THE APPLICATION OF ROBUST BAYESIAN ANALYSIS TO HYPOTHESIS TESTING AND OCCAM'S RAZOR *

by

James O. Berger and William H. Jefferys
Purdue University University of Texas

Technical Report #91-04

Department of Statistics
Purdue University

February 1991

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

THE APPLICATION OF ROBUST BAYESIAN ANALYSIS TO HYPOTHESIS TESTING AND OCCAM'S RAZOR

*James O. Berger, Purdue University**
William H. Jefferys, University of Texas

1. Introduction

There has recently been considerable interest in the development of the robust Bayesian approach to statistics. Berger (1990) presents a reasonably current review. The basic idea is to replace the common single prior distribution in Bayesian analysis by a wide (often nonparametric) class of priors.

We will be concerned with inference concerning an unknown parameter θ, assumed to lie in the parameter space Θ, with the experimental evidence about θ being provided by the observed likelihood function

$$\ell(\theta) = f(x|\theta);$$

here $f(x|\theta)$ is the density of the observed data x, which will usually be suppressed in the notation.

For a prior distribution π, the posterior distribution of θ is then given (under mild conditions) by

$$\pi^*(d\theta) = \pi(d\theta)\ell(\theta)/m,$$

where $m = \int \ell(\theta)\pi(d\theta)$. In most of our examples, Θ will be Euclidean and π will be assumed to have a density w.r.t. Lebesgue measure. For simplicity in such cases we will let the prior density be denoted by $\pi(\theta)$.

Instead of supposing the specification of a single prior π_0, suppose we know only that $\pi \in \Gamma$, a class of distributions on Θ. This class could arise in at least two ways:

(i) Γ could be used to represent uncertainty in the prior elicitation process;

(ii) Γ could consist of the differing prior distributions of a group of individuals.

In either case, there will be some posterior quantity $\rho(\pi)$ of interest (e.g. the posterior mean, posterior variance, posterior probability of a credible region or

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

hypothesis, or posterior expected loss) and we will seek

$$\underline{\rho}_\Gamma = \inf_{\pi \in \Gamma} \rho(\pi), \quad \overline{\rho}_\Gamma = \sup_{\pi \in \Gamma} \rho(\pi).$$

The hope, of course, is that the range $(\underline{\rho}_\Gamma, \overline{\rho}_\Gamma)$ is small enough that the indeterminacy in the prior is deemed to be essentially irrelevant, allowing a claim of robustness with respect to the prior, or alternatively that either $\underline{\rho}_\Gamma$ or $\overline{\rho}_\Gamma$ alone conveys an interesting message.

We will focus on the application of robust Bayesian analysis to hypothesis testing, with Γ chosen to reflect a wide range of differing prior distributions (as in case (ii) above). The reason is that rather startling conclusions can be obtained, contradicting certain commonly held attitudes towards hypothesis testing. Also, this approach can be seen to yield a quantified Occam's razor, i.e., a theorem establishing that a simpler model is more likely to be true than a complicated model when both models are reasonably compatible with the data.

Section 2 develops the notation for hypothesis testing and reviews and extends a key robust Bayesian result. Section 3 considers the application to Occam's razor, and Section 4 presents an application to multinomial testing.

2. Hypothesis Testing

2.1 Setup

We will consider testing of a simple hypothesis $H_0: \theta = \theta_0$ versus a composite hypothesis $H_1: \theta \in \Theta_1$. Often, but not always, $\Theta_1 = \{\theta \in \Theta: \theta \neq \theta_0\}$. When considering "Occam's razor," H_0 will represent the "simple" model and H_1 the "complicated" model.

In this situation, a prior distribution is specified by π_0, the prior probability that H_0 is true, and $g(\theta)$, the prior density of θ given that H_1 is true. The *posterior probability that H_0 is true given the data is then*

$$\Pr(H_0|x) = \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{1}{B_g}\right]^{-1}, \tag{2.1}$$

where the *Bayes factor*, B_g is given by

$$B_g = \ell(\theta_0) / \int_{\Theta_1} \ell(\theta) g(\theta) d\theta. \tag{2.2}$$

In the absence of specific prior probabilities of the hypotheses, it is common to choose $\pi_0 = \frac{1}{2}$, or to simply use the Bayes factor to measure the evidence against

H_0. Note the simplicity of interpretation: $\Pr(H_0|x)$ is easily understandable by even nontechnical people, and B_g, which can be thought of as the odds for H_0 to H_1 in light of the data, is also readily interpretable. Contrast these to P–values, which are frequently misinterpreted.

