Bayesian Robustness
IMS Lecture Notes - Monograph Series (1996) Volume 29

A PRACTICAL, ROBUST METHOD FOR BAYESIAN MODEL SELECTION: A CASE STUDY IN THE ANALYSIS OF CLINICAL TRIALS

JOEL GREENHOUSE AND LARRY WASSERMAN
Carnegie Mellon University

We present a method for model selection based on a proper reference prior. The choice of prior is somewhat arbitrary so Bayesian sensitivity analysis plays an important role in the analysis. We illustrate the methods in the context of a case study. We consider survival times (e.g., time to recurrence of depression) from a clinical trial. Because of the nature of the application we consider a mixture model that allows for a "surviving fraction." A Bayesian treatment of this model has been considered previously by Chen, Hill, Greenhouse and Fayos (1985), Greenhouse and Paul (1995) and Stangl (1991). In this paper, we are concerned with the question: does treatment effect both the probability of being a survivor and the survival times of "non-survivors"? The question is cast as a model selection problem. Reference priors give rise to improper posteriors and, moreover, do not lead to well defined Bayes factors. We adapt the idea of Kass and Wasserman (1995) who proposed "unit information priors." These priors are somewhat ad-hoc. To address this concern, we perform a sensitivity analysis with respect to the priors. We also consider case influence. Our conclusion is that treatment is important for determining long term survival but, among short term survivors, treatment may be less predictive of survival time.

1. Introduction: The Scientific Problem and Previous Analyses. This paper is about model selection for clinical trials data. We present a modest case study to illustrate a general strategy for Bayesian model selection. We suggest a simple method for constructing proper reference priors. The argument for this prior might be considered tenuous but we address the arbitrariness of the prior by performing sensitivity analysis. The calculations are performed using a combination of asymptotic approximations and Markov chain Monte Carlo. We will analyze survival data from a randomized controlled clinical trial but the methods we present are applicable to many model selection problems. There is some debate about whether model selection is appropriate in Bayesian inference. Some authors have argued that many model selection problems should be treated as estimation problems. There is much virtue in these arguments but we do not wish to enter into this debate here. We shall begin by assuming that the model selection is appropriate for our problem. The current problem provides an interesting case study and is a chance to explore the methodology we are proposing.

1.1 Background on the Clinical Trial. It is now widely recognized that the majority of patients who have had an episode of major depression will more than likely suffer a recurrence of their illness (Kupfer et al., 1985). Therefore, a major concern in the treatment of depression has focused on therapeutic interventions for the prevention or delay of the occurrence of subsequent episodes. These interventions have focused primarily on treating non-symptomatic patients with maintenance doses of pharmacotherapies, such as imipramine, that have been shown to be effective in the treatment of acute episodes of depression. Clinical trials for the assessment of the efficacy of such interventions are called maintenance therapy clinical trials. In the late 1970's the National Institute of Mental Health (NIMH) sponsored one of the most important trials of this type (Prien et al. 1984). An objective of this paper is to re-analyze the results of this trial using modern Bayesian methods.

The design of the NIMH study is as follows. Patients in an acute episode of depression who had experienced at least one previous episode of depression in the previous 2 1/2 years were eligible to participate in the maintenance trial if (i) they responded to imipramine for treatment of the acute illness, and ii) once stabilized remained symptom free for a period of eight consecutive weeks. Eligible patients were then randomly assigned either to receive maintenance doses of imipramine or placebo. During the maintenance phase, patients were followed for two years, or until they had a recurrence of depression. There were 78 patients randomized to the imipramine group, and 69 patients in the placebo group. The objective of the NIMH study was to determine whether imipramine at maintenance levels prevents or delays a recurrence of depression. For more details see Greenhouse, Stangl, Kupfer and Prien (1991).

Figure 1 presents the Kaplan-Meier survival curves for the time to recurrence of depression for patients in each treatment group. A comparison of the survival curves suggests that patients assigned to the imipramine group had fewer recurrences than patients who received placebo ("off-imipramine"). Prien et al. (1984) using the Mantel-Cox test for the equality of the two survival distributions found a highly statistically significant difference ($p < 0.001$). An interesting feature of these curves is that after a period of time a fraction of patients have a much lower if not negligible risk of a recurrence of depression. This feature of the survival distribution characterized by the survival curve "flattening-out" at a non-zero value often occurs in practice. We note that approximately 58% of the patients in the imipramine group survive the full two years of the maintenance phase without experiencing a recurrence of depression compared to about 30% of the patients in the placebo group.

A class of biologically plausible models for such survival distributions

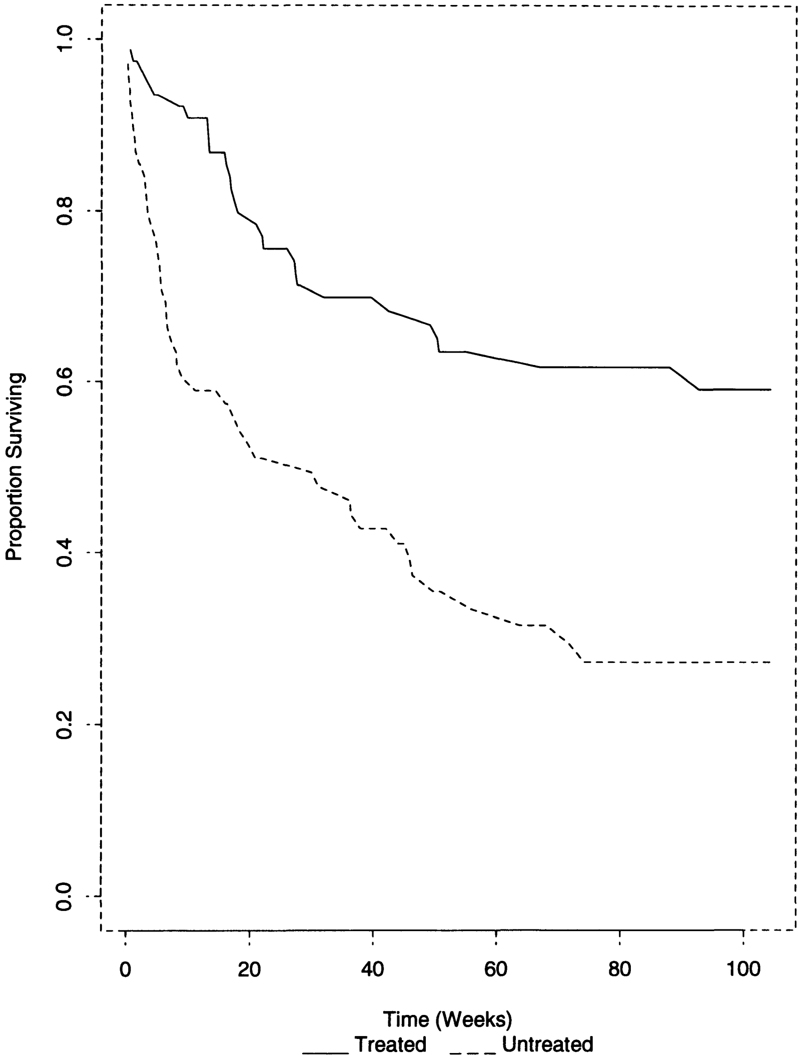


Figure 1: Kaplan-Meier Curves.

consists of a mixture of two populations. This model assumes that a fraction of the patients will experience the event of interest, e.g., a recurrence of depression, and the remaining patients will survive for a very long time without experiencing the event. In some applications this latter group, the surviving fraction, are considered to be "cured" (Boag 1949; Berkson and Gage 1952). Our goal in this paper is to use the mixture survival model to investigate the efficacy of maintenance doses of imipramine to prevent or delay the recurrence of depression. Our methodological goal is to illustrate a general strategy for Bayesian model selection that can be implemented in practice.

