APPLICATIONS AND LIMITATIONS OF ROBUST BAYESIAN BOUNDS AND TYPE II MLE

by

M. J. Bayarri
University of Valencia

and

James O. Berger
Purdue University

Technical Report #93–11C

Department of Statistics
Purdue University

January 1993

◇

Applications and Limitations of Robust Bayesian Bounds and Type II MLE*

M. J. Bayarri[1] and James O. Berger[2]

[1] University of Valencia
[2] Purdue University

Abstract. Three applications of robust Bayesian analysis and three examples of its limitations are given. The applications that are reviewed are the development of an automatic Ockham's Razor, outlier detection, and analysis of weighted distributions. Limitations of robust Bayesian bounds are highlighted through examples that include analysis of a paranormal experiment and a hierarchical model. This last example shows a disturbing difference between actual hierarchical Bayesian analysis and robust Bayesian bounds, a difference which also arises if, instead, a Type II MLE or empirical Bayes analysis is performed.

1 Introduction

1.1 Basic Elements

There has recently been considerable interest in the development of the robust Bayesian approach to statistics. Berger (1990) and Wasserman (1992) present reviews. The basic idea is to replace the common single model and/or single prior distribution in Bayesian analysis by wide (often nonparametric) classes of models and/or priors.

If $f(x|\theta)$ stands for the density of the data X given the unknown parameter θ, and $\pi(\theta)$ is the prior density of θ, then (under mild conditions) the posterior density of θ is

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

where

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

is the marginal density of X. Of interest is typically some functional $\psi(f,\pi)$ (e.g., the posterior mean or a Bayes factor).

Suppose now that it is only known that

$$f \in \mathcal{F} \qquad \text{and/or} \qquad \pi \in \Gamma,$$

where $\mathcal{F}$ and Γ are classes of densities. The most common robust Bayesian approach is to then compute

$$\underline{\psi} = \inf_{f\in\mathcal{F},\pi\in\Gamma} \psi(f,\pi), \quad \overline{\psi} = \sup_{f\in\mathcal{F},\pi\in\Gamma} \psi(f,\pi),$$

* Research supported by the National Science Foundation, Grant DMS-8923071.

and report $(\underline{\psi}, \overline{\psi})$ as the range of possible answers. It may well be that knowledge of this range suffices for answering the problem, and hence more detailed elicitation of f and/or π can be avoided.

As a notational convenience, we will use the symbol $N(x|\mu, \sigma^2)$ to stand for a normal density in the indicated variable (here x), with mean μ and variance σ^2.

1.2 Hypothesis Testing and Type II MLE

Several of the examples in the paper will involve hypothesis testing, and these will connect with the common empirical Bayesian technique of *Type II Maximum Likelihood.* Testing examples will involve expressing model and/or prior uncertainty through a parameter τ, and will reduce to analysis of a Bayes factor (of H_0 to H_1) given by

$$B = \frac{\int f_0(x|\theta)\pi_0(\theta)d\theta}{\int f(x|\theta,\tau)\pi(\theta|\tau)g(\tau)d\theta d\tau}.$$

Here, π_0 and $\pi(\theta|\tau)$ will be specified, and g will be assumed to belong to a class $\mathcal{G}$ of possible distributions. Robust Bayesian analysis in this scenario reduces to determination of

$$\underline{B} = \inf_{g\in\mathcal{G}} B = \frac{\int f_0(x|\theta)\pi_0(\theta)d\theta}{\sup_{g\in\mathcal{G}} \int f(x|\theta,\tau)\pi(\theta|\tau)g(\tau)d\theta d\tau},$$

and the associated $\overline{B} = \sup_g B$. (In our examples, $\overline{B}$ will typically be infinite, and hence of little interest.)

It is often convenient to choose

$$\mathcal{G} = \mathcal{G}_A = \{\text{all distributions}\},$$

since then

$$\underline{B} = \inf_{g\in\mathcal{G}_A} B = \frac{\int f_0(x|\theta)\pi_0(\theta)d\theta}{m^*(x|\hat{\tau})},$$

where

$$m^*(x|\tau) \equiv \int f(x|\theta,\tau)\pi(\theta|\tau)d\theta,$$

and $\hat{\tau}$ is that value which maximizes $m^*(x|\tau)$. Such a $\hat{\tau}$ is called a Type II maximum likelihood estimate, and so $\underline{B}$ computed in this way could also be called the "Type II MLE Bayes factor." Lower bounds on Bayes factors are thus often equivalent to Bayes factors computed under the Type II MLE approach.

1.3 Preview

This paper has two goals. The first is to review the types of applications of robust Bayesian analysis that are possible. The specific applications considered (in Section 2) are the development of a Bayesian Ockham's razor, detection of outliers, and analysis of selection models (i.e., models where $f(x|\theta)$ depends on the mechanism by which the data is selected and/or reported).