The uncertain part of the above program is choice of g. Jeffreys (1961) recommends a particular, not unreasonable, choice, but the conclusion is often sensitive to this choice. It is thus of interest to take a robust Bayesian approach to the problem, which begins by considering a class G of possible choices for g and then determining

$$\underline{B} = \inf_{g \in G} B_g \quad \text{and} \quad \underline{\Pr}(H_0|x) = \inf_{g \in G} \Pr(H_0|x). \tag{2.3}$$

These give lower bounds on the amount of evidence against H_0 that is provided by the data. Upper bounds are also possible — cf., Edwards, Lindman, and Savage (1963), but we will focus on the lower bounds for reasons that will become clear.

A number of somewhat tangential issues can be raised concerning testing of hypotheses such as $H_0: \theta = \theta_0$. One such issue is that it is virtually never the case that one entertains the possibility that $\theta = \theta_0$ exactly; rather, one believes that θ might be "close" to θ_0. Conditions under which a point null can be used to approximate this more realistic hypothesis are given in Berger and Delampady (1987). Discussion of other somewhat controversial issues concerning such testing can be found in Edwards, Lindman, and Savage (1963), Berger and Sellke (1987) and Berger and Delampady (1987). Related work includes Casella and Berger (1987), DeGroot (1973), Delampady (1989a, 1989b), Dempster (1973), Dickey (1977), Good (1983, 1984), Jeffreys (1961), Lindley (1957), Shafer (1982), Smith and Spiegelhalter (1980), and Zellner (1984).

2.2 The Robust Bayesian Methodology

We will utilize in our applications only one of the simplest of the robust Bayesian technical results, the proof of which is standard.

Lemma 1. Suppose that the class of conditional priors g is

$$G = \{g(\theta) = \int_{\Theta_1} g_\tau(\theta) dF(\tau) : F \text{ is any c.d.f. on } [0, \infty)\}. \tag{2.4}$$

Then

$$\underline{B} = \inf_{g \in G} B_g = \frac{\ell(\theta_0)}{\sup_{\tau \geq 0} \int_{\Theta_1} \ell(\theta) g_\tau(\theta) d\theta}. \tag{2.5}$$

Example 1. Suppose $\theta \in \mathbb{R}^1$ and

$$\ell(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(x-\theta)^2/(2\sigma^2)}. \tag{2.6}$$

Consider

$$G_{\theta_1} = \{\text{all } g(\theta) \text{ that are symmetric about } \theta_1 \text{ and nonincreasing in } |\theta - \theta_1|\}. \tag{2.7}$$

Then a standard argument yields that G_{θ_1} is as in (2.4), with the $g_r(\theta)$ being the Uniform $(\theta_1 - r,\ \theta_1 + r)$ densities. Thus (2.5) yields

$$\underline{B} = \ell(\theta_0) / \sup_{r \geq 0} \frac{1}{2r} \int_{\theta_1 - r}^{\theta_1 + r} \ell(\theta) d\theta. \tag{2.8}$$

An iterative expression for computing (2.8) is given in Berger and Sellke (1987), but a quite accurate closed form approximation is available. Indeed, the approximation

$$\hat{\underline{B}} = \sqrt{\frac{2}{\pi}}\, e^{-t_0^2/2} [t_1 + \sqrt{2 \log(t_1 + 1.2)}], \tag{2.9}$$

where $t_0 = |x - \theta_0|/\sigma$ and $t_1 = |x - \theta_1|/\sigma$, turns out to be within $o(1)$ of $\underline{B}$ as $t_1 \to \infty$ and always accurate within 1% if $t_1 > 1.4$. This lower bound corresponds to a maximizing r in (2.8) of approximately

$$\hat{r} = t_1 + [2 \log(t_1 + \sqrt{2 \log(t_1 + 1.2)}) - \log(2\pi)]^{1/2}. \tag{2.10}$$

2.3 Comparison with P–values

The first use of these results, in Berger and Sellke (1987), was in comparing $\underline{B}$ with the corresponding P–value in testing $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$. Here is an interesting example.

