RECENT DEVELOPMENTS AND APPLICATIONS OF BAYESIAN ANALYSIS

by

James O. Berger
Purdue University

Technical Report #95-17C

Department of Statistics
Purdue University

April 1995

RECENT DEVELOPMENTS AND APPLICATIONS OF BAYESIAN ANALYSIS*

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

1. INTRODUCTION

1.1 Goals of the Paper

Bayesian analysis has experienced spectacular growth in recent years, so much so that it cannot be adequately reviewed in a single article. (Recent general books or reviews of special areas of Bayesian statistics are listed in the references.) This article has the considerably more modest goal of illustrating three of the reasons why this spectacular growth is occurring.

One reason is a dramatic increase in computational capability. Ten years ago, Bayesian analysis was limited to comparatively simple problems, because of the difficulties in performing high dimensional integration. The common complaint about Bayesian analysis, in those days, was that, while intuitively appealing, it could not be used for problems of real complexity. The story today is just the opposite: Bayesian analysis is now the preferred (and often the only) method of analyzing highly complex problems. The reason is the advent of Markov Chain Monte Carlo computational tools, as well as access to sufficient computing power for implementation of these tools. This issue will be discussed in Section 2, with a prototypical example being given for illustration.

The second reason for growth in Bayesian statistics is best described as foundational, in the sense of being related to basic justifications of the Bayesian approach. Three of these justifications are: (i) Bayesian answers are particularly easy to interpret, which is important in a world where statistical answers are routinely misinterpreted; (ii) Bayesian

*Presidential Invited Paper for the 50th Session of the International Statistical Institute.

methods have great flexibility in operation; and (iii) Bayesian methods and "good" classical methods tend to agree. In Section 3, we illustrate how these ideas are today coming together, using a clinical trial example for illustration. The example demonstrates the simplicity and flexibility of the Bayesian approach, and also provides a vehicle for discussion of recent theoretical developments that surprisingly show the Bayesian procedures to be superior frequentist procedures.

The final reason that we discuss, for the upsurge in interest in Bayesian methods, can be termed methodological, in the sense of introduction of new Bayesian methodologies. One such recent methodology is 'robust Bayesian analysis,' which is concerned with problems in which there is a multitude of prior opinions that must simultaneously be accommodated. We illustrate this new methodology in Section 4 on the problem of quantifying Ockham's razor, which is the heuristic scientific principle that, if two models are equally well supported by the data, the simpler model is to be preferred.

1.2 Notation

Notation will be kept basic. The (entire) data will be denoted by X, assumed to have a conditional density $f(x|\theta)$, given an unknown parameter $\theta \in \Theta$, the parameter space. A prior density for θ will be denoted by $\pi(\theta)$, with the posterior density being

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

here $m(x) = \int f(x|\theta)\pi(\theta)d\theta$ is the marginal or predictive density. The focus of Bayesian analysis is typically computation of various posterior expectations

$$E^*[g(\theta)] = \int g(\theta)\pi(\theta|x)d\theta;$$

the "$*$" as a superscript to E will always denote posterior expectation, and if $g(\theta)$ is a vector or matrix, the expectation is to be taken componentwise. Common choices of g include

$g(\theta) = \theta$, since then $E^*[\theta] =$ *posterior mean*;

$g(\theta) = (\theta - E^*[\theta])(\theta - E^*[\theta])^t$, since then $E^*[g(\theta)] =$ *posterior covariance matrix*;

$g(\theta) = 1_C(\theta) = \begin{cases} 1 & \text{if } \theta \in C \\ 0 & \text{otherwise,} \end{cases}$ since then $E^*[g(\theta)] =$ *posterior probability of* C.

2. COMPUTATIONAL DEVELOPMENTS IN BAYESIAN ANALYSIS

A major development that has greatly contributed to the upsurge in use of Bayesian methods is the development of computational tools that allow analysis of highly complex and nonstandard models. Indeed, for complicated models, Bayesian analysis has now arguably become the *simplest* (and often only possible) method of analysis. We illustrate this point by presenting an example in Section 2.1, followed by a brief review in Section 2.2 of recent computational advances.

