The Annals of Statistics
1985, Vol. 13, No. 2, 626–637

ROBUST BAYES DECISION PROCEDURES: GROSS ERROR IN THE DATA DISTRIBUTION[1]

BY YA'ACOV RITOV

The Hebrew University of Jerusalem

We consider a standard Bayes decision situation except that with small probability the data may be irrelevant to the parameter of interest (i.e. the experiment is "biased"). The minimax solution under quite general assumptions is described and discussed.

1. Introduction.

Consider the following situation. A manager of a company gets some information about his competitor's company (company B). The information is indirectly related to what he is really concerned about. For example, suppose he wants to know how many devices of a particular type will be assembled by company B, while his information X is how many of a type of transistors commonly used in these types of devices were ordered by company B from company C. Suppose our manager has previous independent information and subjective conjectures about the problem, all formulated in a Bayesian fashion. If this is the case, he takes the quantity of interest, θ, to have a well-defined (for him!) distribution. He believes that he knows how X is distributed for any given value of θ. Finally he defines a loss function.

For the Bayesian, the solution of this problem is, theoretically speaking, straightforward. However, suppose the statistician is willing to consider the possibility that the information is irrelevant, i.e. the prior opinion about model linking information to parameters of interest may be wrong. In the above example, it may be that the transistors sold by company C were actually purchased by other companies for other purposes. It is then appropriate to assume that the distribution of X may be arbitrary and independent of θ.

Let us be more formal. Suppose we observe X and have to choose an action "a" out of a set of possible actions $\mathbf{A}$. There is a third variable Θ which defines both the distribution of X and the "loss" caused by using "a". We take Θ to have a well-defined distribution (defined either as a limit of frequencies or a subjective distribution). The distribution of X is defined

(i) with probability $1 - \varepsilon$, by the known prior and a known Markov kernel.

(ii) with probability ε, by an unknown arbitrary distribution H.

Typically ε will be a small number.

Received May 1983; revised August 1984.

[1] Part of this work was done while the author was at the University of California, Berkeley. The work was partially supported by the Lady Davis Fellowship Trust and by Office of Naval Research Grant N00014-80-C-0163.

AMS 1980 *subject classifications.* Primary 62F35, secondary 62C99.

Key words and phrases. Robust Bayes, bias, minimax.

EXAMPLES.

(i) We have a real "one observation" case, and we are afraid of a "gross" error, e.g. a typographical error, in the data.

(ii) *Measurement of irrelevant parameters.* The assumption that H (as defined above) can be arbitrary excludes all cases where there are structural constraints on H, e.g., when given $\Theta = \theta$ the observations are i.i.d. Nevertheless, we claim that the above formulation is a reasonable approximation to the following situation. Suppose we want to estimate the mean I.Q. of a particular population. Here the I.Q. is defined by a standard test, but as this test is time- and money-consuming, we use another quick test. It is known that for most populations the mean scores in the "quick" and "standard" tests are approximately the same. If we will take a large sample, the average "quick" score may be a reasonable estimate of the mean of the population I.Q. as measured by the quick test. The main problem is that this population is quite special, e.g. students in an art academy, and it may be that these two tests are unrelated for this particular population. If the sample is large we may assume, therefore, that the sample average (which is approximately the population mean) may have, with a small probability an "almost" arbitrary distribution.

This situation exemplifies a quite general situation, when the main problem is whether we are measuring the "true thing" or another parameter which is usually, but not always, closely related to the "true-parameter of interest" (i.e. the experiment may be biased).

(iii) *i.i.d. observations with a common error.* This example is, mathematically speaking, similar to the previous one, although the mechanism is different. In telemetry we may fear that our remote instrument sends us all n-observations with a consistent error (e.g. in the main digit).

(iv) *i.i.d. observations with a contaminated model.* Here the formulation of our problem is clearly inadequate, as H has to be a product distribution on R^n. When n is small, the solution appropriate to taking H arbitrary may be a reasonable approximation which can be traced analytically; see Marazzi (1980).

REMARK. The case where Θ is a location parameter and $\mathscr{L}\{X - \Theta\} \in \{(1-\varepsilon)F + \varepsilon H;\ H \text{ arbitrary}\}$ for a known c.d.f. F, is mathematically equivalent to the same problem but with a known model and an ε-contaminated prior (Ritov, 1983).

The Bayesian decision procedure is based upon the knowledge of the prior distribution Π and the model $\{F_\theta\}$. For the "orthodox Bayesian" they are completely known (see for example De Finetti, 1961). On the other hand, there exists a robust Bayesian viewpoint. Much work was done about ε-contaminated priors and other types of uncertainties in the prior distribution. See for example Hodges and Lehmann (1952), Blum and Rosenblatt (1967), Berger and Berliner (1983), Efron and Morris (1971), Marazzi (1980, 1982) and Ritov (1983). There are, however, situations where the prior information is defined in quite a reliable way, while it is felt that the relation between the data and the parameter is not given exactly by the specified Markov kernel. One possible way to deal with this problem is making the model more complex (i.e. introducing more parameters and priors; see, for example, Box and Tiao, 1973). The analysis becomes then more complex and in some sense more arbitrary. We feel that our assumption that with a small probability the data is completely irrelevant, although it may be arbitrary in some cases, it may describe other situations quite adequately. (Sure, it may be inconvenient to admit that this is the case!) See Berger (1979) for discussion of a similar situation when Θ can get a finite number of values. The real test of a statistical attitude is the procedures it generates. We believe that our results are simple and intuitive modifications of the basic Bayes procedure. Our result can be compared to Box (1980). When the observation is surprising given our (unexact) assumptions, one should examine the assumptions. Unlike Box (1980) we take the prior to be reliable, so we should be careful in the way we use the data, when it is in discrepancy (in a defined way) with the prior.