Example 1(a). In Jefferys (1990), a large experiment involving psychokinesis was reanalyzed from the Bayesian perspective. When expressed in the canonical form of Example 1, the null hypothesis of "no psychokinesis present" is $H_0: \theta = 0$, while the alternative hypothesis of "psychokinesis present" is $H_1: \theta \neq 0$. The sample size was huge ($n = 104{,}490{,}000$) so that normality could be assumed, and the standardized test statistic was

$$t_1 = \frac{|x - 0|}{\sigma} = 3.614.$$

The corresponding P–value was about .0003, which was argued in the original analysis by the experimenters to provide very strong evidence in favor of H_1.

Jefferys's Bayesian analysis indicated that quite the opposite conclusion probably holds. Indeed, he argued that for reasonable conditional priors, $g(\theta)$, the Bayes

factor is actually greater than one, indicating the data supports H_0, not H_1. Of interest for this paper is the lower bound, $\underline{B}$, on the Bayes factor (obtainable from (2.8)), since this can be viewed as the maximum possible support for H_1 that the data provides. (One can argue that the class G_0 in (2.7) includes all reasonably "objective" g.)

In this example, (2.9) applies with $\theta_0 = \theta_1 = 0$ and $t_0 = t_1 = 3.614$, yielding $\underline{\hat{B}} = 1/159$. The interpretation of a Bayes factor of $1/159$ is that H_1 is supported 159 times more than H_0, but recall that this is the lower bound on B_g. It is achieved at the Uniform $(-.00022, .00022)$ conditional prior distribution $g(\theta)$ (the value .00022 being computed from (2.10)), and most people would view a prior that was so narrowly constrained about zero to be unreasonable (hence the argument by Jefferys (1990) that "reasonable" $g(\theta)$ show that the data actually support H_0). In any case, it is of interest that the very small P–value of .0003 "translates" into at most $1/159$ evidence against H_0.

In Berger and Sellke (1987) it is similarly shown that a P–value of 0.05 translates into at most $1/2\frac{1}{2}$ evidence against H_0, and a P–value of 0.01 into at most $1/8$ evidence against H_0.

3. Occam's Razor

3.1 Background

Occam's razor, that is, the principle that an explanation of the facts should be no more complicated than necessary, is an accepted principle in science. Over the years it has proven to be an effective tool for weeding out unprofitable hypotheses, and scientists use it every day, even when they do not cite it explicitly.

Occam's razor is usually thought of as an heuristic, that experience has shown to be an effective tool. It is less widely known, however, that under some circumstances it can be regarded as a *consequence* of deeper principles. This fact is implicit in Jeffrey's book on probability (1939), and has more recently been emphasized by Jaynes (1979), Smith and Spiegelhalter (1980), Gull (1988), and Loredo (1989).

Jeffreys (1939) considered the problem of fitting observed data to an empirical function. Considering a falling body, he considers the law

$$s = a + ut + \frac{1}{2}gt^2, \tag{3.1}$$

where a, u and g are adjustable parameters. So far this is only a standard problem in estimation theory. However, there are infinitely many possible laws that can represent the data set. For example, Jeffreys considers alternative laws of the form

$$s = a + ut + \frac{1}{2}gt^2 + a_3t^3 + \ldots + a_nt^n, \tag{3.2}$$

where n is greater than the number of observations and all coefficients are adjustable. Given such a law, there are infinitely many choices of the parameters that will exactly fit the data, and the question is, why do we prefer (1) over (2)? The easy answer, given by Occam's razor, is that we ought to prefer (1) to (2), assuming that (1) adequately represents the observed data, on the grounds that (2) is unnecessarily complicated. On the other hand, (2) can actually represent the observed data points better than (1), since it can be arranged to pass exactly through each data point. So there must be something other than the ability to fit the data that leads us to prefer the simpler law to the more complex.

