The Annals of Statistics
1990, Vol. 18, No. 1, 454–464

PRIOR ENVELOPES BASED ON BELIEF FUNCTIONS[1]

BY LARRY ALAN WASSERMAN[2]

University of Toronto

We show that the mathematical structure of belief functions makes them suitable for generating classes of prior distributions to be used in robust Bayesian inference. In particular, the upper and lower bounds of the posterior probability content of a measurable subset of the parameter space may be calculated directly in terms of upper and lower expectations (Theorem 4.1). We also extend an integral representation given by Dempster to infinite sets (Theorem 2.1).

1. Introduction. Upper and lower probabilities induced from set-valued mappings were introduced by Dempster (1967, 1968). Shafer (1973, 1976, 1979) called these lower probabilities *belief functions* and generalized Dempster's theory. Associated with a belief function is a convex set of probability measures of which the belief function is a lower bound. (On the other hand, the lower bound of a convex set of probability measures is not necessarily a belief function.) We will show that robust Bayesian inference based on convex sets of prior distributions takes on a particularly tractable form if the set of priors is generated by a belief function (Theorem 4.1).

The theory of belief functions is briefly reviewed in Section 2. Some technical details are omitted. These may be found in Wasserman (1988). Bayesian inference based on envelopes, or sets of probability measures, is discussed in Section 3. Following this, the theory of prior envelopes derived from belief functions is developed in Section 4. In particular, we show that robust Bayesian inference using these envelopes is mathematically tractable and amenable to straightforward numerical approximations. In Section 5 we show that some classes of priors that are commonly used may be generated by belief functions. Section 6 develops a general class of belief function envelopes that can be used to model local uncertainty about a prior distribution. Finally, Section 7 contains a discussion of the results.

2. Belief functions. Let Θ be a Polish space with Borel σ-algebra $\mathscr{B}(\Theta)$ and let X be a convex, compact, metrizable subset of a locally convex topological vector space with Borel σ-algebra $\mathscr{B}(X)$. Let μ be a probability measure on $(X, \mathscr{B}(X))$ and let Γ be a map taking points in X to nonempty, closed subsets of Θ. If X is countable, then the restriction that each $\Gamma(x)$ be closed may be

Received August 1987; revised May 1989.
[1]Research supported by a studentship from the Medical Research Council of Canada.
[2]Now at Carnegie-Mellon University.
AMS 1980 *subject classifications*. Primary 62A15; secondary 62F15.
Key words and phrases. Belief functions, Choquet capacities, Markov kernel, robust Bayesian inference.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*.
www.jstor.org

dropped. For each $A \subset \Theta$, define

$$A_* = \{x \in X;\ \Gamma(x) \subset A\}$$

and

$$A^* = \{x \in X;\ \Gamma(x) \cap A \neq \varnothing\}.$$

Γ is called *strongly measurable* if $A \in \mathscr{B}(\Theta)$ implies $A^* \in \mathscr{B}(X)$. [It follows that $A_* \in \mathscr{B}(X)$ as well.] Natural conditions, such as upper or lower semicontinuity, may be imposed on Γ to guarantee measurability; see Matheron (1975) and Wasserman (1988). Define Bel and Pl on $(\Theta, \mathscr{B}(\Theta))$ by

$$\mathrm{Bel}(A) = \mu(A_*)$$

and

$$\mathrm{Pl}(A) = \mu(A^*).$$

Bel is called a *belief function* and Pl is called a *plausibility function*. The four-tuple $(X, \mathscr{B}(X), \mu, \Gamma)$ is called a *source* for Bel. Bel and Pl are related by $\mathrm{Bel}(A) = 1 - \mathrm{Pl}(A^c)$. An intuitive explanation of Bel and Pl is as follows. Draw x randomly according to μ. Then $\mathrm{Bel}(A)$ is the probability that the random set $\Gamma(x)$ is contained in A and $\mathrm{Pl}(A)$ is the probability that the random set $\Gamma(x)$ hits A. Note that $\mathrm{Bel}(\varnothing) = \mathrm{Pl}(\varnothing) = 0$, $\mathrm{Bel}(\Theta) = \mathrm{Pl}(\Theta) = 1$ and $\mathrm{Bel}(A) \leq \mathrm{Pl}(A)$ with equality if and only if Bel is a probability measure. Thus, belief functions contain probability measures as a special case.

