Robust Bayes and Empirical Bayes Analysis with ε-Contaminated Priors[1]

by

James O. Berger and L. M. Berliner

Purdue University
Ohio State University

Technical Report #83-35

Department of Statistics
Purdue University
West Lafayette, IN 47907

September, 1983

[1] Research supported by the National Science Foundation, Grant MCS-8101670A1.

Abstract

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, and (ii) 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 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.

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, hierarchial priors.

Table of Contents

Robust Bayes and Empirical Bayes Analysis with ε-Contaminated Priors

1. Introduction.

1.1 The Robust Bayesian Viewpoint.

The most frequent criticism of subjective Bayes 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 (1983). (We also do not mean to imply that the single prior Bayesian approach is necessarily bad; it usually works very well.)

Very related to this are various forms of empirical Bayes analysis (c.f. Morris (1983) for discussion and review), in which the prior distribution is also assumed to belong to some class Γ of distributions. Perhaps the major perceived difference between empirical Bayes and robust Bayes analysis is that the former mostly deals with situations in which the data provides the bulk of the information about which priors should be used (the situation is such that the prior can actually be estimated from the data), while the latter usually considers situations in which subjective information is the greatest component of prior information. Since the situations we consider will mostly be situations in which the subjective component dominates, we will call the analysis robust Bayes. Empirical Bayes terminology could just as well have been used, however. Indeed, Section 4 considers some familiar empirical Bayes problems.

There is a wide variety of methods for implementing the robust Bayesian viewpoint. These methods are mainly distinguished by two features: (i) the form of the class Γ considered, and (ii) the method of utilizing Γ to arrive at a conclusion.

Before discussing these two points, 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 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 predictive or 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 $\mathbb{P}$ denote the space of all probability distirbutions on Θ.

1.2 Classes of Priors.

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

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

where $0 \leq \varepsilon \leq 1$ is given, π_0 is a particular prior distribution, and $\mathcal{Q}$ is some subset of $\mathcal{P}$. 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 (c.f. Kadane, et.al. (1980)) makes specification of an initial believable prior, π_0, an attractive starting point. (Because of the subsequent robustification, π_0 can often be chosen to be of some convenient functional form: for example, a conjugate prior.) 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. Also, the commonly considered class, Γ, of all conjugate priors, is usually absurdly large, including many completely implausible distributions (from a subjective viewpoint). Ideally one wants a Γ which includes every prior considered plausible (i.e. close to π_0) and none that are implausible. The use of Γ as in (1.1) is a good starting point.

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 $\mathcal{Q}$. In this paper we will restrict consideration to three interesting choices of $\mathcal{Q}$. First, in Section 2, the choice $\mathcal{Q} = \mathcal{P}$ (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, $\mathcal{Q}$, 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 an appealing class could be worked with and provide reasonably simple answers. Finally, in Section 4, we consider $\mathcal{Q}$ 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 (1980), and Berger (1982, 1983). 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, and with the very related idea of "upper" and "lower" probabilities. Most of the work with classes of priors considers either the "moment conditions" or "conjugate priors" classes mentioned earlier. One exception is DeRobertis and Hartigan (1981). They obtained interesting results for the class of prior densities

$$(1.2) \qquad \Gamma = \{\pi : \pi(\theta_1)/\pi(\theta_2) \le g_1\,(\theta_1)/g_2\,(\theta_2), \text{ for all } \theta_1, \theta_2 \in \Theta\},$$

where $g_1 \le g_2$ are given, positive functions. We prefer the class in (1.1) for intuitive content and ease of analysis.

For a discussion and references concerning upper and lower probabilities, see Berger (1983). The basic idea behind them (that of generalizing probability distributions to functions which can reflect uncertainty in prior probabilities) is interesting, but we feel that there are considerable conceptual and manipulative advantages to sticking with probability distributions and incorporating uncertainty in prior probabilities through classes of prior distributions.

1.3 Robust Bayesian Methodology

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 (1983) - 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 $\mathfrak{D}$ = {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. Furthermore, the "second-level" problem of finding the smallest sized set with posterior probability at least $1 - \alpha$ (for all $\pi \in \Gamma$) is solved. The general theory is then applied to the univariate normal mean problem.

Unfortunately, there are certain inadequacies in assuming that $\mathfrak{D}$ = {all distributions} (see Section 2.3), and attempting the above program with more reasonable Γ (such as that in Section 3) becomes quite difficult (though perhaps not impossible). Also, it may frequently be the case that a proposed inference is not "robust" with respect to all $\pi \in \Gamma$, particularly when $\mathfrak{D}$ is chosen for technical convenience and includes some unreasonable distributions. 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 "optimal" prior, decision, or inference. The following are the five major such methods:

(i) Put a prior distribution on Γ itself, and carry out a formal Bayesian analysis.

(ii) Use minimax type criteria on posterior measures (i.e., posterior expected losses) for $\pi \in \Gamma$.

(iii) Use frequentist measures to select a "good" procedure compatible with $\pi \in \Gamma$.

(iv) Use some measure of "information" to select a prior in Γ, such as a "maximum entropy" prior (c.f. Jaynes (1968)) or a "reference" prior (c.f. Bernardo (1979)).

(v) Use maximum likelihood methods, essentially choosing the prior π which maximizes the predictive density $m(x|\pi)$ over Γ.

Discussion and further references for all of these methods can be found in Berger (1983). In this paper we will mainly utilize method (v), but first pause for a short discussion of (i). 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)) is very natural from a Bayesian viewpoint. 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 can be very hard to put a reasonable prior on complicated Γ, such as those in Sections 2 and 3, 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 be, to a large extent, arbitrary.

Method (v) is probably the simplest method of dealing with Γ. For $\pi = (1-\varepsilon)\,\pi_0 + \varepsilon q$, $q \in \mathfrak{Q}$, 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) $\hat{q} \in \mathfrak{Q}$, we will then suggest formally using the estimated prior $\hat{\pi}$, given by

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

(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. (The name "ML-II" essentially comes from Good (1965), who calls the process Type II maximum likelihood.)

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 apriori, 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, non-rigorous, and the ultimate justification for proceeding in this way is simply that it seems to work. Attempts to demonstrate this will be made throughout the paper. Of course, there is already substantial evidence in the literature attesting to the success of the method, both in the Bayesian literature (c.f. Jeffreys (1961), Good (1965, 1980), Box and Tiao (1973), Bishop, Fienberg and Holland (1975), and Zellner (1982)), and in the empirical Bayesian literature (c.f. 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 $\hat{\pi}$. 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).

1.4 Relationship to Model Robustness and Selection.

The ideas discussed so far have several intimate relationships to the areas of model selection and model robustness, in which f itself is considered to be partially unknown; for instance, perhaps it is felt that $f(x|\theta) = h(x-\theta)$, where $h = (1-\varepsilon)\ h_0 + \varepsilon g$, h_0 being a standard normal density and g being in some class G of possible contaminations. Model robustness does fall, formally, within the Bayesian robustness situation outlined earlier, as is seen by adopting the simple expedient of letting θ also contain any of the uncertainty in f (such as g above). Indeed Bayesians often see little reason to distinguish between the "model" f and a prior π on parameters of the model, in that choice of a model is often a (perhaps extreme) use of subjective information. Although the methods discussed here may work well on model robustness problems, to keep the paper contained (and avoid a number of technical difficulties) we will only consider examples in which f is known up to θ. (Also, there are certainly situations in which knowledge of f is much more precise than knowledge of θ; robustness with respect to the prior on θ is then of most interest.)