2.1 An Illustration

The best way to illustrate the power of current Bayesian methods is to present an example. The following example is based on a problem studied in Andrews, Berger, and Smith (1993), though certain features of the problem are here simplified or enriched for purposes of exposition. We do not actually analyze this example here; our goal is simply to indicate the potential of current Bayesian methods.

The problem was to determine the effect of certain automotive technologies, such as fuel injection, on the fuel efficiency of automobiles. After certain transformations of the data and variables, the base model became

$$Y_{ijk} = \beta^t X_{(ijk)} + \alpha^t X^*_{(ij)} + \varepsilon_{ijk};$$

here

Y_{ijk} = log (fuel efficiency in MPG) of a vehicle;

$X_{(ijk)}$ = a vector of the vehicle characteristics, including indicators of presence of technologies of interest;

β = a vector of unknown "fixed effects";

$X^*_{(ij)}$ = an indicator vector specifying the vehicle model and manufacturer;

α = a vector of unknown "random effects";

$i = 1, \ldots, I$, denoting the manufacturer;

$j = 1, \ldots, J_i$, denoting the vehicle model for manufacturer i;

$k = 1, \ldots, N_{ij}$, denoting a particular vehicle of model j from manufacturer i.

Error distributions deemed possible are $\varepsilon_{ijk} \overset{i.i.d.}{\sim}$ Normal $(0, \sigma^2)$, with σ^2 unknown, and $\varepsilon_{ijk} \overset{i.i.d.}{\sim} t$-distribution with median 0, unknown scale σ, and unknown degrees of freedom ν. The model is complicated by being unbalanced (the J_i and the N_{ij} are highly variable) and there is considerable missing data.

For the fixed effects, β, certain sign and order restrictions are known; indeed, it is known that

$$\beta \in \Omega = \{\beta: \beta_{10} > 0, \beta_{15} > 0, \beta_{18} > 0, \beta_4 \leq \beta_5 \leq \beta_6\}.$$

The car model effects, α, are modeled as

$$\alpha_{ij} \overset{i.i.d.}{\sim} \text{Normal } (\mu_i, V_i), \quad j = 1, \ldots, J_i,$$

where μ_i and V_i are the overall mean and variance for manufacturer i. It is believed, however, that there is a time trend to the overall manufacturer means; this is modeled by the AR (1) process

$$\mu_i(t) = \rho_i \mu_i(t-1) + \gamma_{it}, \quad i = 1, \ldots, I,$$

where t denotes the year of vehicle manufacture, and the γ_{it} are i.i.d. Normal $(0, \gamma)$ errors. Finally, the unknown ρ_i and V_i are also modeled as random effects from the population of all manufacturers, with the ρ_i being i.i.d. Beta (λ, τ) and the V_i being i.i.d. Inverse Gamma (ξ, η).

The desired goal of the analysis is to predict fuel efficiencies, Y, but at uncertain (i.e., random) future vehicle configurations (X, X^*). Estimates, standard errors, and confidence (credible) sets for Y are desired, as well as tests for certain of the $\beta_i = 0$.

There are numerous features of this problem that would virtually preclude the possibility of a classical analysis. Having both fixed and random effects in an unbalanced situation, even with normal errors, is by itself enough to almost require a Bayesian analysis (to produce reasonable standard errors and credible sets). Adding the complications of t-errors, restrictions on the parameters, time series structures for some of the random effects, and the desire to predict Y at random future X creates a problem of almost unapproachable complexity from a classical perspective.