Section 3 explores the limitations of the robust Bayesian/Type II MLE approach. The main limitation is simply that the bounds may be too extreme, resulting in an indeterminate answer. Several examples are given of this, including an example where the bounds are shown to behave very differently than the corresponding posterior quantity for any fixed prior distribution. The example involves a hierarchical model, and also exposes a serious limitation of the Type II MLE method.

2 Applications of Robust Bayesian Bounds

2.1 Model Selection and Ockham's Razor

One of the most interesting applications of robust Bayesian bounds has been the development of an "automatic" Ockham's razor (Berger and Jefferys, 1992, and Jefferys and Berger, 1992). The scenario is that of comparing two models for data $X \sim f(x|\theta)$. The "simpler" model is $M_0: \theta = \theta_0$, while the "complex" model is $M_1: \theta$ arbitrary. Also, under M_1, beliefs about θ are symmetric about 0 (which is typically the value corresponding to existing theory), with larger values of $|\theta|$ being no more credible than smaller values.

Example 1. One of the great scientific projects in the late 1800s and early 1900s was to explain the anomalous behavior of the motion of the perihilion of Mercury (the point at which Mercury is closest to the sun in its orbit). For our purposes, it is sufficient to suppose the data was $X \sim \mathcal{N}(\theta, (2)^2)$ (in units of seconds of arc per century), where Newtonian physics with known astronomical objects predicted $\theta = 0$. The actual data was $x = 41.6$, clearly calling for an alternative explanation.

A number of alternative theories were proposed. Perhaps the two best theories were general relativity (by Einstein) which predicted $\theta = 42.9$ (call this M_0, i.e., $\theta_0 = 42.9$), and a theory of Newcomb that gravity follows an inverse $(2+\varepsilon)$ law (instead of inverse square). Converting ε to θ (a deterministic transformation), and noting that Newcomb's theory equally allows negative and positive values of ε (corresponding to negative and positive values of θ) and that existing scientific belief would not have caused more prior weight to be given to larger values of $|\varepsilon|$ (equivalently, $|\theta|$) than smaller values, we see that Newcomb's model can be described statistically by the model M_1 defined at the beginning of the section.

The Bayes factor of M_0 to M_1 in this situation is given by

$$B(g) = \frac{f(x|\theta_0)}{\int f(x|\theta)g(\theta)d\theta},$$

where $g(\theta)$ represents the prior density of θ under model M_1. Because of the prior beliefs about θ, it is reasonable to restrict g to be in the class

$$g \in \mathcal{G}_{SU} = \{g(\theta) = h(|\theta|), \text{ where } h \text{ is nonincreasing}\}.$$

Then, it is easy to show that a lower bound on $B(g)$ is

$$\underline{B} = \inf_{g \in \mathcal{G}_{SU}} B(g) = \frac{f(x|\theta_0)}{\sup_{r>0} \frac{1}{2r} \int_{-r}^{r} f(x|\theta)d\theta}. \tag{2.1}$$

For the special case where $X \sim \mathcal{N}(\theta, \sigma^2)$ (σ^2 known), $\underline{B}$ can only be computed as the iterative solution to an equation (see Berger and Sellke, 1987), but can be approximated by

$$\hat{\underline{B}} = \sqrt{\frac{2}{\pi}} \exp\{-\frac{1}{2}\, t_0^2\} \left[t_1 + \sqrt{2\log(t_1 + 1.2)}\right], \tag{2.2}$$

where $t_0 = |x - \theta_0|/\sigma$ and $t_1 = |x|/\sigma$. (This approximation is within $o(1)$ of $\underline{B}$ as $t_1 \to \infty$, and is accurate within 1% if $t_1 > 1.4$).

In the perihilion example,

$$t_0 = |41.6 - 42.9|/2.0 = 0.65 \quad \text{and} \quad t_1 = |41.6|/2.0 = 20.8,$$

so

$$\hat{\underline{B}} = \sqrt{\frac{2}{\pi}} \exp\{-\frac{1}{2}\,(0.65)^2\} \left[20.8 + \sqrt{2\log(20.8 + 1.2)}\right] = 15.04.$$

Thus the data supports the simpler law (general relativity) over the more complex law by *at least* a factor of 15 to 1. That this happens in spite of the fact that the data "fits" M_1 even better than it fits M_0, shows that the Bayes factor clearly acts as an Ockham's razor.

Study of (2.2) shows that the Bayesian Ockham's razor operates very sensibly. If the simpler model does not adequately fit the data, i.e. if t_0 is large, then $\hat{\underline{B}}$ will be small, indicating substantial odds against the simpler model. But if the simpler model adequately fits the data, the more complex model becomes penalized by a factor of (roughly) t_1, which can be thought of as the (scaled) amount by which the free parameter in the complex model has to be adjusted (using the current theory as a baseline) to fit the data.

