Test (1994)
Vol. **3**, *No.* 1, *pp.* 5–124

An Overview of Robust Bayesian Analysis*

JAMES O. BERGER
Department of Statistics, Purdue University
West Lafayette, IN 47907-1399, U.S.A.

[Read before the Spanish Statistical Society at a meeting organized by the Universidad Autónoma de Madrid on Friday, November 19, 1993]

SUMMARY

Robust Bayesian analysis is the study of the sensitivity of Bayesian answers to uncertain inputs. This paper seeks to provide an overview of the subject, one that is accessible to statisticians outside the field. Recent developments in the area are also reviewed, though with very uneven emphasis. The topics to be covered are as follows:

Received November 93; Revised February 94.
* Research supported by the National Science Foundation, Grants DMS–8923071 and DMS 93-03556.

1. INTRODUCTION

1.1. *Motivation*

Robust Bayesian analysis is the study of the sensitivity of Bayesian answers to uncertain inputs. These uncertain inputs are typically the model, prior distribution, or utility function, or some combination thereof. Informal or adhoc sensitivity studies have long been a part of applied Bayesian analysis (cf. Box, 1980), but recent years have seen an explosion of interest and literature on the subject. There are several reasons for this interest:

Foundational Motivation: There is a common perception that foundational arguments lead to subjective Bayesian analysis as the only coherent method of behavior. Non-Bayesians often recognize this, but feel that the subjective Bayesian approach is too difficult to implement, and hence they ignore the foundational arguments. Both sides are partly right. Subjective Bayesian analysis is, indeed, the only coherent mode of behavior, *but only if it is assumed that one can make arbitrarily fine discriminations in judgment about unknowns and utilities.* In reality, it is very difficult to discriminate between, say, 0.10 and 0.15 as the subjective probability, $P(E)$, to assign to an event E, much less to discriminate between 0.10 and 0.100001. Yet standard Bayesian axiomatics assumes that the latter can (and will) be done. Non-Bayesians intuitively reject the possibility of this, and hence reject subjective Bayesian theory.

It is less well known that *realistic* foundational systems exist, based on axiomatics of behavior which acknowledge that arbitrarily fine discrimination is impossible. For instance, such systems allow the possibility that $P(E)$ can only be assigned the range of values from 0.08 to 0.13; reasons for such limitations range from possible psychological limitations to constraints on time for elicitation. The conclusion of these

foundational systems is that a type of *robust Bayesian analysis* is the coherent mode of behavior. Roughly, coherent behavior corresponds to having *classes* of models, priors, and utilities, which yield a range of possible Bayesian answers (corresponding to the answers obtained through combination of all model-prior-utility triples from the classes). If this range of answers is too large, the question of interest may not, of course, be settled, but that is only realistic: if the inputs are too uncertain, one cannot expect certain outputs. Indeed, if one were to perform ordinary subjective Bayesian analysis without checking for robustness, one could be seriously misled as to the accuracy of the conclusion.

Extensive developments of such foundational systems can be found in Walley (1991), Ríos Insúa (1990, 1992) and Ríos Insúa and Martín (1994); see also Girón and Ríos (1980) and Kouznetsov (1991). I. J. Good (cf., Good, 1983a) was the first to extensively discuss these issues. Other earlier references can be found in Berger (1984, 1985) and in Walley (1991); this latter work is particularly to be recommended for its deep and scholarly study of the foundations of imprecision and robustness. Recent developments in some of the interesting theoretical aspects of the foundations can be found in Wasserman and Kadane (1990, 1992b) and Wasserman and Seidenfeld (1994).

Practical Bayesian Motivation: Above, we alluded to the difficulty of subjective elicitation. It is so difficult that, in practice, it is rarely done. Instead, noninformative priors or other approximations (e.g., BIC in model selection) are typically used. The chief difficulties in elicitation are (i) knowing the degree of accuracy in elicitation that is necessary; (ii) knowing what to elicit. Robust Bayesian analysis can provide the tools to answer both questions.

As an example of (i), one might be able to quickly determine that $0.05 \leq P(E) \leq 0.15$, but then wonder if more accurate specification is needed. Robust Bayesian methods can operate with such partial specifications, allowing computation of the corresponding range of Bayesian answers. If this range of answers is small enough to provide an answer to the question of interest, then further elicitation is unnecessary. If, however, the range is too large to provide a clear answer, then one must attempt finer elicitation (or obtain more data or otherwise strengthen the information base).

Knowing what to elicit is even more crucial, especially in higher dimensional problems where it is completely infeasible to elicit everything that is possibly relevant. Suppose, for instance, that one believes in a 10-dimensional normal model, but that the mean vector and covariance matrix are unknown. Then there are 65 unknown parameters, and accurate elicitation of a 65-dimensional distribution is impossible (unless one is willing to introduce structure that effectively greatly reduces the number of parameters). But many of these parameters may be accurately determined by the data, or the question of interest may not depend on accurately knowing many of the parameters. In fact, there may only be a few crucial quantities that need to be elicited. Robust Bayesian techniques can help to identify these quantities.

Acceptance of Bayesian Analysis: Rightly or wrongly, the majority of the statistical world resists use of Bayesian methods. The most often vocalized reason is fear of using a subjective prior, because of a number of perceived dangers. While we do not view this fear as being particularly reasonable (assumptions made in other parts of the analysis are usually much more influential and questionable), we recognize its existence. Robust Bayesian methods, which can operate with a wide class of prior distributions (reflecting either the elicitor's uncertainty in the chosen prior or a range of prior opinions of different individuals), seems to be an effective way to eliminate this fear.

Non-Bayesian Motivation: Many classical procedures work well in practice, but some standard procedures are simply illogical. Robust Bayesian analysis can be used to determine which procedures are clearly bad. Consider, for instance, the following example:

Example 1.. A series of clinical trials is performed, with trial i testing drug D_i versus a placebo. Each clinical trial is to be analyzed separately, but all can be modelled as standard normal tests of $H_0: \theta_i = 0$ versus $H_1: \theta_i \neq 0$, where θ_i is the mean effect of D_i minus the mean effect of the placebo. Suppose we know, from past experience, that about 1/2 of the drugs that are tested will end up being ineffective; i.e., will have $\theta_i = 0$. (This assumption is not essential; it merely provides a mental reference for the ensuing understanding.)

We will focus on the meaning of P-values that arise in this sequence of tests. Table 1 presents the first twelve such P-values. Consider, first,

those tests for which the P-value is approximately 0.05; D_2 and D_8 are examples. A crucial question is: among the drugs for which the P-value of the test is approximately 0.05, what fraction are actually ineffective (i.e., correspond to true H_0)? Likewise, consider those D_i for which the P-value is approximately 0.01 (D_5 and D_{10} are examples) and ask: what fraction are actually ineffective?

DRUG	D_1	D_2	D_3	D_4	D_5	D_6
P-Value	0.41	**0.04**	0.32	0.94	**0.01**	0.28
DRUG	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}
P-Value	0.11	**0.05**	0.65	**0.009**	0.09	0.66

Table 1. *P-values resulting from the first twelve clinical trials, testing H_0: D_i has no effect vs. H_1: D_i has an effect.*

The answers to these questions are, of course, indeterminate. They depend on the actual sequence of $\{\theta_i\}$ that arises. However, using robust Bayesian techniques one can find lower bounds on the answers that are valid for *any* sequence $\{\theta_i\}$. These can be computed as in Berger and Sellke (1987, Section 4.3), and are 0.24 for the first question and 0.07 for the second.

This is quite startling, since most statistical users would believe that, when the P-value is 0.05, H_0 is very likely to be wrong and, when the P-value is 0.01, H_0 is almost certain to be wrong. The actual truth is very different. And since 0.24 and 0.07 are lower bounds that are actually difficult to attain, the fractions of true H_0 encountered in practice would typically be much larger (on the order of 50% and 15%, respectively). Thus we have a situation where the standard classical method, or at least its standard interpretation, is highly misleading.

There is also a more subtle potential use of robust Bayesian analysis within frequentist statistics, arising from the fact that "optimal" frequentist procedures are virtually always Bayes (or generalized Bayes) procedures. Note that this, by itself, is not a compelling reason for a frequentist to adopt the Bayesian viewpoint, because the prior distribution that is used to develop the frequentist procedure can be considered merely to be a mathematical artifact, with no inherent meaning. (Using

a prior to develop the procedure but ignoring its Bayesian implications may appear to be rather myopic, but it is not illogical to do so from the frequentist perspective.)

When the statistical problem becomes even moderately difficult, however, in the sense that the frequentist accuracy or performance measure is not constant over the unknown parameters, it can become very difficult for the frequentist to recommend a particular procedure. A very appealing possibility is to then use the Bayesian perspective to choose the prior, and to consider the resulting Bayes procedure from the frequentist perspective. If the Bayesian procedure is a robust Bayesian procedure, there are numerous indications that it will have excellent frequentist properties. See Berger (1984, 1985), DasGupta and Studden (1988a, 1989), Berger and Robert (1990), Robert (1992), Mukhopadhyay and DasGupta (1993) and DasGupta and Mukhopadhyay (1994), for such arguments in general; here we content ourselves with an interesting example, from Berger, Brown, and Wolpert (1994).

Example 2.. Suppose $X_1, X_2, \ldots$ are i.i.d. $\mathcal{N}(\theta, 1)$ and that it is desired to test H_0: $\theta = -1$ versus H_1: $\theta = 1$. If the hypotheses have equal prior probability, the Bayesian inference, after stopping experimentation at sample size N, will be to (i) compute the posterior probability of H_0, which can be seen to be (defining $\overline{x}_N = \sum_{i=1}^{N} x_i/N$)

$$\begin{aligned} P(H_0|x_1, \ldots, x_N) &= 1/[1 + \exp\{2N\overline{x}_N\}] \\ &= 1 - P(H_1|x_1, \ldots, x_N); \end{aligned}$$

(ii) choose the hypothesis with larger posterior probability (assuming the utility structure is symmetric); and (iii) report the posterior probability of the rejected hypothesis as the error probability.

There would seem to be no problem here for a frequentist: simply choose the most powerful Neyman-Pearson test with, say, equal error probabilities. But the situation is not so clear. First, this could have been a sequential experiment (e.g., the SPRT) with N being the stopping time, and stopping rules can have a dramatic effect on classical testing. Second, even if N is fixed, the most powerful test has strange properties. For instance, if $N = 4$, the frequentist error probabilities corresponding to the test "reject if $\overline{x}_4 \geq 0$ and accept otherwise" would be 0.025, and

this would be the reported error for either $\overline{x}_4 = 0$ or $\overline{x}_4 = 1.5$; this is very strange because $\overline{x}_4 = 0$ would seem to indicate no evidence for H_0 versus H_1 (since 0 is equidistant between $\theta = -1$ and $\theta = +1$), while $\overline{x}_4 = 1.5$ would indicate overwhelming evidence for H_1 (it being 5 standard errors from H_0).

When standard frequentist procedures behave unnaturally, frequentists turn to conditional frequentist procedures (cf., Kiefer, 1977). But in this problem there are a plethora of possible conditional frequentist tests, and it is unclear how one should be chosen. Also, the interpretation of conditional tests and conditional error probabilities can be very difficult for practitioners.

Now look back at the simple Bayes test described at the beginning of the example. It is easy to use; it does not depend on the stopping rule in a sequential setting; it avoids the intuitive objections to the Neyman-Pearson test (when $\overline{x} = 0$, one reports $P(H_0|x_1, \ldots, x_4) = 0.5$ and, when $\overline{x} = 1.5$, one reports $P(H_0|x_1, \ldots, x_4) \cong 6 \times 10^{-6}$); and it has a simple interpretation. This test would be delightful for a frequentist, if only it could be given a frequentist interpretation. But it can! Indeed, in Berger, Brown, and Wolpert (1994), it is shown that this is a valid conditional frequentist test, with conditional error probabilities being given by the posterior probabilities.

Because this situation involved only the testing of simple hypotheses, the choice of the prior was not particularly relevant, and hence Bayesian robustness was not a factor. In testing of composite hypotheses, however, it appears to be necessary to utilize robust Bayesian procedures if one seeks to have sensible tests with a conditional frequentist interpretation. This work is currently under development.

1.2. *Preview*

First, this is not exactly a review paper. More formal and thorough reviews can be found in Berger (1984, 1990, and, to a lesser extent, 1985) and in Wasserman (1992b). We will make a somewhat uneven effort to indicate the literature that has arisen since these review papers, but there will be only moderate discussion of this literature.

The primary goals of the paper are, instead, to provide a fairly accessible discussion of Bayesian robustness for statisticians not in the

field, and to summarize our views on some of the important issues and considerations in Bayesian robustness.

Section 2 considers the idea of choosing models and priors that are inherently robust. The idea is that it is perhaps easier to build robustness into the analysis at the beginning, than to attempt to verify robustness at the end.

Section 3 briefly discusses diagnostics, influence, and sensitivity. Our review of this material is admittedly too brief; it is deserving of much more coverage.

Section 4 spends a perhaps inordinate amount of space on the issue of global robustness: finding the range of Bayesian answers as the Bayesian inputs vary. This area has experienced by far the most active development in recent years.

Uses of computing in Bayesian robustness are discussed in Section 5; perhaps of particular interest is the possibility of using Bayesian robustness to enhance interactive elicitation. Section 6 summarizes some thoughts about the future.

There is one major aspect of Bayesian robustness that is essentially ignored in the paper, namely robustness with respect to the utility or loss function. This mirrors a similar avoidance of the issue in the literature. There are, perhaps, three reasons for this avoidance. First, formal statistical decision analysis is not often done in practice (at least by statisticians), because of the extreme difficulty in eliciting utilities. (But perhaps Bayesian robustness is, for this reason, even more compelling in decision problems.) Second, modelling uncertainty in utility functions is often more awkward, and more case-specific, than modelling uncertainty in distributions. Finally, robust Bayesian analysis involving utility functions can be technically more difficult than other types of Bayesian robustness. A few references to robustness involving the utility are Kadane and Chuang (1978), Moskowitz (1992), Ríos Insúa (1990, 1992), Ríos Insúa and French (1991), Drummey (1991), Basu and DasGupta (1992), and Ríos Insúa and Martín (1994).

We will also ignore several other important robustness issues for reasons of space. One such is the issue of model selection and Bayesian prediction in the face of model uncertainty. For discussion and references see Draper (1992), Kass and Raftery (1992), Berger and Pericchi (1993), and Pericchi and Pérez (1994).

We also will not discuss the huge literature on gamma minimax estimation, which is the frequentist version of robust Bayesian analysis. Extensive discussion of this approach, and its relationship to the posterior robust Bayesian approach discussed here, can be found in Berger (1984, 1985), which also contain numerous references. Recent references include Ickstadt (1992), Vidakovic (1992), and Eichenauer-Herrmann and Ickstadt (1993).

Finally, there have been numerous Bayesian robustness investigations in particular problems or situations. A partial list of recent works is Kass and Greenhouse (1989), Lavine and Wasserman (1992), Berger and Chen (1993), Goldstein and Wooff (1994), and O'Hagan (1994).

1.3. *Notation*

The entire data set will be denoted by X, which will be assumed to arise from a density $f(x|\theta_f)$ (w.r.t. a fixed dominating measure), with θ_f denoting unknown parameters of f. A prior density for θ_f will be denoted by $\pi(\theta_f)$; we will assume that this is a density w.r.t. Lebesgue measure, for notational convenience.

Key Bayesian quantities are

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

which is the marginal or predictive density of X, and

$$\pi(\theta_f|x, f) = f(x|\theta_f)\pi(\theta_f)/m(x|\pi, f)$$

which, assuming the denominator is nonzero, is the posterior density of θ_f. We explicitly retain f in the notation to allow for discussion of robustness w.r.t. f. For analyses in which f is fixed, we will simply drop f from the notation. Finally, we define $\psi(\pi, f)$ (suppressing x) to be the posterior (or other) quantity of interest. Typically,

$$\psi(\pi, f) = \int h(\theta_f)\pi(\theta_f|x)d\theta_f = \frac{\int h(\theta_f)f(x|\theta_f)\pi(\theta_f)d\theta_f}{\int f(x|\theta_f)\pi(\theta_f)d\theta_f}.$$

For instance, $h(\theta_f) = \theta_f$ yields the posterior mean and $h(\theta_f) = 1_C(\theta_f)$ (the indicator function on the set C) yields the posterior probability of C. Other types of $\psi(\pi, f)$ are, however, possible: for instance, posterior quantiles or $m(x|\pi, f)$ itself.

2. DEVELOPMENT OF INHERENTLY ROBUST PROCEDURES

2.1. *Introduction*

Choices of the functional form of the statistical model or prior distribution are frequently quite arbitrary.

Example 3.. Suppose $X_1, \ldots, X_n$ are felt to be i.i.d. observations from the measurement error model $X_i = \mu + \varepsilon_i$, where the measurement errors, ε_i, have a symmetric, unimodal distribution with unknown variance σ^2. Very little is known about σ^2, but the unknown μ is felt, apriori, to be $0 \pm \sqrt{2.19}$; we will interpret this to mean that 0 and $\sqrt{2.19}$ are the prior mean and prior standard error, respectively.

The "standard" analysis here would be to choose $f(x_i|\mu, \sigma)$ to be $\mathcal{N}(\mu, \sigma^2)$, and to choose $\pi(\mu, \sigma) = \frac{1}{\sigma} \cdot \pi_1(\mu)$, where $\pi_1(\mu)$ is $\mathcal{N}(0, 2.19)$. (The unknown σ is here given the usual noninformative prior. Sometimes $\pi_1(\mu|\sigma) = \mathcal{N}(0, (2.19)\sigma^2)$ is used in place of $\pi_1(\mu)$.)

While various arguments can be given for such standard choices, the fact remains that they are often quite arbitrary. Furthermore, standard choices such as these often result in models from the exponential family and conjugate priors, both of which are known to be nonrobust in various ways: models in the exponential family are very sensitive to outliers in the data, and conjugate priors can have a pronounced effect on the answers even if the data is in conflict with the specified prior information. (This last is not always bad, but most users prefer to "trust the data" in such situations.) Further discussion and other references can be found in Berger (1984, 1985).

2.2. USE OF FLAT-TAILED DISTRIBUTIONS

Considerable evidence has accumulated that use of distributions with flat tails tends to be much more robust than use of standard choices, such as those discussed in Section 2.1. See Dawid (1973), Box and Tiao (1973), Berger (1984, 1985), O'Hagan (1988, 1990), Angers and Berger (1991), Fan and Berger (1992), Geweke (1992), and Lucas (1992).

Example 3 (continued).. Suppose, instead, that $f(x_i|\mu, \sigma)$ is chosen to be a t-distribution with, say, 4 degrees of freedom. One might actually want to introduce the degrees of freedom, α, as an unknown parameter (see Chib, Osiewalski, and Steel, 1991, for a recent study), but that is

more a model elaboration than a model robustification. Also, $\pi_1(\mu)$ could be chosen to be Cauchy(0, 1) (which matches the quartiles of a $\mathcal{N}(0, 2.19)$).

This analysis would be robust in two respects. First, if there are outliers in the data, they will automatically be filtered out of the analysis. Second, if the prior information about μ turns out to be very inaccurate (due, say, to the all-too-common problem that elicitors typically choose prior variances that are much smaller than their real uncertainties), then it is automatically discounted in the analysis. Neither of these robust behaviors occurs with the standard analysis.

The price to be paid for utilization of inherently robust procedures is computational; closed form calculation is no longer possible. Today, however, computational schemes exist for performing robust Bayes computations routinely. For instance, any situation involving normal models and normal priors that is to be analyzed with Gibbs sampling can, instead, be done with t-distributions (cf, Verdinelli and Wasserman, 1991; Geweke, 1992; and Datta and Lahiri, 1992).

Example 3 (continued).. Saying that $X_i \sim \mathcal{T}_4(\mu, \sigma^2)$ is equivalent to saying that, given τ_i, $X_i \sim \mathcal{N}(\mu, \sigma^2/\tau_i)$, where $\tau_i \sim$ Gamma $(2, \frac{1}{2})$. Likewise, saying that $\mu \sim \mathcal{C}(0, 1)$ is equivalent to saying that, given τ_0, $\mu \sim \mathcal{N}(0, 1/\tau_0)$, where $\tau_0 \sim$ Gamma $(\frac{1}{2}, \frac{1}{2})$. By introducing the τ_i as random unknowns, it is possible to write the conditional posterior distributions of each unknown, given the others, as simple normal, gamma, or inverse gamma distributions, allowing for straightforward Gibbs sampling.

While the above example indicates that, in principle, robustification is always possible for normal models, the computational cost may still be severe. For instance, the original two unknowns, (μ, σ), above are replaced by the unknowns $(\mu, \sigma, \tau_0, \tau_1, \ldots, \tau_n)$. When n is large, the Gibbs sampling simulation can be very expensive.

Introducing such robustifications in hierarchical Bayes scenarios is often much more cost effective. For instance, replacing the standard hierarchical Bayes model, $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$ and $\theta_i \sim \mathcal{N}(\mu, A)$, for $i = 1, \ldots, p$, by the model $X_i \sim \mathcal{T}_4(\theta_i, \sigma^2)$ and $\theta_i \sim \mathcal{C}(\mu, A)$, and introducing τ_i to convert the latter model to a normal and inverse gamma model, would only increase the number of parameters from $p + 3$ to $3p + 3$. A factor of 3 in Gibbs sampling is not severe.

Note that analytic methods for doing computations in certain of these hierarchical models exist. See Spiegelhalter (1985), Fan and Berger (1990, 1992), Angers and Berger (1991), Angers (1992), and Angers, MacGibbon, and Wang (1992).

A somewhat more modest type of robust prior has long existed in multivariate problems. Suppose $\boldsymbol{X} = (X_1, \ldots, X_p)^t \sim \mathcal{N}(\boldsymbol{\theta}, \mathfrak{T})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^t$ is unknown and $\mathfrak{T}$ is given. The conjugate prior for $\boldsymbol{\theta}$ is a $\mathcal{N}(\boldsymbol{\mu}, A)$ prior, for which the posterior mean is $\boldsymbol{\delta}^{\pi}(\boldsymbol{x}) = \boldsymbol{x} - \mathfrak{T}\,(\mathfrak{T} + A)^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$. A variety of arguments suggest that it is more robust to use "shrinkage" versions of $\boldsymbol{\delta}^{\pi}$; among the many approaches to developing such are minimax theory, ridge regression, empirical Bayes analysis, and BLUP theory. But the best robust alternatives to $\boldsymbol{\delta}^{\pi}$ are, arguably, the robust Bayes alternatives, in which the $\mathcal{N}(\boldsymbol{\mu}, \mathfrak{T})$ prior is replaced by a $\mathcal{T}_{\alpha}(\boldsymbol{\mu}, \mathfrak{T})$ prior (for, say, $\alpha = 4$) or something similar. Extensive discussion of one such alternative, that is particularly easy to work with, can be found in Berger (1985, Section 4.7.10), which also has many references. See, also, Zellner (1976) and Berger and Robert (1990).

The reasons this latter type of robustness is more limited than the earlier type are: (i) model robustness is not involved; (ii) one achieves robustness to prior misspecification only in the overall sense that if the prior and data clash, the entire prior is discounted. The earlier discussed use of independent t-distributions would allow discounting of only part of the prior.

2.3. *Use of Noninformative and Partially Informative Priors*

That noninformative priors often yield automatically robust answers was recognized as early as Laplace (1812). Indeed, his development of the Central Limit Theorem was essentially a demonstration that, for large sample sizes, the posterior distribution of an unknown model parameter θ is essentially the same asymptotic normal distribution for any nonzero prior density. (See Ghosh, Ghosal, and Samanta, 1994, for recent developments and references.) For this and various intuitive reasons, Laplace felt that simply using $\pi(\theta) = 1$, as the prior, would give quite robust answers.

Another sense in which use of $\pi(\theta) = 1$ is robust was formalized by L. J. Savage as the theory of *precise measurement*. See Edwards, Lind-

man, and Savage (1963), Moreno and Pericchi (1993b), Mukhopadhyay and DasGupta (1993), and Pericchi and Pérez (1994).

Modern noninformative prior theory takes this one step further. Noninformative priors are specifically constructed so as to have minimal influence, in some sense, on the answer. The sense in which this engenders robustness is rather weak: it seems to ensure that the Bayesian analysis, for small or moderate samples, is not affected by unintended properties of the prior. For instance, in Example 3 we saw that standard conjugate choices of the tail of the prior (or likelihood) could have a dramatic unintended effect on the posterior. In multivariate situations, the potential for such unintended effects is particularly large, since few features of the prior will actually be subjectively elicited and there is a substantial possibility that mistakes can "accumulate" across the dimensions.

The two most extensively developed noninformative prior theories of this type are the *reference prior* theory (cf., Bernardo, 1979; Berger and Bernardo, 1992; and Bernardo and Smith, 1994), and the *maximum entropy* approach (cf., Jaynes, 1983, and Fougere, 1990). Other approaches are discussed in the excellent review paper Kass and Wasserman (1993).

Partially informative priors are also of considerable interest from the robustness perspective. These priors are of two types. The first type is for use in problems where there are, say, "parameters of interest" and "nuisance parameters." The parameters of interest are basically given subjectively elicited prior distributions, perhaps with associated robustness investigations being performed, while the nuisance parameters are given noninformative priors. The idea here is that elicitation of priors for nuisance parameters is likely to be difficult and a less valuable use of available elicitation time, and that attempting formal robustness studies with respect to the nuisance parameters is likely to be ineffective. For examples and further discussion of this general notion, see Liseo (1993) and Berger and Mortera (1994).

The second type of commonly used partially informative prior is a constrained maximum entropy prior. The idea here is that one specifies certain features of the prior (or model) and then chooses that prior (or model) which maximizes entropy subject to the specified constraints. The hope is that the resulting prior (or model) will have the specified features, but be robust (in the noninformative prior sense) with respect to un-

specified features. For further discussion see Jaynes (1983) and Fougere (1990). Somewhat different approaches are considered in Casella and Wells (1991) and Bernardo and Smith (1994).

2.4. *Nonparametric and Infinite Parametric Bayes Procedures*

Bayesian nonparametrics can be considered to be an approach to automatic robustness with respect to model choice. A large nonparametric class of models is entertained and given a prior distribution so that, hopefully, the data will cause the analysis to automatically adapt to the true model.

The majority of the work on Bayesian nonparametrics has involved use of the Dirichlet process prior on the space of all probability distributions. Recent references include Brunner and Lo (1989), Lo and Weng (1989), Gasparini (1990), Ferguson, Phadia, and Tiwari (1992), Tamura (1992), and Doss (1994).

Dirichlet process priors have a number of potentially unappealing features, such as the fact that they give probability one to the set of discrete probability measures. Hence there has been considerable effort expended to develop priors that are supported on continuous densities, such as Gaussian process priors. An example of such a prior, for the space of continuous densities, $f(t)$, on $[0, T]$, is to let

$$f(t) = \exp\{X(t)\} / \int_0^T \exp\{X(t)\} dt,$$

where $X(t)$ is the sample path of a Gaussian process. This and other such priors are studied in Leonard (1978), Lenk (1988), Angers and Delampady (1992), and Zidek and Weerahandi (1992). Computations with such priors are more difficult than with Dirichlet process priors, but the recent new Bayesian computational tools should enhance the utilization of these alternative nonparametric priors.

In regards to Gaussian process priors, the Bayesian interpretation of smoothing splines should also be mentioned. Smoothing splines can be developed as Bayesian function estimates for certain Gaussian process priors on derivates of functions. This interpretation has been important in deriving accuracy estimates for smoothing splines (utilizing the associated posterior covariance function). See Kohn and Ansley (1988), Wahba (1990) and Gu and Wahba (1993). There is considerable

promise in further exploiting this relationship for higher dimensional smoothing splines, especially if structural assumptions on the function are made, such us

$$f(x_1, \ldots, x_p) = \sum_{i=1}^{p} f_i(x_i).$$

Finally, other very promising nonparametric Bayes approaches are being developed, such as Lavine (1992b) and West (1992).

While Bayesian nonparametrics strives to produce inherently robust procedures, there have been a number of recent developments which suggest that caution must be exercised. For instance, a "minimal" robustness condition, that one would hope would be satisfied by any Bayes procedure, is consistency: as the sample size grows to infinity, the Bayes estimates of quantities of interest should converge to the true values. It has been discovered, however, that this need not be the case in Bayesian nonparametrics; see Diaconis and Freedman (1986), Ghosh (1993), and Berliner and MacEachern (1993). The following infinite parametric example is a very simple illustration of the phenomenon.

Example 4.. J. K. Ghosh (personal communication, 1992) has studied an interesting variant of the Neyman-Scott problem. Suppose we observe (all independently) $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), i = 1, \ldots, p$ and $j = 1, 2$. It is desired to estimate σ^2. A simple consistent estimator, as $p \to \infty$, is $\hat{\sigma}^2 = \sum_{i=1}^{p} (x_{i1} - x_{i2})^2/(2p)$.

Now suppose a Bayesian were to proceed by choosing independent proper priors for all parameters $\{\sigma^2, \mu_1, \mu_2, \ldots, \mu_p\}$. Then, for "almost all" sequences $\{\mu_1, \mu_2, \ldots\}$, the Bayes estimator of σ^2 seems to be inconsistent. ("Almost all" here is in a topological sense, not probabilistic; the Bayes estimator is consistent for almost all sequences $\{\mu_1, \mu_2, \ldots\}$ in probability under the prior, but the set of such sequences becomes vanishingly small. Conditions on the priors and sequences are needed for the proof of inconsistency, but the result is probably true generally.)

Determining the extent to which such possible inconsistencies are a practical concern for Bayesians will be an important task for the future. At the very least, these concerns should significantly influence the types of priors chosen for these problems (cf., Ghosh, 1994, in regards to the above example).

3. DIAGNOSTICS, INFLUENCE, AND SENSITIVITY

3.1. *Diagnostics*

An important aspect of robustness is developing methods of detecting when a robustness problem exists and suggesting where the difficulty might lie. Examples include the detection of outliers and the detection of a lack of model fit.

Virtually all Bayesian diagnostic techniques are based on some type of utilization of $m(x|\pi, f)$. Interestingly, some suggested utilizations are non-Bayesian in character. For instance, Box (1980) suggests determining the adequacy of an assumed model, f_0, by choosing a noninformative prior, π_0, and then conducting a classical significance test with the null distribution being $m(x|\pi_0, f_0)$. The formal Bayesian approach would be to, instead, embed f_0 in a larger class of models $\mathcal{F}$, choose a prior distribution on $\mathcal{F}$, and infer the adequacy of f_0 relative to other distributions in $\mathcal{F}$ (through, say, Bayes factors or predictive measures). While we prefer the formal Bayesian approach if feasible, the purpose of diagnostics is often to provide an initial indication that something is wrong, and so suggest that the more formal Bayesian approach be undertaken. Evidence obtained from such initial pseudo-Bayesian diagnostics should not be trusted too far, however, and should be confirmed by the formal Bayesian approach before being considered conclusive. For further discussion of this issue, with examples, see Berger (1985, section 4.7.2.).

We do not have space to review the huge literature on Bayesian diagnostics. A few recent references are Smith (1983), Pettit (1988, 1992), Guttman and Peña (1988), Poirier (1988), Kass, Tierney and Kadane (1989), West and Harrison (1989), Carlin and Polson (1991), Verdinelli and Wasserman (1991), Geisser (1992), Kass and Slate (1992), Peña and Tiao (1992), Weiss (1992, 1993), Peña and Guttman (1993), and Meng (1994). Note that global robust Bayesian methods (see Section 4) have begun to themselves be applied to diagnostics; see Bayarri and Berger (1993b, 1994) for an application to outlier detection.

