The Annals of Statistics
1986, Vol. 14, No. 2, 461–486

ROBUST BAYES AND EMPIRICAL BAYES ANALYSIS WITH ε-CONTAMINATED PRIORS

BY JAMES BERGER[1] AND L. MARK BERLINER[2]

Purdue University and Ohio State University

For Bayesian analysis, an attractive method of modelling uncertainty in the prior distribution is through use of ε-contamination classes, i.e., classes of distributions which have the form $\pi = (1 - \varepsilon)\pi_0 + \varepsilon q$, π_0 being the base elicited prior, q being a "contamination," and ε reflecting the amount of error in π_0 that is deemed possible. Classes of contaminations that are considered include (i) all possible contaminations, (ii) all symmetric, unimodal contaminations, and (iii) all contaminations such that π is unimodal.

Two issues in robust Bayesian analysis are studied. The first is that of determining the range of posterior probabilities of a set as π ranges over the ε-contamination class. The second, more extensively studied, issue is that of selecting, in a data dependent fashion, a "good" prior distribution (the Type-II maximum likelihood prior) from the ε-contamination class, and using this prior in the subsequent analysis. Relationships and applications to empirical Bayes analysis are also discussed.

1. Introduction. The most frequent criticism of subjective Bayesian analysis is that it supposedly presumes an ability to completely and accurately quantify subjective information in terms of a single prior distribution. However, there has long existed [at least since Good (1950)] a *robust Bayesian viewpoint* which assumes only that subjective information can be quantified in terms of a class Γ of possible distributions. The goal is then to make inferences or decisions which are *robust* over Γ, i.e., which are relatively insensitive (or at least are satisfactory) to deviations as the prior distribution varies over Γ. We will not consider the philosophical or pragmatic reasons for adopting this viewpoint. Such a discussion, along with a review of the area, may be found in Berger (1984). (We also do not mean to imply that the single prior Bayesian approach is necessarily bad; it usually works very well.) Related to this are various forms of empirical Bayes analysis [cf. Morris (1983) for discussion and review], in which the prior distribution is also assumed to belong to some class Γ of distributions. Indeed, Section 5 considers some familiar empirical Bayes problem from our perspective. Also, see Berger and Berliner (1984).

Before discussing implementation of the robust Bayesian viewpoint, some notation is helpful. Let X denote the observable random variable (or vector), which will (for simplicity) be assumed to have a density $f(x|\theta)$ (w.r.t. some

Received November 1983; revised October 1985.

[1] Research supported by the National Science Foundation under Grants MCS-8101670A1 and DMS-8401996.

[2] Research supported by the Office of Naval Research under Contract N00014-84-K-0422.

AMS 1980 *subject classifications.* Primary 62A15; secondary 62F15.

Key words and phrases. Robust Bayes, empirical Bayes, classes of priors, ε-contamination, type II maximum likelihood, hierarchical priors.

measure), where θ is an unknown parameter lying in a parameter space Θ. A prior distribution on Θ will be denoted by π (later, in examples, π will be used to denote either a prior or its corresponding density), and the resulting marginal density of X is given by

$$m(x|\pi) = E^{\pi} f(x|\theta) = \int_{\Theta} f(x|\theta)\pi(d\theta).$$

The posterior distribution of θ given x (assuming it exists) will be denoted by $\pi(\cdot|x)$ and, in nice situations, is defined by

$$\pi(d\theta|x) = f(x|\theta)\pi(d\theta)/m(x|\pi).$$

Finally, let $\mathscr{P}$ denote the space of all probability distributions on Θ.

The class, Γ, of prior distributions to be considered in this paper, is the ε-contamination class; namely,

$$\Gamma = \{\pi: \pi = (1-\varepsilon)\pi_0 + \varepsilon q,\ q \in \mathscr{Q}\}, \tag{1.1}$$

where $0 \le \varepsilon \le 1$ is given, π_0 is a particular prior distribution, and $\mathscr{Q}$ is some subset of $\mathscr{P}$. There are several reasons for consideration of this class. First, and foremost, it is a sensible class to consider in light of the prior elicitation process. The extensive and rapidly developing methodology on prior elicitation [cf. Kadane et al. (1980)] makes specification of an initial believable prior, π_0, an attractive starting point. However, in determining π_0 sensibly, one will make *probability* judgements about subsets of Θ, judgements which could be in error by some amount ε. Stated another way, further reflection might lead to alterations of probability judgements by an amount ε. Hence, possible priors involving such alterations should be included in Γ.

Many classes of priors which have been considered are not sensible from the above viewpoint. For instance, classes of priors involving restrictions on moments force severe restrictions on the allowable prior tails. This makes little sense from the elicitation viewpoint, since the tails of a prior involve very small probabilities and are, therefore, nearly impossible to determine. Similarly, classes of conjugate priors are too limited, particularly in their inflexible tail behavior [cf. Berger (1985)].

Two other major reasons for choosing Γ as in (1.1) are (i) such Γ are (as we shall see) surprisingly easy to work with; and (ii) such Γ are very flexible through choice of $\mathscr{Q}$. In this paper we will restrict consideration to four interesting choices of $\mathscr{Q}$. First, in Section 2 the choice $\mathscr{Q} = \mathscr{P}$ (all distributions) will be considered. This choice is easy to work with and is, in some sense, conservative. In Section 3 we consider the class, $\mathscr{Q}$, of all contaminations which are symmetric and unimodal. This is again very easy to work with. In Section 4 we consider the class of all contaminations such that *the resulting* π is unimodal (assuming that π_0 is unimodal). It came as a great surprise to us that such a complicated class could be worked with and provide reasonably simple answers. Finally, in Section 5 we consider $\mathscr{Q}$ that are mixtures of various classes. The purpose of the section is to show how easily mixed contaminations can be dealt with and also to apply the methodology in some typical empirical Bayes situations.

Other articles that have used ε-contamination classes of priors include Schneeweiss (1964), Blum and Rosenblatt (1967), Huber (1973), Marazzi (1985), Bickel (1984), and Berger (1982, 1984). Except for Huber (1973), these articles work within the frequentist Bayesian framework, whereas our approach will be almost entirely conditional Bayesian. Huber (1973) is discussed below and in Section 2.4. There is a substantial literature working with other types of classes of priors [cf. Leamer (1978) and DeRobertis and Hartigan (1981)], and with the very related idea of "upper" and "lower" probabilities. Berger (1984) contains considerable review and discussion of this literature. We strongly prefer the class in (1.1) for intuitive content and ease of analysis.

The ideal analysis, to a robust Bayesian, is one in which it can be shown that the inference or decision to be made is essentially the same for any prior in Γ. [Indeed, it can be argued—see Berger (1984, 1985)—that this is the only way in which a statistical conclusion can claim to be ultimately sound.] What is needed, to provide such conclusions, is essentially the ability to find minimums and maximums of criterion functions as π ranges over Γ. We illustrate this approach in Section 2.4, where, for $\mathcal{Q}$ = {all distributions), the range of posterior probabilities of a (fixed) set C is given [essentially following Huber (1973)]. This allows finding the range of posterior probabilities of confidence sets and the range of posterior probabilities of hypotheses, for such Γ.

Unfortunately, there are certain inadequacies in assuming that $\mathcal{Q}$ = {all distributions} (see Section 2.3), and attempting the above program with more reasonable Γ (such as that in Section 4) becomes more difficult. A number of alternative approaches to the problem of dealing with classes of priors have thus been proposed, essentially leading to the choice of a single "robust" prior, decision, or inference. Berger (1984, 1985) discusses various of these methods, including the appealing technique of putting a prior distribution on Γ. [Such a prior is called a hyperprior or a Type II probability distribution by Good (1965, 1980) and a second-stage prior in certain situations by Lindley and Smith (1972).] Of course, this corresponds to using a certain single prior (the "average" over Γ), but one would suspect that the resulting Bayes rule would be quite robust with respect to Γ. The difficulty in doing this is mainly technical: it is essentially impossible to put a reasonable prior on complicated Γ, such as those in Sections 2–4, and carry out the Bayesian calculations. Note also that, ideally, most of the prior information available will have been exhausted in constructing Γ. Hence, any prior distribution placed on Γ will to a large extent, be arbitrary.

Instead, we will consider the simplest and most commonly used method of selecting a hopefully robust prior in Γ, namely choice of that prior π which maximizes the marginal $m(x|\pi)$ over Γ. This process is called Type II maximum likelihood by Good (1965). For $\pi = (1 - \varepsilon)\pi_0 + \varepsilon q$, $q \in \mathcal{Q}$, maximizing $m(x|\pi) = (1 - \varepsilon)m(x|\pi_0) + \varepsilon m(x|q)$ over π is clearly done by maximizing $m(x|q)$ over q. Assuming that the maximum of $m(x|q)$ is attained at (a unique) $\hat{q} \in \mathcal{Q}$, we will then suggest formally using the estimated prior $\hat{\pi}$, given by

$$\hat{\pi} = (1 - \varepsilon)\pi_0 + \varepsilon\hat{q}. \tag{1.2}$$

(Of course, $\hat{\pi}$ thus depends on x.) Throughout the paper, $\hat{\pi}$ will be called the

ML-II prior. Also, any quantities derived from $\hat{\pi}$ will appear with the modifier "ML-II" for clarity.

Choosing a prior with the help of the data always engenders controversy. Several justifications for doing so can be given, however. First, if $m(x|\pi)$ is "small," it is simply unlikely that such a π could be "true," and hence worrying about such π is counterproductive. Recall that (supposedly) all $\pi \in \Gamma$ are deemed to be reasonable representations of priors beliefs, so $\hat{\pi}$ is simply the prior which is most plausible, in light of prior opinions and the data. A more formal way of saying this is that, if all $\pi \in \Gamma$ are roughly equally likely a priori, then $\hat{\pi}$ is the "posterior mode" of the "uniform" distribution on Γ, and might often be expected to yield a posterior distribution that is close to the true posterior distribution for such a "uniform" distribution on Γ.

The preceding argument for $\hat{\pi}$ is, of course, nonrigorous, and the ultimate justification for proceeding in this way is simply that it can give reasonable answers. Of course, there is already substantial evidence in the literature attesting to the success of the method, both in the Bayesian literature [cf. Jeffreys (1961), Good (1965, 1980), Box and Tiao (1973), Bishop, Fienberg, and Holland (1975), and Zellner (1985)], and in the empirical Bayesian literature [cf. Maritz (1970) and Morris (1983)]. Indeed, note that the "standard" empirical Bayes methodology is to choose Γ to be a class of conjugate priors and then to estimate the "hyperparameters" of the prior by maximizing $m(x|\pi)$, yielding $\hat{\pi}$. Also the related use of the marginal in Bayesian model robustness investigations is well established [cf. Box and Tiao (1973), Dempster (1975), and Box (1980)]. When all is said and done, however, we recognize that the ML-II technique is not foolproof and can produce bad answers, particularly when Γ includes unreasonable distributions. (The basic problem with the ML-II technique is that ensuing calculations of variability do not take into account the "error" of the ML-II estimation; see Section 2.3 for an extreme example of this problem.) In Section 6 we give a general discussion of the success of the method for the situations discussed in the paper.