The paper continues with a formal definition of the problem and the intuitive solution. In the third section the main results are given with some examples. The outline of the proofs is given in the fourth section.

2. Notation and problem definition. We consider a decision problem with three Polish spaces and Borel fields: $\boldsymbol{\Theta}$, $B(\boldsymbol{\Theta})$ the parameter space, $\mathbf{X}$, $B(\mathbf{X})$ the sample space, and, finally the decision space $\mathbf{A}$, $B(\mathbf{A})$. Alternatively we will speak about the random variables Θ and X^* taking values in $\boldsymbol{\Theta} \times \mathbf{X}$ and having a joint distribution defined by the "prior" Π—a probability measure on $\boldsymbol{\Theta}$, $B(\boldsymbol{\Theta})$ and the Markov kernel $\{F_\theta: \theta \in \boldsymbol{\Theta}\}$. Let F be the marginal distribution of X^*, and assume $F_\theta \ll F$ for all θ. Finally, let $\Pi_x(\cdot)$ be the conditional distribution of Θ given $X^* = x$. To complete the definition of the decision situation, we need a measurable loss function L: $\boldsymbol{\Theta} \times \mathbf{A} \to [0, \infty)$. We assume that $\mathbf{A}$, $B(\mathbf{A})$ has a compactification $\mathbf{A}^k$, $B(\mathbf{A}^k)$ and L has an extension to $\boldsymbol{\Theta} \times \mathbf{A}^k$ such that $L(\theta, \cdot)$ is lower semicontinuous for all $\theta \in \boldsymbol{\Theta}$.

A decision procedure δ is a Markov kernel from $\mathbf{X}$ to $\mathbf{A}^k$. For any $x \in \mathbf{X}$, $\delta(\cdot \mid x)$ is a probability measure on $\mathbf{A}$, while $\delta(U \mid \cdot)$ is a measurable function of x for any $U \in B(\mathbf{A}^k)$.

Suppose now that the actual observation X is equal to X^* only with probability $1 - \varepsilon$ $(0 < \varepsilon < 1)$, while with probability ε it is taken from a completely unknown distribution H, i.e.

$$X \sim (1 - \varepsilon)F + \varepsilon H, \quad H \in \mathscr{P} = \{\text{all distribution on } B(\mathbf{X})\}.$$

Π, $\{F_\theta; \theta \in \boldsymbol{\Theta}\}$, H and δ together define the mean risk:

$$L_\varepsilon(\delta, H) = \int \int \int L(\theta, a)\delta(da \mid x)\{(1 - \varepsilon)F_\theta(dx) + \varepsilon H(dx)\}\Pi(d\theta).$$

We want to solve:

P$_1$: Find $\delta_\varepsilon \in \mathbf{D}$ such that

$$\sup_{H\in\mathscr{P}} L_\varepsilon(\delta_\varepsilon, H) = \inf_{\delta\in D}\sup_{H\in\mathscr{P}} L_\varepsilon(\delta, H)$$

where $\mathbf{D}$ is the set of all possible decision procedures.

We can use Fubini's theorem to simplify the definition of $L_\varepsilon(\delta, H)$:

$$(2.1)\quad L_\varepsilon(\delta, H) = (1-\varepsilon)\int r[\delta(\cdot\mid x), x]F(dx) + \varepsilon\int r_0[\delta(\cdot\mid x)]H(dx)$$

where:

$$r[\delta(\cdot\mid x), x] = \int\int L(\theta, a)\delta(da\mid x)\Pi_x(d\theta)$$

$$r_0[\delta(\cdot\mid x)] = \int\int L(\theta, a)\delta(da\mid x)\Pi(d\theta).$$

We can use these formulas to attack the problem from another direction. Take the manager's problem as it is defined in the beginning of the paper. It seems reasonable not to base on the suspected observation a decision which is too risky. Here the riskiness of a decision procedure is measured by the prior risk, or formally, by $r_0[\delta(\cdot\mid\cdot)]$. Hence, the following problem is natural:

P$_2$: Find $\delta \in D(s) = \{$all decision procedures δ such that $\sup_x r_0[\delta(\cdot\mid x)] \le s\}$ such that $\int r[\delta(\cdot\mid x), x]F(dx)$ is minimized.

Actually here we have finished. Leaving measurability problems aside, solving P$_1$ amounts of solving P$_2$ for some s. To see this, note that if $\sup_{H\in\mathscr{P}} L_\varepsilon(\delta, H) < \infty$ then $\sup_x r_0[\delta(\cdot\mid x)] \le s < \infty$ for some s, as H is arbitrary. Conversely, if P$_2$ can be solved for all s, we denote each solution by δ_s, and to solve P$_1$ we will minimize $(1-\varepsilon)\int r[\delta_s(\cdot\mid x), x]F(dx) + \varepsilon s$. The advantage of P$_2$ is that it can be solved pointwise, i.e. for each $x \in \mathbf{X}$ one minimizes $r[\delta(\cdot\mid x), x]$ subject to $r_0[\delta(\cdot\mid x)] \le s$. Typically this minimization problem can be solved using Lagrange multipliers.

