

Default Priors for Robust Bayesian Inference

George Casella¹ and Martin T. Wells

Cornell University

BU-1124-MA

January 1992

Summary

Robust Bayesian inference involves examining the performance of Bayes rules from a class of prior distributions. However, given a specified class of priors, there is no mechanism for choosing a reasonable default option, that is, a robust prior that is somehow noninformative. We show that, by applying results of classical robustness theory, such priors can be easily defined. These priors display attractive robustness properties and also provide a means for "touring" through a class of priors.

¹Research supported by National Science Foundation Grant #DMS9100839 and National Security Agency Grant No. 90F073.

1. Introduction

One attractive feature of the Bayesian approach to statistical inference is its simplicity. Once the loss function and prior distribution have been chosen the calculation of the Bayes rule is a rather straightforward procedure. This simplicity makes Bayesian statistics quite appealing. However, when there is little or no prior information available the Bayesian is at a loss. In this case a Bayesian sometimes constructs a default or “noninformative prior,” that is, a prior that should contain no information about the parameter of interest. The problem of constructing noninformative priors has always attracted a fair amount of attention. For example, see Berger (1985), Box and Tiao (1973), Jeffreys (1961), and Zellner (1987). Perhaps most importantly Bernardo (1979) introduced the concept of reference priors, reviewed in Berger and Bernardo (1991).

A procedure for generating default priors was introduced by Jeffreys (1946). Assume that the probability measures in some class \mathfrak{P} are absolutely continuous with respect to a dominating measure μ , and let $f(x|\theta)$ denote the density of $P_\theta \in \mathfrak{P}$ at the observed value x . Jeffreys recommended the prior

$$\pi_J(\theta) = \left(\det I(\theta) \right)^{\frac{1}{2}} = |I(\theta)|^{\frac{1}{2}} \quad (1.1)$$

where

$$I(\theta) = E_\theta \left[\left(\nabla \log f(x|\theta) \right) \left(\nabla \log f(x|\theta) \right)^T \right]$$

is the Fisher information matrix at θ . Here $\nabla \log f(x|\theta)$ is the vector of first partial derivatives of $\log f(x|\theta)$ evaluated at θ . It may be shown that π_J has the desirable property of parameter invariance.

Jeffreys' original derivation of π_J was based on its connection to the Kullback-Leibler divergence metric (Kullback, 1959)

$$D_K(P_\theta, P_\phi) = \int f(x|\theta) \log[f(x|\theta)/f(x|\phi)] d\mu(x)$$

as well as the squared Hellinger distance

$$D_H(P_\theta, P_\phi) = \int \left[\sqrt{f(x|\theta)} - \sqrt{f(x|\phi)} \right]^2 d\mu(x) ,$$

where μ is a dominating measure. Essentially, Jeffreys' deviation consisted of showing that D_K and D_H behave locally like $I(\theta)$. This can be seen by observing that for $D = D_H$ or $D = D_K$

$$I(\theta) = \nabla^2 D(P_\theta, P_\phi) \Big|_{\phi=\theta}, \quad (1.2)$$

where ∇^2 is the Hessian with respect to ϕ . Since

$$\nabla D(P_\theta, P_\phi) \Big|_{\phi=\theta} = 0. \quad (1.3)$$

it follows that D_H and D_K behave locally like Fisher information. Also, since $I(\theta)$ is positive definite, (1.2) and (1.3) imply that π_J is the solution to a minimization problem.

Jeffrey's prior can also be derived as the solution to an asymptotic minimization problem, as shown by Clarke and Barron (1990). (See also Polson, 1988.) This development is similar in spirit to the reference prior approach of Bernardo (1979), and establishes the asymptotic expansion

$$E[D_K(\pi(\theta | x), \pi(\theta))] = \frac{k}{2} \log \frac{n}{2\pi\epsilon} + \int \pi(\theta) \log \frac{|I(\theta)|^{\frac{1}{2}}}{\pi(\theta)} d\theta + o(1) \quad (1.4)$$

for the Kullback-Leibler divergence between the posterior and prior distribution. Maximizing (1.4) over all priors π should yield a prior that imparts the least information, and Clarke and Barron (1990) show that a limiting form of (1.4) is maximized by $\pi_J = |I(\theta)|^{\frac{1}{2}}$. Extensions of this result to the case of nuisance parameters have been given by Clarke and Wasserman (1991).

These derivations of π_J leads us to interpret it as a "noninformative" prior. However, in the absence of good prior information it is not enough to require only that a default prior be noninformative. Since the Bayesian decision rules are fundamentally dependent on the prior information there must be concern about the influence of the prior. Such concerns have prompted Bayesians to construct procedures which are robust with respect to the specification of the prior distribution. The groundwork for robust Bayes analysis was laid out by Berger (1985), and more recently surveyed by Berger (1990) and Wasserman (1991).

A natural means to investigate robustness with respect to a prior is to specify a class Γ of plausible prior distributions and see how the choices among the priors in Γ affect the constructed decision procedures. Let $\rho(\pi)$ be some posterior quantity of interest (such as a posterior mean,

variance, or credible set probability). For the class of priors Γ much research in robust Bayesian analysis has concentrated on evaluating

$$\rho_{\Gamma}^L = \inf_{\pi \in \Gamma} \rho(\pi) \quad , \quad \rho_{\Gamma}^U = \sup_{\pi \in \Gamma} \rho(\pi) \quad .$$

If the range $(\rho_{\Gamma}^L, \rho_{\Gamma}^U)$ is small enough that the prior is deemed to be noninfluential the prior is called robust.

Perhaps the most common Γ that have been considered are neighborhood classes, one example of which is the ϵ -contamination class of priors. For a single elicited (baseline) prior π_0 , the class is

$$\Gamma_{\epsilon}^c(\pi_0) = \{ \pi : \pi(\theta) = (1 - \epsilon)\pi_0(\theta) + \epsilon q(\theta) , \quad q \in Q \} \quad ,$$

Here ϵ ($0 < \epsilon \leq 1$) reflects the amount of uncertainty in the elicited prior π_0 , and Q is a class of densities determining the amount of contamination which is to be mixed with π_0 . The choice of Q usually reflects the researcher's notion of a reasonable prior. Possibilities for Q include

$$\begin{aligned} Q_A &= \{ \text{all distributions } q \} \quad , \\ Q_U &= \{ \text{all distribution } q \text{ with mode } \theta_0 \text{ (fixed)} \} \quad , \\ Q_{SU} &= \{ \text{all symmetric unimodal distributions } q \text{ with mode } \theta_0 \text{ (fixed)} \} . \end{aligned}$$

The class with Q_A is usually the easiest to work with since there are simple characterizations of ρ^L and ρ^U for various statistical functions (see Huber, 1973). Results for the classes Q_U and Q_{SU} are contained in Sivaganisan and Berger (1989).