2.2 Outlier Detection

Bayesian detection of outliers has been extensively studied; references can be found in Bayarri and Berger (1992). The most common approach uses a model for outliers, which is usually taken to be a generalization of the original model, involving an extra parameter. The original model is then typically a particular case of the contaminating model corresponding to some specific value of the extra parameter, and testing for outliers is reduced to testing for this specific value of the extra parameter. Since little is usually known about the contaminating distribution, robust Bayesian methods are a very natural way to approach outlier detection.

Here, we limit ourselves to discussion of the simplest case, that in which the goal is to detect whether a specific observation, x_0, is an outlier, assuming that all the rest are not. We assume that the data are observations of $n + 1$ independent random variables $X_0, X_1, \ldots, X_n$, originally modelled as $X_i \sim f_0(x_i|\theta)$, but we recognize that there is a (small) probability, ε, that X_0 is an outlier generated from the contaminating density $f(x|\theta, \tau)$; hence,

$$X_0 \sim (1 - \varepsilon) f_0(x_0|\theta) + \varepsilon f(x_0|\theta, \tau).$$

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ denote the non-outlying observations and $\ell(\theta) = \prod_{i=1}^{n} f_0(x_i|\theta)$ denote the likelihood function for θ based solely on $\boldsymbol{x}$.

Testing H_0: x_0 *is not an outlier*, versus H_1: x_0 *is an outlier* is, with the above formulation, equivalent to testing $H_0: X_0 \sim f_0(x_0|\theta)$ versus $H_1: X_0 \sim f(x_0|\theta,\tau)$, so that the expressions in Sect. 1.2 apply. Here, the Bayes factor in favor of x_0 being non-outlier can be written

$$B = \frac{\int \ell(\theta) f_0(x_0|\theta)\pi_0(\theta)d\theta}{\int\int \ell(\theta) f(x_0|\theta,\tau)\pi(\theta|\tau)g(\tau)d\theta d\tau} = \frac{m_0(x_o|\boldsymbol{x})}{m(x_0|\boldsymbol{x})}, \tag{2.3}$$

where m_0 and m are the (posterior) predictive densities at x_0 under the models f_0 and f, respectively, i.e.,

$$m_0(x_0|\boldsymbol{x}) = \int f_0(x_0|\theta)\pi_0(\theta|\boldsymbol{x})d\theta,$$

$$m(x_0|\boldsymbol{x}) = \int\int f(x_0|\theta,\tau)\pi(\theta|\boldsymbol{x},\tau)g(\tau|\boldsymbol{x})d\theta d\tau.$$

Example 2 (Scale Contamination). A widely used model for outliers is that in which the base model, f_0, is the $\mathcal{N}(\theta,\sigma^2)$ distribution (initially assume that σ^2 is known), and the distribution of the possible outlier X_0 is

$$(1-\varepsilon)N(x_0|\theta,\sigma^2) + \varepsilon N(x_0|\theta,\tau\sigma^2),$$

with $\tau \geq 1$ unknown. We assume that θ and τ are independent a priori, so that $\pi(\theta,\tau) = \pi(\theta)g(\tau)$, and $\pi(\theta) = \pi_0(\theta) = \pi(\theta|\tau)$ is taken to be a $N(\theta|m_0,\sigma_0^2)$ density. The Bayes factor (2.3) can then be shown to be

$$B = \frac{N(x_0|m_1,\sigma_1^2+\sigma^2)}{\int N(x_0|m_1,\sigma_1^2+\tau\sigma^2)g(\tau)d\tau}, \tag{2.4}$$

where $m_1 = \lambda\overline{x} + (1-\lambda)m_0$, $\sigma_1^2 = \lambda\sigma^2/n$, and $\lambda = n\sigma_0^2/(\sigma^2+n\sigma_0^2)$. (Using the non-informative prior $\pi(\theta) = 1$ is equivalent to taking $\lambda = 1$ above.)

The infimum, $\underline{B}$, of this Bayes factor over the class $\mathcal{G}_A = \{\text{all distributions}\}$ of prior distributions for τ is then given by

$$\underline{B} = \frac{N(x_0|m_1,\sigma_1^2+\sigma^2)}{\sup_\tau N(x_0|m_1,\sigma_1^2+\tau\sigma^2)} = \sqrt{e}\, z\exp\{-z^2/2\}, \text{ for } z > 1, \tag{2.5}$$

and $\underline{B} = 1$ for $z \leq 1$, where $z = |x_0 - m_1|/s_1$, and m_1 and $s_1 = \sqrt{\sigma_1^2+\sigma^2}$ are the mean and standard deviation, respectively, of the predictive distribution $m_0(x_0|\boldsymbol{x})$ of x_0 under the non-outlier model.