Remark. With some abuse of notation, we will denote nonrandomized procedures by small latin letters, so $a(\cdot)$ is the procedure δ such that $\delta(\cdot\mid x)$ is a point mass at $a(x)$.

To avoid trivialities we assume that $0 < s_0 = \inf\{s: D(s) \ne \phi\} < \infty$. Moreover, $D(s_0)$ contains no Bayes procedure (for $\varepsilon = 0$, i.e. for the "basic" model). Clearly, when one tries to solve either P$_1$ or P$_2$, one restricts himself to the "safe" action set, i.e. to the set:

$$A_0 = \{a: a \in \mathbf{A} \text{ and } E_\Pi L(\theta, a) < \infty\}$$

(as one should use decisions with $r_0[\delta(\cdot\mid x)] < \infty$ for all $x \in \mathbf{X}$). Therefore, without any loss of generality (with regard to the set of possible decision procedures) we assume that $\mathbf{A} = \mathbf{A}_0$. We conclude this section with two more technical definitions:

For a $\delta \in \mathbf{D}$ we denote the Bayes risk of δ under the basic model by

$$L_0(\delta) = \int r[\delta(\cdot \mid x), x]F(dx).$$

For any s we define

$$R(s) = \inf_{\delta \in D(s)} L_0(\delta).$$

3. Main results.

THEOREM 1. *For all $s \geq s_0$ there is a decision procedure which solves P_2.* □

THEOREM 2. *For any $0 < \varepsilon < 1$, P_1 has a solution. Moreover, $L_\varepsilon(\cdot\,,\,\cdot)$ has a saddle point, i.e., there are δ_ε, H_ε such that*

$$L_\varepsilon(\delta_\varepsilon, H_\varepsilon) = \inf_{\delta \in \mathbf{D}} L_\varepsilon(\delta, H_\varepsilon) = \sup_{H \in \mathscr{P}} L_\varepsilon(\delta_\varepsilon, H)$$

and $H_\varepsilon \ll F$. □

So, let us define $\bar{L}(\varepsilon) = L_\varepsilon(\delta_\varepsilon, H_\varepsilon)$.

The existence of a least favorable distribution gives us some measure of how conservative the min-max procedure δ_ε is. If H_ε is a "reasonable"distribution, it is reasonable to be "protected" from it. Technically, this means that δ_ε is a Bayes procedure when the prior is Π and the family of distribution is $\{(1-\varepsilon)F_\theta + \varepsilon H_\varepsilon;\ \theta \in \Theta\}$. Moreover, if $h_\varepsilon = dH_\varepsilon/dF$, then for all x, δ_ε is a minimum point of $(1-\varepsilon)r[\delta(\cdot \mid x), x] + \varepsilon h_\varepsilon(x) r_0[\delta(\cdot \mid x)]$. We will be able to use these facts to conclude that: (i) P_1 and P_2 are "equivalent" and (ii) P_2 (and therefore P_1) can be solved pointwise. That is, we can solve for each x: $r[\delta(\cdot \mid x), x] = \min!$ subject to $r_0[\delta(\cdot \mid x)] \leq s$ by using a Lagrange multiplier (which is equal to $\varepsilon h_\varepsilon(x)/(1-\varepsilon)$!). Let us be more exact.

THEOREM 3. *δ solves P_1 for ε, $0 < \varepsilon \leq 1$ if and only if δ solves P_2 for some s, $s_0 \leq s < \infty$.*

THEOREM 4. (i) *It is possible to solve P_2 ($s_0 < s$) by solving the following problem for every $x \in C$, where C is a measurable set and $F(C) = 1$: Find $\lambda_s(x)$, $0 \leq \lambda_s(x) < \infty$ and $\delta_s(\cdot \mid x)$ a probability measure on $\mathbf{A}^k$ such that: $\delta_s(\cdot \mid x)$ minimizes $r[\delta_s(\cdot \mid x), x] + \lambda_s(x) r_0[\delta_s(\cdot \mid x)]$ over $\mathbf{D}$; $r_0[\delta_s(\cdot \mid x)] \leq s$ and $\lambda_s(x)\{r_0[\delta_s(\cdot \mid x)] - s\} = 0$. Extend $\delta_s(\cdot \mid \cdot)$ to $\mathbf{X} - C$ so it will be equal there to any member of $D(s)$. It is possible to find $\lambda_s(\cdot)$ and $\delta_s(\cdot \mid \cdot)$ such that $\lambda_s(\cdot)$ is measurable and $\delta_s \in \mathbf{D}$.*

(ii) *Any $\lambda_s(\cdot)$ which satisfies* (i) *satisfies also $\int \lambda_s(x)F(dx) < \infty$.*

(iii) *δ_s solves P_1 for ε given by $\varepsilon/(1-\varepsilon) = \int \lambda_s(x)F(dx)$ (with $dH\varepsilon/dF = (1-\varepsilon)\varepsilon^{-1}\lambda_s$). Alternatively, find ε by solving $-R'_+(s) \leq \varepsilon/(1-\varepsilon) \leq -R'_-(s)$ where $R'_\pm(\cdot)$ are the right and left derivatives of $R(\cdot)$.* □

REMARK. The use of a randomized procedure may be unavoidable. For example, this will be the case when we have a normal prior, $X - \Theta \sim N(0, 1)$ and the loss function is

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

Then for large values of x, $\delta_s(\cdot \mid x)$ is a randomized procedure which puts mass of $\alpha_s \simeq (s - 2\Phi(-1))/(1 - 2\Phi(-1))$ near $x/2$ and mass of $1 - \alpha_s$ near 0 (the a priori decision). For more details see Ritov (1983).