The two cornerstones of the technique discussed in this paper actually arose in model robustness research. The use of ε-contamination regions has long been a feature of many frequentist studies on model robustness (c.f. Huber (1981) and Marazzi (1980)). And, Bayesian model robustness studies often make heavy use of some form of the predictive density m. (A portion of the literature on Bayesian model robustness is Jeffreys (1961), Box and Tiao (1962) deFinetti (1961), Good (1965, 1967, 1980), Lindley (1966), Dempster (1976), and - especially comprehensive and thorough - Box and Tiao (1973), Dempster (1975) and Box (1980).) Indeed, there is no real conceptual difference between the use of m in robustness investigations and the use of m in Bayesian model selection: if considering various models, a Bayesian calculates the predictive density for each model, and bases decisions on the relative magnitudes of these predictive densities (at the observed x). (Some relevant references here, in addition to those listed above, are Roberts (1965), DeGroot (1970), Dickey (1971, 1975), Dempster (1971), Zellner (1971, 1982), deFinetti (1972), Hill (1974, 1980), Leamer (1978), Davis (1979), Geisser and Eddy (1979), and Zellner and Siow (1980). Jeffreys (1961) was the first to extensively develop these ideas.)

Finally, it should be mentioned that the predictive density has a great many other important roles to play in Bayesian analysis. A skimpy list of references concerning these roles is Roberts (1965) (which has some earlier history), Guttman (1967), Dempster (1971), Geisser (1971), deFinetti (1972), Aitchinson and Dunsmore (1975), Davis (1979), Geisser and Eddy (1979), and Kadane, et.al. (1980).

1.5 Useful Formulas and Notation

For priors of the form

$$\pi(d\theta) = (1-\varepsilon)\ \pi_0(d\theta) + \varepsilon q(d\theta), \tag{1.4}$$

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

$$m(x|\pi) = (1-\varepsilon)\ m(x|\pi_0) + \varepsilon m(x|q), \tag{1.5}$$

and

$$\pi(d\theta|x) = \lambda(x)\ \pi_0\ (d\theta|x) + (1-\lambda(x))\ q\ (d\theta|x), \tag{1.6}$$

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

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

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

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

2.1. The ML-II Prior and Posterior.

For Γ defined as in (2.1), the ML-II prior and corresponding posterior are often quite simple to obtain.

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

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

Proof: Straightforward.||

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 Section 1.5)

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

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 is 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

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

where

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

Note that $\hat{\lambda}$ is an exponentially decreasing function of $|x-\mu|^2$, so that $\hat{\delta}^{\pi}(x) \to x$ 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 $\hat{\delta}^{\pi}$ 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 $\rho_0(d\theta) = (2\pi^2)^{p/2}\ d\theta$. Hence, we have that following result.

<u>Theorem 2.2</u>. For all $0 < \varepsilon \leq 1$, $\hat{\delta}^{\pi}$ is inadmissible if $p \geq 3$. For all $0 \leq \varepsilon \leq 1$, $\hat{\delta}^{\pi}$ is admissible if $p < 3$. (Obviously, if $\varepsilon = 0$, $\hat{\delta}^{\pi}$ is admissible.)

Proof: The proof follows directly from Corollary 6.4.1 of Brown (1971).||

Although $\delta^{\hat{\pi}}$ is inadmissible for $p \geq 3$, it is unlikely that there exists any significantly better estimator.

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

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

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 apriori. Working with Γ as in Section 3, which do not allow such implausible contaminations, would eliminate this problem.

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 $\mathfrak{D}$ is the class of all distributions, it often becomes feasible to check this. The basic tool is a result of Huber (1973), concerning the range of posterior probabilities of a set. This result is given in Section 2.4.1. In Section 2.4.2, the result is applied to testing of hypotheses. In Section 2.4.3, the result is utilized to solve the problem of finding the set of minimum size which has posterior probability at least 1 - α of containing θ, for all $\pi \in \Gamma$.

2.4.1 Range of Posterior Probabilities of a Set

The following theorem is given without proof in Huber (1973). Though easy, we include a proof here, since the reasoning is similar to reasoning used in more complicated situations later.

Theorem 2.3. 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.,

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

Then

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

and

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

Proof. Let $\overline{C}$ denote the complement of C. Also, for any $q \in \mathcal{Q} = \mathcal{P}$, let

$$z_q(A) = \int_A f(x|\theta)\, q(d\theta).$$

Clearly

$$(2.8) \qquad P^{\pi}(\theta \in C | X = x) = \frac{(1-\varepsilon) m(x|\pi_0)\, \beta_0 + \varepsilon\, z_q(C)}{(1-\varepsilon)\, m(x|\pi_0) + \varepsilon\, z_q(C) + \varepsilon\, z_q(\overline{C})} .$$

Consider the function

$$h(z) = (K_1 + z)/(K_2 + z + g\,(z)).$$

It is straightforward to check that h is increasing in $z \geq 0$ when $K_2 \geq K_1 \geq 0$ and g is a positive, decreasing function of z. Setting $K_1 = (1-\varepsilon)\, m(x|\pi_0)\, \beta_0$, $K_2 = (1-\varepsilon)\, m(x|\pi_0)$, $z = \varepsilon\, z_q(C)$, and

$$g(z) = \varepsilon \int f(x|\theta)\, q(d\theta) - z = \varepsilon\, z_q(\overline{C}),$$

it follows that (2.8) is minimized when $z = z_q(C) = 0$. Thus

$$\inf_{\pi \in \Gamma} P^{\pi}(\theta \in C | X=x) = \inf_{\{q: z_q(C)=0\}} \frac{(1-\varepsilon)\, m(x|\pi_0)\, \beta_0}{(1-\varepsilon)\, m(x|\pi_0) + \varepsilon\, z_q(\overline{C})}$$

$$= \frac{(1-\varepsilon)\, m(x|\pi_0)\, \beta_0}{(1-\varepsilon)\, m(x|\pi_0) + \varepsilon \sup_{\{q: z_q(C)=0\}} z_q(\overline{C})} .$$

But

$$\sup_{\{q:z_q(C)=0\}} z_q(\overline{C}) = \sup_{\theta \in \overline{C}} f(x|\theta),$$

and (2.6) follows. Formula (2.7) is established similarly.||

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

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

The usual 100 × (1-α)% 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 (1-α/2) upper percentile of the standard normal distribution.

To investigate the robustness of C, we use (2.6) of Theorem 2.3. Note that

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

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

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

and, for x ∈ C,

$$\inf_{\pi \in \Gamma} P\ (\theta \in C|X=x)$$

$$= (1-\alpha)\left\{1 + \frac{\varepsilon(1+\tau^2/\sigma^2)^{1/2}}{(1-\varepsilon)} \exp\left[\frac{(x-\mu)^2 - (|x-\mu|V/\tau - z_{\alpha/2}\tau)^2}{2(\sigma^2+\tau^2)}\right]\right\}^{-1} .$$

As a concrete example, suppose that $\sigma^2 = 1$, $\tau^2 = 2$, $\mu = 0$, and $\varepsilon = .2$. First, suppose x = .5 is observed. Then the usual 95% Bayes credible interval for θ is (-1.27, 1.93).

$$\inf_{\pi \in \Gamma} P^{\pi}(-1.27 < \theta < 1.93|X=.5) = .868$$

and

$$\sup_{\pi \in \Gamma} P^{\pi}(-1.27 < \theta < 1.93|X=.5) = .966.$$

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 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

$$\inf_{\pi \in \Gamma} P^{\pi}(1.07 < \theta < 4.27|X=4) = .1426$$

and

$$\sup_{\pi \in \Gamma} P^{\pi}(1.07 < \theta < 4.27|X=4) = .99.$$

Since the posterior probability can get as low as .1426 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 (.1426) 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 apriori. Using a more reasonable Γ might result in robustness. Also, more robust credible sets can be found - see Section 2.4.3. In any case, the use of $\mathcal{D} = \mathcal{P}$ and Theorem 2.3 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.

2.4.2 Hypothesis Testing.

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

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

Letting $C = \Theta_0$, Theorem 2.3 immediately yields the following:

Corollary 2.1. For Γ as in (2.1),

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

and

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

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 0_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 $\mathcal{D} = \mathcal{P}$ very attractive for robustness investigations in testing.