A probability measure P is said to be *compatible* with Bel and Pl if for every $A \in \mathscr{B}(\Theta)$, $\mathrm{Bel}(A) \leq P(A) \leq \mathrm{Pl}(A)$. Let Π be the set of all probability measures compatible with Bel and Pl. It can be shown that Π is nonempty and that for each $A \in \mathscr{B}(\Theta)$,

$$\mathrm{Bel}(A) = \inf_{P \in \Pi} P(A)$$

and

$$\mathrm{Pl}(A) = \sup_{P \in \Pi} P(A).$$

This fact implies that Bel and Pl may be the thought of as the lower and upper bounds of a class of probability measures. This is the interpretation we shall emphasize in this paper. We should point out that in general, Bel is treated as an object of interest itself, rather than as a lower bound of a set of probability measures.

Now we define the notion of upper and lower expectation. For each $\mathscr{B}(\Theta)$-measurable, bounded, real-valued function f on Θ, define the upper expectation $E^*(f)$ and the lower expectation $E_*(f)$ by

$$E^*(f) = \sup_{P \in \Pi} E_P(f) \quad \text{and} \quad E_*(f) = \inf_{P \in \Pi} E_P(f),$$

where $E_P(f) = \int f(\theta) P(d\theta)$. It follows easily that $P \in \Pi$ if and only if $E_P(f) \leq$

$E^*(f)$. It can be shown that

$$E^*(f) = \int f^*(x)\mu(dx) \quad \text{and} \quad E_*(f) = \int f_*(x)\mu(dx),$$

where f^* and f_* are defined by $f^*(x) = \sup_{\theta \in \Gamma(x)} f(\theta)$ and $f_*(x) = \inf_{\theta \in \Gamma(x)} f(\theta)$. This fact has important implications for computation because it reduces the problem of calculating extrema over the set Π to that of finding extrema of f over subsets of Θ followed by a single integral over X. It may be verified that the strong measurability of Γ implies that f^* and f_* are $\mathscr{B}(X)$-measurable. Conversely, if f^* and f_* are $\mathscr{B}(X)$-measurable whenever f is $\mathscr{B}(\Theta)$-measurable, then Γ is strongly measurable.

Now we show how a theorem of Strassen's may be applied to derive an integral representation for the set of compatible measures of a belief function. A proof of this representation in the case where Θ is finite is given in Dempster (1967).

THEOREM 2.1. *$P \in \Pi$ if and only if there exists, for μ-almost all x, a probability measure π_x on $\mathscr{B}(\Theta)$ supported by $\Gamma(x)$ such that*

$$P(A) = \int_X \pi_x(A)\mu(dx)$$

for each $A \in \mathscr{B}(\Theta)$.

PROOF. Suppose $P \in \Pi$. Recall that function π on $\mathscr{B}(\Theta) \times X$ is a Markov kernel if $\pi_x(\cdot)$ is a probability measure on $\mathscr{B}(\Theta)$ for each $x \in X$ and $\pi_{\cdot}(A)$ is $\mathscr{B}(X)$-measurable for each $A \in \mathscr{B}(\Theta)$. An application of Theorem 3 of Strassen (1965) shows that for each $P \in \Pi$, there is a Markov kernel π such that $P(A) = \int \pi_x(A)\mu(dx)$ for each $A \in \mathscr{B}(\Theta)$ and such that $\pi_x \in K_x$ for μ-almost all x, where K_x is the set of all probability measures on $\mathscr{B}(\Theta)$ with support in $\Gamma(x)$. (The facts stated earlier guarantee that the appropriate conditions necessary to apply Strassen's theorem hold.) On the other hand, if the integral equation holds, it follows easily that $P \in \Pi$.

Theorem 2.1 may be given the following interpretation. To construct a compatible measure, draw points from Θ randomly in the following fashion. First draw x randomly from X. Next draw θ randomly from $\Gamma(x)$ according to π_x.