Suppose $L(\theta, \cdot)$ is a convex function for any $\theta \in \boldsymbol{\Theta}$. In this case there is no need for randomization. Theorem 4 defines then the form of the solution. Let $b(\cdot)$ be the Bayes rule (i.e. the solution of P_1 for $\varepsilon = 0$). For a fixed ε, the optimal decision will be the Bayes rule, on $r_0[b(x)] \le s$ for some s. Elsewhere the solution will satisfy $r_0[a(x)] = s$. When the action space is R we can be more specific and claim that the intuitive rule

$$a(x) = \begin{cases} M_1 & b(x) < M_1 \\ b(x) & M_1 \le b(x) \le M_2 \\ M_2 & b(x) > M_2 \end{cases}$$

is optimal for some interval $[M_1, M_2]$ which includes the minimum point of $E_\Pi L(\Theta, a)$ (the "no observation decision"). For a quadratic loss function we may be more specific:

COROLLARY 4.1. *Let* $\boldsymbol{\Theta} = \mathbf{A}_0 = R^k$ *and let* $L(\theta, a) = g(\theta)(\theta - a)^T Q(\theta - a)$, *where* Q *is a symmetric positive definite matrix of order* $k \times k$ *and* $g(\cdot)$ *is a positive real function such that* $\int \|\theta\|^2 g(\theta)\Pi(d\theta) < \infty$, $\int \theta g(\theta)\Pi(d\theta) = 0$ *and* $\int g(\theta)\Pi(d\theta) = 1$. *Let* $b(\cdot)$ *be the nonrandomized Bayes rule and*

$$V(M) = \{x \colon x \in \mathbf{X},\ b^T(x)Qb(x) \le M^2\}.$$

Let $0 < \varepsilon < 1$ *and define* $M(\varepsilon)$ *by*

$$\varepsilon/(1 - \varepsilon) = \int_{x \notin V(M(\varepsilon))} [(b^T(x)Qb(x))^{1/2}/M(\varepsilon) - 1]F(dx). \tag{3.1}$$

Then P_1 *is solved for* ε *by*

$$a_\varepsilon(x) = \begin{cases} b(x) & x \in V(M(\varepsilon)) \\ b(x)M(\varepsilon)/[b^T(x)Qb(x)]^{1/2} & x \notin V(M(\varepsilon)) \end{cases}$$

and the least favorable distribution H_ε *is given by*

$$H_\varepsilon(dx) = (1 - \varepsilon)/\varepsilon[(b^T(x)Qb(x))^{1/2}/M(\varepsilon) - 1]I\{x \notin V(M(\varepsilon))\}F(dx). \quad \square$$

REMARK. The definition of $M(\varepsilon)$ is legitimate as the RHS of (3.1) is monotone and continuous.

PROOF OF THE COROLLARY. The proof of the corollary follows by direct calculations if we note that

$$r[a(x), x] = [a(x) - b(x)]^T Q[a(x) - b(x)] + r[b(x), x]$$

$$r_0[a(x)] = a^T(x)Qa(x) + \int \theta^T Q\theta g(\theta)\Pi(d\theta). \quad \square$$

REMARK. Marazzi (1980) obtained this solution for a normal prior and quadratic loss function.

EXAMPLE 1. Let $\mathbf{A} = \{a_0, a_1\}$. Let ε, $0 < \varepsilon < 1$ be fixed and $f_\theta(\cdot)$ be the model density relative to some common measure. Suppose $\int L(\theta, a_0)\Pi(d\theta) < \int L(\theta, a_1)\Pi(d\theta)$. Then the "no observation" decision is a_0. Let

$$\mathscr{X}_1 = \left\{x: \int L(\theta, a_0)f_\theta(x)\Pi(d\theta) > \int L(\theta, a_1)f_\theta(x)\Pi(d\theta)\right\}$$

(i.e. the region in $\mathbf{X}$ where the Bayes rule $b(x)$ should be equal to a_1). Let $h(\cdot)$ be the least favorable density. Clearly it concentrates on $\mathscr{X}_1$. Now, after observing $X = x$ we should decide a_0 if

$$(1 - \varepsilon)\int L(\theta, a_0)f_\theta(x)\Pi(d\theta) + \varepsilon h(x)\int L(\theta, a_0)\Pi(d\theta)$$

$$< (1 - \varepsilon)\int L(\theta, a_1)f_\theta(x)\Pi(d\theta) + \varepsilon h(x)\int L(\theta, a_1)\Pi(d\theta)$$

and a_1 if the inequality holds in the opposite direction. Define ε_0 by

$$\frac{\varepsilon_0}{1 - \varepsilon_0} = \frac{\int [L(\theta, a_0) - L(\theta, a_1)]F_\theta(\mathscr{X}_1)\Pi(d\theta)}{\int [L(\theta, a_1) - L(\theta, a_0)]\Pi(d\theta)}. \tag{3.2}$$