Other common neighborhood classes are the Kolmogorov and Levy neighborhoods given by

$$\begin{aligned} \Gamma_{\epsilon}^{KS}(\pi_0) &= \{ P : P_0(\theta) - \epsilon \leq P(\theta) \leq P_0(\theta) + \epsilon , \text{ for all } \theta \} , \\ \Gamma_{\epsilon, \delta}^L(\pi_0) &= \{ P : P_0(\theta - \delta) - \epsilon \leq P(\theta) \leq P_0(\theta + \delta) + \epsilon , \text{ for all } \theta \} , \end{aligned}$$

respectively, where $\pi_0 = dP_0/d\mu$. Although Levy neighborhood structure is an important structure in robust estimation theory (the neighborhoods are based on the Levy distance, which metrizes the weak topology) calculations of ρ^L and ρ^U have not been carried out for these neighborhoods. The Kolmogorov and Levy neighborhood classes are related to each other because $\Gamma_{\epsilon}^{KS} = \Gamma_{\epsilon, 0}^L$. Also, they

are related to the density and distribution bounded class, which has been extensively studied by Lavine (1991a,b). Given two specified bounding functions L and U , the density bounded class is given by

$$\Gamma_{L,U}^B = \left\{ P : L(\theta) \leq P(\theta) \leq U(\theta) , \text{ for all } \theta \right\} ,$$

Note that for the choice $L(\theta) = P_0(\theta - \delta) - \epsilon$ and $U(\theta) = P_0(\theta + \delta) + \epsilon$, the Levy class is equal to the distribution bounded class. This identification reduces the complexity of the density bounded class dramatically. For details on other classes see the review articles by Berger (1990) and Wasserman (1991).

Now the question arises, "How can one be a robust Bayesian while at the same time try to use a prior that is somewhat noninformative?" That is, given a class of priors Γ , what is the single prior $\pi \in \Gamma$ such that, among the class Γ , π is the least informative. At first thought this seems like a difficult constrained optimization problem; however, we are able to apply results from the classical parametric robustness literature, specifically applications of results of Huber (1973), to solve these problems and construct solutions. In the next section we will survey the relevant results and give some application to the construction of robust priors.

The remainder of the paper is organized as follows. In Section 2 we discuss methods for deriving robust default priors, and show connections with Bayesian decision theory. Section 3 gives details for many common classes of priors, and exhibits the form of a robust default prior and investigates some of the behavior. Section 4 has a concluding discussion.

2. Robust Default Priors

In this section we outline a strategy for constructing a robust default prior based on a Bayesian decision theoretic argument. That is, we look for priors that (approximately) maximize a Bayes risk. We also show that, in many common problems, the decision-theoretic answer can also be arrived at through classical robustness theory. These problems are ones whose information is constant in the parameter of interest, and include location and regression problems.

2.1 A Decision-Theoretic Approach

One possible method for generating a robust default prior is to find a prior that maximizes the Bayes risk of a procedure over some class, Γ , of priors. This is the Γ -minimax approach to robustness. (Berger (1985) gives a full review of the literature.) The main difficulty encountered in the Γ -minimax approach is that of solving the variational problem which produces the default prior. This problem is remedied by using an approximation of the Bayes risk due to Brown (1988) and Brown and Gajek (1990). Once this approximation is developed it yields a more straightforward Γ -minimax problem where the prior solves a second order linear differential equation restricted to the class Γ . Fortunately, solutions of these equations exist in the literature.

We will first develop the approximation to the Bayes risk. Let X be an observable random variable with probability density $f(x|\theta)$ relative to some σ -finite measure μ . Assume $\theta \in \Theta$, where $\Theta \subset \mathbb{R}$ is a possibly infinite interval. Suppose it is desired to estimate θ by $\delta \in \mathbb{R}$ under the loss $L(\theta, \delta) = m(\theta)(\delta - \theta)^2$, where $m > 0$ is a specified weight function. Let $R(\theta, \delta) = E_{\theta}L(\theta, \delta(X))$ denote the risk function of the non-randomized estimator δ , and let π be a prior probability density with respect to Lebesgue measure on Θ . For any estimator δ define the integrated risk and Bayes risk as

$$r(\pi, \delta) = \int R(\theta, \delta)\pi(d\theta) \quad \text{and} \quad r(\pi) = \inf_{\delta} r(\pi, \delta),$$

respectively. Let $V(\theta) = I(\theta)^{-1}$, the inverse of Fisher information. Brown and Gajek (1990) proved, under certain regularity conditions, that

$$r(\pi) \geq \frac{C^2}{C+D} \tag{2.1}$$

where

$$C = \int V(\theta)h(\theta)d\theta$$

$$D = \int \frac{[(Vh)'(\theta)]^2}{h(\theta)} d\theta$$

$$h(\theta) = m(\theta)\pi(\theta).$$

Simulation results indicate that this bound becomes sharp as the sample size tends to infinity. Note that this inequality is a function of the loss function, the prior and sampling density, three elements in any Bayesian decision theory problem. Also, the inequality is invariant under transformations on the

sample space since the sampling density enters only through Fisher information. If $f(x|\theta)$ is in the exponential family this inequality is related to Brown's (1971) heuristic method of proving admissibility.

We shall proceed as follows. First, we will assume that the inequality in (2.1) is sharp enough to be a good approximation of the Bayes risk. Next we will find the prior that maximizes (2.1) over the class of priors of interest. This prior will give us an approximation of the Γ -minimax Bayes risk. Since this approach is approximate, we will need to examine the default priors to check for sensibility. In examples in the next section, the default priors turn out to be quite reasonable. Most have the properties that the posterior behaves like the base prior in center of the distribution and like the likelihood in the tails. Also, by construction, the default priors will have the desired invariance property of a Jeffrey's prior.

As an alternative to working with the approximation (2.1), we could use an asymptotic approximation of $r(\pi, \delta)$. The article by Ghosh et al. (1982) gives such an approximation as well as a review of the research in this direction. Examination of the rather complex expansion of Ghosh et al. will show that the prior enters there in the form of C and D in (2.1). Hence, this alternate route will yield the same results as the maximization of (2.1). This is related to the approach proposed by Clarke and Barron (1990).