3.2. *Influence and Sensitivity*

Whereas diagnostics is oriented towards detecting that a problem exists with an analysis, influence and sensitivity seeks to determine which features of the model, prior, or utility, or which data, have a large effect on

the answer. There are many parametric analyses of this type, including Guttman and Peña (1988, 1993), Kass, Tierney, and Kadane (1989), McCulloch (1989), Meczarski and Zieliński (1991), Geisser (1992), Lavine (1992d), and Basu and Jammalamadaka (1993).

A recent interesting approach to investigating sensitivity to the prior, in a nonparametric fashion, is to consider functional derivatives of the Bayes operator $\psi(\pi, f)$ with respect to π. (One could, similarly, take derivatives w.r.t. f, but this is usually more involved.) These derivatives, evaluated at a base prior π_0 and in "direction" g, indicate how sensitive $\psi(\pi, f)$ is to local changes in π_0. Besides indicating local sensitivity, these derivatives can be used to construct quite accurate global robustness bounds. The rapidly growing literature on functional derivatives in Bayesian robustness includes Diaconis and Freedman (1986), Cuevas and Sanz (1988), Srinivasan and Truszczynska (1990, 1993), Ruggeri and Wasserman (1991, 1993), Boratyńska and Zielińska (1991), Fortini and Ruggeri (1992, 1994), Sivaganesan (1993c), Basu, Jammalamadaka and Liu (1993a, 1993b), Gustafson and Wasserman (1993), Delampady and Dey (1994), and Salinetti (1994).

4. GLOBAL ROBUSTNESS

4.1. *Introduction*

In Bayesian robustness it is frequently assumed that $f \in \mathcal{F}$ and that $\pi(\theta_f) \in \Gamma_f$, where $\mathcal{F}$ and Γ_f are classes of densities. (Frequently, Γ_f will be enlarged to include distributions that do not have densities with respect to Lebesgue measure; we will abuse notation when this is needed.) If $\psi(\pi, f)$ is the posterior functional of interest (e.g., the posterior mean), global robustness is concerned with computation of

$$\underline{\psi} = \inf_{f \in \mathcal{F}} \inf_{\pi \in \Gamma_f} \psi(\pi, f), \quad \overline{\psi} = \sup_{f \in \mathcal{F}} \sup_{\pi \in \Gamma_f} \psi(\pi, f). \tag{4.1}$$

One then reports $(\underline{\psi}, \overline{\psi})$ as the range of possible answers. If this range is small enough for the conclusion to be clear, the conclusion is declared to be robust. If not, further elicitation, data collection, or analysis is necessary.

4.2. *Parametric Classes*

Historically, global robustness has been investigated using parametric classes of likelihoods and priors.

Example 5.. In the situation of Example 3, instead of considering the $\mathcal{N}(\mu, \sigma^2)$ density for the i.i.d. $X_1, \ldots, X_n$, one could consider the class

$$\mathcal{F} = \{\mathcal{T}_\alpha(\mu, K_\alpha \sigma^2) \text{ densities for the } X_i, \alpha \geq 1\}, \tag{4.2}$$

where $\sqrt{K_\alpha} = (0.674)/q_\alpha$, with q_α being the third quartile of the $\mathcal{T}_\alpha(0, 1)$ distribution. K_α is introduced because the $\mathcal{T}_\alpha(\mu, K_\alpha \sigma^2)$ distribution will then have the same quartiles as the $\mathcal{N}(\mu, \sigma^2)$ distribution, so that μ and σ^2 will have comparable meanings across all distributions. If the restriction $\alpha > 2$ were employed, one could instead choose K_α so that all distributions have the same mean and variance, but we generally prefer scaling by quartiles.

Suppose μ and σ^2 are thought to be independent apriori, with μ having unimodal density with quartiles $-1, 0, 1$ and nothing being known about σ^2. Then the prior, $\pi(\mu, \sigma^2)$, might be assigned the class

$$\Gamma = \{\pi(\mu, \sigma^2) = \pi_1(\mu)\pi_2(\sigma^2)\text{: } \pi_1 \text{ is } \mathcal{T}_\nu(0, q_\nu^{-2}), \nu \geq 1,$$

$$\text{and} \qquad \pi_2(\sigma^2) = (\sigma^2)^a, -2 \leq a \leq 0.\}$$

The $\mathcal{T}_\nu(0, q_\nu^{-2})$ distributions are a fairly wide class of unimodal distributions with quartiles $-1, 0, 1$, and might appropriately represent the specified information about μ. Since nothing is specified about σ^2, it would be typical to use a range of noninformative priors as the relevant class (but see Pericchi and Walley, 1991; and Walley, 1991, for other suggestions). Note that, because of the scaling of the $f \in \mathcal{F}$ to preserve the meaning of μ and σ^2, it is not necessary to write (μ_f, σ_f^2) and define Γ_f depending on f.

For any functional $\psi(\pi, f)$ of interest, one can now compute $(\underline{\psi}, \overline{\psi})$ by minimizing and maximizing $\psi(\pi, f)$ with respect to (α, ν, a).

There are two main reasons that parametric robustness is attractive. The first is that computations are relatively straightforward. For instance, in Example 5, the maximizations are only three-dimensional. Of course, computation of the $\psi(\pi, f)$ will involve two-dimensional integration (over μ and σ^2), so the computation is not trivial (see, also, Section 5.1).

The second attractive feature of parametric classes is that they can allow for convenient communication of robust Bayesian conclusions. An example is given in Section 4.10.2 of Berger (1985).

The main disadvantage of parametric classes is that they may fail to capture realistic possible deviations from the base model or prior. Thus, in Example 5, we have robustified against normality, in the sense of allowing flatter tails for the distributions, but no allowance for, say, possible skewness has been made. Ideally, one will construct $\mathcal{F}$ and/or the Γ_f to reflect all deviations that are deemed to be possible, but it is unfortunately all-too-common to fail to anticipate the actual deviations that arise.

Recent references utilizing parametric classes of priors include Leamer (1982), Polasek (1985), Good and Crook (1987), Polasek and Pötzelberger (1988, 1994), DasGupta and Studden (1988a, 1988b, 1989, 1991), Drummey (1991), Pötzelberger and Polasek (1991), Coolen (1993), and Dette and Studden (1994).

4.3. *Nonparametric Classes of Priors*

The majority of recent papers on Bayesian robustness deal with a fixed likelihood and nonparametric classes of priors. This is an important problem, for several reasons. First, there are many situations in which priors are less well known than the model. Second, the major objection of non-Bayesians to Bayesian analysis is uncertainty in the prior, so eliminating this concern can make Bayesian methods considerably more appealing. Third, serious inadequacies in certain classical methods can be revealed by Bayesian prior robustness (see Sections 1.2 and 4.3.3). Finally, conclusions must frequently be reached by a group of people with differing prior opinions, and robust Bayesian analysis, with Γ equal to the class of prior opinions, can then have a variety of uses.

That said, the main reason researchers have concentrated on global prior robustness is probably its mathematical elegance. There is nothing wrong with this, of course, as long as we remember that global prior robustness is only one piece of the robustness puzzle.

In the remainder of this subsection, f will be considered fixed, so we write just θ for the unknown parameters, Γ for the class of priors being considered, and $\psi(\pi)$ (instead of $\psi(\pi, f)$) as the criterion functional.

4.3.1. *Factors Involved in Choosing a Class.*

Several discussions and reviews concerning choice of good classes of priors already exist, including Berger (1990), Sivaganesan (1990), Lavine (1991), Pericchi and Walley (1991), Walley (1991), Moreno and Pericchi (1992a), and Wasserman (1992b). The following issues should be kept in mind in choosing a class:

(i) The class should be as easy to elicit and interpret as possible. Recall that a prime reason for considering Bayesian robustness is the difficulty of eliciting a prior; making the class, Γ, difficult to elicit would thus be self-defeating.

(ii) The class should be as easy to handle computationally as is possible. The usual computational technique is to identify "extreme points" of Γ (relative to $\psi(\pi)$) and perform maximizations over these extreme points. Typically, the extreme points will be in a low-dimensional subset, Γ^*, of Γ, so the maximizations are over a low-dimensional set. The dimension of Γ^* will depend on several factors, but primarily on the dimension of θ and the number of elicited features of the prior. Hence, rather paradoxically, the more features one elicits, the harder the robust Bayesian computation is likely to become. Part of the computability issue is also having a class, Γ, which is compatible with model and/or utility robustness.

(iii) The size of Γ should be appropriate, in the sense of being a reasonable reflection of prior uncertainty. If Γ is too small, one might fear being erroneously led to a conclusion of robustness. If Γ is too large, in the sense of containing many prior distributions that are clearly unreasonable, then one might conclude that robustness is lacking when, in fact, a reasonable Γ would imply robustness. For detecting this latter problem, it is useful to determine the $\pi \in \Gamma$ at which $\underline{\psi}$ or $\overline{\psi}$ is attained, and judge if such a π is reasonable. If not, one should try to refine Γ to eliminate such π.

(iv) Γ should be extendable to higher dimensions and adaptable in terms of allowing incorporation of constraints (e.g., shape constraints, independence, etc.). The point here is that eventual methodological implementations will need to be based on at most a few "standard" classes (for elicitational, computational,

and interpretational reasons), and so these classes need to be flexible enough to handle a very wide range of problems.

The following simple example illustrates several of the above ideas.

Example 6.. Suppose prior beliefs about a real parameter θ are symmetric about 0, with the third quartile, q_3, being between 1 and 2. Consider

$$\Gamma_1 = \{\mathcal{N}(0, \tau^2) \text{ priors}, 2.19 < \tau^2 < 8.76\},$$

$$\Gamma_2 = \{\text{all symmetric priors with } 1 < q_3 < 2\}.$$

Both classes are easy to elicit (i.e., easy to specify from the given information; the range of τ^2 in Γ_1 yields q_3 between 1 and 2). Also, both are easy to handle computationally; indeed, maximization over Γ_2 will often only involve maximization over the "extreme points"

$$\Gamma_2^* = \{\text{distributions giving probability } \tfrac{1}{2} \text{ each to } \pm q_3\colon\ 1 < q_3 < 2\}.$$

Although Γ_1 can be appropriate for some situations, it will often be considered "too small" because of its specified prior shape and because it has only sharp-tailed distributions. In contrast, Γ_2 will typically be a "too big" reflection of the prior information, in the sense that it contains prior distributions which, upon reflection, are probably unreasonable.

Very sensible classes can be formed by taking "too large" classes, such as Γ_2, and adding shape constraints. For instance, if it is also believed that the prior density is unimodal, then one obtains

$$\Gamma_3 = \{\text{unimodal, symmetric densities with } 1 < q_3 < 2\}.$$

Such classes are often very sensible, in that they are large enough to include all reasonable priors compatible with prior information, but small enough that unreasonable priors are excluded.

4.3.2. *Common Classes.*

We briefly review the common classes of priors that are used. For extensive discussion, comparisons, and examples, see the references listed under each class.

Classes of Given Shape or Smoothness: An example of a class based on shape is $\Gamma = \{\text{all symmetric, unimodal priors}\}$. Such classes have

interesting uses in hypothesis testing (see Section 4.3.3). Usually, however, shape is used as an additional constraint in one of the other classes (cf., Example 6), so as to eliminate unreasonable priors from the class. Note that general shape features are often relatively easy to elicit, even in higher dimensions.

Smoothness constraints typically limit the rate of change of the prior density. (Note that requiring only continuity adds nothing, because arbitrary distributions can typically be approximated, arbitrarily well, by continuous densities.) Although one could define a class of priors based only on smoothness, it is typically used, instead, as a supplemental constraint for other classes (cf., Bose, 1990, 1994).

Moment Class: This is defined as the set of all priors with a specified collection of moments. Analysis using probabilistic moment theory is typically straightforward. See Sivaganesan and Berger (1989, 1993), Goutis (1991), Betró, Meczarski and Ruggeri (1994), and Sivaganesan (1992).

Moments are quite difficult to elicit. For this reason, moment conditions are also typically used merely as additional constraints in other classes, in the hope that misspecification of moments will then have a reduced effect.

Contamination Class: This is defined by

$$\Gamma = \{\pi = (1-\varepsilon)\pi_0 + \varepsilon q, \;\; q \in \mathcal{Q}\}, \tag{4.3}$$

where π_0 is a base prior (for instance, the prior elicited in a standard Bayesian analysis), ε is the perceived possible error in π_0, and $\mathcal{Q}$ is the allowed class of contaminations. In terms of the four criteria of Section 4.3.1, this class is easy to elicit; computation is relatively easy for many reasonable choices of Q; and the class can easily incorporate additional constraints and be used in higher dimensions. The class can be "too big" if $\mathcal{Q}$ is "too big" and ε is appreciable. In one dimension this is rarely a problem, but it can be a severe problem in higher dimensions. References include Berger and Berliner (1986), Sivaganesan (1988, 1989, 1993a), Sivaganesan and Berger (1989), Moreno and Pericchi (1990, 1991), Dey and Birmiwal (1991), Gelfand and Dey (1991), Boratyńska (1991), Lavine (1991b), Moreno and Cano (1991), and Bose (1994).

Density Ratio (or Density Band) Class: This is defined by

$$\Gamma = \{\text{generalized } \pi\text{: } L(\theta) \leq \pi(\theta) \leq U(\theta)\}. \tag{4.4}$$

(A "generalized" prior is one which does not integrate to 1; typically the posterior will, nevertheless, be proper.) Often this class is the simplest to handle computationally, and is reasonable in higher dimensions. Its main disadvantage is that it is very hard to elicit; choosing L and U appropriately can be quite difficult.

A useful modification of this class is the *Density Bounded Class*, which is as in (4.4), but with the additional constraint that π must be proper. The class then becomes much easier to elicit and interpret, but can be more challenging computationally.

References working with these classes include DeRobertis (1978), DeRobertis and Hartigan (1981), Hartigan (1983), Lavine (1991a, 1991b, 1992c), Ruggeri and Wasserman (1991), Wasserman (1991, 1992a, 1992b, 1992c), Moreno and Pericchi (1992b), and Sivaganesan (1994).

Quantile Class: This is defined by

$$\Gamma = \{\pi\colon\ \alpha_i \leq \Pr(\theta \in \Theta_i) \leq \beta_i,\ \ i = 1, \ldots, m\},$$

where the Θ_i are specified subsets of Θ. (Usually $\{\Theta_i;\ i = 1, \ldots, m\}$ is a partition of Θ.) This class is probably the most natural of all from the viewpoint of elicitation, and is computationally manageable. It tends to be "too big" in higher dimensions, however, unless additional shape constraints are added. References to this class include Cano, Hernandez, and Moreno (1985), Berger and O'Hagan (1988), O'Hagan and Berger (1988), Moreno and Cano (1989), Moreno and Pericchi (1990), Ruggeri (1990, 1991, 1992), and Sivaganesan (1991).

Mixture Classes: These are of the form

$$\Gamma = \{\pi(\theta) = \int \pi(\theta|\alpha) dG(\alpha),\ G \in \mathcal{G}\}. \tag{4.5}$$

Most other classes are actually themselves mixture classes.

Example 7.. Suppose $\theta \in R^p$, and the prior distribution is known to depend only on $|\theta|$. The class, Γ_S, of all such priors is typically too big, in the sense of containing many unreasonable distributions. Often, however, unimodality is also believed to hold, leading to Γ_{US}, the class of unimodal spherically symmetric priors. Interestingly, this class can be written as

$$\Gamma_{US} = \{\pi(\theta) = \int_0^\infty 1_{(0,\alpha)}(|\theta|)V_\alpha^{-1}dG(\alpha), G \text{ any c.d.f. on } [0,\infty)\}, \tag{4.6}$$

where V_α is the volume of the ball in R^p of radius α. This is a much smaller class than Γ_S, and would be reasonable for most purposes, but it may be possible to refine the class even further. In particular, if prior beliefs are felt to be "bell-shaped," a class such as

$$\Gamma_{NS} = \{\pi(\theta) = \int_0^\infty (2\pi\alpha)^{-p/2}e^{-|\theta|^2/(2\alpha)}dG(\alpha),$$
$$G \text{ any c.d.f. on } [0,\infty)\}$$

could be employed. This is easily seen to be a subset of Γ_{US} that contains only bell-shaped distributions (though admittedly not all bell-shaped distributions). Recall that we earlier encountered such priors in Section 2.2, as being "inherently robust" for certain G.

Example 8.. An archeological artifact is θ years old, θ unknown. It could have been produced by any one of 3 civilizations that occupied the given site. For civilization i, a $\mathcal{N}(\mu_i, A_i)$ distribution (to be denoted π_i) is thought to describe the likelihood of artifact production at any given time. (All μ_i and A_i are assumed to be specified.)

Several experts are asked to classify the object, based on its style. They do not agree completely, but conclude that all their opinions are contained in

$$\mathcal{G} = \{g = (g_1, g_2, g_3)\colon\ 0.1 \le g_1 \le 0.2,\ 0.6 \le g_2 \le 0.7\},$$

where $g_i = \Pr$ (the artifact is from civilization i), and $g_1 + g_2 + g_3 = 1$. Then $\pi(\theta)$, the overall prior distribution for θ, is in

$$\Gamma = \left\{\pi(\theta) = \sum_{i=1}^{3} g_i\pi_i(\theta)\colon\ g \in \mathcal{G}\right\},$$

which is of the form (4.5) with $\alpha = i$ and G being discrete.

Mixture classes will play a very prominent role in Bayesian robustness because of several key properties:

(i) Mixture classes are often computationally simple. In Example 7, for instance, maximization over Γ_{US} or Γ_{NS} will typically reduce to maximization over

$$\pi(\theta|\alpha) = 1_{(0,\alpha)}(|\theta|)V_\alpha^{-1}$$

or

$$\pi(\theta|\alpha) = (2\pi\alpha)^{-p/2}\exp\{-|\theta|^2/(2\alpha)\},$$

respectively, both of which are simple one-dimensional maximizations (over α).

(ii) Mixture classes can flexibly represent prior information about structure or shape, as in Example 7, or information arising from several sources, as in Example 8.

(iii) Mixture classes are often not "too big," in the sense of containing unreasonable distributions. This is particularly crucial for multivariate θ, where the range of Bayesian answers as π varies over Γ will typically be huge, unless Γ is somehow constrained so as not to contain unreasonable distributions. Operating with mixture classes seems to be the only effective way of avoiding the problem (other than using parametric classes, of course).

As a final comment about mixture classes, note that they can also arise as refinements in the elicitation process. In Example 8, for instance, suppose $X \sim \mathcal{N}(\theta, \sigma^2)$ is observed (say, X is a radiocarbon dating of the artifact). One might first consider just $\Gamma = \{\pi_1, \pi_2, \pi_3\}$, and compute the Bayesian answer (e.g., posterior mean of θ) for each prior in Γ. If the range of answers is small enough, there would be no need to look further. If, however, there are substantial differences between the answers, then one might go to the next "level" of elicitation, obtaining g. In this example, uncertainty thus ends up residing in the higher level elicitation (see also Good, 1983b; and Pericchi and Nazaret, 1988). Note that there could easily be virtually complete robustness with respect to $g \in \mathcal{G}$, even if there is no robustness with respect to the π_i in Γ. For other examples of mixture classes, see Bose (1990, 1993), Cano (1993), Moreno and Pericchi (1993a), and Liseo, Petrella, and Salinetti (1993). West (1992) discusses their uses in modelling.

Marginal and Independence Classes: When $\theta = (\theta_1, \ldots, \theta_p)$ is multidimensional, elicitation of $\pi(\theta)$ is particularly difficult. One could hope that elicitation of, say, the marginal densities, $\pi_i(\theta_i)$, could be effective, in the sense that robustness over

$$\Gamma = \{\pi(\theta) \text{ having the specified marginals}\} \tag{4.7}$$

would often obtain. Alas, this is not the case, as was dramatically shown by Lavine, Wasserman, and Wolpert (1991); the class in (4.7) is so large that the range of resulting Bayesian answers is typically enormous. See, also, Moreno and Cano (1992) for related results.

Of course, if the θ_i were, apriori, judged to be independent, then one would simply have the single prior

$$\pi(\theta) = \prod_{i=1}^{p} \pi_i(\theta_i).$$

It is sometimes possible to make the judgment of independence, and it is then natural to consider its effect on other classes. This is studied for contamination classes and density ratio classes in Berger and Moreno (1994), where it is shown that the assumption of independence of coordinates does have a dramatic effect on robustness; the range of Bayesian answers can decrease dramatically. Independence is admittedly a strong assumption, but one typically must make strong assumptions in multidimensional problems to obtain a moderate range of Bayesian answers.

Other Classes: An interesting alternative to the Density Ratio class for one-dimensional θ is the *Distribution Band* class of all priors whose c.d.f. lies between two nondecreasing functions. This is studied in Basu (1992a, 1992b) and Basu and DasGupta (1992).

Neighborhood classes can be defined by choosing a "distance measure," $d(\pi_1, \pi_2)$, between priors (it need not be a true distance function), and defining $\Gamma = \{\pi\colon\ d(\pi, \pi_0) \leq \varepsilon\}$, where π_0 is again a "base" prior. Related classes can be developed using "concentration functions"; see Regazzini (1992) and Fortini and Ruggeri (1990, 1992, 1994).

Belief Function classes use belief functions (a type of generalization of probability) to generate the class of priors. See Wasserman (1990) for an example.

Classes based on *Choquet Capacities* are defined and studied in Wasserman and Kadane (1990, 1992a) and Wasserman (1992b). Two-alternating capacities have particularly attractive theoretical and computational properties.

Pericchi and Walley (1991) and Walley (1991) (see also Sansó and Pericchi, 1992) propose *Near Ignorance* classes of priors to provide a robust Bayesian alternative to noninformative priors. Their approach provides an interesting contrast to typical constructions of Γ, which seek to construct Γ so as to contain all the "nice," believable priors. Instead, Pericchi and Walley argue that one can construct a "nice" class from a collection of "not nice" or not compatible priors, and that there can be positive advantages in doing so. In part, this notion arises from the axiomatic development in Walley (1991), which effectively shows that rationality corresponds to operating with some Γ, but does not require or imply that the priors in Γ correspond to actual subjective beliefs. Although counterintuitive to standard Bayesian thinking, this approach should not be casually dismissed. Its counterintuitive nature, however, poses real difficulties for elicitation of Γ.

Conclusions: No single class of priors is likely to dominate robust Bayesian analysis. Our personal favorites are contamination, quantile, and mixture classes, with shape and structural restrictions as appropriate. We prefer the contamination and quantile classes because they are easiest to elicit and interpret. They can be considerably more difficult than, say, the density ratio class in terms of computation, but computations will eventually just be hidden within software. The important issue will be how user-friendly is the software in terms of choice of Γ, so the most easily elicitable classes are to be preferred.

The argument for mixture classes is somewhat different, although they too are often natural from an elicitation viewpoint. The argument is simply the necessity, in multi-dimensional problems, of doing something fairly drastic to reduce the size of Γ in order to avoid excessively large ranges of Bayesian answers. It is important to be clear here: in very low dimensional problems one can often verify Bayesian robustness, even when the prior inputs are very weak. In high-dimensional problems this is typically impossible, and one must accept the need to make rather strong and "dangerous" assumptions if an answer is to be obtained. The point, of course, is to only make those strong assumptions which seem

plausible. The next section contains an example illustrating some of these notions.

4.3.3. *Application to Hypothesis Testing and Ockham's Razor.*

Some of the most interesting applications of robust Bayesian analysis have been to hypothesis testing, and related model selection ideas. We review these here, in part as an illustration of points made in the preceding section.

Suppose $\boldsymbol{X} = (X_1, \ldots, X_p) \sim \mathcal{N}_p(\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$, σ^2 known, and that model M_1 specifies $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, while model M_2 has $\boldsymbol{\theta}$ unrestricted. Under M_2, consider the following two classes of priors for $\boldsymbol{\theta}$:

$$\Gamma_A = \{\text{all prior distributions}\},$$
$$\Gamma_{\boldsymbol{\mu}} = \{\text{all prior densities of the form } \pi(\boldsymbol{\theta}) = h(|\boldsymbol{\theta} - \boldsymbol{\mu}|), \; h \text{ nonincreasing}\}.$$

Here $\boldsymbol{\mu}$ is fixed, corresponding to a prior "most likely" value of $\boldsymbol{\theta}$ under M_2 or, perhaps, to the "center of symmetry" of π under M_2. Often, $\boldsymbol{\mu}$ will equal $\boldsymbol{\theta}_0$, but other values are possible.

The Bayes factor of M_1 to M_2, corresponding to a prior density $\pi(\boldsymbol{\theta})$ under M_2, is

$$B(\pi) = f(\boldsymbol{x}|\boldsymbol{\theta}_0) / \int f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Define $\underline{B}_A$ and $\underline{B}_{\boldsymbol{\mu}}$ as the lower bounds on $B(\pi)$ as π ranges over Γ_A and $\Gamma_{\boldsymbol{\mu}}$, respectively.

For the case $\boldsymbol{\mu} = \boldsymbol{\theta}_0$, which arises naturally in testing H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus H_1: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ with prior opinions under H_1 being symmetric about $\boldsymbol{\theta}_0$, Table 2 gives values of $\underline{B}_A$ and $\underline{B}_{\boldsymbol{\theta}_0}$; instead of presenting the values as a function of $\boldsymbol{x}$, we state them as a function of the P-value associated with $\boldsymbol{x}$. (See Delampady, 1989, for computation of $\underline{B}_{\boldsymbol{\theta}_0}$.)

This table reveals the, by now familiar, discrepancy between P-values and Bayes factors. In one dimension for instance, when the P-value is 0.05 the lower bound on the Bayes factor over all symmetric (about $\boldsymbol{\theta}_0$) unimodal priors is 0.409, and the lower bound over *all* priors is 0.146. This proves that a P-value of 0.05, in this situation, is at best quite weak evidence against H_0.

P-value		dimension p 1	2	3	4	5	10	15	20	40
0.05	$\underline{B}_A$	.146	.050	.020	.009	.004	10^{-4}	10^{-6}	10^{-7}	10^{-12}
	$\underline{B}_{\theta_0}$	.409	.348	.326	.314	.307	.293	.288	.284	.279
0.01	$\underline{B}_A$	.036	.010	.003	.001	.0005	10^{-5}	10^{-7}	10^{-8}	10^{-14}
	$\underline{B}_{\theta_0}$	.123	.098	.090	.085	.082	.078	.075	.074	.073
0.001	$\underline{B}_A$	.004	.001	.0003	.0001	10^{-5}	10^{-7}	10^{-8}	10^{-10}	10^{-16}
	$\underline{B}_{\theta_0}$	.018	.014	.012	.011	.010	.009	.009	.009	.009

Table 2. *Lower Bounds, $\underline{B}_A$ and $\underline{B}_{\theta_0}$, corresponding to various P-values for testing H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$.*

A secondary point is the demonstration that classes of priors which are "too big" fail to give useful bounds in high dimensions. Thus $\underline{B}_A$ becomes uselessly small as the dimension increases. In contrast, $\underline{B}_{\theta_0}$ is very stable as the dimension increases. And note that, as in (4.6), Γ_{θ_0} can be written as a one-dimensional mixture class; this is an example of why we view mixture classes as promising in high dimensions.

For $p = 1$ and general μ, a very accurate approximation to $\underline{B}_\mu$ is

$$\underline{B}_\mu = 2\varphi(d_0)[d_1 + \sqrt{2\log(d_1 + 1.2)}], \tag{4.8}$$

where $d_0 = |x - \theta_0|/\sigma$, $d_1 = |x - \mu|/\sigma$, and φ is the standard normal density. This is argued, in Berger and Jefferys (1992) and Jefferys and Berger (1992), to be a "Bayesian Ockham's razor" for comparing models M_1 and M_2. For instance, those papers discussed the situation of comparing, in about the year 1920, M_1: *Einstein's general relativity* versus M_2: *Newcomb's gravity theory*, based on data from unexplained perturbations in the orbit of Mercury. The situation fits our framework with $\theta_0 = 42.9$ (the perturbation predicted by M_1); $\mu = 0$ (Newcomb's theory made no prediction about the size or sign of the perturbation, so centering prior opinion at zero is natural); and $x = 41.6$ (the observed perturbation, with a standard error of $\sigma = 2$). Computation yields $\underline{B}_\mu = \underline{B}_0 = 15.04$; since this is a lower bound over $\Gamma_\mu = \Gamma_0$, we can conclude that the evidence favors M_1 by *at least* 15 to 1. This relates to Ockham's razor because M_1 was the "simple" model, in the sense of

having no free parameter (it specified $\theta_0 = 42.9$), while M_2 allowed θ to float freely. "Ockham's razor" argues that one should prefer a simple model that adequately explains the data to a complex model that does so, which is precisely what $\underline{B}_\mu$ established quantitatively.

As a final point, it must be remembered that the $\underline{B}$ above are lower bounds on the Bayes factor, and can be much lower than actual reasonable Bayes factors (cf, Bayarri and Berger, 1994). If the lower bound, itself, answers the question of interest, then all is well. If not, substantial refinement of Γ is needed. Note that, in contrast to estimation problems, there are not (in general) "robust" noninformative priors for testing or model selection problems. See Kass and Raftery (1992) and Berger and Pericchi (1993) for discussion and default methods of proceeding.

Robust Bayesian analysis of testing problems can be found in Edwards, Lindman, and Savage (1963), Berger and Sellke (1987), Berger and Delampady (1987), Casella and Berger (1987), Delampady (1989a, 1989b), Moreno and Cano (1989), DasGupta and Delamapady (1990), Delampady and Berger (1990), Berger and Mortera (1991, 1994), Berger (1992), and Berger and Jefferys (1992).

4.4. *Nonparametric Classes of Likelihoods*

Dealing with likelihoods via the global robustness approach varies from trivial to nearly impossible. One approach is to take the observed likelihood function, $\ell_0(\theta) = f_0(x|\theta)$, for a hypothesized model f_0 and given data, and embed it in a larger class of likelihoods, such as $\mathcal{F}_\varepsilon = \{\ell(\theta) = (1-\varepsilon)\ell_0(\theta) + \varepsilon q(\theta),\ q \in \mathcal{Q}\}$. Since $\ell(\theta)$ and $\pi(\theta)$ operate interchangeably in Bayesian computations, this approach to likelihood robustness is equivalent to the global prior robustness approach (with the contamination class).

The difficulty with this approach is that such classes of likelihoods do not reflect typical types of uncertainty in f. For instance, if $X_1, \ldots, X_n$ are i.i.d. $g(x_i|\theta)$, uncertainty would typically reside in g, reflected by, say,

$$\mathcal{F}_1^g = \{g = (1-\varepsilon)g_0(x_i|\theta) + \varepsilon q(x_i|\theta),\ q \in \mathcal{Q}\} \text{ or}$$
$$\mathcal{F}_2^g = \{\text{densities } g\colon\ g_1(x_i|\theta) \le g(x_i|\theta) \le g_2(x_i|\theta)\}.$$

(Even these may not be completely natural, but they suffice for making the point.) The resulting classes of likelihoods are

$$\mathcal{F}_j = \left\{ \ell\colon\ \ell(\theta) = \prod_{i=1}^{n} g(x_i|\theta),\ g \in \mathcal{F}_j^g \right\},\ \ j = 1, 2, \tag{4.9}$$

and these are very complex and difficult to work with. For instance, the relevant subclasses of extreme points are typically at least n-dimensional, which can become prohibitively expensive computationally for large n.

A second difficulty with classes of likelihoods, as in (4.9), is that they can be too large, unless the $\mathcal{F}_j^g$ are very small. One approach that does seem to give useful answers is that of Lavine (1991a, 1991b, and 1994).

For special or restricted problems, robustness analysis can be much easier. Robustness among certain generalized elliptical distributions is studied in Osiewalski and Steel (1993a, b, c), and Fernández, Osiewalski, and Steel (1993). The following example is from another special situation, studied in Bayarri and Berger (1993a).