3.2 Quantification of Occam's Razor

Consider the situation where we can identify the simpler law with $H_0: \theta = \theta_0$ (θ_0 specified) and the more complex law with $H_1: \theta \in \Theta_1$. The law H_1 is more complex because it has a "free parameter" θ. Then if data x is collected according to $f(x|\theta)$, Bayesian reasoning states that the Bayes factor, B_g, reflects the comparative support of the data for H_0 and H_1. If, furthermore, only $g \in G$ need to be considered, then

$$\underline{B} = \inf_{g \in G} B_g$$

becomes a lower bound on the comparative evidence for H_0 to H_1. If this lower bound happens to be large, we can conclude that the evidence strongly supports the simpler model H_0.

Example 1(b). If $f(x|\theta)$ is Normal (θ, σ^2), and it is reasonable to consider only $g \in G_{\theta_1}$ from (2.7) (as is commonly the case when $\Theta_1 = \{\theta: \theta \neq \theta_1\}$), then (2.8) (or (2.9)) provides the quantification of Occam's razor.

3.3 An Example: The Motion of Mercury's Perihelion

Ever since Leverrier's work in the early 19th century, astronomers were aware of a serious problem with the theory of Mercury's motion. Newtonian theory, which had been extraordinarily successful in accounting for most of the motions in the solar system, had run up against a small discrepancy in the motion of Mercury that it could not explain easily. After all of the perturbing effects of the planets had been taken into account, there remained an unexplained residual motion of Mercury's perihelion (the point in its orbit where the planet was closest to the Sun) in the amount of approximately 43 seconds of arc per century.

Clearly, it seemed as if something had been overlooked. It was known that physical mechanisms existed that might explain the discrepancy. One that seemed particularly appealing in the light of recent experience was the possibility that another planet might exist, closer to the Sun than Mercury. The reason that this idea was so appealing was that Leverrier himself, along with the English astronomer

Adams, had recently (in 1846) met with brilliant success by predicting that a previously unknown planet was responsible for the known discrepancies in the motion of Uranus; not only did Leverrier and Adams hypothesize that such a planet existed, but they also suggested where it might be found, and indeed, when J.G. Galle looked for it, the planet Neptune was discovered in the predicted place. It certainly seemed possible that a similar phenomenon might explain the anomaly in Mercury's motion.

Indeed, a number of astronomers duly set out to find the new planet, dubbed "Vulcan" in anticipation of its discovery, and some sightings were announced. However, the sightings could not be confirmed, and over time interest in the Vulcan hypothesis waned.

Other mechanisms that might explain the anomaly were also proposed. It was suggested that rings of material around the Sun could, if massive enough, produce the observed effect; or, the Sun itself might be slightly oblate, due to its rotation on its axis; or, finally, the law of gravity itself might not be exactly right. The astronomer Simon Newcomb, for example, proposed that the exponent in Newton's law of gravity might not be exactly 2, but instead might be $2 + \epsilon$, although other modifications to the law of gravity were also possible.

All these hypotheses had one characteristic in common: they possessed a parameter that could be adjusted to agree with whatever data on the motion of Mercury existed. In modern parliance, we would call this a "fudge factor." For example, the Vulcan hypothesis had the mass of the putative planet; the ring hypothesis had the mass of the ring of material; the solar oblateness hypothesis had the unknown amount of the oblateness; and all the hypotheses that modified Newton's law of gravity had an adjustable parameter (like Newcomb's ϵ) that could be chosen at will.

Not all the hypotheses were equally probable, however (Roseveare, 1982). As we stated above, sightings of "Vulcan" were never confirmed, for example. As time went on, the hypothesis of matter rings of sufficient density became less and less likely (Jeffreys, 1921) although some still believed in them (Poor, 1921). A solar oblateness of sufficient size probably would have been detectable with 19th century techniques. However, the hypothesis that Newton's law of gravity needed an arbitrary adjustment to fit the data could not be ruled out.

What happened historically is well known. In 1915, Einstein announced his theory of general relativity, one of the consequences of which was that there should be an excess advance in the perihelion motion of the planets that was largest for Mercury. After some confusion (Roseveare, 1982: pp. 154–159) it soon became clear that the amount of the advance predicted by general relativity was very close to the unexplained discrepancy in Mercury's motion. The amazing thing was that

the predicted value (42.98"/century using modern values, Nobili & Wills, 1986) was *not* some kind of fudge factor, but instead was an inevitable consequence of Einstein's theory!