The inequality in (2.1) has interesting implications for conjugate priors. Suppose $m \equiv 1$ and $f(x|\theta)$ is an exponential family with expectation parameter θ and π is a conjugate prior. Then (and only then), subject to mild regularity conditions, (2.1) is actually an equality. This follows since both the Bayes rule and $(V\pi)'/\pi$ are linear. Hence the Bayes risk attains its lower bound and is therefore minimized. If $m \neq 1$, then equality holds if and only if π is proportional to a conjugate prior density. Again the Bayes risk attains its minimum. As we are searching for the prior that maximizes the Bayes risk to generate robust priors, the prior that minimizes the Bayes risk will lack robustness. The fact that conjugate priors lack robustness has been noted earlier by Berger (1984).

Example 2.1: To illustrate the proposed technique consider the problem of observing $X \sim \text{binomial}(n, p)$, and estimating $\theta = \log(p/1-p)$ under the normalized loss $L(\theta, \delta) = I(\theta)(\theta - \delta)^2$

where $I(\theta) = ne^{\theta}/(1+e^{\theta})^2$ is the Fisher information. Maximization of the bound in (2.1) requires finding the prior π which minimizes D . The solution to this calculus-of-variations problem follows by setting $\sqrt{\pi} = u$ and solving the Euler equation $(u'/I)' - \lambda u = 0$, where λ is a Lagrange multiplier used to incorporate the constraint that π is a proper density. The solution of the Euler equation, in terms of π , is $\pi(\theta) = 6e^{2\theta}/(1+e^{\theta})^4$. This corresponds to a prior, in terms of p , of $\pi(p) = 6p(1-p)$. Note that the Jeffrey's prior for this problem is $\pi_J(p) \propto (p(1-p))^{-\frac{1}{2}}$. The Jeffrey's prior does not take the weighted loss function into account where the other prior does. It is interesting to note the $\pi(p)$ puts the most mass near one-half whereas $\pi_J(p)$ puts more mass near zero and one. Also $\pi(p)$ is symmetric and concave whereas $\pi_J(p)$ is symmetric and convex.

2.2 Connections with Classical Robustness

We now investigate the parallels between the problem at hand and Huber's minimax variance theory. Let X_1, \dots, X_n be a random sample from a population with distribution function $F(x - \theta)$, where θ is an unknown location parameter. It is assumed that F is an unknown member of a specified convex vaguely compact neighborhood, \mathcal{F} , of a fixed baseline (or ideal) distribution G . Suppose $\{T_n\}$ is a sequence of estimators of θ such that $\sqrt{n}(T_n - \theta)$ converges in distribution to the normal law with mean equal to zero and variance $V(T, F)$. Let F_0 be the distribution which minimizes Fisher information, $I(F)$, over all $F \in \mathcal{F}$. If T_0 denotes an estimator which is asymptotically efficient at F_0 , that is

$$V(T_0, F_0) = I^{-1}(F_0),$$

then the minimum value of $\sup\{V(T, F) ; F \in \mathcal{F}\}$ is $I^{-1}(F_0)$, which is attained at T_0 . Thus the problem of finding the form of the minimax variance estimator corresponding to particular neighborhood models may be solved by finding the distribution, F_0 , with minimum Fisher information in \mathcal{F} .

The link between classical robustness and Bayesian robust is most easily seen in problems where the Fisher information is constant in θ (such as location and regression problems). The classical

statistician is looking for the distribution with minimum Fisher information, which is the distribution that minimizes

$$I(F) = \int \frac{(f'(x))^2}{f(x)} dx \quad (2.2)$$

over a specified class of distributions. If $m(\theta) = 1$ in (2.1), then minimization of (2.2) is exactly equivalent to maximization of (2.1), the goal of Bayesian robustness. Thus, in these cases, we can use the minimizers of (2.2) as our robust default priors.

In the classical robustness literature many classes of distributions have been studied, and distributions with minimum Fisher information have been explicitly found in many cases. Perhaps the neighborhood class which has been most thoroughly studied is the ϵ -contamination model, $\Gamma_\epsilon^C(G)$ where G is a fixed distribution symmetric about zero. Huber (1981) gives a complete solution for strongly unimodal G . When the unimodality of G is no longer assumed the results are no longer as complete. However, Collins and Wiens (1985) find the minimum information distribution F_0 for a quite general base distribution G .

Another neighborhood model which has received attention is the Kolmogorov neighborhood model $\Gamma_\epsilon^{KS}(G)$ where G is a fixed symmetric distribution. The most complete results pertain to the special case where G is the normal distribution. In this case the minimum information distribution may be found using results of Huber (1964) for $\epsilon < .0303$ and by Sacks and Ylvisacker (1972) for $\epsilon \geq .0303$. Calculations by Wiens (1986) may be used to obtain the minimum information distribution of non-normal G , subject to various regularity conditions. The Kolmogorov neighborhood may be extended to cover the important Levy neighborhood structure, $\Gamma_{\epsilon,\delta}^L(G)$. The minimum information distribution is given in Collins and Wiens (1989).

In the next Section we give several examples of prior distributions which maximize (2.1), hence are approximately Γ -minimax, for several classes of neighborhood priors. These will give examples of robust default priors.

When working in some class of priors one must distinguish between the central portion of the prior and the tail. Most Bayes procedures will be somewhat robust in the central portion of the prior, but it is rare that they are robust to changes in the tail areas. Furthermore, it is quite difficult to elicit

prior information in the tails of the prior. If the likelihood gives most of its weight to the tails, this is strong evidence that the prior has probably been misspecified. Rubin (1977) also shows that risk robustness tends to be much worse if the tails are influential.

3. Examples

3.1 The ϵ -contamination class

As the first example consider the ϵ -contamination class $\Gamma_\epsilon^c(\Phi)$, where Φ is the standard normal distribution. We can use the results of Huber (1973) to maximize (2.1) and hence find that the robust noninformative prior $\pi_c^*(\theta)$, for specified ϵ , is given by

$$\pi_c^*(\theta) = \begin{cases} \frac{1-\epsilon}{\sqrt{2\pi}} \exp\left\{-\frac{\theta^2}{2}\right\} & |\theta| \leq k \\ \frac{1-\epsilon}{\sqrt{2\pi}} \exp\{k^2/2 - k|\theta|\} & |\theta| > k \end{cases} \quad (3.1)$$

where k is a function of ϵ chosen so that $\pi_c^*(\theta)$ integrates to 1. (Table 4.5.1 of Huber (1981) gives values of k .) By varying ϵ , we can examine a variety of these priors and effectively tour the class. Note that as ϵ increases the prior information becomes more diffuse and hence the procedure becomes more robust. These priors are illustrated in Figure 3.1 for a number of values of ϵ .