3. Robust Bayesian inference. Thorough discussions of robust Bayesian inference may be found in Berger (1984, 1985). A related approach based on upper and lower probabilities is discussed in Walley (1981). We begin by giving a precise statement of Bayes' theorem; See DeRobertis and Hartigan (1981) for details. Let $(Y, \mathscr{B}(Y))$ be a sample space and let $f(y|\theta)$, $\theta \in \Theta$, be a class of densities on Y with respect to a dominating σ-finite measure ν, where $(\Theta, \mathscr{B}(\Theta))$ is the parameter space. If P is a prior probability measure on $(\Theta, \mathscr{B}(\Theta))$ and $f(y|\theta)$ is $\mathscr{B}(Y) \times \mathscr{B}(\Theta)$-measurable, then there is a unique probability measure on $(Y \times \Theta, \mathscr{B}(Y) \times \mathscr{B}(\Theta))$ with θ-marginal P and whose conditional distribution

given $\mathscr{B}(\Theta)$ has density $f(y|\theta)$ with respect to $\nu \times P$. The regular conditional probability of this measure, given $\mathscr{B}(Y)$, has density $f(y|\theta)/\int f(y|\theta)P(d\theta)$ with respect to P, given y. This measure is called the posterior distribution given the data. If more than one observation is sampled independently, we can use the product measure on the sample space. Thus, no loss of generality occurs by assuming a single observation.

Now suppose a prior P cannot be accurately specified. Then we might consider using a class of priors Π. We call a nonempty convex set of probability measures, an *envelope*. Each prior P in the envelope Π may be updated by Bayes' theorem to produce a posterior measure P_y. Denote this class of posterior probability measures by Π_y. Define $P_*(A) = \inf_{P\in\Pi} P(A)$ and $P^*(A) = \sup_{P\in\Pi} P(A)$. These functions are related by $P_*(A) = 1 - P^*(A^c)$. Similarly define $P_{y*}(A) = \inf_{P_y\in\Pi_y} P_y(A)$ and $P_y^*(A) = \sup_{P_y\in\Pi_y} P_y(A)$.

A robust Bayesian analysis proceeds by reporting $P_{y*}(A)$ and $P_y^*(A)$ rather than a single posterior probability. The difference between the lower and upper bound is an indication of the robustness of the analysis to the specification of the prior. It is therefore of interest to develop simple methods for computing these bounds. The next section considers this issue in the context of belief functions.

4. Envelopes based on belief functions. Let Bel be a belief function on $(\Theta, \mathscr{B}(\Theta))$ with source $(X, \mathscr{B}(X), \mu, \Gamma)$ and let Π be the convex class of measures compatible with Bel. Here we are viewing Bel as a convenient method for constructing a class of priors. Let $L_A(\theta) = L(\theta)I_A(\theta)$ where $L(\theta) = f(y|\theta)$ is the likelihood function and $I_A(\theta)$ is the indicator function for the set A. Now we state our main result which shows that the bounds of the posterior probability derived from a belief function envelope take a special form.

THEOREM 4.1. *If $L(\theta)$ is bounded, then for any $A \in \mathscr{B}(\Theta)$,*

$$P_{y*}(A) = \frac{E_*(L_A)}{E_*(L_A) + E^*(L_{A^c})} = \frac{E_\mu(L_{A*})}{E_\mu(L_{A*}) + E_\mu(L_{A^c}^*)}$$

and

$$P_y^*(A) = \frac{E^*(L_A)}{E^*(L_A) + E_*(L_{A^c})} = \frac{E_\mu(L_A^*)}{E_\mu(L_A^*) + E_\mu(L_{A^c*})}.$$

PROOF. We shall prove the formula for $P_y^*(A)$. The proof for the lower bound is similar. The strategy of the proof is to define π_x to be a probability measure supported by $\Gamma(x)$ that puts all its mass in the location that maximizes the posterior probability. If $\Gamma(x)$ has a nonempty intersection with A, this means putting the mass where the likelihood is greatest. Otherwise, we put the mass where the likelihood is smallest. We now proceed with the details.

For any $P \in \Pi$,

$$P_y(A) = \frac{\int_A f(y|\theta)P(d\theta)}{\int_\Theta f(y|\theta)P(d\theta)} = \frac{E_P(L_A)}{E_P(L_A) + E_P(L_{A^c})}.$$

Clearly

$$\frac{E^*(L_A)}{E^*(L_A) + E_*(L_{A^c})}$$

is an upper bound for $P_y^*(A)$ and by our remarks in Section 2, this upper bound is equal to

$$\frac{E_\mu(L_A^*)}{E_\mu(L_A^*) + E_\mu(L_{A^c*})}.$$

We shall now show that there is a net of measures $Q^\gamma \in \Pi$ such that $Q_y^\gamma(A)$ converges to this upper bound, implying that this is the least upper bound.