When σ^2 is unknown, and the usual Normal-Gamma prior for (θ,σ^2) is taken, that is, $\theta|\delta \sim \mathcal{N}(m_0,(h_0\delta)^{-1})$, $\delta \sim Ga(a_0,b_0)$, where $\delta = 1/\sigma^2$, the infimum $\underline{B}$ of the Bayes factor over the same class $\mathcal{G}_A$ of priors for τ can be computed to be (see Bayarri and Berger, 1992)

$$\underline{B} = z\left(\frac{\alpha+1}{\alpha+z^2}\right)^{(\alpha+1)/2} \qquad \text{for } z > 1, \tag{2.6}$$

and $\underline{B} = 1$ for $z \leq 1$, where $\alpha = 2(a_1-1)$, $a_1 = a_0+(n-1)/2$, and $z = |x_0-m_1|/s_1$ where, here, $m_1 = \lambda\overline{x}+(1-\lambda)m_0$, $\lambda = n/(h_0+n)$, $s_1^2 = [(1+h_1)/h_1][b_1/(a_1-1)]$,

$h_1 = h_0 + n$, and $b_1 = (2b_0 + \sum_{i=1}^{n}(x_i - \overline{x})^2 + \lambda h_o(\overline{x} - m_0)^2)/2$. For the non-informative prior $\pi(\theta, \sigma^2) = 1/\sigma^2$, these become $\alpha = n - 3$, $a_1 = (n-1)/2$, $m_1 = \overline{x}$, $h_1 = n$, and $b_1 = \sum_{i=1}^{n}(x_i - \overline{x})^2/2$.

In Table 1, we give $\underline{B}$ for certain values of z, both in the case of known σ^2 and unknown σ^2 (for $\alpha = 20$). Note that z does not numerically represent the same quantity in both cases; in each case it is, however, the standardized distance between x_0 and its predicted value m_1 under the predictive distribution of X_0 derived under the non-outlier hypothesis. In Fig. 1, the lower bound $\underline{B}$ for the case of unknown σ^2 is graphed, as a function of z, for $\alpha = 10, 20, 40, 100$. ($\underline{B}$ for known σ^2 is indistinguishable from that corresponding to $\alpha = 100$.) Both Table 1 and Fig. 1 were derived using the non-informative priors.

Table 1. Lower bounds on the Bayes factor of non-contamination to scale contamination.

z	1.5	2	2.5	3	3.5	4	4.5
σ^2 known	.8029	.4463	.1811	.0549	.0126	.0022	.0003
σ^2 unknown	.8174	.4922	.2401	.1012	.0387	.0139	.0049

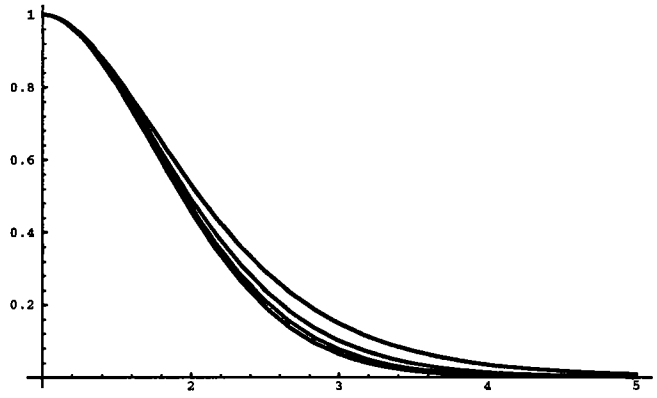

Fig. 1. $\underline{B}$ for σ^2 unknown as a function of z for $\alpha = 10, 20, 40, 100$.

To interpret these results, recall that $\underline{B}$ is the lower bound on the Bayes factor for x_0 being a nonoutlier to x_0 being an outlier. Hence, when $z = 2$, the evidence that x_0 is an outlier is no stronger than about 1 to 2. Even when $z = 3$, so that x_0 is three predictive standard deviations from its predicted location, the evidence for x_0 being an outlier is no stronger than 1 in 20 (known σ^2) or 1 in 10 (unknown σ^2,

$\alpha = 20$). This suggests that outlier rejection standards need to be set higher than is commonly perceived. Unless z exceeds, say, 3 or 3.5 one should not consider the evidence that x_0 is an outlier to even possibly be strong.

2.3 Weighted Distributions

Assume that the random variable $X \in \mathbf{R}^1$ is distributed over some population of interest according to $f(x|\theta)$, $\theta \in (r, s)$, a (possibly infinite) interval in $\mathbf{R}^1$, 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

$$f_w(x|\theta) = \frac{w(x)f(x|\theta)}{\nu_w(\theta)}, \tag{2.7}$$

where $\nu_w(\theta) = E_\theta[w(X)]$. There is, actually, no reason to require $w(x)$ to be a probability; all we require is that w be nonnegative and that $E_\theta[w(X)] < \infty$ for all θ. Then w can be interpreted as a weight function that distorts (multiplies) the probability or density $f(x|\theta)$ that observation x gets selected. 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 (2.7), so that the likelihood function for θ is

$$L_w(\theta) \propto l(\theta)[\nu_w(\theta)]^{-n}, \tag{2.8}$$

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

$$\pi(\theta|x, w) = \frac{l(\theta)[\nu_w(\theta)]^{-n}\pi(\theta)}{\int l(\theta)[\nu_w(\theta)]^{-n}\pi(\theta)d(\theta)}, \tag{2.9}$$

assuming π is such that the denominator is finite. Expression (2.9) 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

$$\mathcal{W} = \{\text{nondecreasing } w\text{: } w_1(x) \leq w(x) \leq w_2(x)\}, \tag{2.10}$$