It should be clear that Theorem 2.3 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.

2.4.3 Optimal Robust Credible Regions.

Definition: A measurable subset C of Θ is a level $1 - \alpha$, Γ-credible region for θ if

$$(2.9) \qquad \inf_{\theta \in \Gamma} P_\pi(\theta \in C|X=x) \geq 1 - \alpha.$$

We seek the $1 - \alpha$, Γ-credible region of smallest Lebesgue measure, ν, for Γ as in (2.1). The characterization given in the following theorem for this optimal credible region is surprisingly simple.

Theorem 2.4. Suppose that $\Theta \subseteq \mathbb{R}^P$ and that π_0 has a density w.r.t. Lebesgue measure. Consider sets A_η, B_ψ, and D_ψ, defined by

$$A_\eta = \{\theta: f(x|\theta) > \eta\},$$

and

$$(2.10) \qquad B_\psi = \{\theta\colon \pi_0(\theta|x) > \psi\}, \quad D_\psi = \{\theta\colon \pi_0(\theta|x) = \psi\}.$$

Let

$$\eta_0 = \inf\{\eta\colon \inf_{\pi\in\Gamma} P^\pi(\theta \in A_\eta|X=x) \leq 1-\alpha \text{ and } \inf_{\pi\in\Gamma} P^\pi(\theta\in\overline{A}_\eta|X=x) \geq 1-\alpha\},$$

where $\overline{A}_\eta$ denotes the closure of A_η. Then the level $1-\alpha$, Γ credible region of smallest Lebesgue volume is of the form

$$(2.11) \qquad C^* = A_\eta \cup B_{\psi(\eta)} \cup D$$

for some $\eta \geq \eta_0$, where $0 \leq \psi(\eta) \leq \infty$ is defined by

$$(2.12) \qquad \psi(\eta) = \sup\{\psi\colon P^{\pi_0}(\theta \in A_\eta \cup B_\psi \cup D_\psi|X=x) \geq (1-\alpha)\,[1+\varepsilon\eta/(1-\varepsilon)m(x|\pi_0)]\}$$

and $D \subset D_{\psi(\eta)}$ is an arbitrary set such that

$$P^{\pi_0}(\theta \in C^*|X=x) = (1-\alpha)[1+\varepsilon\eta/(1-\varepsilon)m(x|\pi_0)].$$

Proof. First note that $\psi(\eta)$ is well defined for $\eta \geq \eta_0$ and is decreasing in η. Next, denote the desired optimal set by C^*, and define η^* as

$$\eta^* = \sup_{\theta\notin C^*} f(x|\theta).$$

Then, by definition, $A_{\eta *} \subseteq C^*$. Furthermore, it must be the case that $\eta^* \geq \eta_0$. To see this, note that, if $\eta^* < \eta_0$, then

$$\inf_{\pi \in \Gamma} P^{\pi}(\theta \in C^*|X=x) \geq \inf_{\pi \in \Gamma} P^{\pi}(\theta \in A_{\eta *}|X=x) > 1 - \alpha,$$

using the definition of η_0. Since the last inequality is strict, it is easy to show that $\overline{A}_{\eta_0}$, or some subset thereof, is level $1 - \alpha$, Γ credible and has smaller volume that A^*_{η}..

Now, let $B^* = C^* - A_{\eta *}$. If B^* is empty, (2.11) is trivally satisfied with $\eta = \eta^* = \eta_0$ and $\psi(\eta) = \infty$. Hence, suppose that B^* is non-empty and not of the form $B_{\psi(\eta^*)} \cup D$. Let $\nu^* = \nu(B^*)$, and consider a set of the form

$$B^{**} = B_K \cup D, \quad D \subseteq D_{\psi} \,,$$

where K and D are chosen so that $(B^{**} \cap A^{\text{complement}}_{\eta *}) = \nu^*$. Let $C^{**} = B^{**} \cup A_{\eta *}$. Then

$$P^{\pi_0}(\theta \in B^{**}|X=x) > P^{\pi_0}(\theta \in B^*|X=x)$$

and $\sup_{\theta \notin C^{**}} f(x|\theta) \leq \eta^*$. Therefore, by Theorem 2.3,

$$\inf_{\pi \in \Gamma} P^{\pi}(\theta \in C^{**}|X=x) > \frac{P^{\pi_0}(\theta \in A_{\eta *}|X=x)+P^{\pi_0}(\theta \in B^*|X=x)}{1 + \varepsilon\eta^*/[(1-\varepsilon)m(x|\pi_0)]}$$

$$= \inf_{\pi \in \Gamma} P^{\pi}(\theta \in C^*|X=x) \geq 1 - \alpha.$$

Since the first inequality is strict, we could increase K or decrease D in the definition of B** to yield

$$\inf_{\pi \in \Gamma} P^{\pi}(\theta \in C^{**} | X=x) = 1 - \alpha,$$

and the resulting C** must have smaller Lebesgue volume than C*, a contradiction.||

2.4.4 Optimal Robust Credible Regions for a Normal Mean

Assume that $X \sim \mathcal{N}(\theta, \sigma^2)$, σ^2 known, and that π_0 is $\mathcal{N}(\mu, \tau^2)$. By a simple linear transformation, it is sufficient to consider the case $\sigma^2 = 1$ and $\mu = 0$. Recall that, then, $m(\cdot|\pi_0)$ is $\mathcal{N}(0,1+\tau^2)$ and $\pi_0(\cdot|x)$ is $\mathcal{N}(\delta,V^2)$, where $\delta = V^2x$ and $V^2 = \tau^2/(1+\tau^2)$.

To apply Theorem 2.4 in this situation, note that D_ψ has measure zero for all ψ, and hence D can be ignored in (2.11). For similar reasons, η_0 is defined by

$$(2.13) \qquad 1 - \alpha = \inf_{\pi \in \Gamma} P^{\pi}(\theta \in A_{\eta_0} |X-x) = P^{\pi_0}(\theta \in A_{\eta_0} |X=x) \left\{1 + \frac{\varepsilon\eta_0}{(1-\varepsilon)m(x|\pi_0)}\right\}^{-1},$$

and $\psi(\eta)$ is defined by

$$(2.14) \qquad p^{\pi_0}(\theta \in A_\eta \cup B_{\psi(\eta)} |X=x) = (1-\alpha) \left[1 + \frac{\varepsilon\eta}{(1-\varepsilon)m(x|\pi_0)}\right].$$

Also,

$$A_\eta = (x - \sqrt{2 \log (\eta a)^{-1}}\ ,\ \ x + \sqrt{2 \log (\eta a)^{-1}}\ )$$

and

$$B_\psi = (\delta - V\sqrt{2 \log (\psi a V)^{-1}},\ \delta + V\sqrt{2 \log (\psi a V)^{-1}}\ )\ ,$$

where $a = (2\pi)^{1/2}$. We will give only the results for $x > 0$. Results for $x < 0$ are entirely analogous.

Three possible cases may arise: either

<u>Case 1</u>: $C^* = A_{\eta_0}$;

<u>Case 2</u>: $C^* = A_{\eta^*} \cup B_{\psi(\eta^*)}$, where $\nu(A_{\eta^*} \cap B_{\psi(\eta^*)}) > 0$;

<u>Case 3</u>: $C^* = A_{\eta^*} \cup B_{\psi(\eta^*)}$, where $\overline{A}_{\eta^*} \cap \overline{B}_{\psi(\eta^*)} = \phi$.

(The case where A_{η^*} and $B_{\psi(\eta^*)}$ have only a common boundary can be shown to occur with measure zero, and hence will be ignored in this discussion.)