Solving this problem from the Bayesian perspective is comparatively straightforward. One must first place a prior distribution on all unknown parameters that do not already

have a "random effects" distribution. The simplest possibility is the constant density on these parameters (restricted to Ω, of course), i.e.,

$$\pi(\beta, \gamma, \lambda, \tau, \xi, \eta, \sigma^2, \nu) = 1_\Omega.$$

(Choosing improper "noninformative" densities such as this can be justified from a number of perspectives; see Berger and Bernardo (1992).) Using the techniques discussed in the next section, one can then compute posterior means and variances (and other desired posterior expectations) for any of the unknown parameters, or for future Y.

There are, of course, the usual variety of concerns with the above analysis, centering around issues of sensitivity to, and plausibility of, assumptions (including choice of the prior density). Also, the computation required is far from trivial. The point to be stressed, however, is that with the Bayesian approach the statistician has complete freedom to utilize whatever models, structures, or restrictions seem reasonable for a particular problem, while maintaining the capability to compute answers. There is no need to force the problem into a standard mold by oversimplification.

2.2 Markov Chain Simulation Techniques

The newest techniques to be extensively utilized for numerical Bayesian computations are Markov Chain Simulation Techniques, including the popular Gibbs Sampling. (Certain of these techniques are actually quite old – see, e.g., Hastings (1970); it is their application and adaption to Bayesian problems that is new.) A brief generic description of these methods is as follows:

Step 1. Select a "suitable" Markov chain on Θ, with $p(\cdot, \cdot)$ being the transition probability density (i.e., $p(\theta, \theta^*)$ gives the transition density for movement of the chain from θ to θ^*). Here "suitable" means primarily that $\pi(\theta|x)$ is a stationary distribution of the Markov chain, which can be assured in a number of ways.

Step 2. Starting at a point $\theta^{(0)} \in \Theta$, generate a sequence of points $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)}$ from the chain.

Step 3. Then, for large m, $\theta^{(m)}$ is (approximately) distributed as $\pi(\theta|x)$ and

$$\frac{1}{m} \sum_{i=1}^{m} g(\theta^{(i)}) \cong E^*[g(\theta)].$$

The main strengths of Markov chain methods for computing $E^*[g(\theta)]$ are:

(i) Many different g can simultaneously be handled via Step 3, once the sequence $\theta^{(1)}, \ldots, \theta^{(m)}$ has been generated.

(ii) Programming tends to be comparatively simple.

(iii) Methods of assessing convergence and accuracy exist and/or are being developed.

The main weaknesses of the Markov chain methods are:

(i) They can be quite slow. It is not uncommon in complicated problems to need m to be in the hundreds of thousands, requiring millions of random variable generations if the dimension of θ is appreciable.

(ii) One can be misled into prematurely judging that convergence has obtained.

The more common Markov chain methods, corresponding to different choices of $p(\cdot, \cdot)$, will briefly be discussed.

Metropolis-Hastings Algorithm: One generates a new θ^* based on a "probing" distribution, and then moves to the new θ^* or stays at the old θ according to certain "accept-reject" probabilities. See Hastings (1970).

Gibbs Sampling: The Markov chain moves from $\theta^{(i)}$ to $\theta^{(i+1)}$ one coordinate at a time (or one group of coordinates at a time), the transition density being the conditional posterior density of the coordinate(s) being moved given the other coordinates. This is a particularly attractive procedure in many Bayesian scenarios, such as analysis of hierarchical models, because the conditional posterior density of one parameter given the others is often relatively simple (or can be made so with the introduction of auxiliary variables). Extensive discussion and illustration of Gibbs sampling can be found in Gelfand and Smith (1990), Gelman and Rubin (1992), Raftery (1992), and Smith and Gelfand (1992).

Hit and Run Sampling: The idea here is roughly that one moves from $\theta^{(i)}$ to $\theta^{(i+1)}$ by choosing a random direction and then moving in that direction according to the appropriate conditional posterior distribution. This method is particularly useful when Θ is a sharply constrained parameter space. Extensive discussion and illustration can be found in Belisle, Romeijn, and Smith (1993), and Chen and Schmeiser (1993).

