Implications of Reference Priors for Prior Information and for Sample Size

Bertrand CLARKE

Here we use posterior densities based on relative entropy reference priors for two purposes. The first purpose is to identify data implicit in the use of informative priors. We represent an informative prior as the posterior from an experiment with a known likelihood and a reference prior. Minimizing the relative entropy distance between this posterior and the informative prior over choices of data results in a data set that can be regarded as representative of the information in the informative prior. The second implication from reference priors is obtained by replacing the informative prior with a class of densities from which one might wish to make inferences. For each density in this class, one can obtain a data set that minimizes a relative entropy. The maximum of these sample sizes as the inferential density varies over its class can be used as a guess as to how much data is required for the desired inferences. We bound this sample size above and below by other techniques that permit it to be approximated.

KEY WORDS: Asymptotic normality; Experimental design; Information; Relative entropy.

1. INTRODUCTION

One way to evaluate a subjectively chosen prior is to assess the information that it contains. There are various ways to do this. One might search for an optimization procedure that resulted in that prior and then regard the prior as reflecting the information in the constraints following Soofi, Ebrahimi, and Habibullah (1995) (see also Soofi 1995). Alternatively, one might compare the subjectively chosen prior to other priors, including "objective" priors. This was proposed by Bernardo (1979) when he introduced reference priors. Despite the terminology, the goal of comparing subjective priors to reference priors so as to evaluate their information content can actually be done better by using posteriors, see Berger and Bernardo (1992). Here we compare a subjectively chosen density for a parameter to posteriors formed by updating a reference prior. In particular, we seek the data set that will make such a reference posterior as close as possible to the subjectively chosen density. The resulting data set can be regarded as representative of the information in the subjective density.

There are two reasons for identifying the posterior that is closest to an arbitrary density for the parameter. First, the arbitrary density might be an informative prior. In this case we identify the data implicit in the informative prior. If this data set is unreasonable, then the prior does not summarize the experimenter's beliefs well. Second, the arbitrary density might be a target density for inference. In this case we identify the data that would best permit the desired inferences. If the distance from the best approximation to the target density is large, then the inferences represented by the target density are unobtainable. Furthermore, we can consider a class of target densities. Because a sample can be associated to each member of the class, we can choose the largest of these sample sizes as a way to ensure that we will collect enough data to make inferences from a posterior likely to be close to one of the target densities. This can be regarded as a Bayesian analog of sample size calculations for confidence intervals of prescribed length done from a frequentist standpoint by guessing the standard deviation.

To fix ideas, suppose that we try to assess the information content of a subjective prior relative to iid data from a Poisson(θ) distribution. To make calculations easy, let the conjugate prior $\Gamma(1,1)$ play the role of the objective prior and choose another conjugate prior, say a $\Gamma(10,10)$, to be the subjectively chosen informative prior. If we form the posterior from n outcomes of a Poisson(θ) random variable and the $\Gamma(1,1)$ prior, it is seen to be a $\Gamma(\sum x_i + 1, n+1)$ distribution. To evaluate the information in the $\Gamma(10,10)$ prior, we might find the data set that would make the posterior as close as possible to the subjective prior. Thus we solve the pair of equations $10 = \sum x_i + 1$ and $10 = n + 1$ to see that $\sum x_i = 9$ and $n = 9$. The information content of the $\Gamma(10,10)$ prior relative to a Poisson(θ) density and a $\Gamma(1,1)$ prior is seen to be nine data points that sum to nine. This is not surprising given the shapes of the $\Gamma(1,1)$ and $\Gamma(10,10)$ densities; the former is an Exponential(1) density and the latter is peaked around 1 (see Bickel and Doksum 1977). We see that the data set is not unique, as any data set of size nine summing to nine from a Poisson(θ) will do because the sum is a sufficient statistic for θ.

It is instructive also to consider the reverse case. That is, suppose that we let the $\Gamma(10,10)$ prior play the role of the objective prior and we let the $\Gamma(1,1)$ prior be the subjective prior. Then we try to solve the pair of equations $1 = 10 + \sum x_i$ and $1 = n + 10$. But this gives $\sum x_i = -9$ and $n = -9$, neither of which is possible, because both n and $\sum x_i$ must be positive. This means that there is no data set from a Poisson(θ) random variable that can be used to update a $\Gamma(10,10)$ prior so as to make it closer to a $\Gamma(1,1)$ prior than it already is. In short, a $\Gamma(10,10)$ is already more informative than a $\Gamma(1,1)$, so we cannot make it less informative by adding information.