Suppose $0 < \varepsilon < \varepsilon_0$. Then the solution according to Theorem 4 will be given by $s = \int L(\theta, a_1)\Pi(d\theta)$,

$$h(x) = \frac{1 - \varepsilon_0}{\varepsilon_0}\frac{\int [L(\theta, a_0) - L(\theta, a_1)]f_\theta(x)\Pi(d\theta)}{\int [L(\theta, a_1) - L(\theta, a_0)]\Pi(d\theta)}, \quad x \in \mathscr{X}_1$$

and the "conservative" decision is just the Bayes decision. The opposite extreme case will be when $\varepsilon > \varepsilon_0$. In that case one should ignore the observed X and use the "no observation" decision—a_0. In particular, if the denominator in (3.2) is zero then $\varepsilon_0 = 1$, i.e., if there is a priori "indifference" between a_0 and a_1 then we should always use the Bayes rule. To be more specific, suppose Θ has a $N(\theta_0, \tau^2)$, $\theta_0 > 0$ a priori distribution while $X - \Theta$ has a $N(0, \sigma^2)$ distribution. As a loss function we take $L(\theta, a_0) = -\theta + p\theta^2$ and $L(\theta, a_1) = \theta + p\theta^2$ for a small value of p. (The $p\theta^2$ term was added for the sake of the technical definition of the loss function as a bounded from below function and can be ignored from any practical point of view. This loss function may seem appropriate, e.g. if we have to choose between two varieties of apples and θ is the difference between their mean yields.)

For this example we get that $\mathscr{X}_1 = \{x: x < -\theta_0\sigma^2/\tau^2\}$ and after some calculations:

$$\frac{\varepsilon_0}{1-\varepsilon_0} = -\theta_0^{-1}\tau^{-1} \int \theta\Phi(-\sigma^1(\theta + \theta_0\sigma^2/\tau^2))\varphi(\tau^{-1}(\theta - \theta_0))\, d\theta$$
$$= k^{-1}\varphi(k) - \Phi(-k)$$

where $k = (1 + \sigma^2/\tau^2)^{1/2}\theta_0/\tau$, $\Phi(\cdot)$ the standard normal c.d.f. and $\varphi(\cdot)$ its density. Typically when one does an experiment like that σ^2 is much smaller than τ^2 while θ_0/τ is relatively small—otherwise the experiment is not "needed". Therefore, k is relatively small and ε_0 is not too small. On the other hand, if the experimenter has relatively high a priori confidence that Θ is positive (i.e. k is large) then the data is useful only when he is quite sure that the experiment is relevant to his "real problem". (For $k = 0.5, 1.0, 1.5, 1{,}96$ we get $\varepsilon_0 = 0.28, 0.077, 0.019$ and 0.0048 respectively.)

EXAMPLE 2. Suppose $\boldsymbol{\Theta} = \mathbf{A} = \mathbf{X} = R^k$, $\Theta \sim N(0, \sum_\theta)$, $X - \Theta \sim N(0, \sum_x)$ and $L(\theta, a) = \| \theta - a \|^2$. Then by Corollary 4.1 we ought to use the Bayes rule: $b(x) = \sum_\theta (\sum_x + \sum_\theta)^{-1}x$ when $b(x)$ is in the ellipsoid $\| b \| \le M(\varepsilon)$ and $M(\varepsilon)b(x)/\| b(x) \|$ outside it.

Consider again the I.Q. example (example ii in Section 1). Suppose that our prior distribution for this particular population is $N(\theta_0, \tau^2)$ while the mean of the sample is distributed according to $X - \Theta \sim N(0, \sigma^2)$. Corollary 4.1 defines $M(\varepsilon)$ by

$$\frac{\varepsilon}{1-\varepsilon} = 2\int_{x>M(\tau^2+\sigma^2)/\tau^2} \left(\frac{\tau^2}{\tau^2+\sigma^2}\frac{x}{M} - 1\right) d\Phi\left(\frac{x}{\sqrt{\tau^2+\sigma^2}}\right)$$
$$= 2\left[\frac{\tau^2}{M\sqrt{\tau^2+\sigma^2}}\,\varphi\left(\frac{M\sqrt{\tau^2+\sigma^2}}{\tau^2}\right) - \Phi\left(-\frac{M\sqrt{\tau^2+\sigma^2}}{\tau^2}\right)\right].$$

Now, for $\varepsilon = 0.085$ and 0.038 we get that the Bayes estimator should be truncated to $\theta_0 \pm 1.2\tau$ and $\theta_0 \pm 1.5\tau$ respectively.

In practice this result may be used by a non-Bayesian. The experiment can be done such that a small subsample will be tested by the standard test, and then the result of the majority of the sample will be used to refine the result of the subsample.

REMARK. This result complements Efron and Morris (1971) which looked on the same basic situation but when the statistician tries to be protected from "irrelevant prior". □

Our procedure may seem to be too conservative. The Bayes regret is one possible measure for how conservative a robust procedure is (Anscombe, 1960). This is defined as the difference between the expected risk of a procedure and the Bayes risk, assuming the ideal model (i.e. $\varepsilon = 0$) is true. One can take this measure as the "premium" one pays in order to be insured against the least favorable deviation from the model he believes to be true. Technically we define

$$\Delta(\varepsilon) = \inf_{\delta\in D_\varepsilon} L_0(\delta) - \inf_{\delta\in\mathbf{D}} L_0(\delta), \quad 1 \ge \varepsilon \ge 0$$

where D_ε is the set of all decision procedures which solve P_1 for the particular value of ε. The following result shows that the minimax rule is not too conservative (at least for small values of ε). For more details consult Ritov (1983).

THEOREM 5.