We conclude this section with useful formulas and notation. For priors of the form

$$\pi(d\theta) = (1 - \varepsilon)\pi_0(d\theta) + \varepsilon q(d\theta), \tag{1.3}$$

computations give [assuming the existence of the posterior distributions $\pi_0(d\theta|x)$ and $q(d\theta|x)$]

$$m(x|\pi) = (1 - \varepsilon)m(x|\pi_0) + \varepsilon m(x|q) \tag{1.4}$$

and

$$\pi(d\theta|x) = \lambda(x)\pi_0(d\theta|x) + (1 - \lambda(x))q(d\theta|x), \tag{1.5}$$

where $\lambda(x) \in [0, 1]$ is given by

$$\lambda(x) = (1 - \varepsilon)m(x|\pi_0)/m(x|\pi). \tag{1.6}$$

Furthermore, the posterior mean, δ^π, and posterior variance, V^π, can be written

(assuming they exist) as

$$\delta^{\pi}(x) = \lambda(x)\delta^{\pi_0}(x) + (1 - \lambda(x))\delta^{q}(x) \tag{1.7}$$

and

$$V^{\pi}(x) = \lambda(x)V^{\pi_0}(x) + (1 - \lambda(x))V^{q}(x) + \lambda(x)(1 - \lambda(x))(\delta^{\pi_0}(x) - \delta^{q}(x))^2. \tag{1.8}$$

Part of the appeal of the ε-contamination class, Γ, is the simplicity of these formulas.

2. Analysis for arbitrary contaminations. A natural suggestion for a class of contaminations of a fixed, elicited prior π_0 is the class of all possible contaminations. In this section we will examine inferences, including point estimation, testing, and credible regions, for such a class, i.e., for

$$\Gamma = \{\pi: \pi = (1 - \varepsilon)\pi_0 + \varepsilon q, q \in \mathscr{P}\}. \tag{2.1}$$

In a number of respects this is too *large* a class of priors, including many priors that are unreasonable. And it will be seen that this can lead to serious difficulties in some situations (although for certain purposes no problems are encountered). We give a fairly detailed analysis of this situation because its relative simplicity allows easy comprehension of important concepts (*including* the difficulties of using too large a Γ), and because *some* useful robustness results do emerge. All proofs are easy and are omitted.

2.1. The ML-II prior and posterior. For Γ defined as in (2.1), the ML-II prior and corresponding posterior are as follows.

THEOREM 2.1. *Assume X has a density $f(x|\theta)$ w.r.t. some dominating measure on the sample space of X. Assume that the usual maximum likelihood estimator for θ, say $\hat{\theta}(x)$, exists and is unique. For Γ defined as in* (2.1), *the ML-II prior is given by*

$$\hat{\pi}(\cdot) = (1 - \varepsilon)\pi_0(\cdot) + \varepsilon\hat{q}_x(\cdot), \tag{2.2}$$

where $\hat{q}_x$ assigns probability one to the point $\theta = \hat{\theta}(x)$. The ML-II posterior is given by

$$\hat{\pi}(\cdot|x) = \hat{\lambda}(x)\pi_0(\cdot|x) + (1 - \hat{\lambda}(x))\hat{q}_x(\cdot), \tag{2.3}$$

where

$$\hat{\lambda}(x) = (1 - \varepsilon)m(x|\pi_0)/[(1 - \varepsilon)m(x|\pi_0) + \varepsilon f(x|\hat{\theta}(x))]. \tag{2.4}$$

2.2. The ML-II posterior mean. Under the assumptions of Theorem 2.1, the ML-II posterior mean of θ is given by [see (1.7)]

$$\delta^{\hat{\pi}}(x) = \hat{\lambda}(x)\delta^{\pi_0}(x) + (1 - \hat{\lambda}(x))\hat{\theta}(x). \tag{2.5}$$

As an estimator of θ, $\delta^{\hat{\pi}}$ is intuitively appealing, in that it is a reasonable data dependent mixture of δ^{π_0} and $\hat{\theta}$. When the data are consistent with π_0, $m(x|\pi_0)$

will be reasonably large and $\hat{\lambda}(x)$ close to one (for small ε), so that $\delta^{\hat{\pi}}$ will essentially equal δ^{π_0}. When the data and π_0 are not compatible, however, $m(x|\pi_0)$ will be small and $\hat{\lambda}(x)$ near zero; $\delta^{\hat{\pi}}$ will then be approximately equal to the m.l.e. $\hat{\theta}$.

The following example presents $\delta^{\hat{\pi}}$ in an important situation. Some properties of the estimator are discussed which give a degree of "outside validation" to the estimator.

EXAMPLE 1. Let $X = (X_1, \ldots, X_p)^t \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$, where $\theta = (\theta_1, \ldots, \theta_p)^t$ is unknown and σ^2 is known. Suppose the elicited prior, π_0, for θ is $\mathcal{N}_p(\mu, \tau^2 I_p)$. (Thus μ and τ^2 are specified.) Since the usual maximum likelihood estimator of θ is $\hat{\theta}(x) = x$, and $\delta^{\pi_0}(x) = x - [\sigma^2/(\sigma^2 + \tau^2)](x - \mu)$, formula (2.5) reduces to

$$\delta^{\hat{\pi}}(x) = \big(1 - \hat{\lambda}(x)\sigma^2/(\sigma^2 + \tau^2)\big)(x - \mu) + \mu,$$

where

$$\hat{\lambda}(x) = \Big[1 + (\varepsilon/(1 - \varepsilon))\big(1 + \tau^2/\sigma^2\big)^{p/2} \exp\big\{|x - \mu|^2/2(\sigma^2 + \tau^2)\big\}\Big]^{-1}.$$

Note that $\hat{\lambda}$ goes to 0 exponentially fast in $|x - \mu|^2$, so that $\delta^{\hat{\pi}}(x) \to x$ quite rapidly as $|x - \mu|^2$ gets large. Because of this, one might conjecture that the estimator is minimax, in a frequentist decision-theoretic sense under, say, quadratic loss. Unfortunately, this turns out not to be the case, although the deviation from minimaxity is usually fairly slight. It is also interesting to note that $\delta^{\hat{\pi}}$ happens to coincide with the generalized Bayes estimator corresponding to the formal prior

$$\rho(d\theta) = (1 - \varepsilon)\pi_0(d\theta) + \varepsilon\rho_0(d\theta),$$

where $\rho_0(d\theta) = (2\pi\sigma^2)^{p/2}\, d\theta$. Note that priors of a similar form were considered by, for instance, Leonard (1974). The development here can be viewed as proposing a reasonable method for choosing the relative weights of π_0 and $(d\theta)$.

2.3. The ML-II posterior variance. To determine the estimation error in using $\delta^{\hat{\pi}}$, it is natural to look at the posterior variance, $V^{\hat{\pi}}$. From (1.8) it follows that

$$V^{\hat{\pi}}(x) = \hat{\lambda}(x)\Big[V^{\pi_0}(x) + (1 - \hat{\lambda}(x))\big(\delta^{\pi_0}(x) - \hat{\theta}(x)\big)^2\Big].$$

It will typically be the case (as in Example 1) that, as $\hat{\lambda}(x) \to 0$, $V^{\hat{\pi}}(x)$ will also go to zero. Indeed, $\hat{\pi}$ will usually "converge" to a point mass at $\hat{\theta}(x)$. This is clearly inappropriate; although data incompatible with π_0 can be cause for preference of $\hat{\theta}(x)$ to $\delta^{\pi_0}(x)$, it does not cause one to think that θ equals $\hat{\theta}(x)$ *exactly.*

The trouble here is caused by the fact that Γ contains unrealistic distributions. We *may* feel that π_0 could be in error, but surely a point mass at $\hat{\theta}(x)$ (when far from the center of π_0) is not usually a reasonable contamination to expect a priori. Working with Γ as in Sections 3 and 4, which do not allow such implausible contaminations, substantially alleviates this problem. (See also Section 6.)

2.4. Robustness as π ranges over Γ. As mentioned in the introduction, the ideal goal for a robustness study would be to show that a decision or inference being contemplated is satisfactory for all $\pi \in \Gamma$. When $\mathcal{Q}$ is the class of all distributions, it often becomes feasible to check this. The basic tool is the following result of Huber (1973), concerning the range of posterior probabilities of a set.

THEOREM 2.2 [Huber (1973)]. *Suppose $\mathcal{Q} = \mathcal{P}$. Let C be a measurable subset of Θ, and define β_0 to be the posterior probability of C under π_0, i.e.,*

$$\beta_0 = P^{\pi_0}(\theta \in C|X = x).$$

Then

$$\inf_{\pi \in \Gamma} P^{\pi}(\theta \in C|X = x) = \beta_0\left\{1 + \frac{\varepsilon \sup_{\theta \notin C} f(x|\theta)}{(1 - \varepsilon)m(x|\pi_0)}\right\}^{-1}, \tag{2.6}$$

and

$$\sup_{\pi \in \Gamma} P^{\pi}(\theta \in C|X = x) = \frac{(1 - \varepsilon)m(x|\pi_0)\beta_0 + \varepsilon \sup_{\theta \in C} f(x|\theta)}{(1 - \varepsilon)m(x|\pi_0) + \varepsilon \sup_{\theta \in C} f(x|\theta)}. \tag{2.7}$$

EXAMPLE 2. Assume that $X \sim \mathcal{N}(\theta, \sigma^2)$, σ^2 known, and that π_0 is $\mathcal{N}(\mu, \tau^2)$. It is well known that $\pi_0(d\theta|x)$ is $\mathcal{N}(\delta(x), V^2)$, where

$$\delta(x) = x - \left(\sigma^2/(\sigma^2 + \tau^2)\right)(x - \mu), \qquad V^2 = \sigma^2\tau^2/(\sigma^2 + \tau^2).$$

The usual $100(1 - \alpha)\%$ Bayes credible region for θ is

$$C = \{\theta: \delta(x) - K < \theta < \delta(x) + K\},$$

where $K = z_{\alpha/2}V$, $z_{\alpha/2}$ being the $100\,(1 - \alpha/2)$ upper percentile of the standard normal distribution.