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

$$\psi(w) = \int \xi(\theta)\pi(\theta|x, w)d\theta \tag{2.11}$$

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:

$$w(x) = \begin{cases} w_1(x) & \text{if } r < x \leq a \\ w_2(x) & \text{if } a < x < s \end{cases}, \tag{2.12}$$

$$w(x) = \begin{cases} w_2(x) & \text{if } r \leq x < h_2(c) \\ c & \text{if } h_2(c) < x < h_1(c) \\ w_1(x) & \text{if } h_1(c) < x < s \end{cases}, \tag{2.13}$$

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

Example 3. 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 (2.10), 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 (2.11) over w of the form (2.12) and (2.13), 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.

3 Limitations of Robust Bayesian Bounds

Robust Bayesian bounds are, of course, just bounds, and their usefulness depends on how close they are to the "real" Bayesian answer (that which would hypothetically arise from infinite reflection on subjective elements of the problem). If the class of priors (or models) is too big, then the bounds can be expected to be poor. We illustrate this here with three examples. The examples also highlight the strong role of the data in determining whether or not a class is "too large." Two of the examples also indicate potential inadequacies of the Type II MLE approach.

3.1 Outlier Rejection with Location Contamination

A commonly used alternative to the scale-contamination model for outliers that was considered in Sect. 2.2 is the location contamination model. As in Sect. 2.2, assume that $x_1, x_2, \ldots, x_n$ are non-outlying observations from a $\mathcal{N}(\theta, \sigma^2)$ distribution, but that x_0 is generated by the mixture

$$(1 - \varepsilon)N(x_o|\theta, \sigma^2) + \varepsilon N(x_0|\theta + \tau, \sigma^2).$$

We limit ourselves, here, to the case in which σ^2 is known. (Lower bounds when σ^2 is unknown are derived in Bayarri and Berger, 1992.) The infimum of the Bayes

factor (2.3) over the class $\mathcal{G}_A = \{\text{all priors}\}$ for the location parameter τ is given by

$$\underline{B}_A = \frac{N(x_0|m_1, s_1^2)}{\sup_\tau N(x_0|m_1+\tau, s_1^2)} = \exp\{-z^2/2\},$$

where z, m_1 and s_1^2 are defined following (2.4) and (2.5). $\underline{B}_A$ differs from $\underline{B}$ in (2.5) by a factor of $\sqrt{e}\ z$. For z in the "interesting" range 2 to 4, this is a factor of from 3.3 to 6.6, a dramatic difference.

When a lower bound seems to be too small to be useful, one should investigate whether the class, $\mathcal{G}$, of priors that is used is too large, in the sense of containing unreasonable prior distributions that could be removed from consideration. In the case of $\mathcal{G}_A$, which was used to derive $\underline{B}_A$, the minimizing prior is a point mass (at $\tau = x_0 - m_1$). Point mass priors are typically quite unreasonable as reflections of prior belief. A much smaller class of "reasonable" priors is

$$\mathcal{G}_{SU} = \{\text{densities } g(\tau) = h(|\tau|), \text{ where } h \text{ is nonincreasing}\}.$$

This class represents symmetry in prior beliefs about the contamination, and the belief that larger absolute contaminations should receive no more prior weight than smaller absolute contaminations.

In Bayarri and Berger (1992), lower bounds on B over $\mathcal{G}_{SU}$ were obtained. For the σ^2 known case, the lower bound is

$$\underline{B}_{SU} = \frac{2\exp\{-z^2/2\}}{\exp\{-(z+\gamma)^2/2\} - \exp\{-(z-\gamma)^2/2\}}, \text{ for } z > 1,$$

and $\underline{B}_{SU} = 1$ for $z \leq 1$, where z is as in $\underline{B}_A$ above, and γ is the unique solution of

$$\gamma[\phi(z+\gamma) + \phi(z-\gamma)] = \Phi(z+\gamma) - \Phi(z-\gamma),$$

where ϕ, Φ represent, as usual, the standard normal p.d.f. and c.d.f., respectively.

Table 2 gives $\underline{B}_A$ and $\underline{B}_{SU}$ for certain values of z, all obtained with the non-informative prior $\pi(\theta) = 1$.

Table 2. Lower bounds on the Bayes factor of non-contamination to location contamination.

z	1.5	2	2.5	3	3.5	4	4.5
$\underline{B}_A$	.3247	.1353	.0439	.0111	.0022	.0003	.0000
$\underline{B}_{SU}$	.7493	.3835	.1458	.0420	.0093	.0016	.0002

Clearly $\underline{B}_{SU}$ is much larger than $\underline{B}_A$. It is also very similar to the bound $\underline{B}$ in (2.5) (see, also, the first row of Table 1) that was obtained for the scale-contamination model. This is evidence that $\mathcal{G}_A$ is simply too large to be useful when considering location contaminations.