(i) $$\lim_{\varepsilon\to 0}\Delta(\varepsilon) = 0 \quad \textit{and} \quad \lim_{\varepsilon\to 0}\bar{L}(\varepsilon) = \bar{L}(0).$$

(ii) *Suppose there exists a Bayes procedure* β *such that* $\sup_{x\in\mathbf{X}} r_0[\beta(\cdot \mid x)] < \infty$. *Then*: $\Delta(\varepsilon) = o(\varepsilon)$ *as* $\varepsilon \to 0$, *and there is a Bayes procedure* β' *such that* $\sup_{H\in\mathscr{P}} L_\varepsilon(\beta', H) = \bar{L}(\varepsilon) + o(\varepsilon)$ *as* $\varepsilon \to 0$. □

4. Appendix. We give here only an outline of the proofs. For details the reader may refer to Ritov (1983).

LEMMA 1. *Let* $\{\delta_n\}$ *be a sequence of decision procedures. Then it has a subsequence* $\{\delta n_i\}$ *and there is a decision procedure* δ^* *such that*

$$L_0(\delta^*) \le \liminf_{i\to\infty} L_0(\delta n_i)$$

$$\sup_{x\in\mathbf{X}} r_0[\delta^*(\cdot \mid x)] \le \liminf_{i\to\infty}\sup_{x\in\mathbf{X}} r_0[\delta n_i(\cdot \mid x)]. \quad □$$

The proof of the lemma follows essentially the argument in Farrell (1967).

In the following $\bar{\delta}(\cdot \mid \cdot)$ will be any procedure such that $\bar{\delta}(\cdot \mid x) = \bar{\delta}(\cdot \mid x')$ and $r_0[\bar{\delta}(\cdot \mid x)] = s_0$ for all $x, x' \in \mathbf{X}$.

PROOF OF THEOREM 1. For any $s \ge s_0$ we can find a sequence $\{\delta_i\} \subset D(s)$ and $L_0(\delta_i) \to R(s)$. By Lemma 1 there is a procedure $\delta \in D(s)$ with $R(s) \le L_0(\delta) \le R(s)$. □

PROOF OF THEOREM 2. Let $\mathbf{X}^k$, $B(\mathbf{X}^k)$ be $\mathbf{X}$, $B(\mathbf{X})$ in case $\mathbf{X}$ is compact and its one point compactification otherwise. Let $\mathscr{P}^k$ be the set of all substochastic measures on $\mathbf{X}^k$, $B(\mathbf{X}^k)$ with the weak topology. Fix any $\varepsilon \in [0, 1)$ and define $K_\varepsilon\colon \mathbf{D} \times \mathscr{P}^k \to [0, \infty]$ by $K_\varepsilon(\delta, H) = \limsup_{G\to H} L_\varepsilon(\delta, G)$.

Now, K_ε is the closure of L_ε as a concave function and, therefore it inherits from $L_\varepsilon(\cdot, \cdot)$ its convex-concave-like characteristics (Rockafellar, 1971, page 115). $\mathscr{P}^k$ is compact and hence by Sion's theorem (Sion, 1958) K_ε has a saddle value.

The compactness of $\mathscr{P}^k$ and the upper-semicontinuity of $K_\varepsilon(\delta, \cdot)$ imply that nature has a least favorable distribution (for K_ε) H_ε which is clearly a probability measure. On the other hand, the statistician has by Lemma 1 a minimax procedure δ_ε. Hence $(\delta_\varepsilon, H_\varepsilon)$ is a saddle point for K_ε. Note that for any $\delta \in \mathbf{D}$, $\sup_{H\in\mathscr{P}^k} K_\varepsilon(\delta, H) = \sup_{H\in\mathscr{P}^k} L_\varepsilon(\delta, H)$. Hence δ_ε is a solution of P_1 for this ε. $(\delta_\varepsilon, H_\varepsilon)$ is also a saddle point for $L_\varepsilon(\cdot\,, \cdot)$, i.e. $\inf_{\delta\in\mathbf{D}} L_\varepsilon(\delta, H_\varepsilon) = L_\varepsilon(\delta_\varepsilon, H_\varepsilon) = \sup_{H\in\mathscr{P}} L_\varepsilon(\delta_\varepsilon, H)$. Suppose that this is not the case and $L_\varepsilon(\delta_\varepsilon, H_\varepsilon) < \sup_{H\in\mathscr{P}} L_\varepsilon(\delta_\varepsilon, H) = K_\varepsilon(\delta_\varepsilon, H_\varepsilon)$. Then there exists a closed set C and a positive number η such that $H_\varepsilon(C) > 0$ and

$$\sup_{x\in C} r_0[\delta_\varepsilon(\cdot \mid x)] \le \sup_{x\in \mathrm{X}} r_0[\delta_\varepsilon(\cdot \mid x)] - \eta = s - \eta.$$

Let C_n be open sets $\bigcap_{n=1}^{\infty} C_n = C$ and define

$$\delta_n = \begin{cases} \bar{\delta} & x \in C_n - C \\ \delta_\varepsilon & x \notin C_n - C. \end{cases}$$

Then $L_0(\delta_n) \to L_0(\delta_\varepsilon)$ and $\lim \sup_n K_\varepsilon(\delta_n, H_\varepsilon) \le K_\varepsilon(\delta_\varepsilon, H_\varepsilon) - \eta H_\varepsilon(C)$ and we have a contradiction.