Since for any measurable function f, $\sup_{P \in \Pi} E_P(f) = E^*(f)$, there is a net P^α such that $E_{P^\alpha}(L_A) \to E^*(L_A)$. Corresponding to P^α, there is a net of Markov kernels π_x^α such that $P^\alpha(B) = \int_X \pi_x^\alpha(B)\mu(dx)$ for every $B \in \mathscr{B}(\Theta)$. Similarly, there are nets P^β and π_x^β such that $E_{P^\beta}(L_{A^c}) \to E_*(L_{A^c})$ and $P^\beta(B) = \int_X \pi_x^\beta(B)\mu(dx)$. We define a net q_x^α by

$$q_x^\beta(B) = \begin{cases} \pi_x^\beta(B), & \text{if } \pi_x^\beta(A) = 0, \\ \dfrac{\pi_x^\beta(B \cap A)}{\pi_x^\beta(A)}, & \text{if } \pi_x^\beta(A) > 0. \end{cases}$$

It is straightforward to show that for each α, q_x^α is a Markov kernel. This defines a net Q^α in Π by $Q^\alpha(B) = \int_X q_x^\alpha(B)\mu(dx)$. In a similar way, define nets q_x^β and Q^β by

$$q_x^\beta(B) = \begin{cases} \pi_x^\beta(B), & \text{if } \pi_x^\beta(A) = 0, \\ \dfrac{\pi_x^\beta(B \cap A)}{\pi_x^\beta(A)}, & \text{if } \pi_x^\beta(A) > 0, \end{cases}$$

and $Q^\beta(B) = \int_X q_x^\beta(B)\mu(dx)$. Now define a net q_x^γ by

$$q_x^\gamma = \begin{cases} q_x^\alpha, & \text{if } x \in A^*, \\ q_x^\beta, & \text{if } x \notin A^*. \end{cases}$$

It can be verified that the q_x^γ are also Markov kernels and define a net Q^γ by way of $Q^\gamma(B) = \int_X q_x^\gamma(B)\mu(dx)$. Since $L_A(\theta)$ is nonnegative and vanishes outside A, and since Q^α dominates P^α for subsets of A,

$$E_{P^\alpha}(L_A) \le E_{Q^\alpha}(L_A) \le E^*(L_A).$$

The convergence of the quantity on the left-hand side to the quantity on the right implies that $E_{Q^\alpha}(L_A) \to E^*(L_A)$. Also, $L_{A^c}(\theta)$ is nonnegative and vanishes outside A^c and P^β dominates Q^β on A^c. Hence,

$$E_{P^\beta}(L_{A^c}) \ge E_{Q^\beta}(L_{A^c}) \ge E_*(L_{A^c}).$$

The convergence of $E_{P^\beta}(L_{A^c})$ to $E_*(L_{A^c})$ implies that $E_{Q^\beta}(L_{A^c}) \to E_*(L_{A^c})$.

Now, applying Theorem 2.1 and using the fact that L_A vanishes on the support of q_x^β for each $x \in (A^*)^c$,

$$E_{Q^\gamma}(L_A) = \int_\Theta L_A(\theta) Q^\gamma(d\theta) = \int_X \int_\Theta L_A(\theta) q_x^\gamma(d\theta)\mu(dx)$$

$$= \int_{A^*} \int_\Theta L_A(\theta) q_x^\alpha(d\theta)\mu(dx) = E_{Q^\alpha}(L_A).$$

Now,

$$E_{Q^\gamma}(L_{A^c}) = \int_\Theta L_{A^c}(\theta) Q^\gamma(d\theta) = \int_X \int_\Theta L_{A^c}(\theta) q_x^\gamma(d\theta)\mu(dx)$$

$$= \int_{(A^*)^c} \int_\Theta L_{A^c}(\theta) q_x^\beta(d\theta)\mu(dx).$$

This last equality follows from the fact that $q_x^\alpha(A^c) = 0$ for each $x \in A^*$. Also,

$$E_{Q^\beta}(L_{A^c}) = \int_X \int_\Theta L_{A^c}(\theta) q_x^\beta(d\theta)\mu(dx)$$

$$= \int_{(A^*)^c} \int_\Theta L_{A^c}(\theta) q_x^\beta(d\theta)\mu(dx) = E_{Q^\gamma}(L_{A^c}).$$

