BAYES ESTIMATION WITH CONVEX LOSS

By M. H. DeGroot[1] and M. M. Rao[2]

Carnegie Institute of Technology

1. Introduction and summary. Let X be a generalized random variable taking values in an abstract set $\mathfrak{X}$ on which is defined an appropriate σ-field of subsets. Suppose that the distribution of X depends on a real parameter Θ and that it is desired to estimate the value of Θ from an observation on X.

Let $W(\cdot)$ be a sufficiently smooth, non-negative, symmetric, convex function defined on the real line. Suppose that when the true value of Θ is θ and the estimated value is δ, the loss incurred is $W(\theta - \delta)$.

For a given prior distribution of Θ and any $x \varepsilon \mathfrak{X}$, let $F(\cdot \mid x)$ be the posterior distribution function of Θ when the observed value of X is x. A *Bayes estimate*, for the given value of x, is a number δ^* such that

$$\int_{-\infty<\theta<\infty} W(\theta - \delta^*)\, dF(\theta \mid x) = \inf_{-\infty<\delta<\infty} \int_{-\infty<\theta<\infty} W(\theta - \delta)\, dF(\theta \mid x). \tag{1.1}$$

Thus, for each given x, the problem of finding a Bayes estimate reduces to the problem of minimizing the integral

$$\int_{-\infty<\theta<\infty} W(\theta - \delta)\, dF(\theta), \tag{1.2}$$

where $F(\cdot)$ is a specified distribution function.

In Section 2 the solution of this minimization problem is presented and some properties of the minimizing values of δ are discussed. In Section 3 it is shown that a Bayes estimator $\delta^*(\cdot)$ satisfying (1.1) for all $x \varepsilon \mathfrak{X}$ can be chosen so that it is a measurable function of x. In Section 4 the question of evaluating the expectation of $W[\Theta - \delta^*(X)]$ is considered and lower bounds for this quantity are presented.

The special problem in which $W(\cdot)$ is of the form $W(t) = |t|^k$, $-\infty < t < \infty$, $k \geqq 1$, is considered in some detail. It is known (see e.g., [1], p. 302) that for $k = 1$ the integral (1.2) is minimized when δ is a median of the distribution function $F(\cdot)$, and for $k = 2$ it is minimized when δ is the mean of $F(\cdot)$. The solution of the minimization problem presented in Section 2 is a generalization of these familiar results.

2. The minimization problem. For any continuous convex function $g(\cdot)$ defined

Received May 21, 1962; revised March 1, 1963.

[1] This research was supported in part by the National Science Foundation under grants NSF-G 9662 and NSF-G 25087.

[2] This research was supported in part by the National Science Foundation under grant NSF-G 14832.

on the real line, let

$$g^R(t) = \lim_{\epsilon \to 0^+} [g(t + \epsilon) - g(t)]/\epsilon, \qquad -\infty < t < \infty, \tag{2.1}$$

and let $g^L(t)$ be defined similarly as the limit of the difference quotient on the right side of (2.1) when $\epsilon \to 0^-$. It is well-known that these right-hand and left-hand derivates exist. Furthermore, for each fixed t, the difference quotient on the right side of (2.1) is a non-decreasing function of ϵ for $\epsilon > 0$ and, hence, so also is $[g(t - \epsilon) - g(t)]/\epsilon$. These properties are needed in the following development.

Suppose that $W(\cdot)$ is a continuous non-negative, convex function defined on the real line such that $W(t) = W(-t)$ for all real numbers t. To avoid a triviality we assume that $W(\cdot)$ is not identically a constant. Furthermore, we assume that $W(\cdot)$ is differentiable at all values of t with the possible exception of $t = 0$.

Let $\xi(t)$ denote the derivative of $W(\cdot)$ at t for $t \neq 0$, and to be specific, define $\xi(0) = W^R(0)$.