1.2 Overview. In section 2 we present the model, the prior, and we discuss how the computations are done. The computations are non-trivial and we discuss three methods for doing them. The results are presented at the end of section 2. In section 3 we carry out a sensitivity analysis. We find that the results in section 2 are quite robust. More importantly, we discover to which prior the posterior is most sensitive. We also consider case influence. Based on the results in section 3 we select one model. In section 4 we briefly discuss some inferences from the selected model. We make some closing remarks in section 5.

2. Model Specification, Computations, and Analysis.

2.1 The Model. We will speak of "survival times" throughout this paper. It should be understood that "survival" in this context means "non-recurrence of depression." Let T_i be the survival time of subject i and let C_i be the censoring time for subject i. Let $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ otherwise. We observe $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ where $Y_i = \min\{T_i, C_i\}$. We use a model for T_i with the following survival function:

$$S_{T_i}(t) = p_i + (1 - p_i)G(t|\theta_i) \tag{1}$$

where $p_i \in [0, 1]$ is the probability of being a survivor (i.e. of being "cured") and $G(t|\theta)$ is the survival distribution for the patients who are "not cured." For the application to the NIMH study we take $G(t|\theta) = exp(-\theta t)$, an exponential distribution with hazard rate θ. In the literature (1) is called a "surviving fraction model" or a "cure model" since a fraction of the population p do not have a recurrence. A number of authors motivate and discuss derivations for (1) (see for example, Farewell 1982; Greenhouse and Wolfe 1984; Chen, Hill, Greenhouse, Fayos 1985). Stangl (1991) was the first to use the mixture survival model in a Bayesian analysis of the results of the NIMH study to investigate heterogeneity of treatment effects across participating centers (see also Stangl and Greenhouse, 1995).

Let $X_i = 1$ if patient i was treated with imipramine and let $X_i = 0$ otherwise. We relate the covariate X to the parameters p and θ in (1) by

TABLE 1. *The Four Models.*

Model	γ_0	γ_1	β_0	β_1
M_1	●	●	●	●
M_2	●		●	●
M_3	●	●	●	
M_4	●		●	

assuming that

$$\log(p_i/(1-p_i)) = \gamma_0 + \gamma_1 X_i \tag{2}$$

$$\log(\theta_i) = \beta_0 + \beta_1 X_i. \tag{3}$$

The "surviving fraction" model is unrealistic since survivors are assumed to have an infinite survival time. We view the notion of an infinite survival time merely as an approximation for some longer survival time well beyond the two year limit of this study. In other words, we refer to a patient as being a "survivor" if the patient's probability of recurrence during the two year study period is negligible.

Our model selection problem is concerned with the specification and assessment of the treatment effect in model (1). There are four models of interest:

M_1: $(\gamma_0, \gamma_1, \beta_0, \beta_1) \in \mathbb{R}^4$,
M_2: $(\gamma_0, \beta_0, \beta_1) \in \mathbb{R}^3$ and $\gamma_1 = 0$,
M_3: $(\gamma_0, \gamma_1, \beta_0) \in \mathbb{R}^3$ and $\beta_1 = 0$ and
M_4: $(\gamma_0, \beta_0) \in \mathbb{R}^2$ and $\gamma_1 = \beta_1 = 0$.

Schematically, the models may be represented as in Table 1.

In M_1, treatment determines both whether someone is a survivor and also affects the hazard rate among non-survivors. In M_2, treatment only affects the hazard rate among non-survivors. In M_3, treatment only affects the probability of being a survivor. In M_4, treatment has no effect. To better understand the models, Figure 2 shows survival functions for treated and untreated patients for three hypothetical situations corresponding to Models 1, 2 and 3 respectively. We see that, for Model 1, the curves separate quickly and have different asymptotes. In Model 2, the curves separate quickly but have the same asymptotes. In Model 3, the curves separate slowly but have different asymptotes. In Model 4, not shown, the two curves would be identical.

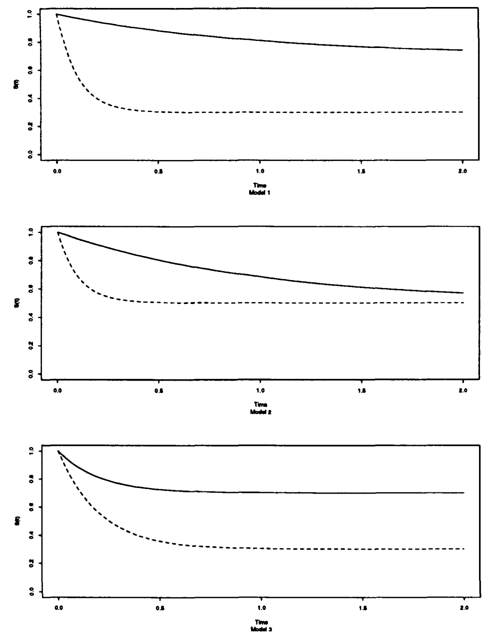


Figure 2: *Hypothetical survival curves for Models 1, 2 and 3. The solid line is the treated group; the dashed line is the untreated group.*

We will determine

$$Pr(M_i|data) = \frac{m_i}{\sum_j m_j}$$

where $m_j = \int L(\omega_j)\pi_j(\omega_j)d\omega_j$ and ω_j represents the vector of parameters for model M_j (see Jeffreys, 1961; Cornfield, 1966; Kass and Raftery, 1995) For another example of model selection in survival models, the reader is referred to Raftery, Madigan and Volinsky (1995).