Finally, this leads to

$$\lim_\gamma Q_y^\gamma(A) = \lim_\gamma \frac{E_{Q^\gamma}(L_A)}{E_{Q^\gamma}(L_A) + E_{Q^\gamma}(L_{A^c})} = \frac{\lim_\alpha E_{Q^\alpha}(L_A)}{\lim_\alpha E_{Q^\alpha}(L_A) + \lim_\beta E_{Q^\beta}(L_{A^c})}$$

$$= \frac{E^*(L_A)}{E^*(L_A) + E_*(L_{A^c})}. \qquad \square$$

This theorem shows that when a class of priors can be generated by a belief function, the upper and lower bounds of the posterior probability for a subset of the parameter space take on a tractable form. There always exists, at least in principle, a Monte Carlo method for estimating P_{y*} and P_y^*. One may simply draw x's randomly from X and estimate the expectations with the sample average. Thus the problem of finding extrema over Π may be approached by sampling from a single measure μ on X. This is what distinguishes sets of probability measures that are compatible with belief functions from other sets of probability measures.

For a general class of priors Π, not necessarily generated by a belief function, we can still define an upper and lower expectation via $E^*(f) = \sup_{P \in \Pi} E_P(f)$ and $E_*(f) = \inf_{P \in \Pi} E_P(f)$. It is tempting to use the formula given in Theorem 4.1 in this case. To see that the formula does not hold in general, consider the following example. Let $\Theta = \{1, 2, 3, 4\}$ and let the vector of likelihood values on Θ be (a, b, c, d). Suppose that a, b, c and d are all positive and that $a > c$, $b > d$ and $ad > bc$. Now let P be a probability measure with values $(\frac{1}{2}, \frac{1}{2}, 0, 0)$ on the singletons of Θ and let Q have values $(0, 0, \frac{1}{2}, \frac{1}{2})$. Finally, let Π be the convex closure of P and Q. It may be verified that P^* is not a plausibility

function where P^* is the upper probability generated by Π. Let $A = \{1,3\}$. Then $P_y^*(A) = a/(a + b)$. However, $E^*(L_A)/(E^*(L_A) + E_*(L_{A^c})) = a/(a + d) > P_y^*(A)$, so equality in the theorem fails.

Recall that a normed set function P^*, defined on the subsets of a finite set Θ, is an alternating Choquet capacity of order n if, for each $A_1, \ldots, A_n$,

$$P^*(\cap A_i) \leq \sum_i P^*(A_i) - \sum_{i \neq j} P^*(A_i \cup A_j) + \cdots + (-1)^{n+1} P^*(A_1 \cup \cdots \cup A_n).$$

P^* is alternating of order ∞ if it is alternating of order n for each n. See Choquet (1953) and Huber and Strassen (1973) for details. It can be shown that a function is alternating of order ∞ if and only if it can be represented as a plausibility function [Matheron (1975) and Shafer (1979)]. A version of Theorem 4.1 will be proved in a forthcoming paper for capacities of order 2. [See Wasserman and Kadane (1990). Also, see Walley (1981).] The proof of that result is considerably different than the proof of Theorem 4.1. In particular, there is no notion of sampling random sets, nor do we have the luxury of a Markov kernel representation in that case. In this paper, we have exploited the properties that distinguish belief functions from other lower probabilities. This is particularly important for gaining insights into the mathematical structure of the compatible class Π and for suggesting methods of computation and approximation.

5. Some examples. In this section we show that two well-known classes of probability measures may be generated by belief functions.

EXAMPLE 5.1 (Probabilities on partitions). Let $h = (h_1, \ldots, h_k)$ be a partition of Θ and suppose that one only specifies the prior probability content p_i of each h_i. Let $\Pi = \{P;\ P(h_i) = p_i\}$. Then Π is generated by a belief function with source $(X, 2^X, \mu, \Gamma)$ where $X = \{x_1, \ldots, x_k\}$, $\mu(\{x_i\}) = p_i$ and $\Gamma(x_i) = h_i$. Applying Theorem 4.1, we see that

$$P_y^*(h_i) = \frac{p_i L_i^*}{p_i L_i^* + \sum_{j \neq i} p_j L_{j*}}$$

and

$$P_{y*}(h_i) = \frac{p_i L_{i*}}{p_i L_{i*} + \sum_{j \neq i} p_j L_j^*},$$

where $L_{i*} = \inf_{\theta \in h_i} L(\theta)$ and $L_i^* = \sup_{\theta \in h_i} L(\theta)$. These bounds were given in Berliner and Goel (1986) using a different argument.