Throughout this section, $F(\cdot)$ denotes a given distribution function on the real line. It is assumed that

$$\int_{-\infty<\theta<\infty} W(\theta - \delta)\, dF(\theta) < \infty, \qquad -\infty < \delta < \infty. \tag{2.2}$$

This, in turn, implies that

$$\begin{aligned} &\int_{\theta>\delta} \xi(\theta - \delta)\, dF(\theta) < \infty, \qquad -\infty < \delta < \infty, \\ &\int_{\theta<\delta} \xi(\delta - \theta)\, dF(\theta) < \infty, \qquad -\infty < \delta < \infty. \end{aligned} \tag{2.3}$$

The finiteness of the first integral in (2.3) is verified by noting that since $W(\cdot)$ is convex,

$$c \int_{\theta>\delta} \xi(\theta - \delta)\, dF(\theta) \leqq \int_{\theta>\delta} [W(\theta - \delta + c) - W(\theta - \delta)]\, dF(\theta) < \infty, \tag{2.4}$$

for any $c > 0$. The second result in (2.3) can be verified similarly.

Now, define

$$U(\delta) = \int_{-\infty<\theta<\infty} W(\theta - \delta)\, dF(\theta), \qquad -\infty < \delta < \infty. \tag{2.5}$$

In the remainder of this section we will characterize the set I of values of δ at which the function $U(\cdot)$ is minimized.

Since $W(\cdot)$ is convex (2.5) implies that $U(\cdot)$ is also convex. Furthermore, since $W(\cdot)$ is symmetric and not identically constant, then $\lim_{t\to\pm\infty} W(t) = \infty$, and it follows from (2.5) that $\lim_{\delta\to\pm\infty} U(\delta) = \infty$. Under these conditions, the results summarized in the next lemma are immediate.

LEMMA 1. *The set I of values of δ at which $U(\delta)$ is minimized is a non-empty,* *bounded, closed interval. For any real number* δ, $\delta \varepsilon I$ *if and only if* $U^R(\delta) \geqq 0$ *and* $U^L(\delta) \leqq 0$.

The next lemma gives expressions for $U^R(\cdot)$ and $U^L(\cdot)$ in terms of the derivative $\xi(\cdot)$ defined earlier in this section.

LEMMA 2. *For all* δ, $-\infty < \delta < \infty$,

$$\begin{aligned} U^R(\delta) &= \int_{\theta \leqq \delta} \xi(\delta - \theta)\, dF(\theta) - \int_{\theta > \delta} \xi(\theta - \delta)\, dF(\theta), \\ U^L(\delta) &= \int_{\theta < \delta} \xi(\delta - \theta)\, dF(\theta) - \int_{\theta \geqq \delta} \xi(\theta - \delta)\, dF(\theta). \end{aligned} \tag{2.6}$$

PROOF. Let $\{\epsilon_n ; n = 1, 2, \cdots\}$ be a decreasing sequence of positive numbers converging to 0. For any fixed number δ, consider the sequence of functions $\{G_n(\cdot); n = 1, 2, \cdots\}$ defined by

$$G_n(\theta) = [W(\theta - \delta - \epsilon_n) - W(\theta - \delta)]/\epsilon_n, \quad -\infty < \theta < \infty. \tag{2.7}$$

It follows from the discussion contained in the first paragraph of this section that this is a non-increasing sequence of functions, and hence, by the Lebesgue monotone convergence theorem,

$$\int_{-\infty<\theta<\infty} [\lim_{n\to\infty} G_n(\theta)]\, dF(\theta) = \lim_{n\to\infty} \int_{-\infty<\theta<\infty} G_n(\theta)\, dF(\theta). \tag{2.8}$$

However, (2.7) implies that

$$\lim_{n\to\infty} G_n(\theta) = -W^L(\theta - \delta), \tag{2.9}$$

and that the right-hand side of (2.8) is

$$\lim_{n\to\infty} [U(\delta + \epsilon_n) - U(\delta)]/\epsilon_n = U^R(\delta). \tag{2.10}$$

Thus, it follows from (2.8) that

$$U^R(\delta) = -\int_{-\infty<\theta<\infty} W^L(\theta - \delta)\, dF(\theta). \tag{2.11}$$