To demonstrate the robustness of these priors we consider the example treated by Berger and Berliner (1986) and Wasserman (1989). Berger and Berliner investigated the robustness of the HPD region using a normal prior, and Wasserman looked at the robustness of the likelihood region. In each case, robustness was investigated using the class $\Gamma_\epsilon^c(\pi_0)$, where $\pi_0 = \text{normal}(\mu, \tau^2)$. The robustness of an interval $C(x)$ is quantified by calculating

$$\delta(x) = \sup_{\pi \in \Gamma_\epsilon^c} P(\theta \in C(x) | x) - \inf_{\pi \in \Gamma_\epsilon^c} P(\theta \in C(x) | x) . \quad (3.2)$$

Calculation of $\delta(x)$ is quite simple, as Huber (1973) provides a formula (also used by Wasserman, 1989).

We consider $x \sim \text{normal}(\theta, \sigma^2)$ and $\theta \sim \text{normal}(\mu, \tau^2)$, with $\sigma^2 = 1$, $\tau^2 = 2$ and $\mu = 0$. The

following small table shows the robustness behavior of the three intervals.

Table 1
Comparison of Credible Regions ($1 - \alpha = .9$, $\epsilon = .25$)

	$x = 0.5$		$x = 4.0$	
	Region	δ	Region	δ
HPD normal	(-1.01, 1.68)	0.246	(1.33, 4.01)	0.893
Likelihood	(-1.15, 2.15)	0.146	(2.36, 5.65)	0.756
HPD robust	(-1.03, 1.78)	0.224	(1.89, 5.18)	0.821

The HPD robust region is calculated by finding the $1 - \alpha$ HPD region of the posteriors arising from the prior (3.1). Notice that, in each case, the likelihood region is the most robust, the HPD normal region is least robust, while the HPD robust region is in between. Moreover, the HPD robust region displays exactly the behavior that we want. When the data and prior agree ($x \approx 0$) the HPD robust region acts like the HPD normal region. However, when the data and prior disagree (x far from 0) the HPD robust region behaves like the likelihood region (which is maximally robust, as shown by Wasserman, 1989). This behavior is illustrated in Figure 3.2, which shows the interval endpoints for a range of x . It can be seen that for x near 0 the HPD robust region behaves similar to the HPD normal region. However, as x gets far from 0, the HPD robust region gets closer to the likelihood region, while the HPD normal region goes off on its own.

It is interesting to note that the HPD robust region will not converge to the likelihood region as $x \rightarrow \infty$. This is because the tails of $\pi_c^*(\theta)$, which are exponential, will always maintain some influence. This point is also made by Hwang and Casella (1991), where it is shown that if $C_E(x)$ is the HPD region using an exponential (β) prior, then

$$\lim_{x \rightarrow \infty} P(\theta \in C_E(x) | x) = P\left(-Z_{\alpha/2} - \frac{1}{\beta} \leq Z \leq Z_{\alpha/2} - \frac{1}{\beta}\right)$$

where Z is a standard normal random variable and $Z_{\alpha/2}$ is the upper $\alpha/2$ cutoff. Upon reflection, such behavior is actually desirable if the prior is to be regarded as noninformative in any way. This is

because if we truly are to have no information about θ , then we should never allow the prior to be totally ignored. Even if $x \rightarrow \infty$, a noninformative prior should still leave doubts about the finiteness of θ . If not, then it is somehow informative.

Extensions to other ϵ -contamination models may also be developed, including multivariate and scale problems. When working on the multivariate problem one usually assumes that the distributions under consideration are from a spherically symmetric distribution; that is, having a density of the form $\pi(\theta) = \pi(|\theta|)$ where $\theta \in \mathbb{R}^d$ and $|\cdot|$ is the Euclidean norm. The problem may be transformed into a univariate problem by defining $\eta = |\theta|$. The density of η is given by $\nu(\eta) = dC_d |\eta|^{m-1} \pi(\eta)$, where C_d denotes the volume of the unit sphere in \mathbb{R}^d . When working on a scale problem the problem may also be transformed to a location problem by a logarithmic transformation.

3.2 The Kolmogorov-Smirnov Class

Now we turn our attention to the Kolmogorov-Smirnov neighborhood structure, $\Gamma_\epsilon^{\text{KS}}$. The class $\Gamma_\epsilon^{\text{KS}}$ has not received much attention in Bayesian robustness despite the fact that it has great intuitive appeal. Recall that the base prior density in $\Gamma_\epsilon^{\text{KS}}(\pi_0)$ is π_0 , which has distribution function equal to P_0 . Define $\xi(\theta) = -\pi'_0(\theta) / \pi_0(\theta)$ and $J(\theta) = 2\xi(\theta) - \xi_\theta^2$. If $\pi_0(\theta)$ is symmetric, $J(\theta)$ is strictly decreasing on $[0, \infty)$ and continuously differentiable on $(0, \infty)$, then by applying the results of Wiens (1986) and Collins and Wiens (1989) it may be shown that for some $\epsilon_0(P_0)$ the robust noninformative prior for this class, with $\epsilon \in (\epsilon_0(P_0), \frac{1}{2})$, is given by

$$\pi_{\text{KS}}^*(\theta) = \begin{cases} \frac{\pi_0(a) \cos^2(\lambda_1 \theta / 2)}{\cos^2(\lambda_1 a / 2)} & \text{on } [0, a] \\ \pi_0(b) \exp\{-\lambda(\theta - a)\} & \text{on } [a, \infty) \end{cases} \quad (3.3)$$

where $\lambda = \lambda_1 \tan(\lambda_1 a / 2)$. The constants are determined by (i) $P_{\text{KS}}^*(a) = P_0(a) - \epsilon$, (ii) $P_{\text{KS}}^*(\infty) = 1$. The constant a will also satisfy $-\pi_{\text{KS}}^{*1}(a) / \pi_{\text{KS}}^*(a) \leq \xi(a)$. In Figures 3.3 and 3.4 these priors are graphed for a selection of values of epsilon. Figure 3.3 has the normal as a root prior, while Figure 3.4 has the Cauchy as a root prior.

Based on the figures alone, it seems a difficult task to choose between (3.2) or (3.3) as the appropriate form for a noninformative prior. Since examination of the graphs gives little guidance, the experimenter must rely on the interpretation of the neighborhoods Γ_ϵ^c and Γ_ϵ^{KS} . Fortunately, this interpretation is rather straightforward, and should allow the experimenter to make a somewhat informative choice of the appropriate robust prior based on the form and degree of contamination.

The density given in (3.3) is the “large ϵ ” solution of Collins and Wiens (1989), the “small ϵ ” solution containing another term. However, our computations indicate that “small ϵ ” is very small, and for most values of ϵ that will be considered (say $\geq .05$), the “large ϵ ” solution is appropriate.