EXAMPLE 5.2 (Contaminated priors). Suppose we have an initial prior π on Θ. We are not completely confident about the prior so we form the following class of priors:

$$\Pi = \{P;\ P = (1 - \varepsilon)\pi + \varepsilon Q,\ Q \in \mathcal{Q}\},$$

where ε is a fixed number between 0 and 1 and $\mathcal{Q}$ is the set of all probability measures on $\mathcal{B}(\Theta)$. This is the class of ε-contaminated priors considered by

Huber (1973) and Berger and Berliner (1986). The corresponding belief function has source $(X, \mathscr{B}(X), \mu, \Gamma)$ where $X = \Theta \cup \{x_0\}$, $\mathscr{B}(X) = \mathscr{B}(\Theta) \times \{x_0\}$, $\mu = (1 - \varepsilon)\pi' + \varepsilon\delta$ and $\Gamma(x) = \{x\}$ if $x \in \Theta$ and equals Θ if $x = x_0$. Here, π' is a probability measure on X that gives zero probability to x_0 and is identical to π on $X - \{x_0\}$ while δ is a point mass on x_0. In words, we draw $\{\theta\}$ with probability $(1 - \varepsilon)\pi(d\theta)$ and draw Θ with probability ε. We then have that

$$P_y^*(A) = \frac{(1 - \varepsilon)\int_A L(\theta)\pi(d\theta) + \varepsilon a}{(1 - \varepsilon)\int_\Theta L(\theta)\pi(d\theta) + \varepsilon a}$$

and

$$P_{y*}(A) = \frac{(1 - \varepsilon)\int_A L(\theta)\pi(d\theta)}{(1 - \varepsilon)\int_\Theta L(\theta)\pi(d\theta) + \varepsilon b},$$

where $a = \sup_{\theta \in A} L(\theta)$ and $b = \sup_{\theta \in A^c} L(\theta)$. This result was stated in Huber (1973) and Berger and Berliner (1986).

6. Local perturbations to priors. We can exploit the rich mathematical structure of belief functions to define new interesting classes of probability measures. A general class of measures is examined in this section. Suppose we begin with a prior measure π on $\mathscr{B}(\Theta)$. Not being completely confident about the prior, we wish to consider a class of priors Π. It may be that we are less certain about some parts of the prior than others. For example, we may feel confident about the central portion of the prior but not the tails. This suggests that we model our uncertainty locally. Thus, to each $\theta \in \Theta$, we attach a subset $\Gamma(\theta) \subset \Theta$ such that $\theta \in \Gamma(\theta)$. Any probability measure that results by starting with π and moving the mass at θ to any point of $\Gamma(\theta)$ will be called a local perturbation of π. Formally, we call a belief function Bel a *local perturbation* of π if Bel has source $(\Theta, \mathscr{B}(\Theta), \pi, \Gamma)$ and $\theta \in \Gamma(\theta)$ for each $\theta \in \Theta$. $\Gamma(\theta)$ is taken to be a large subset in a region of high uncertainty and is taken to be small in a region of low uncertainty. Theorem 4.1 can then be applied to find bounds on the posterior probabilities. Consider an example.

Suppose Y is normally distributed with mean θ and variance 1. Let our initial prior π be normal with mean 0 and variance 2. Set $\Gamma_c(\theta) = [\theta - c, \theta + c]$. Let P_c^* be the upper prior probability based on Γ_c. The first graph in Figure 1 shows $P_c^*(\{\theta\})$ for $c = 0.01$, 0.1 and 1. Let P_ε^* be the upper probability derived from the ε-contaminated class of priors. For comparison, $P_\varepsilon^*(\{\theta\})$, with $\varepsilon = 0.1$, is also graphed. This graph allows us to inspect the local behavior of the upper probability. A more vivid comparison between the local perturbation method and the ε-contaminated model is given in the second graph of Figure 1 which plots $P_c^*(\{\theta\})/\pi(\theta)$ and $P_\varepsilon^*(\{\theta\})/\pi(\theta)$ which we call the relative upper probability. Here we see that $P_\varepsilon^*(\{\theta\})$ allows too much mass in the tails. In particular, an ε point mass is permitted at any point θ. On the other hand, the local perturbation model restricts the mass that may travel to the tail. This model takes into account the topological structure of Θ. The ε-contaminated model ignores the information that certain points are closer together than others.