It is interesting that (2.5), which seems reasonable, was itself obtained using $\mathcal{G}_A$ (but for the scale-contamination model). In Bayarri and Berger (1992) it is shown that going from $\mathcal{G}_A$ to a smoother class has little effect on $\underline{B}$ in the scale-contamination model. Hence $\mathcal{G}_A$ does not appear to be too large when scale-contaminations are being considered. Indeed, we recommended (2.5) (or (2.6)) for actual use precisely because it seems not to be "too small," while at the same time being easy to compute (because of the simplicity in using $\mathcal{G}_A$).

3.2 A Paranormal Example

Jefferys (1990) analyzes an experiment of Jahn, Dunne, and Nelson (1987) in paranormal phenomena. The statistical question is to test $H_0: \theta = 0.5$ versus $H_1: \theta \neq 0.5$, on the basis of $x = 52,263,471$ successes out of $n = 104,490,000$ Bernoulli (θ) trials. Here $\theta = 0.5$ corresponds to "no paranormal phenomenon present." The interest in this particular example is that the (one-tailed) P-value is less than 0.00015 which, from a classical perspective, would indicate extremely significant evidence against H_0.

A robust Bayesian analysis of this example might reasonably consider, as a class of prior densities on $H_1: \theta \neq 0.5$,

$$\mathcal{G}_{SU} = \{\text{densities } g(\theta) = h(|\theta - 0.5|), \text{ where } h \text{ is nonincreasing}\}. \qquad (3.1)$$

Symmetry is a typically-made assumption in paranormal research (to avoid the appearance of prejudging the conclusion), and it is reasonable apriori to give no more weight to large differences from $\theta = 0.5$ than to small differences.

As in (2.1), it follows directly that $\underline{B}$, the minimum Bayes factor over $\mathcal{G}_{SU}$, is attained at a Uniform $(0.5 - r,\ 0.5 + r)$ distribution. It can be computed that $\underline{B} = 0.0064$, attained at $r \cong 0.00025$. Jefferys also shows that, for even moderately large r, say $r \geq 0.55$, it will be the case that $B \geq 1$ which would imply that the evidence actually *favors* H_0 for such r.

Note, first, that even the lower bound $\underline{B} = 0.0064$ is much larger than the P-value (even the two-tailed P-value is just 0.0003). Hence $\underline{B}$ is not nearly as misleading as the P-value. However, a strong case can be made that $\underline{B}$ is itself misleading, with "natural" values of r (or natural priors) giving much larger Bayes factors.

The problem here is, again, that $\underline{B}$ seeks out an unusual prior. It would be unusual for someone to specify, apriori, that their prior beliefs about θ are concentrated fairly uniformly over the interval (0.49975, 0.50025), and if a much larger (or, for that matter, much smaller) interval were specified, the Bayes factor would be large enough to *favor* H_0.

Part of the interest here is that $\mathcal{G}_{SU}$ in (3.1) is usually considered to be a "nice", reasonably restricted class of priors (compared to, say, $\mathcal{G}_A = \{\text{all priors}\}$). But even a nice class can be unreasonably large in the face of a large amount of data.

A final point of interest is that the Uniform (0.49975, 0.50025) prior is the Type II MLE prior in $\mathcal{G}_{SU}$, i.e., is the prior in $\mathcal{G}_{SU}$ most supported by the data. However, the "likelihood" in, say, r, namely

$$m(x|r) = \frac{1}{2r} \int_{0.5-r}^{0.5+r} f(x|\theta) d\theta,$$

is sharply peaked about $r = .00025$, but virtually all the "mass" of the likelihood is removed from this peak. This is the type of likelihood for which use of the MLE is highly questionable.

3.3 Hierarchical Bayesian Analysis

An interesting discrepancy between robust Bayesian and actual Bayesian analysis arises in hierarchical models. To take the simplest case, suppose $X \sim \mathcal{N}_p(\theta, I)$ and $\theta \sim \mathcal{N}_p(0, \tau^2 I)$, and that we wish to test $H_0: \theta = 0$ versus $H_1: \theta \neq 0$. Interest will focus on how uncertainty about τ^2 is handled. In Sect. 3.3.1 the robust Bayesian approach is illustrated, while Sect. 3.3.2 presents the strict hierarchical Bayesian approach. The conflict between the two is discussed in Sect. 3.3.3.