The Levy neighborhood $\Gamma_{\epsilon,\delta}^L(\pi_0)$ is a generalization of Γ_ϵ^{KS} neighborhood structure. In this class an additional location parameter, δ , indexes the family. The necessary calculation for the construction of the robust noninformative prior is given in Collins and Wiens (1989). The form of these priors is similar to $\pi_{KS}^*(\theta)$; however, now the location parameter, δ , is added.

3.3 The Quantile Class

Elicitation of prior information can sometimes involve specification of a finite number of features of the prior distribution. Suppose the parameter space Θ is partitioned into intervals $I_i = [\xi_i, \xi_{i+1}]$ $i=0, \dots, k$ where $-\infty = \xi_0 < \xi_1 < \dots < \xi_k < \xi_{k+1} = \infty$. Let the prior probability assigned to I_i equal p_i $i=1, \dots, k$. Now define the quantile class of prior distributions to be

$$\Gamma_Q = \left\{ \pi : p_i = \int_{I_i} \pi(d\theta) \right\} .$$

This is a natural class based on some elicitation mechanisms. See Berger (1990) for more on this class.

Consider estimating the mean of a normal $(\theta, 1)$ distribution under squared error loss and with prior information given by Γ_Q . Again we follow the proposed methodology and maximize the bound in (2.1) to find the default prior. The results of Huber (1974) may be extended to show that default prior distribution function P_0 is such that

- (i) $P_0(\xi_i) = p_i$ $i=0, 1, \dots, k+1$;
- (ii) P_0 is twice continuously differentiable;

- (iii) the density $\pi_0 = P'_0$ is strictly positive, except that it vanishes on those intervals $[\xi_i, \xi_{i+1}]$ for which $p_i = p_{i+1}$;
- (iv) on each interval $[\xi_i, \xi_{i+1}]$ the function $\sqrt{\pi_0''}/\sqrt{\pi_0}$ is constant $= \lambda_i$, that is

$$\sqrt{\pi_0(\theta)} = \begin{cases} a_i e^{\lambda_i \theta} + b_i e^{-\lambda_i \theta} & \text{if } \lambda_i > 0 \\ a_i \theta + b_i & \text{if } \lambda_i = 0 \\ a_i \cos|\lambda_i|\theta + b_i \sin|\lambda_i|\theta & \text{if } \lambda_i < 0. \end{cases}$$

In the case where θ is constrained to lie within two points, say $-L$ and L , one can define the class

$$\Gamma_L = \{\pi : |\theta| < L\}.$$

This is a very special case where the first and last constraint quantiles are set at L and $-L$ respectively, and no other quantiles are constrained. In this case it can be shown that the distribution that maximizes (2.1) equals

$$\pi_0(\theta) = \frac{1}{L} \cos^2\left(\frac{\pi}{2L} \theta\right) \quad |\theta| < L.$$

This is the solution of the calculus of variations problem solved by Bickel (1981) in the context of the bounded normal mean problem.

4. Discussion

In this article we considered the construction of default robust priors. These priors give the Bayesian a way to specify a particular class of priors and then, using an automatic mechanism, choose a noninformative member of that class from which to make inferences. Often in robust Bayesian analysis, one has an idea about the range of posterior decisions, but not about a particular prior with which to work. The present article takes a step toward remedying this.

The default robust prior has a number of advantages associated with it that make it a good candidate for a default prior. Firstly, it is a proper prior, which eliminates any possible incoherent inferences. Secondly, it has a meaningful interpretation in the context of a class of priors. Thirdly, it

is robust in the sense that it tends to favor the data over the prior when there is disagreement between the two. This was illustrated in Figure 3.2, and is also seen in Figure 4.1. There we see that as prior information and data become more discrepant, the robust posterior moves toward the data and flattens, mimicking the likelihood function. However, as is also illustrated in Figures 3.2 and 4.1, when the data and prior agree, the robust posterior behaves like a conjugate posterior, providing narrow confidence intervals and precise inferences. As this happens, however, the default prior retains its noninformative interpretation by always retaining a slight amount of influence.

The methodology that started these investigations, that of Jeffreys' outlined in the introduction, provides a method for constructing default priors when no partial prior information (in the form of prior classes) is available. It would then be possible to construct a Jeffreys' prior from a class of sampling distributions. In this case one would need to solve for the distribution in the class which has the minimum Fisher information, as is usually done in classical robustness theory. Once this distribution has been found one can compute the Fisher information of this distribution and take the square root of the Fisher information as the Jeffreys prior. Note in this discussion the class of interest is the class of sampling distributions. One could envision *super*-robustness methods where it is assumed that both the sampling distributions of the prior distribution are in classes. However, this would be extremely complicated to deal with analytically.

Lastly, one might be curious as to why we are using a frequentist measure (Bayes risk) to generate a Bayesian object (the prior). The Bayesian point of view is that one should condition on the data at hand then choose the optimal decision procedure with respect to that data only. Unfortunately, this is a somewhat local perspective on the construction of actions. To study the global properties of a procedure one must find optimal actions over a class of possible actions. In this case the goodness of the procedure should be measured by the Bayes risk. Note that if a procedure has a large Bayes risk, the repeated use of the procedure will give poor long run risk properties.

This is not to say that Bayes risk is necessarily a good measure for a particular data set. As it is a frequentist assessment, its worth can only be judged when averaged over all data sets. However, frequentist measures have a role in robust Bayesian statistics. Looking at the behavior of a Bayes rule

for a variety of data points (often extreme) may point out unsuspected and unacceptable features of the subjectively chosen prior. These extreme points may be the data points which one particularly wishes to guard against. Hence, considering the possibility of their existence will certainly increase the robustness of a procedure.

One goal of robust Bayesian statistics should be to develop Bayesian procedures which are automatically robust. This will help users of Bayesian methods who may not be sophisticated users, and may not be experienced in the use of the methods. Such users will surely not carry out a sensitivity analysis of the procedures. In such cases procedures which are automatically robust will be extremely helpful.