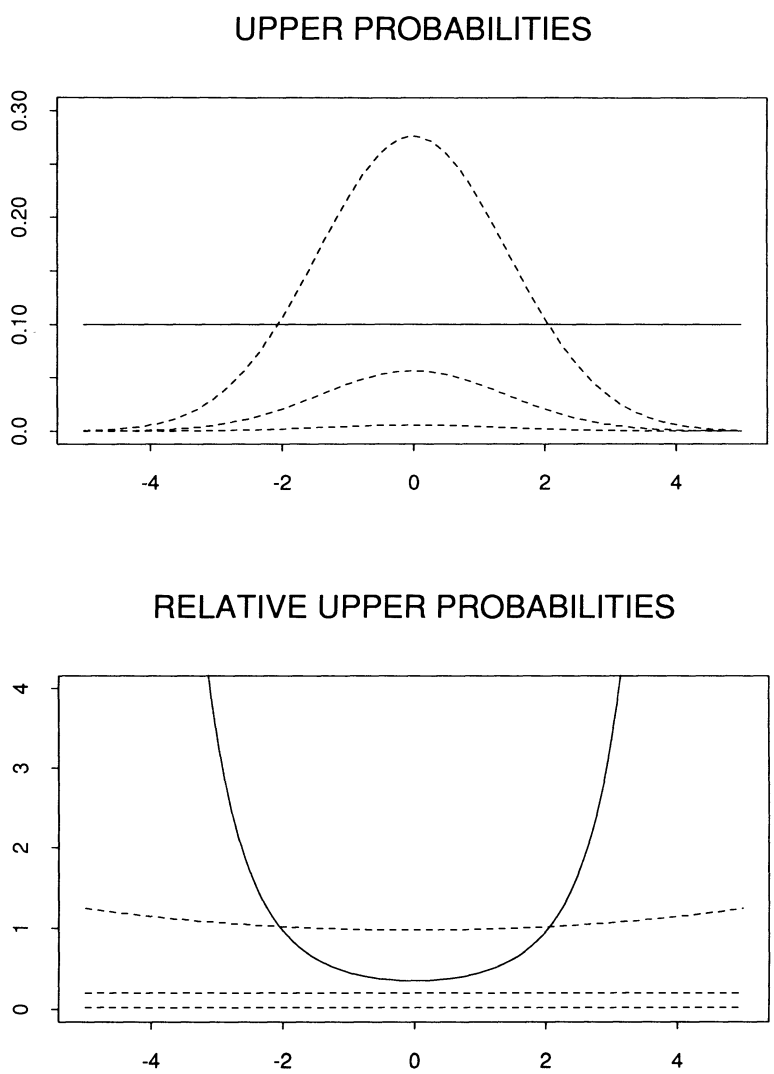

FIG. 1. *The upper graph gives the values of* $P^*(\{\theta\})$. *The lower graph gives the values of* $P^*(\{\theta\})/\pi(\theta)$. *In each graph, the solid line is based on the* ε*-contaminated prior with* $\varepsilon = 0.1$ *and the remaining curves, in increasing order, are based on* Γ_c *for* $c = 0.01, 0.1$ *and* 1.

Now consider two numerical examples. A reasonable way to choose c is to set an upper bound on the probability of some reference set R. Consider the set $R = (-\sqrt{2}, \sqrt{2})$ which has a prior probability of 0.683 under π. Suppose we felt that an upper probability of 0.714 is reasonable. An ε-contaminated class of priors with $\varepsilon = 0.1$ gives $P_\varepsilon^*(R) = 0.714$. Choosing $c = 0.095$ also gives $P_c^*(R) = 0.714$. Let us proceed with these chosen values for ε and c. Suppose