Bertrand S. Clarke is an Assistant Professor, Department of Statistics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2. This research was partially supported by Natural Science and Engineering Research Council Operating Grant 5-81506. The author would like to express his gratitude to two anonymous referees whose comments greatly improved this article, to A. Yuan for a careful reading of this manuscript, and to L. A. Wasserman for helpful discussions.

Journal of the American Statistical Association
March 1996, Vol. 91, No. 433, Theory and Methods

In this example we began by regarding the $\Gamma(1,1)$ as noninformative and the $\Gamma(10,10)$ as informative. However, there is no reason to regard the $\Gamma(1,1)$, or any member of the $\Gamma(p,\lambda)$ family, as noninformative. We have merely assessed the information in an informative density for the parameter relative to a posterior formed from another density for the parameter that happened to be less informative. To have a notion of the informativity of a subjective density for the parameter, we must use an objective density that is never more informative than any other density. That is, the objective prior must be noninformative in some meaningful sense, particularly because we expect a prior to contain information representable by a small sample size.

It is only for likelihoods in exponential form that the use of conjugate priors is possible and, even in this setting, noninformative priors need not be conjugate to the likelihood. Moreover, the subjective density that we assessed, the $\Gamma(10,10)$, need not be one that really interested us. We might be interested in a density that is not readily approximable by any member of the conjugate family. For instance, one might wish to evaluate the information in a uniform, or a density that had a polynomial tail. Such densities cannot be well approximated by an element of the $\Gamma(p,\lambda)$, because a polynomial tail reflects less information than an exponentially decreasing tail. Indeed, the tails of conjugate densities typically decrease exponentially, which is informative. Subjectively chosen priors usually correspond to less information, which is better represented by a polynomial tail.

From an experimental design perspective, we may want to identify the data that would best support inferences from a specific density for the parameter. Indeed, this can be done for classes of inferential densities, each member of which achieves a specified degree of concentration at a parameter value. Recall that posteriors are typically asymptotically normal independently of the likelihood and the prior and that the normal family is the easiest to use for experimental design purposes. Now, from an experimental design perspective, we virtually always want to know how much data must be collected when the subjective distribution varies over the normal family for a range of means and variances.

This complicates the problem of representing a subjective density as a posterior from a noninformative prior, because we can no longer assume that either density for the parameter is in the conjugacy class of the likelihood. Thus in general we cannot find a data set by solving equations and cannot obtain an exact match between the subjective and objective densities for the parameter. But we can still carry out this matching approximately by using techniques as described later.

The article is organized as follows. In Section 2 we give a general framework in which one can match subjective priors to data sets by minimizing the relative entropy distance between the subjective prior and a posterior formed from a noninformative prior. This minimization is over potential data sets. In Section 3 we verify that minimization is possible in general, then present computations to show how the minimum behaves for several choices of the subjective density in the context of a Poisson(θ) family. In Section 4 we turn to implications for experimental design. We perform the same minimization over a class of densities for the parameter and choose the largest of the sample sizes that result. After demonstrating existence of this maximal sample size for certain classes of inferential densities, we give bounds on it in terms of two other methods for selecting a sample size. In Section 5 we summarize our results and give some possible extensions of the line of inquiry pursued here.

2. STATEMENT OF THE PROBLEM

We begin by choosing an objective prior to represent the absence of information. Compatible with this choice, we must match the information in a subjective prior to a posterior formed from this noninformative prior so as to find a data set that makes them close. We begin by setting up the problem and then turn to formalities.

2.1 The Setting

The notion of noninformativity for a prior can be interpreted in diverse ways yielding diverse priors, each with some claim to noninformativity (see, e.g., Datta and Ghosh 1994a,b, Ghosh and Mukerjee 1992, and Kass and Wasserman 1993 for critical surveys). Here we choose the Jeffreys prior as noninformative, because it can be justified by four distinct arguments. The first is invariance; the Jeffreys prior was originally proposed on the basis of invariance (see Jeffreys 1961), even though this does not characterize the Jeffreys prior. George and McCulloch (1992) obtain the Jeffreys prior uniquely only by imposing an extra condition. The second argument in favor of the Jeffreys prior is that using it results in estimation which corresponds to frequentist coverage probabilities (see Hartigan 1983). The third argument in favor of the Jeffreys prior is that the formal sense in which it is noninformative can be interpreted in terms of the rate of information accumulation (see Bernardo 1979 and Clarke and Barron 1994). Finally, the fourth argument is that the Jeffreys prior performs well in a practical sense (see Berger and Bernardo 1989).