To find C*, we consider each of the three cases separately, and find all the potential "candidates" for optimality. Among the candidates generated will be the optimal credible set, but our solution process can generate bogus candidates, i.e., sets of the wrong form (for the case being considered), or sets failing to satisfy the requirement (2.9) (or (2.14)). Hence any apparent solution pair, $(\eta^*, \psi(\eta^*))$, that is generated should be checked for correctness of form and satisfaction of (2.14). All satisfactory solutions can then be compared for size, the smallest being C*.

For notational convenience, denote $m(x|\pi_0)$ simply by m, and define R by

$$R = (1-\varepsilon)m/[\varepsilon(1-\alpha)].$$

Case 1. Direct application of (2.13) shows that η_0 is the largest solution to the equation

$$(2.15) \qquad \eta = R\{\Phi([(1+\tau^2)^{-1}x + (2 \log (\eta a)^{-1})^{1/2}] V^{-1}) - \Phi([(1+\tau^2)^{-1}x - (2 \log (\eta a)^{-1})^{1/2}] V^{-1}) + \alpha - 1\},$$

where Φ denotes the c.d.f. of a standard normal random variable.

Case 2. It is shown, in the Appendix, that η^* is a solution to the equation

$$(2.16) \qquad \eta = R\{\Phi([(1+\tau^2)^{-1}x + (2 \log (\eta a)^{-1})^{1/2}]/V) - \Phi(-(2 \log (\psi(\eta)aV)^{-1})^{1/2}) + \alpha - 1\},$$

where

$$(2.17) \qquad \psi(\eta) = R^{-1} \eta(2 \log (\eta a)^{-1})^{1/2} + (aV)^{-1} \exp\{-[(1+\tau^2)^{-1}x + (2 \log (\eta a)^{-1})^{1/2}]^2(2V^2)^{-1}\}.$$

Case 3. It is shown, in the Appendix, that η^* is a solution to the equation

$$(2.18)\qquad \eta = R\{2\Phi([2\log(\psi(\eta)aV)^{-1}]^{1/2}) + \Phi([(1+\tau^2)^{-1}x + (2\log(a\eta)^{-1})^{1/2}]/V)$$
$$- \Phi([(1+\tau^2)^{-1}x - (2\log(a\eta)^{-1})^{1/2}]/V) + \alpha - 2\},$$

where

$$(2.19)\qquad \psi(\eta) = (2R)^{-1}\eta(2\ \log(a\eta)^{-1})^{1/2}$$
$$+ (2aV)^{-1}\{\exp\{-[(1+\tau^2)^{-1}x + (2\ \log(\eta a)^{-1})^{1/2}]^2(2V^2)^{-1}\}$$
$$+ \exp\{-[(1+\tau^2)^{-1}x - (2\log(\eta a)^{-1})^{1/2}]^2\ (2V^2)^{-1}\}\}.$$

As an aid to finding the simultaneous solutions of (2.16) and (2.17), or (2.18) and (2.19), note that η^* must satisfy

$$\eta_0 \leq \eta^* \leq \min\ \{a^{-1}, \alpha R\}.$$

The lower bound was established in Theorem 2.4, and the upper bound follows from the fact that the right hand side of (2.14) must be less than one. It should also be noted that Case 3 usually does not occur (i.e., usually the optimal confidence set does not consist of disjoint intervals). Indeed, it only seems possible for Case 3 to occur when $\alpha < \varepsilon$ and τ^2 is small (compared to σ^2), both of which will be rare in practice. Furthermore, the "pattern" will be that for $|x-\mu|$ small, it will be a Case 1 situation; for $|x-\mu|$ larger it will become a Case 2 situation; then a Case 3 (if it occurs at all); and finally a returning to Case 2 and then Case 1 as $|x-\mu|$ gets larger yet.

It is interesting that Case 1 becomes the limiting case as $|x-\mu| \to \infty$. Indeed it is not hard to show that the appropriate value of η_0, as $|x-\mu| \to \infty$, is

$$\eta_0 = \{\alpha(1-\varepsilon)/[\varepsilon(1-\alpha)m(x|\pi_0)]\}(1+o(1)).$$

(This equation will actually hold in great generality, not just in the normal case.) Since $m(x|\pi_0) \to 0$ as $|x-\mu| \to \infty$, it follows that η_0 will go to zero and A_{η_0} will become arbitrarily large, certainly a not very appealing result. This is another indication that letting $\mathcal{Q}$ contain all distributions will often be inappropriate.

3. Unimodality Preserving Contaminations

If the elicited prior, π_0, is unimodal with unique mode θ_0, an extremely attractive class Γ to consider is

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

When π_0 is unimodal, it will frequently be the case that the only plausible priors are those which are close to π_0 and also unimodal, requirements reflected very well in the above Γ. (Often one might want to choose $\varepsilon' = \varepsilon$, for reasons of symmetry, but the analysis is the same for general ε'.) Any prior in Γ will typically be plausible (in contrast to the case when all contaminations are allowed), and Γ will contain any plausible prior (for large enough ε). It came as a great surprise to us that such a reasonable Γ could be worked with and provide relatively simple answers.

The situation we consider here is where $\Theta \subseteq \mathbb{R}^1$ and the likelihood function $f(x|\theta)$ is also unimodal (as a function of θ, of course) with unique mode $\hat{\theta}(x)$. (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 positive and strictly monotonic on each side of the modes. More general cases could be handled, but the results get messier. Finally, we will only present the results when $\hat{\theta}(x) \geq \theta_0$. The case $\hat{\theta}(x) < \theta_0$ is essentially identical.

Under the above assumptions, we first determine $\hat{\pi}$, the prior maximizing

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

among all π in Γ. Those interested only in the basic ideas may wish to skip the next section.

3.1 Preliminaries and Notation

For $-\varepsilon' \leq \rho \leq \varepsilon$, define $v(\rho) \geq \theta_0$, implicitely, by

$$(3.1) \qquad \pi_0(\theta_0)\ (1-\rho)\ (v(\rho) - \theta_0) - (1-\varepsilon) \int_{\theta_0}^{v(\rho)} \pi_0(\theta)\ d\theta = \varepsilon,$$

and define

$$(3.2) \qquad V(\rho) = f(x|v(\rho))(v(\rho) - \theta_0) - \int_{\theta_0}^{v(\rho)} f(x|\theta)\ d\theta\ .$$

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

$$(3.3) \qquad (1-\varepsilon)\ \pi_0(\theta)(w(\theta) - \theta) - (1-\varepsilon) \int_{\theta}^{w(\theta)} \pi_0(\xi)\ d\xi = \varepsilon,$$

and define

$$(3.4) \qquad w(\theta) = f(x|w(\theta))\ (w(\theta) - \theta) - \int_{\theta}^{w(\theta)} f(x|\xi)\ d\xi.$$

Lemma 3.1.

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

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

Proof. (a) At $v = \theta_0$, the left hand side of (3.1) is zero. As $v \to \infty$, the left hand side of (3.1) goes to ∞. Finally, since π_0 is decreasing for $\theta > \theta_0$, the derivative, with respect to v, of the left hand side of (3.1) is easily seen to be strictly positive. A solution to (3.1) thus exists and is unique.

To show that $v(\rho)$ is strictly increasing, one can differentiate both sides of (3.1) with respect to ρ and solve for $v'(\rho)$ (i.e., $\frac{d}{d\rho} v(\rho)$), 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 $f'(x|\theta) = \frac{d}{d\theta} f(x|\theta)$, calculation gives

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

Since f is decreasing for $\theta > \hat{\theta}(x)$, the monotonicity result follows from part (a). If $V(\rho) = 0$, the unimodality of f ensures that $v(\rho) > \hat{\theta}(x)$ (for otherwise the right hand side of (3.2) is positive). The strict monotonicity of V for such ρ ensures that any solution to $V(\rho) = 0$ must be unique.