Suppose now that there is a decision procedure δ such that $\eta = L_\varepsilon(\delta_\varepsilon, H_\varepsilon) - L_\varepsilon(\delta, H_\varepsilon) > 0$. We will prove that this implies the existence of a decision procedure $\tilde{\delta}$ such that $K_\varepsilon(\tilde{\delta}, H_\varepsilon) < L_\varepsilon(\delta_\varepsilon, H_\varepsilon)$, which is a contradiction. Let $\mu = (1 - \varepsilon)F + \varepsilon H_\varepsilon$, $f = dF/d\mu$ and $h = dH_\varepsilon/d\mu$. Define

$$\begin{aligned} V &= \{x\colon (1-\varepsilon) r[\delta(\cdot \mid x), x] f(x) + \varepsilon r_0[\delta(\cdot \mid x)] h(x) \\ &\qquad\qquad < (1-\varepsilon) r[\delta_\varepsilon(\cdot \mid x), x] f(x) + \varepsilon r_0[\delta_\varepsilon(\cdot \mid x)] h(x) - \eta\} \\ s^* &= \inf\{s\colon \mu[V \cap \{x\colon r_0[\delta(\cdot \mid x)] < s\}] > 0\} \\ S' &= V \cap \{x\colon r_0[\delta(\cdot \mid x)] < s^* + \eta/2\}. \end{aligned}$$

Let S be a compact subset of S' with $\mu(S) > 0$ and C an open set containing S such that

$$\int_{C-S} \{(1-\varepsilon) r[\bar{\delta}(\cdot \mid x), x] f(x) + \varepsilon s_0\}\, d\mu(x) < \eta\mu(S)/2.$$

Define now

$$\tilde{\delta} = \begin{cases} \delta(\cdot \mid x) & x \in S \\ \bar{\delta}(\cdot \mid x) & X \in C - S \\ \delta_\varepsilon(\cdot \mid x) & x \notin C. \end{cases}$$

Let

$$h_\pm(x) = \begin{cases} s^* + \eta/4 \pm \eta/4 & x \in S \\ s_0 & x \in C - S \\ \sup_{x \in \mathrm{X}^k} r_0[\delta(\cdot \mid x)] & x \notin C. \end{cases}$$

$h_+(\cdot)$ and $h_-(\cdot)$ are bounded upper-semicontinuous functions and satisfy:

$$\int h_-(x) H_\varepsilon(dx) \le \int r_0[\tilde{\delta}(\cdot \mid x)] H_\varepsilon(dx) \quad \text{and} \quad r_0[\tilde{\delta}(\cdot \mid \cdot)] < h_+(\cdot).$$

Hence

$$\begin{aligned} K_\varepsilon(\tilde{\delta}, H_\varepsilon) - L_\varepsilon(\tilde{\delta}, H_\varepsilon) &= \varepsilon \lim \sup_{H \to H_\varepsilon} \int r_0[\tilde{\delta}(\cdot \mid x)][H(dx) - H_\varepsilon(dx)] \\ &\le \varepsilon \lim \sup_{H \to H_\varepsilon} \left[\int h_+(x) H(dx) - \int h_-(x) H_\varepsilon(dx)\right] \\ &\le \varepsilon \left[\int h_+(x) H_\varepsilon(dx) - \int h_-(x) H_\varepsilon(dx)\right] \\ &\le \eta\mu(S)/2. \end{aligned}$$

This result together with $L_\varepsilon(\tilde{\delta}, H_\varepsilon) < L_\varepsilon(\delta_\varepsilon, H_\varepsilon) - \mu(S)\eta/2$ give us the desired contradiction.

Finally, we have to prove that $H_\varepsilon \ll F$. Suppose that there is a measurable set U such that $F(U) = 0 < H_\varepsilon(U)$. As the statistician may use $\tilde{\delta}$ on U without increasing his loss, this means that $\delta_\varepsilon \in D(s_0)$. Now $H_\varepsilon(U) = 1$ means that there is a Bayes procedure in $D(s_0)$; hence $H_\varepsilon(U) < 1$. Therefore nature may use H_ε concentrated on U^c as well. □

PROOF OF THEOREM 3. The fact that any solution of P_1 should be a solution of P_2 for some s is quite clear and was mentioned before. Suppose now that δ^* is a solution of P_2 for some s^*. If $s^* = s_0$ then δ^* is the solution of P_1 for $\varepsilon = 1$. Suppose $s^* > s_0$. Now, by the first part there is a function $s(\varepsilon)$ (not necessarily unique) such that

$$\bar{L}(\varepsilon) = \inf_{\delta\in D}\sup_{H\in\mathscr{S}} L_\varepsilon(\delta, H) = (1-\varepsilon)R(s(\varepsilon)) + \varepsilon s(\varepsilon)$$
$$\leq (1-\varepsilon)R(s) + \varepsilon s; \quad 1 > \varepsilon > 0$$

for any other s. This implies that $R(\cdot)$ has a supporting line at $s(\varepsilon)$ with a slope of $-\varepsilon/(1-\varepsilon)$. Using an argument similar to that of Hodges and Lehmann (1952), we can prove that $R(\cdot)$ is concave, and, hence, there is ε^* such that $-\varepsilon^*/(1-\varepsilon^*)$ is the slope of a supporting line $R(\cdot)$ at s^*. But $-\varepsilon^*/(1-\varepsilon^*)$ is also the slope of a supporting line of $R(\cdot)$ at $s(\varepsilon^*)$, or $R(s(\varepsilon^*)) = R(s^*) - \varepsilon^*[s(\varepsilon^*) - s^*]/(1-\varepsilon^*)$. Hence

$$\bar{L}(\varepsilon^*) = (1-\varepsilon^*)R(s(\varepsilon^*)) + \varepsilon^* s(\varepsilon^*) = (1-\varepsilon^*)R(s^*) + \varepsilon^* s^*$$

i.e., δ^* is a solution of P_1 for ε^*. □

PROOF OF THEOREM 4. Fix any $s > s_0$. We will prove that (a) such a solution exists, (b) (i) + (ii) ⇒ (iii), and (c) (i) ⇒ (ii).