A similar argument, using the non-increasing sequence of functions $\{H_n(\cdot); n = 1, 2, \cdots\}$ defined by

$$H_n(\theta) = [W(\theta - \delta + \epsilon_n) - W(\theta - \delta)]/\epsilon_n, \quad -\infty < \theta < \infty, \tag{2.12}$$

yields

$$U^L(\delta) = -\int_{-\infty<\theta<\infty} W^R(\theta - \delta)\, dF(\theta). \tag{2.13}$$

Finally, it follows from the symmetry and differentiability of $W(\cdot)$ that $W^R(t) = W^L(t) = \xi(t)$ for all $t \neq 0$; that $\xi(t) = -\xi(-t)$ for $t \neq 0$; and that $W^R(0) = -W^L(0) = \xi(0)$. If we make use of these relations, together with the integrability assumed in (2.3), the Equations (2.11) and (2.13) become those given in the statement of the lemma, and the proof is complete.

Together, Lemmas 1 and 2 immediately yield

THEOREM 1. *The set of values of* δ *that minimize the function* $U(\cdot)$ *is a non-empty, bounded, closed interval* I *such that* $\delta \varepsilon I$ *if and only if*

$$\int_{\theta \leqq \delta} \xi(\delta - \theta)\, dF(\theta) \geqq \int_{\theta > \delta} \xi(\theta - \delta)\, dF(\theta) \tag{2.14}$$

and

$$\int_{\theta < \delta} \xi(\delta - \theta)\, dF(\theta) \leqq \int_{\theta \geqq \delta} \xi(\theta - \delta)\, dF(\theta). \tag{2.15}$$

We conclude this section with two corollaries that further characterize the interval I in some special situations.

COROLLARY 1. *If* $W(\cdot)$ *is strictly convex then* I *contains exactly one point.*

PROOF. If $W(\cdot)$ is strictly convex, then it can easily be shown from the definition (2.5) that $U(\cdot)$ is also strictly convex and, hence, its minimum can be attained at only one point.

COROLLARY 2. *Suppose that either* $\xi(0) = 0$ *or* $F(\cdot)$ *is continuous. Then* I *is the set of values of* δ *such that*

$$\int_{\theta < \delta} \xi(\delta - \theta)\, dF(\theta) = \int_{\theta > \delta} \xi(\theta - \delta)\, dF(\theta). \tag{2.16}$$

PROOF. It is readily seen that under either of the conditions of this corollary, the pair of inequalities, (2.14) and (2.15), reduces to (2.16).

3. Bayes estimators. The interpretation of the results of Section 2 in the context of Bayes estimation will now be given. For a given prior distribution of Θ, let π be the marginal distribution of the observation X. For any $x \varepsilon \mathfrak{X}$, the sample space, let $F(\cdot \mid x)$ be the posterior distribution function of Θ given that the observed value of X is x. Let D be the class of all estimators $\delta(\cdot)$ of Θ. That is, D is the class of all measurable, real-valued functions $\delta(\cdot)$ on $\mathfrak{X}$. The assumptions made about $W(\cdot)$ in Section 2 are still in force when $F(\cdot)$ is replaced by $F(\cdot \mid x)$ for any $x \varepsilon \mathfrak{X}$. Throughout the rest of the paper, all expectations are taken with respect to the joint distribution of Θ and X with the exception of those that are explicitly indicated as being conditional expectations given $X = x$.

A Bayes estimator is a function $\delta(\cdot) \varepsilon D$ that minimizes $E[W(\Theta - \delta(X))]$. In the remainder of this section we will give a characterization of the class of Bayes estimators.

For each $x \varepsilon \mathfrak{X}$ and $-\infty < \delta < \infty$, let $U(\delta \mid x)$ be defined by (2.5) with $F(\cdot \mid x)$ instead of $F(\cdot)$, and let $I(x)$ be the interval of values of δ for which $U(\cdot \mid x)$ is a minimum. Moreover, from our assumption on the existence of the conditional distributions it follows that for each real number δ, $U(\delta \mid \cdot)$ can be taken as a measurable function of x([2], p. 27). From now on we assume that this has been done; i.e., $U(\delta \mid \cdot)$ is measurable in x.