Bibliography

- Berger, J.O. (1984). The Robust Bayesian Viewpoint (with discussion). In *Robustness of Bayesian Analysis*, J. Kadane (Ed.). North Holland, Amsterdam.
- Berger, J.O. (1985). *Statistical Decision Theory* (2nd Edition). Springer-Verlag: NY.
- Berger, J.O. (1990). Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Plann. Inference* 25, 303-328.
- Berger, J.O. and Berliner, M. (1986). Robust Bayes and empirical Bayes analysis with ϵ -contaminated priors. *Ann. Statist.* 14, 461-486.
- Berger, J. O., and Bernardo, J. M. (1991). On the Development of the Reference Prior Method. To appear in the Proceedings of the Fourth Valencia International Meeting on Bayesian Statistics, Peniscola, Spain, April 1991.
- Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc.* 41, 113-147.
- Bickel, P.J. (1981). Minimax estimation of a mean of normal distribution when the parameter space is restricted. *Ann. Statist.* 9, 1301-1309.
- Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison Welsey, Reading, MA.
- Brown, L.D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary-value problems. *Ann. Math Statist.* 42, 855-903.
- Brown, L. D. (1988). The Differential Inequality of a Statistical Estimation Problem. In *Statistical Decision Theory & Related Topics III*, S.S. Gupta and J. Berger (Eds.). Springer-Verlag, New York.
- Brown, L.D., and Gajek, L. (1990). Information Inequalities for the Bayes Risk. *Ann. Statist.* 18, 1578-1594.
- Clarke, B. and Barron, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory* 36, 453-471.
- Clarke, B. and Wasserman, L. (1991). Noninformative priors and nuisance parameters. Carnegie Mellon Stat. Tech Report #535.
- Collins, J.R. and Wiens, D.P. (1985). Minimax variance M-estimators in ϵ -contamination models. *Ann. Statist.* 13, 1078-1096.
- Collins, J.R. and Wiens, D.P. (1989). Minimax properties of M-, R-, and L- estimators of location in Levy neighborhoods. *Ann Statist.* 17, 327-336.
- Ghosh, J.K., Sinhu, B.K. and Joshi, S.N. (1982). Expansions for posterior probability integrated Bayes risk. In *Statistical Decision Theory & Related Topics III*, S.S. Gupta and J. Berger (Eds.). Academic Press, New York.
- Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* 35, 73-101.

- Huber, P.J. (1973). The use of Choquet capacities in statistics. *Bull. Inst. Internat. Statist.* 45, 181-191.
- Huber, P.J. (1974). Fisher information & spline interpolation. *Ann. Statist.* 2, 1029-1034.
- Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.
- Hwang, J.T. and Casella, G. (1991). Frequentist Priors. Technical report BU-924-MA, Biometrics Unit, Cornell University.
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London Ser. A.* 186, 453-461.
- Jeffreys, H. (1961). *Theory of Probability* (3rd Edition). Oxford, Oxford University.
- Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.
- Lavine, M. (1991a). Sensitivity in Bayesian Statistics: The Prior and the Likelihood. *J. Amer. Statist. Assoc.* 86, 396-399.
- Lavine, M. (1991b). An Approach to Robust Bayesian Analysis for Multidimensional Parameter Spaces. *J. Amer. Statist. Assoc.* 86, 400-403.
- Polson, N. (1988). Bayesian perspectives on statistical modeling. Ph.D. dissertation, Department of Mathematics, University of Nottingham.
- Rubin, H. (1977). Robust Bayesian estimation. In *Statistical Decision Theory and Related Topics II*, S.S. Gupta and D.S. Moore (Eds.). Academic Press, New York.
- Sacks, J. and Ylvisaker, D. (1972). A note on Huber's robust estimation of a location parameter. *Ann. Math. Statist.* 43, 1068-1075.
- Sivaganesan, S. and Berger, J.O. (1989). Ranges of posterior measures for priors with unimodal contaminations. *Ann. Statist.* 17, 868-889.
- Wasserman, L. (1989). A robust Bayes interpretation of likelihood regions. *Ann. Statist.* 17, 1387-1393.
- Wasserman, L. (1991). Recent Methodological Advances in Robust Bayesian Inference. To appear in the Proceedings of the Fourth Valencia International Meeting on Bayesian Statistics, Peniscola, Spain, April 1991.
- Wiens, D. (1986). Minimax variance M-estimators of location in Kolmogorov neighborhoods. *Ann. Statist.* 13, 724-732.
- Zellner, A. (1987). *An Introduction to Bayesian Inference in Econometrics*. Krieger, Florida.

Figure 3.1: Prior distributions from equation (3.1), using a normal root prior, for $\epsilon = 0$ (solid line), .05 (long dashes), .25 (dots) and .4 (short dashes).