Example 9.. (*Weighted Distributions*): Assume that the random variable $X \in \Re^1$ is distributed over some population of interest according to $f(x|\theta)$, $\theta \in (r, s)$, a (possibly infinite) interval in $\Re^1$, but that, when $X = x$, the probability of recording x (or the probability that x is *selected* to enter the sample) is $w(x)$. Then the true density of an actual observation is

$$f_w(x|\theta) = \frac{w(x)f(x|\theta)}{\nu_w(\theta)}, \tag{4.10}$$

where $\nu_w(\theta) = E_\theta[w(X)]$. Selection models occur often in practice (Rao, 1985; Bayarri and DeGroot, 1992).

Often the specification of $w(\cdot)$ is highly subjective. It is thus of considerable interest to study the robustness of the analysis to choice of w. The problem becomes particularly important in the multi-observational setting, because the effect of the weight function can then be extremely dramatic. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from the density (4.10), so that the likelihood function for θ is

$$L_w(\theta) \propto \ell(\theta)[\nu_w(\theta)]^{-n}, \tag{4.11}$$

where $\ell(\theta) \propto \prod_{i=1}^{n} f(x_i|\theta)$ would be the likelihood function for the unweighted base density. If $\pi(\theta)$ is the prior density for θ, the posterior density is then

$$\pi(\theta|x, w) = \frac{\ell(\theta)[\nu_w(\theta)]^{-n}\pi(\theta)}{\int \ell(\theta)[\nu_w(\theta)]^{-n}\pi(\theta)d(\theta)}, \tag{4.12}$$

assuming π is such that the denominator is finite. Expression (4.12) suggests that, at least for large n, the weight function w can have a much more significant effect on $\pi(\theta|x, w)$ than might the prior π. Hence we will treat $\pi(\theta)$ as given here; for instance, it might be chosen to be a noninformative prior for the base model $f(x_i|\theta)$.

In Bayarri and Berger (1993), this problem is studied for the class of weight functions

$$\mathcal{W} = \{\text{nondecreasing } w\text{: } w_1(x) \leq w(x) \leq w_2(x)\}, \tag{4.13}$$

where w_1 and w_2 are specified nondecreasing functions representing the extremes of beliefs concerning w. Posterior functionals

$$\psi(w) = \int \xi(\theta)\pi(\theta|x, w)d\theta \tag{4.14}$$

are studied for a variety of shapes of the target $\xi(\theta)$. When $\xi(\theta)$ is monotonic (e.g., $\xi(\theta) = \theta$ or $\xi(\theta) = 1_{(c,\infty)}(\theta)$), the extreme points in $\mathcal{W}$ at which $\overline{\psi} = \sup_w \psi(w)$ and $\underline{\psi} = \inf_w \psi(w)$ are attained were shown to have one of the following two forms:

$$w(x) = \begin{cases} w_1(x) & \text{if } r < x \leq a \\ w_2(x) & \text{if } a < x < s \end{cases}, \tag{4.15}$$

$$w(x) = \begin{cases} w_2(x) & \text{if } r \leq x < h_2(c) \\ c & \text{if } h_2(c) < x < h_1(c), \\ w_1(x) & \text{if } h_1(c) < x < s \end{cases} \tag{4.16}$$

where $h_1(c) = \inf\{x\text{: } w_1(x) \leq c\}$ and $h_2(c) = \sup\{x\text{: } w_2(x) \geq c\}$. The condition needed for this result is primarily that $f(x|\theta)$ have monotone likelihood ratio.

As a specific example, suppose $f(x_i|\theta) = \theta \exp\{-\theta x_i\}$ for $i = 1, \ldots, n$, where $x_i > 0$ and $\theta > 0$. Any x_i that is less than a value T_1 is, however, not observed. Any x_i that is greater than T_2 is observed. For $T_1 \leq x_i \leq T_2$, the probability of its being observed is not known, but the probability is known to be nondecreasing. This specifies the class of weight functions in (4.13), with $w_1(x) = 1_{(T_2,\infty)}(x)$ and $w_2(x) = 1_{(T_1,\infty)}(x)$.

Suppose $\xi(\theta) = \theta$ is of interest, so that $(\underline{\psi}, \overline{\psi})$ is the range of the posterior mean as w ranges over $\mathcal{W}$. Then one can explicitly minimize and maximize (4.14) over w of the form (4.15) and (4.16), obtaining $\underline{\psi} = 1/(\overline{x} - T_1)$ and $\overline{\psi} = 1/(\overline{x} - T_2)$. Whether or not robustness is achieved is thus easy to determine. Note that it depends on the size of $\overline{x}$ as well as the closeness of T_1 and T_2.

4.5. *Limitations of Global Robustness*

Global robustness ignores a very important quantity, namely $m(x|\pi, f)$, which can be considered to be the "likelihood" of π and/or f. A full Bayesian analysis automatically takes this into account.

Example 10.. Suppose $X_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \ldots, p$. The class of prior distributions under consideration for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is

$$\Gamma = \{\pi(\boldsymbol{\theta})\text{: the } \theta_i \text{ are i.i.d. } \mathcal{N}(\mu, 1),\ -8 < \mu < 12\}.$$

Suppose we are interested in the posterior mean for θ_1. This is given by $\hat{\theta}_1 = (x_1 + \mu)/2$. Thus the range of posterior means, as π varies over Γ, is $(\frac{1}{2}x_1 - 4, \frac{1}{2}x_1 + 6)$.

Calculation shows that, here,

$$m(\boldsymbol{x}|\pi) = \frac{1}{(4\pi)^{p/2}} \exp\left\{-\tfrac{1}{4}[p(\overline{x} - \mu)^2 + \sum_{i=1}^{p}(x_i - \overline{x})^2]\right\}.$$

Thus values of μ close to $\overline{x}$ are far more "likely" than values far from $\overline{x}$. For instance, if $p = 8$ and $\overline{x} = 3$, this likelihood is a normal likelihood with mean 3 and variance $1/4$, so that only the μ between 2 and 4 have appreciable likelihood. Note that, if a full Bayesian analysis were done with, say, μ being given the noninformative prior $\pi(\mu) = 1$, then the

posterior mean for θ_1 would be

$$\begin{aligned}\hat{\theta}_1^* &= \int \tfrac{1}{2}(x_1+\mu)\pi(\mu|\overline{x})d\mu \\ &= \int \tfrac{1}{2}(x_1+\mu)\frac{1}{\sqrt{2\pi(2/p)}}\exp\{-\tfrac{p}{4}(\overline{x}-\mu)^2\}d\mu \\ &= \tfrac{1}{2}(x_1+\overline{x}),\end{aligned}$$

which has effectively "weighted" the $(x_1+\mu)/2$ by the likelihood of μ.

The message here is that a global robustness analysis might erroneously indicate a lack of robustness, erroneous in the sense that, were $m(x|\pi, f)$ taken into account, robustness might obtain. There are two possibilities for formally investigating if this is so. The first is to go to a "higher level" Bayesian robustness investigation, as in Cano (1993) and Moreno and Pericchi (1993a).

Example 10 (continued).. It is determined that $(\frac{1}{2}x_1 - 4, \frac{1}{2}x_1 + 6)$ is too large an interval to reach a conclusion. Prior information about μ, the presumed common mean of the θ_i, is thus considered. A "best guess" for μ is 2, but there is considerable uncertainty in this guess. It is decided that the standard error of this guess is at least 2, but that finer elicitation would be difficult. This information can reasonably be modeled by the class of priors (for μ)

$$\Gamma^* = \{\mathcal{N}(2, \tau^2) \text{ densities}, \tau^2 \geq 4\}.$$

For given τ^2, an easy computation yields that the posterior mean for θ_1 is

$$\hat{\theta}_1^{**} = \tfrac{1}{2}(x_1+\overline{x}) + \tfrac{1}{(p\tau^2+2)}(2-\overline{x}).$$

The range of possible posterior means as $\pi(\mu)$ varies over Γ^* (i.e., for $\tau^2 \geq 4$) is thus (if, say, $2 - \overline{x} \leq 0$)

$$(\tfrac{1}{2}(x_1+\overline{x}) + \tfrac{1}{(4p+2)}(2-\overline{x}),\ \tfrac{1}{2}(x_1+\overline{x})).$$

For the case $p = 8$ and $\overline{x} = 3$, this range is

$$(\tfrac{1}{2}(x_1+3) - \tfrac{1}{34},\ \tfrac{1}{2}(x_1+3)), \tag{4.17}$$

which would typically be considered to be a highly robust conclusion.

The second possibility for utilizing $m(x|\pi)$ (similar ideas apply if dealing with $m(x|\pi, f)$) is to replace Γ by

$$\Gamma^* = \{\pi \in \Gamma\colon\ m(x|\pi) \geq K\}; \tag{4.18}$$

here K could be chosen by likelihood or noninformative prior Bayesian methods. (See Sivaganesan and Berger, 1993, for development of this approach.)

Example 10 (continued).. For the case $p = 8$ and $\overline{x} = 3$, we observed that $m(x|\pi)$ is essentially a $\mathcal{N}(3, \frac{1}{4})$ likelihood for μ. Likelihood or Bayesian noninformative prior methods would suggest that (2, 4) is a "95% confidence or credible set" for μ, so we might replace Γ by

$$\Gamma^* = \{\pi \in \Gamma\colon\ 2 \leq \mu \leq 4\}$$

(which can easily be seen to be of the form (4.18)). The range of the posterior mean for θ_1, as π ranges over Γ^*, is $(\frac{1}{2}x_1 + 1, \frac{1}{2}x_1 + 2)$, which might well be small enough to claim that the conclusion is reasonably robust.

This second method of incorporating $m(x|\pi)$ (or $m(x|\pi, f)$) is appealing because it seems to avoid the need to put "priors on priors", etc. It also is related to empirical Bayes techniques; indeed, empirical Bayes analysis can be thought of as simply replacing Γ by the prior $\pi^* \in \Gamma$ for which $m(x|\pi)$ is maximized (clearly the degenerate limit of (4.18)). Unfortunately, this second method can give the wrong answer (as can empirical Bayes analysis). A rather startling example of this is given in Bayarri and Berger (1994). Hence we cannot definitively recommend this second method.

4.6. *Optimal Robust Procedures*

Global Bayesian robustness lends itself naturally to defining notions of optimality. Here is an example, from Sivaganesan, Berliner, and Berger (1993).

Example 11.. We observe $X \sim$ Cauchy $(\theta, 1)$. Elicitation yields –0.3, 0.0, 0.3 as the prior quartiles for θ. The usual "inherently robust" prior density for θ would be the Cauchy(0, 0.3) density; call this π_0. Even though one expects considerable inherent robustness in this situation, it is decided to formally consider global robustness with respect to the contamination class of priors Γ, in (4.3), with $\varepsilon = 0.01$.

Suppose now that a credible set, C, for θ, is desired and that it is (conservatively) decided to require that the posterior probability of C satisfy

$$\Pr(\theta \in C|x, \pi) \geq 0.90 \text{ for all } \pi \in \Gamma. \tag{4.19}$$

Under this condition, one can be assured that C is a 90% credible set.

A natural notion of optimality, here, is to define C^* as optimal if C^* has minimal size (e.g., Lebesgue measure) among all C satisfying (4.19). In Sivaganesan, Berliner, and Berger (1993), it is shown how to find such optimal C^* for quite general problems of this type. For the specific case considered here, and when $x = 6$ is observed, the optimal C^* is $C^* = (-1.22, 2.70) \cup (3.56, 8.43)$; note that this is the union of two intervals, one where the likelihood is large and one where the prior is large.

Many other notions of optimality w.r.t. Γ in global robustness are discussed in Berger (1985), Wasserman (1989), Li and Saxena (1990), DasGupta (1991), Meczarski (1991), Basu (1992c), De la Horra and Fernandez (1993, 1994), and Sivaganesan (1993b).

Optimal global robustness is potentially useful. For instance, if C^* in Example 11 is deemed to be a small enough set for practical purposes then, in light of (4.19), one can be quite satisfied. It can even be possible to design the experiment so as to achieve this with high predictive probability (cf, Mukhopadhyay and DasGupta, 1993; and DasGupta and Mukhopadhyay, 1994).

There is a serious danger with some optimality notions, however: the optimal procedure can be terrible from a "real" Bayesian perspective. This is because, as discussed in Section 4.5, it can be important to take $m(x|\pi, f)$ into account. (See, also, Berger, 1985; DasGupta and Studden, 1988; Sivaganesan and Berger, 1993; Zen and DasGupta, 1993; and Bayarri and Berger, 1994.)

Example 10 (continued).. The initial global robustness analysis yielded $(\frac{1}{2}x_1 - 4, \frac{1}{2}x_1 + 6)$ as the range of posterior means for θ_1. Many notions of optimality would suggest that the midpoint of this interval, $\frac{1}{2}x_1 + 1$, is optimal. However, this corresponds to the value $\mu = 2$, which has very low likelihood, $m(\boldsymbol{x}|\mu)$; indeed, we saw that $\mu = 2$ is at the edge of the "95% noninformative prior credible set" for μ. Furthermore, $\frac{1}{2}x_1 + 1$ is well outside the interval of possibilities in (4.17) that was obtained by a "higher level" robustness analysis.

5. COMPUTING

5.1. *Computational Issues*

In discussing creation of classes of likelihoods or priors, it was observed that computational considerations are crucial. Here we briefly review several generally useful computational techniques.

Linearization: It is easy to see that, under mild conditions, $\overline{\psi}$ (see (4.1)) is the solution to

$$0 = \sup_{f \in \mathcal{F}} \sup_{\pi \in \Gamma_f} \int [h(\theta_f) - \overline{\psi}] f(x|\theta_f) \pi(\theta_f) d\theta_f. \tag{5.1}$$

The point here is that maximization over $\psi(\pi, f)$ is a non-linear operation, but it can be converted, via (5.1), to a linear maximization together with a root-finding operation. This can be a useful simplification. (However, if one can theoretically determine the relevant functional extreme points of the class, (5.1) is unnecessary.) Development and discussion of this algorithm can be found in DeRobertis and Hartigan (1981), Lavine (1991b), Lavine, Wasserman, and Wolpert (1993), Wasserman (1992b), and Wasserman and Kadane (1992a). The latter two papers discuss computation of $\overline{\psi}$, via (5.1), in the important case when it is necessary to utilize Monte-Carlo techniques for computation of the integral.

Reweighting: When computing Bayesian integrals via Monte-Carlo techniques, there are opportunities for relatively easy robustness investigations. To take the simplest case, suppose we approximate $\psi(\pi)$ by

$$\psi(\pi) = \int h(\theta) \pi(\theta|x) d\theta$$

$$\cong \frac{\sum_{i=1}^{N} h(\theta^{(i)}) f(x|\theta^{(i)}) \omega_\pi(\theta^{(i)})}{\sum_{i=1}^{N} f(x|\theta^{(i)}) \omega_\pi(\theta^{(i)})}, \tag{5.2}$$

where

$$\omega_\pi(\theta^{(i)}) = \pi(\theta^{(i)})/g(\theta^{(i)}),$$

and $\theta^{(1)}, \ldots, \theta^{(N)}$ is an i.i.d. sample from the "importance function" g. (See Berger, 1985, for background.) Then switching from one prior to another simply requires recomputing the "weights" $\omega_\pi(\theta^{(i)})$, a relatively simple operation. Indeed, a scheme such as this is virtually necessary for efficient maximization of $\psi(\theta)$, since (5.2) provides a well-defined function to maximize over π. (The alternative, of, say, trying to maximize over π with $\psi(\pi)$ being computed anew by numerical integration at each step, is very unstable.) For formal discussion as to when this scheme for maximization is convergent, see Salinetti (1994).

Reweighting schemes are, unfortunately, not useful if too wide a range of π is being considered. This is because the approximation in (5.2) need not be accurate if the weights can be extremely large. If, however, g can be chosen so that $\omega_\pi(\theta) \leq K$ (moderate) for all π under consideration, then (5.2) can be extremely effective.

Reweighting schemes are also possible for more complicated Markov Chain simulation procedures. See Stephens and Smith (1992) for discussion.

5.2. *Interactive Robustness*

In the Introduction, the possibility of using Bayesian robustness to guide the elicitation process was mentioned. Developing computer-interactive methods of doing this is particularly appealing. Ultimately, one could hope to have a robust Bayesian computer package that processed any given partial information, provided the implied range of Bayesian answers (see Moskowitz, 1992, for a description of such a system for discrete problems) and suggested what additional elicitations would be most desirable, if needed to increase robustness. Here is a simple example we are in the process of developing.

Example 12.. Suppose that elicited information, at stage m of the interactive elicitation process for a real-valued parameter θ, will be a set of quantiles $q_1 < q_2 < \ldots < q_m$, with elicited $p_i = \Pr(\theta \in (q_i, q_{i+1}])$, $i = 1, \ldots, m$. (Allowing for uncertainty in the p_i would be an easy modification.) Also, suppose that the prior distribution for θ is felt to be unimodal. Then, at stage m, one has effectively specified the class of priors

$$\Gamma_m(\boldsymbol{q}, \boldsymbol{p}) = \{\text{all unimodal distr. with the given quantiles}\}. \quad (5.3)$$

Suppose $\psi(\pi)$ is the posterior functional of interest (the likelihood is being considered fixed), and that the degree of robustness is reasonably measured by

$$\overline{\psi}_m - \underline{\psi}_m = \sup_{\pi \in \Gamma_m} \psi(\pi) - \inf_{\pi \in \Gamma_m} \psi(\pi).$$

The problem of computing $\overline{\psi}_m$ and $\underline{\psi}_m$ is discussed in Berger and O'Hagan (1988), and O'Hagan and Berger (1988).

Suppose $\overline{\psi}_m - \underline{\psi}_m$ is deemed to be too large, and that additional refinement of the prior is needed. Since we are eliciting in terms of quantiles, this means that a new quantile, q^*, must be chosen, with the associated p^* (for the new interval created) being elicited. Which q^* should be chosen? It is quite natural to make the choice so that the ensuing $\overline{\psi}_{m+1} - \underline{\psi}_{m+1}$ is likely to be smallest; this would make q^* maximally efficient in terms of Bayesian robustness.

A reasonable scheme for implementing this idea is to consider each possible candidate location, b, for q^*. If $q_i < b < q_{i+1}$, one could assume that the "least informative" elicitation will be done, resulting in

$$p(b) = \Pr(\theta \in (q_i, b]) = \frac{(b - q_i)}{(q_{i+1} - q_i)} \cdot p_i.$$

(This just assumes that the prior probability p_i, assigned to $(q_i, q_{i+1}]$, is distributed uniformly over the interval.) Special adjustments have to be made for $b < q_1$ or $b > q_m$. Assuming b and $p(b)$ specify the new quantile, one would have the new class $\Gamma_{m+1}(\{\boldsymbol{q} \cup b\}, \{\boldsymbol{p} \cup p(b)\})$ as in (5.3), and could compute the corresponding range $(\overline{\psi}_{m+1}(b) - \underline{\psi}_{m+1}(b))$. Minimizing over b would yield the quantile that, if elicited, would be

most likely to result in a substantial gain in robustness. (Of course, once $q^* = b$ is chosen, the actual p^* would be elicited; $p(b)$ would not be used.)

Schemes for interactive elicitation could also be developed based on notions of "most sensitive direction in prior space," as discussed in Section 3.2. The difficulty with such an approach is that the optimal direction in which to focus elicitation efforts may not correspond to quantities that are easy to elicit. Hence we prefer to consider the types of allowed elicitations (e.g., quantiles) as being specified in advance, at which point it is probably easier to consider $\overline{\psi} - \underline{\psi}$ directly, rather than look at local sensitivity.

6. FUTURE DIRECTIONS

Many of the theoretical and methodological directions in which Bayesian robustness is developing were discussed in the paper. Rather than attempting to summarize that discussion, it is useful to focus here on the *types* of statistical problems in which Bayesian robustness can be most usefully applied.

Statistical problems fall into several different categories. The most difficult are problems in which it is a challenge to perform any Bayesian analysis whatsoever. For such problems it will inherently be the case that formal Bayesian robustness cannot be investigated; at best, the informal "try a few models and priors" will be done.

The next category consists of those problems in which subjective Bayesian analysis is feasible, but objective (noninformative) Bayesian methods are also available (and perhaps classical methods that are very similar to the objective Bayesian methods). For such problems, subjective Bayesian analysis is typically performed only when the subjective information is quite influential, relative to the information in the data. Until subjective Bayesian methods become more widely used in these problems, the scope for utilization of formal Bayesian robustness methods will be limited. Of course, as mentioned in the Introduction, the capability to routinely augment a subjective analysis with robustness determinations might increase the willingness to use subjective methods. Probably the most immediate contribution of Bayesian robustness for this class of problems is the possibility of using relatively sophisti-

cated automatically robust methods, such as those discussed in Section 2, as an alternative to the standard methodology.

The third category of statistical problems consists of problems for which objective Bayesian methods do not exist (i.e., the answers typically depend significantly on prior opinions) or are to be avoided (see below). One example that we have discussed is precise hypothesis testing, where the prior on the parameter space of the alternative hypothesis always has a significant effect. Other examples are mentioned below.

For problems in this category, it can be argued that robust Bayesian analysis is *required*; since the answer depends strongly on prior opinions, it is important to show that any conclusions are valid over the range of sensible prior opinions. Problems in this category also typically lack sensible classical answers, so that many non-Bayesians are more willing to consider Bayesian approaches to these problems. It is thus this third category of problems that promises to provide the most immediate applications of robust Bayesian theory. A brief, partial listing of these problems follows.

Similarly to precise hypothesis testing, in *Model Selection* from among models of differing complexity, the prior distributions always have a significant effect. The challenge here, for the robust Bayesian approach, is to choose classes of priors that are appropriately "tied together" for the differing models (see Berger and Pericchi, 1993, for discussion of what this means). Simply having unrelated classes of priors is likely to result in uselessly wide ranges of answers (see Berger and Mortera, 1994, for an illustration in a simple setting).

Extrapolation beyond the range of the data inherently involves subjective opinion, and is very non-robust. Hence it is a natural problem in which to consider Bayesian robustness. See Berger and Chen (1993) for an example.

A common aspect of *Meta-Analysis* is the need to relate the various studies or experiments that are to be combined; the protocol, populations studied, and experimental conditions will often vary from study to study, requiring adjustments if the studies are to be combined. In Bayesian analysis, these adjustment factors, which are typically highly subjective, are built in through the prior distributions (cf, DuMouchel and Harris, 1983; Morris and Normand, 1992; and Wolpert and Warren Hicks, 1992). Since the adjustments are typically highly uncertain, robust Bayesian

analysis is natural. (See Berger and Mortera, 1991, for study of one such situation.)

We have already illustrated robust Bayesian analysis for *Selection Models* or *Weighted Distributions*. Because the selection or weighting mechanism can have an enormous effect and is often uncertain, there is clear motivation for studying Bayesian robustness in these problems.

Clinical Trials provide a natural domain for various types of Bayesian robustness investigations. The reason is that attention is increasingly being paid to conducting clinical trials in a fashion that is as ethical as possible towards the patients in the trial. There are two aspects of this that are particularly relevant to Bayesian robustness. First, the trials may assign patients to treatments in a partially non-random way that involves medical opinion. Second, prior opinion may be used to allow the trial to stop earlier, not only because of the effect of the additional information, but also because Bayesian sequential trials will naturally stop earlier (since repeated looks at the data are not penalized). In both cases, there are typically a variety of prior opinions that must be taken into account, and so some type of Bayesian robustness investigation is needed. For discussion and examples, see Berry, Wolf, and Sack (1992), Carlin and Louis (1993), Carlin, Chaloner, Louis, and Rhame (1993), Sedransk (1993), and Kadane (1994).

Group Decision Making is a related domain in which there naturally exist a variety of prior opinions. Group decision making often begins by seeing if there is a possible action that is simultaneously optimal for all members of the group. This would involve a type of robust Bayesian computation. Only if there is not an optimal answer, in this sense, would more involved group decision making techniques be utilized. (Note, however, that it is not necessarily correct to behave in this way; indeed, it is easy to construct examples where every member of the group initially thinks that a certain action is optimal but that, after sharing information, a different action is seen to be optimal. Ideally, therefore, complete information sharing should be done before applying the robust Bayesian methods.) See Genest and Zidek (1986) and Van Eeden and Zidek (1994) for discussion and references.

REFERENCES

Angers, J. F. (1992). Use of the student-t prior for the estimation of normal means: a computational approach. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 567–576.

Angers, J. F. and Berger, J. O. (1991). Robust hierarchical Bayes estimation of exchangeable means. *Canadian J. Statist.* **19**, 39–56.

Angers, J. F. and Delampady, M. (1992). Hierarchical Bayesian curve fitting and smoothing. *Canadian J. Statist.* **20**, 35–49.

Angers, J. F., MacGibbon, B. and Wang, S. (1992). A robust Bayesian approach to the estimation of intra-block exchangeable normal means with applications. *Tech. Rep.* **92-7**, Université de Montréal.

Basu, S. (1992a). Variations of posterior expectations for symmetric unimodal priors in a distribution band. *Tech. Rep.* **214**, Department of Statistics and Applied Probability, University of California, Santa Barbara.

Basu, S. (1992b). Ranges of posterior probability: symmetry, unimodality and the likelihood. *Tech. Rep.* **215**, University of California, Santa Barbara.

Basu, S. (1992c). A new look at Bayesian point null hypothesis testing: HPD sets, volume minimizing sets, and robust Bayes. *Tech. Rep.* **216**, University of California, Santa Barbara

Basu, S. and DasGupta, A. (1992). Bayesian analysis with distribution bands: the role of the loss function. *Tech. Rep.* **208**, University of California, Santa Barbara.

Basu, S. and Jammalamadaka, S. R. (1993). Local posterior robustness with parametric priors: maximum and average sensitivity. *Tech. Rep.* **239**, University of California, Santa Barbara.

Basu, S., Jammalamadaka, S. R., and Liu, Wei (1993a). Local posterior robustness: total derivatives and functional derivatives. *Tech. Rep.* **239**, University of California, Santa Barbara.

Basu, S. Jammalamadaka, S. R., and Liu, W. (1993b). Qualitative robustness and stability of the posterior distributions and posterior quantities. *Tech. Rep.* **238**, University of California, Santa Barbara.

Bayarri, M. J. and Berger, J. (1993a). Robust Bayesian analysis of selection models. *Tech. Rep.* **93-6**, Purdue University, W. Lafayette.

Bayarri, M. J. and Berger, J. (1993b). Robust Bayesian bounds for outlier detection. *Proceedings of the* 4^{th} *International Meeting of Statistics in the Basque Country-IMSIBAC4* (M. L. Puri and J. P. Vilaplana, eds.). Amsterdam: North-Holland.

Bayarri, M. J. and Berger, J. (1994). Applications and limitations of robust Bayesian bounds and type II MLE. *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer, 121–134.

Bayarri, M. J. and DeGroot, M. (1992). A BAD view of weighted distributions and selection models. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 17–34.

Berger, J. (1984). The robust Bayesian viewpoint. *Robustness of Bayesian Analysis* (J. B. Kadane, ed.). Amsterdam: North-Holland, 63–124, (with discussion).

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* New York: Springer–Verlag.

Berger, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Planning and Inference* **25**, 303–328.

Berger, J. (1992). A comparison of minimal Bayesian tests of precise hypotheses. *Rassegna di Metodi Statistici ed Applicazioni* **7**, Pitagora Editrice, Bologna, 43–78

Berger, J. O. and Berliner, L. M. (1986). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. *Ann. Statist.* **14**, 461–486.

Berger, J. and Bernardo, J. M. (1992). On the development of the reference prior method. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 35–60.

Berger, J. O., Brown, L. and Wolpert, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. *Ann. Statist.* (to appear).

Berger, J. O. and Chen, M. H. (1993). Determining retirement patterns: prediction for a multinomial distribution with constrained parameter space. *The Statistician* **42**, 427–443.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.* **2**, 317–352, (with discussion).

Berger, J. O. and Jefferys, W. (1992). The application of robust Bayesian analysis to hypothesis testing and Occam's razor. *J. It. Statist. Soc.* **1**, 17–32.

Berger, J. O. and Moreno, E. (1994). Bayesian robustness in bidimensional models: prior independence. *J. Statist. Planning and Inference.* (to appear).

Berger, J. O. and Mortera, J. (1991). Interpreting the stars in precise hypothesis testing. *Internat. Statist. Rev.* **59**, 337–353.

Berger, J. O. and Mortera, J. (1994). Robust Bayesian hypothesis testing in the presence of nuisance parameters. *J. Statist. Planning and Inference.* (to appear).

Berger, J. O. and O'Hagan, A. (1988). Ranges of posterior probabilities for unimodal priors with specified quantiles. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 45–66.

Berger, J. O. and Pericchi, L. R. (1993). The intrinsic Bayes factor for model selection and prediction. *Tech. Rep.* **93-43C**, Purdue University, W. Lafayette.

Berger, J. O. and Robert, C. (1990). Subjective hierarchical Bayes estimation of a multivariate normal mean: on the frequentist interface. *Ann. Statist.* **18**, 617–651.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of significance levels and evidence. *J. Amer. Statist. Assoc.* **82**, 112–122, (with discussion).

Berliner, L. M. and MacEachern, S. N. (1993). Examples of inconsistent Bayes procedures based on observations on dynamical systems. *Statistics and Probab. Letters* **17**, 355–360.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. A* **41**, 113–147, (with discussion).

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.

Berry, D., Wolff, M. C. and Sack, D. (1992). Public health decision making: a sequential vaccine trial. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 79–96.

Betrò, B., Meczarski, M., and Ruggeri, F. (1994). Robust Bayesian analysis under generalized moment conditions. *J. Statist. Planning and Inference*, (to appear).

Boratyńska, A. (1991). On Bayesian robustness with the ε–contamination class of priors. *Tech. Rep.*, University of Warsaw

Boratyńska, A. and Zieliński, R. (1991). Infinitesimal Bayes robustness in the Kolmogorov and the Lévy metrics. *Tech. Rep.*, University of Warsaw

Bose, S. (1990). *Bayesian Robustness with Shape-constrained Priors and Mixtures of Priors*. Ph.D. Thesis, Purdue University.

Bose, S. (1993). Bayesian robustness with mixture classes of priors. *Tech. Rep.* **93-1**, George Washington University.

Bose, S. (1994). Bayesian robustness with more than one class of contaminations. *J. Statist. Planning and Inference*, (to appear).

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

Box, G. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Brunner, L. and Lo, A. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.* **17**, 1550–1566.

Cano, J. A. (1993). Robustness of the posterior mean in normal hierarchical models. *Comm. Statist. Theory and Methods* **22**, 1999–2014.

Cano, J. A., Hernández, A. and Moreno, E. (1985). Posterior measures under partial prior information. *Statistica* **2**, 219–230.

Carlin, B. P., Chaloner, K. M., Louis, T. A., and Rhame, F. S. (1993). Elicitation, monitoring, and analysis for an AIDS clinical trial. *Tech. Rep.*, University of Minnesota

Carlin, B. P. and Louis, T. A. (1993). Identifying prior distributions that produce specific decisions, with application to monitoring clinical trials. *Bayesian Inference in Econometrics and Statistics*, (to appear).

Carlin, B. P., and Polson, N. G. (1991). An expected utility approach to influence diagnostics. *J. Amer. Statist. Assoc.* **86**, 1013–1021.

Casella, G. and Berger, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* **82**, 106–111, (with discussion).

Casella, G. and Wells, M. (1991). Noninformative priors for robust Bayesian inference. *Tech. Rep.*, Cornell University

Chib, S., Osiewalski, J. and Steel, M. F. J. (1991). Posterior inference on the degrees of freedom parameter in multivariate-t regression models. *Economics Letters* **37**, 391–397.

Coolen, F. P. A. (1993). Imprecise conjugate prior densities for the one-parameter exponential model. *Statistics and Probability Letters* **16**, 337–342.

Cuevas, A. and Sanz, P. (1988). On differentiability properties of Bayes Operators. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 569–577.

DasGupta, A. (1991). Diameter and volume minimizing confidence sets in Bayes and classical problems. *Ann. Statist.* **19**, 1225–1243.