Hybrid Methods: Complex problems will typically require a mixture of the above (and other) methods. Here is an example, from Mueller (1991), the purpose of which is to do Gibbs sampling when the posterior conditionals (e.g., $\pi(\theta_i|x, \text{other } \theta_k)$) are not "nice":

Step 1. Each step of the Markov chain will either

- generate $\theta_j^{(i)}$ from $\pi(\theta_j|x, \text{other } \theta_k^{(i)})$ if the conditional posterior is "nice" or
- generate $\theta_j^{(i)}$ by employing one or several steps of the Metropolis-Hastings algorithm if the conditional is not nice.

Step 2. For the probing function in the Metropolis-Hastings algorithm, use the relevant conditional distribution from a global multivariate normal (or t) importance function, as typically developed in Monte Carlo importance sampling.

Step 3. Adaptively update the importance function periodically, using estimated posterior means and covariance matrices.

Other discussions or instances of use of hybrid methods include Geyer (1992), Gilks and Wild (1992), Tanner (1991), Smith and Roberts (1993), Berger and Chen (1993), and Tierney (1994).

3. FOUNDATIONAL ISSUES AND APPLICATION TO CLINICAL TRIALS

We illustrate a standard 'default' Bayesian analysis in the context of a particular clinical trial. For pedagogical reasons the example is artificial, but it is typical of many real examples. Discussion of the example, together with its foundational implications, will follow.

3.1 The Clinical Trial

Treatment 1, consisting of use of Drug A, is to be compared with Treatment 2, which utilizes both Drug A and Drug B. The design is a paired comparison design, resulting in observation of independent $X_i \sim \mathcal{N}(\theta, 1)$, $i = 1, 2, \ldots$, where θ is the mean difference in effect between Treatment 2 and Treatment 1. Typically, one tests $H_0: \theta = 0$ versus $H_a: \theta \neq 0$, although a more reasonable formulation might be to test $H_0: \theta = 0$ versus $H_1: \theta < 0$ versus $H_2: \theta > 0$.

A common default Bayesian analysis of this problem assigns H_0 and H_a equal prior probabilities of $1/2$, and chooses the conditional prior density on $\theta \neq 0$ to be a $\mathcal{N}(0,2)$ density (see below for discussion). Straightforward computations then yield, as the posterior probabilities of the hypotheses,

$$\begin{aligned}
\Pr(H_0 \text{ is true } | X_1, \ldots, X_n) &= B_n/(1+B_n), \\
\Pr(H_a \text{ is true } | X_1, \ldots, X_n) &= 1/(1+B_n), \\
\Pr(H_1 \text{ is true } | X_1, \ldots, X_n) &= \frac{\Phi(-\sqrt{n}\overline{X}_n/\sqrt{1+1/(2n)})}{1+B_n} \\
\Pr(H_2 \text{ is true } | X_1, \ldots, X_n) &= \frac{\Phi(\sqrt{n}\overline{X}_n/\sqrt{1+1/(2n)})}{1+B_n},
\end{aligned}$$

where $\overline{X}_n$ is the sample mean, Φ is the standard normal c.d.f., and

$$B_n = \sqrt{1+2n}\exp\{-\frac{1}{2}n\overline{X}_n^2/(1+\frac{1}{2n})\}.$$

Suppose the data, given in Table 1, arrives sequentially. Then, after each new data point, one can compute the above posterior probabilities; these are also given in Table 1.