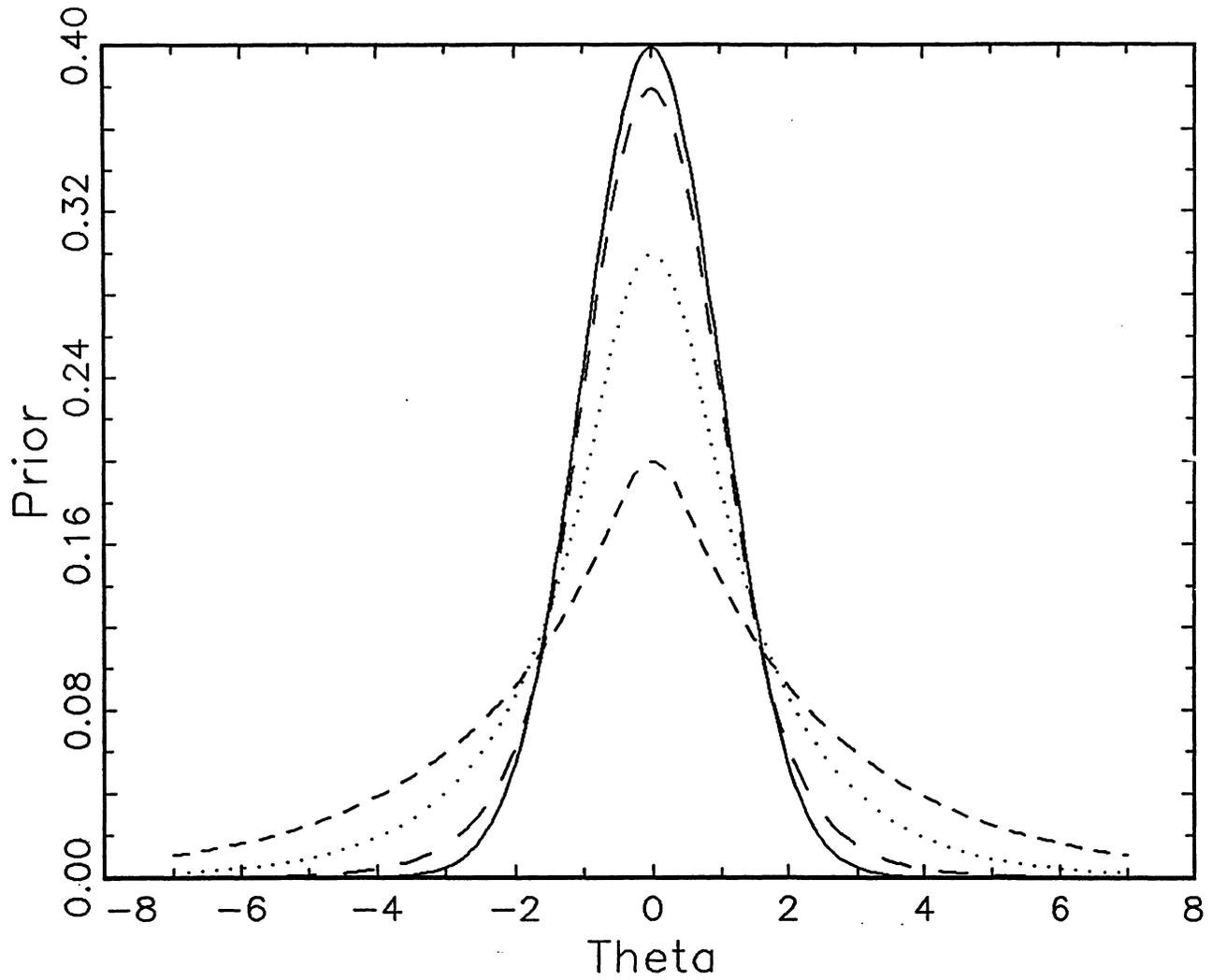


Figure 3.2: Upper and lower endpoints of the .9 credible interval from a normal prior (dotted line), the robust prior of (3.1) with $\epsilon = .25$ (solid line) and the likelihood interval (dashed line).

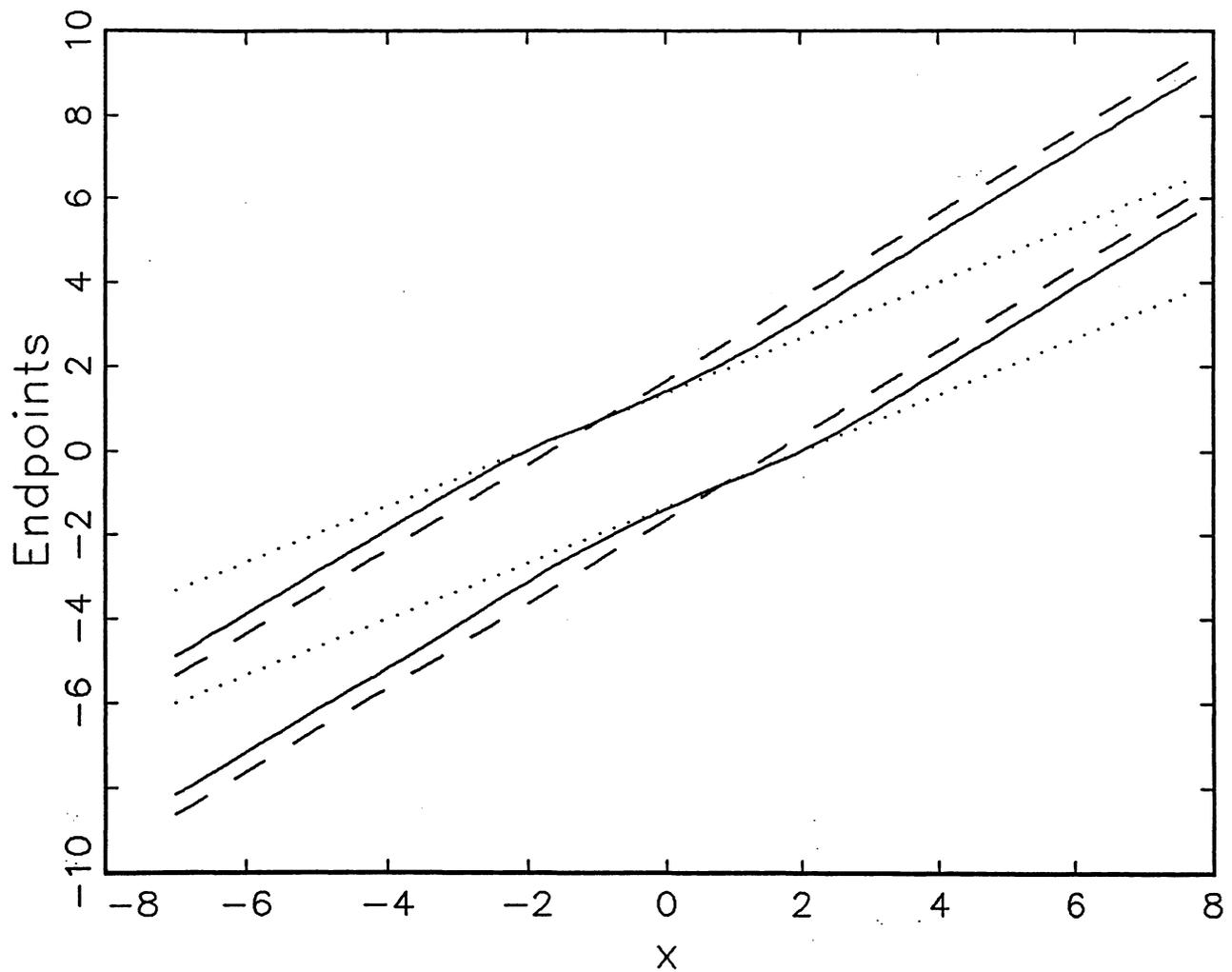


Figure 3.3: Prior distributions from equation (3.3), using a normal root prior, for $\epsilon = 0$ (solid line), .05 (long dashes), .25 (dots) and .4 (short dashes).