To investigate the robustness of C, we use (2.6) of Theorem 2.2. Note that

$$\sup_{\theta \notin C} f(x|\theta) = \begin{cases} (2\pi\sigma^2)^{-1/2} & \text{if } x \notin C, \\ (2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left(|x - \delta(x)| - K\right)^2\right\} & \text{if } x \in C. \end{cases}$$

Thus (2.6) becomes, for $x \notin C$,

$$\inf_{\pi \in \Gamma} P^{\pi}(\theta \in C|X = x) = (1 - \alpha)\left\{1 + \frac{\varepsilon\left(1 + \tau^2/\sigma^2\right)^{1/2}}{(1 - \varepsilon)}\exp\left[\frac{(x - \mu)^2}{2(\sigma^2 + \tau^2)}\right]\right\}^{-1},$$

and, for $x \in C$,

$$\inf_{\pi \in \Gamma} P(\theta \in C|X = x) = (1 - \alpha)\left\{1 + \frac{\varepsilon\left(1 + \tau^2/\sigma^2\right)^{1/2}}{(1 - \varepsilon)} \times \exp\left[\frac{(x - \mu)^2 - \left(|x - \mu|V/\tau - z_{\alpha/2}\tau\right)^2}{2(\sigma^2 + \tau^2)}\right]\right\}^{-1}.$$

As a concrete example, suppose that $\sigma^2 = 1$, $\tau^2 = 2$, $\mu = 0$, and $\varepsilon = 0.2$. First, suppose $x = 0.5$ is observed. Then the usual 95% Bayes credible interval for θ is $(-1.27, 1.93)$. Calculation gives

$$\inf_{\pi \in \Gamma} P^{\pi}(-1.27 < \theta < 1.93 | X = 0.5) = 0.817,$$
$$\sup_{\pi \in \Gamma} P^{\pi}(-1.27 < \theta < 1.93 | X = 0.5) = 0.966.$$

Hence, the standard credible set is reasonably robust. On the other hand, suppose $x = 4$ is observed. [Note that, since $m(x|\pi_0)$ is $\mathcal{N}(0, 3)$, this is not an "outrageous" observation.] Then the usual 95% credible set is $(1.07, 4.27)$. However, in this case we have that

$$\inf_{\pi \in \Gamma} P^{\pi}(1.07 < \theta < 4.27 | X = 4) = 0.1355,$$
$$\sup_{\pi \in \Gamma} P^{\pi}(1.07 < \theta < 4.27 | X = 4) = 0.99.$$

Since the posterior probability can get as low as 0.1355 for $x = 4$, robustness is not present.

Two interesting general points emerge from the previous example. First, robustness with respect to Γ will usually depend significantly on the x observed. Second, a lack of robustness may be due to the fact that Γ is "too large." When $x = 4$, for instance, the low probability of coverage (0.1355) is achieved when the contamination, q, is a point mass at 4.27. The resulting prior would probably not have been deemed to be reasonable a priori. Using a more reasonable γ might result in robustness. Also, more robust credible sets can be found—indeed Berger and Berliner (1983) determine the optimal $1 - \alpha$ robust credible set, optimal in the sense of having smallest size (Lebesgue measure) subject to its posterior probability being at least $1 - \alpha$ for all π in Γ. In any case, the use of $\mathcal{Q} = \mathcal{P}$ and Theorem 2.2 is conservative, in that, if robustness of a credible set is achieved for such Γ, one knows that robustness is also present for the more reasonable, smaller Γ.

Theorem 2.2 can also be used for hypothesis testing. Thus suppose we desire to test the hypothesis H_0: $\theta \in \Theta_0$ versus the alternative H_1: $\theta \in \Theta - \Theta_0$. For a fixed prior π, the usual Bayesian test is based on the posterior odds ratio $O_{\pi}(x)$, defined by

$$O_{\pi}(x) = P^{\pi}(\theta \in \Theta_0 | X = x) / [1 - P^{\pi}(\theta \in \Theta_0 | X = x)].$$

Letting $C = \Theta_0$, Theorem 2.2 immediately yields the following:

COROLLARY 2.1. *For* Γ *as in* (2.1),

$$\inf_{\pi \in \Gamma} O_{\pi}(x) = O_{\pi_0}(x) \left\{ 1 + \frac{\varepsilon \sup_{\theta \notin \Theta_0} f(x|\theta)}{(1 - \varepsilon)(1 - \beta_0) m(x|\pi_0)} \right\}^{-1},$$

and

$$\sup_{\pi \in \Gamma} O_{\pi}(x) = O_{\pi_0}(x) \left\{ 1 + \frac{\varepsilon \sup_{\theta \in \Theta_0} f(x|\theta)}{(1 - \varepsilon)(1 - \beta_0) m(x|\pi_0)} \right\},$$

where $\beta_0 = P^{\pi_0}(\theta \in \Theta_0 | X = x)$.

In testing, it will usually be much easier to achieve robustness using this "too large" Γ, since extreme x [i.e., x for which $m(x|\pi_0)$ is small], which lead to the unrealistic point mass contaminations, will usually provide extreme evidence for, or against, Θ_0. (The difference between the inf and sup of O_π may be substantial, but they will both be substantially less than one or substantially greater than one.) Together with the simplicity of the results in Corollary 2.1, this makes the use of $\mathcal{Q} = \mathcal{P}$ very attractive for robustness investigations in testing.

It should be clear that Theorem 2.2 is also immediately applicable to the testing of several hypotheses and to classification problems. Lower and upper bounds on the posterior probabilities of all hypotheses can be obtained.

3. Symmetric unimodal contaminations. A natural, yet remarkably tractable, class of priors to consider when $\Theta \subseteq \mathbb{R}^1$, is the ε-contamination class defined (for fixed θ_0) by

$$\mathcal{Q} = \{\text{densities of the form } q(|\theta - \theta_0|),\ q \text{ nonincreasing}\}. \tag{3.1}$$

This class is particularly reasonable when π_0 itself is unimodal and symmetric about θ_0. Note that under such circumstances, the resulting contaminated priors π display the desirable properties that (i) values of θ far from θ_0 cannot be given overwhelming weight (unlike the possibilities observed in Section 2), but (ii) priors with tails larger than π_0 are considered.

The considerable simplicity of working with (3.1) accrues from the fact that in much of the analysis, (3.1) can be replaced by

$$\mathcal{Q}' = \{\text{Uniform } (\theta_0 - a, \theta_0 + a) \text{ densities},\ a \geq 0\}, \tag{3.2}$$

where the "density" when $a = 0$ is a point mass at θ_0. Required optimizations thus involve only the variable a. Preliminary analyses of the type discussed in Section 2.4 have been carried out in this manner, but are a bit involved and will be reported elsewhere. However, the ML-II prior is quite simple to present:

THEOREM 3.1. *For the ε-contamination class with $\mathcal{Q}$ as in* (3.1), *an ML-II prior is*

$$\hat{\pi} = (1 - \varepsilon)\pi_0 + \varepsilon\hat{q},$$

where $\hat{q}$ is Uniform $(\theta_0 - \hat{a}, \theta_0 + \hat{a})$, $\hat{a}$ being the value of a which maximizes

$$m(x|a) = \begin{cases} (2a)^{-1}\int_{\theta_0 - a}^{\theta_0 + a} f(x|\theta)\, d\theta, & a > 0, \\ f(x|\theta_0), & a = 0. \end{cases} \tag{3.3}$$

PROOF. The proof follows trivially after noting that (i) any prior in (3.1) is a mixture of priors in (3.2), and (ii) $m(x|\pi)$ is a linear functional of π. □

Theorem 3.1 is an adaptation of a result in Berger and Sellke (1984), who utilized the fact that $m(x|\hat{a})$ is an upper bound on $m(x|q)$, q in $\mathcal{Q}$ given by (3.1), to establish startling lower bounds on posterior probabilities of point null

hypotheses that are an order of magnitude larger than classical significance levels or P-values. Here we utilize the theorem to calculate the ML-II posterior mean and variance in an illustrative example:

EXAMPLE: ESTIMATING A NORMAL MEAN. Suppose $X \sim \mathcal{N}(\theta, \sigma^2)$, σ^2 known, and π_0 is $\mathcal{N}(\theta_0, \tau^2)$, θ_0 and τ^2 given. Let $\mathcal{Q}$ be as in (3.1) and define $z = (x - \theta_0)/\sigma$. Following Berger and Sellke (1984) [see also Berger (1985)], $\hat{a}$ of Theorem 3.1 is

$$\hat{a} = \begin{cases} 0 & \text{if } z \le 1.65, \\ a^*\sigma & \text{if } z > 1.65, \end{cases} \tag{3.4}$$

where a^* satisfies the equation

$$\hat{a}^* = |z| + \Big[-2\log\Big(\sqrt{2\pi}\,\{[\Phi(a^* - |z|) - \Phi(-(a^* + |z|))]/a^* - \phi(-(a^* + |z|))\}\Big)\Big]^{1/2}, \tag{3.5}$$

Φ and ϕ denoting the standard normal c.d.f. and density, respectively. Equation (3.5) can be solved (usually very quickly) by standard fixed point iteration, starting on the right-hand side with initial value $a^* = |z|$.

CASE 1: $\hat{a} > 0$. It can be shown that the ML-II posterior mean and variance corresponding to the Uniform $(\theta_0 - \hat{a}, \theta_0 + \hat{a})$ prior are, respectively,

$$\delta^{\hat{q}}(x) = x - (\sigma/a^*)\tanh(za^*)$$

and

$$V^{\hat{q}}(x) = \sigma^2\big[z - (\sigma^2/a^*)\tanh(za^*)\big]\big[a^{*-1}\tanh(za^*)\big].$$

Furthermore,

$$\hat{\lambda}(x) = \Big[1 + \big(0.5\varepsilon(1 + \tau^2/\sigma^2)^{1/2}/(1 - \varepsilon)\big)(1 + \exp(-2za^*)) \times \exp\big\{-0.5\big((\tau^2z^2/(\sigma^2 + \tau^2)) + a^{*2} - 2za^*\big)\big\}\Big]^{-1},$$

so that the ML-II posterior mean and variance are [see (1.7) and (1.8)]

$$\delta^{\hat{\pi}}(x) = \hat{\lambda}(x)\big(x - (\sigma^2/(\sigma^2 + \tau^2))(x - \theta_0)\big) + (1 - \hat{\lambda}(x))\delta^{\hat{q}}(x) \tag{3.6}$$

and

$$V^{\hat{\pi}}(x) = \hat{\lambda}(x)\tau^2\sigma^2/(\tau^2 + \sigma^2) + (1 - \hat{\lambda}(x))V^{\hat{q}}(x) + \hat{\lambda}(x)(1 - \hat{\lambda}(x))\big[(\sigma/a^*)\tanh(za^*) - \sigma^2(x - \theta_0)/(\sigma^2 + \tau^2)\big]^2. \tag{3.7}$$

CASE 2: $\hat{a} = 0$. If $\hat{a} = 0$, then $\delta^{\hat{q}}(x) = \theta_0$ and $V^{\hat{q}}(x) = 0$. The formulas for $\hat{\lambda}$, $\delta^{\hat{\pi}}$, and $V^{\hat{\pi}}$ are then easily obtained and omitted here. In this case, however, though x is close to θ_0, it is probably undesirable to allow $\hat{\pi}$ to be more concentrated at θ_0 then π_0. The natural "fix-up" is to simply replace $(\delta^{\hat{\pi}}, V^{\hat{\pi}})$ by $(\delta^{\pi_0}, V^{\pi_0})$. In fact we generally recommend this modification whenever $V^{\hat{\pi}} < V^{\pi_0}$.

TABLE 1
ML-II results for unimodal, symmetric contaminations

x	$\hat{a}$	δ^{π_0}	$\hat{\lambda}$	$\delta^{\hat{\pi}}$	$V^{\hat{\pi}}$
3.00	4.13	2.000	0.661	2.257	0.796
5.00	6.43	3.333	0.166	4.594	1.054
7.00	8.61	4.667	4.7×10^{-3}	6.873	0.842
10.00	11.78	6.667	1.3×10^{-6}	9.915	0.851

As a specific example, suppose $\sigma^2 = 1$, $\theta_0 = 0$, $\tau^2 = 2$, and $\varepsilon = 0.2$. Values of ML-II quantities and $\delta^{\pi_0}(x) = 2x/3$ for various x are given in Table 1. Further numerical results are given in Section 6.