(a) Theorem 3 implies that the solution of P_2 for this s is a solution of P_1 for some $\varepsilon(s)$. Theorem 2 implies that there is such a solution which is Bayes for a $H_s^* \ll F$. Take λ_s^* to be a version of $(1-\varepsilon)^{-1}\,\varepsilon\, dH_s^*/dF$. H_s^*, as a saddle point strategy of nature is supported on $\{x: r_0[\delta_s^*(\cdot\mid x)] = s\}$. Hence $\lambda_s^*(s)[r_0[\delta_s^*(\cdot\mid x)] - s] \equiv 0$ and $\delta^*(\cdot\mid x)$ minimizes $r[\delta(\cdot\mid x), x] + \lambda_s^*(x)r_0[\delta(\cdot\mid x)]$ for almost all $X(F)$.

(b) Define $H_s(dx) = (1-\varepsilon)/\varepsilon\ \lambda_s(x)F(dx)$. Then by construction

$$\inf_{\delta\in D} L_\varepsilon(\delta, H_s) = L_\varepsilon(\delta_s, H_s) = \sup_{H\in\mathscr{S}} L_\varepsilon(\delta_s, H).$$

(c) Suppose (i) is true while $\int \lambda_s(x)F(dx) = \infty$. Then there is a sequence of measurable sets $\{B_n\}$, $B_n \in B(\mathbf{X})$, $n = 1, 2, \cdots$ such that

$$\int_{B_n} \lambda_s(x)F(dx) < \infty \quad \text{while} \quad \lim_{n\to\infty} \int_{B_n} \lambda_s(x)F(dx) = \infty.$$

Let λ_s^* δ_s^* and H_s^* be defined as in part (a) of the proof. Define now for

$n = 1, 2, \cdots$

$$\lambda_{s,n}(x) = \begin{cases} \lambda_s(x) & x \in B_n \\ \lambda_s^*(x) & x \notin B_n \end{cases} \qquad \delta_{s,n} = \begin{cases} \delta_s(x) & x \in B_n \\ \delta_s^*(x) & x \notin B_n. \end{cases}$$

Part (b) implies that $\delta_{s,n}$ is a solution of P_1 for a sequence $\{\varepsilon_n\}$. $\varepsilon_n/(1 - \varepsilon_n) = \int \lambda_{s,n}(x)F(dx) \longrightarrow \infty$, or $\varepsilon_n \longrightarrow 1$. But for any function $s(\varepsilon)$ defined as in the proof of Theorem 3, $\lim_{\varepsilon \to 1} s(\varepsilon) = s_0$, contradicting the assumption that $s > s_0$. □

Acknowledgment. This work is based on a chapter in the author's Ph.D. thesis submitted to the Hebrew University of Jerusalem. I wish to express my gratitude to my thesis advisors, Professor P. J. Bickel and Professor J. Yahav. Due to helpful remarks of an associate editor, referees and Professor P. J. Bickel, this is a remote version of a previous effort.

REFERENCES

ANSCOMBE, F. J. (1960). Rejection of outlier. *Technometrics* **9** 123–147.

BERGER, J. O. and BERLINER, L. M. (1983). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. Technical Report #83-35, Department of Statistics, Purdue University.

BERGER, R. L. (1979). Gamma minimax robustness of Bayes rules. *Comm. Statist.* **A-8** 543–560.

BLUM, J. R. and ROSENBLATT, J. (1967). On partial a priori information in statistical inference. *Ann. Math. Statist.* **38** 1671–1678.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. A* **143** 383–430.

BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading.

EFRON, B. and MORRIS, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators—part I: The Bayes case. *J. Amer. Statist. Assoc.* **66** 807–815.

FARRELL, R. (1967). Weak limits of Bayes procedures in estimation theory. *Proc. 5th Berkeley Symp. Math. Statist. Probab.* **1** 83–112.

HODGES, J. L., JR. and LEHMANN, E. L. (1952). The use of previous experience in reaching statistical decisions. *Ann. Math. Statist.* **23** 396–407.

MARAZZI, A. (1980). Robust Bayesian estimation for the linear model. Research Report no. 27, Fachgruppe fuer Statistik, Eidgenoessische Technische Hochschule, Zuerich.

MARAZZI, A. (1982). On constrained minimization of the Bayes risk for the linear model. Research Report no. 34, Fachgruppe fuer Statistik, Eidgenoessische Technische Hochschule, Zuerich.

RITOV, J. Y. (1983). Robust quasi-Bayesian inference. Ph.D. thesis, The Hebrew University of Jerusalem.

ROCKAFELLAR, R. T. (1971). Saddle points and convex analysis. In H. W. Kuhn and G. P. Szego (eds.) *Differential Games and Related Topics.* North Holland, Amsterdam.

SION, M. (1958). On general minimax theorems. *Pacific J. Math.* **8** 171–176.

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
91905 JERUSALEM
ISRAEL