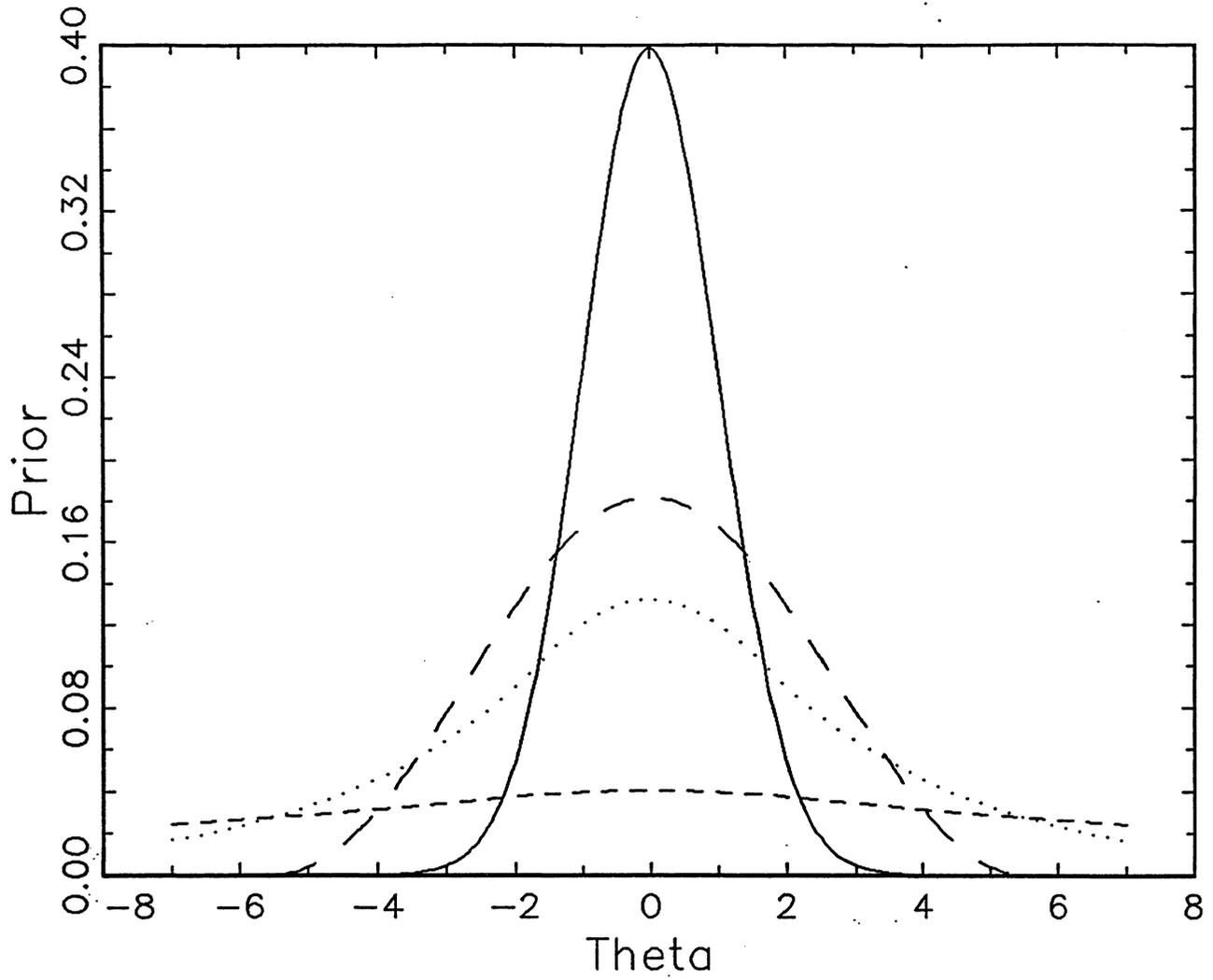


Figure 3.4: Prior distributions from equation (3.3) using a Cauchy root prior, for $\epsilon = 0$ (solid line), .05 (long dashes), .1 (dots) and .25 (short dashes).

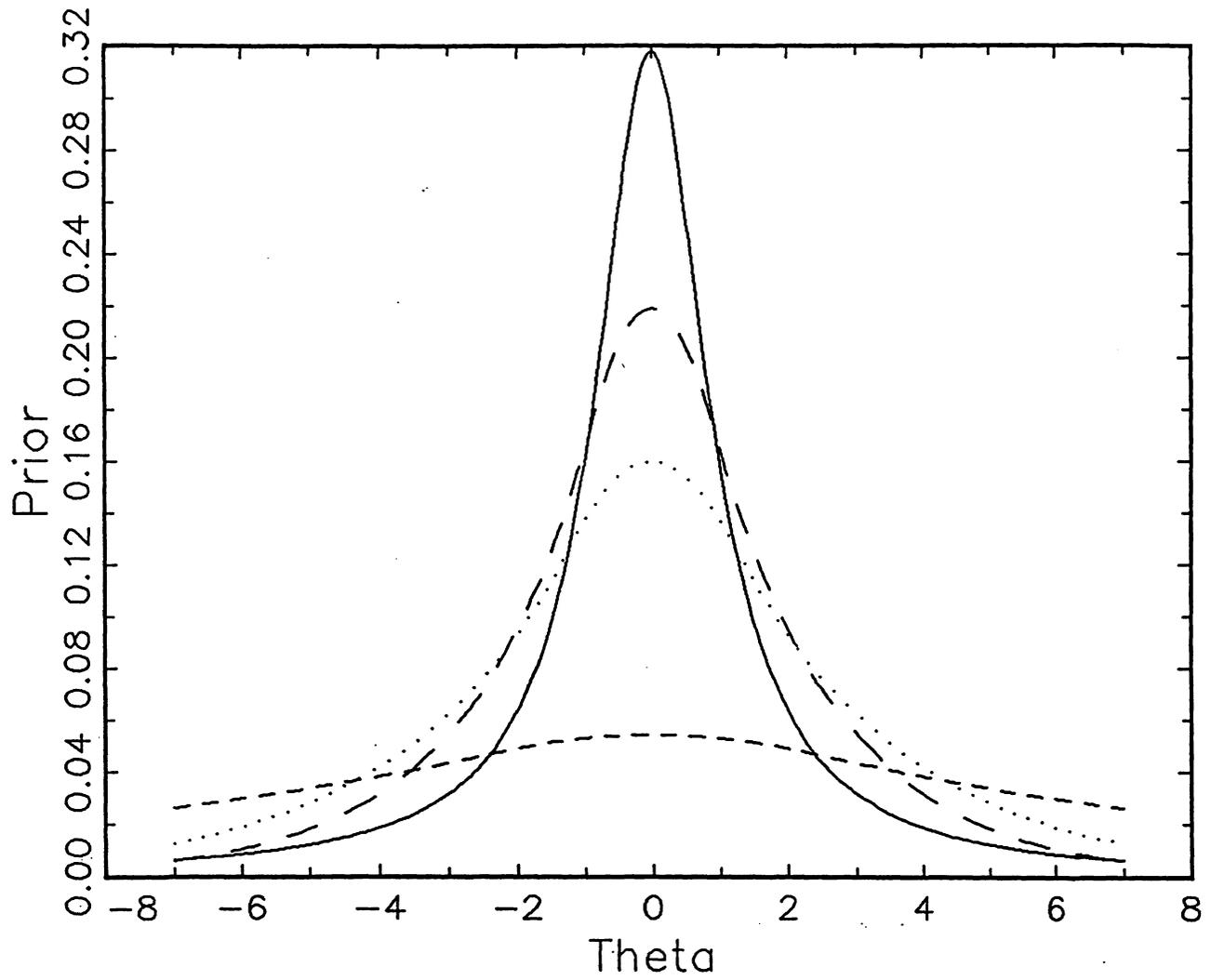


Figure 4.1: Posterior distributions, for $x = 0, 2.5, 5$, using a normal prior (dashed lines) and the robust prior of (3.1) with $\epsilon = .25$ (solid lines).