DasGupta, A. and Delampady, M. (1990). Bayesian testing with symmetric and unimodal priors. *Tech. Rep.* **90-47**, Purdue University.

DasGupta, A. and Mukhopadhyay, S. (1994). Uniform and subuniform posterior robustness: the sample size problem. *J. Statist. Planning and Inference,* (to appear).

DasGupta, A. and Studden, W. J. (1988a). Frequentist behavior of robust Bayes procedures: new applications of the Lehmann-Wald minimax theory to a novel geometric game. *Tech. Rep.* **88-36C**, Purdue University.

DasGupta, A. and Studden, W. J. (1988b). Robust Bayesian analysis in normal linear models with many parameters. *Tech. Rep.* **88-14**, Purdue University.

DasGupta, A. and Studden, W. J. (1989). Frequentist behavior of robust Bayes estimates of normal means. *Statistics and Decisions* **7**, 333–361.

DasGupta, A. and Studden, W. J. (1991). Robust Bayesian experimental designs in normal linear models. *Ann. Statist.* **19**, 1244–1256.

Datta, G. and Lahiri, P. (1992). Robust hierarchical Bayes estimation of small area characteristics in presence of covariates. *Tech. Rep.* **92-28**, University of Georgia.

Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika* **60**, 664–667.

De la Horra, J. and Fernandez, C. (1993). Bayesian analysis under ε-contaminated priors: a trade-off between robustness and precision. *J. Statist. Planning and Inference* **38**, 13–30.

De la Horra, J. and Fernández, C. (1994). Bayesian robustness of credible regions in the presence of nuisance parameters. *Communications in Statistics* **23**. (to appear).

Delampady, M. (1989a). Lower bounds on Bayes factors for interval null hypotheses. *J. Amer. Statist. Assoc.* **84**, 120–124.

Delampady, M. (1989b). Lower bounds on Bayes factors for invariant testing situations. *J. Multivariate Analysis* **28**, 227–246.

Delampady, M. and Berger, J. (1990). Lower bounds on Bayes factors for Multinomial and chi-squared tests of fit. *Ann. Statist.* **18**, 1295–1316.

Delampady, M. and Dey, D. (1994). Bayesian robustness for multiparameter problems. *J. Statist. Planning and Inference,* (to appear).

DeRobertis, L. (1978). The use of partial prior knowledge in Bayesian inference. Ph.D. Thesis, Yale University.

DeRobertis, L. and Hartigan, J. A. (1981). Bayesian inference using intervals of measures. *Ann. Statist.* **1**, 235–244.

Dette, H. and Studden, W. J. (1994). A geometric solution of the Bayesian E-optimal design problem. *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer, 157–170.

Dey, D. and Birmiwal (1991). Robust Bayesian analysis using entropy and divergence measures. *Tech. Rep.*, University of Connecticut

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1–67.

Doss, H. (1994). Bayesian estimation for censored data: an experiment in sensitivity analysis. *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer, 171–182.

Draper, D. (1992). Assessment and propagation of model uncertainty. *Tech. Rep.*, University of California

Drummey, K. W. (1991). *Robust Bayesian Estimation in the Normal, Gamma, and Binomial Probability Models: a Computational Approach.* Ph.D. Thesis, University of Maryland.

DuMouchel, W. and Harris, J. (1983). Bayes methods for combining the results of cancer studies in human and other species. *J. Amer. Statist. Assoc.* **78**, 293–315, (with discussion).

Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Rev.* **70**, 193–242.

Eichenauer-Herrmann, J. and Ickstadt, K. (1993). A saddle point characterization for classes of priors with shape-restricted densities. *Statistics and Decisions* **11**, 175–179.

Fan, T. H. and Berger, J. O. (1990). Exact convolution of t–distributions, with application to Bayesian inference for a normal mean with t prior distributions. *J. Statist. Computation and Simulation* **36**, 209–228.

Fan, T. H. and Berger, J. O. (1992). Behavior of the posterior distribution and inferences for a normal mean with t prior distributions. *Statistics and Decisions* **10**, 99-120.

Ferguson, T.S., Phadia, E.G., and Tiwari, R.C. (1992). Bayesian nonparametric inference. *Current Issues in Statistical Inference: Essays in Honor of D. Basu* (M. Ghosh, and P. K. Pathak, eds.). Hayward CA.: IMS, 127–150.

Fernández, C., Osiewalski, J., and Steel, M.F.J. (1993). Marginal equivalence in ν-spherical models. *Tech. Rep.*, Universidad Autónoma de Madrid

Fortini, S. and Ruggeri,F. (1990). Concentration function in a robust Bayesian framework. *Tech. Rep.* **90.6**, CNR-IAMI, Milano.

Fortini, S. and Ruggeri, F. (1992). On defining neighborhoods of measures through the concentration function. *Sankhyā A.* (to appear).

Fortini, S. and Ruggeri, F. (1994). Concentration functions and Bayesian robustness. *J. Statist. Planning and Inference*, (to appear).

Fougere, P. (Ed.) (1990). *Maximum Entropy and Bayesian Methods.* Dordrecht: Kluwer

Gasparini, M. (1990). Nonparametric Bayes estimation of a distribution function with truncated data. *Tech. Rep.* **182**, University of Michigan.

Geisser, S. (1992). Bayesian perturbation diagnostics and robustness. *Bayesian Analysis in Statistics and Econometrics* (P. K. Goel and N. S. Iyengar, eds.). Berlin: Springer, 289–302.

Gelfand, A. and Dey, D. (1991). On Bayesian robustness of contaminated classes of priors. *Statistics and Decisions* **9**, 63–80.

Genest, C., and Zidek, J. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statist. Sci.* **1**, 114–135.

Geweke, J. (1992). Priors for macroeconomic time series and their application. *Proceedings of the Conference on Bayes Methods and Unit Roots*, a special issue of *Econometric Theory*, (to appear).

Ghosh, J. K., Ghosal, S., and Samanta, T. (1994). Stability and convergence of the posterior in non-regular problems. *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer, 183–200.

Ghosh, M. (1993). Inconsistent maximum likelihood estimators for the Rasch model. *Tech. Rep.*, University of Florida.

Ghosh, M. (1994). On some Bayesian solutions of the Neyman-Scott problem. *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer, 267–276.

Girón, F. J. and Ríos, S. (1980). Quasi-Bayesian behavior: a more realistic approach to decision making? *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 17–38.

Goldstein, M. and Wooff, D.A. (1994). Robustness measures for Bayes linear analyses. *J. Statist. Planning and Inference*, (to appear).

Good, I. J. (1983a). *Good Thinking: The Foundations of Proability and Its Applications.* University of Minnesota Press, Minneapolis, MN.

Good, I. J. (1983b). The robustness of a hierarchical model for multinomials and contingency tables. *Scientific Inference, Data Analysis and Robustness* (G. E. P. Box, T. Leonard and C. F. Wu, eds.). New York: Academic Press.

Good, I. J. and Crook, J. F. (1987). The robustness and sensitivity of the mixed-Dirichlet Bayesian test for 'independence' in contingency tables. *Ann. Statist.* **15**, 694–711.

Goutis, C. (1991). Ranges of posterior measures for some classes of priors with specified moments. *Tech. Rep.* **70**, University College London.

Gu, C. and Wahba, G. (1993). Smoothing spline ANOVA with component-wise Bayesian confidence intervals. *J. Comput. and Graphical Stat.* **2**, 97–117.

Gustafson, P. and Wasserman, L. (1993). Local sensitivity diagnostics for Bayesian inference. *Tech. Rep.* **574**, Carnegie-Mellon University.

Guttman, I. and Peña, D. (1988). Outliers and influence: evaluation by posteriors of parameters in the linear model. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 631–640.

Guttman, I. and Peña, D. (1993). A Bayesian look at diagnostics in the univariate linear model. *Statistica Sinica* **3**, 367–390.

Hartigan, J.A. (1983). *Bayes Theory*. New York: Springer-Verlag.

Ickstadt, K. (1992). Gamma-minimax estimators with respect to unimodal priors. In *Operations Research '91* (P. Gritzmann, *et al.*, eds.). Heidelberg: Physica-Verlag.

Jaynes, E. T. (1983). *Papers on Probability, Statistics and Statistical Physics*, (R. Rosenkrantz, ed.), Dordrecht: Reidel

Jefferys, W. and Berger, J. (1992). Ockham's razor and Bayesian analysis. *Amer. Scientist* **80**, 64–72.

Kadane, J. (1994). An application of robust Bayesian analysis to a medical experiment. *J. Statist. Planning and Inference*, (to appear).

Kadane, J. B. and Chuang, D. T. (1978). Stable decision problems. *Ann. Statist.* **6**, 1095–1110.

Kass, R. and Greenhouse, J. (1989). Investigating therapies of potentially great benefit: A Bayesian perspective. Comments on "Investigating therapies of potentially great benefit: ECMO," by J. H. Ware. *Statist. Sci.* **4**, 310–317.

Kass, R. E. and Raftery, A. (1992). Bayes factors and model uncertainty. *Tech. Rep.* **571**, Carnegie-Mellon University.

Kass, R. E., and Slate, E. H. (1992). Reparametrization and diagnostics of posterior non-normality. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 289–306.

Kass, R. E., Tierney, L. and Kadane, J. B. (1980). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika* **76**, 663–674.

Kass, R. E. and Wasserman, L. (1993). Formal rules of selecting prior distributions: a review and annotated bibliography. *Tech. Rep.*, Carnegie-Mellon University.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators (theory and methods). *J. Amer. Statist. Assoc.* **72**, 789–827.

Kohn, R. and Ansley, C. F. (1988). The equivalence between Bayesian smoothness priors and optimal smoothing for function estimation. In *Bayesian Analysis of Time Series and Dynamic Models* (J. C. Spall, ed.). New York: Marcel Dekker.

Kouznetsov, V. P. (1991). *Interval Statistical Models*. Moscow: Radio and Communication.

Laplace, P. S. (1812). *Theorie Analytique des Probabilites*. Courcier, Paris.

Lavine, M. (1989). The boon of dimensionality: how to be a sensitive multidimensional Bayesian. *Tech. Rep.* **89-14**, Duke University.

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 with multidimensional spaces. *J. Amer. Statist. Assoc.* **86**, 400–403.

Lavine, M. (1992a). Sensitivity in Bayesian statistics: the prior and the likelihood. *J. Amer. Statist. Assoc.* **86**, 396–399.

Lavine, M. (1992b). Some aspects of Polya tree distributions for statistical modelling. *Ann. Statist.* **20**, 1222–1235.

Lavine, M. (1992c). A note on bounding Monte Carlo variances. *Comm. Statist. Theory and Methods* **21**, 2855–2860.

Lavine, M. (1992d). Local predictive influence in Bayesian linear models with conjugate priors. *Commun. Statist.-Simulation* **21**, 269–283.

Lavine, M. (1994). An approach to evaluating sensitivity in Bayesian regression analysis. *J. Statist. Planning and Inference*. (to appear).

Lavine, M. and Wasserman, L. (1992). Can we estimate N? *Tech. Rep.* **546**, Carnegie-Mellon University.

Lavine, M., Wasserman, L. and Wolpert, R. (1991). Bayesian inference with specified prior marginals. *J. Amer. Statist. Assoc.* **86**, 964–971.

Lavine, M., Wasserman, L., and Wolpert, R. (1993). Linearization of Bayesian robustness problems. *J. Statist. Planning and Inference* **37**, 307–316.

Leamer, E. E. (1982). Sets of posterior means with bounded variance prior. *Econometrica* **50**, 725–736.

Lenk, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.* **83**, 509–516.

Leonard T. (1978). Density estimation, stochastic processes, and prior information. *J. Roy. Statist. Soc. B* **40**, 113–146.

Li, Y. and Saxena, K. M. L. (1990). Optimal robust Bayesian estimation. *Tech. Rep.* **4**, University of Nebraska.

Liseo, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika* **80**, 295–304.

Liseo, B., Petrella, L., and Salinetti, G. (1993). Block unimodality for multivariate Bayesian robustness. *J. Ital. Statist. Soc*, (to appear).

Lo, A.Y. and Weng, C. S. (1989). On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. *Ann. Inst. Statist. Math.* **41**, 227–245.

Lucas, T. W. (1992). When is conflict normal? *Tech. Rep.*, The Rand Corporation.

Meczarski, M. (1991). Stable Bayesian estimation in the Poisson model: a nonlinear problem. *Tech. Rep.* **91.14**, CNR-IAMI, Milano.

Meczarski, M. and Zieliński, R. (1991). Stability of the Bayesian estimator of the Poisson mean under the inexactly specified gamma prior. *Statist. Prob. Letters* **12**, 329–333.

Meng, X. L. (1994). Posterior predictive P-values. *Ann. Statist.*, (to appear).

McCulloch, R. E. (1989). Local model influence. *J. Amer. Statist. Assoc.* **84**, 473–478.

Moreno, E. and Cano, J. A. (1989). Testing a point null hypothesis: Asymptotic robust Bayesian analysis with respect to the priors given on a subsigma field. *Int. Statist. Rev.*, 57, 221–232.

Moreno, E. and Cano, J. A. (1991). Robust Bayesian analysis for ε-contaminations partially known. *J. Roy. Statist. Soc. B* **53**, 143–155.

Moreno, E. and Cano, J. A. (1992). Classes of bidimensional priors specified on a collection of sets: Bayesian robustness. *Tech. Rep.*, Universidad de Granada.

Moreno, E. and Pericchi, L. R. (1990). Sensitivity of the Bayesian analysis to the priors: structural contaminations with specified quantiles of parametric families. *Actas III Cong. Latinoamericano Probab. Estad. Mat.*, 143–158.

Moreno, E. and Pericchi, L. R. (1991). Robust Bayesian analysis for ε-contaminations with shape and quantile constraints. *Proc. Fifth Inter. Symp. on Applied Stochastic Models and Data Analysis*. World Scientific Publ., 454–470.

Moreno, E. and Pericchi, L. R. (1992a). Subjetivismo sin dogmatismo: análisis Bayesiano robusto (with discussion). *Estadist. Española* **34**, 5–60.

Moreno, E. and Pericchi, L. R. (1992b). Bands of probability measures: A robust Bayesian analysis. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 707–714.

Moreno, E. and Pericchi, L. R. (1993a). Bayesian robustness for hierarchical ε-contamination models. *J. Statist. Planning and Inference* **37**, 159–168.

Moreno, E. and Pericchi, L. R. (1993b). Precise measurement theory: robust Bayesian analysis. *Tech. Rep.*, Universidad de Granada.

Morris, C. N. and Normand, S. L. (1992). Hierarchical models for combining information and for meta-analyses. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 321–344.

Moskowitz, H. (1992). Multiple-criteria robust interactive decision analysis for optimizing public policies. *Eur. J. Oper. Res.* **56**, 219–236.

Mukhopadhyay, S. and DasGupta, A. (1993). Uniform approximation of Bayes solutions and posteriors: frequentistly valid Bayes inference. *Tech. Rep.* **93-12C**, Purdue University.

O'Hagan, A. (1988). Modelling with heavy tails. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 345–360.

O'Hagan, A. (1990). Outliers and credence for location parameter inference. *J. Amer. Statist. Assoc.* **85**, 172–176.

O'Hagan, A. (1994). Robust modelling for asset management. *J. Statist. Planning and Inference*, (to appear).

O'Hagan, A. and Berger, J. O. (1988). Ranges of posterior probabilities for quasi-unimodal priors with specified quantiles. *J. Amer. Statist. Assoc.* **83**, 503–508.

Osiewalski, J. and Steel, M. F. J. (1993a). Robust Bayesian inference in ℓ_q-spherical models. *Biometrika*, (to appear).

Osiewalski, J. and Steel, M. F. J. (1993b). Robust Bayesian inference in elliptical regression models. *J. Econometrics* **57**, 345–363.

Osiewlaski, J. and Steel, M. F. J. (1993c). Bayesian marginal equivalence of elliptical regression models. *Journal of Econometrics*, (to appear).

Peña, D. and Guttman, I. (1993). Comparing probabilistic methods for outlier detection in linear models. *Biometrika* **80**, 603–610.

Peña, D. and Tiao, G. C. (1992). Bayesian outlier functions for linear models. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 365–388.

Pericchi, L. R. and Nazaret, W. (1988). On being imprecise at the higher levels of a hierarchical linear model. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 361–376.

Pericchi, L. R. and Pérez, M. E. (1994). Posterior robustness with more than one sampling model. *J. Statist. Planning and Inference.*, (to appear).

Pericchi, L. R. and Walley P. (1991). Robust Bayesian credible intervals and prior ignorance. *Internat. Statist. Rev.* **58**, 1–23.

Pettit, L. I. (1988). Bayes methods for outliers in exponential samples. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 731–740.

Pettit, L. (1992). Bayes factors for outlier models using the device of imaginary observations. *J. Amer. Statist. Assoc.* **87**, 541–545.

Poirier, D. J. (1988). Bayesian diagnostic testing in the general linear normal regression model. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 725–732.

Polasek, W. (1985). Sensitivity analysis for general and hierarchical linear regression models. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.). Amsterdam: North-Holland

Polasek, W. and Pötzelberger, K. (1988). Robust Bayesian analysis in hierarchical models. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 377–394.

Polasek, W. and Pötzelberger, K. (1994). Robust Bayesian methods in simple ANOVA models. *J. Statist. Planning and Inference*, (to appear).

Pötzelberger, K. and Polasek, W. (1991). Robust HPD-regions in Bayesian regression models. *Econometrica* **59**, 1581–1590.

Rao, C. R. (1985). Weighted distributions arising out of methods of ascertainment: what population does a sample represent? *A Celebration of Statistics: The ISI Centenary Volume* (A. G. Atkinson and S. Fienberg, eds.). Berlin: Springer.

Regazzini, E. (1992). Concentration comparisons between probability measures. *Sankhyā B* **54**, 129–149.

Ríos-Insúa, D. (1990). *Sensitivity Analysis in Multiobjective Decision Making*. Berlin: Springer.

Ríos-Insúa, D. (1992). Foundations for a robust theory of decision making: the simple case. *Test* **1**, 69–78.

Ríos-Insúa, D. and French, S. (1991). A framework for sensitivity analysis in discrete multiobjective decision making. *Eur. J. Oper. Res.* **54**, 176–190.

Ríos-Insúa, D. and Martín, J. (1994). Robustness issues under precise beliefs and preferences. *J. Statist. Planning and Inference*, (to appear).

Robert, C. (1992). *L'Analyse Statistique Bayesienne*. Paris: Economica.

Ruggeri, F. (1990). Posterior ranges of functions of parameters under priors with specified quantiles. *Comm. Statist. Theory and Methods* **19**, 127–144.

Ruggeri, F. (1991). Robust Bayesian analysis given a lower bound on the probability of a set. *Comm. Statist. Theory and Methods* **20**, 1881–1891.

Ruggeri, F. (1992). Bounds on the prior probability of a set and robust Bayesian analysis. *Theory of Probability and Its Applications* **37**.

Ruggeri, F. and Wasserman, L. (1991). Density based classes of priors: Infinitesimal properties and approximations. *Tech. Rep.* **528**, Carnegie-Mellon University.

Ruggeri, F. and Wasserman, L. (1993). Infinitesimal sensitivity of posterior distributions. *Canadian J. Statist.* **21**, 195–203.

Salinetti, G. (1994). Stability of Bayesian decisions. *J. Statist. Planning and Inference*, (to appear).

Sansó, B. and Pericchi, L. R. (1992). Near ignorance classes of log-concave priors for the location model. *Test* **1**, 39–46.

Sedransk, N. (1993). Admissibility of treatment. *Clinical Trials: Bayesian Methods and Ethics* (J. Kadane ed.). New York: Wiley.

Sivaganesan, S. (1988). Range of posterior measures for priors with arbitrary contaminations. *Comm. Statst.* **17**, 1591–1612.

Sivaganesan, S. (1989). Sensitivity of posterior mean to unimodality preserving contaminations. *Statistics and Decisions* **7**, 77–93.

Sivaganesan, S. (1990). Sensitivity of some standard Bayesian estimates to prior uncertainty – a comparison. *J. Statist. Planning and Inference* **27**, 85–103.

Sivaganesan, S. (1991). Sensitivity of some posterior summaries when the prior is unimodal with specified quantiles. *Canadian J. Statist.* **19**, 57–65.

Sivaganesan, S. (1992). An evaluation of robustness in binomial empirical Bayes testing. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 783–790.

Sivaganesan, S. (1993a). Range of the posterior probability of an interval for priors with unimodality preserving contaminations. *Ann. Inst. of Statist. Math.* **45**, 187–199.

Sivaganesan, S. (1993b). Optimal robust sets for a density bounded class. *Tech. Rep.*, University of Cincinnati.

Sivaganesan, S. (1993c). Robust Bayesian diagnostics. *J. Statist. Planning and Inference* **35**, 171–188.

Sivaganesan, S. (1994). Bounds on posterior expectations for density bounded classes with constant bandwidth. *J. Statist. Planning and Inference.*, (to apppear).

Sivaganesan, S. and Berger, J. (1989). Ranges of posterior measures for priors with unimodal contaminations. *Ann. Statist.* **17**, 868–889.

Sivaganesan, S. and Berger, J. (1993). Robust Bayesian analysis of the binomial empirical Bayes problem. *Canadian J. Statist.* **21**, 107–119.

Sivaganesan, S., Berliner, L. M., and Berger, J. (1993). Optimal robust credible sets for contaminated priors. *Statist. and Probab. Letters*, (to appear).

Smith, A. F. M. (1983). Bayesian approaches to outliers and robustness. *Specifying Statistical Models* (J. P. Florens et.al. eds.). Berlin: Springer.

Spiegelhalter, D. J. (1985). Exact Bayesian inference on the parameters of a Cauchy distribution with vague prior information. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 743–749.

Srinivasan, C. and Truszczynska, H. (1990). Approximation to the range of a ratio linear posterior quantity based on Frechet derivative. *Tech. Rep.* **289**, University of Kentucky.

Srinivasan, C. and Truszczynska, H. (1993). Ranges of non-linear posterior quantities. *Ann. Statist.*, (to appear).

Stephens, D. A. and Smith, A. F. M. (1992). Sampling-resampling techniques for the computation of posterior densities in normal means problems. *Test* **1**, 1–18.

Tamura, H. (1992). Robust Bayesian auditing. *Tech. Rep.*, University of Washington.

VanEeden, C. and Zidek, J. V. (1994). Group Bayes estimation of the exponential mean: a retrospective view of the Wald theory. *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer, 35–50.

Verdinelli, I. and Wasserman, L. (1991). Bayesian analysis of outlier problems using the Gibbs sampler. *Statist. Computing* **1**, 105–117.

Vidakovic, B. (1992). A study of properties of computationally simple rules in estimation problems. Ph.D. Thesis, Purdue University.

Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM.

Walley, P. (1991). *Statistical Reasoning With Imprecise Probabilities.* London: Chapman and Hall.

Wasserman, L. (1989). A robust Bayesian interpretation of likelihood regions. *Ann. Statist.* **17**, 1387–1393.

Wasserman, L. (1990). Prior envelopes based on belief functions. *Ann. Statist.* **18**, 454–464.

Wasserman, L. (1992a). The conflict between improper priors and robustness. *Tech. Rep.* **559**, Carnegie-Mellon University.

Wasserman, L. (1992b). Recent methodological advances in robust Bayesian inference. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 483–502.

Wasserman, L. (1992c). Invariance properties of density ratio priors. *Ann. Statist.* **20**, 2177–2182.

Wasserman, L. and Kadane, J. (1990). Bayes' theorem for Choquet capacities. *Ann. Statist.* **18**, 1328–1339.

Wasserman, L. and Kadane, J. B. (1992a). Computing bounds on expectation. *J. Amer. Statist. Assoc.* **87**, 516–522.

Wasserman, L. and Kadane, J. (1992b). Symmetric upper probabilities. *Ann. Statist.* **20**, 1720–1736.

Wasserman, L. and Seidenfeld, T. (1994). The dilation phenomenon in robust Bayesian inference. *J. Statist. Planning and Inference*, (to appear).

Weiss, R. E. (1992). Influence diagnostics with the Gibbs sampler. *Computing Science and Statistics: Proceedings of the* 24^{th} *Symposium on the Interface* (H. J. Newton, ed.). Alexandria: ASA.

Weiss, R. (1993). Identification of outlying perturbations. *Tech. Rep.* **594**, University of Minnesota.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74**, 646–648.

West, M. (1992). Modeling with mixtures. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 503–524.

West, M. and Harrison, J. (1989). *Bayesian Forecasting and Dynamic Models*. Berlin: Springer.

Wolpert, R. L. and Warren Hicks, W. J. (1992). Bayesian hierarchical logistic models for combining field and laboratory survival data. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 525–546.

Zellner, A. (1976). Bayesian and non-Bayesian analysis of regression models with multivariate student-t error terms. *J. Amer. Statist. Assoc.* **71**, 400–405.

Zen, M. M. and DasGupta, A. (1993). Estimating a binomial parameter: is robust Bayes real Bayes? *Statistics and Decisions* **11**, 37–60.

Zidek, J. V. and Weerahandi, S. (1992). Bayesian predictive inference for samples from smooth processes. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 547–566.

DISCUSSION

ELÍAS MORENO *(Universidad de Granada)*

This is an excellent and stimulating paper on Robust Bayesian Analysis and James Berger is to be congratulated for it. I will focus my comments on two points, (i) elicitation of contamination classes, and (ii) limitations of global robustness.

1. *Elicitation of contamination classes.* The ε-contamination class of priors

$$\Gamma = \{\pi(\theta) : (\theta) = (1-\varepsilon)\pi_0(\theta) + q(\theta), \quad q \in Q\},$$

has been considered in Bayesian statistics to model uncertainty on the prior distribution in the following scenario. A prior $\pi_0(\theta)$ for θ is elicited. Some prior beliefs are accurately stated, for instance the probabilities of some sets $\{C_i, i \geq 1\}$ which form a partition of the parameter space Θ,

$$Pr\{\theta \in C_i\} = \alpha_i, \quad i \geq 1, \tag{1}$$

where $\alpha_i = \int_{C_i} \pi_0(d\theta), i \geq 1$. A constant $\varepsilon, 0 < \varepsilon < 1$, reflecting our overall degree of uncertainty on the form of $\pi_0(\theta)$ on the sets $C_i, 1 \geq 1$, is specified. Finally, a class Q of possible priors compatible with (1), is then chosen. In case of condition (1), a sensible class Q would be

$$Q = \left\{ q(\theta) : \int_{C_i} q(\theta) d\theta = \alpha_i, \quad i \geq 1 \right\}.$$

For a given quantity of interest $\varphi(\theta)$, and a sample observation x robustness of $E^{\pi}(\varphi(\theta)|x)$ as π ranges over Γ is then studied.

When Γ is not *a posteriori* robust, the efforts in eliciting more prior information on θ have the result of reducing the class Q under consideration, and robustness of the corresponding class Γ is again investigated. For instance, in Berger and O'Hagan (1988), O'Hagan and Berger (1988), Moreno and Pericchi (1991) shape constraints are added to conditions (1) to define a smaller class Q.

In this interactive process of refinement of the class Γ until posterior robustness is achieved, ε has not played a relevant role even when the size of the class Γ depends on it, and also when to add shape conditions to the quantile constraints results in a quite tractable analysis in one dimensional problems but possibly intractable in a multidimensional setting. To do something to reduce the class in the latter context is absolutely necessary.

However, ε might be of interest in this process by simply observing that we are less confident of the form of the tails of $\pi_0(\theta)$ than we are of the body. In other words, we are able to make more accurate statements about the form of $\pi_0(\theta)$ in its central part than we are in its tails. This might result in robust answers without more refinement of the class Q. Note that ε should then be a function of θ capable of discriminating between uncertainty in the tails and body of $\pi_0(\theta)$.

The question arising is what kind of $\varepsilon(\theta)$'s are appropriated to model a different degree of confidence in different parts of $\pi_0(\theta)$, but retaining the prior beliefs given in (1). The next theorem addresses this issue.

Theorem. *(Moreno, Martínez, and Cano, 1993)*

Let $(\Theta, \mathcal{A})$ be the measurable parameter space of the statistical problem, and let $\mathcal{B}$ be a sub sigma field of $\mathcal{A}$. Consider the class

$$\Gamma_{\mathcal{B}} = \{\pi(\theta) : \pi(\theta) = (1 - \varepsilon(\theta))\pi_0(\theta) + \varepsilon(\theta)q(\theta), \quad q \in Q_{\mathcal{B}}\}$$

where $\varepsilon(\theta)$, $0 \leq \varepsilon(\theta) \leq 1$, *is an* $\mathcal{A}$*-measurable function, and*

$$Q_{\mathcal{B}} = \left\{ q(\theta) : \int_B q(\theta)d\theta = \int_B \pi_0(\theta), \quad B \in \mathcal{B} \right\}.$$

Then, (i) if $\varepsilon(\theta)$ *is a* $\mathcal{B}$ *measurable function it follows that any* $\pi \in \Gamma_{\mathcal{B}}$ *satisfies* $\int_B \pi(\theta)d\theta = \int_B \pi_0(\theta)$ *for any* $B \in \mathcal{B}$. *(ii) Conversely, if* $\mathcal{B}$ *is such that for any set* $A \in \mathcal{A} - \mathcal{B}$ *there exists* $q_1, q_2 \in Q_{\mathcal{B}}$ *such that*

$$\int_A q_1(\theta)d\theta = (\pi_0^{\mathcal{B}})_*(A), \quad \int_A q_2(\theta)d\theta = (\pi_0^{\mathcal{B}})^*(A),$$

where $(\pi_0^{\mathcal{B}})_*, (\pi_0^{\mathcal{B}})^*$ *are the inner and outer measures of* π_0 *with respect to* $\mathcal{B}$ *respectively, and any* $\pi \in \Gamma_{\mathcal{B}}$ *satisfies* $\int_B \pi(\theta)d\theta = \int_B \pi_0(\theta)d\theta$, $B \in \mathcal{B}$, *then* $\varepsilon(\theta)$ *is a* $\mathcal{B}$ *measurable function.*

Corollary. *Let* $\mathcal{B} = \sigma(C_i, i \geq 1)$ *be the sub sigma field of* $\mathcal{A}$ *generated by a partition* $\{C_i, i \geq 1\}$. *Then, any* $\pi \in \Gamma_{\mathcal{B}}$ *satisfies* $\int_B \pi(\theta)d\theta = \int_B \pi_0(\theta)d\theta, B \in \mathcal{B}$, *if and only if* $\varepsilon(\theta)$ *is a* $\mathcal{B}$*-measurable function.* ◁

This theorem means that the elicitor can chose $\varepsilon(\theta)$ in the class of $\mathcal{B}$-measurable functions to express his degrees of uncertainty on different parts of $\pi_0(\theta)$, while keeping fixed the probabilities of the sets of $\mathcal{B}$.

Example 1. Let X be a random variable $N(\theta, 1)$ distributed. Suppose we are interested in testing $H_0 : \theta \leq 0$. It is elicited that the distribution of θ is approximately symmetric in a neighborhood around zero and that the probabilities of the sets $C_1 = (-\infty, -0.9539], C_2 = (-0.9539, 0.9539]$, and $C_3 = [0.9539, \infty)$ are

$$\int_{-\infty}^{-.9539} \pi(\theta)d\theta = 0.25, \int_{-.9539}^{-.9539} q(\theta)d\theta = 0.5,$$
$$\int_{-.9539}^{\infty} \pi(\theta)d\theta = 0.25.$$

The base prior $\pi_0(\theta) = N(0, 2)$ is typically being used and the class Γ_0

$$\Gamma_0 = \{\pi(\theta) : \pi(\theta) = (1 - \varepsilon)\pi_0(\theta) + \varepsilon q(\theta), q \in Q\},$$

where $\varepsilon = 0.2$ and

$$Q = \left\{ \int_{-\infty}^{-.9539} \pi(\theta)d\theta = 0.25, \int_{-.9539}^{-.9539} q(\theta)d\theta = 0.5, \int_{-.9539}^{\infty} \pi(\theta)d\theta = 0.25 \right\}$$

is the usual ε-contamination class.