We will now show that for each $x \varepsilon \mathfrak{X}$ it is possible to choose $\delta(x) \varepsilon I(x)$ such that the resulting function $\delta(\cdot)$ is a measurable function of x, and hence $\delta(\cdot) \varepsilon D$. It then follows from the results of Section 2 that it is a Bayes estimator.

THEOREM 2. *For each $x \varepsilon \mathfrak{X}$, there exists a measurable function $\delta(\cdot)$ such that $\delta(x) \varepsilon I(x)$ where $I(x)$ is the nonempty closed interval of values of δ that minimize $U(\cdot \mid x)$. One such function is defined by taking $\delta(x)$ to be the left end point of $I(x)$.*

PROOF. For definiteness, we consider the choice of $\delta(\cdot)$ as indicated in the theorem. Now for each fixed $x \varepsilon \mathfrak{X}$, $U(\cdot \mid x)$ is a convex function with all of the properties listed in Lemma 1 and the paragraph preceding it. Since $\delta(x)$ is the left end point of $I(x)$, it follows that, for any real number c, $\delta(x) \leqq c$ if and only if $U^R(c \mid x) \geqq 0$.

Let $\{\epsilon_n, n = 1, 2, \cdots\}$ be a decreasing sequence of positive numbers converging to zero. Then

$$U^R(c \mid \cdot) = \lim_{n\to\infty}[U(c + \epsilon_n \mid \cdot) - U(c \mid \cdot)]/\epsilon_n . \tag{3.1}$$

Since each difference quotient on the right side of (3.1) is a measurable function of x, so also is $U^R(c \mid \cdot)$. Thus, for each real number c, $\{x:\delta(x) \leqq c\} = \{x: U^R(c \mid x) \geqq 0\}$ is a measurable subset of $\mathfrak{X}$, and hence $\delta(\cdot)$ is a measurable function of x.

Now that we have established the existence of at least one estimator $\delta(\cdot) \varepsilon D$ such that $\delta(x) \varepsilon I(x)$ for all $x \varepsilon \mathfrak{X}$, the next theorem follows immediately from Theorem 1 and its corollaries.

THEOREM 3. *An estimator $\delta(\cdot) \varepsilon D$ is a Bayes estimator if and only if it satisfies the following inequalities* a.e. (π):

$$\begin{aligned} &\int_{\theta \geqq \delta(x)} \xi(\theta - \delta(x))\, dF(\theta \mid x) \geqq \int_{\theta < \delta(x)} \xi(\delta(x) - \theta)\, dF(\theta \mid x), \\ &\int_{\theta > \delta(x)} \xi(\theta - \delta(x))\, dF(\theta \mid x) \leqq \int_{\theta \leqq \delta(x)} (\delta(x) - \theta)\, dF(\theta \mid x). \end{aligned} \tag{3.2}$$

Moreover, if $W(\cdot)$ is strictly convex there is, for each $x \varepsilon \mathfrak{X}$, a unique value $\delta(x)$ satisfying (3.2). *If $\xi(0) = 0$, the inequalities* (3.2) *reduce to the equation*

$$\int_{\theta > \delta(x)} \xi(\theta - \delta(x))\, dF(\theta \mid x) = \int_{\theta < \delta(x)} \xi(\delta(x) - \theta)\, dF(\theta \mid x). \tag{3.3}$$

This theorem characterizes the Bayes estimators $\delta^*(\cdot)$. The value of the Bayes risk (i.e., the minimum expected loss that can be attained) is also of interest and in the next section we consider the problem of obtaining lower bounds for this quantity.