As is well known, Einstein's theory made two other major predictions in addition to Mercury's perihelion motion (gravitational bending of light, and the slowing down of clocks in a gravitational field). There has been a lively debate over the years as to how important each has been in convincing scientists that general relativity was the correct theory of gravity (Brush, 1989). In this paper we will not go into this argument, but will instead try to put ourselves into the mindset of a Bayesian observer in the early 1920s, who is trying to weigh the evidence of Mercury's motion.

An interesting pair of papers was published in 1921 (Poor, 1921; Jeffreys, 1921). Poor was an astronomer at Columbia University, who had not been convinced that general relativity was correct and still clung to the matter ring theory. Unfortunately, he also made some serious errors in his assessment of the evidence as regards the other inner planets. Jeffreys, in response, argued persuasively that the ring theory was not viable because sufficient matter did not exist. This paper was published before Jeffreys made his major contributions to probability theory, and he does not, ironically, make the Bayesian argument that we discuss.

Poor gives the data $x = 41.6" \pm 1.4"$ for the centennial anomalous motion of Mercury. The uncertainty is undoubtedly a "probable error" (conforming to convention at the time) so the standard deviation would be $\sigma = 2.0"$. Poor also gives $\theta_0 = 42.9"$ as the amount predicted by Einstein's theory, which is very close to the modern value. Thus, in our setup, general relativity would correspond to the hypothesis $H_0: \theta = 42.9"$, where θ refers to the true perihilion advance.

Specifying the alternative hypothesis, H_1, is not as difficult as it might at first appear to be. Adopting the "prior–to–data" perspective, one can ask — what value of θ would one anticipate, conditional on any of the "fudged Newton" hypotheses being true? Recall that we have scaled so that $\theta = 0$ corresponds to Newtonian theory, so a first effort might be simply to define $H_1: \theta \neq 0$. Also, large deviations from Newtonian theory would likely have seemed less plausible apriori than small deviations, and most of the alternative theories (with the possible exception of "Vulcan") would have equally allowed positive or negative θ. Thus the class of priors, G, specified by (2.7) with $\theta_1 = 0$ would have seemed very reasonable. (Further restrictions could be imposed upon G by incorporating other prior information — for instance, any law resulting in $|\theta| > 100$ would have caused anomalies in the orbits of other planets that would quite likely have been detected — but we shall see that there is no need to further refine G.)

Applying the quantification of Occam's razor in (2.9) (implicitly assuming a normal

error for the data), we obtain

$$t_0 = \frac{|41.6 - 42.9|}{2.0} = 0.65 \quad \text{and} \quad t_1 = \frac{|41.6 - 0|}{2.0} = 20.8,$$

and

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

Thus the data supports the simpler law (general relativity) over any of the more complex laws by at least a factor of 15 to 1. The $g(\theta)$ at which this minimum is attained is the Uniform ($-44.87''$, $44.87''$) density (determined from (2.10).) This is not an unreasonable $g(\theta)$ (in contrast with the minimizing g in Example 1(a)) but it still is the g that is "most favorable" to H_1. Hence the 15 to 1 odds in favor of general relativity should be thought of as probably too low, i.e., the evidence for the simpler general relativity is actually probably quite a bit stronger.

Note that this happened in spite of the fact that the data is, if anything, more compatible with H_1 than H_0 (since H_1 can fit the data exactly). The Occam's razor effect is thus dramatic.

3.4 Conclusions

The situation in the example of Mercury's perihilion is not uncommon, and one can restate the conclusion as follows:

Occam's Razor (Quantified): Suppose data conflicts with an "established" theory by t_1 standard deviations. Two alternative theories are proposed. Theory 1 has no additional parameters and conflicts with the data by t_0 standard deviations. Theory 2 has an additional free parameter and so can exactly accommodate the data. Then the odds for Theory 1 over Theory 2 are at least $\underline{\hat{B}}$ (see 2.9)).

As a final comment, it should be observed that there are other "Occam's razors." One such is that simpler models may well be assigned larger prior probabilities. (Jeffreys (1939) seems to argue for such.) Note that, by operating only with Bayes factors, we circumvented the issue of what prior probabilities to assign hypotheses. The Occam's razor we developed does not depend on prior probabilities of the hypotheses.