In the context of the third argument, Bernardo (1979) proposed reference priors obtained from a noninformativity principle based on maximizing the Shannon mutual information between the parameter and the sample over possible choices of the density for the parameter. This can be expressed as the expected relative entropy between a posterior and the prior from which it was formed. The prior asymptotically achieving the maximal value is the prior that permits the most rapid accumulation of information from the data, on average. This sense of optimality is consistent with the way we assess the information content of an arbitrary density for the parameter below.

The Jeffreys prior is the normalized square root of the Fisher information. In the Poisson(θ) example, for instance, the Jeffreys prior is proportional to $1/\sqrt{\theta}$, which must be truncated to a compact set to yield a proper prior. No member of the conjugate family $\Gamma(p,\lambda)$ is the Jeffreys prior. Thus assessing the objective information content of a $\Gamma(10, 10)$, or of a normal distribution, requires comparing it to the posterior formed from a Poisson(θ) likelihood and the Jeffreys prior. Unfortunately, comparing either of these densities for the parameter to the posterior formed using the Jeffreys prior does not reduce in any obvious way to solving equations. Nevertheless, we show that one can use approximate techniques to identify a data set from a Poisson(θ) that can be used to update the Jeffreys prior so that it is as close as possible to the $\Gamma(10, 10)$ or a given normal.

In relative entropy we approximate an arbitrary density for the parameter by a posterior formed from a likelihood and a Jeffreys prior. Relative entropy is a strong measure of distance and thus may be difficult to use. But it is a natural choice for two reasons. First, the relative entropy has physical interpretations in terms of the redundancy of code lengths, and in terms of the rate of transmission across an information-theoretic channel (see Clarke and Barron 1994). Thus the Jeffreys prior provides maximal data compression and the highest rate of data transmission, both statistically desirable properties. Second, the relative entropy is the natural distance to use with the Jeffreys prior, because it is the measure of distance with respect to which the Jeffreys prior is the reference prior. We remark that although the L^1 distance may seem a more natural choice, no reference prior under L^1 distance has been derived.

Here we examine the relative entropy distance between the posterior formed from the Jeffreys prior and a known likelihood and an arbitrary density for the parameter. We seek the data set that makes this relative entropy distance as small as possible. The data set achieving this minimum represents the extra information in the arbitrary density.

2.2 The Minimization

Suppose that we have a parameterized class of densities $p_\theta(x)$ with respect to Lebesgue measure for a real, possibly vector-valued, random variable and that the experimenter wishes to estimate the parameter θ, assumed to vary over a parameter space Ω contained in R^d. If the only information that an experimenter is willing to assume is the appropriateness of the reference prior method, then the Jeffreys prior, denoted here by w_N, must be used as the objective prior. Denote the arbitrary informative density elicited from the experimenter by $w(\theta)$, with respect to d-dimensional Lebesgue measure.

Given the likelihood p_θ, we update w_N with a data set $x^n = (x_1, \ldots, x_n)$ to form the reference posterior $w_N(\theta|x^n)$. If we choose w to be a subjective prior with the same support as w_N, then we can formally diagnose what p_θ data is implicit in w by minimizing

$$D(w_N(\cdot|x^n)\|w(\cdot)) \tag{1}$$

over possible data sets x^n. Here D is the relative entropy, which for densities p, q with a common dominating measure is defined by $D(p\|q) = \int p\, ln(p/q)$; x^n is an iid sample of size n drawn from p_θ; and $w_N(\cdot|x^n)$ is the posterior density equal to $w_N(\theta)p(x^n|\theta)/m(x^n)$ in which $m(x^n) = \int_\Omega w_N(\theta)p(x^n|\theta)\,d\theta$ is the marginal density for the data. Here ln denotes the natural logarithm.

Data sets x^n achieving the minimum in (1) update w_N by p_θ so as to get a density for θ as close as possible to w. It is in this sense that x^n is representative of the information implicit in w, the experimenter's subjective beliefs. This is a general form of the Poisson(θ) $-$ $\Gamma(p, \lambda)$ example examined in closed form earlier. A posterior obtained by minimizing (1) does not generally equal w; it is only the posterior that is closest to w.