3.3.1 Robust Bayesian Approach For illustrative purposes, it suffices to consider simply the class

$$\mathcal{G}_N = \{\mathcal{N}_p(0, \tau^2 I) \text{ distributions}, \tau^2 > 0\},$$

as prior distributions for θ. This is studied in Edwards, Lindman, and Savage (1963), where it is shown that, for observed x, the lower bound on the Bayes factor, over $\mathcal{G}_N$, is

$$\begin{aligned} \underline{B} &= \inf_{\tau^2} \frac{N_p(x|0, I)}{\int N_p(x|\theta, I) N_p(\theta|0, \tau^2 I) d\theta} \\ &= \frac{N_p(x|0, I)}{N_p(x|0, (1+\hat{\tau}^2) I)} \\ &= (1+\hat{\tau}^2)^{\frac{p}{2}} \exp\{-\frac{p}{2}\hat{\tau}^2\}, \end{aligned}$$

where $\hat{\tau}^2 = \max\{0, \frac{1}{p}|x|^2 - 1\}$ is the Type II MLE for τ^2.

Interest often focuses on values of x for which the classical P-value for testing $H_0: \theta = 0$ is approximately some specified α. Since $|X|^2$ is chi-squared with p degrees of freedom under H_0, it is the case that the P-value is α when (as $p \to \infty$)

$$|x|^2 = p + z_\alpha \sqrt{2p} + O(1), \tag{3.2}$$

where z_α is the $(1-\alpha)$ quantile of the standard normal distribution.

Lemma 1. *For $|x|^2$ as in (3.2) and $\alpha > \frac{1}{2}$,*

$$\lim_{p \to \infty} \underline{B} = \exp\{-\frac{1}{2} z_\alpha^2\}. \tag{3.3}$$

Proof. Observe that $|x|^2 > p$ for large enough p (since $\alpha > \frac{1}{2}$), so that then

$$\log \underline{B} = \left(\frac{p}{2}\right) \{\log(\frac{|x|^2}{p}) + 1 - \frac{|x|^2}{p}\}. \tag{3.4}$$

Expanding this in a Taylors series in $|x|^2/p$ about 1, and using (3.2), yields the conclusion immediately.

3.3.2 The Strict Hierarchical Bayesian Approach A strict Bayesian deals with uncertainty about τ^2 by placing a prior distribution, $g(\tau^2)$, on τ^2. Then the Bayes factor is

$$B(\pi) = N_p(x|0, I)/m(x),$$

where

$$m(x) = \int_0^\infty N_p(x|0, (1+\tau^2)I)g(\tau^2)d\tau^2.$$

Lemma 2. *Suppose $g(\tau^2)$ is bounded and that $g(\tau^2) \propto C(\tau^2)^a$ as $\tau^2 \to 0$, for constants $C > 0$ and $a \geq 0$. Then, for $|x|^2$ as in (3.2) and $\alpha > \frac{1}{2}$,*

$$B(\pi) = Kp^{(a+1)/2} \exp\{-\frac{1}{2}\, z_\alpha^2\}(1 + o(1)) \tag{3.5}$$

as $p \to \infty$, where

$$K = \{2^{(a+2)/2}\sqrt{\pi}\; C\; E^Z[(z_\alpha - Z)^a 1_{(-\infty, z_\alpha)}(Z)]\}^{-1},$$

with Z being standard normal.

Proof. Transforming to $\delta = (1+\tau^2)^{-1}$ yields

$$m(x) = \int_0^1 (2\pi)^{-p/2}\delta^{p/2} \exp\{-\frac{1}{2}\, \delta|x|^2\}g(\frac{1}{\delta} - 1)d\delta.$$

Application of Laplace's Method to the term $\delta^{p/2} \exp\{-\frac{1}{2}\delta|x|^2\}$ yields, as $p \to \infty$,

$$m(x) = \frac{\sqrt{2}\exp\{-p/2\}}{\sqrt{p}(2\pi)^{(p-1)/2}} \cdot \left(\frac{p}{|x|^2}\right)^{(p+2)/2} \int_0^1 N(\delta|\frac{p}{|x|^2}, \frac{2p}{|x|^4})g(\frac{1}{\delta} - 1)d\delta(1 + O(\frac{1}{\sqrt{p}})). \tag{3.6}$$

The indicated bound on the error term arising from Laplace's method is obtained by observing, from (3.2), that

$$\frac{p}{|x|^2} = 1 - z_\alpha\sqrt{\frac{2}{p}} + O(\frac{1}{p}). \tag{3.7}$$

Since, then,

$$\frac{2p}{|x|^4} = \frac{2}{p}(1 + O(\frac{1}{\sqrt{p}})),$$

it is immediate from the assumptions on g and $\alpha > \frac{1}{2}$ that

$$\begin{aligned}
&\int_0^1 N(\delta|\frac{p}{|x|^2}, \frac{2p}{|x|^4})g(\frac{1}{\delta} - 1)d\delta \\
&= \int_{z_\alpha - \sqrt{\frac{p}{2}}}^{z_\alpha} N(z|0,1)g(1/[(\sqrt{\frac{2}{p}}(z_\alpha - z))^{-1} - 1])dz \\
&= \int_{-\infty}^{z_\alpha} N(z|0,1)C[\sqrt{\frac{2}{p}}(z_\alpha - z)]^a dz(1 + o(1)) \\
&= \left(\frac{2}{p}\right)^{a/2} CE^Z[(z_\alpha - Z)^a 1_{(-\infty, z_\alpha)}(Z)](1 + o(1)).
\end{aligned} \tag{3.8}$$