4. Unimodality preserving contaminations.

4.1. Introduction. Despite the simplicity of the analyses in Sections 2 and 3, there are some objections. In Section 2 we saw that choosing $\mathcal{Q} = \mathcal{P} = \{\text{all distributions}\}$ can cause serious problems, due to the unrealistic nature of some of the resultant priors in Γ. On the other hand, the restriction to symmetric unimodal contaminations in Section 3 could be criticized for not allowing certain plausible contaminations, particularly when π_0 is not symmetric. When $\Theta \subset \mathbb{R}^1$ and π_0 is unimodal, as will be assumed throughout this section, perhaps *the most appealing* $\mathcal{Q}$ is that which contains those contaminations which preserve the unimodality of $\pi = (1 - \varepsilon)\pi_0 + \varepsilon q$ (note that q need not be unimodal). *Any* such π would typically be plausible a priori, and virtually *all* π deemed reasonable a priori are in this class. Successfully working with such "minimally complete" Γ is a very desirable goal in Bayesian robustness.

The main goals of this section are to indicate that such complex classes can, surprisingly, be analyzed and to provide some basic mathematical techniques likely to be encountered in such analyses. The presentation here is restricted to the determination of the ML-II prior and the ML-II posterior mean and variance, a detailed example concerning the mean of a normal distribution, plus some discussion of results later in Section 6. We are currently investigating the more fundamental question of determining ranges of posterior measures as π varies over Γ.

The exact class Γ that will be considered is (where θ_0 denotes the mode of π_0, which we assume to be unique)

$$\Gamma = \{\pi = (1 - \varepsilon)\pi_0 + \varepsilon q\colon\ q \in \mathcal{Q},\ \text{the set of all probability densities for which } \pi \text{ is unimodal with (not necessarily unique) mode } \theta_0, \text{ and } \pi(\theta_0) \leq (1 + \varepsilon')\pi_0(\theta_0)\}.$$

The final inequality in the definition of Γ specifies the reasonable constraint that q not be allowed to concentrate too sharply near θ_0. (Usually it would be reasonable to select $\varepsilon' = \varepsilon$, but this is not necessary. Indeed the choice $\varepsilon' = 0$

might sometimes be desired, it having the attractive property of ensuring that the ML-II posterior variance never drops much below that of π_0 because of excessive prior concentration near θ_0.) We will also assume that the likelihood function $f(x|\theta)$ is unimodal (as a function of θ, of course) with unique mode $\hat{\theta}$. [Of course, x is fixed, so $f(x|\theta)$ need only be unimodal for the observed x, not for all x.] It will also be technically convenient to restrict consideration to π_0 and f which are nonzero and are strictly monotonic on each side of the modes. More general cases could be handled, but the results get messier. We also assume, without loss of generality, that $\hat{\theta} \geq \theta_0$.

4.2. The form of the ML-II prior and posterior. The formal calculation, in Sections 4.4 and 4.5, of the ML-II prior, $\hat{\pi}$, is complicated by the need to consider several different cases. The result, however, is always of the quite simple form

$$\hat{\pi}(\theta) = \begin{cases} K & \text{for } \theta \in B, \\ (1-\varepsilon)\pi_0(\theta) & \text{for } \theta \notin B, \end{cases} \tag{4.1}$$

i.e., $\hat{\pi}$ is uniform over B (which will be an interval about $\hat{\theta}$), and equals $(1-\varepsilon)\pi_0(\theta)$ outside of B. (Note that K is implicitly defined by the constraint that $\hat{\pi}$ have mass one.) Thus the ML-II posterior will be

$$\hat{\pi}(\theta|x) = \hat{\lambda}(x)\pi_0(\theta|x) + (1-\hat{\lambda}(x))\hat{q}(\theta|x), \tag{4.2}$$

where [letting $I_B(\theta)$ denote the usual indicator function on B]

$$\begin{aligned} \hat{q} &= \varepsilon^{-1}\left[K - (1-\varepsilon)\pi_0(\theta)\right] I_B(\theta), \\ \hat{\lambda}(x) &= (1-\varepsilon)m(x|\pi_0)/\left[(1-\varepsilon)m(x|\pi_0) + \varepsilon m(x|\hat{q})\right], \\ m(x|\hat{q}) &= \varepsilon^{-1}\int_B \left[K - (1-\varepsilon)\pi_0(\theta)\right] f(x|\theta)\, d\theta. \end{aligned} \tag{4.3}$$

The interesting case (Case 1 in Section 4.5) is that in which $|\hat{\theta} - \theta_0|$ is moderately large (i.e., where prior and likelihood are not in close agreement), since it is only in this case that the choice of $\pi \in \Gamma$ will have a substantial effect. As $|\hat{\theta} - \theta_0| \to \infty$, it can typically be shown that the uniform piece of $\hat{\pi}$ dominates, in the sense that

$$\hat{\pi}(\theta|x) \to f(x|\theta) \Big/ \int f(x|\theta)\, d\theta,$$

which would be the posterior for the noninformative uniform prior. This kind of behavior can be labelled "robust" from a number of viewpoints [cf. Berger (1984)], and is certainly more pleasing than the limiting behavior of $\hat{\pi}(\theta|x)$ in Section 2, which collapsed to a point at $\hat{\theta}$ in the limit. Discussion of the degree of "robustness" which is attained by $\hat{\pi}$ is delayed until Section 6.

4.3. The normal distribution. We present here an example of the overall theory, using the formulas from Sections 4.4 and 4.5. The example considered is that in which X is $\mathcal{N}(\theta, 1)$ and π_0 is $\mathcal{N}(0, \tau^2)$. [The more general case where $X \sim \mathcal{N}(\theta, \sigma^2)$ and π_0 is $\mathcal{N}(\mu, \tau^2)$, σ^2, τ^2, and μ all known, can be reduced to

this case by a linear transformation.] Only Case 1 will be considered (which here means that $|x|$ is larger than a certain constant depending on ε and τ) and we assume that $x > 0$. Let Φ denote the standard normal c.d.f., and ϕ the standard normal density function.

The set B in (4.1) is the interval $[\theta^*, w(\theta^*)]$, the endpoints being implicitly defined by the equations [noting $\theta^* < x < w(\theta^*)$]

$$(4.4)\quad \begin{aligned}&(1-\varepsilon)\phi\left(\frac{\theta^*}{\tau}\right)\left[\frac{w(\theta^*)}{\tau}-\frac{\theta^*}{\tau}\right]-(1-\varepsilon)\left[\Phi\left(\frac{w(\theta^*)}{\tau}\right)-\Phi\left(\frac{\theta^*}{\tau}\right)\right]=\varepsilon,\\ &\phi(w(\theta^*)-x)[w(\theta^*)-\theta^*]-[\Phi(w(\theta^*)-x)-\Phi(\theta^*-x)]=0.\end{aligned}$$

These equations can be easily solved by iteration. Calculation then gives that the ML-II posterior is

$$\hat{\pi}(\theta|x)=\begin{cases}\hat{\lambda}(x)C_0 f(x|\theta) & \text{if } \theta\in B,\\ \hat{\lambda}(x)\pi_0(\theta|x) & \text{if } \theta\notin B,\end{cases}$$

where

$$\pi_0(\theta|x) \text{ is } \mathcal{N}(\delta, V^2), \qquad \delta=\frac{\tau^2 x}{1+\tau^2}, \qquad V^2=\frac{\tau^2}{1+\tau^2}$$

(this notation is more convenient here than the previous $\delta^{\pi_0}, V^{\pi_0}$),

$$\hat{\lambda}(x)=\left[1-B_0+C_0\phi(w(\theta^*)-x)(w(\theta^*)-\theta^*)\right]^{-1},$$

$$C_0=\frac{(1+\tau^{-2})^{1/2}\phi(\theta^*/\tau)}{\phi\left(x[1+\tau^2]^{-1/2}\right)}, \quad \text{and} \quad B_0=\Phi\left(\frac{w(\theta^*)-\delta}{V}\right)-\Phi\left(\frac{\theta^*-\delta}{V}\right).$$

The ML-II posterior mean and variance are given, respectively, by

$$\delta^{\hat{\pi}}=\hat{\lambda}(x)\delta+[1-\hat{\lambda}(x)]\delta^{\hat{q}}$$

and

$$V^{\hat{\pi}}=\hat{\lambda}(x)V^2+[1-\hat{\lambda}(x)]V^{\hat{q}}+\hat{\lambda}(x)[1-\hat{\lambda}(x)][\delta-\delta^{\hat{q}}]^2,$$

where

$$\delta^{\hat{q}}=x+(C_0D_0-E_0+(x-\delta)B_0)\frac{\hat{\lambda}}{(1-\hat{\lambda})},$$

$$\begin{aligned}V^{\hat{q}}=\left[\hat{\lambda}(x)^{-1}-1\right]^{-1}\Big\{&-C_0D_0(2\delta^{\hat{q}}-x-\theta^*)+(x-\delta^{\hat{q}})^2\left[\hat{\lambda}(x)^{-1}-1+B_0\right]\\ &-(\theta^*+\delta-2\delta^{\hat{q}})E_0-\left[V^2+(\delta-\delta^{\hat{q}})^2\right]B_0\\ &+V[w(\theta^*)-\theta^*]\phi([w(\theta^*)-\delta]/V)\Big\},\end{aligned}$$

$$D_0=\phi(\theta^*-x)-\phi(w(\theta^*)-x),$$

and

$$E_0=V\left\{\phi\left(\frac{\theta^*-\delta}{V}\right)-\phi\left(\frac{w(\theta^*)-\delta}{V}\right)\right\}.$$

TABLE 2
ML-II quantities for various x

x	B	δ	$\hat{\lambda}$	$\delta^{\hat{\pi}}$	$V^{\hat{\pi}}$
1.75	(0, 2.53)	1.167	0.609	1.425	0.599
3.00	(1.39, 3.73)	2.000	0.375	2.581	0.616
5.00	(2.25, 6.08)	3.333	0.052	4.735	0.666
7.00	(2.64, 8.35)	4.667	1.5×10^{-3}	6.827	0.735
10.00	(2.97, 11.61)	6.667	4.5×10^{-7}	9.880	0.797

As a specific example, suppose $\tau^2 = 2$ and $\varepsilon = 0.2$. Then one can show [using (4.6)] that Case 1 occurs providing $|x| \geq 1.75$. Table 2 gives the relevant quantities above for various x.

The behavior alluded to earlier clearly obtains: as $|x|$ gets large, $\hat{\lambda} \to 0$, and the uniform part of the prior (on B) dominates. Also, $\delta^{\hat{\pi}}(x) \to x$ and $V^{\hat{\pi}}(x) \to 1$. Indeed, the following theorem gives large x approximations to the key quantities, approximations which are accurate, in the above example, for $x \geq 7$ and which show that the domination of the uniform portion occurs at an exponential rate. (The proof of the theorem is routine and will be omitted.)