(c) Letting $w'(\theta) = \frac{d}{d\theta} w(\theta)$, 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 $W(\theta_0) \geq 0$ and $W(\hat{\theta}(x)) < 0$. Since $v(\varepsilon) = w(\theta_0)$, it follows that $W(\theta_0) = V(\varepsilon) \geq 0$ (by assumption). That $W(\hat{\theta}(x)) < 0$ follows from (3.4) and an easy application of the mean value theorem (since $f(x|\theta)$ decreases for $\theta > \hat{\theta}(x)$).

<u>Lemma 3.2</u>. Suppose $V(\varepsilon) \geq 0$, and let $\theta_0 \leq \theta^* \leq \hat{\theta}(x)$ 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

$$(3.5) \qquad \int_{\theta^*}^{w(\theta^*)} g(\theta)\ f(x|\theta)\ d\theta \leq 0.$$

<u>Proof</u>. (a) Clearly $f(x|\theta^*) < f(x|w(\theta^*))$, for otherwise the integrand in (3.4) 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 $w(\theta^*) > \hat{\theta}(x)$, for otherwise (3.4) 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 $h(\theta) = -\frac{d}{d\theta} g(\theta)$ (note $h \geq 0$) and writing

$$g(\theta) = K - \int_{\theta^*}^{\theta} h(\xi)\, d\xi,$$

where

$$(3.6) \qquad K = \frac{1}{[w(\theta^*) - \theta^*]} \int_{\theta^*}^{w(\theta^*)} \int_{\theta^*}^{\eta} h(\xi)\, d\xi d\eta,$$

$$= \frac{1}{[w(\theta^*) - \theta^*]} \int_{\theta^*}^{w(\theta^*)} (w(\theta^*) - \xi) h(\xi)\, d\xi,$$

we obtain from Fubini's theorem

$$(3.7) \qquad \int_{\theta^*}^{w(\theta^*)} g(\theta) f(x|\theta) d\theta = K \int_{\theta^*}^{w(\theta^*)} f(x|\theta) d\theta$$

$$- \int_{\theta^*}^{w(\theta^*)} h(\xi) \int_{\xi}^{w(\theta^*)} f(x|\theta)\, d\theta\, d\xi.$$

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

$$(3.8) \qquad \psi(\xi) \equiv \int_{\xi}^{w(\theta^*)} f(x|\theta)\, d\theta \geq (w(\theta^*) - \xi)\, f(x|w(\theta^*)).$$

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

$$(3.9) \qquad \psi(\theta^*) = \int_{\theta^*}^{w(\theta^*)} f(x|\theta)\, d\theta = (w(\theta^*) - \theta^*)f(x|w(\theta^*))$$

(since $W(\theta^*) = 0$). Hence, $\psi(\xi)$ must lie above the line $(w(\theta^*) - \xi)\, f(x|w(\theta^*))$, establishing (3.8).

Using (3.8) in (3.7) we get that

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

the right hand side of which is zero by (3.6) and (3.9).||

3.2 The ML-II Prior

Define $\hat{\pi}$ as follows:

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

$$(3.10) \qquad \hat{\pi}(\theta) = \begin{cases} (1-\varepsilon)\, \pi_0\, (\theta^*) & \text{for } \theta^* \leq \theta \leq w(\theta^*), \\ (1-\varepsilon)\, \pi_0\, (\theta) & \text{otherwise.} \end{cases}$$

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

$$(3.11) \qquad \hat{\pi}(\theta) = \begin{cases} (1-\rho^*)\, \pi_0\, (\theta_0) & \text{for } \theta_0 \leq \theta \leq v(\rho^*) \\ (1-\varepsilon)\, \pi_0\, (\theta) & \text{otherwise.} \end{cases}$$

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

$$(3.12) \qquad \hat{\pi} = \begin{cases} (1+\varepsilon')\ \pi_0(\theta_0) & \text{for } \theta' \leq \theta \leq \theta'' \\ (1-\varepsilon)\ \pi_0(\theta) & \text{otherwise,} \end{cases}$$

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

$$(3.13) \qquad f(x|\theta') = f(x|\theta''),$$

$$(1+\varepsilon')\ \pi_0(\theta_0)(\theta''-\theta') - (1-\varepsilon) \int_{\theta'}^{\theta''} \pi(\theta)\ d\theta = \varepsilon.$$

Lemma 3.1 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 the 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 3.1. The $\hat{\pi}$ defined in (3.10) through (3.12) 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

$$(3.14) \qquad m(x|\pi) - m(x|\hat{\pi}) = \int [\pi(\theta) - \hat{\pi}(\theta)] \, f(x|\theta) \, d\theta \leq 0$$

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

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

(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 3.2 (a) and (i) show that

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

Lemma 3.2 (b) and (ii) imply that

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

Thus

$$(3.15) \qquad \int g(\theta)f(x|\theta)\ d\theta < f(x|w(\theta^*))(-K) + \frac{K}{[w(\theta^*)-\theta^*]} \int_{\theta^*}^{w(\theta^*)} f(x|\theta)\ d\theta.$$

Since $W(\theta^*) = 0$, the right hand side of (3.15) is zero, and (3.14) follows.||

Comments: 1. The key step in the proof of Theorem 3.1 is really Lemma 3.2 (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 the constraints (i) π is nonnegative, (ii) π has mass one (iii) $(1-\varepsilon)\ \pi_0 \leq \pi \leq (1+\varepsilon')\ \pi_0$, and (iv) π is unimodal. Calculus of variations with such side constraints is quite difficult.

3.3 The ML-II Posterior and Robustness

The prior $\hat{\pi}$ can be written as

$$\hat{\pi} = \begin{cases} K_i & \text{for } \theta \in B_i \\ (1-\varepsilon)\ \pi_0(\theta) & \text{for } \theta \notin B_i \end{cases}$$

where i refers to Case 1, 2, 3, or 4, B_i is the appropriate interval, and K_i is the appropriate constant. Alternatively, $\hat{\pi}$ can be written as

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

where the ML-II contamination, $\hat{q}$, is given by

$$\hat{q} = \varepsilon^{-1}\ [K_i - (1-\varepsilon)\ \pi_0(\theta)]\ I\ (\theta \in B_i).$$

Thus the ML-II posterior is

$$(3.16) \qquad \hat{\pi}(\theta|x) = \hat{\lambda}(x)\ \pi_0(\theta|x) + (1-\hat{\lambda}(x))\ \hat{q}\ (\theta|x),$$

where

$$(3.17) \qquad \hat{\lambda}(x) = (1-\varepsilon)\ m\ (x|\pi_0)/[(1-\varepsilon)\ m\ (x|\pi_0) + \varepsilon m(x|\hat{q})]$$

and

$$(3.18) \qquad m(x|\hat{q}) = \varepsilon^{-1} \int_{B_i} [K_i - (1-\varepsilon)\ \pi_0\ (\theta)]\ f(x|\theta)\ d\theta.$$

Of interest in evaluating the robustness of $\hat{\pi}$ is Case 1; only when π_0 and $f(x|\theta)$ are moderately separated will answers vary insignificantly for different $\pi \in \Gamma$. In this case,

$$(3.19) \qquad \hat{q} = \varepsilon^{-1}(1-\varepsilon)\ [\pi_0(\theta^*) - \pi_0(\theta)]\ I\ (\theta \in [\theta^*, w(\theta^*)]).$$

As $|\hat{\theta}(x) - \theta_0| \to \infty$, it will typically happen that $\theta^* \to \infty$, $(\hat{\theta}(x) - \theta^*) \to \infty$,

$(w(\theta^*) - \hat{\theta}(x)) \to \infty$, and $\pi_0(\theta^*)/m(x|\pi_0) \to \infty$. It is then easy to see that $\hat{\lambda}(x) \to 0$ and

$$(3.20) \qquad \hat{q}(\theta|x) \to f(x|\theta)/\int f(x|\theta)\, d\theta.$$

In other words, $\hat{\pi}$ behaves essentailly as a uniform prior over θ. (The uniform part of $\hat{\pi}$, that on $[\theta^*, w(\theta^*)]$, comes to dominate, and the interval $[\theta^*,w(\theta^*)]$ expands to span all except a negligible part of the support of $f(x|\theta)$.) As the data and π_0 come into conflict, therefore, the ML-II prior will effectively ignore π_0 and act as a (noninformative) uniform prior. This kind of behavior can be labelled "robust" from a number of viewpoints (c.f. Berger (1983)). Also, this limiting behavior is much more pleasing than that of $\hat{\pi}$ in Section 2, which collapsed to a point mass at $\hat{\theta}(x)$ in the limit. Note, finally, that $\hat{\pi}$ is always a proper prior, so that, although its limiting nature is that of a noninformative prior, it should not succumb to any problems related to the impropriety of typical noninformative priors.