2.2 The Prior. We are interested in using a reference prior for this problem such as $\pi(\beta_0, \beta_1, \gamma_0, \gamma_1) \propto 1$. However, it is well known that mixture models can lead to improper posteriors if improper priors are used. To see this, let $\gamma = (\gamma_0, \gamma_1)$, let $\beta = (\beta_0, \beta_1)$ and consider the latent variable $Z = (Z_1, \ldots, Z_n)$ where $Z_i = 1$ if subject i is a survivor and $Z_i = 0$ else. The posterior can be written as

$$p(\gamma, \beta|data) = \sum_Z p(\gamma, \beta|data, Z)p(Z|data) \tag{4}$$

the sum being over all 2^n configurations of Z. There is one term in the sum (4) corresponding to $Z = (0, \ldots, 0)$. For this term there is no information about γ in the data. Hence, the posterior for γ is equal to the prior and if the prior is improper, so is the posterior. Of course, since we are interested in hypothesis testing, we will need proper priors anyway. It is worth noting that intrinsic Bayes factors (Berger and Pericchi 1994) and fractional Bayes factors (O'Hagan 1995) are not useful here since the posterior, under the reference prior, is improper for any sample size. Instead we seek a proper reference prior.

For simplicity, we shall use independent, normal priors on all the parameters:

$$(5) \qquad \gamma_0 \sim N(\gamma_0^*, a_0^2),\ \gamma_1 \sim N(0, a_1^2),\ \beta_0 \sim N(\beta_0^*, b_0^2),\ \beta_1 \sim N(0, b_1^2).$$

A priori we assume no treatment effect. We need to choose γ_0^*, β_0^*, a_0, a_1, b_0 and b_1. We shall take γ_0^* and β_0^* equal to their maximum likelihood estimates under the full model. This makes the prior data dependent but since β_0 and γ_0 are common to all models under consideration, these are not the essential priors and we expect this choice to be of little consequence. (This is discussed more fully in Section 3.1). The more serious matter is choosing a_0, a_1, b_0 and b_1. Kass and Wasserman (1995) have suggested that in testing problems, it is reasonable to use "unit information priors." These are priors whose concentration is about the same as the concentration of the likelihood after one observation. Kass and Wasserman (1995) argue that using a prior whose information content is about that of one observation seems often to lead to reasonable inferences. A similar idea is used in Spiegelhalter and Smith (1982). A crude method for determining the variance of a unit information prior for a parameter ω is to compute the asymptotic standard error $s.e.(\hat{\omega})$ of the maximum likelihood estimate and define the standard deviation of the prior to be $s = \sqrt{n}[s.e.(\hat{\omega})]$. Again, this introduces some data dependence into the prior (except in the special case where the standard error does not involve an estimate of the parameter). But in well behaved problems, s will converge to a fixed constant almost surely so the data dependence vanishes asymptotically. To implement this idea here, we set $a_0 = \sqrt{n}[s.e.(\hat{\gamma}_0)]$, $a_1 = \sqrt{n}[s.e.(\hat{\gamma}_1)]$, $b_0 = \sqrt{n}[s.e.(\hat{\beta}_0)]$ and $b_1 = \sqrt{n}[s.e.(\hat{\beta}_1)]$ where the standard errors are based on the usual asymptotic approximations from the full model.

Clearly this choice of prior is open to many criticisms. The unit information idea seems intuitively reasonable but could certainly be questioned. Moreover, the notion of sample size is a foggy issue. We shall take n to be the number of cases but one could reasonably argue that censored observations contribute less than one unit of observation (see for example Raftery, Madigan, Volinsky 1995). All this suggests that some sensitivity analysis is in order. Given the complexity of the computations involved, we will need to stick to simple robustness calculations. In section 3 we use simple Bayesian robustness methods to at least partially address these concerns about the arbitrariness of the prior.

2.3 The Computations. We need to find $Pr(M_i|data)$ for $i = 1, 2, 3, 4$. To begin, for each of the four models, a sample from the posterior was obtained using a Markov chain Monte Carlo. We used a "Metropolis within Gibbs" scheme driven by a Gaussian random walk (Tierney 1995). We drew 10,000 samples from each of the four posteriors. Using a proper prior turns out to be very important in this problem. If we had used a flat prior then the posterior would have a flat spot far in the tail, corresponding to the improper component. Figures 3a and 3b show the joint likelihood function and log-likelihood function, respectively, for γ_1 and β_1 with γ_0 and β_0 held fixed at their maximum likelihood estimates. In Figure 3a, it appears that the likelihood function is very well-behaved. Yet in Figure 3b, the plot of the log-likelihood function shows the flat spot in the tail quite clearly. If the Markov chain is run for a short time there will be no problem. But if the chain is run a long time, then eventually it will visit the tail. When it does so, the chain behaves essentially like a random walk, moving erratically in the flat region with no hope of moving back to a region of high probability. Thus, the proper reference prior serves the dual purpose of providing a basis for testing and improving the performance of the Markov chain Monte Carlo. Presumably a very flat proper prior will also produce erratic chains since it will mimic the behavior of the posterior when the prior is flat. For this reason we recommend against proper but very diffuse priors in these models. We emphasize that this is not a problem with multimodality or slow convergence of the Markov chain. Paradoxically, long chains create more problems than short chains in this case.

For each model M_i, we need to calculate $m_i = \int L(\omega_i)\pi(\omega_i)d\omega_i$, $i = 1, 2, 3, 4$ where $L(\cdot)$ is the likelihood, π_i is the prior and ω_i is the set of parameters for model M_i. Unfortunately, the output from a posterior simulation does not provide a direct estimate of m_i. There is a quickly growing literature on computing the normalizing constant m_i from simulation; some recent papers include Carlin and Chib (1995), Chib (1994), DiCiccio, Kass, Raftery and Wasserman (1995), Gelfand and Dey (1994), Green (1995), Kass and Wasserman (1992), Lewis and Raftery (1994), Meng and Wong (1993), Raftery (1994) and Verdinelli and Wasserman (1995). Many of these methods require evaluating $L(\omega_i)$ for each sampled value ω_i, for each of the four models. But there are two methods which are quite quick and simple: the "simulation Laplace method" (Lewis and Raftery 1994, Kass and Wasserman 1992, DiCiccio, Kass, Raftery and Wasserman 1995) and the "generalized Savage-Dickey" method (Verdinelli and Wasserman 1995). We now briefly describe these two.