THEOREM 4.1. *As* $x \to \infty$ (*log denotes natural logarithm*),

$$\theta^* = \left[2\tau^2 \log x\right]^{1/2} + o(1), \qquad w(\theta^*) = x + \left[2\log\left(x/\sqrt{2\pi}\right)\right]^{1/2} + o(1),$$

$$\hat{\lambda}(x) = (\varepsilon^{-1} - 1)x\left[(1+\tau^2)2\pi\right]^{-1/2}\exp\left\{-x^2/\left[2(1+\tau^2)\right]\right\}(1 + o(1)),$$

$$\delta^{\hat{q}}(x) = x - x^{-1}(1 + o(1)), \quad \text{and} \quad V^{\hat{q}}(x) = 1 - x^{-1}\left[2\log x\right]^{1/2}(1 + o(1)).$$

4.4. Preliminaries and notation for the general theory. For $-\varepsilon' \leq \rho \leq \varepsilon$, define $v(\rho) \geq \theta_0$, implicitly, by

$$\pi_0(\theta_0)(1-\rho)(v(\rho) - \theta_0) - (1-\varepsilon)\int_{\theta_0}^{v(\rho)} \pi_0(\theta)\, d\theta = \varepsilon, \tag{4.5}$$

and define

$$V(\rho) = f(x|v(\rho))(v(\rho) - \theta_0) - \int_{\theta_0}^{v(\rho)} f(x|\theta)\, d\theta. \tag{4.6}$$

For $\theta_0 \leq \theta$, define $w(\theta) \geq \theta$, implicitly, by

$$(1-\varepsilon)\pi_0(\theta)(w(\theta) - \theta) - (1-\varepsilon)\int_{\theta}^{w(\theta)} \pi_0(\xi)\, d\xi = \varepsilon, \tag{4.7}$$

and define

$$W(\theta) = f(x|w(\theta))(w(\theta) - \theta) - \int_{\theta}^{w(\theta)} f(x|\xi)\, d\xi. \tag{4.8}$$

LEMMA 4.1. (a) *The quantities* $v(\rho)$ *and* $w(\theta)$ *are well defined, unique, continuous, and strictly increasing for* $-\varepsilon' \leq \rho \leq \varepsilon$ *and* $\theta \geq \theta_0$.

(b) *If* $v(\rho) > \hat{\theta}$, *then* $V(\rho)$ *is decreasing in* ρ. *Furthermore,* $V(\rho) = 0$ *has at most one solution.*

(c) *If* $\theta_0 \le \theta \le \hat{\theta}$ *and* $w(\theta) > \hat{\theta}$, *then* $W(\theta)$ *is decreasing at* θ. *Furthermore, if* $V(\varepsilon) \ge 0$, *then* $W(\theta) = 0$ *has a unique solution* $\theta_0 \le \theta^* < \hat{\theta}$.

PROOF. (a) At $v = \theta_0$, the left-hand side of (4.5) is zero. As $v \to \infty$, the left-hand side of (4.5) goes to ∞. Finally, since π_0 is decreasing for $\theta > \theta_0$, the derivative, with respect to v, of the left-hand side of (4.5) is easily seen to be strictly positive. A solution to (4.5) thus exists and is unique.

To show that $v(\rho)$ is strictly increasing, one can differentiate both sides of (4.5) with respect to ρ and solve for $v'(\rho)$ [i.e., $d/d\rho\, v(\rho)$], obtaining

$$v'(\rho) = \pi_0(\theta_0)(v(\rho) - \theta_0)/[\pi_0(\theta_0)(1 - \rho) - (1 - \varepsilon)\pi_0(v(\rho))].$$

Since $v(\rho) > \theta_0$, $\rho < \varepsilon$, and π_0 is decreasing for $\theta > \theta_0$, it is clear that $v'(\rho) > 0$. The verification for $w(\theta)$ is very similar.

(b) Letting $f'(x|\theta) = d/d\theta\, f(x|\theta)$, calculation gives

$$\frac{d}{d\theta} V(\rho) = f'(x|v(\rho))v'(\rho)(v(\rho) - \theta_0).$$

Since f is decreasing for $\theta > \hat{\theta}$, the monotonicity result follows from part (a). If $V(\rho) = 0$, the unimodality of f ensures that $v(\rho) > \hat{\theta}$ [for otherwise the right-hand side of (4.6) is positive]. The strict monotonicity of V for such ρ ensures that any solution to $V(\rho) = 0$ must be unique.

(c) Letting $w'(\theta) = d/d\theta\, w(\theta)$, calculation gives

$$\frac{d}{d\theta} W(\theta) = f'(x|w(\theta))w'(\theta)(w(\theta) - \theta).$$

The monotonicity of f and part (a) show that this is negative. Using this, to show that $W(\theta) = 0$ has a unique solution, it is only necessary to show that $W(\theta_0) \ge 0$ and $W(\hat{\theta}) < 0$. Since $v(\varepsilon) = w(\theta_0)$, it follows that $W(\theta_0) = V(\varepsilon) \ge 0$ (by assumption). That $W(\hat{\theta}) < 0$ follows from (4.8) and an easy application of the mean value theorem [since $f(x|\theta)$ decreases for $\theta > \hat{\theta}$]. □

LEMMA 4.2. *Suppose* $V(\varepsilon) \ge 0$, *and let* $\theta_0 \le \theta^* \le \hat{\theta}$ *be the solution to* $W(\theta) = 0$. *Then*

(a) $f(x|\theta) < f(x|w(\theta^*))$ *for* $\theta \notin [\theta^*, w(\theta^*)]$.

(b) *For any nonincreasing integrable function* g *such that* $\int g(\theta)\, d\theta = 0$, *it follows that*

$$\int_{\theta^*}^{w(\theta^*)} g(\theta) f(x|\theta)\, d\theta \le 0. \tag{4.9}$$

PROOF. (a) Clearly $f(x|\theta^*) < f(x|w(\theta^*))$, for otherwise the integrand in (4.8) would be everywhere larger than $f(x|w(\theta^*))$ and $W(\theta^*)$ would be nonzero, a contradiction. The unimodality of f thus gives the result for $\theta < \theta^*$. Now $w(\theta^*) > \hat{\theta}$, for otherwise (4.8) could again be used to contradict $W(\theta^*) = 0$. The unimodality of f thus also gives the result for $\theta > w(\theta^*)$.

(b) Note first that it suffices to prove the result for differentiable g. Letting $h(\theta) = -d/d\theta\, g(\theta)$ (note $h \geq 0$) and writing $g(\theta) = K - \int_{\theta^*}^{\theta} h(\xi)\,d\xi$, where

$$
\begin{aligned}
K &= \frac{1}{[w(\theta^*) - \theta^*]} \int_{\theta^*}^{w(\theta^*)} \int_{\theta^*}^{\eta} h(\xi)\, d\xi\, d\eta \\
&= \frac{1}{[w(\theta^*) - \theta^*]} \int_{\theta^*}^{w(\theta^*)} (w(\theta^*) - \xi) h(\xi)\, d\xi,
\end{aligned} \tag{4.10}
$$

we obtain from Fubini's theorem

$$
\begin{aligned}
\int_{\theta^*}^{w(\theta^*)} g(\theta) f(x|\theta)\, d\theta &= K \int_{\theta^*}^{w(\theta^*)} f(x|\theta)\, d\theta \\
&\quad - \int_{\theta^*}^{w(\theta^*)} h(\xi) \int_{\xi}^{w(\theta^*)} f(x|\theta)\, d\theta\, d\xi.
\end{aligned} \tag{4.11}
$$

Next we show that, for $\theta^* < \xi < w(\theta^*)$,

$$
\psi(\xi) \equiv \int_{\xi}^{w(\theta^*)} f(x|\theta)\, d\theta \geq (w(\theta^*) - \xi) f(x|w(\theta^*)). \tag{4.12}
$$

For $\xi \geq \hat{\theta}$ this is a trivial consequence of the monotonicity of f. For $\theta^* < \xi < \hat{\theta}$, note that $\psi(\xi)$ is concave [$f(x|\xi)$ is increasing here] and that

$$
\psi(\theta^*) = \int_{\theta^*}^{w(\theta^*)} f(x|\theta)\, d\theta = (w(\theta^*) - \theta^*) f(x|w(\theta^*)) \tag{4.13}
$$

[since $W(\theta^*) = 0$]. Hence, $\psi(\xi)$ must lie above the line $(w(\theta^*) - \xi) f(x|w(\theta^*))$, establishing (4.12). Using (4.12) in (4.11) we get that

$$
\begin{aligned}
&\int_{\theta^*}^{w(\theta^*)} g(\theta) f(x|\theta)\, d\theta \\
&\quad \leq K \int_{\theta^*}^{w(\theta^*)} f(x|\theta)\, d\theta - f(x|w(\theta^*)) \int_{\theta^*}^{w(\theta^*)} (w(\theta^*) - \xi) h(\xi)\, d\xi,
\end{aligned}
$$

the right-hand side of which is zero by (4.10) and (4.13). □

4.5. The ML-II prior. Define $\hat{\pi}$ as follows:

CASE 1. If $V(\varepsilon) \geq 0$, and $\theta^* \in [\theta_0, \hat{\theta}]$ is the solution to $W(\theta) = 0$, let

$$
\hat{\pi}(\theta) = \begin{cases} (1 - \varepsilon)\pi_0(\theta^*) & \text{for } \theta^* \leq \theta \leq w(\theta^*), \\ (1 - \varepsilon)\pi_0(\theta) & \text{otherwise.} \end{cases} \tag{4.14}
$$

CASE 2. If $V(\varepsilon) < 0$ but $V(-\varepsilon') \geq 0$, find $\rho^* \in [-\varepsilon', \varepsilon]$ so that $V(\rho^*) = 0$, and let

$$
\hat{\pi}(\theta) = \begin{cases} (1 - \rho^*)\pi_0(\theta_0) & \text{for } \theta_0 \leq \theta \leq v(\rho^*), \\ (1 - \varepsilon)\pi_0(\theta) & \text{otherwise.} \end{cases} \tag{4.15}
$$

CASE 3. If $V(-\varepsilon') < 0$ and $f(x|\theta_0) \leq f(x|v(-\varepsilon'))$, let $\hat{\pi}$ be as in Case 2 with $\rho^* = -\varepsilon'$.

CASE 4. If $V(-\varepsilon') < 0$ and $f(x|\theta_0) > f(x|v(-\varepsilon'))$, let

$$\hat{\pi}(\theta) = \begin{cases} (1+\varepsilon')\pi_0(\theta_0) & \text{for } \theta' \le \theta \le \theta'', \\ (1-\varepsilon)\pi_0(\theta) & \text{otherwise,} \end{cases} \tag{4.16}$$

where θ' and θ'' are the (unique) solutions to the equations

$$\begin{aligned} &f(x|\theta') = f(x|\theta''), \\ &(1+\varepsilon')\pi_0(\theta_0)(\theta''-\theta') - (1-\varepsilon)\int_{\theta'}^{\theta''} \pi(\theta)\,d\theta = \varepsilon. \end{aligned} \tag{4.17}$$

Lemma 4.1 establishes that all quantities involved in the definition of $\hat{\pi}$ are well defined and unique. (The existence and uniqueness of θ' and θ'' in Case 4 is easy to establish.) Observe that, in all cases, $\hat{\pi}$ has a very simple and easy to work with form of being uniform in a certain interval, and otherwise being equal to $(1-\varepsilon)\pi_0$. Case 1 corresponds to the situation where the elicited prior, π_0, and the likelihood function, $f(x|\theta)$, are moderately separated, Case 2 to the situation where they are fairly close, and Cases 3 and 4 to situations where they are very close.