Observe, next, that Lemma 1 can be restated as (see (3.4))

$$\left(\frac{p}{|x|^2}\right)^{p/2} = \exp\left\{\frac{p}{2}\left(1-\frac{|x|^2}{p}\right)+\frac{1}{2}z_\alpha^2\right\}(1+o(1)).$$

Combining this with (3.6) and (3.8) yields

$$B(\pi) = \frac{(2\pi)^{-p/2}\exp\{-\frac{1}{2}|x|^2\}}{\frac{\sqrt{2}\exp\{-p/2\}}{\sqrt{p}(2\pi)^{(p-1)/2}}\cdot\frac{p}{|x|^2}\cdot\exp\{\frac{p}{2}(1-\frac{|x|^2}{p})+\frac{1}{2}z_\alpha^2\}\cdot\frac{k_\alpha^{-1}}{2\sqrt{\pi}p^{a/2}}(1+o(1))}.$$

Upon simplification, this is equal to the right hand side of (3.5), except for an extra factor of $(|x|^2/p)$. But this factor goes to one by (3.7), completing the proof.

Priors, $g(\tau^2)$, that are typically considered in hierarchical modelling, such as Gamma $(a+1, b)$ priors or any continuous bounded prior that is positive at zero, satisfy the conditions of the lemma. In the latter case (i.e., $a = 0$), note that the constant in (3.5) can be written $K = 1/[2\sqrt{\pi}(1-\alpha)\pi(0)]$.

3.3.3 The Discrepancy From (3.3) and (3.5), it is clear that, for large p,

$$\frac{B(\pi)}{\underline{B}} = Kp^{(a+1)/2}(1+o(1)).$$

Since $a \geq 0$, this demonstrates that $\underline{B}$ can underestimate the actual Bayes factor, $B(\pi)$, in hierarchical models by an enormous factor if p is large. And recall that $a \geq 0$ includes virtually all standard proper priors for τ^2. Hence the indication is that robust Bayesian bounds based only on first stage priors can be excessively small in high dimensional hierarchical models.

The discrepancy observed here is even more surprising when one considers that $\hat{\tau}^2 \to \tau^2$ as $p \to \infty$. And the prior at which $\underline{B}$ is attained is the $\mathcal{N}_p(0, \hat{\tau}^2 I)$ prior, which is actually thus converging to the true prior. Nevertheless, the answer obtained by using $\hat{\pi}$ can be dramatically wrong. The moral here obviously applies as well to the Type II MLE approach, which here is also called the empirical Bayes approach.

There is a natural and simple robust Bayesian solution to this problem; simply work with a class of prior distributions on τ^2. This has been studied in Sánchez (1990).

4 Conclusion

We make no effort to review the various insights that arose through study of the examples in the paper. Overall, we feel that a cautious optimism towards use of robust Bayesian bounds is warranted. They are effective in a number of scenarios for settling concerns about robustness, and can provide useful quantitative measures for phenomena such as Ockham's Razor and outlier detection.

At the same time, one must be wary of using the lower bounds themselves as a substitute for "real" Bayesian measures. Real Bayesian measures can differ substantially from the bounds, even when the priors on which the bounds are based seem to be extremely close to the "true" prior. The examples also indicate that care must be taken when using the Type II MLE or empirical Bayes approach.

References

Bayarri, M. J. and Berger, J. (1992). Robust Bayesian bounds for outlier detection. Technical Report #92–43C, Department of Statistics, Purdue University.

Bayarri, M. J. and Berger, J. (1993). Robust Bayesian analysis of selection models. Technical Report #93–6, Department of Statistics, Purdue University.

Bayarri, M. J. and DeGroot,M. (1992). A BAD view of weighted distributions and selection models. In *Bayesian Statistics IV*, J. M. Bernardo, et. al. (Eds.). Oxford University Press, London.

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

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

Berger, J. and Sellke, T. (1987). Testing of a point null hypothesis: the irreconcilability of significance levels and evidence. *J. Amer. Statist. Assoc.* **82**, 112–139.

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

Jahn, R. G., Dunne, B. J., and Nelson, R. D. (1987). Engineering anomalies research. *J. of Scientific Exploration* **1**, 21–50.

Jefferys, W. (1990). Bayesian analysis of random event generator data. *J. of Scientific Exploration* **4**, 153–169.

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

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

Sánchez, J. A. C. (1990). Robustness of the posterior mean in normal hierarchical models. Technical Report, Departamento de Matemática Aplicada y Estadística, Universidad de Murcia, Spain.

Wasserman, L. (1992). Recent methodological advances in robust Bayesian inference. In *Bayesian Statistics IV*, J. M. Bernardo, et. al. (Eds.). Oxford University Press, London.

This article was processed using the LaTeX macro package with LLNCS style