The posterior imprecision of H_0 with respect to that class for various values of x, is displayed in the second column of Table 1. This imprecision is defined as

$$\Delta_{\Gamma_0} P^{\pi}(H_0|x) = \sup_{\pi \in \Gamma_0} P^{\pi}(H_0|x) - \inf_{\pi \in \Gamma_0} P^{\pi}(H_0|x).$$

If, in the class Γ_0, ε is replaced by $\varepsilon(\theta) = 01_{C_2}(\theta) + 0.51_{C_2^c}(\theta)$, where C_2^c denotes the complement of C_2, only an uncertainty of 0.5 in the tail of $\pi_0(\theta)$ is allowed. Let us denote by $\Delta_{\Gamma_{.5}} P^{\pi}(H_0|x)$ the posterior imprecision of H_0 with respect to the class associated with this $\varepsilon(\theta)$. Values of this imprecision for various observations x are given in the third column of Table 1.

x	$\Delta_{\Gamma_0} P^{\pi}(H_0\|x)$	$\Delta_{\Gamma_{.5}} P^{\pi}(H_0\|x)$
0	0.21	0.13
0.5	0.18	0.12
1.0	0.14	0.09
1.5	0.09	0.04

Table 1. *Posterior imprecisions of* H_0 *for* Γ_0 *and* $\Gamma_{.5}$

Table 1 shows that a significant reduction of posterior imprecision is obtained if only uncertainty in the tails of $\pi_0(\theta)$ is considered, even when this uncertainty is as big as 0.5. Probably the situation considered in the last column is a better reflection of our uncertainty and therefore it is a more realistic measurement of robustness.

2. *Choosing a base prior.* It is usually the case that the prior beliefs we have used to elicit the base prior $\pi_0(\theta)$ are satisfied for more than one distribution having different tail behaviour. The imprecision in the tails from the associated contamination classes can help to choose one.

Example 2. Consider the situation stated in Example 1. The three base priors, Normal, $\pi_{01}(\theta) = N(0, 2)$; Intrinsic (see Berger and Pericchi (1993) for a genesis of it), $\pi_{02}(\theta) = (1 - \exp(-\theta^2/.9174^2))/(2\sqrt{\pi}\theta^2/.9174^2))$; and Cauchy $\pi_{03}(\theta) = C(0, 0.9539)$ have on the set $(-.9539, .9539)$ a rather similar shape, and they also satisfy the accurate prior beliefs stated in (1). Hence, the three classes

$$\Gamma_{0i} = \{\pi(\theta) : \pi(\theta) = (1 - \varepsilon)\pi_{0i}(\theta) + \varepsilon q(\theta), q \in Q\}, i = 1, 2, 3$$

might be considered.

The posterior imprecision of H_0 with respect to these classes for various values of x and ε, are displayed in Table 2.

x	$\Delta_{\Gamma_{01}} P^\pi(H_0\|x)$		$\Delta_{\Gamma_{02}} P^\pi(H_0\|x)$		$\Delta_{\Gamma_{03}} P^\pi(H_0\|x)$	
	$\varepsilon = .2$	$\varepsilon = .5$	$\varepsilon = .2$	$\varepsilon = .5$	$\varepsilon = .2$	$\varepsilon = .5$
0	0.21	0.51	0.23	0.55	0.22	0.53
0.5	0.18	0.47	0.22	0.53	0.20	0.49
1.0	0.14	0.38	0.20	0.48	0.16	0.41
1.5	0.09	0.26	0.15	0.39	0.12	0.31

Table 2. *Posterior imprecisions of* H_0 *for* Γ_{01}, Γ_{02} *and* Γ_{03}

Table 2 shows that the posterior imprecisions of H_0 in the usual contamination class with the Normal base prior are similar these with the Cauchy prior. Posterior imprecisions with the Intrinsic base prior are slightly bigger. For the three classes, however, the imprecisions are very big. Hence, more effort in prior elicitation is required.

If we recognize again that our uncertainty in the base prior is essentially uncertainty in their tails, then a small ε for the body and another bigger ε for the tails should be chosen. In particular, let us take $\varepsilon = 0$

for the central part of $\pi_{0i}(\theta)$, that is

$$\Gamma_{ti} = \Big\{\pi(\theta) : \pi(\theta) = \Big[1 - \varepsilon \mathbf{1}_{C_2^c}(\theta)\Big]\pi_{0i}(\theta) + \varepsilon q(\theta) \mathbf{1}_{C_2^c}(\theta), q \in Q\Big\}, \quad i = 1, 2, 3$$

where now ε represents the uncertainty on $\pi_{0i}(\theta)\ \mathbf{1}_{C_2^c}(\theta)$. The corresponding posterior imprecisions of H_0 with respect these classes are given in Table 3.

x	$\Delta_{\Gamma_{t1}} P^\pi(H_0\|x)$			$\Delta_{\Gamma_{t2}} P^\pi(H_0\|x)$			$\Delta_{\Gamma_{t3}} P^\pi(H_0\|x)$		
	$\varepsilon = .2$	$\varepsilon = .5$	$\varepsilon = 1$	$\varepsilon = .2$	$\varepsilon = .5$	$\varepsilon = 1$	$\varepsilon = .2$	$\varepsilon = .5$	$\varepsilon = 1$
0	0.05	0.13	0.27	0.06	0.14	0.26	0.03	0.08	0.15
0.5	0.05	0.12	0.25	0.05	0.13	0.25	0.03	0.08	0.16
1.0	0.03	0.09	0.21	0.04	0.10	0.22	0.03	0.07	0.15
1.5	0.02	0.04	0.13	0.03	0.07	0.18	0.02	0.06	0.14

Table 3. *Posterior imprecisions of* H_0 *for* Γ_{t1}, Γ_{t2} *and* Γ_{t3}

This table shows that:

1. Posterior imprecisions have been substantially reduced by considering only uncertainty in the tails. Therefore, to concentrate efforts on accurately eliciting many features in the body of the base prior is important because it might result in a robust answer. The hope is that this kind of information can be obtained from the experts.
2. The posterior ranges of H_0 with the Cauchy base prior are, for any of the considered values for ε, smaller than those given by the Normal and Intrinsic base priors. The fourth, seventh and tenth columns give the biggest possible posterior imprecision when the tails of Normal, Intrinsic and Cauchy distributions are allowed to vary respectively, in Q. The reduction in the range of the posterior probability when using Cauchy is a factor of 0.55. The analysis suggests recommending use of the Cauchy prior distribution as the base prior for this problem.

3. *Limitations of Global Robustness.* Given the likelihood $f(x|\theta)$, the prior $\pi(\theta|\mu)$, the hyperprior $h(\mu)$ and the quantity of interest $\varphi(\theta)$, the posterior expectation is given by

$$E_{\boldsymbol{x}}^{\pi,h}\varphi(\theta) = \frac{\int\int \varphi(\theta) f(\boldsymbol{x}|\theta)\pi(\theta|\mu)h(\mu)d\mu d\theta}{\int\int f(\boldsymbol{x}|\theta)\pi(\theta|\mu)h(\mu)d\mu d\theta}.$$

(It has been assumed that $\boldsymbol{x} \perp \mu|\theta$). The problem can be considered as one with the elements:

$$\left\{ f(\boldsymbol{x}|\theta), \pi_h(\theta)\left(= \int \pi(\theta|\mu)h(\mu)d\mu\right), \varphi(\theta)\right\},$$

with global robustness quantified by the range

$$\left(\inf_{\pi,h} E_{\boldsymbol{x}}^{\pi,h}\varphi(\theta), \sup_{\pi,h} E_{\boldsymbol{x}}^{\pi,h}\varphi(\theta)\right),$$

where the inf and sup is over some class of distributions $\pi(\theta|\mu)$ and $h(\mu)$; see for instance Moreno and Pericchi (1993). The idea stated by James Berger in Section 4.5 is to look at the problem as one with the elements:

$$\left\{ m(\boldsymbol{x}|\mu)\left(= \int f(\boldsymbol{x}|\theta)\pi(\theta|\mu)d\theta\right), h(\mu), \phi(\mu)\right\},$$

where $\phi(\mu) = E^{\pi(\theta|\boldsymbol{x},\mu)}\varphi(\theta)$. Global robustness is now quantified by the range

$$\left(\inf_{\pi,h} E^h\phi(\mu), \sup_{\pi,h} E^h\phi(\mu)\right),$$

where, as above, inf and sup is over some class of priors $\pi(\theta|\mu)$ and $h(\mu)$.

Under mild conditions it is clear that $E^h\phi(\mu) = E_x^{\pi,h}\varphi(\theta)$, so that both viewpoints of the problem are processing the same inputs. In example 10 we have

$$E_x^{\pi,h}(\theta_1) = \frac{\int \frac{1}{2}(x_1+\mu)m(\boldsymbol{x}|\mu)h(\mu)d\mu}{\int m(\boldsymbol{x}|\mu)h(\mu)d\mu},$$

where $\pi(\theta|\mu) = N(\theta|\mu, 1)$ and $h(\mu)$ is in the class

$$\Gamma_1 = \left\{ h(\mu) : \int_{-8}^{-12} h(\mu)d\mu = 1 \right\}.$$

Consequently, the range of the posterior mean for θ_1, say R_1, is

$$R_1 = \left(\inf_{\mu\in(-8,12)} \frac{1}{2}(x_1 + \mu), \sup_{\mu\in(-8,12)} \frac{1}{2}(x_1 + \mu) \right)$$

$$= \left(\frac{1}{2}x_1 - 4, \frac{1}{2}x_1 + 6 \right).$$

Here, $m(\boldsymbol{x}|\mu)$ has effectively been ignored. But we observe that it is caused by the class Γ_1 (the extreme priors concentrate mass on one point) and not by the underlying philosophy of global robustness. In fact, if Γ_1 is replaced by the class

$$\Gamma_2 = \left\{ h(\mu) : \int_{-8}^{12} h(\mu)d\mu = 1, \quad \int_{-8}^{3} h(\mu)d\mu = 0.5 \right\},$$

which is suggested by the likelihood $m(\boldsymbol{x}|\mu)$ for $\overline{x} = 3$, then the corresponding range of the posterior expectation of θ_1, say R_2, is

$$R_2 = \left(\inf_{\substack{-8 \le a_1 \le 3 \\ 3 \le a_2 \le 12}} g(a_1, a_2), \sup_{\substack{-8 \le a_1 \le 3 \\ 3 \le a_2 \le 12}} g(a_1, a_2) \right),$$

where $g(a_1, a_2)$ is given by

$$g(a_1, a_2) = \frac{(x_1 + a_1)\exp\left\{-\frac{p}{4}(\overline{x} - a_1)^2\right\} + (x_1 + a_2)\exp\left\{-\frac{p}{4}(\overline{x} - a_2)^2\right\}}{\exp\left\{-\frac{p}{4}(\overline{x} - a_1)^2\right\} + \exp\left\{-\frac{p}{4}(\overline{x} - a_2)^2\right\}}.$$

Note that now $m(\boldsymbol{x}|\mu)$ has been taken into account.

Something similar happens if $h(\mu)$ is unimodal with mode at μ_0. In this case the range of the posterior expectation for θ_1 is given by the total variation of the function

$$k(a) = \frac{\int_{\mu_0}^{\mu_0+a} \frac{1}{2}(x_1 + \mu)m(\boldsymbol{x}|\mu)h(\mu)d\mu}{\int_{\mu_0}^{\mu_0+a} m(\boldsymbol{x}|\mu)h(\mu)d\mu},$$

and $m(x|\mu)$ is taken into account with the only exception when $a = 0$. Also in this case all the mass has been concentrated in one point. This point is here the mode. Except for this situation I guess that the only class of priors for μ that ignores $m(x|\mu)$ in the analysis is the class of all prior distributions on a given range (μ_1, μ_2).

I have learnt from the arguments stated by the author in section 4.5 the following:

1. When using hyperparameters, it is of interest to consider the likelihood at the second stage. It can help us in eliciting hyperpriors.
2. Even at a higher level of the hierarchy (example 10 considers a second level) caution in using a very big class for the hyperprior parameter is important (see also Moreno and Pericchi, 1993).

Finally, I should say that I am resistant to admit, for the being time, limitations of global robustness, although I admit difficulties that are being solved step by step. This paper by James Berger is a proof of this.

LUIS RAUL PERICCHI *(Universidad Simon Bolivar)*

In the first phrase of his paper Prof. Berger defines, "Robust Bayesian analysis is the study of the sensitivity of Bayesian answers to uncertain inputs". This paper is an incisive and insightful review of the subject defined above, which makes a delightful reading. It is also a "hijacker's guide to the galaxy" of Prof. Berger's thinking, who has recently produced key contributions in areas like global robustness, inherent robust methods, reference priors and automatic model selection, among others. This paper shows the deep link of his thinking, very much concerned with the correct foundations but also aware of the compromises that statistical practice dictates. This review demonstrates the good health of the subject, and develops the implications of a broad and realistic foundational system. In (1986) in a meeting with J. W. Tukey I mentioned a Robust Bayesian method in medical diagnosis and then he asked me, "Is there a Robust Bayesian approach?" This review clearly states an affirmative answer in 1993.

Somewhat outside the automation of the sensitivity analysis interpretation, which is one of the main themes of this review of Robust Bayes, one still wonders if a fully automatic Robust Bayesian analysis is possible, free from the first stage of informative elicitation. For ex-

ample, some "reference" Bayesian analyses are clearly more robust in some sense (particularly when the likelihood is very informative), than other reference analyses. But, do we have a reliable measure of the difference in robustness? Or in other words, is a Robust Bayes analysis possible which is based essentially on likelihood assumptions? Admittedly, "Near Ignorance Classes" is a limited suggestion in that direction and it is not elicitation free. However I think that this general question is quite relevant.

Turning to specific areas, I will concentrate on model choice. I just want to expand on some of Prof. Berger's comments. Bayesian thinking on model selection is necessary. For this class of problems frequentist and Bayesian measures of evidence are typically ever more in conflict as data accumulates. One of the major achievements of Robust Bayes has been to show transparently that P-values are misleading, and this is even of more concern in large samples, where significance testing typically rejects the simpler model. Quite differently, Bayesian methods embody an automatic "Ockham's razor", the well accepted scientific principle that if two models fit the data approximately equally well, the simpler one is to be preferred. To be more specific consider asymptotic approximations. Regarding estimation problems and assuming mild regularity conditions on likelihood and prior, the posterior density $\pi(\theta|x)$ is approximately multivariate-Normal $N_p(\hat{\theta}, [I(x)]^{-1})$, where $\hat{\theta}$ is the Maximum Likelihood Estimator, and $I(X)$ is the observed Fisher's Information Matrix evaluated at $\hat{\theta}$. Thus for regular likelihood the posterior is asymptotically independent of the prior. Here it is the *smoothness* of the prior that matters for the prior to fade away. This is the basis of the claim of some frequentist statisticians that they can just work with Maximum Likelihood, since it is approximately a Bayesian answer, without taking the trouble of specifying a prior, or a class of priors.

However the situation in Model Selection is completely different. The asymptotic Bayes Factor in some generality is, for

$$M_i : f_i(x|\theta_i) \quad \text{vs} \quad M_j : f_j(x|\theta_j),$$

$$B_{ij} \approx \frac{f_j(x|\hat{\theta}_j)|I_j(x)|^{-1/2}(2\pi)^{pj/2}\pi_j(\hat{\theta}_j)}{f_j(x|\hat{\theta}_i)|I_i(x)|^{-1/2}(2\pi)^{pj/2}\pi_i(\hat{\theta}_i)}. \tag{1}$$

Approximation (1) shows that the influence of the priors remain, in the form of the ratio of priors evaluated at Maximum Likelihood estimates.

Here it is the *size* of the priors what matters. This expose two facts. The first in that usual automatic reference prior analyses do not work since the arbitrary constants in the priors do not cancel out. The second is that the influence of the priors, in the form of their relative sizes remain asymptotically, and thus Robust Bayes in model selection problems is unavoidable. In summary, there is a vast room for developing new automatic and robust Bayesian procedures for comparing models. In my view this is the most exciting and promising arena for Bayesian approaches. Note that seemingly different problems, like detection of outliers or density estimation can be encompassed in the umbrella of model selection problems.

But, how to tackle model selection, in a Bayesian way? Prof. Berger rightly emphasized that the classes of priors should be somehow "tied together" for different models. Berger and Pericchi (1993) have proposed the "Intrinsic Bayes Factor" that seems to automatically match predictives. The Intrinsic Bayes Factor is based on taking minimal training samples and taking averages of the resulting Bayes Factors. If for instance we are comparing any two location-scale likelihoods, then minimal training samples match the $1/3$ and $2/3$ quantiles of the predictive distribution of a future observation. See Berger and Pericchi (1993) for other examples. Turning to classes of priors, Pericchi and Pérez (1994) consider a finite set of separate likelihoods: $M_1, M_2, \ldots, M_j$. There it is supposed that we are willing to assume values for $P(M_j), j = 1, \ldots, J$. In this way the marginal of observations $m(x|\pi, M_j)$ naturally arises. (This is another interesting theme of Prof. Berger's paper). In Pericchi and Pérez (1994) it is suggested to take the "common class of predictives" approach. That is to consider classes of priors that are such that the predictives for a future observation across different models, simultaneously obey some conditions. For example that a set of predictive quantiles are the same across models. How difficult it is to work with such classes is still an open question. We also provide an example to warn about the illusion of robustness (or lack of it) when considering a class of priors, but only one model.

Summing up, Prof. Berger has put together a masterful review of the most successful (in my opinion) framework of statistics, on which his own influence has been crucial.

M. JESÚS BAYARRI *(Universitat de València)*

It is indeed a honor and a pleasure for me to have the opportunity to comment on the great paper by Professor Berger. The world of Bayesian robustness has grown so fast and in so many directions that this expert, insightful overview will undoubtedly be a very illuminating and valuable contribution to the area. Unfortunately, the better the paper the harder the task of the discussant, and this one is so thoughtful that also includes a critical discussion of the roles and limitations of robustness analysis. My discussion will be reduced to a comment of support and a comment of warning, the first of which intends to emphasize the need of robust Bayes analyses with respect to changes in the likelihood function as exemplified in a situation involving weighted distributions, whereas the second one intends to be a warning against indiscriminate, naïve use of robust Bayes method as exemplified in a situation involving high dimensions.

1. *Weighted Distributions.* Section 4.4 in the paper addresses the important issue of studying robustness with respect to changes in the likelihood function by using nonparametric classes of likelihoods. The question then might arise as to whether non-parametric classes are really needed. The following example shows that in some situations studying robustness within a parametric class of likelihoods might not suffice. The example is in the framework of weighted distributions, as presented in Example 9, and it is inspired in a real problem analyzed in Nair and Wang (1989), and West (1993), although what appears here is a very simplified version that only bears a remote resemblance with the original analysis.

The example is as follows. Assume that in a problem of searching for oil pools in a certain (very large) oil field, n pools have already been discovered and are under exploitation. (We do not take here into account the finite nature of the total population of pools.) Assume that some measure of surface, X_i of the pools is supposed to have an exponential distribution with mean μ. Assume that, for the $n = 23$ pools already discovered, $\bar{x} = 150$ (in the appropriate units) and that interest is in estimating μ with the usual posterior mean.

A naïve analysis ignoring the size-bias effect, would take the distribution for the X_i's at face value and would therefore assume that the

model generating the data has density

$$f(x|\mu) = \frac{1}{\mu} e^{-x/\mu}. \tag{B.1}$$

It then follows that, with the non-informative prior $\pi(\mu) \propto 1/\mu$, the posterior distribution of μ is an inverse gamma, $\pi(\mu|\boldsymbol{x}) = Ga^{-1}(\mu|n, n\bar{x})$, with mean $n\bar{x}/(n-1)$, so that for the given data

$$E^{\text{naïve}}(\mu|\boldsymbol{x}) = 156.82. \tag{B.2}$$

Assume now that the discovery process of the pools is such that the larger the surface of the pool, the more likely it is to be discovered. In this case the analysis above is flawed since (B.1) is not the density generating the actual data that we get to observe. Instead, a weighted version of it

$$f_w(x|\mu) \propto w(x) f(x|\mu), \tag{B.3}$$

should be used, with a weight function $w(x)$ which is an non decreasing function of x. Using $w(x) = x^a$ with $a \geq 0$, results in a f_w which is the density of a gamma distribution $Ga(a+1, 1/\mu)$, so that the non-informative prior results in the posterior $Ga^{-1}(\mu|n+na, n\bar{x})$, with mean $n\bar{x}/(n-1+na)$. Of course, the real analyses of the data mentioned above did take the size-bias effect into account and used a weight function which could be roughly compatible with taking, in our simplified version,

$$w(x) = x^{0.8}. \tag{B.4}$$

Hence, the estimate of μ now becomes

$$E_{0.8}(\mu|\boldsymbol{x}) = 85.34. \tag{B.5}$$

The large difference between the two estimates, (B.2) and (B.5) dramatically emphasizes the need for taking into account the size-bias effect. Besides, such a large difference also demonstrates that this effect is very important in this case, so that it would be wise to investigate how $E(\mu|\boldsymbol{x})$ changes as $w(x)$ varies from its assumed from (B.4).

A natural thing for a Bayesian to do would be to assume $w_a(x) = x^a$ with $a \geq 0$ and to put a prior $\pi(a)$ on a. The posterior mean of μ then becomes

$$E(\mu|\boldsymbol{x}) = n\bar{x} \int_0^\infty \frac{1}{n-1+na} \pi(a) da. \tag{B.6}$$

A prior of the form $\pi(a) = Ga(a|\alpha, 5\alpha/4)$ has $E(a) = 0.8$ and variance $\text{Var}(a) = 0.16$ if $\alpha = 4$, and $\text{Var}(a) = 0.064$ if $\alpha = 10$. With this two values of α, the estimates of μ as given by (B.6) are

$$E_{\alpha=4}(\mu|\boldsymbol{x}) = 89.51, \qquad E_{\alpha=10}(\mu|\boldsymbol{x}) = 87.12. \tag{B.7}$$

Therefore, if one is fairly sure about the form of $w(x) = x^a$ with a not differing much from 0.8, and if a gamma prior for a is used, it would be concluded that the estimate (B.5) is fairly robust.

Assume, on the other hand, that the investigators, while believing that the weight function behaves roughly as x^a, with a close to 0.8, do not feel very confident about precisely assessing a gamma prior on a, so that an exploratory robustness analysis is in order. The easiest one would study the range of posterior means as $w(x)$ ranges over a parametric family of the form

$$\mathcal{W}^P = \{w(x) = y^a : a_0 \leq a \leq a_1\}. \tag{B.8}$$

For $a_0 = 0.7$, $a_1 = 0.9$, and the given data, the range of $E(\mu|\boldsymbol{x})$ as $w(\cdot)$ ranges over $\mathcal{W}^P$ is easily seen to be

$$80.0 \leq E_a(\mu|\boldsymbol{x}) \leq 90.55, \tag{B.9}$$

and robustness might still be claimed. Nevertheless, in this example, as it is often the case in size-biased problems, the very same *form* of the weight function, x^a, is highly subjective. Hence, it would be more appropriate to only assume that $w(x)$ is a non-decreasing function of x, and study robustness of the estimate as $w(x)$ ranges over the *non-parametric* class

$$\mathcal{W} = \{\text{non-decreasing } w : \min\{x^{0.7}, x^{0.9}\} \leq w(x) \leq \max\{x^{0.7}, x^{0.9}\}\}. \tag{B.10}$$

By using the results in Bayarri and Berger (1993), it can be shown that the infimum of $E(\mu|\boldsymbol{x})$ as w ranges over $\mathcal{W}$ is attained at a weight function $w(x)$ of the form (4.15) with $r = 0$, $a = 153.6$, $s = \infty$, and the supremum at a $w(x)$ of the form (4.16) with $r = 0$, $h_2(c) = 81.9$, $c = (81.9)^{.9}$, $h_1(c) = (81.9)^{9/7}$, and $s = \infty$. (Notice that $w_1(x) = x^{0.9}$ for $0 \leq x \leq 1$, and $w_1(x) = x^{0.7}$ for $x > 1$, and that $w_2(x) = x^{0.7}$

for $0 \leq x \leq 1$, and $w_2(x) = x^{0.9}$ for $x > 1$). The resulting range of posterior means is

$$65.99 \leq E(\mu|x) \leq 111.4, \qquad (B.11)$$

which is four and a half times as large as the range (B.9) obtained with the parametric class (B.8). Robustness may very well not be claimed here, thus revealing how sensitive $E(\mu|x)$ is with respect to the assumed form x^a for the weight function.

2. *High Dimensions.* While the previous comment intended to emphasize the need of robustness analyses, the purpose of this one is to warn against their indiscriminate, naïve use, specially when dealing with very high dimensions. The scenario is that of Example 10 and assumes

$$\begin{aligned} X_i &\sim N(\theta_i, 1)\,, \qquad i = 1, 2, \cdots, p \\ \theta_i &\sim N(\mu, 1)\,, \qquad i = 1, 2, \cdots, p \\ \mu &\sim N(2, \tau^2)\,. \end{aligned} \qquad (B.12)$$

The difficulties in dealing with a very large p are already hinted in the estimation problem treated in the paper. Thus, for instance, assume that the goal is to estimate θ_1 and that, instead of fully assessing $\mu \sim N(2, \tau^2)$ as in (B.12), the class in (4.18) is used with k selected so that μ varies in a "95% confidence or credible set" for μ computed by using likelihood or Bayesian non-informative prior methods, as suggested in the paper. It can be checked that the class (4.18) is then

$$\Gamma^* = \{\pi \in \Gamma : \bar{x} - 2\sqrt{\frac{2}{p}} \leq \mu \leq \bar{x} + 2\sqrt{\frac{2}{p}}\}. \qquad (B.13)$$

which is the class used in Example 10 (continued) for the case $p = 8$ and $\bar{x} = 3$. The range of the posterior mean of θ_1 as π ranges over Γ^* is

$$\frac{x_1 + \bar{x}}{2} \pm \sqrt{\frac{2}{p}}\,, \qquad (B.14)$$

so that the answer gets more and more robust as p grows. It should be kept in mind, however, that the class Γ^* gets *very* small for large values of p.

The difficulties of robust Bayes analyses in high dimensions are more evident in testing problems, specially when using priors that are not as naturally robust as the two-level hierarchical prior above. To be more explicit, consider the following variation of Example 10:

$$\begin{aligned} X_i &\sim N(\theta_i, 1)\,, \qquad i = 1, 2, \cdots, p \\ \theta_i &\sim N(2, \tau^2) \qquad i = 1, 2, \cdots, p\,. \end{aligned} \tag{B.15}$$

Let $\theta = (\theta_1, \theta_2, \cdots, \theta_p)^t$ and assume that the goal is to test $H_0 : \theta = \mathbf{21}_p$ versus $H_A : \theta \neq \mathbf{21}_p$. A full bayesian analysis procceeds then by assessing a prior $g(\tau^2)$ for τ^2 and computing, say, the Bayes factor $B(g)$ for H_0. On the other hand, assume that, to avoid the specifications of g the infimum of the Bayes factor, $\underline{B}$ over $\tau^2 \geq 0$ is computed instead, and conclusions are based on $\underline{B}$. The warning is that $\underline{B}$ can be a horrible substitute for $B(g)$ when p is very large, as the following simple example demonstrates:

Assume that $g(\tau^2)$ is such that $t = 1/(1+\tau^2)$ is distributed according to

$$g^*t) \propto t^{a-1}e^{-bt}, \quad 0 \leq t \leq 1, \tag{B.16}$$

and take, for instance, $a = 2$ and $b = 12$ so that $E(t) = 0.167$ (recall that in Example 10, moderate values of τ^2 were contemplated.) The following table shows the ratio $B(g)/\underline{B}$ for several values of p:

p	4	8	20	50	1000
$B(g)/\underline{B}$	3820	28813	479259	5.25×10^6	2.6×10^8

Thus, the larger p the worse $\underline{B}$, even though the larger p the better we can estimate θ. This phenomena is, at the very least, disturbing, and it is not a peculiarity of the particular g selected. Similar effects occur with virtually any proper g (Bayarri and Berger, 1994).

Of course, this not need to be the case in *every* testing situation involving high dimensions. Let us revisit Example 10 once more, using now a full two-level hierarchical prior:

$$\begin{aligned} X_i &\sim N(\theta_i, 1)\,, \qquad i = 1, 2, \cdots, p \\ \theta_i &\sim N(\mu, 1)\,, \qquad i = 1, 2, \cdots, p \\ \mu &\sim g(\mu)\,. \end{aligned} \tag{B.17}$$

Assume now that the goal is to test $H_0 : \theta_1 = 2$ versus $H_A : \theta_1 \neq 2$. Again, we can compute the Bayes factor $B(g)$ for some g, or avoid the specification of g and compute the infimum of Bayes factor, $\underline{B}$ over all values of μ. For the non-informative prior $g(\mu) \propto constant$, the ratio $B(g)/\underline{B}$ can be computed to be

$$\frac{B(g)}{\underline{B}} = \frac{p}{p-1} \exp\left\{\frac{1}{4}\frac{p-1}{p}(\bar{x}^* - x_1)^2\right\} \qquad (B.18)$$

where $\bar{x}^*$ is the mean of $x_2, x_3, \cdots, x_p$. Some particular values of this ratio are:

p	$B(g)/\underline{B}$
8	$1.143 \exp\{0.219(\bar{x}^* - x_1)^2\}$
50	$1.020 \exp\{0.245(\bar{x}^* - x_1)^2\}$
1000	$1.001 \exp\{0.25(\bar{x}^* - x_1)^2\}$

which does behave as it could be expected.

As an overall conclusion of this comment, we could say that, when dealing with very high dimensions, while we have been warned that large classes are useless, it might be the case that sensible classes are much too small. Besides, robust answers may be sensible approaches to a full Bayesian analysis or they can be clearly inappropriate, and whether one or the other applies is not always clear without assessing the prior. Thus, in certain problems in high dimensions, it might not be possible to avoid strong, detailed prior specifications, no matter how difficult or time consuming it can be.

JOSÉ M. BERNARDO *(Universidad de Valencia)*

I would like to suggest the exploration of an information-based class of priors to study robustness.

Let $p_0(\theta)$ be a prior distribution for the quantity of interest θ which may be claimed to have some priviledged position, either as a good approximation to an honest subjective prior, or as some accepted standard. Then, there are a number of foundationally based arguments (see e.g., Bernardo and Smith, 1994, pp. 154–160) to consider the class

$$\Gamma_1 = \left\{p(\theta);\ \int_\Theta p_0(\theta) \log_2 \frac{p_0(\theta)}{p(\theta)}\, d\theta < \epsilon\right\}$$

for some $\epsilon > 0$. Indeed,

(i) the logarithmic divergence has an interesting interpretation as the expected loss to be suffered if $p(\theta)$ is used instead of $p_0(\theta)$, when preferences are described by a proper, local score rule,

(ii) the utility constant ϵ may be given an information-theoretical interpretation as the number of *bits* of information which are necessary to recover $p_0(\theta)$ from $p(\theta)$ (Renyi, 1962/1970, p. 564), and

(iii) the results would be invariant under one-to-one reparametrizations of the parameter of interest.

Better still, if $p(x \mid \theta)$ is the model to be used, one might consider the class

$$\Gamma_2 = \left\{p(\theta);\ \int_X m_0(x) \log_2 \frac{m_0(x)}{m(x)}\, dx < \epsilon\right\},$$

where

$$m(x) = \int_\Theta p(x \mid \theta) p(\theta)\, d\theta, \quad m_0(x) = \int_\Theta p(x \mid \theta) p_0(\theta)\, d\theta.$$

In this case,

(iv) the results would be robust with respect to the dimension of θ, and

(v) one explicitly considers prediction robustness.