THEOREM 4.2. *The $\hat{\pi}$ defined in* (4.14)–(4.16) *is the ML-II prior in* Γ.

PROOF. We only present the argument for Case 1, the other cases being very similar. The goal is to show that

$$m(x|\pi) - m(x|\hat{\pi}) = \int [\pi(\theta) - \hat{\pi}(\theta)]\, f(x|\theta)\,d\theta \le 0 \tag{4.18}$$

for all $\pi \in \Gamma$. Letting $g(\theta) = \pi(\theta) - \hat{\pi}(\theta)$, note that

(i) $g(\theta) \ge 0$ for $\theta \notin [\theta^*, w(\theta^*)]$, since $\hat{\pi}(\theta) = (1-\varepsilon)\pi_0(\theta)$ here and $\pi(\theta) \ge (1-\varepsilon)\pi_0(\theta)$;

(ii) $g(\theta)$ is nonincreasing on $[\theta^*, w(\theta^*)]$, since $\hat{\pi}(\theta)$ is uniform on this interval and so $\pi(\theta) = g(\theta) + \hat{\pi}(\theta)$ would have a secondary mode were $g(\theta)$ somewhere increasing;

(iii) $K \equiv \int_{\theta^*}^{w(\theta^*)} g(\theta)\,d\theta = -\int_{[\theta^*, w(\theta^*)]^c} g(\theta)\,d\theta$.

Lemma 4.2(a) and (i) show that

$$\int_{[\theta^*, w(\theta^*)]^c} g(\theta) f(x|\theta)\,d\theta < f(x|w(\theta^*))(-K).$$

Lemma 4.2(b) and (ii) imply that

$$\int_{\theta^*}^{w(\theta^*)} \left(g(\theta) - \frac{K}{[w(\theta^*) - \theta^*]} \right) f(x|\theta)\,d\theta \le 0.$$

Thus

$$\int g(\theta) f(x|\theta)\,d\theta < f(x|w(\theta^*))(-K) + \frac{K}{[w(\theta^*) - \theta^*]} \int_{\theta^*}^{w(\theta^*)} f(x|\theta)\,d\theta. \tag{4.19}$$

Since $W(\theta^*) = 0$, the right-hand side of (4.19) is zero, and (4.18) follows. □

COMMENTS. 1. The key step in the proof of Theorem 4.2 is really Lemma 4.2(b), which shows that one cannot improve on a uniform $\hat{\pi}$ on $[\theta^*, w(\theta^*)]$. 2. The problem might be susceptible to attack through calculus of variations, since one is trying to maximize an expression involving an integral of π over a class of π. The difficulty is that the $\pi \in \Gamma$ satisfy a large number of inequality and differential inequality constraints. Calculus of variations with such side constraints is quite difficult.

5. Hierarchical classes of priors.

5.1. Introduction. Hierarchical priors are typically employed when $\boldsymbol{\theta}$ is a vector $(\theta_1, \theta_2, \ldots, \theta_p)$, and the θ_i are thought to be independent realizations from a common prior distribution g. Typically g is assumed to lie in some class $\Gamma_1 = \{g_\omega\colon\ \omega \in \Omega\}$ of distributions, often the class of conjugate priors, and a "second stage" prior h_0 is placed on this class, i.e., on ω. Such a hierarchical prior can, of course, be written as a single prior, namely

$$\pi_0(\boldsymbol{\theta}) = \int_\Omega \left[\prod_{i=1}^{p} g_\omega(\theta_i)\right] h_0(\omega)\, d\omega. \tag{5.1}$$

(We restrict ourselves to densities in this section, for convenience, and also will not consider hierarchical priors with more than two stages.) Development of and references for this approach can be found in Good (1980), Lindley and Smith (1972), Morris (1983), and Berger (1985).

There are three possible robustness concerns in working with (5.1). One could question the assumptions (i) that the θ_i are i.i.d.; (ii) that the prior g belongs to Γ_1; and (iii) that h_0 is specified correctly. Each of these concerns deserves careful consideration separately but in the following we will simply deal with uncertainty in the second stage (i.e., h_0), or in both the first and second stages together.

Simultaneous uncertainty in different stages or aspects of a prior can often be expressed most simply by allowing more than one contamination in the ε-contamination model. For instance, one could consider

$$\Gamma = \{\pi\colon\ \pi = (1 - \varepsilon_1 - \varepsilon_2)\pi_0 + \varepsilon_1 q_1 + \varepsilon_2 q_2,\ q_1 \in \mathcal{Q}_1,\ q_2 \in \mathcal{Q}_2\}, \tag{5.2}$$

where $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are appropriate possible classes of contaminations. Such an extension of the ε-contamination model vastly increases its flexibility while causing no real hardship in many applications, because the important formulas (1.4), (1.7), and (1.8) become simply

$$m(x|\pi) = (1 - \varepsilon_1 - \varepsilon_2)m(x|\pi_0) + \varepsilon_1 m(x|q_1) + \varepsilon_2 m(x|q_2), \tag{5.3}$$

$$\delta^\pi(x) = \left[1 - \lambda_1(x) - \lambda_2(x)\right]\delta^{\pi_0}(x) + \lambda_1(x)\delta^{q_1}(x) + \lambda_2(x)\delta^{q_2}(x), \tag{5.4}$$

$$\begin{aligned} V^\pi(x) = {} & (1 - \lambda_1 - \lambda_2)V^{\pi_0} + \lambda_1 V^{q_1} + \lambda_2 V^{q_2} + \lambda_1\lambda_2(\delta^{q_1} - \delta^{q_2})^2 \\ & + (1 - \lambda_1 - \lambda_2)\lambda_1(\delta^{\pi_0} - \delta^{q_1})^2 + (1 - \lambda_1 - \lambda_2)\lambda_2(\delta^{\pi_0} - \delta^{q_2})^2, \end{aligned} \tag{5.5}$$

where $\lambda_i(x) = \varepsilon_i m(x|q^i)/m(x|\pi)$ for $i = 1, 2$. Thus one can find the ML-II prior

by separately maximizing $m(x|q_1)$ and $m(x|q_2)$ in (5.3) (unless $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are related in some fashion) and then easily calculate the resultant ML-II posterior mean and variance.

Before proceeding, it is worthwhile to note that Γ of the form (5.2) might be of interest in other than hierarchical prior situations. Indeed, whenever one has several possible models in mind for the contamination, or even for π_0 itself, the uncertainty can be reasonably represented by such a Γ.

5.2. Second stage uncertainty. Suppose, in the situation of Section 5.1, that only h_0 is deemed uncertain. (Knowledge at higher levels of hierarchical priors will often be more vague than at lower levels.) An ε-contamination model for h would be

$$h(\omega) = (1 - \varepsilon)h_0(\omega) + \varepsilon s(\omega), \qquad s \in \mathcal{S}. \tag{5.6}$$

The resulting prior for θ is

$$\pi(\boldsymbol{\theta}) = \int \left[\prod_{i=1}^{p} g_\omega(\theta_i) \right] h(\omega)\, d\omega = (1 - \varepsilon)\pi_0(\boldsymbol{\theta}) + \varepsilon q(\boldsymbol{\theta}), \tag{5.7}$$

where

$$\pi_0(\boldsymbol{\theta}) = \int \left[\prod_{i=1}^{p} g_\omega(\theta_i) \right] h_0(\omega)\, d\omega \quad \text{and} \quad q(\boldsymbol{\theta}) = \int \left[\prod_{i=1}^{p} g_\omega(\theta_i) \right] s(\omega)\, d\omega.$$

Letting $\mathcal{Q} = \{q\colon s \in \mathcal{S}\}$, it follows that the uncertainty in π can be expressed by $\Gamma = \{\pi\colon \pi = (1 - \varepsilon)\pi_0 + \varepsilon q,\ q \in \mathcal{Q}\}$.

In determining the ML-II prior for this situation, it will be convenient to define

$$m(x|\omega) = \int f(x|\boldsymbol{\theta}) \left[\prod_{i=1}^{p} g_\omega(\theta_i) \right] d\boldsymbol{\theta},$$

which is clearly the marginal distribution of X under the assumption that the prior for $\boldsymbol{\theta}$ is $[\prod_{i=1}^{p} g_\omega(\theta_i)]$. Note that

$$m(x|\pi) = (1 - \varepsilon)m(x|\pi_0) + \varepsilon \int m(x|\omega)s(\omega)\, d\omega. \tag{5.8}$$

When $\mathcal{S} = \mathcal{P} = \{\text{all distributions}\}$, it is clear from (5.8) that

$$\sup_{\pi \in \Gamma_1} m(x|\pi) = (1 - \varepsilon)m(x|\pi_0) + \varepsilon \sup_{\omega} m(x|\omega).$$

Assuming that $m(x|\omega)$ has a maximum at $\hat{\omega}$, it follows that the ML-II prior is

$$\hat{\pi}(\boldsymbol{\theta}) = (1 - \varepsilon)\pi_0(\boldsymbol{\theta}) + \varepsilon \left[\prod_{i=1}^{p} g_{\hat{\omega}}(\theta_i) \right],$$

for which analysis is usually quite straightforward.

EXAMPLE 3. Suppose that $X = (X_1, \ldots, X_p) \sim \mathcal{N}_p(\boldsymbol{\theta}, \sigma^2 I_p)$, σ^2 known, and that the first-stage prior information is that the θ_i are independent with a

common $\mathcal{N}(\mu, \tau^2)$ distribution, to be denoted g_ω, with $\omega = (\mu, \tau^2)$ unknown. Note that $m(x|\omega)$ is $\mathcal{N}_p(\mu\mathbf{1}, (\sigma^2 + \tau^2)I_p)$, where $\mathbf{1} = (1, \ldots, 1)$. It is easy to check that $m(x|\omega)$ is maximized at

$$\hat{\omega} = (\hat{\mu}, \hat{\tau}^2) = \left(\bar{x}, \max\left[0, \frac{1}{p}\sum_{i=1}^{p}(x_i - \bar{x})^2 - \sigma^2\right]\right).$$

Hence, with contaminated second-stage prior as in (5.6) and $\mathcal{S} = \mathcal{P}$, the ML-II prior is $\hat{\pi}(\boldsymbol{\theta}) = (1 - \varepsilon)\pi_0(\boldsymbol{\theta}) + \varepsilon\hat{q}(\boldsymbol{\theta})$, where $\hat{q}$ is $\mathcal{N}_p(\hat{\mu}\mathbf{1}, \hat{\tau}^2 I_p)$.

