mid \theta)$ *and (possibly improper) prior* $\pi(\theta)$ *the Bayes estimator,* C_x^π, *against the loss*

$$L_S(\theta, C) = S[\mathrm{vol}(C)] - \mathbb{I}(\theta \in C) ,$$

satisfies

$$C_x^\pi \subset \{\theta : \pi(\theta \mid x) \geq k\} \quad \textit{or} \quad \{\theta : \pi(\theta \mid x) \geq k\} \subset C_x^\pi \tag{A.1}$$

for every $k > 0$. *Furthermore, if* C_x^π *satisfies*

$$\{\theta : \pi(\theta \mid x) \geq k_1\} \subset C_x^\pi \subset \{\theta : \pi(\theta \mid x) \geq k_2\}$$

where $k_1 > k_2$ *and* $\{\theta : \pi(\theta \mid x) = k\} = \emptyset$ *for* $k_2 < k < k_1$, *then there exists* $k^* \in [k_2, k_1]$ *such that*

$$C_x^\pi = \{\theta : \pi(\theta \mid x) \geq k^*\}$$

and k^* *minimizes*

$$S\Big(\mathrm{vol}\{\theta : \pi(\theta \mid x) \geq k\} - \int_{\{\theta : \pi(\theta \mid x) \geq k\}} \pi(\theta \mid x) d\theta .$$

Proof. Since C_x^π is a Bayes set it minimizes the posterior loss

$$L\left(C_x^\pi \mid x\right) = S[\mathrm{vol}(C)] - \int_C \pi(\theta \mid x) d\theta .$$

If (A.1) is not satisfied, there exists $k \geq 0$ such that

$$C_x^\pi \cap \{\theta : \pi(\theta \mid x) < k\} \neq \emptyset \quad \text{and} \quad \left(C_x^\pi\right)^c \cap \{\theta : \pi(\theta \mid x) \geq k\} \neq \emptyset ,$$

the intersections being different from zero (recall that we are working with sets defined only up to sets of Lebesgue measure zero). Therefore, there exists sets A_x and B_x such that

$$A_x \subset C_x^\pi \cap \{\theta : \pi(\theta \mid x) < k\}, \quad B_x \subset \left(C_x^\pi\right)^c \cap \{\theta : \pi(\theta \mid x) \geq k\}$$

and $\mathrm{vol}(A_x) = \mathrm{vol}(B_x) > 0$. If we now define $C_x^* = \left(C_x^\pi - A_x\right) \cup B_x$, it follows that

$$L\left(C_x^\pi \mid x\right) > L\left(C_x^* \mid x\right) ,$$

as $\mathrm{vol}\left(C_x^\pi\right) = \mathrm{vol}\left(C_x^*\right)$ and $\int_{A_x} \pi(\theta \mid x) d\theta < \int_{B_x} \pi(\theta \mid x) d\theta$. Thus we have a contradiction, so C_x^π must satisfy (A.1). The second part of the theorem follows immediately by continuity. □