Recall that the Laplace estimate (Kass, Tierney and Kadane 1989) of an integral of the form $m = \int L(\omega)\pi(\omega)d\omega$ is given by $\hat{m} = L(\hat{\omega})\pi(\hat{\omega})(2\pi)^{d/2}|V|^{1/2}$ where $\hat{\omega}$ is the posterior mode, d is the dimension of ω and V is minus the

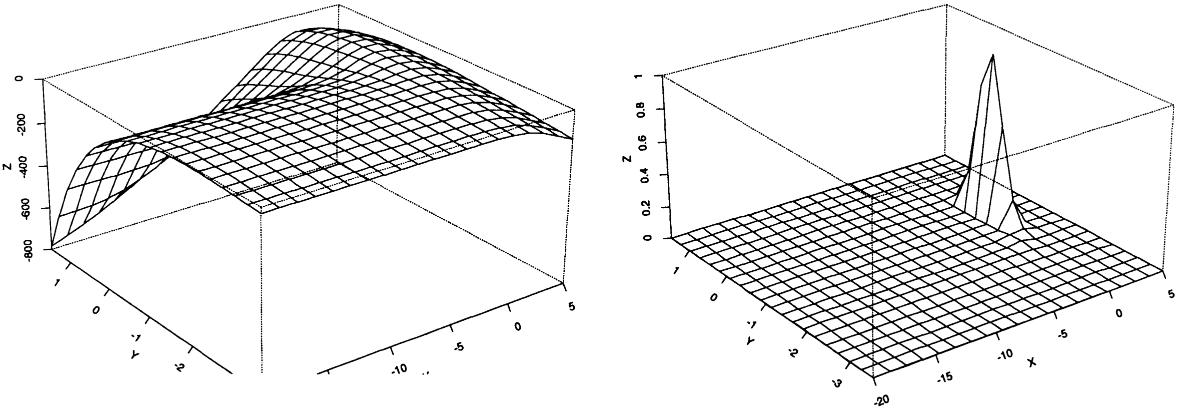


Figure 3: *Likelihood and Log-Likelihood for γ_1 and β_1 with $\gamma_0 = \hat{\gamma}_0$ and $\beta_0 = \hat{\beta}_0$.*

inverse of the Hessian. The estimate is accurate to order $O(n^{-1})$. The simulated version merely estimates $\hat{\omega}$ and V through the simulated values. For this we can resort to standard maximization routines. A cruder but simpler estimate is to substitute the posterior mean for $\hat{\omega}$ and the posterior covariance for V. This is appealing for its simplicity but our experience has been that this leads to inaccurate estimates. A compromise is to use a robust estimate of location and scale. Lewis and Raftery (1994) suggest the minimum volume ellipsoid estimator as implemented in Rousseuw and van Zomeren (1990). We have found this to work well and we shall adopt it here. We emphasize that a great virtue of the simulation Laplace method is that it requires only a single evaluation of $L(\cdot)$. This is a non-trivial consideration in even moderately complicated models.

Before describing the numerical results, we also mention the generalized Savage-Dickey method as developed in Verdinelli and Wasserman (1995). To describe the method, consider the full model M_1 and consider computing the Bayes factor

$$B = \frac{Pr(M_i|data)}{Pr(M_1|data)} \div \frac{Pr(M_i)}{Pr(M_1)} \tag{6}$$

for any submodel M_i, $i = 2, 3, 4$. We can recover the $m_i's$ given these Bayes factors. Let us write the parameter for the full model as $\omega = (\psi, \phi)$ where the submodel corresponds to $\psi = \psi_0$. Let $p_1(\psi|data)$ be the marginal posterior for ψ under M_1. Similarly, let $\pi_1(\psi, \phi)$ be the prior under M_1 and let $\pi_i(\phi)$

be the prior under M_i. Verdinelli and Wasserman show that

$$B = p_1(\psi_0|data)E_{\phi|\psi_0,data}\left(\frac{\pi_i(\phi)}{\pi_1(\psi_0,\phi)}\right). \tag{7}$$

The first quantity can be estimated directly from the simulated values from the full model using standard density estimation techniques. We generally use kernel density estimation based on a normal kernel using the bandwidth suggested in Silverman (1986, page 86). The second term can be estimated by simulating with ψ fixed at ψ_0. However, in the special case where $\pi_1(\phi|\psi_0) = \pi_i(\phi)$, which holds in our case, (7) reduces to Dickey's original formula given by

$$B = \frac{p_1(\psi_0|data)}{\pi_1(\psi_0)} \tag{8}$$

requiring only a simple density estimation. Remarkably, we can then recover all the $m_i^\prime s$ in this case using only a simulation from the full model so there is no need to simulate from M_2, M_3 and M_4. This advantage is balanced by the fact that the answers can be sensitive to the choice of bandwidth in the density estimator. Another advantage of this method is that it is "simulation exact", i.e., it converges almost surely to the true answer as the simulation size increases while the simulation Laplace estimator has an error of $O(n^{-1})$.

Finally, we mention the Schwarz approximation (Schwarz 1978) in which m_i is approximated by $n^{d_i/2}L(\hat{\omega}_i)$. Generally, this leads to only a O(1) approximation to Bayes factors and posterior probabilities of models. Kass and Wasserman (1995) show that under certain conditions, the Schwarz approximation is accurate to order $O(n^{-1/2})$. To get this accuracy one needs certain regularity conditions to hold. Moreover, one must use a unit information prior of a slightly different form than that used here. Nonetheless, we shall include the Schwarz calculations as well. Similar calculations are considered in Greenhouse and Paul (1995).

2.4 Results. Taking $Pr(M_1) = Pr(M_2) = Pr(M_3) = Pr(M_4) = 1/4$ we obtained estimates of $Pr(M_i|data)$ using the three methods. The results are in Table 2.

There is some discrepancy between the methods but the main conclusions from these calculations are consistent. Model 3 seems to be greatly preferred followed by Model 1 and then Model 2. The interpretation of Model 3 is that treatment effects the probability of being cured. It does not effect (or has little effect) on the hazard rate of those who are not cured. One model not mentioned here is the model in which $\gamma_0 = \gamma_1 = 0$. In such a model, there would be no surviving fraction. This model turned out to have essentially zero posterior probability and was not considered in further calculations.

TABLE 2. *Posterior Probabilities of the Models.*

Method	$Pr(M_1 \vert data)$	$Pr(M_2 \vert data)$	$Pr(M_3 \vert data)$	$Pr(M_4 \vert data)$
Simulation Laplace	.17	.10	.73	.01
Savage-Dickey	.19	.13	.68	.00
Schwarz	.17	.06	.76	.01

The question we now turn to is: how sensitive is the conclusion that Model 3 is superior to the choice of prior?

3. Sensitivity Analysis.

3.1. Sensitivity to the prior variances. As we mentioned in Section 2, we believe the priors we have chosen are a good starting point but we are concerned about the arbitrariness in our choices. Our main concern is the choice of a_0, a_1, b_0, b_1. To address this concern we will perturb each variance by a factor c where $1/10 \leq c \leq 10$ and we re-compute the posterior probabilities. Let $\tilde{\pi}$ be the new prior under consideration and let $\tilde{m} = \int L(\omega)\tilde{\pi}(\omega)d\omega$ be the new value of m. One way to compute $\tilde{m}$ is to re-weight the output of the Markov chain. A simpler approach is to use the $O(n^{-1})$ approximation $\tilde{m} = \hat{m}\tilde{\pi}(\hat{\omega})/\pi(\hat{\omega})$. The results for perturbations of size 1/10 and 10 are in Table 3. Figure 4 shows $Pr(M_i|data)$ as a function of $\log c$ for each M_i.