Outside the case where w_N and w are conjugate to p_θ, (1) will be strictly positive and bounded away from zero. Also, the minimal value of (1) is less than or equal to $D(w_N\|w)$. If equality is achieved, then no data can be added to the reference prior so as to give a posterior closer to w than w_N already is. The smaller the minimal value of (1) is, the better w can be approximated by a posterior based on w_N and p_θ, and the more representative the minimizing x^n will be of the information implicit in w. If the minimal value is too large, then one cannot approximate w well by any posterior based on p_θ and w_N. In this case no information obtainable from p_θ can justify the beliefs implicit in w.

Note that the data implicit in w need not have originally come from p_θ. In particular w may be a posterior based on a data set from a different likelihood and prior. The minimization in (1) still converts the information in w into equivalent data from p_θ.

In an experimental design context, one regards w as a target inferential density for θ. Data sets x^n that minimize (1) identify the data that would best permit the inferences from w. If the minimal value of (1) were too large, then the inferences represented by w cannot be justified from p_θ. In practice, w is not known, and one only has the goal of obtaining a small credible region that has high posterior probability. Thus one would be equally content to make inferences from any member of a class Γ of densities for θ, where all the densities in Γ achieve a prescribed degree of concentration and could be well matched by minimizing (1).

For instance, let Γ be the collection of densities for θ with variance between two preassigned values. In this case one could minimize (1) for each w in Γ and so obtain a collection of data sets, say $x^n(w)$ for w ranging over Γ. Let n^* be the largest value of the $n(w)$'s. Using n^* as a sample size ensures that for each element w of Γ there is at least one data set that might be realized and would give inferences as good as or better than w. Because we do not expect n^* to be small, we can choose Γ to be the collection of Normal(μ, σ^2) distributions with σ^2 between two preassigned values. This is natural, because the posterior converges to a normal distribution concentrating at its mean at rate $O(1/\sqrt{n})$ for large sample sizes.

3. ASSOCIATING A SAMPLE TO AN INFORMATIVE PRIOR

In this section we begin by demonstrating that in general one can find a data set that minimizes (1). First, consider a simple algorithm that could perform the minimization. Choose a fixed noninformative reference prior w_N and a likelihood and suppose that we are given a subjective density w. Then we know the relative entropy between them, $D(w_N\|w)$. We can consider adding one data point x_1 to w_N to form $w_N(\theta|x_1)$. Now we can evaluate $D(w_N(\cdot|x_1)\|w(\cdot))$ for each x_1 and compare it to $D(w_N\|w)$. If $D(w_N(\cdot|x_1)\|w(\cdot))$ is larger than $D(w_N\|w)$ for all values of x_1, then we know that adding data to w_N cannot move w_N closer to w than it already is. So w_N and w represent roughly equivalent information. If there is a value of x_1, say x_1^*, that minimizes $D(w_N(\cdot|x_1)\|w(\cdot))$ to give a value less than $D(w_N\|w)$, then we consider adding another data point x_2. We evaluate $D(w_N(\cdot|x_1,x_2)\|w(\cdot))$ for all pairs (x_1,x_2) to search for a pair $(x_1,x_2)^*$ that minimizes it. If this minimal value is greater than $D(w_N(\cdot|x_1^*)\|w(\cdot))$, then we say that x_1^* is the data set implicit in w. Otherwise, we continue and minimize $D(w_N(\cdot|x_1,x_2,x_3)\|w(\cdot))$ over all triples. We proceed in this way to find a value of x^n that minimizes (1).

The intuition that this sort of minimization procedure will stop at a minimum is to recall that if n increases too much, then the posterior $w_N(\cdot|x^n)$ converges to a normal with variance decreasing at rate $O(1/n)$. If this convergence obtains, then it forces the relative entropy between $w_N(\cdot|x^n)$ and w to become arbitrarily large, because the limiting normal is far from any fixed w. This is not a surprise, because the expected value of the quantity in (1) under the marginal distribution of X^n is increasing with rate $O(ln\, n)$ (see Clarke and Barron 1994). The consequence is that a minimizing data set x^n must be obtained for a value of n that occurs before convergence of the posterior to normality pushes the posterior too far away from w. Our result shows that we can find a minimizing data set x^n with high probability, as assessed in the marginal distribution for the data.

The second part of this section examines the behavior of $D(w_N(\cdot|x^n)\|w(\cdot))$ when the likelihood is Poisson(θ) and w_N is the corresponding Jeffreys prior. With this fixed prior-likelihood pair, we regard $D(w_N(\cdot|x^n)\|w(\cdot))$ as a real-valued function of n and the sufficient statistic $\sum x_i$ and generate three-dimensional graphs for various choices of w. These graphs reveal the typical shape of such surfaces and permit one to identify a minimizing data set.