Table 1. Data and Posterior Probabilities for the Clinical Trial

			Posterior Probabilities of			
Pair	X_i	$\overline{X}_n$	H_0	H_a	H_1	H_2
1	1.63	1.63	.417	.583	.054	.529
2	1.03	1.33	.352	.648	.030	.618
3	0.19	0.95	.453	.547	.035	.512
4	1.51	1.09	.266	.734	.015	.719
5	−0.21	0.83	.409	.591	.023	.568
6	0.95	0.85	.328	.672	.016	.657
7	0.64	0.82	.301	.699	.013	.686
8	1.22	0.87	.220	.780	.009	.771
9	0.60	0.84	.177	.823	.006	.817
10	1.54	0.91	.082	.918	.003	.915

If one were observing this data, and the associated posterior probabilities, two primary features would affect the decision as to whether or not to stop the experiment at a given observation. The first is the posterior probability of H_0, which decreases quite slowly. Indeed, even after ten observations, this probability is not small enough to conclusively

reject the hypothesis of no effect (i.e., that Drug B has no effect). The other significant feature of the analysis is that the posterior probability of H_1 is remarkably small throughout the analysis. This means that, even early in the experiment, it becomes clear that Treatment 1 is not superior to Treatment 2. But whether Treatment 2 is actually better, or just equivalent, remains uncertain. The time at which one might choose to stop the experiment would depend on the relative importance of these two factors (and also the magnitude of the effect, if present, which is determined from $\overline{X}_n$).

The above analysis does contain several somewhat arbitrary elements, which could be subject to criticism or alteration. One is the specification of a 'precise' null hypothesis. Rarely is a hypothesis of *exactly* zero treatment difference formally correct. However, it is quite plausible that the addition of Drug B has essentially no effect, and it can be shown that H_0 is then satisfactory as an approximation to this 'essentially no effect' hypothesis, unless the sample size is very large (cf, Berger and Delampady (1987)).

Other concerns about the above analysis center on the specifications of the prior distribution. Choosing the prior probabilities of H_0 and H_a to be 1/2 each is natural, but obviously could be changed if desired. Likewise choosing a prior distribution for θ, under H_a, to be $\mathcal{N}(0, 2)$ can be challenged. The centering of the prior at zero is rather innocuous, and would typically be done to provide an appearance of 'fairness'. The choice of a normal distribution is not essential, and is done here mainly for convenience in presenting the answers. (Indeed, Jeffreys (1961), who first carefully discussed this issue, actually preferred a Cauchy distribution to a normal.) The choice of a variance of 2 is the most arbitrary feature of the analysis. It was chosen, here, because use of the resulting prior is then similar in effect to use of the Cauchy(0,1) distribution that Jeffreys (1961) proposes (based on extensive arguments). The issue here is more one of standardization than of 'correctness.' Anyone desiring to use an actual subjective Bayesian analysis, with subjectively specified probabilities and distributions, is welcome to do so. But to provide a standard analysis for purposes of general communication, use of agreed-upon default choices has considerable appeal.

3.2 Standard Bayesian Motivations

There are several aspects of the Bayesian analysis of this clinical trial that are worthy

of note. The first is the ease of interpretation of the answers. One simply states the probability that each hypothesis is true. The classical alternatives, such as frequentist error probabilities or P-values, are not only much harder to interpret, but their interpretation is fraught with danger (cf, Edwards, Lindman, and Savage (1963), Berger and Sellke (1987), and Berger and Delampady (1987)).

The second appealing aspect of the Bayesian analysis here is the ease with which three hypotheses are handled. In particular, it is crucial in problems such as this to separate the evidence of no effect (i.e., the evidence for H_0) from the evidence for H_1 or H_2. In contrast, classical methods of dealing with these three hypotheses are awkward and cannot effectively capture the difference between rejection of H_0 and rejection of H_1. Of course, one might also want to produce other inferences, such as confidence (credible) sets for θ, conditional on H_a being true. This is trivial to do using Bayesian analysis.

The third appealing feature of Bayesian analysis, here, is the ability to ignore the 'stopping rule.' One can simply compute the Bayesian answers sequentially, as the data arrives, and stop the experiment whenever the evidence is sufficient. It is not necessary to formally specify a pre-experimental stopping rule, and the Bayesian answers do not depend on any stopping rule chosen. Classical analyses, in contrast, not only must pre-specify a stopping rule, but the answers depend dramatically on this rule, leading to complicated discussions such as how to 'spend α' for looks at the data. Furthermore, deviation from the pre-specified stopping rule can invalidate the analysis. Extensive discussion of Bayesian analysis and stopping rules can be found in Berger and Berry (1988).