The results confirm that the priors on γ_0 and β_0 have very little effect. The prior on β_1 has a substantial effect but not in any way that affects our conclusions. The prior on γ_1 on the other hand, has a more complicated effect. When c gets very small or very large, $Pr(M_3|data)$ gets smaller and $Pr(M_2|data)$ gets larger. Eventually, these quantities cross and M_2 becomes the favored model. We do not yet have an intuitive explanation for why this happens when perturbing $\pi(\gamma_1)$ but not when perturbing $\pi(\beta_1)$. For a large range of values of c, Model 3 continues to dominate Model 2.

What we have learned from this sensitivity analysis is (i) M_3 appears to be the favored model and (ii) the crucial prior is the prior on γ_1. If further work is to be invested in the construction of priors, it should focus first on $p(\gamma_1)$. Incidentally, we believe that this illustrates a general point. Even die-hard subjectivists who dislike the idea of using reference priors (even proper reference priors are offensive to some) can still find it useful to begin an analysis with the reference priors. The reference prior analysis together with simple sensitivity tools leads to an understanding of which priors are

TABLE 3. *Sensitivity to Perturbations in Prior Variances.*

Perturbation	$Pr(M_1\|data)$	$Pr(M_2\|data)$	$Pr(M_3\|data)$	$Pr(M_4\|data)$
none	.17	.10	.73	.01
0.1 a_0	.14	.09	.76	.00
10 a_0	.13	.10	.77	.01
0.1 b_0	.21	.16	.63	.01
10 b_0	.13	.10	.76	.01
0.1 a_1	.26	.41	.31	.03
10 a_1	.07	.50	.40	.04
0.1 b_1	.42	.00	.57	.01
10 b_1	.02	.01	.96	.01

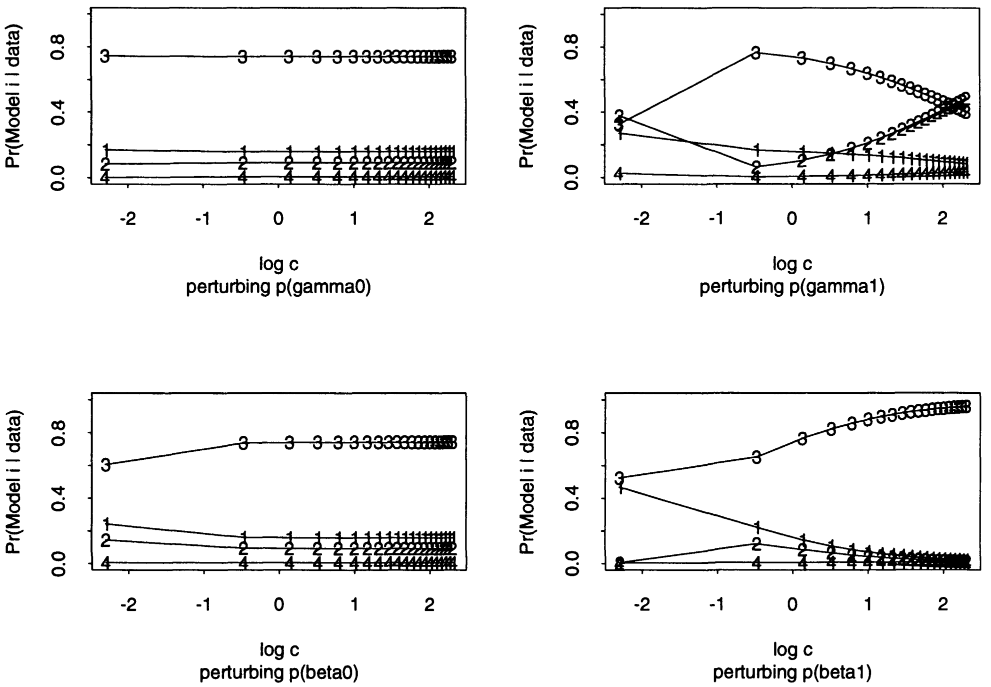


Figure 4: *Sensitivity plots: Each plot shows that posterior probabilities of the various models as a function of* $\log c$.

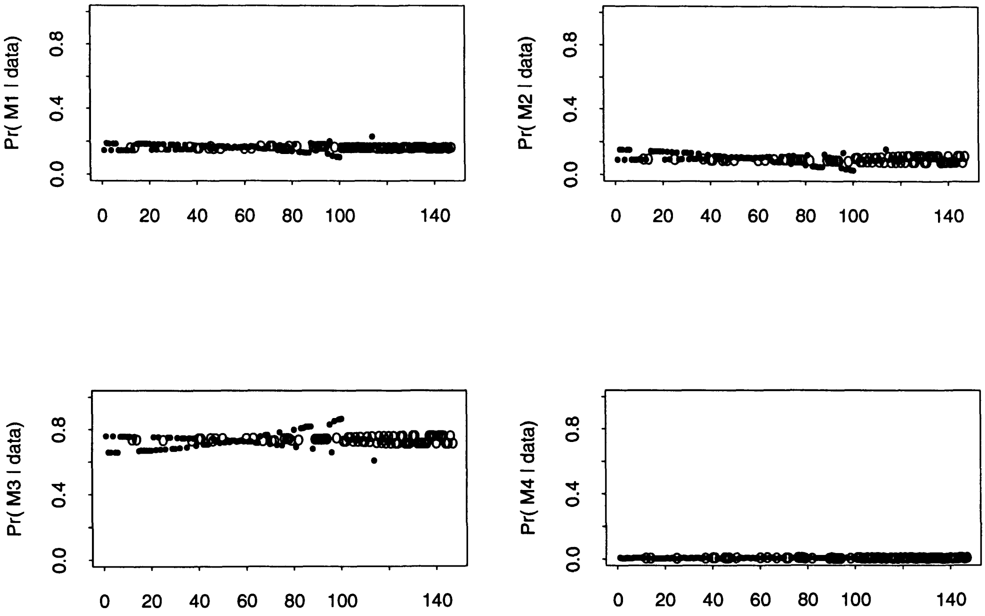


Figure 5: *Influence Diagnostics*

important in the problem.

3.2. Case Influence. Consider the effect of dropping the j^{th} observation from the analysis. It is possible to re-weight the samples from the posterior to re-compute the posterior probabilities. Weiss (1993) discusses this idea in detail. Again, we find it much easier to use the approximation $m_{(j)} = \hat{m}/L_j(\hat{\omega})$ where the subscript (j) refers to removing the j^{th} observation and L_j is the likelihood based on the j^{th} observation. Figure 5 shows $Pr_{(j)}(M_i|data)$ where the observations have been ordered from smallest to largest. Circles indicate censored observations. We see that no single observation significantly affects the posterior probabilities. One can also examine deletions of 2 or more observations. We do not pursue this here.