As a very special case, suppose h_0 is a point mass at (μ_0, τ_0^2), so that π_0 is simply $\mathcal{N}_p(\mu_0\mathbf{1}, \tau_0^2 I_p)$. Then the ML-II posterior is

$$\hat{\pi}(\boldsymbol{\theta}|x) = \hat{\lambda}(x)\pi_0(\boldsymbol{\theta}|x) + (1 - \hat{\lambda}(x))\hat{q}(\boldsymbol{\theta}|x),$$

where $\pi_0(\boldsymbol{\theta}|x)$ is $\mathcal{N}_p(\delta^{\pi_0}(x), v_0 I_p)$, $\hat{q}(\boldsymbol{\theta}|x)$ is $\mathcal{N}_p(\delta^{\hat{q}}, \hat{v} I_p)$, $v_0 = \sigma^2\tau_0^2/(\sigma^2 + \tau_0^2)$, $\hat{v} = \sigma^2\hat{\tau}^2/(\sigma^2 + \hat{\tau}^2)$,

$$\delta^{\pi_0}(x) = x - \frac{\sigma^2}{\sigma^2 + \tau_0^2}(x - \mu_0\mathbf{1}), \qquad \delta^{\hat{q}}(x) = x - \frac{\sigma^2}{\sigma^2 + \hat{\tau}^2}(x - \hat{\mu}\mathbf{1}),$$

and

$$\begin{aligned}\hat{\lambda}(x) &= (1 - \varepsilon)m(x|\pi_0)/[(1 - \varepsilon)m(x|\pi_0) + \varepsilon m(x|\hat{q})] \\ &= \left[1 + \frac{\varepsilon}{(1 - \varepsilon)} \cdot (\sigma^2 + \tau_0^2)^{p/2}\exp\left\{\sum_{i=1}^{p}(x_i - \mu_0)^2/[2(\sigma^2 + \tau_0^2)]\right\}\rho(x)\right]^{-1},\end{aligned}$$

where

$$\rho(x) = \begin{cases} \sigma^{-p}\exp\left\{-\sum_{i=1}^{p}(x_i - \bar{x})^2/2p\right\} & \text{if } \sum(x_i - \bar{x})^2 < p\sigma^2, \\ \left[\frac{1}{p}\sum_{i=1}^{p}(x_i - \bar{x})^2\right]^{-p/2}\exp\{-\tfrac{1}{2}p\} & \text{otherwise.}\end{cases}$$

Note that δ^{π_0} is the usual conjugate prior estimate of θ, while $\sigma^{\hat{q}}$ is the usual empirical Bayes estimate of θ. The overall posterior mean [see (1.7)] is thus

$$\delta^{\hat{\pi}} = \hat{\lambda}(x)\delta^{\pi_0}(x) + (1 - \hat{\lambda}(x))\delta^{\hat{q}}(x),$$

which will be close to δ^{π_0} if the x_i are close to μ_0, and close to $\delta^{\hat{q}}$ if the x_i are similar but far from μ_0.

Of course, only rarely will it be appropriate to choose h_0 to be a point mass. More natural would be a choice such as $h_0(\mu, \tau^2) = w(\mu)v(\tau^2)$, where $w(\mu)$ is $\mathcal{N}(\mu_0, A)$ and v is, say, a gamma distribution. Although the ML-II posterior, is no longer expressible in closed form for such a situation, the posterior mean and variance can be written in a form involving a single numerical integral over τ^2 [see, e.g., Lindley (1971)].

Several features of the above example are worth noting. First, the strong relationship of the ML-II theory with standard empirical Bayes analysis is apparent. Indeed, if one were to choose $\varepsilon = 1$, the standard empirical Bayes

situation would result. As mentioned in the introduction, we much prefer the analysis with reasonably small ε, the choice $\varepsilon = 1$ resulting (typically) in there being a large number of unrealistic priors in Γ. Of course, the choice $\mathcal{Q} = \mathcal{P}$ also suffers somewhat from this deficiency, as discussed in Section 2.3. An appealing possibility in the above example is, therefore, to attempt to apply the ideas of Section 3 (or possibly Section 4) and work with more reasonable $\mathcal{Q}$. For instance, if independence of μ and τ^2 can be assumed, so that $h(\mu, \tau^2) = w(\mu)v(\tau^2)$, one could elicit w_0 and v_0, consider

$$\mathcal{W} = \{w = (1 - \varepsilon_1)w_0 + \varepsilon_1 q_w\text{: } q_w \text{ is unimodal, symmetric about the mode (or perhaps median) of } w_0\}$$

$$\mathcal{V} = \{v = (1 - \varepsilon_2)v_0 + \varepsilon_2 q_v\text{: } q_v \text{ is unimodal, symmetric about the mode (or median) of } v_0\},$$

and apply the ideas of Section 3. We do not attempt the analysis here, because nothing new conceptually is involved and the argument would be moderately lengthy.

5.3. First and second stage uncertainty. The simplest modification of (5.7) that introduces uncertainty in the first stage of the prior is simply to add an arbitrary overall contamination. Thus we consider

$$\pi(\theta) = (1 - \varepsilon_1 - \varepsilon_2)\pi_0 + \varepsilon_1 q_1 + \varepsilon_2 q_2,$$

where $q_2 \in \mathcal{Q}_2 = \mathcal{P}$,

$$q_1 = \int \left[\prod_{i=1}^{p} g_\omega(\theta_i)\right] s(\omega)\, d\omega \in \mathcal{Q}_1 = \{q\text{: } s \in \mathcal{S}\},$$

and π_0, s, and $\mathcal{S}$ are as in (5.7). In other words, q_1 arises from possible second stage prior uncertainty, while q_2 allows for basic error in the empirical Bayes model.

Allowing arbitrary q_2 is again, probably excessively crude. In particular, complete abandonment of the empirical Bayes structure may be unrealistic. For illustrative purposes, however, this is convenient.

As mentioned in Section 5.1, the ML-II prior can be found (here, at least) by separately maximizing $m(x|q_1)$ and $m(x|q_2)$. Maximization of $m(x|q_1)$ was discussed in the previous section. And $m(x|q_2)$ will simply by maximized when q_2 is a unit point mass at $\hat{\boldsymbol{\theta}}$, the maximum likelihood estimate. Thus the ML-II prior is [assuming $\mathcal{S} = \mathcal{P}$ and letting $I(\hat{\boldsymbol{\theta}})$ denote a unit point mass at $\hat{\boldsymbol{\theta}}$]

$$\hat{\pi}(\boldsymbol{\theta}) = (1 - \varepsilon_1 - \varepsilon_2)\pi_0(\boldsymbol{\theta}) + \varepsilon_1\left[\prod_{i=1}^{p} g_{\hat{\omega}}(\theta_i)\right] + \varepsilon_2 I(\hat{\boldsymbol{\theta}}).$$

Formulas (5.3)–(5.5) can now easily be employed to give desired conclusions. In the situation of Example 3, for instance, all calculations can be carried out explicitly; indeed, the needed modifications to the formulae there are very minor and so will be omitted. The behavior of $\delta^{\hat{\pi}}$, the ML-II posterior mean, is worth

mentioning, however. If the data are compatible with π_0 (i.e., are near μ_0) then the conjugate prior posterior mean δ^0 will dominate; if the data are similar but not near μ_0, then $\delta^{\hat{\pi}}$ will be close to the natural empirical Bayes rule $\hat{\delta}^q$; and if the data are not compatible with the empirical Bayes model, then $\delta^{\hat{\pi}}$ will be close to the maximum likelihood estimate, $\hat{\theta} = x$.

6. Discussion. We view this paper as a hopeful first step in the development of systematic robust Bayesian analyses for rich classes of priors. The original goals of the paper were (i) to demonstrate that it is possible to work with complex classes of priors (as in Section 4), and to indicate mathematical techniques for doing so; (ii) to point out the numerous intuitive and calculational reasons for approaching robustness through consideration of ε-contamination classes; and (iii) to exhibit the value of the ML-II approach in obtaining "robust priors." Through tenacious prodding from skeptical referees and the associate editor, we have become more cautious in our assessment of success in goal (iii). A brief discussion of this issue is in order.

The key to obtaining a robust prior appears to be the selection of a prior with tails that are much flatter than the tails of the likelihood function [see Berger (1984, 1985) for discussion and references]. Unfortunately, this observation does not provide a readily implementable "solution" to robustness questions. Basic difficulties include (i) the uncertainty as to the choice of robust prior tails and (ii) the calculational complexities that can result. In addition to the already established computational simplicity, our hope for the ML-II technique was that it would *automatically* provide a "prior" robust against the type of deviations considered plausible.

It is important to explain our reasons for believing that ML-II would succeed when applied to ε-contamination classes. Suppose that ε is fairly small and that Γ contains all plausible priors, but none that are terribly implausible. Consider first the case where $m(x|\pi_0)$ is large, i.e., the data is compatible with the nominally specified π_0. This is a situation of nearly automatic robustness, in that, since the central portions of all π in Γ (and, hence, $\hat{\pi}$) will be similar to that of π_0, the conclusions will be very similar for all π in Γ. On the other hand, consider the case where x is such that $m(x|\pi_0)$ is small. This is precisely the situation in which the prior tail is highly influential and the use of a large prior tail is desirable. The ML-II technique will naturally select a prior with a large tail, since such priors are those most compatible with the data. Opposing the above encouraging tendencies toward robustness, is the danger that data-selected priors will tend to over-concentrate about the likelihood function, thereby yielding error estimates that are too small. (Such dangers lurk in the shadows of much of empirical Bayesian analysis and for that matter the use of data-selected models.) When Γ contains unreasonable priors, as in Section 2, this problem can completely dominate. However, the hope was that reasonable Γ, such as those in Sections 3 and 4, by limiting the possible concentration about the likelihood function, would not succumb to this danger to a serious extent.

To examine the degree to which this hope was realized, we return to the example discussed in Sections 2–4: $X \sim \mathcal{N}(\theta, 1)$, π_0 is $\mathcal{N}(0, 2)$, and $\varepsilon = 0.2$. The

four estimators of θ presented were (changing notation for convenience):

δ_1, the Bayes estimator with respect to π_0;
δ_2, the ML-II posterior mean for $\mathcal{Q}_2 = \{\text{all distributions}\}$;
δ_3, the ML-II posterior mean for $\mathcal{Q}_3 = \{\text{all symmetric, unimodal distributions}\}$;
δ_4, the ML-II posterior mean for $\mathcal{Q}_4 = \{q \text{ such that } \pi \text{ is unimodal}\}$.