3.4 The Normal Distribution

As an example of the preceding theory, we consider the case where X is $\mathcal{N}(\theta,1)$ and π_0 is $\mathcal{N}(0,\tau^2)$. (The more general case where $X \sim \mathcal{N}(\theta, \tau^2)$ and π_0 is $\mathcal{N}(\mu, \tau^2)$, σ^2, μ, and τ^2 all known, can be reduced to this case by a linear transformation.) We will determine $\hat{\pi}$, and also give explicit expressions for the posterior mean and variance of $\hat{\pi}$. Only Case 1 will be considered, it being the most interesting (and difficult). As before, Φ will denote the standard normal c.d.f., and ϕ the standard normal density function.

3.4.1 The ML-II Posterior

For completeness, we rewrite the defining equations for θ^* and $w(\theta^*)$ as

$$(3.21) \qquad (1-\varepsilon)\,\phi\left(\frac{\theta^*}{\tau}\right)\left(\frac{w(\theta^*)}{\tau} - \frac{\theta^*}{\tau}\right) - (1-\varepsilon)\left[\Phi\left(\frac{w(\theta^*)}{\tau}\right) - \Phi\left(\frac{\theta^*}{\tau}\right)\right] = \theta,$$

$$(3.22) \qquad \phi(w(\theta^*) - x)\,(w(\theta^*) - \theta^*) - [\Phi(w(\theta^*) - x) - \Phi(\theta^*-x)]=0.$$

These equations can be easily solved for θ^* and $w(\theta^*)$, recalling that $\theta^* < x < w(\theta^*)$. From (3.16) through (3.19) we get after some algebra (and using (3.22))

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

where

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

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

$$C_0 \equiv \sqrt{1+\tau^{-2}}\quad \phi\left(\frac{\theta^*}{\tau}\right)/\phi\left(\frac{x}{\sqrt{1+\tau^2}}\right),$$

$$B_0 \equiv P^{\pi_0(\theta|x)}([\theta^*,w(\theta^*)]) = \Phi\left(\frac{w(\theta^*)-\delta}{V}\right) - \Phi\left(\frac{\theta^*-\delta}{V}\right),$$

and

$$\hat{q}(\theta|x) = \frac{I(\theta \in [\theta^*, w(\theta^*)])}{(\hat{\lambda}(x)^{-1} - 1)} \{C_0\ f(x|\theta) - \pi_0(\theta|x)\}.$$

An alternate formula for $\hat{\pi}(\theta|x)$, more useful for purposes such as obtaining credible sets, is

$$\hat{\pi}(\theta|x) = \hat{\lambda}(x)\ \{\pi_0(\theta|x)\ I\ (\theta \notin [\theta^*, w(\theta^*)]) + C_0\ f(x|\theta)\ I\ (\theta \in [\theta^*,w(\theta^*)])\}.$$

(Though this posterior looks as though it could be bimodel, it will in reality always be unimodal.)

3.4.2 The ML-II Posterior Mean and Variance

The posterior mean can be calculated (using calculus and (3.22)) to be

$$\begin{aligned}\delta^{\hat{\pi}} &= \int \theta\hat{\pi}(\theta|x)\ d\theta \\ &= \hat{\lambda}(x)\ \delta + (1-\hat{\lambda}(x))\ \delta^{\hat{q}},\end{aligned}$$

where

$$\delta^{\hat{q}} = x + \frac{C_0D_0 - E_0 + (x-\delta)\ B_0}{-B_0 + C_0\ \phi(w(\theta^*)-x)\ [w(\theta^*)-\theta^*]},$$

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

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

Also, the posterior variance can be calculated to be (see (1.8))

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

where

$$V^{\hat{q}} = [\hat{\lambda}(x)^{-1}-1]^{-1}\{-C_0D_0(2\delta^{\hat{q}}-x-\theta^*) + C_0(x-\delta^{\hat{q}})^2[w(\theta^*)-\theta^*]\phi(w(\theta^*)-x)$$

$$- (\theta^*+\delta-2\delta^{\hat{q}})\, E_0 - [V^2+(\delta-\delta^{\hat{q}})^2]\, B_0 + V[w(\theta^*)-\theta^*]\, \phi([w(\theta^*)-\delta]/V)\}.$$

3.4.3 Limiting Behavior

Though easy to calculate, the formulas for $\delta^{\hat{\pi}}$ and $V^{\hat{\pi}}$ are too involved to be easily understood intuitively. The formulas simplify greatly for large x. We present the limiting behavior here, partly to allow intuitive consideration of the results, and partly to show that the "robust" behavior discussed in Section 3.3 does hold. The proof, though lengthy, is routine and will be omitted.

Theorem 3.2. As $x \to \infty$,

(i) $\theta^* = \sqrt{2\tau^2 \log x} + o(1)$,

(ii) $w(\theta^*) = x + \sqrt{2\log(x/\sqrt{2\pi})} + o(1)$,

(iii) $\hat{\lambda}(x) = \dfrac{(1-\varepsilon)x}{\varepsilon\sqrt{(1+\tau^2)2\pi}}\, e^{-x^2/[2(1+\tau^2)]}\, (1+o(1))$,

(iv) $\delta^{\hat{q}} = x - \frac{1}{x} + o(x^{-1})$

(v) $v^{\hat{q}} = 1 - x^{-1} \sqrt{2 \log x} \, (1+o(1))$.

It can also be shown that $\hat{\pi}(\theta|x)$ does become essentially the uniform prior concentrated on $[\theta^*, w(\theta^*)]$. The rapid (exponential) rate at which $\hat{\lambda}(x)$ goes to zero means that this uniform portion can quickly become dominant.

4. Hierarchical Classes of Priors

4.1 Introduction

Hierarchical priors are typically employed when θ 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 $\Gamma_1 = \{g_\omega : \omega \in \Omega\}$ 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

$$(4.1) \qquad \pi_0(\theta) = \int_\Omega [\prod_{i=1}^{p} g_\omega(\theta_i)]\, h_0(\omega)\, d\omega.$$

(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), and Morris (1983).

There are three possible robustness concerns in working with (4.1). 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 stage 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

$$(4.2) \qquad \Gamma = \{\pi = (1-\varepsilon_1-\varepsilon_2)\,\pi_0 + \varepsilon_1 q_1 + \varepsilon_2 q_2,\ q_1 \in \mathcal{Q}_1,\ q_2 \in \mathcal{Q}_2\},$$

where $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are appropriate possible classes of contamination. Such an extension of the ε-contamination model vastly increases its flexibility while causing no real hardship in many applications, because the important formulas (1.5), (1.7), and (1.8) become simply

$$(4.3) \qquad m(x|\pi) = (1-\varepsilon_1-\varepsilon_2)\ m(x|\pi_0) + \varepsilon_1\ m(x|q_1) + \varepsilon_2\ m(x|q_2),$$

$$(4.4) \qquad \delta^{\pi}(x) = [1-\lambda_1(x)-\lambda_2(x)]\delta^{\pi_0}(x) + \lambda_1(x)\ \delta^{q_1}(x) + \lambda_2(x)\delta^{q_2}(x),$$

$$(4.5) \qquad V^{\pi}(x) = (1-\lambda_1-\lambda_2)V^{\pi_0} + \lambda_1 V^{q_1} + \lambda_2 V^{q_2} + \lambda_1\lambda_2(\delta^{q_1}-\delta^{q_2})^2$$
$$+ (1-\lambda_1-\lambda_2)\lambda_1(\delta^{\pi_0}-\delta^{q_1})^2 + (1-\lambda_1-\lambda_2)\lambda_2(\delta^{\pi_0}-\delta^{q_2})^2,$$

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 (4.3) (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 (4.2) 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.