Another "Occam's razor" is that simpler models may be more useful for reasons of parsimony. They may be just as good for actual predictive purposes as a "true" but more complex model, and their simplicity would then make them attractive for practical use. The Occam's razor in this paper bears no obvious relationship to such a "parsimonious Occam's razor."

4. A Multi–dimensional Example: Multinomial Testing

Similar ideas and results apply in higher dimensions. As an illustration, we consider the multinomial testing problem. (See Good, 1967, and Berger and Delampady, 1990) for related results.) Thus, suppose $X = (X_1, \ldots, X_{p+1}) \sim$ Multinomial (n, θ), where the x_i are nonnegative integers, $\sum_{i=1}^{p+1} x_i = n$, and

$$\theta \in \Theta = \{(\theta_1, \ldots, \theta_p) : 0 < \theta_i < 1 \text{ for } i = 1, \ldots, p, \text{ and } \sum_{i=1}^{p} \theta_i < 1\}.$$

Defining $\theta_{p+1} \equiv 1 - \sum_{i=1}^{p} \theta_i$, the resulting likelihood for θ given x is then proportional to

$$\ell(\theta) = \prod_{i=1}^{p+1} \theta_i^{x_i}, \tag{4.1}$$

and the Bayes factor (2.2) for testing

$$H_0 : \theta = \theta^0 = (\theta_1^0, \ldots, \theta_0^0) \quad \text{vs.} \quad H_1 : \theta \neq \theta^0$$

is

$$B_g = \ell(\theta^0) / \int_\Theta \ell(\theta) g(\theta) d\theta. \tag{4.2}$$

It is not easy to define a sensible G here, for which computation of the Bayes factor and lower bounds are relatively easy. One possibility is to first transform to "centered" log odds

$$\xi_i(\theta) = \log \frac{\theta_i}{\theta_{p+1}} - \log \frac{\theta_i^0}{\theta_{p+1}^0}. \tag{4.3}$$

It will frequently be natural to consider coordinatewise symmetry in the $\xi_i(\theta)$, leading to natural "base" conditional priors

$$g_r(\underset{\sim}{\xi}) = \frac{1}{(2r)^p} \prod_{i=1}^{p} I_{(-r,r)}(\xi_i),$$

where $\underset{\sim}{\xi} = (\xi_1, \ldots, \xi_p)$ and $I_{(-r,r)}(\xi_i)$ stands for the indicator function on $(-r, r)$. A wide class, G, of "plausible" conditional priors is then given by mixing over r, as in (2.4).

As in Lemma 1, it then follows that

$$\underline{B} = \inf_{g \in G} B_g = \inf_{r > 0} B_r,$$

where (transforming variables)

$$B_r = (2r)^k / \int_{(-r,r)^p} \ell(\underset{\sim}{\xi}) d\underset{\sim}{\xi}; \tag{4.4}$$

here $(-r,r)^p = (-r,r) \times (-r,r) \times \ldots (-r,r)$, and

$$\ell(\underset{\sim}{\xi}) = \left[\theta^0_{p+1} + \sum_{i=1}^{p} \theta^0_i e^{\xi_i}\right]^{-n} \exp\left\{\sum_{i=1}^{p} \xi_i x_i\right\}. \tag{4.5}$$

A Monte-Carlo approximation to the integral in (4.4) (see the Appendix for discussion) is, for $r > 0$,

$$\hat{B}_r = \frac{(2r)^p m \left(\prod_{i=1}^{p} \tau_i d_i(r)\right)^{-1}}{\sum_{j=1}^{m} \left[\ell(\underset{\sim}{\xi}^{(j)}(r)) \prod_{i=1}^{p} (1 + |\xi_i^{(j)}(r) - \hat{\xi}_i|/\tau_i)^2\right]}, \tag{4.6}$$

where, for $i = 1, \ldots, p$,

$$\hat{\xi}_i = \log\left(\frac{x_i \theta^0_{p+1}}{x_{p+1}\theta^0_i}\right), \qquad \tau_i = (.675)\left(\frac{1}{x_i} + \frac{1}{x_{p+1}}\right)^{1/2}, \tag{4.7}$$

$$c_i(r) = -\frac{(r + \hat{\xi}_i)}{(\tau_i + |r + \hat{\xi}_i|)}, \qquad d_i(r) = \frac{(r - \hat{\xi}_i)}{(\tau_i + |r - \hat{\xi}_i|)} - c_i(r),$$

$$\xi_i^{(j)}(r) = \hat{\xi}_i + \tau_i W_{i,j}(1 - |W_{i,j}|)^{-1}, \; W_{i,j} = c_i(r) + d_i(r) U_{i,j}; \tag{4.8}$$

here the $\{U_{i,j}\}$ are i.i.d. $\mathcal{U}(0,1)$ random variables.