4. Inference for the Selected Model, M_3.

Our analysis so far suggests that Model 3 is the "best" model. Here we consider some analysis based on this model. First we report the posterior means and 95 per cent credible intervals for the three parameters:

$E(\gamma_0|data) = -1.15$; 95 per cent interval: (-1.87 , -.53)
$E(\beta_0|data) = -3.22$; 95 per cent interval: (-3.57,-2.95)
$E(\gamma_1|data) =$ 1.52; 95 per cent interval: (.72 , 2.39)

Now let $\gamma_t^0 = Pr(T > t|\text{untreated}) = S(t|X = 0)$, $\gamma_t^1 = Pr(T > t|\text{treated}) = S(t|X = 1)$. Note that $S(\infty) = p$, the probability of being cured. Figure 6 shows the posteriors γ_t^i for 3 months, 6 months, 9 months and 12 months for both untreated and treated individuals. The treatment appears to have a substantial (in the clinical sense) effect. At early times, the effect is not so great but becomes more noticeable for longer times. Treatment significantly increases the probability of being "cured." Patients who receive imipramine are approximately 4.5 times as likely to not have a recurrence than patients who receive placebo. Figure 7 shows the estimated survival curves for the two groups with 95 per cent intervals. The curves are similar to the curves in Figure 1 suggesting, at least informally, that there is not a lack of fit. Again, we see that the effect of treatment seems to be mainly in long term rather than short time survival.

5. Conclusions.

We have outlined a strategy for model selection that consists of two steps. First, construct "unit information priors" and compute $Pr(M_i|data)$ for each model M_i. Second, perform a sensitivity analysis to see how the conclusions depend on the prior. If the conclusions are robust to the choice of prior then the unit information prior suffices; otherwise the more effort needs to be put into prior construction or, perhaps, it should simply be reported that the data do not support strong conclusions.

In our case study, one model seemed to stand out. According to this model, treatment effects whether a patient is a long term survivor but not the survival time of short term survivors. We also found that the Schwarz approximation was reasonably accurate despite the fact that the conditions for $O(n^{-1/2})$ accuracy outlined in Kass and Wasserman (1995) are not met. This is consistent with our experience with the Schwarz approximation in other cases; however, Berger and Pericchi (1995) report less success with this approximation.

Our analysis leaves many open questions some of which we outline here.

1. It may make better sense to use correlated priors on the parameters.

2. There are other relevant covariates that should be included in the analysis. For example, see Greenhouse and Paul (1995) who incorporate additional covariates in the exponential mixture survival model using the regression models in (2) and (3) and use a model selection criteria

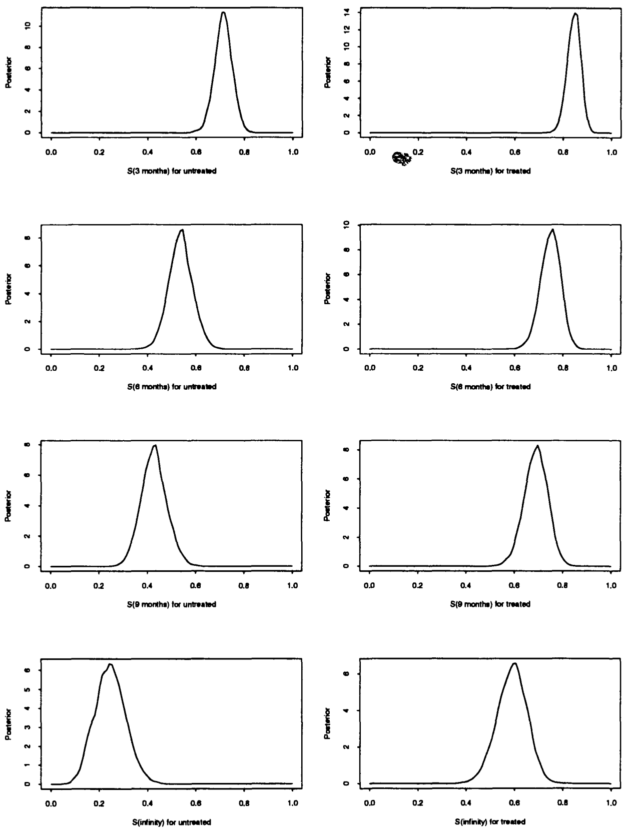


Figure 6: *Posterior probability of $S(t)$ for various t under Model 3. The left column is for untreated patients; the right column is for treated patients.*

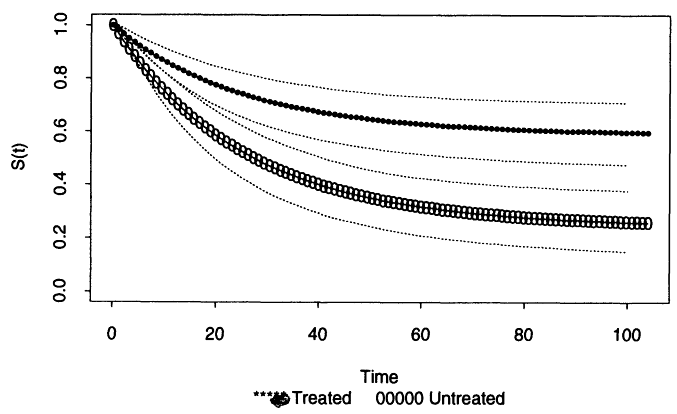


Figure 7: *Fitted survival curves under Model 3.*

based on the Schwarz approximation.

3. The sensitivity analysis should be expanded to include joint perturbations of the priors and to include nonparametric perturbations.

4. The mixture model should be expanded to include two exponential components rather than one exponential component plus a surviving fraction. (However, the parameters of the second component might be close to non-identifiable.)

5. The appropriateness of the exponential assumption deserves investigation.

Acknowledgements This work is supported in part by grants from the National Cancer Institute, CA54852 (JG,LW), the National Institute of Mental Health, MHCRC30915 (JG), and the National Science Foundation DMS-9303557 (LW) and DMS-9357646 (LW).

REFERENCES

Berger, J. and Pericchi, L. (1994). The intrinsic Bayes factor for model selection and prediction. Technical report, Department of Statistics, Purdue University.

Berger, J. and Pericchi, L. (1995). The intrinsic Bayes factor for linear models. To appear: *Bayesian Statistics 5.*

Berkson, J. and Gage, R. P. (1952). Survival curves for cancer patients following treatment, *Journal of the American Statistical Society*, **47**, 501-515.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society, Series B*, **11**, 15-44.