4.2 Second Stage Uncertainty

Suppose, in the situation of Section 4.1, 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

$$(4.6) \qquad h(\omega) = (1-\varepsilon)\ h_0(\omega) + \varepsilon s(\omega), \quad s \in \mathcal{S}.$$

The resulting prior for θ is

$$(4.7) \qquad \pi(\theta) = \int [\prod_{i=1}^{p} g_\omega(\theta_i)] \; h(\omega) \; d\omega$$

$$= (1-\varepsilon) \; \pi_0(\theta) + \varepsilon q(\theta),$$

where

$$\pi_0(\theta) = \int [\prod_{i=1}^{p} g_\omega(\theta_i)] \; h_0(\omega) \; d\omega$$

and

$$q = \int [\prod_{i=1}^{p} \; q_\omega(\theta_i)] \; s(\omega) \; d\omega.$$

Letting $\mathcal{Q} = \{q\colon s \in \mathcal{S}\}$, it follows that the uncertainty in π can be expressed by

$$\Gamma \; = \{\pi = (1-\varepsilon) \; \pi_0 + \varepsilon q, \; q \in \mathcal{Q}\}.$$

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

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

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

$$(4.8) \qquad m(x|\pi) = (1-\varepsilon)\ m(x|\pi_0) + \varepsilon\int m(x|\omega)s(\omega)\ d\omega.$$

When $\mathcal{S} = \wp$ = {all distributions}, it is clear from (4.8) 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

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

for which analysis is usually quite straightforward.

Example 3. Suppose that $X = (X_1,\ldots,X_p) \sim \mathcal{N}_p(\theta,\sigma^2 I_p)$, 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 $\mathcal{N}_p(\mu\underset{\sim}{1},\ (\sigma^2+\tau^2)I_p)$, where $\underset{\sim}{1} = (1,\ldots,1)$. It is easy to check that $m(x|\omega)$ is maximized at

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

Hence, with contaminated second stage prior as in (5.6) and $\mathcal{S} = \wp$, the ML-II prior is

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

where $\hat{q}$ is $\mathcal{N}_p(\hat{\mu}\underset{\sim}{1},\hat{\tau}^2 I_p)$.

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

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

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

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

and

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

where

$$\delta(x) = \begin{cases} \sigma^{-p}\exp\{-\sum_{i=1}^{p}(x_i-\overline{x})^2/2p\} & \text{if } \Sigma(x_i-\overline{x})^2 < p\sigma^2 \\ [\frac{1}{p}\sum_{i=1}^{p}(x_i-\overline{x})^2]^{-p/2}\ \exp\{-\frac{1}{2}p\} & \text{otherwise.} \end{cases}$$

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

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

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 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

$$\mathcal{W} = \{w = (1-\varepsilon_1)\ w_0 + \varepsilon_1\ q_w\text{:}\quad w \text{ is unimodal}\},$$

$$\mathcal{V} = \{v = (1-\varepsilon_2)\ v_0 + \varepsilon_2\ q_v\text{:}\quad v \text{ is unimodal}\},$$

and apply the ideas of Section 3, first maximizing $m(x|\pi)$ over $\mathcal{V}$ and then over $\mathcal{W}$(the given order making the necessary unimodality verifications easier). We do not attempt the analysis here, because nothing new conceptually is involved and the argument would be moderately lengthy.

4.3 First and Second Stage Uncertainty

The simplest modification of (4.7) 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 $q_2 \in \mathcal{Q}_2 = \mathcal{P}$,

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

and π_0, s, and $\mathcal{S}$ are as in (4.7). 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 4.1, 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 be maximized when q_2 is a unit point mass at $\hat{\theta}$, the maximum likelihood estimate. Thus the ML-II prior is (assuming $\mathcal{S} = \wp$ and letting $I(\hat{\theta})$ denote a unit point mass at $\hat{\theta}$)

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

Formulas (4.3) through (4.5) can now easily be employed to give desired conclusions. In the situation of Example 3, for instance, all calculations can be carried out explicitely; 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 $\delta^{\hat{q}}$; and if the data are not compatible with the empirical Bayes model, then $\delta^{\hat{\pi}}$ will be close to the maximum likelihood estimate, $\hat{\theta} = x$.

5. Conclusions

The basic message of the paper is that a wide variety of explicit analyses concerning Bayesian robustness can be implemented. For very reasonable classes of priors, one can find the ML-II prior and find the range of posterior quantities of interest as the prior varies. This possibility brings the philosophically compelling robust Bayesian viewpoint (c.f. Berger (1983)) closer to the domain of practical statistics, and also introduces exciting new theoretical problems. Generalizations and applications of many kinds suggests themselves. One of the most challenging generalizations would be the combining of uncertainty in the prior with uncertainty in the model for the data.

The success of the approach taken here is based on the use of ε-contamination classes of priors. It is worthwhile to summarize the benefits of the use of such classes, since other classes are possible and have been considered.

I. Subjective Interpretation: Prior uncertainty should be reflected in uncertain probabilities, and it is very appealing to model uncertainty by choosing all priors close to an elicited π_0, since very little extra elicitation effort is then needed: just the determination of ε and the types of contaminations to be allowed. (Of course, the ε-contamination approach could also be taken with "objective priors" as the base, π_0.) The need to consider only actual probability distributions, and not unfamiliar constructs based on interval valued set functions, is another attractive aspect of the ε-contamination approach.

II. Flexibility: The flexibility of the ε-contamination class lies in the wide range of choices for the contaminations q. One can work very easily with $\mathcal{Q}$ equal to all priors or with $\mathcal{Q}$ equal to all conjugate priors (which we actually did not do in the paper), and can also work successfully with appealing classes such as all q for which the resulting prior is unimodal. Finally; hierarchical priors can be easily dealt with, as can the often related possibility of incorporating several different types of contaminations.

III. Calculation: The above properties would be of only theoretical interest were it not for the surprising ease with which maximizations and minimizations over the various classes can be performed. This ease is due, in part, to the fact that, frequently, one need only maximize or minimize the desired quantity separately over $\mathfrak{Q}$ (or separately over the different $\mathfrak{Q}_i$, if more than one is involved). Also, quantities of interest, such as the ML-II posterior, posterior mean, and posterior variance can all be expressed as simple weighted averages of the same quantities for each component of π (i.e., π_0 and the q_i). The advantages of this for calculation and interpretation are considerable. Finally, the types of robustness analyses envisaged here are easily implementable on the computer, providing a very attractive systematic alternative to the usual sensitivity analysis of merely trying a few different priors, which (especially for the nonexpert) can run a serious risk of not being extensive enough.

As a final point, many of the above advantages of the ε-contamination class also make it attractive if one takes a frequentist Bayes approach to Bayesian robustness, i.e., if one works with frequentist Bayes measures that average over both x and θ. For instance, Γ-minimax problems (even for Γ such as in Section 3) often reduce to restricted risk Bayes minimization problems (of the Hodges-Lehmann (1952) type), which can frequently be given reasonably simple approximate solutions (c.f. Berger (1982, 1983)).