As examples, consider the three situations

(i) $p = 2, n = 13, x_1 = 7, x_2 = 5, x_3 = 1, \theta^0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$;

(ii) $p = 3, n = 11, x_1 = 7, x_2 = 2, x_3 = x_4 = 1, \theta^0 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$;

(iii) $p = 3, n = 14, x_1 = 9, x_2 = 3, x_3 = x_4 = 1, \theta^0 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.

The lower bounds $\underline{B}$ can be found by minimization of $\hat{B}_r$ in (4.6) over r. (In carrying out these minimizations, one should fix the random variates used in the Monte–Carlo computation.) These lower bounds, and the corresponding P–values for the three cases, turn out to be

(i) $\underline{B} = 0.752$, P-value $= 0.100$;

(ii) $\underline{B} = 0.257$, P-value $= 0.053$;

(iii) $\underline{B} = 0.065$, P-value $= 0.008$.

A rather strong "Occam's razor" effect is evidenced in all cases. For instance, in case (i) the data appears to fit the more complex model H_1 quite a bit better than the simpler model H_0, but the Bayes factor is not much smaller than 1 for any prior $g \in G$. In all cases the Bayesian Occam's razor shifts the weight of evidence towards the simpler model.

Appendix

That $\hat{B}_r$ in (4.6) is a Monte-Carlo approximation to B_r follows directly from the observation that the $\xi_i^{(j)}(r)$ in (4.8) are independent random variables with density

$$g_i(\xi_i^{(j)}(r)) = 1/\left[\tau_i d_i(r)(1 + |\xi_i^{(j)}(r) - \hat{\xi}_i|/\tau_i)^2\right], \qquad (A.1)$$

so that the expression in (4.6) is the usual Monte-Carlo approximation with importance function $\prod_{i=1}^{p} g_i(\cdot)$ (cf. Berger, 1985, for discussion). The reason for choosing this importance function is that it is easily computable, easy to generate random variables from, has fat tails, and mimics the likelihood function on the domain of integration.

In elaboration of this last point, note that the usual "observed likelihood" approximation to $\ell(\underset{\sim}{\xi})$ is proportional to a $\mathcal{N}_p(\hat{\underset{\sim}{\xi}}, \mathfrak{V})$ density, where $\hat{\underset{\sim}{\xi}} = (\hat{\xi}_1, \ldots, \hat{\xi}_p)$ with the $\hat{\xi}_i$ defined in (4.7), and

$$\mathfrak{V} = \text{ diag } \left\{\frac{1}{x_1}, \ldots, \frac{1}{x_p}\right\} + \frac{1}{x_{p+1}} \ (1),$$

with diag { } denoting a diagonal matrix with the given diagonal entries, and (1) denoting the $p \times p$ matrix of all ones. Here $\hat{\underset{\sim}{\xi}}$ is the m.l.e. of $\ell(\underset{\sim}{\xi})$, and $\mathfrak{V}$ is the inverse of the observed information corresponding to $\ell(\underset{\sim}{\xi})$ (cf. Berger, 1985). Because fatter tails than normal are desirable for an importance function, we consider, for ξ_i, the importance function

$$g_i^*(\xi_i) = \frac{1}{2\tau_i(1 + |\xi_i - \hat{\xi}_i|/\tau_i)^2}, \qquad (A.2)$$

with quartiles chosen to match the normal density (hence the choice of τ_i in (4.7)). For simplicity, we consider the ξ_i to be independent in this new importance function.

The final alteration needed arises because the domain of integration is $\xi_i \in (-r, r), i = 1, \ldots, p$. One of the pleasant features of g_i^* in (A.2) is that the conditional density obtained by conditioning on $\xi_i \in (-r, r)$ becomes the also simple

(A.1). Hence the densities in (A.1) define the actual importance function used. Note that, if one attempted to incorporate into the importance function the dependence among the original ξ_i (or that in the $N_p(\hat{\xi}, \hat{\Sigma})$ approximation), the ability to transform the importance function to have range precisely equal to the domain of integration would be lost.