4. Lower bounds for the Bayes risk. Of central importance in the derivation of lower bounds for the Bayes risk is the following class M of integrable functions $m(\cdot)$ defined on the product space of Θ and X:

$$M = \{m(\cdot):E[m(\Theta, X) \mid x] = 0 \text{ a.e.}(\pi)\}. \tag{4.1}$$

The next result, which provides a useful class of lower bounds, appears as Theorem 7 in [3] in a somewhat different context. The utilization of this result in the problem being considered here requires essentially nothing more than a change in notation and viewpoint. A sketch of the proof is given for completeness.

THEOREM 4. *Suppose that for some value of $k(k > 1)$ the function $W^{1/k}(\cdot)$ is convex. Let $k' = k/(k-1)$. Then for any $m(\cdot) \varepsilon M$ such that $0 < E[|m(\Theta, X)|^{k'}] < \infty$ and any estimator $\delta(\cdot) \varepsilon D$,*

$$E(W(\Theta - \delta(X))) \geqq W[E(\Theta m(\Theta, X))/E(|m(\Theta, X)|)] \cdot [E(|m(\Theta, X)|)/E^{1/k'}(|m(\Theta, X)|^{k'})]^k. \tag{4.2}$$

For $k = 1$, (4.2) still holds if $E^{1/k'}(|m(\Theta, X)|^{k'})$ is interpreted as the essential supremum of $|m(\Theta, X)|$.

PROOF. Let $\bar{W}(\cdot) = W^{1/k}(\cdot)$. Then $\bar{W}(\cdot)$ is convex and has the properties that $\bar{W}(t)$ is non-decreasing for $t \geqq 0$ and $\bar{W}(t) = \bar{W}(-t)$ for all t. It follows from these properties and Jensen's inequality that

$$\begin{aligned} &E[|m(\Theta, X)|\bar{W}(\Theta - \delta(X))] \\ &\geqq E(|m(\Theta, X)|)\bar{W}[E(|(\Theta - \delta(X))m(\Theta, X)|)/E(|m(\Theta, X)|)] \\ &\geqq E(|m(\Theta, X)|)\bar{W}[E[(\Theta - \delta(X))m(\Theta, X)]/E(|m(\Theta, X)|)]. \end{aligned} \tag{4.3}$$

But, since $m(\cdot) \varepsilon M$,

$$E[(\Theta - \delta(X))m(\Theta, X)] = E[\Theta m(\Theta, X)] - E[\delta(X)E(m(\Theta, X)|X)] = E[\Theta m(\Theta, X)]. \tag{4.4}$$

Thus,

$$E[|m(\Theta, X)|\bar{W}(\Theta - \delta(X))] \geqq E(|m(\Theta, X)|)\bar{W}[E(\Theta m(\Theta, X))/E(|m(\Theta, X)|)]. \tag{4.5}$$

For $k > 1$, the Hölder inequality states that

$$E[|m(\Theta, X)|\bar{W}(\Theta - \delta(X))] \leqq E^{1/k}[\bar{W}^k(\Theta - \delta(X))]E^{1/k'}[|m(\Theta, X)|^{k'}]. \tag{4.6}$$

Since $\bar{W}^k(\cdot) = W(\cdot)$, the desired result (4.2) follows immediately from (4.5) and (4.6). For $k = 1$, $\bar{W}(\cdot) = W(\cdot)$ and, again, the desired result follows from (4.5) and the appropriate modification of (4.6).

A useful property of (4.2) is that the lower bound is valid for all estimators $\delta(\cdot) \varepsilon D$ and does not involve $\delta(\cdot)$. However, one important question raised by this result is that of choosing $m(\cdot) \varepsilon M$ and k so that one obtains the best possible lower bound. A helpful result (as shown in [3]) is that for any fixed $m(\cdot) \varepsilon M$, the right-hand side of (4.2) is a non-decreasing function of k. The conditions for equality in (4.2) are relatively complicated. We will forego further discussion at this level of generality and turn to the special but important situation where

$$W(t) = |t|^k, \qquad -\infty < t < \infty, \tag{4.7}$$

for some value of $k \geqq 1$. Specializing the above theorem for this $W(\cdot)$, and utilizing the well-known condition for equality in the Hölder inequality, yields