Appendix

Derivations of (2.16) - (2.19):

Case 2. The probability condition (2.14) reduces to

(A.1) $(1-\alpha)(1+\varepsilon\eta/(1-\varepsilon)m) =$

$$\Phi([(1+\tau^2)^{-1}x+(2\log(\eta a)^{-1})^{1/2}]V^{-1}) - \Phi(-(2\log(\psi aV)^{-1})^{1/2}).$$

Also, the length, L, of the corresponding interval is given by

(A.2) $L = [x+(2\log(\eta a)^{-1})^{1/2} - \delta + V(2\log(\psi aV)^{-1})^{1/2}].$

Since we are assuming that the minimum sized set actually occurs in Case 2, it must satisfy

(A.3) $\frac{dL}{d\eta} = 0.$

Equation (A.3) simplifies to

(A.4) $-\psi'(\eta)/\psi(\eta) = (\eta V)^{-1}(\log(\psi aV)^{-1})^{1/2}(\log(a\eta)^{-1})^{-1/2},$

where $\psi'(\eta) = d\psi(\eta)/d\eta$.

Next, differentiation of (A.1) w.r.t. η yields

(A.5) $R^{-1} = -\exp\{-[(1+\tau^2)^{-1}x+(2\log(\eta a)^{-1})^{1/2}]^2(2V^2)^{-1}\}\ (aV\eta(2\log(\eta a)^{-1}))^{1/2})^{-1}$

$+ (-\psi'(\eta)/\psi(\eta))\{V\psi(\eta)\ (2\log(\psi aV)^{-1})^{-1/2}\}.$

Substitution of (A.4) into (A.5) and some algebra yields (2.17). Simplication of (A.1) yields (2.16).

Case 3. In this case, (2.14) reduces to

$$(\text{A.6}) \qquad (1-\alpha)(1+\varepsilon\eta/(1-\varepsilon)m) =$$

$$2\Phi([2\log(\psi aV)^{-1}]^{1/2})-1 + \Phi([(1+\tau^2)^{-1}x+(2\log(\eta a)^{-1})^{1/2}]V^{-1})$$

$$- \Phi([(1+\tau^2)^{-1}x-(2\log(\eta a)^{-1})^{1/2}]V^{-1}).$$

(Note that the identity $\Phi(z) = 1 - \Phi(-z)$ was used.) The length, L, of the corresponding interval is given by

$$L = 2V(2\log(\ aV)^{-1})^{1/2} + 2(2\log(\eta a)^{-1})^{1/2}.$$

Differentiation of L w.r.t. η and setting the result equal to zero again implies (A.4). Differentiation of (A.6) yields

$$(\text{A.7}) \qquad R^{-1} = - [\exp\{-[(1+\tau^2)^{-1}x+(2\log(\eta a)^{-1})^{1/2}]^2(2V^2)^{-1}\}$$

$$+ \exp\{-[(1+\tau^2)^{-1}x-(2\log(\eta a)^{-1})^{1/2}]^2(2V^2)^{-1}\}] [aV\eta(2\log(a\eta)^{-1})^{1/2}]^{-1}$$

$$+ (-\psi'(\eta)/\psi(\eta))\{2V\psi(\eta)(2\log(\psi aV)^{-1})^{-1/2}\}.$$

Substitution of (A.4) into (A.7) and simplification yields (2.19). Formula (2.18) follows from (A.6).

References

Aitchinson, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis. University Press, Cmabridge.

Berger, J. (1980). Statistical Decision Theory: Foundations, Concepts, and Methods. Springer-Verlag, New York.

Berger, J. (1982). Bayesian robustness and the Stein effect. J. Amer. Statist. Assoc. 77, 358-368.

Berger, J. (1983), The robust Bayesian viewpoint (with Discussion). To appear in Robustness in Bayesian Statistics, ed. J. Kadane. North-Holland, Amsterdam.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with Discussion). J. Roy. Statist. Soc. B 41, 113-147.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.H. (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. A 143, 383-430.

Box, G.E.P. and Tiao, G.C. (1962). A further look at robustness in Bayes' theorem. Biometrika 49, 419-432.

Box, G.E.P. and Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading.

Brown, L.D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. Ann. Math. Statist. 42, 855-904.

Davis, W.A. (1979). Approximate Bayesian predictive distributions and model selection. J. Amer. Statist. Assoc. 74, 312-317.

de Finetti, B. (1961). The Bayesian approach to the rejection of outliers. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1, 199-210. University of California Press, Berkeley.

de Finetti, B. (1972). Probability, Induction and Statistics. Wiley, New York.

DeGroot, M. H. (1970). Optimal Statistical Decisions. McGraw-Hill, New York.

Dempster, A.P. (1971). Model searching and estimation in the logic of inference. In Foundations of Statistical Inference, eds. V.P. Godambe and D.A. Sprott, Holt, Rinehart, and Winston, Toronto.

Dempster, A.P. (1975). A subjectivist look at robustness. Bulletin of the International Statistical Institute 46, 349-374.

Dempster, A.P. (1976). Examples relevant to the robustness of applied inferences. In Statistical Decision Theory and Related Topics III, eds. S.S. Gupta and J. Berger. Academic, New York.

DeRobertis, L. and Hartigan, J.A. (1981). Bayesian inference using intervals of measures. Ann. Statist. 9, 235-244.

Dickey, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. Ann. Math. Statist. 42, 204-223.

Dickey, J.M. (1975). Bayesian alternatives to the F-test and least-squares estimates in the normal linear model. In Studies in Bayesian Econometrics and Statistics in Honor of L.J. Savage, eds. S.E. Fienberg and A. Zellner, 515-554. North-Holland, Amsterdam.

Geisser, S. (1971). The inferential use of predictive distributions (with Discussion). In Foundations of Statistical Inference, eds. V.P. Godambe and D.A. Sprott, 458-469. Holt, Rinehart, and Winston, Toronto.

Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. J. Amer. Statist. Assoc. 74, 153-160.

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. (1967). A Bayesian significance test for multinomial distributions. J. Roy. Statist. Soc. B 29, 399-431.

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

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. J. Roy. Statist. Soc. B 29, 83-10.

Hill, B. (1974). On coherence, inadmissibility and inference about many parameters in the theory of least squares. In Studies in Bayesian Econometrics and Statistics, eds. S.E. Fienberg and A. Zellner, 555-584. North-Holland, Amsterdam.

Hill, B. (1980). Robust analysis of the random model and weighted least squares regression. In Evaluation of Econometric Models, 197-217. Academic, New York.

Hodges, J.L. and Lehmann, E.L. (1952). The use of previous experience in reaching statistical decisions. Ann. Math. Statist. 23, 396-407.

Huber, P.J. (1973). The use of Choquet capacities in statistics. Bulletin of the International Statistical Institute 45, 181-191.

Huber, P.J. (1981). Robust Statistics. Wiley, New York.

Jaynes, E.T. (1968). Prior probabilities. IEEE Transactions on Systems Science and Cybernetics SSC-4, 227-241.

Jeffreys, H. (1961). Theory of Probability (3rd Edition). 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.

Lindley, D.V. (1961). The robustness of interval estimates. Bulletin of the International Statistical Institute 38, 209-220.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model. J. Roy Statist. Soc. B 34, 1-41.

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

Marazzi, A. (1980). Robust Bayesian estimation for the linear model, Research Report No. 27, Fachgruppe fuer Statistik, Eigenoessische Technische Hochschule, Zurich.

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.

Roberts, H.V. (1965). Probabilistic prediction. J. Amer. Statist. Assoc. 60, 50-62.

Robbins, H. (1955). An empirical Bayes approach to statistics. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1, 157-164. University of California Press, Berkeley.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. Ann. Math. Statist. 35, 1-20.

Schneeweiss, H. (1964). Eine Entscheidungsregel für den Fall partiell bekannter Wahrscheinlichkeiten, Unternehmensforschung 8; no. 2, 86-95.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. Wiley, New York.

Zellner, A. (1982). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. Report of the H.G.B. Alexander Research Foundation, Graduate School of Business, University of Chicago.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In Bayesian Statistics, eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, 585-603. University Press, Valencia.