CARLIN, B. AND CHIB, S. (1995). Bayesian model choice via Markov chain Monte Carlo. To appear: *J. Roy. Statist. Soc. B.*

CHEN, W. C., HILL, B. M., GREENHOUSE, J. B., AND FAYOS, J. V. (1985). Bayesian Analysis of Survival Curves for Cancer Patients Following Treatment, *Bayesian Statistics 2: Proceedings of the 2nd Valencia International Meeting*, eds. J. M. Bernardo et al., North-Holland.

CHIB, S. (1994). Marginal likelihood from the Gibbs output. Unpublished manuscript, Olin School of Business, Washington University.

CORNFIELD, J. (1966). A Bayesian analysis of some classical hypotheses- with applications to sequential clinical trials. *J. Amer. Statist. Assoc.*

DICICCIO, T., KASS, R.E., RAFTERY, A. AND WASSERMAN, L. (1995). Computing Bayes factors by combining simulation and asymptotic approximations.

FAREWELL, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics*, **38**, 1041-1046.

GELFAND, A.E. AND DEY, D.K. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. B.*, **56**, 501-514.

GREEN, P. (1995). Reversible jump MCMC computation and Bayesian model determination. Technical report, Department of Mathematics, University of Bristol.

GREENHOUSE, J. B. AND PAUL, N. (1995). Applications of a mixture survival model with covariates to the analysis of a depression prevention trial. Submitted *Statistics in Medicine.*

GREENHOUSE, J.B., STANGL, D., KUPFER, D. J., AND PRIEN, R. F. (1991). Methodologic issues in maintenance therapy clinical trials, *Archives of General Psychiatry*, **48**, 313-318.

GREENHOUSE, J. B. AND WOLFE, R. (1984). A Competing Risk Derivation of a Mixture Model for the Analysis of Survival Data, *Communications in Statistics: Theory and Methods*, **13**, 3133-3154.

JEFFREYS, H. (1961). *Theory of Probability, 3rd ed.*, Oxford: Oxford University Press.

KASS, R.E. AND RAFTERY, A.E. (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association*, in press.

KASS, R.E. AND WASSERMAN, L. (1992). Improving the Laplace approximation using posterior simulation. Technical report #566, Department of Statistics, Carnegie Mellon University.

KASS, R.E. AND WASSERMAN, L. (1995). A reference Bayesian test and its relationship to the Schwarz criterion. To appear: *J. Amer. Statist. Assoc.*

KASS, R.E., TIERNEY, L. AND KADANE, J.B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika*, **76**, 663-674.

KUPFER, D.J, BERGER, P.A., CONGER J.J. ET AL. (1985). NIMH/NIH consensus development conference statement. Mood disorders: Pharmacologic prevention of recurrences. *American Journal of Psychiatry*, 142:469-476.

LEWIS, S. AND RAFTERY, A. (1994). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. Technical report 279, Department of Statistics, University of Washington.

MENG, X.L. AND WONG, W.H. (1993). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Technical report 365, Department of Statistics, University of Chicago.

NEWTON, M.A. AND RAFTERY, A.E. (1991). Approximate Bayesian Inference by the Weighted Likelihood Bootstrap. Technical Report 199, Department of Statistics, University of Washington.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. R. Statist. Soc. B.* **57**, 99-138.

POCOCK, S. J., GORE, S. M. AND KERR, G. R. (1982). "Long term survival analysis: the curability of breast cancer", *Statistics in Medicine*, **1**, 93-106.

PRIEN, R. F., KUPFER, D. J., MANSKY, P. A., SMALL, J. G., TUASON, V. B., VOSS, C. B. AND JOHNSON, W. E. (1984). Drug therapy in the prevention of recurrences in unipolar and bipolar affective disorders: Report of the NIMH collaborative study group comparing lithium carbonate, imipramine, and a lithium carbonate-imipramine combination, *Archives of General Psychiatry*, **41**, 1096-1104.

RAFTERY, A. (1994). Hypothesis testing and model selection via posterior simulation. In *Practical Markov Chain Monte Carlo.* (W. Gilks, S. Richardson, D.J. Spiegelhalter), London: Chapman and Hall.

RAFTERY, A.D., MADIGAN, D. AND VOLINSKY, C.T. (1995). Accounting for model uncertainty in survival analysis improves predictive performance. In *Bayesian Statistics 5* (J.M. Bernardo et al. eds), to appear.

ROUSSEEUW, P.J. AND VAN ZOMEREN, B.C. (1990). Unmasking multivariate outliers and leverage points (with discussion). *J. Amer. Statist. Assoc.* **85**, 633-651.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

SPIEGELHALTER, D.J. AND SMITH, A.F.M. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. R. Statist. Soc. B.* **44**, 377-387.

STANGL, D. (1991). Modeling heterogeneity in multi-center clinical trials using Bayesian hierarchical survival models. Unpublished doctoral dissertation. Department of Statistics, Carnegie Mellon University.

STANGL, D. AND GREENHOUSE, J. (1995). Assessing placebo response using Bayesian hierarchical survival models. Submitted *J. Amer. Statist. Assoc.*

TIERNEY, L. (1995). Markov chains for exploring posterior distributions. To appear: *Ann. Statist.*

TIERNEY, L., KASS, R. AND KADANE, J. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.* **84**, 710-716.

VERDINELLI, I. AND WASSERMAN, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. To appear: *J. Amer. Statist. Assoc.*

WEISS, R. (1993). Bayesian sensitivity analysis using divergence measures. Technical report, Department of Biostatistics, UCLA School of Public Health.

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PA 15213

A Practical, Robust Method for Bayesian Model Selection: A Case Study in the Analysis of Clinical Trials

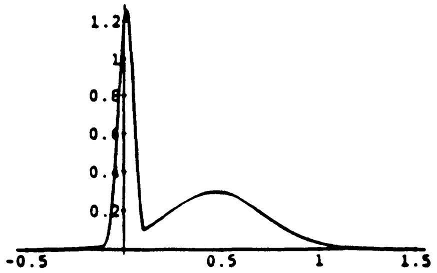
discussion by
M.J. BAYARRI
Universitat de València

This is a very nice paper indeed. It does carry a *true* sensitivity analysis in a *true* real problem. The scenario is that of a meaningful (as opposite to artificial) model selection, and the authors have addressed the touchy issue of assessing a"default" proper prior as well as imaginatively solved the formidable computation task in a problem where naïve Gibbs sampling would fail. All this makes the discussion a very pleasant but, alas, very difficult task. My discussion focused on a couple of robustness issues that the authors did not address, namely robustness with respect to the form of analysis (or the old issue of estimation versus testing) and with respect to the form of the survival function for survivors (of which the constant survival term in the mixture (1) is only an approximation); it also pointed to some few facts in the numerical example that could potentially look somehow odd. Due to severe space limitations, this written version will entirely skip the robustness with respect to the model issue an d will only sketch the rest of the discussion, with basically no derivations nor numerical details.