first the $Y = 0.5$ is observed. Then the usual 95% credible region based on π is $T = [-1.267, 1.93]$. Let Q_c^* be the upper posterior corresponding the prior P_c^* and let Q_ε^* denote the upper posterior corresponding to P_ε^*. The lower posteriors are denoted in a similar way. We find that $Q_\varepsilon^*(T) = 0.958$ and $Q_{\varepsilon *}(T) = 0.887$ while $Q_c^*(T) = 0.970$ and $Q_{c*}(T) = 0.923$ so that the local perturbation method gives slightly tighter bounds on the posterior. Now suppose we have the more extreme observation $Y = 4$. The 95% credible region is $T = [1.07, 4.27]$. Then $Q_\varepsilon^*(T) = 0.987$ and $Q_{\varepsilon *}(T) = 0.259$ while $Q_c^*(T) = 0.972$ and $Q_{c*}(T) = 0.917$. The large point mass permitted in the ε-contaminated model makes the posterior bounds overly sensitive to extreme observations.

The problems with ε-contaminated priors are well known and much effort has gone into finding restrictions on this class of priors to overcome these difficulties; see Berger and Berliner (1986), for example. We propose that the local perturbation method might be a reasonable alternative class.

Future research on these classes should focus on choosing the shape and size of the sets $\Gamma(\theta)$. In a multiparameter problem, we might construct Γ to reflect a greater uncertainty about nuisance parameters. Also, efficient computation and theoretical study of the behavior of Q_c^* are issues deserving further attention.

7. Discussion. We have shown that robust Bayesian inference based on classes of priors generated by belief functions enjoy special properties. In particular, the problem of finding extrema over the set of priors is reduced to that of maximizing and minimizing the likelihood over sets in the parameter space followed by an integration. This is not an easy task, but sometimes might be simpler than maximizing and minimizing over the set of priors.

Much remains to be done from a practical point of view. For example, methods of choosing prior belief functions such as those introduced in Section 6 need to be studied and exemplified. Other useful belief functions need to be investigated as well. Also note that Theorem 4.1 is a vehicle with which theoretical properties of posterior probability bounds may be studied.

Another problem that might be fruitfully studied is that of approximation. That is, given a class of probability measures Π, does there exist a way of approximating this class with a set of measures generated by a belief function? Computationally, such an approximation might make otherwise intractable problems accessible.

Finally, it would be useful to know when the lower posterior probability can be represented as a belief function. This could simplify the process of computing posterior expectations and would also be useful for sequentially updating the posterior probabilities.

Acknowledgments. I would like to thank Mike Evans for invaluable discussions on this problem and for helping me with the proof of Theorem 4.1. I would also like to thank Rob Tibshirani for his comments on an earlier draft of this paper. Finally, I am grateful to an Associate Editor and a referee for many useful suggestions.

REFERENCES

BERGER, J. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness in Bayesian Statistics* (J. Kadane, ed.) 63–144. North-Holland, Amsterdam.

BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BERGER, J. and BERLINER, M. (1986). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. *Ann. Statist.* **14** 641–486.

BERLINER, L. M. and GOEL, P. K. (1986). Incorporating partial prior information: Ranges of posterior probabilities. Technical Report No. 357, Dept. Statist., Ohio State Univ.

CHOQUET, G. (1953). Theory of capacities. *Ann. Inst. Fourier (Grenoble)* **5** 131–295.

DEMPSTER, A. (1967). Upper and lower probabilities induced from a multivalued mapping. *Ann. Math. Statist.* **38** 325–339.

DEMPSTER, A. (1968). A generalization of Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **30** 205–247.

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

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

HUBER, P. J. and STRASSEN, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist.* **1** 251–263.

MATHERON, G. (1975). *Random Sets and Integral Geometry*. Wiley, New York.

SHAFER, G. (1973). Allocations of probability: A theory of partial belief. Ph.D. dissertation, Princeton Univ.

SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Univ. Princeton Press, Princeton, N.J.

SHAFER, G. (1979). Allocations of probability. *Ann. Prob.* **7** 827–839.

STRASSEN, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.* **36** 423–439.

WALLEY, P. (1981). Coherent lower (and upper) probabilities. Statistics research report, Univ. of Warwick, Coventry.

WASSERMAN, L. (1988). Some applications of belief functions to statistical inference. Ph.D. dissertation, Univ. of Toronto.

WASSERMAN, L. and KADANE, J. (1990). Bayes' theorem for Choquet capacities. *Ann. Statist.* To appear.

DEPARTMENT OF STATISTICS
CARNEGIE-MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213-3890