THEOREM 5. *Suppose $k > 1$ and let $k' = k/(k-1)$. Then, for any $m(\cdot) \in M$ such that $0 < E[|m(\Theta, X)|^{k'}] < \infty$ and any estimator $\delta(\cdot) \in D$,*

$$E(|\Theta - \delta(X)|^k) \geqq [|E[\Theta m(\Theta, X)]|/E^{1/k'}[|m(\Theta, X)|^{k'}]]^k. \tag{4.8}$$

Equality holds in (4.8) if and only if there exists a constant c such that, for almost all values of θ and x, $|\theta - \delta(x)|^k = c\,|m(\theta, x)|^{k'}$ and $(\theta - \delta(x))m(\theta, x)$ is of constant sign. For $k = 1$, (4.8) still holds if $E^{1/k'}[|m(\Theta, X)|^{k'}]$ is interpreted as the essential supremum of $|m(\Theta, X)|$.

Through use of the condition for equality given in Theorem 5 and the known form of the Bayes estimators, it can be shown that the bounds given in (4.8) is always attained by the Bayes estimator with an appropriate choice of $m(\cdot)$ from M. This result is summarized in the next theorem. Its proof is omitted.

THEOREM 6. *Suppose $k \geqq 1$. Let k' and the right-hand side of (4.9) below be as defined in Theorem 5. Then there exists a function $m^*(\cdot) \in M$ such that*

$$\inf_{\delta(\cdot)\in D} E[|\Theta - \delta(X)|^k] = \{|E[\Theta m^*(\Theta, X)]|/E^{1/k'}[|m^*(\Theta, X)|^{k'}]\}^k. \tag{4.9}$$

It is easy to construct functions belonging to the class M. Indeed, if $g(\cdot)$ is any integrable function of θ and x then $m(\theta, x) = g(\theta, x) - E[g(\Theta, X) \mid x]$ defines a function in M. However, it would be of great interest to construct functions $m(\cdot) \in M$ that maximize the right-hand side of (4.8), without making explicit use of the Bayes estimator $\delta^*(\cdot)$. We will now briefly illustrate how this might be done.

Suppose that there exists a real-valued sufficient statistic X with an absolutely continuous distribution. Let $p(\cdot)$ denote the prior density function of Θ, $\varphi(\cdot \mid \theta)$ the conditional density function of X for each given θ, and $f(\cdot \mid x)$ the posterior density of Θ for each given $X = x$. Then, under the standard regularity conditions justifying the interchange of differentiation and integration, it follows that

$$\frac{\partial \log f(\theta \mid x)}{\partial x} = \frac{\partial \log \varphi(x \mid \theta)}{\partial x} - \frac{\int_{-\infty}^{\infty} [\partial\varphi(x \mid \theta)/\partial x]\, p(\theta)\, d\theta}{\int_{-\infty}^{\infty} \varphi(x \mid \theta)\, p(\theta)\, d\theta}, \tag{4.10}$$

and if $m(\theta, x)$ is defined as either side of (4.10), then $m(\cdot) \in M$. If $f(\theta \mid x)$ is of the form $\beta(x)e^{\theta x}h(\theta)$, then $m(\theta, x) = \theta + \beta'(x)/\beta(x) = \theta - E(\Theta \mid x)$. When $k = 2$, it follows from Theorem 5 and the well-known fact that $\delta^*(\cdot)$ given by $\delta^*(x) = E(\Theta \mid x)$ is a Bayes estimator, that this $m(\cdot)$ maximizes the right-hand side of (4.8).

5. Acknowledgment. We are indebted to the referee, whose comments led to a substantial revision of an earlier version of this paper. In particular, the present format of Sections 2 and 3 has resulted from his suggestions.

REFERENCES

[1] Blackwell, David and Girshick, M. A. (1954). *Theory of Games and Statistical Decisions*. Wiley, New York.

[2] Doob, J. L. (1953). *Stochastic Processes*. Wiley, New York.

[3] Rao, M. M. (1961). Theory of lower bounds for risk functions in estimation. *Math. Ann.* **143** 379–398.