3.3 Frequentist Motivation

It has long been known that most commonly used classical procedures have 'default' Bayesian interpretations (and indeed many were originally devised using Bayesian reasoning). Two exceptions are testing of a precise hypothesis and sequential analysis, where typical Bayesian answers often differ markedly from classical answers. This has been something of an embarrassment for the field of statistics, since we cannot present any professionally agreed-upon analyses for these situations.

Recent work in Berger, Brown, and Wolpert (1995) and Berger, Boukai, and Wang (1994) indicates that this disagreement is unnecessary: the Bayesian procedures for test-

ing precise hypotheses and sequential testing can often be shown to also be frequentist procedures. This is a startling development, and indicates that the controversies are not necessarily due to irreconcilable philosophical differences, but may simply be due to current use of the 'wrong' frequentist procedures.

The key to these new developments is adoption of the conditional frequentist paradigm, as formalized in Kiefer (1977). This paradigm is a generalization of the common notion of providing frequentist inferences conditional on an ancillary statistic. The idea is that one considers a partition of the sample space, and develops frequentist error probabilities conditional on elements of the partition. Such conditional inferences have all of the usual frequentist justifications, and are arguably superior to unconditional frequentist analysis in that they better discriminate between data of different evidentiary strength.

For the above clinical trial example, a partition of the sample space can be found such that the posterior probability of H_0 can also be interpreted as the probability of Type I error, conditional on the observed partition element. Furthermore, the posterior probability of H_a is an 'average' Type II error, conditional on the observed partition element; the 'average' is with respect to the posterior distribution of θ (on H_a), conditional on the observed partition element. (Reporting of this 'average' Type II error compares quite favorably with the common classical prescription of reporting Type II error rates at one or two subjectively specified values of θ.) For definition of the partition which achieves this duality of interpretation, see Berger, Boukai, and Wang (1994). Note that the partition can depend on the stopping rule used, but the reported Type I and Type II error probabilities do not depend on the stopping rule.

In conclusion, the Bayesian test, with its intuitively attractive operational properties, can be used with complete frequentist justification. This not only unifies the two paradigms in sequential experiments, but also yields considerably improved frequentist procedures.

4. ROBUST BAYESIAN ANALYSIS WITH APPLICATION TO QUANTIFYING OCKHAM'S RAZOR

There has recently been extensive development of Bayesian methodology for situations in which elements of the Bayesian model are uncertain (cf, Berger (1985, 1990, 1994),

Walley (1991), and Wasserman (1992)). Attention has primarily focused on uncertainty in the prior distribution, in part because non-Bayesians perceive this to be the main difficulty with Bayesian analysis, and in part because there are numerous situations in which one would like to draw a conclusion that is valid for a wide range of prior opinions. We illustrate this latter phenomenon by reviewing the robust Bayesian quantification of Ockham's razor, the heuristic scientific principle which states that, if the data is compatible with two models, then the simpler model is to be preferred.

The situation we consider here has data $X \sim \mathcal{N}(\theta, \sigma^2)$, σ^2 known. Two models are proposed for this data: M_1 asserts that $\theta = \theta_0$, a specified constant, while M_2 places no restriction on θ, but corresponds to prior beliefs about θ that (i) are symmetric about zero, and (ii) view values of θ nearer to zero as more plausible than values of θ farther from zero. This can be formalized by saying that, under M_2, the prior density of θ is a nonincreasing function of $|\theta|$.