3.1 Existence of a Minimum

Suppose that the likelihood $p(x|\theta)$ admits at least two continuous derivatives in θ for almost every x, that the Fisher information matrix $\mathbf{I}(\theta)$ is well defined and positive definite for every θ in $\Omega \subset \mathbb{R}^d$, and that the support of the prior density w is Ω, a compact connected set that is the closure of an open set.

We assume that there is a $\delta = \delta(\theta) > 0$, so that for each i, j from 1 to d we have

$$E_\theta \sup_{\|\theta-\theta'\|\le\delta} \left| \frac{\partial^2}{\partial\theta_i\partial\theta_j}\, ln\, p(X_1|\theta') \right|^2 < \infty, \qquad (2)$$

and that for each i we have

$$E_\theta \left| \frac{\partial}{\partial\theta_i}\, ln\, p(X_1|\theta) \right| < \infty. \qquad (3)$$

Also, we assume that p_θ is soundly parameterized in the sense that convergence of a sequence of parameter values in Euclidean distance in $\mathbb{R}^d$ is equivalent to weak convergence of the distributions they index. Details on this assumption can be found in Clarke and Barron (1990). We write M_n to mean the probability defined by the mixture density m_n. We denote the expectation with respect to $p(\cdot|\theta)$ by E_θ, and the expectation with respect to $m = m_n$ by E_m. For convenience, we write E when no confusion will result.

Theoretical verification that the stepwise search described earlier will be successful rests on the asymptotic normality of the posterior density. Essentially, we show that as n increases, the M_n-probability that $D(w_N(\cdot|x^n)\|w(\cdot))$ is large also increases. That is, we know $D(w_N\|w)$, and our goal is to find a data set x^n to minimize $D(w_N(\cdot|x^n)\|w(\cdot))$ at a value strictly less than $D(w_N\|w)$. Letting K be strictly positive, we restrict attention to data sets x^n for which $D(w_N(\cdot|x^n)\|w(\cdot))$ is less than $D(w_N\|w)+K$. Formally, we establish that there is a critical value n_0, so that for sample sizes $n \ge n_0$, the M_n probability of $D(w_N(\cdot|x^n)\|w(\cdot))$ being larger than $D(w_N\|w)+K$ is at least $1-\varepsilon$, for arbitrary preassigned $\varepsilon > 0$. This means that it is essentially impossible to find data sets of arbitrarily large size that can minimize $D(w_N(\cdot|x^n)\|w(\cdot))$. In particular we restrict our search for a minimizing data set x^n to those data sets for which $n < n_0$, because consideration of sample sizes larger than n_0 is pointless. It is appropriate to use the mixture probability M_n because we choose priors pre-experimentally.

In addition to (2) and (3), we assume the Wald (1949) hypotheses for consistency of the maximum likelihood estimator (MLE). These can be compactly stated as three conditions: First, for each θ_0 there is a $\rho = \rho(\theta_0)$ so large that

$$E_{\theta_0} \sup_{\psi:|\psi-\theta_0|>\rho} ln\, \frac{p(X|\psi)}{p(X|\theta_0)} < 0.$$

Second, for each θ_0 and for any θ, there is a $\delta = \delta(\theta) > 0$ small enough such that

$$E_{\theta_0} ln\, \frac{p(X|\theta_0)}{\sup_{\theta':|\theta-\theta'|<\delta} p(X|\theta')} < \infty.$$

Third, for every $x, p(x|\theta) \to 0$ as $\|\theta\| \to \infty$. Now, our result is the following.

Theorem 3.1. Assume Wald's hypotheses and let $K > 0$. For any preassigned $\varepsilon > 0$, there exists an $N = N_\varepsilon$, so that for all $n \ge N_\varepsilon$,

$$M_n(D(w_N(\cdot|X^n)\|w(\cdot)) \ge K) \ge 1-\varepsilon \qquad (4)$$

where M_n is the probability measure associated to the density $m(\cdot)$, and $w_N(\cdot|X^n)$ is a posterior based on a noninformative prior $w_N(\cdot)$.

Proof. This is a special case of Theorem 4.1.

We remark that any prior satisfying the formal conditions of this result can be used in place of w_N. The difference in interpretation is any other prior ν would correspond to finding a data set that must be added to an informative prior.