For a given sample $\{x_1, \ldots, x_n\}$, one could go ever further and analyse robustness with respect to the model by considering the class Γ_3

$$\{p(x \mid \theta);\ \int_X \pi_0(x \mid x_1, \ldots, x_n) \log_2 \frac{\pi_0(x \mid x_1, \ldots, x_n)}{\pi(x \mid x_1, \ldots, x_n)}\, dx < \epsilon\},$$

where $\pi_0(x \mid x_1, \ldots, x_n)$ is the reference posterior predictive distribution (Bernardo, 1979, Berger and Bernardo, 1992) which corresponds to the hypothesised model $p_0(x \mid \theta)$ and $\pi(x \mid x_1, \ldots, x_n)$ is the reference posterior predictive distribution which corresponds to any other model $p(x \mid \theta)$.

JUAN A. CANO *(Universidad de Murcia)*

First let me compliment James Berger on his paper and talk that interestingly review a number of general issues in Robust Bayesian Analysis. Although, in the last years, there have been several reviews on the Robust Bayesian approach to inference, for instance, Berger (1984, 1990) and Wasserman (1992b), all of them are different from each other because they address different points; Berger (1990) focuses on sensitivity to the prior, Wasserman (1992b) deals with methodological advances up to that date and in this paper the author mainly presents his own opinions on general issues to statisticians not in the field.

The discussion of a review is always a difficult task but in this case it has additional difficulties because when you read the paper you think that all that has been done is said and all that might be done is also said. In spite of this, I will give my own opinions on some issues

My first comment is related to those facts this review addresses to statisticians not in the field. In Subsection 1.1 an overwhelming number of motivations to adopt the robust Bayesian viewpoint are given. Particularly good seems to me the idea of showing the lower bound $\underline{P}$, along with the respective P-value, when teaching and reporting to the users on hypothesis testing problems. A similar good idea would be to present the optimal robust credible set as defined in 4.6 along with the respective confidence interval; obviously, cautions pointed out in 4.6 should be taken into account. On the other hand, it would be appealing for statisticians not in the field, mainly for classical statisticians and even for mathematicians, to undertake bayesian robustness problems where involved mathematics is needed; for instance, in Moreno and Cano (1992) a partial solution of the Monge-Kantorovich problem provides an approximation to the problem of sensitivity with respect to the prior for classes of bidimensional priors having specified marginals.

Second, I specifically focus my attention on the problem of robustness with respect to sampling models. From a mathematical point of view the use of the Dirichlet process and other devices to put priors on the space of all probability distributions is very charming but it has some drawbacks that are being solved as is shown in Lavine (1992b) and references therein. On the other hand, to consider nonparametric classes of likelihoods such as $\mathcal{F}_\varepsilon$ defined in subsection 4.4 reduces the problem to one of sensitivity with respect to the prior but does not reflect typical

types of uncertainty in $g(x_i|\theta)$, the density of the observable random variable X_i, so that the other classes $\mathcal{F}_1^g$ and $\mathcal{F}_2^g$ suggested by the author are more appealing. In Cano and Moreno (1993) a class of likelihoods

$$\mathcal{F} = \left\{ l : l(\theta) = \prod_{i=1}^{n} g(x_i|\theta), g \in \mathcal{F}_1^g \right\}$$

where the class $\mathcal{Q}$ in $\mathcal{F}_1^g$ is a density band class, say, a $\mathcal{F}_2^g$ type class, is considered. This class is not too large because it combines use of $\mathcal{F}_1^g$ and $\mathcal{F}_2^g$ and furthermore it is mathematically tractable. Cano and Moreno (1993) provides a way to compute bounds on posterior quantities with respect to this class in the discrete case; nice maths including Lagrangian multipliers and mathematical programming are needed. It is there shown that new difficulties emerge yielding undesirable results as is shown in the following example.

Example 1. Consider $\mathcal{X} = \{x_1, x_2, x_3\}$ and $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$. Assume the prior $\pi(\theta_i) = 0.2; i = 1, \ldots, 5$. Let $g_0(x_i|\theta_j)$ be as displayed in table 1.

	x_1	x_2	x_3
θ_1	.1	.2	.7
θ_2	.2	.3	.5
θ_3	.4	.3	.3
θ_4	.5	.3	.2
θ_5	.7	.2	.1

Table 1. *Base probability mass function* $g_0(x_i|\theta_j)$.

Consider the class $\mathcal{F}$ with $\varepsilon = 0.2$ where $q(x_i|\theta_j) \in \mathcal{F}_2^g$ with $g_1(x_i|\theta_j) = g_0(x_i|\theta_j) + 0.1$ and $g_2(x_i|\theta_j) = g_0(x_i|\theta_j) + 0.1$. For different samples, the upper and lower bounds of the posterior probabilities of $\{\theta_3\}$ as the sampling model ranges over $\mathcal{F}$, denoted respectively by $\overline{P}$ and $\underline{P}$, are displayed in table 2. Here R denotes the corresponding ranges as given by $\overline{P} - \underline{P}$.

Table 2 shows a surprising drawback: in our example, adding data results in a larger R. Even more, consider the class of probability mass

	x_3	(x_2, x_3)
$\underline{P}$	.15	.17
$\overline{P}$	.18	.22
R	.03	.05

Table 2. *Posterior bounds for different samples.*

functions

$$\mathcal{F}_2^g = \{g : g_0(x_i|\theta_j) - \delta \leq g(x_i|\theta_j) \leq g_0(x_i|\theta_j) + \delta\}.$$

The respective $\underline{P}$ and $\overline{P}$ for the sample x_2 are 0.125 and 0.4 for $\delta = 0.1$. But if we repeatedly observe x_2, an asymptotically vacuous answer, that is $\underline{P} = 0$ and $\overline{P} = 1$, is obtained for any $\delta > 0$. This shows that the nature of the problem of global robustness with respect to the likelihood is very different to that related to the prior.

My conclusion is that robutness with respect to sampling models needs further research and the finding of inherent robust situations in this context is needed.

JULIÁN DE LA HORRA *(Universidad Autónoma de Madrid)*

It is really hard to add something interesting to this complete review of recent developments in Bayesian robustness.

I would like just to point out an idea for the case in which the parameter space is multidimensional. Sometimes, it is relatively easy to specify the prior marginals, but the joint prior distribution may be extremely difficult to elicit. There exists a vast literature about families of multivariate distributions with fixed marginals. These families of distributions are well-studied and this could be an advantage for knowing whether the prior beliefs fit to one of them.

In De la Horra and Fernández (1994), the well-known Farlie-Gumbel-Morgenstern system (see Johnson and Kotz (1975, 1977) for a good description of this family of distributions) is used for carrying out a robustness analysis with respect to departures from the assumption of prior independence. The density functions of the Farlie-Gumbel-Morgenstern

family are given by

$$\pi_\lambda(\theta_1, \theta_2) = f(\theta_1)g(\theta_2)[1 + \lambda(1 - 2F(\theta_1))(1 - 2G(\theta_2))],$$

$$\lambda \in [-1, 1],$$

where $f(\theta_1)$ and $g(\theta_2)$, $F(\theta_1)$ and $G(\theta_2)$ are the marginal densities and the marginal distribution functions, respectively (both fixed).

This model allows for moderate departures from independence and, therefore, could be suitable for situations in which the marginal priors are easy to elicit and we have reasonable confidence in the assumption of prior independence, but we are not absolutely sure about it. The main problem with this class is that it is a parametric family and may be too narrow for some situations. But, even in this case, the simplicity of its analysis still makes this class worthy to be used at a first step of a robustness study, because, if a lack of robustness is found for this class, the same conclusion would be reached for a wider class (possibly, much more difficult to analyze).

Similar studies could be carried out with other well-known families of multivariate distributions.

JACINTO MARTÍN and DAVID RÍOS-INSÚA
(Universidad Politécnica de Madrid)

Our discussion of Berger's excellent paper will be limited to some general remarks and a few technical points.

Traditionally, Bayesian analysis has been criticised on three accounts: i) excessive precision demanded in the judgmental inputs to an analysis; ii) very involved computational problems; iii) failures as a descriptive theory of actual inference and decision making processes.

The first two issues have been dealt with by Berger. The third one has led to an enormous volume of literature in the fields of Economics and Decision Sciences which seems to challenge Berger's suggestion of *a common perception of subjective Bayesian analysis as the only coherent method of behavior*, see e.g., Rios Insua (1994). The good news for robust Bayesians is that, somehow, robust Bayesian analysis may account for those failures. Consider, for example, a subject facing an Allais' type experiment. That is usually a fairly new situation for him and he has to respond without much analysis and guidance. We should expect him to have imprecise preferences, modelled through a

class of utility functions. If he is incoherent in the experiment, we could explain it by arguing that some functions in his class lead to one type of choice in the experiment and other functions lead to the opposite choice. A descriptive account of this type of model may be found in Leland (1992). Our first experimental results, Rios et al (1994), indicate that this might be an alternative explanation of those failures, fitting nicely in the robust Bayesian framework: incoherent performers are just imprecise performers who have not devoted enough time to think about decision making problems.

We believe that most robust Bayesian analysis has been oriented towards inference problems, hence the emphasis on studying robustness mainly with respect to the prior and, to a lesser extent, to the likelihood. However, that theory is insufficient for decision making purposes, since it forgets another essential component, the utility/loss function. Berger identifies three reasons for this, but we are in partial disagreement. First, statisticians have avoided utilities, but perhaps because of traditional training in classical methods: eliciting utilities is not that difficult, and is probably easier than eliciting probabilities. At least that is our applied and didactical experience. Second, the Decision Analysis (e.g., Keeney and Raiffa, 1993) and Stochastic Dominance (e.g., Levy, 1992) literature have identified classes of utility functions, which would allow a theory parallel to that of robustness with respect to classes of priors. Moreover, utility-only studies are probably easier than prior-only robustness studies, since the former case involves linear operators.

We agree, though, with the fact that prior-utility (and model) robustness studies are much more difficult computationally, see Proll *et al.* (1993) for comparisons. But this should be seen just as a challenge: more work should be devoted to this area. Specifically, we believe that the computation of nondominated actions with respect to classes of priors and utilities is a fundamental problem for those studies, which deserves urgent research. Incidentally, we believe that reweighting schemes such as those in section 5.1 might provide a way forward, with switches from (h, π) to (h', π'), rather than from π to π'.

Berger mentions at various points determining whether a range of a certain posterior quantity is large or not, as a way to suggest whether additional analysis is necessary. But when is large large enough? The Robust Bayesian literature remains quite silent about this question, which

deserves additional work, if we do not want to rely on informal methods. Calibration procedures in McCulloch (1989) are probably good ideas.

Finally, a minor technical point. Most robust Bayesian axiomatisations lead to a convex class of priors modelling imprecise beliefs. Thus, a minimal requirement for a class in a robustness study is that it is convex. This would invalidate some classes used in the literature, specially some of the parametric classes.

The following contributions were later received in writting.

BRUNO BETRÒ *(CNR-IAMI, Milano)*

Thanks to professor Berger for his comprehensive review of the rapidly growing field of robust Bayesian analysis.

My only comments concern the *Generalized Moment* Class which in the paper is referred to under the heading of *Moment Class*. Indeed the class defined as

$$\Gamma = \{\pi : \int_{\Theta} H_i(\theta)\pi(d\theta) \leq \alpha_i,\ i = 1, \ldots, n\} \qquad (1)$$

where H_i are given π-integrable functions and α_i, $i = 1, \ldots, n$, are fixed real numbers, includes a great variety of situations, not just sets of priors with a specified collection of moments. Here are some of them:

- given bounds on quantiles ($\Theta = R$, $H_i(\theta) = I_{(-\infty, a_i)}(\theta)$);
- given bounds on prior probabilities ($H_i(\theta) = I_{K_i}(\theta)$, $K_i \subset \Theta$);
- given bounds on marginal probabilities of data ($H_i(\theta) = \int_{X_i} l_x(\theta)\, dx$, $X_i \subset \mathcal{X}$).

Notice that restrictions of the above classes can be accomplished, without exiting the Generalized Moment Classs, if π is taken to belong to a *Contamination Class* or to a *Mixture Class* like the ones of Example 7 in Berger's paper. Indeed it is easily seen that such restrictions can be incorporated without affecting the linearity of the constraints (1) and the linear ratio form of ψ.

Therefore the procedure for optimization within a Generalized Moment Class outlined in Betrò, Męczarski and Ruggeri (now 1994; to appear in *J. Statist. Planning and Inference*) provides a rather general tool for robust Bayesian analysis.

A. DASGUPTA *(Purdue University)*

As always, Professor Berger has made a contribution that is illuminating, informative, very enjoyable, and frequently provoking. Especially useful and gratifying is the extensive bibliography, a point that deserves mention. On my part, I will do two things: I will elaborate a little on a few points made by Professor Berger in his article, but I will devote practically all of my time and space to a number of points not explicitly made in the article.

Exactly what constitutes a study of Bayesian robustness is of course impossible to define. It seems, however, that Bayesians and others alike clearly appreciate the value and importance of a study of Bayesian robustness. A few years ago, after a talk given by Persi Diaconis, Herman Rubin stood up and said that all statisticians should work only on problems of Bayesian robustness. I feel less inclined to go that far, but the comment signifies the importance of research in this area. Importance and usefulness are sometimes completely different things, however. Ultimately the value of statistical work will be judged on the basis of whether people will use methods arising out of this work. It seems natural and actually nearly inevitable on hindsight that work on Bayesian robustness started out in the form of sensitivity analysis. In fact, for a while, that is practically all one saw. These were important on a number of grounds: they certainly helped clarify questions regarding when posterior robustness will usually obtain, occasionally they helped understand the role of the dimension of the data, and to me personally they shook us by our knees and showed that apparently abstract mathematics can provide wonderful tools in obtaining answers: moment theory, Choquet capacities, operator differentials, these have been enormously useful in Bayesian robustness problems. It is not clear, at least to me personally, and on this I think I differ with Professor Berger, that beyond that sensitivity studies have any transparent and concrete use. I have not found a really satisfactory answer in my mind to the question of what should one do with the range of a posterior mean. I know there are plausible answers: continuous refinement, or use of the range as a credible interval by itself, etc. Classical robustness grew into a successful and flourishing area because they were able to answer to at least a reasonable degree the question of what is robust. All of this research would have been most likely much less influential if the results only

went as far as saying the sample mean is not a good thing to use unless you have pretty much normally distributed data. They were successful in providing alternatives that were apparently acceptable: the obvious energy that went into studies of M and L estimates is a testimony to that. Having said that, it is not at all clear what would be a criterion for prescription in our area. In my paper with Mei-mei Zen, I had shown that a posterior minimax choice, coincidentally but fortunately, results again in a Bayes procedure, Bayes with respect to one of the priors one started with. But undoubtedly, we will not see this phenomenon very often in other problems. Professor Berger's due concern about whether "robust choices" are not silly from a "real Bayes" perspective therefore has to be regarded with a lot of seriousness. In spite of that, more effort should probably go into this issue than has so far.

Another point that Professor Berger (implicitly) makes and one with which I fully agree is that it is now time for us to go beyond the canonical problems. These are also the problems that are the hardest to "solve". The frequentists have the blessing that the technology of large sample theory is now so advanced that even the nastiest problem with the dirtiest model is amenable to some structured theory in the form of limit theorems. The issue of finite samples aside, this is nice within the frequentist domain. Models that people really care about: all kinds of censoring, various regressions (Cox models, many more), and the now popular semiparametric models are just a few examples that really do need to be looked at. Will the choice of a link function matter? To what extent? These are entirely different sensitivity questions we can and should ask. It may very well be that no answers are possible: a consequence I will personally find very unfortunate. But we don't know that. There is also the well understood need of a simultaneous likelihood-prior-loss(?) robustness study. But I have my doubts that much structure will ever come out in this problem: we will be only successful in seeing what we knew we will see. I will have something more to say on this later. Let me now touch on a few things that are not explicitly addressed by Professor Berger, but do seem to be natural. I have no real doubts in my mind that questions I ask myself are often "infected" with a dash of frequentism; I therefore caution the typical reader of this article that parts of what I will now show can appear to be grotesque and strange. I will give some precise theorems, mostly without proofs, because of

space and also because they will all appear elsewhere.

1. *Robustness with respect to the likelihood: the role of dependence.* One can make a very short case for this: real data are never really iid. Many questions suggest themselves: is the iid inference reasonably robust against moderate dependence, do noninformative or "robust" priors give some protection, etc. I will only talk about the first issue here, in the form of two results, one of which is rather surprising. For the rest of this part of the discussion, let us have the implicit understanding that we have data coming from a Gaussian process.

Theorem 1. *Consider n observations coming from a weakly stationary Gaussian process with mean of each observation equal to t, and a covariance kernel given by $r(i,j) = r(|i-j|)$; let us pretend as though the covariances are arising from a continuous time function $r(x)$ (as is the case with how L estimates are defined, for instance). For estimating t using mean squared error, let $R(n)$ stand for the ratio of the Bayes risks of the iid case Bayes estimator under the true model and the iid model; assume a standard normal prior for t in this. Then $\lim R(n)$ (as n tends to infinity) exists under (frequently satisfied) conditions, and furthermore the limit equals $1 + 2.\int_0^\infty r(x)dx$.*

Corollary 1. *If data are coming from an Ornstein Uhlenbeck process that we mistakenly think as iid, then we will suffer a Bayes risk 3 times as large as for the iid case even in the limit.*

Proof. The covariance kernel is $r(x) = \exp(-|x|)$. ◁

This is somewhat disappointing; even an exponentially decreasing covariance results in a loss 3 times in magnitude. With slowly varying covariance kernels (see Karamata or Feller) as is the commonly made assumption in the frequentist world, the loss will often be infinitely more! One can state a more general version of this result in terms of two general kernels, not restricting to the case when one of them is the iid case. This result, because it talks about Bayes risk, is half-frequentist. The next result is purely conditional.

Theorem 2. *Consider a* $100(1-\alpha)\%$ *credible interval that is the correct Bayesian interval if the data were iid. Consider the posterior probability of this interval when the process in reality is an AR(1) with parameter* ϕ. *Then, as* n *tends to infinity, this posterior probability converges a.s. to* $2.\Phi((1-\phi).z_{\alpha/2}) - 1$, *it being understood that the underlying probability space uses the true marginal distribution as the measure.*

Corollary 2. *Mild autoregression is not much problem in this case if we have lots of data, but being close to the unit root case is disastrous.*

My first version of Theorem 2 was corrected by N. D. Shyamalkumar, a graduate student at Purdue.

2. *Group decisions: will they usually agree.* Clearly this smacks of frequentism; but the apparently neat nature of the result following might make up for something.

Theorem 3. *Consider estimating a univariate normal mean* t *by using a credible interval with* n *iid observations. Bayesian 1 has* $N(0, \tau_1^2)$ *and Bayesian 2 has* $N(0, \tau_2^2)$ *as prior for* t. *Denote by* $C_1(X)$ *the* $100(1-\alpha)\%$ *interval that Bayesian 1 will use if left alone. Let* $P_2(X)$ *denote the posterior probability of* $C_1(X)$ *if Bayesian 2 is forced to use* $C_1(X)$ *although it is not his Bayes solution. Then, for any* β *such that* $1-\beta < 1-\alpha$, $P_t\{P_2(X) < 1-\beta\}$ *converges to zero for any* t, *as* n *tends to infinity, and in fact*

$$\lim\{n.\exp(n^2\gamma^2/2).P_t\{P_2(X) < 1-\beta\}\} = \sqrt{2/\pi}.1/\gamma,$$

$$\text{where } \gamma = (\beta-\alpha)/\left\{\left(\frac{1}{\tau_1^2} - \frac{1}{\tau_2^2}\right)^2 z_{\alpha/2}\phi(z_{\alpha/2})\right\}.$$

Some remarks are necessary because the statement can be baffling. The Theorem asks how often use of the other Bayesian's inference will lead to bad performance if we have lots of data. The convergence to zero is not surprising. What is surprising is the extraordinary fast convergence and even more the fact that under each t, the limit on the right side is

the same. In other words, so far as pointwise limits are concerned, t vanishes altogether from the field in the long run. Is there any role of t at all? Yes indeed; the convergence is not uniform!!

Table 1 gives some numbers for finite n. They clarify the finite case to some extent. I refrain from discussing.

$$1-\beta = 0.7 \qquad 1-\alpha = 0.9$$
$$\tau_1^2 = 1 \qquad \tau_2^2 = 4$$

θ	$n=2$	$n=5$	$n=10$	$n=20$	$n=30$
0.5	0.00146033	8.1124×10^{-13}	0.	0.	0.
1.0	0.0116026	1.37753×10^{-9}	0.	0.	0.
1.5	0.0590335	6.91185×10^{-7}	0.	0.	0.
2.0	0.196045	0.000103836	0.	0.	0.
2.5	0.440885	0.00477846	6.43929×10^{-15}	0.	0.
3.0	1.0	0.0703178	4.52661×10^{-10}	0.	0.
3.5	1.0	0.361147	2.75817×10^{-6}	0.	0.
4.0	1.0	0.777161	0.00152334	0.	0.
4.5	1.0	0.969992	0.0835089	0.	0.
5.0	1.0	1.0	0.578982	2.22045×10^{-16}	0.
5.5	1.0	1.0	0.962497	2.44078×10^{-9}	0.
6.0	1.0	1.0	0.999612	0.00015012	0.
6.5	1.0	1.0	1.0	0.0839436	0.
7.0	1.0	1.0	1.0	0.80429	1.99618×10^{-12}
7.5	1.0	1.0	1.0	0.99901	0.0000134143
8.0	1.0	1.0	1.0	1.0	0.0721139
8.5	1.0	1.0	1.0	1.0	0.899443
9.0	1.0	1.0	1.0	1.0	0.999971
9.5	1.0	1.0	1.0	1.0	1.0

Table 1. *$P_\theta(P_2(\bar{X}) < 1-\beta)$ for given n and θ*

Here is another result, which follows on use of Dini's theorem, but I am almost certain it follows from known results on the posterior CLT or even the Portmanteau theorem.

Theorem 4. *In the set up of Theorem 3, let $P_1(X)$ and $P_2(X)$ denote the posterior probabilities of the common null hypothesis H_0: $t < 0$ under the two stated priors. Then the P_t joint distribution of $(P_1(X), P_2(X))$ converges weakly, as n goes to infinity, to a*

singular distribution supported on the main diagonal of the unit square.

Remark. Of course this is expected. The result can be stated in far greater generality; consequences of such results are that with a large probability, the two Bayesian's answers hang very close together. The case of k statisticians can be done with appropriate formulation.

The valuable cases are cases where the two Bayesians differ more seriously, like normal vs. t. I believe similar results are valid there and they will appear elsewhere.

3. *Robustness with respect to outliers: Shrinking neighborhoods.* Again, so far as concrete theory goes, the frequentists within their own domain are ahead on this. I will give only one result, purported to show that it is seemingly imperative to use shrinking neighborhoods, at least in this formulation.

Theorem 5. *Consider data that are* $N(t,1)$ $100(1-\varepsilon)\%$ *of the times and the rest of the times we see an outlier at some* x. *In principle,* ε *can depend on the sample size* n. *Consider estimating* t *using a* $N(0,1)$ *prior and mean squared error as criterion. Let* $r(n)$ *denote the Bayes risk of the estimate that would be Bayes in the absence of outliers* ($\varepsilon = 0$). *Then* $r(n)$ *is unbounded unless* $\varepsilon = O(1/n)$; *it converges to zero only if* $\varepsilon = o(1/n)$. *If* ε *is* $O(k/n)$ *(meaning exact order), then* $r(n)$ *converges to* k^2x^2, *and hence with the usual definition, the Influence of an outlier is unbounded.*

Many other questions can be raised here. I would not go into them.

To sum it up, this is another profound contribution by Professor Berger to the profession. This made me think, helped me to understand. In chapter 14 of his book, Nicholas Young (the operator theorist) writes: "a mathematical model never describes the behavior of a system exactly...How well will an aircraft stand up to unpredictable external disturbances — gusts of wind or a stewardess wheeling a drinks trolley down the aisle? One might wonder why the idea (of robust designs) was such a late starter. Part of the answer must be that engineers were unaware of the relevant theorems and operator theorists of the engineering problem. The connection is developing rapidly ...". Perhaps there is some interest in simply exploring as a matter of scientific truth whether

classical and Bayesian robustness will lead to common grounds: can one justify use of M estimates from a robust Bayesian viewpoint? From a strictly likelihood principle point of view, evidently not. But perhaps from another viewpoint. It is my feeling that that can only be good for the community as a whole. I offer my deepest gratitude to Professor Berger for again doing what he always does: open new doors.

PAUL GUSTAFSON and LARRY WASSERMAN
(Carnegie Mellon University)

1. *Introduction.* We congratulate Professor Berger for this superb review of robust Bayesian inference. Professor Berger may rightly be called the leader of this field and this paper is a useful summary of some of the contributions he and others have made. We agree with his main points. We thus take this opportunity to elaborate on some points that he did not have space for. We also raise the question of why robust Bayesian techniques are not in routine use.

2. *Foundational Issues.* Coherence arguments are often used to justify the Bayesian approach. These argument may be relaxed to justify robust Bayesian inference. The most thorough recent account is Walley (1991). So it seems that robust Bayesian inference is on secure ground. There is one annoying problem that arises, however. The problem was dubbed "dilation" by Seidenfeld and Wasserman (1993). The problem is that bounds on probabilities may become uniformly more precise by conditioning. Specifically, consider a set of probability measures Γ, let A be an event and let $\mathcal{B} = \{B_1, \ldots, B_n\}$ be a partition. We say that $\mathcal{B}$ dilates A if for $i = 1, \ldots, n$,

$$\inf P(A|B_i) \leq \inf P(A) \leq \sup P(A) \leq \sup P(A|B_i)$$

with at least one of the outer inequalities being strict for some i. The infimums and supremums are over Γ. When dilation occurs it seems that there is incentive to not observe B_i, which seems counter to the usual Bayesian philosophy. Seidenfeld and Wasserman (1993) showed that dilation is not pathological; most sets of probabilities dilate. We might add that even if the dilation is not uniform, that is, even if the bounds expand only for some B_i, then there is still cause for concern. In a quantile class, for example, we begin with precise probabilities

on given sets $A_1, \ldots, A_k$. It is somewhat embarrassing to start with $\inf P(A_i) = \sup P(A_i) = p_i$, say, and then find after collecting data we have $\inf P(A_i|x) < p_i < \sup P(A_i|x)$ even if this doesn't happen for all x. We wonder if Professor Berger has any advice on what do to in these circumstances.

3. *Local Sensitivity.* We would like to elaborate on Professor Berger's comments regarding the local assessment of sensitivity via functional differentiation. The hope is that local methodology will lead to simple, easily-computed diagnostic measures of the sensitivity of posterior distributions to priors. As a reflection of the overall sensitivity of the posterior distribution to the prior, it seems reasonable to start with the limiting ratio

$$s(\Pi, Q; x) = \lim_{\epsilon \downarrow 0} \frac{d(\Pi^x, Q_\epsilon^x)}{d(\Pi, Q_\epsilon)},$$

where Π is the base prior, Q_ϵ is the ϵ- contamination of Π by Q, superscript x denotes Bayesian updating after observing data x, and d (not necessarily a metric) measures distance between distributions on the parameter space. To permit a variety of deviations from the base prior, we consider

$$s(\Pi, \Gamma; x) = \sup_{Q \in \Gamma} s(\Pi, Q; x), \tag{1}$$

as a measure of local sensitivity, where Γ is a class of priors. This is a nonparametric analogue of the diagnostic proposed by McCulloch (1989), and is similar in spirit to the frequentist diagnostics of Cook (1986). When d is total variation distance, $S(\Pi, Q; x)$ can be interpreted as the restricted norm of a Fréchet derivative. Diaconis and Freedman (1986) considered the total variation case, though not in the context of assessing sensitivity to the prior.

As discussed in Gustafson and Wasserman (1993), there is an asymptotic problem with using (1) as a sensitivity diagnostic. Specifically, under a weak condition on Γ, $s(\Pi, \Gamma; x)$ will diverge to infinity as the sample size increases, even though we know the prior becomes less important as more data are collected. This problem persists if ϕ-divergence (which includes Kullback-Leibler divergence as a special case) replaces total variation distance, or if geometric contamination replaces ϵ-contamination (the geometric contamination of density π by

density q is the density proportional to $p^{1-\epsilon}q^{\epsilon}$). Thus it appears to be very difficult to construct a sensible local sensitivity diagnostic based on the whole posterior distribution, unless we are willing to restrict ourselves to a parametric class of priors.

An alternative strategy explored in Gustafson (1993a) is to restrict attention to a particular posterior quantity, for instance the posterior expectation of a function $g(\theta)$ of the parameter. To use functional differentiation, we must embed the class of priors in a linear space. One way to do so is to consider (unnormalized) prior densities of the form $\pi + u$, with u a nonnegative function. Based on invariance considerations, a sensible way to measure the size of u, or equivalently the discrepancy between π and $\pi + u$, is by

$$\text{size}\,(u) = \|u/\pi; \Pi\|_p = \begin{cases} \left(\int_\Theta (u/\pi)^p d\Pi\right)^{1/p} & : p < \infty, \\ \operatorname*{ess\ sup}_\Theta u/\pi & : p = \infty. \end{cases} \tag{2}$$

With respect to this norm, we can differentiate $T^g u$, the posterior expectation of $g(\theta)$ when the prior density is $\pi(\theta) + u(\theta)$. Under weak conditions, we obtain the Fréchet derivative at 0, $\dot{T}^g(0)$, and its norm:

$$\dot{T}^g(0)u = Cov_\pi^x\left(g(\theta), \frac{u(\theta)}{\pi(\theta)}\right), \tag{3}$$

$$\|\dot{T}^g(0)\| = \max\{\|a^+; \Pi\|_q, \|a^-; \Pi\|_q\}, \tag{4}$$

where

$$a(\theta) = \frac{(g(\theta) - \rho_g)\pi^x(\theta)}{\pi(\theta)}, \tag{5}$$

with $\rho_g = E_\pi^x g(\theta)$, and $a^+ = \max(a, 0)$, $a^- = -\min(a, 0)$. Here q is the extended real number satisfying $p^{-1} + q^{-1} = 1$. We have (4) as a measure of local sensitivity of the posterior expectation to the prior, in analogy to (1).

The choice of p in (2) appears to be very important. In the case $p = 1$, the underlying prior structure is based on ϵ-contamination, and the quantity (4) has been investigated by Ruggeri and Wasserman (1993), Sivaganesan (1993c), and Srinivasan and Trusczcynska (1990). When

$p = \infty$, the underlying distance is the density ratio metric, and there are connections to the work of Ruggeri and Wasserman (1991). The problem in the former case ($p = 1$) is that (4) is of constant (or increasing) order asymptotically, and therefore does not reflect the asymptotically diminishing role of the prior. On the other hand, when $p = \infty$, (4) asymptotically has no dependence upon the base prior. In practice this means the norm does not reflect the degree of data-prior conflict, again in contrast to what we would like to see in a diagnostic. Both these problems are obviated by taking an intermediate value of p; a convenient choice is $p = 2$, at least when the parameter space is one-dimensional. For higher dimensions, we find that the norm vanishes asymptotically only when p is larger than the dimension of the parameter space. This is an instance of the difficulty with high-dimensional parameter spaces alluded to by Professor Berger in Section 4.3.2. For a more detailed discussion of the pertinent asymptotics, see Gustafson (1993b). A practical strategy is to perturb one-dimensional aspects of the prior (marginals or conditionals) in turn; then using $p = 2$ suffices.

4. *Why Isn't Everyone a Robust Bayesian?*. Efron's (1986) paper "Why Isn't Everyone a Bayesian?" raised an important question: why isn't Bayesian inference the dominant form of statistical analysis, given its secure philosophical and logical foundations? Things have changed since 1986 and, while it might not be that Bayesian statistics is *the* dominant mode of inference, it is much more so than seven years ago. This is largely due to advances in statistical computing. Where philosophical arguments fail to convince, ease of use does.