Result 1. For the above situation and if M_1 and M_2 are each assigned prior probability $1/2$ of being correct, then the posterior probability of M_1 is *at least*

$$\underline{P} \equiv \left(1 + \sqrt{\frac{\pi}{2}}\left[d_1 + \sqrt{2\log(d_1 + 1.2)}\right]^{-1} \exp\{\frac{1}{2}d_0^2\}\right)^{-1},$$

providing $d_1 > 1.4$, where $d_0 = |X - \theta_0|/\sigma$ and $d_1 = |x|/\sigma$.

Idea of the Proof: This is a robust Bayesian result, found by computing an approximation to the lower bound of the posterior probability of M_1 over all prior densities for θ (under M_2) that are nonincreasing functions of $|\theta|$; see Berger and Jefferys (1992) and Jefferys and Berger (1992) for details. □

Example. One of the great scientific problems of the latter part of the Nineteenth century and early part of the Twentieth century was finding an explanation for an anomaly in the orbit of Mercury. After taking into account Newtonian theory, the available data exhibited an unexplained residual motion of Mercury's perihilion of 41.6 seconds of arc per century. Statistically, this observation of 41.6 can be considered to have been the realization of $X \sim \mathcal{N}(\theta, 4)$, where θ is the true unexplained residual motion.

Numerous theories were advanced to explain the anomaly, but by 1920 only two remained viable. One was a theory of Newcomb, proposing that gravity followed an inverse

$(2 + \varepsilon)$ law, rather than an inverse square law. It is crucial for our analysis to note that no value of ε was predicted apriori by Newcomb, although ε could, of course, be estimated using the Mercury data. The second theory was Einstein's general relativity, which (independent of the Mercury data) predicted that θ would be 42.9.

Clearly the data is compatible with Einstein's theory. It is also clearly compatible with Newcomb's theory, in that ε could be adjusted to exactly fit the data. Because of this adjustable parameter, however, Newcomb's theory is the more "complex" theory, so that Ockham's razor would suggest that Einstein's theory is to be preferred.

This situation fits the formalism that was described earlier, with $\sigma^2 = 4$, M_1 being Einstein's theory that specifies $\theta_0 = 42.9$, and M_2 being Newcomb's theory. Apriori, there is no reason to favor positive or negative values of ε, and values of ε that are closer to zero are more plausible, in that large $|\varepsilon|$ would have been more likely to have been previously detected. The same can be said about θ (which can be shown to be an approximately linear function of ε that passes through the origin), so that the prior assumptions about M_2, needed for application of Result 1, are satisfied.

Computation yields $d_0 = |41.6 - 42.9|/2 = 0.65$, $d_1 = |41.6|/2 = 20.8$, and $\underline{P} = 0.938$. The posterior probability of M_1 (the "simpler" Einstein model) is thus *at least* 0.938, which is a rather convincing lower bound. Robust Bayesian analysis has thus provided a meaningful quantification of Ockham's razor.

As a final remark, note that the simpler model must be compatible with the data for Ockham's razor to apply. Indeed, if M_1 is not compatible with the data, then d_0 will be large and, hence, $\underline{P}$ will be small.

Acknowledgment: This work was supported by the National Science Foundation, Grant DMS–9303556.

BIBLIOGRAPHY

Andrews, R., J. Berger, and M. Smith (1993), "Bayesian estimation of fuel economy potential due to technology improvements," in C. Gatsonis, J. Hodges, R. Kass, and N. Singpurwalla (editors), *Case Studies in Bayesian Statistics*, Springer-Verlag, New York, pp. 1–77.

Belisle, C., H. E. Romeijn and R. Smith (1993), "Hit-and-run algorithms for generating multivariate distributions," *Mathematics of Operation Research* **18**, 255–266.

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd edition), Springer-Verlag, N.Y.

Berger, J. (1990), "Robust Bayesian analysis: sensitivity to the prior," *J. Statist. Plann. Inf.* **25**, 303–328.

Berger, J. (1994), "An overview of robust Bayesian analysis," *Test* **3**, 5–124.