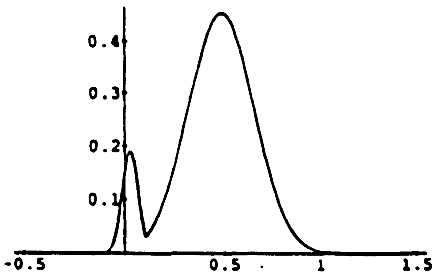
Posing the problem of model selection as one of testing implies a prior that is highly spiked around the point null. (We understand testing of point nulls as approximations to testing small intervals around the point null, as in Berger and Delampady, 1987.) The problem is that spiked priors are *very* stubborn (and stubborn priors are not robust): it usually takes a large n for the likelihood to "swallow" the prior. But if n is very large, then the approximation of a "sharp" null by a point null might not be appropriate (Berger and Delampady, 1987).

We demonstrate our claims in a very simple example. Assume X_1, $X_2, 0D \ldots, X_n$ are i.i.d. $N(\theta, 1)$ and we wish to test $H_0 : |\theta| 0D \leq 0.1$ versus $H_1 : |\theta| > 0.1$. Routine Bayesian testing assumes that the prior is highly spiked on H_0 so that, under some conditions, the problem can be approximated by that of testing $H_0^* : \theta 3D0$ versus $H_1^* : \theta \neq 0$, with a prior that has a point mass at H_0^*. Usually $Pr(H_0)3DPr(H_0^*)3D1/2$. Figure B1 shows the lingering effect of the prior on the posterior as well as the error incurred when approximating Bayes factors corresponding to the spiked, continuous prior by that resulting from the usual point mass at 0. We use a prior which is proportional to a truncated $N(0, 1/25)$ on H_0 and a truncated $N(0, 1)$ on H_1 (so that $Pr(H_0)3D1/2$ and the resulting density is a continuous function). The computations are done for $\bar{x}3D0.5$; the results are even more dramatic for $\bar{x}3D0.2$, where the effect of the prior is larger than that of the likelihood even for $n3D150$, but the error committed is already of 61%; in fact, the likelihood does not begin to "wash out" the prior till n does not get as large as 300, and by then the approximation by a point null clearly fails, with an error on the Bayes factor of 376%.

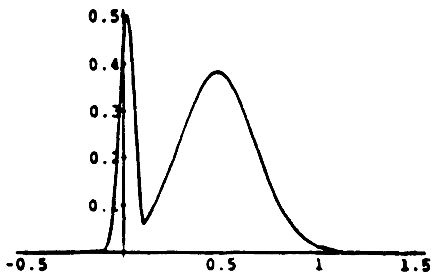
n=15, error=0.7%



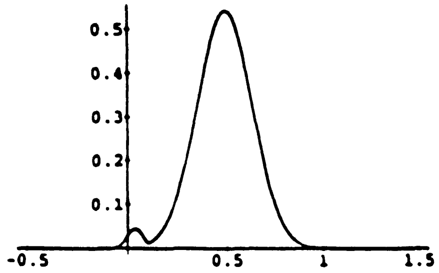
n=35, error=12%



n=25, error=3.4%



n=50, error=35%



The above example aims at showing that even fully recognized testing problems can be very sensitive to both, the "default" $Pr(H_0^*)3D1/2$ and the "default" approximation of an interval null by a point null. On the other hand, *Table B1* aims to showing how sensitive the results can be to whether we pose the problem as one of estimation or as one of testing, that is, to whether we use a smooth (non-informative) prior or a spiked one for solving the testing of an interval null. In that table we assume that in all of the entries $\bar{x}$ is such that the classical point-null testing would produce the (inappropriate) $\alpha 3D0D0.5$. We compute the "real" classical p-value α_ϵ, as well as 20 $Pr(H_0|data)$ for both, a flat prior (estimation prior) and the previous spiked prior. As it can be seen, for $n3D100$, one can reach any conclusion one wishes by merely approaching the problem as estimation or testing.

n	real p-value α_ϵ	diffuse prior $Pr(H_0\|data)$	testing prior $Pr(H_0\|data)$
15	.067	.048	.399
25	.079	.065	.456
50	.109	.101	.551
100	.170	.167	.665

Table B1

In the actual numerical results, we find somehow surprising the behavior of $Pr(Model\ i|data)$ as the prior variances (or c) changes, as reported in Figure 4 and Table 3. For instance, Model 3 corresponds to $\beta_1 3D0$. Hence, since a priori $\beta_1 \sim N(0, (cb_1)^2)$, the larger c the less sure we are a priori that β_1 is close to 0, and therefore, the less probability should be given (a posteriori) to Model 3; but it can be seen that $Pr(Model\ 3|data)$ is, quite surprisingly, *the only one* that increasees with c. Similar behaviour occurs when perturbing the variance of γ_1.

Last, I was curious to find out about the estimate of the mean life time for non-survivors. For the selected model, $T \sim Ex(t|\theta)$, and from the results for confidence intervals, it looks as if an estimate of $1/\theta$ is close to 24 months, which is the observation period. This might suggest that the inferences may be sensitive to the model and to even the precise observation period.

ADDITIONAL REFERENCE

Berger, J.O., and Delampady, M. (1987). Testing Precise Hypothesis, *Statistical Science* **2**, 317-352.

REJOINDER

Joel Greenhouse and Larry Wasserman

Although, we tried to avoid discussing the testing versus estimation controversy, Professor Bayarri has not allowed us to do so. As she notes, a test of a precise hypothesis may fail to approximate a test of an imprecise hypothesis when n is large. This is relevant in our paper since model selection is akin to testing whether regression coefficients are zero.

One can take the precise test at face value and not think of it as approximating an imprecise hypothesis. This was Jeffreys's approach which we find quite compelling. If we do regard the precise null as an approximation, we usually think of the imprecise null as being small *relative to sampling variation.* We can express this by a null of length ϵ_n where $\epsilon_n = o(1/\sqrt{n})$. In this case, the problems discussed by Professor Bayarri are obviated. On the other hand, we agree with her that if a fixed imprecise null is truly of interest, then great care is needed if one uses a precise null as an approximation.

Professor Bayarri expresses surprise with the behavior of the Bayes factors to changes in the prior variance in our sensitivity analysis. Intuition might fail here since the Bayes factor is not monotonic in prior variance. Moreover, changing the prior on β_1 affects the posterior probability of models 1 and 2. How this will ultimately affect the posterior probability of model 3, which depends on how well models 1 and 2 explain the data, is unclear.

Finally, Professor Bayarri asks about the magnitude of the estimate of $E(T|\theta)$. Censoring leads to increased estimates of mean survival time. Since we have a mixture model, the group membership of the censored observations (i.e. "true" survivors or censored non-survivors) is not known. Effectively, we average over all group memberships and some of these groupings lead to large values of mean survival.

We would like to express our appreciation to Professor Bayarri for her careful reading of our paper and to thank her for her thoughtful comments.