So why isn't everyone a robust Bayesian? Or for that matter, why isn't every Bayesian a robust Bayesian? Our impression is that everyone who uses Bayesian methods agrees that some form of sensitivity analysis is important. But few use the formal techniques of robust Bayesian inference. Most often sensitivity analysis is not done or, at best, the data analyst tries out a few priors and stops there. Why haven't robust Bayesian techniques caught on?

We believe there are two reasons. First, there is the computational burden. Robustness is most important in complicated models. Formally constructing classes of priors and carrying out all the computations in complicated models is a serious burden. Even using an ϵ-contaimination

class with unrestricted contaminations usually involves several high-dimensional maximizations.

Second, and perhaps most important, we tend to overlook the fact that people need sensitivity analysis *before* they construct a prior. That is, we need to know when it is worthwhile to even bother constructing a real prior, rather than using some convenient default prior. To answer this question, we need methods for assessing sensitivity to default priors which are usually improper. The usual techniques cannot be used in this case (Wasserman 1993) though ongoing work (Srinivasan and Wasserman) suggests that it is possible to derive new diagnostics for this case.

It may be that local diagnostics are an answer to both problems. Sometimes they are quicker to compute and it may be possible to extend them to handle improper priors. If so, then the development of local sensitivity methods may be an important direction for robust Bayesian inference. In his discussion of Diaconis and Freedman (1986), Professor Berger extolled the potential value of local diagnostics. We are interested to know if he still feels this is a fruitful direction.

JOSEPH B. KADANE* and CID SRINIVASAN
(Carnegie-Mellon University)

Professor Berger is to be congratulated for his excellent review which pulls together so many disparate strands of work. That so much has been accomplished is a great credit to Professor Berger, and the students and collaborators he has inspired.

We wish to share a question about this body of work, however, with a view toward strengthening Bayesian robustness work in the future. We start with a likelihood, prior, and loss function believed to be accurately assessed, but wonder if the quality of the decision about to be made might be severely affected by errors in the inputs. The "typical" way to address this question in the literature Professor Berger surveys is to

* The research of Professor Kadane was supported in part by the following grants; ONR: N0004-89-J-1851, NSF: SES-9123370, DMS-9005858 and DMS-9302557. Professor Srinivasan was supported in part by NSF grants ATM-9108177 and DMS-9204380.

examine the extent to which the posterior mean changes over a class Γ of prior distributions.

We agree that sensitivity with respect to loss functions and likelihoods would also be nice, but concentrate on prior sensitivity, as does the literature. To use the posterior mean suggests that the decision problem is an estimation problem, and that squared error loss, or negative squared error utility, is being used. Given the declaration of such a loss function, it seems to us that what must be of concern is the sensitivity of expected loss (or expected utility), not the sensitivity of the optimal decision. To us, the optimal decision is merely a tool in achieving high expected utility or low expected loss. Two examples illustrate that expected-utility robustness is not the same as decision-robustness.

Example A. Suppose the base prior is normal (0,1), and that the class of priors allows for epsilon symmetric contamination (whose mean exists) of the base prior. Take the likelihood to be flat (*i.e.* no data), and loss to be squared error. Under each member of the contamination class, the posterior (and prior) expectation is zero, so the optimal decision is maximally robust. However, the expected loss, which is the variance, can be arbitrarily large, as can be seen from the contamination

$$X_n = \begin{cases} -n & \text{with probability } 1/n \\ 0 & \text{with probability } 1 - 2/n \\ n & \text{with probability } 1/n \end{cases},$$

so the variance of the contaminated random variable is $(1-\epsilon) + 2\epsilon n$ which goes to ∞ as $n \to \infty$, for every $\epsilon > 0$. Thus decision-robustness does not imply expected-loss robustness. $\triangleleft$

Example B. Suppose again the likelihood is flat, but that loss here is absolute error. Suppose the base prior has the following form: with probability $1/2$ it is uniform on $[0,1]$, and with probability $1/2$ it is uniform on $[2,3]$. Then the optimal decision is the median, any number in the interval $[1,2]$. Now suppose an ϵ-contamination of this prior. By choosing different contaminations, the optimal decision can easily be moved to $1-2\epsilon$ or $2+2\epsilon$, but in neither case will it's effect on the absolute error be substantial. So here there is high sensitivity of the optimal decision, but robustness of expected loss. Hence expected-loss robustness does not imply decision-robustness. $\triangleleft$

There is one case in which the difficulty we perceive does not apply, which is when posterior probabilities of sets are reported. Here the utility function is often 2-valued, say 0 if the set does not contain θ and 1 if it does. Then expected utility is exactly the posterior probability of the set, so robustness of one is the same as robustness of the other. We hope that future work in Bayesian Robustness will take expected-loss robustness more seriously.

MICHAEL LAVINE *(Duke University)*

1. *Overview.* Let $X_1, \ldots, X_n$ be an exchangeable sequence of observations in a space $\mathcal{X}$ and let $\mathcal{F}$ be the set of all probability distributions on $\mathcal{X}$. A prior μ is a probability measure on $\mathcal{F}$. In a parametric analysis a small subset $\mathcal{F}_\Theta \equiv \{F_\theta : \theta \in \Theta\} \subset \mathcal{F}$ is designated and $\mu(\mathcal{F}_\Theta) = 1$. Berger says "Robust Bayesian analysis is the study of the sensitivity of Bayesian answers to uncertain inputs." In this case the inputs are the specification of $\mathcal{X}$, the assumption of exchangeability, the data, μ and possibly a loss function. Berger's paper is primarily devoted to the sensitivity of Bayesian answers to the way μ distributes mass on $\mathcal{F}_\Theta$ without questioning the assumption $\mu(\mathcal{F}_\Theta) = 1$. This is entirely appropriate for a review paper because almost all the work done so far in Bayesian robustness is devoted to the distribution of mass on $\mathcal{F}_\Theta$. However, in my experience and for reasons sketched near Equation (4.12), posteriors are much more sensitive to letting μ put a little bit of mass a little bit away from $\mathcal{F}_\Theta$.

A similar situation holds in regression problems. Let $r(x)$ be the expected value of Y when the covariate is equal to x and let $\mathcal{R}$ be the set of all possible regression functions. A prior μ is a probability measure on $\mathcal{R}$. In a parametric analysis the prior is supported by a small subset of $\mathcal{R}$. Inferences may be very sensitive to moving a little bit of prior mass a little bit away from the subset.

Of course inferences may also be highly sensitive to the specification of $\mathcal{X}$, the assumption of exchangeability, the data and the loss function. Therefore I believe that the major contributions of Bayesian robustness in the future will be in these areas and not in sensitivity to the way mass is distributed on $\mathcal{F}_\Theta$.

2. *Nonparametric Bayes.* When there is no parametric subset $\mathcal{F}_\Theta$ deserving all the prior mass statisticians sometimes resort to nonparametric methods. Berger says "Bayesian nonparametrics can be considered to be an approach to automatic robustness with respect to model choice." Nonetheless, a nonparametric analysis based on a single prior μ does not address the question of sensitivity to μ and still leaves a role to play for Bayesian robustness. Nonparametric priors, including Dirichlet processes and Gaussian processes, have features that are chosen for convenience rather than because they accurately represent subjective prior belief. Therefore it is still desirable to investigate the extent to which inferences vary over small changes to μ or over a class of reasonable priors.

3. *Inherently Robust Procedures.* In Section 2 Berger talks about inherently robust procedures, saying "use of distributions with flat tails tends to be much more robust than use of standard choices." Some practitioners consider this to be a reason to use flat-tailed distributions. However, in problems where robustness is lacking for standard choices, the fact remains that there are some distributions consistent with the elicited inputs that give widely different conclusions than other distributions also consistent with the elicited inputs. The fact that a neighborhood of priors exists over which the conclusions are robust is not comforting if there are other reasonable neighborhoods of priors over which the conclusions are highly variable.

But perhaps part of the input is a statement by the investigator that "my posterior is robust to small changes in the prior or the data." Then it may be sensible to ask whether there really are models and priors for which the posterior is robust and, if there are, to use them. However, in view of the well known inability of humans to estimate posteriors accurately without formal calculations, it is not clear how much weight should be given to the input that posteriors are robust.

4. *Motivation for Bayesian Robustness.* In arguing for Bayesian robustness Berger says "... the question of interest may not depend on accurately knowing many of the parameters. ... there may only be a few crucial quantities that need to be elicited. Robust Bayesian techniques can help to identify these quantities." This does seem to be one of

the promises of robust Bayesian analysis. But Berger cites no examples. Does he know of any?

ANTHONY O'HAGAN *(University of Nottingham)*

Jim Berger has given us an elegant, thoughtful and scholarly review of recent work in robust Bayesian statistics. There is an impressive number of references, and the fact that the great majority have been published in 1990 or later is a powerful testament to the vigour of current research in this field. Yet although there is a lot of work I am just a little disturbed at the lack of progress. One area of little progress is in tackling multivariate problems. Unless one can show how it can be built into more complex problems, a robust analysis of a model with one parameter is a toy, not a practical tool. There are a few references to genuinely multivariate work on robustness, but the vast bulk is still one-dimensional. Another area of slow progress is applications. Toy theory can only lead to toy applications, so the two are linked, but the applications picture is a little better. Particularly in the area of using heavy-tailed modelling, some of the papers cited in section 2.2 are putting these models into practical effect. Here also there is some multivariate work, and I would like to add two more references: O'Hagan and Le (1993) and Le and O'Hagan (1993) study a class of bivariate heavy-tailed models.

The papers of O'Hagan (1994) and Goldstein and Wooff (1994) are cited as examples of applications, but I think it is important to point out that these involve ideas of robustness that are not mentioned in this review, although its coverage is otherwise impressive. In O'Hagan (1994) I stressed the importance of modelling and elicitation to robustness. A major reason for studying robustness is the difficulty of specifying or eliciting prior distributions (or likelihoods or utilities) accurately. Lack of robustness is a function of the inaccuracy or vagueness in those specifications. Surely the primary way of achieving robustness is to improve the elicitation process. The key to this is asking the right questions. Modelling needs to express the problem in terms of parameters that are meaningful, and about which prior information is most easily expressed. Elicitation then needs to ask clear questions about summaries of the prior distributions that express the most firmly held beliefs. There is depressingly little published work on elicitation, but that might change if it were recognised as a vital ingredient in the Bayesian robustness cru-

sade. Garthwaite (1993) and Garthwaite and Dickey (1993) are useful recent references.

Both O'Hagan (1994) and and Goldstein and Wooff (1994) concern the application of Bayes Linear methods, which I believe can also contribute to robustness. The idea here is to use a limited specification of the prior distribution and likelihood, and so to limit the number of summaries that must be elicited. Michael Goldstein has for too long been a lone voice advocating Bayes linear methods. He has been developing it into a complex methodology capable of handling substantive real applications.

Finally, I object to Jim Berger's use of the word 'objective' in connection with vague or improper priors. There never was and never will be anything objective about such a practice. There cannot be a unique improper or 'reference' prior to represent 'ignorance', 'letting the data speak for itself', or whatever your favourite phrase might be to try to justify these distributions. Which 'objective' distribution one uses will always be a matter of subjective choice.

These grumbles apart, let me repeat my congratulations to Jim Berger for an excellent and authoritative paper.

WOLFGANG POLASEK *(University of Basel)*

Jim Berger is not only a leading person in the field of Bayesian robustness, he also has the ability to review the work in this area periodically from many subtle viewpoints. Given the limitation of any article, his review is thorough and therefore I want to concentrate my discussion on rather general issues of Bayesian robustness (B.r.).

1. *The B.r. profile.* Following the structure of the paper of J. Berger, I want to express my personal profile toward B.r. in the following ascending order of importance:

1) Inherent robustness;
2) Local robustness (Diagnostics and Sensitivity);
3) Global robustness;
4) Computing robustness;
5) Strategies for modeling robustness.

This list also shows that robustness covers almost the entire spectrum of applied statistics. Each of the 5 subjects can be further divided in

estimation, testing and prediction. Point 3) was renamed (after a suggestion of Leamer 1978) because of earlier attempts to distinguish different approaches to B.r. Note that only the last point - modeling robustness - is not explicitly discussed in J. Berger's overview, it is only mentioned in the last section with the promising title 'future directions'. The absence is simply explained, since we rather need much more experiences on B.r. techniques and more available software.

In the last 50–70 years statistics as a science has grown substantially. As a newcomer in science of this century others now respect the adolescent strength of statistical methods which gained a lot of smartness using the advances of modern computing. Now as statistics stands on its own feet it is looking for robustness. This means that trying to be good looking to others depends on the ability to withstand harder and lighter blows on the original intention of statistical methods.

The last decade in B.r., a decade which can be also considered to be pretty much the first serious decade in methodological research, was dominated by the advancement of theory. Applications are and have been rather scarce in this decade, with exceptions of a few interesting approaches like Grieve (1985). So we hope to see rather more applications in B.r. for the future but also more interaction between theory and practice of B.r. This does not mean we don't need more theory in B.r.B.r. is now very much in a state of an experimental methodology. I expect more impact on applications if we have more elegant ways to compute and communicate Bayesian or classical robustness.

Mathematically, the most challenging area is the global sensitivity field in B.r. The nerve of any statistical analysis is hit here. As Bayesians we are full aware that any choice of a model is highly subjective which is only justified by computational convenience and driven by a wide acceptance of appliers. Naturally, we want to know if our conclusions hold if we make distributional assumptions in the neighborhood of the chosen one. The neighborhoods of distributions is the key for the range of B.r. If it is to wide, then we get always a wide range of answers. If the class is to narrow then B.r. becomes meaningless. Also the curse of dimensionality is felt here. High dimensional variation might have a tremendous effect. Reading over all theoretical results obtained so far one gets the impression that a theory is waiting out there to be detected. Priors, outliers, likelihood and model choice might be a moving

pointer on a unknown scale. The only thing which I missed in the review is the connection to classical robustness. Since the prior and the likelihood is not a disjoint set of assumption there is a need for some more interactions in robustness in general. In Polasek (1992) I called this combined (classical and Bayesian) robustness effort 'joint sensitivity analysis'.

2. *Why B.r.?*. B.r. right now comes rather as a meta language than a client language in statistics. This means that B.r. now tells the experienced Bayesian analyst what are the pro and cons of a certain Bayes procedure. B.r. is not developed so far to help an average practioner of statistics to find his or her sensitive aspects of an analysis. Therefore I want to express the following list of challenges:

(i). B.r. today is beyond average client apprehension. Most clients want a clear cut answer even they know that the basis of a sampling process is a very random one. If statistics is the essence of modern science then a statistical answer should be straightforward and any if and thens are a sign of a less experienced if not to say low expert science.

(ii). B.r. is a small sample problem, in general a limited d.f. problem. As for classical robustness, large amount of data will bury many considerations of robustness. This statement is only true if the desire and appetite for more elaborate models does not increase with sample size (as it will usually do). More elaborate models come with a larger number of parameters and therefore we might find us again in a small sample situation. This type of behavior I call the curse of large sample modeling or simpler the limited degree of freedom (d.f.) problem. Limited d.f. problems might again need robustness consideration, but when the number of parameters become large the robustness considerations multiply as well.

(iii). B.r. is essentially an 'inference mapping' problem. What do we learn from a sample, i.e. what is the message from the prior to the posterior, and how do we communicate this learning process? Leamer calls this 'The mapping is the message', a certainly true statement if clients would generally like to enjoy statistical sophistications.

(iv). B.r. is a 'meta-methodology' problem. Simple statistical inference depends on the assumption of meta-hypotheses, like normal distribution or independence. Therefore advanced statistical methodology has

to explore the dependency of statistical inferences on assumed meta-hypotheses. Therefore semi- or non-parametric statistics has become an important topic. Despite a bad reputation of statistics as science in general, people like to be fooled by statistical statements if they only would serve their own interests. 'How to lie with statistics' by D. Huff is cited by many people who also enjoy living with the shortcomings of simple statistics, rather than to fuzz around with more scientifically honest and complicated statements which might tire the readers attention and distract from the main focus of an empirical result. Therefore B.r. has the tendency to be a research topic between the priests of the statistical science community than between the priests and the laymen.

(v). B.r. is a certain kind of intellectual refinement (game?) If certain statistical statements can't be made clear enough to be generally accepted what is the use to 'play games with the data', i.e. to find out under what circumstances what conclusions could have been drawn and might lead to different answers? Pushing this aspect too hard might lead to the same line of criticism we encounter in classical statistics where we have to take into account non-observed data.

3. *The future of B.r.* Despite many critical remarks there is a future for B.r.

(i). B.r. is most rewarding for multivariate problems. If one can look at the problem at hand then there is not so much gained by a B.r. analysis. Certainly more can be gained if the sensitive parts of a problem are not so obviously to detect or the parameters are so numerous that it is hard to see through.

(ii). B.r. can be used for quick and dirty Bayesian analysis. By making this statement I don't recommend this. But people who don't want to think about their problems too much in advance and prefer to see 'the data' right away might find some help by getting immediately also first B.r. diagnostics.

(iii). B.r. has a non Bayesian flavor. If the language of uncertainty is probability theory then we should learn to express our modeling uncertainty in an appropriate way. Right now we explore the range of inference by defining convenient prior classes of distributions. Rather than classes it might be more appropriate to define distributions, like

Dirichlet processes or Polya distributions. This needs to be worked out in future.

(iv). B.r. reflects to a certain degree inappropriate current methodology. Certainly we need all these illustrious ideas as how to find out about the weak and the strong aspects of an inference problem. But in future there is a need for more efficient methodology to make B.r. more attractive for daily use.

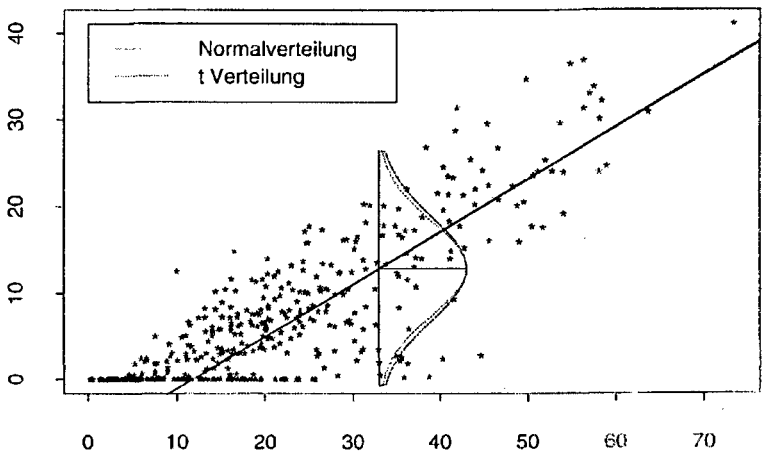
(v). B.r. will change dramatically if Gibbs sampling becomes a standard tool in Bayesian software packages. Most B.r. analyses so far depend very much on mathematical convenience. This might change quite a bit if we get more used and experienced in reading numerical outputs.

4. *Conclusions.* A viewpoint that I missed was the joint or combined view of classical and Bayesian robustness. The work of DasGupta and Studden hints in that direction. Also, I tried to push this area in Polasek (1992, 1993). While J. Berger cites many new ideas for more B.r. methods, communication problems have not been the main focus in Berger's review. A successful B.r. method will be the one which gives immediate insight into the inference sensitivity of a problem and does not need long explanations. Graphical methods would be a useful vehicle for this purpose, unfortunately, suggestions in this direction seem to be rare up to now.

Personally, I have large difficulties with the concept of non-informative priors. Bayesian statistics can be considered as the art of learning or simply as a device for information processing. It is virtually impossible to learn anything from scratch. I have not encountered any empirical analysis so far where I could not elicitate a sensible prior distribution. Therefore I claim that for any statistical analysis there exists a useful prior information. Rather than worrying about what is the most noninformative or maximum ignorance prior it is more sensible to ask what size of prior information do I need to move my conclusion from the present inference position. Therefore, I think the so-called inherent approach is rather counter-productive to B.r. If a result turns out to be extremely sensitive to prior information then there is only one thing to do: forget this! Statistics is simple no cure for ill-posed (empirical) problems. Non-informative priors is a kind of platonic love with Bayesian statistics — it misses the real fun it was designed for. Non-informative priors can

be used as scaling device in the learning process, e.g. what is the zero degree in the temperature scale if one uses a particular viewpoint. But it is not essential for our everyday life. On any scale on this earth we will find out what does it mean if is more colder or warmer or what is an enjoyable temperature.

Also I don't think that a t-likelihood assumption will be a major tool for B.r.. In Krause and Polasek (1992) we have calculated a t-based regression model for the simple but censored regression size of the flower depending on the size of stem and leaves. Figure 1 shows that both models are rather indistinguishable.



The first decade of B.r. has come in a broad spectrum and I expect that the development will continue in this pace. B.r. is a challenging problem, a nice interface to classical statistics, but still a topic in search after a permanent role in applied statistics.

CHRISTIAN P. ROBERT *(Université de Rouen)* and
CONSTANTINOS GOUTIS *(University College London)*

We first want to congratulate Professor Berger on such an extensive coverage of the multiple facets of robust Bayesian analysis and for giving us further insights on the various issues at stake in this fascinating field. We want to take advantage of this tribune to address some directions

only briefly mentioned by Berger. We discuss the ambivalent aspects of robust Bayesian analysis and, as a corollary, the problem of defining a measure of robustness. We then move to a related issue, namely the goal of re-placing robust Bayesian analysis within a decision-theoretic framework with loss functions, merely exhibiting some open problems in this direction. We conclude this discussion with remarks on a robustification of conjugate priors by mixture distributions and their wide diversity, along with indications on the corresponding inference in such setups.

1. *Robustness versus robustness.* Robustness is truly a Loch Ness monster of Bayesian Statistics: often mentioned, sometimes sighted, but never truly exhibited. The *elusive* nature of this notion comes partially from a lack of proper definition. Berger's survey has definitely reduced the area where the monster roams to a single lake (sorry, loch!) but it may be that a further reduction of its habitat would be harmful to the monster or its searchers! In fact, we do perceive two opposite trends on the literature on Bayesian robustness:

(a) a selection of the "most robust" Bayesian procedure – hereafter the champion of the (Bayesian) world – for a given *inferential* problem, with a more or less determined list of desirable, acceptable or loathsome properties;

(b) a determination of a "minimal set" of Bayesian procedures (or of prior distributions) large enough to contain all the possible inferences a sane Bayesian would choose given the information at hand and not larger in order to exclude absurd or evidently biased procedures.

The first goal is to restrict the choice of a prior, whether it is informative or non-informative, by imposing some "robustness" constraints like outlier rejection, prior insensitivity (somewhat of a contradiction with the Bayesian paradigm?), risk minimaxity etc. This is for instance the purpose of gamma-minimax analysis (see Kempthorne, 1988, as an additional reference) or of the asymptotic coverage properties of Stein (1985), for which Tibshirani (1989) showed the relevance of the reference priors (see also Liseo, 1993). The question is then: which prior should we choose among those that satisfy the imposed constraints? There is no answer if we are done with all the restrictions available in the current setup: any acceptable prior will be satisfactory (and the re-

verse!). Moreover, this type of approach is usually quite difficult and, basically, cannot be pursued outside the exponential family setup.

While we can wonder about measures of robustness and the comparison of priors in term of robustness, as we all know, "four legs good, two legs better!" This is supposed to mean that a prior with no crutch is better than a prior with two crutches... But how can we make the desiderata precise? Or, to be more precise ourselves, how can we combine robustness with the prior knowledge?

Consider first a given prior π_0 which fits the best the feelings, intuition, understanding etc. of the experimenter but, unfortunately, does not meet the robustness requirements. The selected "robust" prior could obviously be chosen as the "closest" to π_0 or, if we may speak so, as the projection of π_0 on the set $\mathcal{E}$ of acceptably robust priors. Definition? Given a functional distance d, such as the Kullback-Leibler distance, we select π_0^* which minimizes on $\mathcal{E}$ the distance to π_0. If the information at hand is compatible with a whole class of priors, $\mathcal{I}$, we can extend the previous idea as follows. For every $\pi_0 \in \mathcal{I}$, define π_0^* as the above projection and $d_{\pi_0} = d(\pi_0, \pi_0^*)$. The selected robust prior should then be the least disruptive, i.e. the projection of a prior of $\mathcal{I}$ which minimizes d_{π_0} over $\mathcal{I}$. The implementation of this idea is utterly awesome, of course! In fact, the set $\mathcal{E}$ is usually undetermined and lacks even the most basic structure for being a decent functional space... A relaxing approximation would be to replace $\mathcal{E}$ with a smaller set of manageable priors, such as hierarchical ones. In less chartered regions of the inferential space, it may be appropriate to consider only mixtures of conjugate priors (see below).

The second robustness goal is widely documented in Berger's overview and we only stop at this station to wonder whether ranges of Bayesian answers (and subsequent uses of the width of these ranges as measures of robustness) can serve as proxies for further studies in the determination of a robust prior. This is because intervals (in $\Re$) or sets (in $\Re^k$) are not necessarily available to the layman as a type of true inference while assessment of the size of a set implies a correct perception of "small" and "large" sizes, i.e. a partial knowledge of a loss function related to the problem, which is an issue considered below.

2. *Lost losses.* As mentioned repeatedly and heatedly by H. Rubin in the last thirty years, loss and prior information are indistinguishable. Therefore, a study of Bayesian robustness should be completed by a corresponding study of the other side, as rightly indicated by Berger. In fact, as long as some information is available about loss or utility, it should be translated into restrictions on the prior distribution. Once again, to implement this drift of information in practice may be far beyond our reach but it must be considered, if only for the beauty of the concept.

On a different loss front, while we may switch on the automatic pilot and select the Bayes rule associated with the quadratic loss and the Jeffreys prior, it always pays to think, if only for a while... Like it or not, robustification of a procedure is an action, hence it involves an implicit loss function. Deciding on a procedure always invites some questions. What are the consequences if it is not robust? Why choose this particular posterior measure to measure robustness? Traditionally one examines the range of posterior means or posterior probabilities, but why? What if a class Γ has an acceptable range of posterior means but an appallingly wide (or narrow?) range of posterior probabilities? Should we declare it a Good or a Bad class? One can probably answer these questions by formulating the robustness problem from a decision theoretical point of view. Dare we write down a "robustness loss"? Perhaps not, there is still too much water in Loch Ness to do so. (Still, see Robert, 1993, for a tentative approach to such losses.)

Talking about losses, a somewhat neglected topic is robustness of the reported loss of the decision. If a Bayes rule α^* minimises some posterior expected loss

$$\int L(\theta, \alpha)\pi(\theta|x)d\theta$$

an integral part of the process is the reported loss

$$\int L(\theta, \alpha^*)\pi(\theta|x)d\theta.$$

Now, it is fair to impose on the selection criteria that it should also give a reasonable range for the latter quantity. Pushing it a little bit further, from a decision-theoretic point of view, the best action α^* is

unimportant if, no matter which α we choose, we get the same expected loss. If reported loss and expected loss coincide, that begs the question whether robustness with respect to the loss is more important than with respect to the action itself.

3. *Mixed up priors.* To conclude – as we somehow perceive a growing tenseness bordering on boredom in the most attentive of our readers – we want to mention the issue of *mixtures*. This is somehow our pet topic but it keeps popping up in most setups, and robustness is one of them. Although conjugate priors could and should be dismissed as over-conventional representations of the prior information, they are still used quite a lot, either blatantly or in disguised ways. Convention? Inertia? Maybe... But there is also a definite uncertainty in the information which propels us towards the most convenient choice. Detailed reasons why conjugate priors should not be used in general can be found in the best textbooks (see, e.g., Berger, 1985). But it follows from Diaconis and Ylvisaker (1985) (see also Dalal and Hall, 1983 or even Robert, 1992) that mixtures of conjugate prior distributions are almost as general as one could wish since they approximate almost any distribution. The effect on the posterior inference can be slightly disruptive, as pointed out by Berger (1985b), but these mixtures still improve greatly upon the original conjugate priors.

As also mentioned in Berger's overview, mixtures have a kind of universality if the finiteness of the number of components is dropped, since they encompass most of the classes described in §4.3.2. But, instead of commenting any further on the mixture class as perceived in the overview, we would rather support a "zen" version of it, namely an approximation of the true prior by a mixture of conjugate distributions involving as few components as possible. We even go further and dare to predict that, in a near future, a complete theory of mixtures will come as a competitor of non-parametric techniques such as Parzen-Rosenblatt kernel estimation.

This truly brings us to our last – yes, last ! – remark. Neighbourhood classes as those mentioned in the overview seem (to us) to be the archetype of robustness classes, once a proper distance measure has been selected – the influence of the choice being minor – as they offer a comprehensive view of all the possible uses of the prior – if not the posterior – distribution. The ideas we have been widely casting around

in this discussion thus boil down to the determination of a most influential direction in these neighbourhoods, similar to the developments of Salinetti (1994). Done!

FABRIZIO RUGGERI *(CNR-IAMI and Duke University)*

There are just a few comments I want to make about this extensive, really needed, review of the recent works on Bayesian robustness. The first one refers to the posterior quantity we look at to assess robustness. Many researches have been interested in finding bounds on posterior expectations as the prior probability measures changed in a class Γ, whereas I believe that a different, very interesting, robustness problem could be faced by considering the distance among the posterior probability measures. Such a distance could depend on some topology in the space of the probability measures. Here I am presenting two methods, which check if the distribution function (or the density) of any probability measure, obtained from a prior in Γ, is in a neighbourhood of a "base" distribution (or density). In particular, the first method that I suggest, a new one, is based on the distribution bands, considered by Basu and DasGupta (1992) to define a class of priors and compute bounds on many posterior quantities; it could be worth checking if, given a class Γ of priors, the corresponding posteriors are within a distribution band containing a "base" probability measure. Besides, the width of a band, containing all the posterior distribution functions, could measure the robustness under this criterion.

A different approach is based on the concentration function, defined by Cifarelli and Regazzini (1987), and it was developed by Fortini and Ruggeri (1992, 1994). The density functions are considered to investigate if the posterior probability measures assign too much probability where a "base" probability measure P_0 does not and the concentration function gives a natural tool to inspect it. The robustness criterion is given, in this case, by taking a continuous, convex, monotone nondecreasing function $g : [0,1] \to [0,1]$, with $g(0) = 0$, and asking that any posterior probability measure P satisfies $P(A) \geq g(P_0(A))$, for all measurable subsets A. As an example (see Fortini and Ruggeri, 1994, for others), we could take $g(x) = (1-\varepsilon)x$, $\forall x \in [0,1]$, which corresponds to P being in a ε–contamination class, where the contaminating measure is any probability measure. The above requirement can be re-

stated by asking that the concentration function $\varphi(x)$ of P w.r.t. P_0 is such that $\varphi(x) \geq g(x), \forall x \in [0,1]$.

Like Berger, I have my favourite classes, the ones which have a simple interpretation and could be easily elicited by any unaccustomed user: the quantile class and the one based on conditions on the marginals, considered by Betrò, Męczarski and Ruggeri (1994). Since the former has been described in the paper, I just focus on the latter, which is defined by means of some conditions on the marginal distribution of the r.v. X, e.g. some quantiles. Since X is usually an observable quantity, it should be rather easy to express prior judgements on some of its features, obtaining an appealing class Γ. Besides, the computation of bounds on posterior expectations, as the prior varies in Γ, is quite simple, being the functional optimisation problem transformed into a finite dimensional one.

I believe that an effort should be taken in order to reduce the use of improper priors and the robust Bayesian approach could be very helpful in achieving such a goal, by asking people not to throw away the knowledge, even poor, they have, in favour of a sometimes meaningless, but trouble solving, improper prior.

Finally, I want to mention the conditional Γ–minimax approach as relevant and appropriate from a robust Bayesian point of view, because it chooses an action, among those a priori acceptable, under uncertainty in the prior, specifying a payoff which depends only on the posterior expected loss and the family Γ of priors. As two examples, DasGupta and Studden (1989) and Zen and DasGupta (1993) have considered the actions which minimise the supremum, as the prior varies in Γ, of the posterior expected loss and the posterior regret, respectively.