Let us add to this list δ_5, the posterior mean for a t-prior distribution with 4 degrees of freedom, median zero, and quartiles equal to ± 0.96 (the quartiles of π_0); this corresponds to a scale factor of 1.3 for the t-distribution. The point of including the t-prior is that it is a reasonable robust prior (which happens to be in all the classes Γ for $\mathcal{Q}_2$, $\mathcal{Q}_3$, and $\mathcal{Q}_4$), and provides a benchmark for judging the

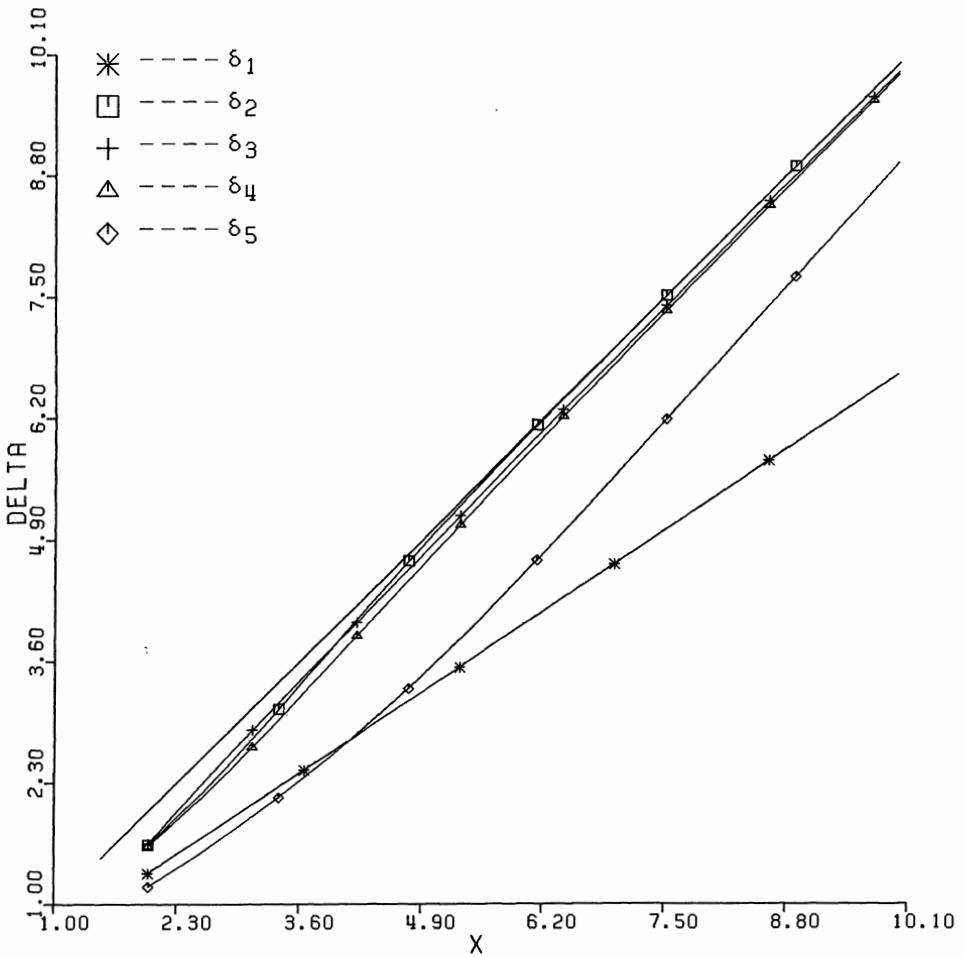

FIG. 1.

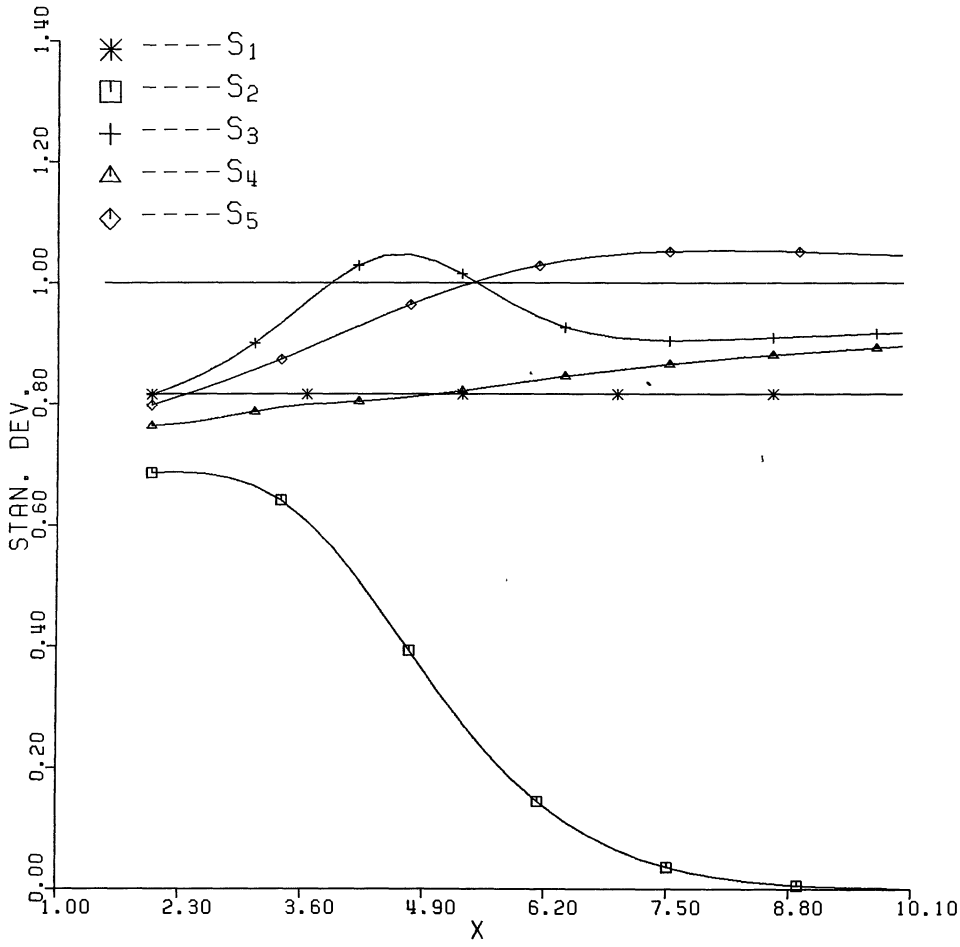

FIG. 2.

performance of ML-II. Figure 1 presents graphs of δ_1 through δ_5 for $2 \leq x \leq 10$ (there is very little difference between the estimators for $x < 2$). Figure 2 presents the corresponding natural error estimates, namely, the posterior (ML-II versions for δ_2, δ_3, and δ_4) standard deviations $s_1(x)$, $s_2(x), \ldots, s_5(x)$.

The most obvious conclusion from Figure 1 is that the ML-II estimates are more conservative (closer to x) than the Bayes estimates. Also, quite naturally, larger $\mathcal{Q}$ result in more conservative estimates (note that $\mathcal{Q}_3 \subset \mathcal{Q}_4 \subset \mathcal{Q}_2$).

Turning to Figure 2, note first that $s_2(x)$ is indeed fairly ridiculous for large x. Also, $s_4(x)$ is moderately smaller than $s_5(x)$. We had hoped that the attractive Γ of Section 4 would not be so large as to result in an excessively small ML-II error estimate, but when compared to s_5, the behavior of s_4 is borderline. On the other hand, the behavior of s_3 is definitely satisfactory. Finally, note that for moderate x, both s_3 and s_5 rise above the conditional standard deviation (1) of X. [The

behavior of s_5 is in fact quite general for robust priors; for example, see O'Hagan, (1981)].

To summarize, we proposed the ML-II technique as a possible automatic "robustifier" of standard (conjugate) priors. We have shown that the ML-II technique is calculationally feasible, and that it can successfully robustify π_0 if the class Γ is sensible (i.e., does not contain silly contaminations such as those in $\mathcal{Q}_2$). Alternatively, in any given problem one could (should?) attempt to construct a robust prior, such as the t-prior in the above example. We are in no way opposed to such efforts, but there are numerous technical and theoretical issues involved in insuring that robustness is obtained; the process is far from automatic. While we are not strong proponents of "automation" in statistics (few Bayesians are), we recognize the forces driving statistics in that direction. Of course, nothing is completely automatic; our ML-II approach does require the imputation of ε and $\mathcal{Q}$. However, it will often be reasonable to choose $\mathcal{Q}$ in a standard or default fashion, say as in Section 3. The quantity ε is reasonably accessible to intuition, relating in a fairly straightforward manner to one's confidence in the specification of π_0.

Acknowledgments. We would like to thank the referees and Associate Editor for very valuable comments and suggestions.

REFERENCES

BERGER, J. (1982). Bayesian robustness and the Stein effect. *J. Amer. Statist. Assoc.* **77** 358–368.

BERGER, J. (1984). The robust Bayesian viewpoint (with Discussion). In *Robustness in Bayesian Statistics* (J. Kadane, ed.). North-Holland, Amsterdam.

BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer, New York.

BERGER, J. and BERLINER, M. (1983). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. Technical Report 83-35, Purdue Univ., West Lafayette, Indiana.

BERGER, J. and BERLINER, M. (1984). Bayesian input in Stein estimation and a new minimax empirical Bayes estimator. *J. Econometrics* **25** 87–108.

BERGER, J. and SELLKE, T. (1984). Testing a point null hypothesis: the irreconcilability of significance levels and evidence. Technical Report 84-27, Purdue Univ., West Lafayette, Indiana.

BICKEL, P. J. (1984). Parametric robustness or small biases can be worthwhile. *Ann. Statist.* **12** 864–879.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis.* M.I.T. Press, Cambridge.

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. Ser. A* **143** 383–430.

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

DEMPSTER, A. P. (1975). A subjectivist look at robustness. *Bull. Inst. Internat. Statist.* **46** 349–374.

DEROBERTIS, L. and HARTIGAN, J. A. (1981). Bayesian inference using intervals of measures. *Ann. Statist.* **9** 235–244.

GOOD, I. J. (1950). *Probability and the Weighing of Evidence.* Griffin, London.

GOOD, I. J. (1965). *The Estimation of Probabilities.* M.I.T. Press, Cambridge.

GOOD, I. J. (1980). Some history of the hierarchical Bayesian methodology. In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.). University Press, Valencia.

HUBER, P. J. (1973). The use of Choquet capacities in statistics. *Bull. Inst. Internat. Statist.* **45** 181–191.

JEFFREYS, H. (1961). *Theory of Probability.* 3rd ed. University Press, Oxford.

KADANE, J. B., DICKEY, J. M., WINKLER, R. L., SMITH, W. S. and PETERS, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *J. Amer. Statist. Assoc.* **75** 845–854.

LEAMER, E. E. (1978). *Specification Searches.* Wiley, New York.

LEONARD, T. (1974). A modification to the Bayes estimate for the mean of a normal distribution. *Biometrika* **61** 627–628.

LINDLEY, D. V. (1971). The estimation of many parameters. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Holt, Rinehart, and Winston, Toronto.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41.

MARAZZI, A. (1985). Robust Bayesian estimation for the linear model. *Statistics and Decisions* **3**.

MARITZ, J. S. (1970). *Empirical Bayes Methods.* Methuen, London.

MORRIS, C. (1983). Parametric empirical Bayes inference: theory and applications (with Discussion). *J. Amer. Statist. Assoc.* **78** 47–65.

O'HAGAN, A. (1981). A moment of indecision. *Biometrika* **68** 329–330.

SCHNEEWEISS, H. (1964). Eine Entscheidungsregel für den Fall partiell bekannter Wahrscheinlichkeiten. *Unternehmensforschung* **8** 86–95.

ZELLNER, A. (1985). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques with Applications* (P. K. Goel and A. Zellner, eds.). North-Holland, Amsterdam.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907

DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
1958 NEIL AVENUE
COLUMBUS, OHIO 43210