References

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

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

BERGER, J. & DELAMPADY, M. (1987). Testing precise hypotheses (with Discussion). *Statistical Science* **2**, 317–352.

BERGER, J. & SELLKE, T. (1987). Testing of a point null hypothesis: the irreconcilability of significance levels and evidence (with Discussion). *Journal of the American Statistical Association* **82**, 112–139.

BRUSH, S. (1989). Prediction and theory evaluation: The case of light bending. *Science* **246**, 1124–1129. See also the responses to this article in *Science* **248**, 422–423.

CASELLA, G. & BERGER, R.L. (1987). Reconciling Bayesian and frequentist evidence in the one–sided testing problem. *J. Amer. Statist. Assoc.* **82**, 106–111.

DEGROOT, M.H. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *J. Amer. Statist. Assoc.* **68**, 966–969.

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

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

DELAMPADY, M. & BERGER, J. (1990). Lower bouunds on posterior probabilities for multinomial and chi–squared tests. *Ann. Statist.* **18**, 1295–1316.

DEMPSTER, A.P. (1973). The direct use of likelihood for significance testing. In *Proc. of the Conference on Foundational Questions in Statistical Inference* (O. Barndorff–Nielsen et al., eds.) 335–352. Dept. Theoretical Statistics, Univ. Aarhus.

DICKEY, J.M. (1977). Is the tail area useful as an approximate Bayes factor? *J. Amer. Statist. Assoc.* **72**, 138–142.

EDWARDS, W., LINDMAN, H. & SAVAGE, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review* **70**, 193–242.

GOOD, I.J. (1967). A Bayesian significance test for the multinomial distribution. *J. Roy. Statist. Soc. Ser. B* **29**, 399–431.

GOOD, I.J. (1983). *Good Thinking: The Foundations of Probability and Its Applications.* Univ. Minnesota Press, Minneapolis.

GOOD, I.J. (1984). Notes C140, C144, C199, C200 and C201. *J. Statist. Comput. Simulation* **19**.

GULL, S. (1988). Bayesian inductive inference and maximum entropy. In G.J. Erickson and C.R. Smith (eds.), *Maximum Entropy and Bayesian Methods in Science and Engineering (Vol. 1)*, 53–74. Dordrecht: Kluwer Academic Publishers.

JAYNES, E.T. (1979). Inference, method, and decision: Towards a Bayesian philosophy of science. *Journal of the American Statistical Association* **74**, 740–741.

JEFFERYS, W.H. (1990). Bayesian analysis of random event generator data. Technical Report, Department of Astronomy, University of Texas at Austin.

JEFFREYS, H. (1921). Secular perturbations of the inner planets. *Science* **54**, 248.

JEFFREYS, H. (1939). *Theory of Probability, Third Edition* (1961). Oxford: Clarendon Press.

LINDLEY, D.V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

LOREDO, T.J. (1990). From Laplace to Supernova 1987A: Bayesian inference in astrophysics. In P. Fougere (ed.), *Maximum Entropy and Bayesian Methods*, 81–142. Dordrecht: Kluwer Academic Publishers.

NOBILI, A.M. & WILL, C.M. (1986). The real value of Mercury's perihelion advance. *Nature* **320**, 39–41.

POOR, C.L. (1921). The motions of the planets and the relativity theory. *Science* **54**, 30–34.

ROSEVEARE, N.T. (1982). *Mercury's Perihelion from Le Verrier to Einstein.* Oxford: Clarendon Press.

SHAFER, G. (1982). Lindley's paradox (with discussion). *J. Amer. Statist. Assoc.*

77, 325–351.

SMITH, A.F.M. & SPIEGELHALTER, D.J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. B* **42**, 213–220.

THORBURN, W.M. (1918). The myth of Occam's razor. *Mind* **27**, 345–353.

ZELLNER, A. (1984). Posterior odds ratios for regression hypotheses: General considerations and some specific results. In *Basic Issues in Econometrics* (A. Zellner, ed.), 275–305.