Berger, J. and J. Bernardo (1992), "On the development of the reference prior method," in J. Bernardo, J. Berger, A. Dawid and A. F. M. Smith (editors), *Bayesian Statistics 4*, Oxford University Press, London.

Berger, J. and D. Berry (1988), "The relevance of stopping rules in statistical inference," in S. Gupta and J. Berger (editors), *Statistical Decision Theory and Related Topics IV*, Springer-Verlag, New York, pp. 29–72.

Berger, J., B. Boukai, and Y. Wang (1994), "The conditional frequentist interpretation of Bayesian testing for a composite alternative hypothesis," Technical Report #94-25C, Department of Statistics, Purdue University.

Berger, J., L. D. Brown and R. L. Wolpert (1994), "A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing," To appear in *Annals of Statistics.*

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

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

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

Berger, J. and T. Sellke (1987), "Testing a point null hypothesis: the irreconcilability of P values and evidence," *J. Amer. Statist. Assoc.* **82**, 112–122.

Bernardo, J., J. Berger, A. F. Dawid, and A. F. M. Smith (1995), editors of *Bayesian Statistics 5*, Oxford University Press, London.

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

Chen, M. H. and B. Schmeiser (1993), "Performance of the Gibbs, hit-and-run, and Metropolis samplers," *Journal of Computational and Graphical Statistics* **2**, 1–22.

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

Fearn, T. and A. O'Hagan (1993, 1994), editors of "Special Issues on Practical Bayesian Statistics," *The Statistician* **42**, Numbers 4 and 5, and **43**, Number 1.

Gatsonis, C., J. Hodges, R. Kass, and N. Singpurwalla (1993), editors of *Case Studies in Bayesian Statistics*, Springer-Verlag, New York.

Gelfand, A. E. and A. F. M. Smith (1990), "Sampling based approaches to calculating marginal densities," *J. Amer. Statist. Assoc.* **85**, 398–409.

Gelman, A. and D. B. Rubin (1992), "On the routine use of Markov Chains for simulation," in J. Bernardo, J. Berger, A. Dawid, and A. F. M. Smith (editors), *Bayesian Statistics 4*, Oxford University Press, London.

Gilks, W. R. and P. Wild (1992), "Adaptive rejection sampling for Gibbs sampling," in J. Bernardo, J. Berger, A. Dawid, and A. F. M. Smith (editors), *Bayesian Statistics 4*, Oxford University Press, London.

Geyer, C. (1992), "Practical Markov chain Monte Carlo," *Statistical Science* **7**, 473–483.

Hastings, W. K. (1970), "Monte-Carlo sampling methods using Markov chains and their applications," *Biometrika* **57**, 97–109.

Jeffreys, H. (1961), *Theory of Probability* (3rd edition), Oxford University Press, London.

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

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

Mueller, P. (1991), "A generic approach to posterior integration and Gibbs sampling," Technical Report 91-09, Department of Statistics, Purdue University.

O'Hagan, A. (1994), *Bayesian Inference*, Edward Arnold, London.

Raftery, A. (1992), "How many iterations in the Gibbs sampler?" in J. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith (editors), *Bayesian Statistics 4*, Oxford University Press, London.

Smith, A. F. M. and A. E. Gelfand (1992), "Bayesian statistics without tears: a sampling-resampling perspective," *American Statistician* **46**, 84–88.

Smith, A. F. M. and G. O. Roberts (1993), "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods," *J. Roy. Statist. Soc. B* **55**, 3–23.

Tanner, M. A. (1991), *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*, Lecture Notes in Statistics **67**, Springer Verlag, New York.

Tierney, L. (1994), "Markov chains for exploring posterior distributions," *Ann. Statist.* **22**.

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

Wasserman, L. (1992), "Recent methodological advances in robust Bayesian inference," in J. M. Bernardo, et. al. (editors), *Bayesian Statistics 4*, 483–502.