GABRIELLA SALINETTI *(Università di Roma "La Sapienza")*

1. *Introduction.* It is difficult to discuss an overview paper, in particular, as in this case, on a subject which has registered an explosion of interest and literature in the last decade. The paper explores the vast areas of the field, exposes the general issues and delineates the technical developments. Looking at these vast areas one feels that approaching and solving a robustness problem is often a difficult task and the literature on the subject has registered a continuously increasing number of works in different specific situations.

On the other hand Bayesian stability and robustness problems are not substantially different, in their mathematical nature, from analogous problems approached and solved in different fields such as, for example, the stochastic optimization and the moment problem.

The objective of this discussion is to point out some these connections trying to illustrate how some specific results in related fields could reveal useful in approaching Bayesian robustness problems and possibly yield an easier computational setting. Major attention will be devoted here to the use of the results on the moment problem in the global robutness with respect to variations of the prior.

However, before approaching this aspect and without entering in the mathematical details, it can be relevant to observe that robustness and stability questions with respect to loss or utility functions, but more generally robustness and stability of Bayesian decision problems, have not received an adequate attention in the Bayesian literature; this avoidance is possibly due, among other reasons illustrated in the text, to the fact that robustness analysis involving loss functions can be technically more difficult than other types of Bayesian robustness. This difficulty mainly depends on the fact that it is required to analyze the behaviour of infima functionals and argmin functionals. From the mathematical point of view these are the same functionals faced in stochastic optimization problems for which stability and robustness are a major concern. Specific and recent literature, based on an appropriate convergence of functionals and their infima, could offer possibly fruitful technical tools to approach the robustness of Bayesian decision problems; examples are in Salinetti (1994).

The rest in the following is more specific and points out the technical connection between moment problem and global robustness in different classes of priors. The potential of this connection does not seem to be adequately explored and often the connection is reduced to the moment class; this is probably due to the fact that the moment problem is linear in nature and the key Bayesian quantities of interest are usually non linear, typically ratios of linear functionals. However the use of the linearization algorithm for ratios, but more generally, convenient "conditioning", make the results of the moment problem theory applicable and reduce the robustness computations to optimization problems in finite dimension.

2. *The Moment Problem.* Consider the moment problem in the general form

$$\inf_{\mu\in\mathcal{M}}\left\{\int h d\mu : \int g_i d\mu \leq \alpha_i, \quad i \in I\right\} \tag{2.1}$$

where $\mathcal{M}$ is a convex set of finite measures on the measurable space $(\Theta, \mathcal{B})$, I is an index set, not necessarily finite, and reasonable assumptions make the quantities well defined; in particular it is assumed that the g_i are integrable on $\mathcal{M}$ for every $\mu \in \mathcal{M}$.

Explicit solutions for classes of problems (2.1) are known. Specific references are Kemperman (1972), Kemperman (1983) and for a more recent review Kemperman (1987) to which we refer here.

For $\mathcal{M} = \mathcal{P} = \{$*class of all probability measures*$\}$ we have:

$$\inf_{\mu\in\mathcal{P}} \int h d\mu = \inf_{\theta\in\Theta} h(\theta) \tag{2.2}$$

For $\mathcal{M}$ *class of mixtures* μ defined by

$$\mu(A) = \int K(u, A)\nu(du) \tag{2.3}$$

where K is a Markov kernel, i.e., for each $u \in U$, $K(u, \cdot)$ is a probability measure on $\mathcal{B}$ and for each A, $K(\cdot, A)$ is measurable, ν is any probability measure on U, we have (Kemperman (1987), Example 2.3):

$$\inf_{\mu\in\mathcal{M}} \int h d\mu = \inf_{u\in U} \int h(\theta) K(u, d\theta). \tag{2.4}$$

For $\mathcal{M}$ *density band class*, i.e., $\mathcal{M}$ class of all measures μ on Θ of the form

$$\mu(A) = \int_A \rho(\theta)\nu(d\theta) \text{ with } a(\theta) \leq \rho(\theta) \leq b(\theta), \forall\theta \in \Theta$$

with ν a σ-finite measure on $\mathcal{B}$ and $0 < a(\theta) \leq b(\theta)$ given measurable functions; we have (Kemperman (1987), Example 2.5):

$$\inf_{\mu\in\mathcal{M}} \int h(\theta)\mu(d\theta) = \int h(\theta)\rho_h(\theta)\nu(d\theta) \tag{2.5}$$

with $\rho_h(\theta) = a(\theta)$ if $h(\theta) > 0$ and $\rho_h(\theta) = b(\theta)$ if $h(\theta) < 0$.

For the so called *main moment problem* (Kemperman 1987)

$$H = \inf_{\mu \in \mathcal{M}} \left[\int h(\theta)\mu(d\theta) : \int g_i(\theta)\mu(d\theta) = \alpha_i, i = 1, 2, \ldots, n \right] \tag{2.6}$$

with $\mathcal{M}$ a given convex set of finite measures on $\mathcal{B}$, under general conditions, we have

$$H = \sup_{d \in \Re^n} \{< d, \alpha > + H(d)\}; \tag{2.7}$$

$d = (d_1, d_2, \ldots, d_n)$, $< d, \alpha >= \sum_{i=1}^n d_i \alpha_i$ and

$$H(d) = \inf_{\mu \in \mathcal{M}} \int \left[h(\theta) - \sum d_i g_i(\theta) \right] \mu(d\theta). \tag{2.8}$$

Observe that if $\mathcal{M} = \mathcal{P}$ then by (2.8) and (2.2), (2.7) reduces to

$$H = \sup_{d \in \Re^n} \inf_{\theta \in \Theta} \left[\sum d_i(\alpha_i - g_i(\theta)) + h(\theta) \right]. \tag{2.9}$$

Same type of results hold with inequalities constraints and many of the results above carry over to an infinite number of constraints; references are Kemperman (1972) and Kemperman (1983).

3. *Global robustness and moment problem.* The robustness of a Bayesian quantity of interest $\psi(\pi)$ over a class Γ of plausible priors π is measured by the interval $(\inf_{\pi \in \Gamma} \psi(\pi), \sup_{\pi \in \Gamma} \psi(\pi))$. Key Bayesian quantities, in addition to the marginal and the posterior, are ratio of linear functionals, typically

$$\psi(\pi) = \frac{\int h(\theta) f(\theta) \pi(d\theta)}{\int f(\theta) \pi(d\theta)},$$

f denoting the likelihood. For this type of functional, under mild conditions the well known *linearization algorithm* states that the value $\inf_{\pi \in \Gamma} \psi(\pi)$ is the unique solution λ_0 of the equation in λ

$$\inf_{\pi \in \Gamma} \int [h(\theta) - \lambda] f(\theta) \pi(d\theta) = 0 \tag{3.1}$$

(and symmetrically for the sup), thus converting the minimization of the ratio to a linear minimization problem together with a root-finding operation.

This relevant simplification still contains a minimization on the functions class Γ, a main source of difficulties in solving (3.1).

For the main classes Γ of interest the moment problem results convert the minimization in (3.1) into an optimization in a finite dimension space yielding either an explicit solution for (3.1) or an equation easier to handle.

For Γ *class of mixtures* as in (2.3), by (2.4) for every λ we have

$$\inf_{\pi\in\Gamma}\int[h(\theta)-\lambda]f(\theta)\pi(d\theta)=\inf_{u\in U}\int[h(\theta)-\lambda]f(\theta)K(u,d\theta),$$

and the equation (3.1) becomes

$$\inf_{u\in U}\left[\int h(\theta)f(\theta)K(u,d\theta)-\lambda\int f(\theta)K(u,d\theta)\right]=0. \qquad (3.2)$$

In this case the same argument of the linearization algorithm shows that the solution of (3.2) is

$$\lambda_0=\inf_{u\in U}\frac{\int h(\theta)f(\theta)K(u,d\theta)}{\int f(\theta)K(u,d\theta)}. \qquad (3.3)$$

Observe that in cases of practical interest U is a finite dimensional euclidean space as in the examples in the text. The generality of (3.3) includes, as particular cases, the class of unimodal contaminations on the real line examined in Sivaganesan and Berger (1989) and the class of the multidimensional block unimodal contaminations examined in Liseo, Petrella and Salinetti (1993).

For the density band class of Section 2 let Γ be the *density bounded class*, the class of probability measures π with bounded density ρ

$$a(\theta)\le\rho(\theta)\le b(\theta),\quad\forall\theta\in\Theta,$$

we have $\Gamma=\{\pi\in\Delta:\int\pi(d\theta)=1\}$ where Δ denotes the density band class.

For every λ, by (2.5) and (2.7) we have

$$\inf_{\pi\in\Gamma}\int[h(\theta)-\lambda]f(\theta)\pi(d\theta)$$
$$=\inf_{\pi\in\Delta}\left\{\int[h(\theta)-\lambda]f(\theta)\pi(d\theta):\int\pi(d\theta)=1\right\}$$
$$=\sup_{d\in\Re}\left\{d+\int_{A(\lambda,d)}([h(\theta)-\lambda]f(\theta)-d)a(\theta)\nu(d\theta)\right.$$
$$\left.+\int_{B(\lambda,d)}([h(\theta)-\lambda]f(\theta)-d)b(\theta)\nu(d\theta)\right.$$

where $A(\lambda,d)=\{\theta\in\Theta:[h(\theta)-\lambda]f(\theta)>d\}$ and $B(\lambda,d)=\{\theta\in\Theta:[h(\theta)-\lambda]f(\theta)<d\}$.

The $\inf_{\pi\in\Gamma}\psi(\pi)$ is solution λ_0 of the equation in λ obtained setting equal to 0 the last member of the above relation. Again the optimization in the equation is on the real line. An equivalent equation in λ, more suitable for computations, could also be obtained still based on the notion of admissible solution for the moment problem.

For the *quantile class*

$$\Gamma=\left\{\pi\in\mathcal{P}:\int_{A_i}\pi(d\theta)=\alpha_i,\qquad i=1,2,\ldots,n\right\}$$

with $\mathcal{P}$ class for all the probability measures and $\{A_i,i=1,2,\ldots,n\}$ partition of Θ, the main moment problem, by (2.9), reduces the equation (3.1) to

$$\sup_{d\in\Re^n}\inf_{\theta\in\Theta}\left[\sum_{i=1}^{n}d_i(\alpha_i-I_{A_i}(\theta))+[h(\theta)-\lambda]f(\theta)\right]=0.$$

Again the optimization in the equation is in finite dimension and the particular nature of the constraints makes the solution of the equation reasonably handleable. In addition it is relevant to observe that, as in the case of mixture classes, the argument of the linearization algorithm can be used to obtain the solution λ_0 as optimal value of a ratio of functions in d and θ.

It is immediate to realize that the same type of result holds for the *moment class*.

It is relevant to emphasize the flexibility of the result (2.7). In fact it allows to deal with more complicated classes Γ, for example classes of mixtures with quantile constraints or more generally linear constraints, once one has a reasonable expression for $H(d)$.

4. *Final remarks.* The use of the moment problem in solving robustness problems has the appeal to restate the problem in finite dimension and seems to delineate easier computational directions. The linearization algorithm reserved to ratios of linear functionals allows its use. An alternative use based on suitable "conditioning" can actually be pursued and extended to more general cases. Thus the robstness of the ratio of linear functionals

$$\inf_{\pi\in\Gamma} \frac{\int h f d\pi}{\int f d\pi}, \tag{4.1}$$

under rather mild conditions, can also be approached as constrained minimization problem

$$\inf_{x\in[D^-,D^+]} \inf_{\pi\in\Gamma} \left\{ \frac{1}{x} \int h f d\pi : \int f d\pi = x \right\}$$

where $D^- = \inf_{\pi\in\Gamma} \int f d\pi, D^+ = \sup_{\pi\in\Gamma} \int f d\pi$.

The main moment problem allows the computation of the range $[D^-, D^+]$ of the denominator $D(\pi) = \int f d\pi$ and of the constrained problem

$$\inf_{\pi\in\Gamma} \left\{ \int h f d\pi : \int f d\pi = x \right\}.$$

This approach is pursued in Perone Pacifico, Tardella and Salinetti (1994) for the density bounded class.

The case where the numerator in (4.1) is not linear, of the type

$$N(\pi) = \int h(\theta, \beta(\pi)) f(\theta) \pi(d\theta)$$

can be approached through the moment problem as a double constrained minimization problem on the possible values of the denominator and the possible values of $\beta(\pi)$ for $\pi \in \Gamma$. The most common example arises

from $h(\theta, \beta(\pi)) = (\theta - E(\pi))^2$ where $E(\pi)$ is the posterior mean and then $\psi(\pi)$ is the posterior variance. It can be convenient to convert the non linear quantity into conditional linear quantities

$$\inf_{\pi \in \Gamma} \left[\frac{1}{x} \int h(\theta, y) d\pi : E(\pi) = y, D(\pi) = x \right]$$

to which applying the moment problem tools and then mininizing over (x, y).

Certainly the idea is not new and in particular it is sketched in Berger (1990). The emphasis here is on the fact that the technical potential of the moment problem in the problems of sensitivity to the prior is particularly rich, only partially explored and perhaps it has not been taken the most of it.

SIVA SIVAGANESAN *(University of Cincinnati)*

There has been an explosion of research in the field of robust Bayesian analysis since Professor Berger's last review, Berger (1990). Despite the enormity of the literature in the field, Professor Berger has given us an excellant overview of not only the past developments, but also of possible future developments in the area. This will undoubtedly serve as a very valuable reference for new and current researchers in this area; we ought to be very grateful to him.

In this discussion, I will focus on two areas, namely: local robustness, and multidimensional priors. In the first, a brief review of the literature is given with some comments. The second topic is relatively large and is well covered by Berger; here I focus only on certain classes of priors. For convenience, I will follow the notation and references in the main article.

1. *Local Robustness.* In local robustness study, one seeks to obtain a measure of sensitivity of $\psi(\pi, f)$ to small deviations from the base model π_0 (and/or f_0). This is achieved by using a (suitable) derivative of $\psi(\pi, f)$ with respect to a convenient measure of deviation (from π_0). For instance, consider the ε-contamination class Γ_ε of priors $\pi = (1 - \varepsilon)\pi_0 + \varepsilon q$. Let $D_q\psi$ be the derivative of $\psi(\pi, f)$ w.r.t. ε evaluated at $\varepsilon = 0$ (i.e., at $\pi = \pi_0$). It is reasonable to use $D_q\psi$ as a measure of sensitivity of $\psi(\pi, f)$ to small deviations in the direction of q. When a class Q of possible deviations is of interest, the (maximum) sensitivity

measure $\bar{D}\psi = \sup D_q\psi$ may be used instead. Alternatively, one may want to consider the derivative $D\bar{\psi}$, at $\pi = \pi_0$, of $\bar{\psi} = \sup_{\pi\in\Gamma} \psi(\pi, f)$. In fact, for density bounded and density ratio classes, this second approach is taken in Ruggeri and Wasserman (1991), which seems to be the only workable approach for these classes. There, the authors consider these two classes where the upper and lower bands are given by $L = \frac{1}{k}\pi_0$ and $U = k\pi_0$ for a fixed prior π_0 and a constant $k > 1$, and calculate the derivatives of $\bar{D}$, w.r.t. k, at $k = 1$ For ε-contamination classes, however, both approaches lead to the same answer in most cases of interest, see Sivaganesan (1993c). One of the advantages of this approach is that the robustness measures proposed here, viz. $\bar{D}\psi$ (or $D\bar{\psi}$), are generally easier to compute than the global robustness measure $\bar{\psi}$ (and $\underline{\psi}$). In particular, this approach is just as easily tractable for investigating the robustness of likelihood (f) using non-parametric deviations with typical uncertainties about f, see Sivaganesan (1993c). However, as Berger describes in Section 4.4, with global robustness approach, this becomes a very complicated problem. Such computational ease of the local robustness approach also makes it appealing in other more complex situations, such as hierarchical prior settings and those where multidimensional parameters are considered, e.g., see Sivaganesan (1993c) and Gustafson and Wasserman (1993). Another possible application arises from the simple notion that the larger the value of $\bar{D}\psi$ is, the more sensitive ψ would be to deviations from the base prior (or base likelihood). Examples of such application can be found in Sivaganesan (1993c). There, local robustness measures $\bar{D}\psi$ are used, on a comparative basis, to determine whether ψ is more sensitive to small deviations from the (base) prior π_0, or to small deviations from the (base) likelihood f_0. In another application, robustness with respect to the specification of hierarchical prior is considered with the goal of determining which stage prior causes most sensitivity in ψ. When there is a lack of robustness, such analyses would be useful in determining where to concentrate the re-elicitation efforts, in order to make most gain in robustness.

There lies, however, a difficulty with the local robustness approach. It is not clear how to interpret a single value of, say, $\bar{D}\psi$, or, how small it should be so that one can be satisfied that robustness exists. The answer to such questions, in general, have to be problem specific. But, even in specific situations, it seems more work would be required to gauge the

values of measures such as $\bar{D}\psi$. Note that global robustness measures such as ranges do not share these difficulties, and are easy to interpret. This said, it is perhaps worth pointing out again that in many complex situations where global robustness approach with realistic priors are hard to carry out, local robustness approach can be useful.

The idea of using derivatives goes back to Diaconis and Freedman (1986), where Fréchet and Gateaux derivatives are suggested as measures of sensitivity. These ideas are followed up in Ruggeri and Wasserman (1993), where these derivatives are derived and explored, while in Gustafson and Wasserman (1993), rates of convergence of certain local robustness measures are studied for different classes of priors. In a related paper, Cuevas and Sanz (1988) investigate the differentiability, in the Fréchet sense, of the Bayes operator (or the posterior distribution) with respect to prior. In Srinivasan and Truszczynska (1990, 1993), these derivatives are used to obtain good approximations to global robustness bounds. Other related papers are Salinetti (1994), where stability properties (or, qualitative aspects) of Bayes decision rules are considered; Basu, Jammalamadakka and Liu (1993a), where local robustness is studied in parametric settings using partial and total derivatives; and Basu, Jammalamadakka and Liu (1993b), where certain classical robustness notions such as qualitative robustness and stability are adapted and investigated in the context of posterior distributions and related quantities.

2. *Multidimensional Priors.* In general, classes which have considerable smoothness constraints would be desirable in high dimensional problems. In this sense, density based classes such as density ratio or density band classes, or mixture classes would seem particularly attractive choices. Marginal and Independence classes are attractive from both elicitational and computational viewpoints. From computational viewpoint, however, many of the aforementioned classes seem to be tractable only for low dimensional problems, e.g., see Lavine (1991b), Lavine, Wasserman and Wolpert (1991), Sivaganesan (1994).

One of the few exceptions is the density ratio class where $U = k\pi_0$ and $L = \frac{1}{k}\pi_0$ for some fixed $k > 1$ and a fixed (base) prior π_0. In a very interesting paper, for arbitrary dimension, DasGupta and Studden (1988b) obtained a closed form solution for the set of posterior means in the normal linear model problem. Later, in a breakthrough paper, Wasserman and Kadane (1992a) developed a method of computing the

bounds for this class (and for some other classes) using Monte Carlo simulation. Although this class has thus been shown to be computationally tractable, its adequacy has been somewhat questioned, e.g., see Berger's discussion to Wasserman (1992b). In the following, an example is given that also seems to support such concerns, although possibly from a different viewpoint. This example illustrates a certain phenomenon which is rather counterintuitive, but is commonly associated with this class.

Suppose that $X_1, ..., X_n$ are i.i.d. $N(\theta, \sigma^2)$, where θ is a real parameter, and $\sigma^2 > 0$. Assume that (θ, σ^2) is given the (base) prior $\pi_0(\theta, \sigma^2) = \pi_0^{(1)}(\theta)\pi_0^{(2)}(\sigma^2)$. Consider the two density ratio classes Γ_1 and Γ_2 given by

$$\Gamma_1 = \{\text{generalized } \pi(\theta, \sigma^2) : \frac{1}{k}\pi_0 \leq \pi \leq k\pi_0\},$$

$$\Gamma_2 = \{\text{generalized } \pi(\theta, \sigma^2) = \pi^{(1)}(\theta)\pi_0^{(2)}(\sigma^2)$$
$$: \frac{1}{k}\pi_0^{(1)} \leq \pi^{(1)} \leq k\pi_0^{(1)}\}.$$

Note that, marginally, these two classes represent the same prior information about θ. However, in terms of how θ and σ^2 are related, and in terms of the prior information about σ^2, these two classes represent very different information. Whereas the class Γ_1 allows θ and σ^2 to be dependent and entertains some uncertainty about the (marginal) prior for σ^2, the class Γ_2 does neither. But, it is easy to see that, for $A \subset R$,

$$\inf_{\pi \in \Gamma_1} P(\theta \in A | \boldsymbol{x}, \pi) = \inf_{\pi \in \Gamma_2} P(\theta \in A | \boldsymbol{x}, \pi),$$

i.e., these two classes yield the same lower bound for the posterior probability that $\theta \in A$. In other words, when considering the parameter θ, the structure of the density ratio class seems to ignore both the nature of the uncertainty about the prior information about the (other) parameter σ^2, and the possible interdependency between the two parameters. This seems counterintuitive, since one would anticipate that the different nature of the prior information about σ^2 (as given above) would yield different bounds for the posterior probability concerning θ. This particular phenomenon of the density ratio class, which holds in any multidimensional problems, may actually be ralated to the invariance

property for this class established by Wasserman and Kadane (1992a). Note that for other classes, e.g., the density band class with the same bands as above, this phenomenon does not hold, and that the bounds for the two analogous classes would, in general, be different. Although the density ratio class has definite advantages in terms of compuatation, in light of the phenomenon above, it seems one ought to be cautious about using this class.

REPLY TO THE DISCUSSION

This is perhaps the finest set of discussions that I have seen for a discussion article. Several discussions are essentially quality articles in their own right, and all convey illuminating viewpoints or developments. Particularly gratifying is that the discussions overcome the limitations of the paper, by elaborating or explaining topics that were inadequately treated in the paper. Taken as a whole, the discussions form essentially a complementary review paper, and one that is fascinating reading. My first attempt at preparing a reply to the discussion was to search for areas of contention and elaborate further on the disagreements. With considerable surprise, however, I found that there were almost no substantive areas of disagreement. Many discussants perceived the 'truth' from a different perspective than mine, and there were naturally quite different predictions as to what is likely to be important in the future, but I found the tolerance among robust Bayesians to be refreshingly sincere; we all seem to recognize that, at this period in the development of the field, scope and support must be given to a great variety of approaches. My second attempt to prepare a reply was to outline or provide an index to the discussions. I quickly realized that this too was essentially impossible, because the discussions were astonishingly different; their scope of coverage and diverse viewpoints virtually defied my attempts to impose structure. Thus I was left with only the smaller job of replying to questions specifically raised in the discussions. There were three such questions.

Drs. Gustafson and Wasserman discuss the "annoying problem" of 'dilation,' and ask if I have any advice about what to do when the problem occurs. I might quibble with the phrase "annoying problem," since I view the dilation phenomenon as one of those delightful facts of our subject that make it subtle enough to be interesting. My original reaction to dila-

tion was much like my reaction to the fact that the posterior variance can actually be larger than the prior variance; the subject suddenly became more interesting. I imagine, however, that Gustafson and Wasserman are asking the serious question of whether certain robust Bayesian approaches might be more resistant than others to dilation phenomena; alas, I do not know.

The second question of Gustafson and Wasserman concerns my current views on the potential value of local diagnostics. I was delighted to see discussion of such diagnostics by many of the discussants, because it was certainly an area that was not adequately covered in the paper. My view today is simultaneously more positive and more negative than it was in the discussion of the Diaconis and Freedman (1986) paper. It is more positive in the sense that I have seen a large variety of fascinating local diagnostic tools developed since then; the negative feelings come from a sense that the tools are hard to use because interpretation is often difficult.

I am not a strong believer in the notion that our profession can learn to 'calibrate' tools whose initial interpretation is unclear; after all, we are still miserable calibrators of P-values. That said, I do believe that local diagnostics are likely to end up being an important part of the robust Bayesian toolkit.

The remaining question, asked by Dr. Lavine, is whether I know of any examples in which robust Bayesian analysis has helped to identify unknown quantities for which subjective elicitation is necessary. My answer is 'yes,' if one is willing to be somewhat generous in the definition of robust Bayesian analysis. First of all, for some general situations I think that robust Bayesian analysis has crystallized the understanding of which features of the analysis are most important in terms of assessment. One such example is the recognized importance of the 'spread' of the priors in hypothesis testing and model comparison; robust Bayesian analysis has aided this understanding, partly by clearly demonstrating the lack of robustness of the answers to such quantities, and partly by showing that alternatives such as P-values (which were even sometimes recommended by Bayesians as a reasonable 'solution' to the problem) are simply not tenable.

On a more practical level, I think we use robust Bayesian reasoning all the time in analyzing data, even if we do not yet routinely use formal

robust Bayesian methods. During an involved data analysis I initially play around with the model and various crude or automatic priors, with the primary goal being to see which elements of the problem remain essentially fixed and which vary considerably. It is upon these latter elements that I then focus attention, with the stable quantities typically just being assigned a noninformative prior. For the unstable quantities, I typically first attempt to impose some reasonable structure, perhaps a hierarchical structure or constraints on, say, ordering or positivity. If such structural assumptions are not warranted or do not seem to be sufficient to provide robustness, then I attempt subjective elicitations, in the usual sense, from the subject-matter client. Of course, the previous steps of the analysis, involving model building, playing around with priors, and imposing structure, are also done with the client, but they will often require far less of the client's time than will formal prior elicitation.

Let me finish by thanking all the discussants for their highly illuminating contributions. Also, I would like to express the thanks of all of us to the Spanish Statistical Society and the Editor and editorial board of TEST for providing our community with the opportunity to review this rapidly growing and exciting field.

ADDITIONAL REFERENCES IN THE DISCUSSION

Berger, J. O. (1985b). Discussion of 'Quantifying prior opinion' by Diaconis and Ylvisaker. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 133-156.

Cano, J. A. and Moreno, E. (1993). Sampling models: a robust Bayesian analysis. *Tech. Rep.*, Universidad de Murcia.

Cifarelli, D. M. and Regazzini, E. (1987). On a general definition of concentration function. *Sankhyā B* **49** 307-319.

Cook, R.D. (1986). Assessment of local influence. *J. Roy. Statist. Soc. B* **48**, 133–169, (with discussion)..

Dalal, S. R. and Hall, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *J. Roy. Statist. Soc. B* **45**, 278-286.

De la Horra, J. and Fernández, C. (1994). Sensitivity to prior independence via Farlie-Gumbel-Morgenstern model. *Tech. Rep.* **94–17**, Tilburg University.

Diaconis, P. and Ylvisaker, D. (1985). Quantifying prior opinion. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 163–175.

Efron, B. (1986). Why Isn't Everyone a Bayesian? *Ann. Statist.* **40**, 1–11.

Feller, W. (1973). *An Introduction to Probability Theory and Applications* **2**, New York: Wiley.

Garthwaite, P. H. (1992). Preposterior expected loss as a scoring rule for prior distributions. *Comm. Statist. Th. Meth.* **21**, 3601–3619.

Garthwaite, P. H. and Dickey, J. M. (1992). Elicitation of prior distributions for variable-selection problems in regression. *Ann. Statist.* **20**, 1697–1719.

Grieve, A. P. (1985). A Bayesian Analysis of the Two-Period Crossover Design for Clinical Trials. *Biometrika* **41**, 979–990.

Gustafson, P. (1993a). Local sensitivity in Bayesian statistics. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh.

Gustafson, P. (1993b). The local sensitivity of posterior expectations. *Tech. Rep.*, Carnegie Mellon University.

Johnson, N. L. and Kotz, S. (1975). On some generalized Farlie-Gumbel-Morgenstern distributions. *Comm. Statist.* **A4**, 415–427.

Johnson, N. L. and Kotz, S. (1977). On some generalized Farlie-Gumbel-Morgenstern distributions II: regression, correlation and further generalizations. *Comm. Statist.* **A6**, 485–496.

Keeney, R. and Raiffa, H. (1993). *Decision Making with Multiple Objectives*, Wiley.

Kemperman, J. H. B. (1972). On a class of moment problem. *Proc. Berkeley Symposium on Math. Statist. and Prob.* **2**, 101–126.

Kemperman, J. H. B. (1983). On the role of duality on the theory of moments. *Semi-infinite programming and applications* (Fiacco, A. V. and Kortaneck, K. O., eds.). Lecture Notes in Economics and mathematical Systems **215**, New York: Springer-Verlag.

Kemperman, J. H. B. (1987). Geometry of the moment problem. *Proceedings of Symposia in Applied mathematics* **37**, 16–53

Krause, A. and Polasek, W. (1992). Approaches to Tobit Models via Gibbs sampling, in: COMPSTAT 1992, Physica Verlalg, 559–564,

Le, H. and O'Hagan, A. (1993). A class of bivariate heavy-tailed distributions. *Tech. Rep.* **93-20**, Nottingham University Statistics Group.

Leamer, E. E. (1978). Specification Searches. New York: Wiley.

Leland, J. (1992). An approximate Expected Utility Theory, *Tech. Rep.*, Carnegie Mellon University.

Levy, H. (1992). Stochastic Dominance and Expected Utility Analysis: A Review, *Mgt. Science* **38**, 555–593.

Moreno, E., Martínez, C. and Cano, J. A. (1993). Elicitation of contamination classes of prior distributions. *Tech. Rep.*, Universidad de Granada.

Nair, V.J., and Wang, P.C.C. (1989). Maximum likelihood estimation under a successive sampling discovery model. *Technometrics* **31**, 423–436.

O'Hagan, A. and Le, H. (1993). Conflicting information and a class of bivariate heavy-tailed distributions. *Aspects of Uncertainty: A Tribute D. V. Lindley* (Smith, A. F. M. and Freeman, P. R. eds.). Wiley: Chichester.

Polasek, W. (1992). Joint sensitivity analysis for covariance matrices in Bayesian linear regression. *Tech. Rep.* **9205**, University of Basel.

Polasek, W. (1993). Bayesian Generalized Errors in Variables (GEIV) models for censored regression. *Tech. Rep.*, University of Basel.

Proll, L., Ríos-Insúa, D. and Salhi, A. (1993). Mathematical Programming and the sensitivity of multicriteria decisions, *Annals of Operations Research* **43**, 109–122.

Renyi, A. (1962/1970). *Wahrscheinlichkeitsrechnung*. Berlin: Deutscher Verlag der Wissenschaften. English translation in 1970 as *Probability Theory*. San Francisco, CA: Holden-Day.

Ríos, S., Ríos-Insúa, S., Ríos-Insúa, D. and Pachón, J. (1994). Experiments in robust decison making (S. Ríos, ed.) *Decision Making and Decision Analysis: Trends and Challenges*, Dorfrecht: Kluwer, 233–242.

Ríos-Insúa, D. (1994). Ambiguity, imprecision and sensitivity in Decison Theory, (Puri and Vilaplana, eds.) *New Progress in Probability and Statistics*, STP. (to appear).

Robert, C. P. (1992). *L'Analyse Statistique Bayesianne*. Paris: Economica.

Robert, C. P. (1993). Intrinsic losses. Rapport techn., URA CNRS 1378, Univ. de Rouen.

Seidenfeld, T. and Wasserman, L. (1993). Dilation for sets of probabilities. *Ann. Statist.* **21**, 1134–1154.

Stein, C. (1985). On the coverage probability of confidence sets based on a prior distribution. In *Sequential Methods in Statistics*, Banach Center Publication **16**, 485–514. Warsaw: PWN Publishers.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604–608.

West, M. (1993). Prediction for finite populations under biased sampling. *ISDS Discussion Paper* **93**, Duke University.

Young, N. (1988). Hilbert Space, Cambridge: University Press.