3.2 The Poisson Example

To understand how the quantity $D(w_N(\cdot|x^n)\|w(\cdot))$ behaves for various choices of prior w as we vary the data set x^n, we produce graphs of $D(w_N(\cdot|x^n)\|w(\cdot))$ as a function of the sample size and a summary statistic for the Poisson(θ) example that cannot be studied in closed form.

Let the X_i's be iid Poisson(θ) and let the parameter space be [1, 11]. Now the reference prior is

$$w_N(\theta) = \frac{1}{c\sqrt{\theta}}\ , \tag{5}$$

where $c = \int_1^{11} 1/\sqrt{\theta}\, d\theta$, (see Bernardo 1979, and Clarke and Barron 1994). The probability mass function for X^n is

$$P_\theta(x^n) = \frac{e^{-n\theta}\theta^{\Sigma x_i}}{\prod x_i!}\ , \tag{6}$$

where the sum and product are over $i = 1, \ldots, n$. Mixing (6) with respect to (5) gives

$$m(x^n) = \int_1^{11} \frac{1}{c} \frac{e^{-n\theta}\theta^{\Sigma x_i - 1/2}}{\prod x_i!}\, d\theta.$$

So the posterior is

$$w_N(\theta|x^n) = w_N\Big(\theta\Big|\sum x_i\Big) = \frac{e^{-n\theta}\theta^{\Sigma x_i - 1/2}}{\int_1^{11} e^{-n\theta}\theta^{\Sigma x_i - 1/2}}\, d\theta,$$

and (1) is

$$D\Big(w_N\Big(\cdot\Big|\sum x_i\Big)\Big\|w(\cdot)\Big) = \int_1^{11} \frac{e^{-n\theta}\theta^{\Sigma x_i - 1/2}}{\int_1^{11} e^{-n\theta}\theta^{\Sigma x_i - 1/2}\, d\theta} \times ln\, \frac{\theta^{\Sigma x_i - 1/2}e^{-n\theta}}{w(\theta)\int_1^{11} e^{-n\theta}\theta^{\Sigma x_i - 1/2}\, d\theta}\, d\theta, \tag{7}$$

where $w(\theta)$ is any prior density continuous with respect to Lebesgue measure and strictly positive on [1, 11].

As a first example, choose w to be a Normal(6, 1) truncated to [1, 11] and renormalized. By direct calculation as described at the beginning of this section, one can use numerical integration to verify that

$$\min_{x^n} D\Big(w_N\Big(\cdot\Big|\sum x_i\Big)\Big\|\ w(\cdot)\Big) = .01260359, \tag{8}$$

and this minimum occurs for $n = 6, \sum x_i = 37$. This means that relative to the reference prior, this truncated normal prior implicitly assumes information equivalent to 6 previous Poisson outcomes summing to 37. We see that $37/6 = 6.1666$ is close to the mean of w. Note that we only have a data set from a Poisson distribution; no statement about the value of the true parameter has been made.

Figure 1 is the graph of (7) as a function of n and $\sum x_i$ when w is a Normal(6, 1), truncated to [1, 11], and renormalized. This gives a surface in three dimensions. The minimum of the surface is attained in the central portion of the valley in the diagram. Beyond this point, the valley is actually increasing.

To verify that the minimizing posterior density matches the subjective density, one can graph them to see how close they are on regions of the parameter space that are most important. Figure 2 shows the plot of the truncated Normal(6, 1) for the parameter as a heavy line and its closest approximation as a thin line. It can be seen that the two are quite close in an L^1 sense.

The height of the surface over the $D = 0$ plane indicates how good the approximation is at each point in the plane in relative entropy. The minimal value of D indicates how well the prior has been approximated. Because the relative entropy is the integral of the logarithm of a density ratio, it is very sensitive to tail behavior and can magnify small differences. This makes the minimal value difficult to interpret. A crude bound on the L^1 distance in terms of the relative entropy is $\|f-g\|_1 \leq \sqrt{2D(f\|g)}$ (see Csiszar 1967). In this example the L^1 distance between the truncated normal and its closest approximation is bounded by .1587677, at worst. But one can see graphically that the agreement between them is much closer than that bound would indicate.

In principle, one can reverse the roles of the Jeffreys prior and the truncated normal prior. Solving the corresponding minimization problem would reveal what data a user of the Jeffreys prior is assuming relative to the normal prior. But as in the $\Gamma(1,1) - \Gamma(10,10)$ example in Section 1, this data set would be void, because the Jeffreys prior is minimally informative. Because the relative entropy is asymmetric, it only makes sense to put the posterior formed from the noninformative prior in the first entry and to put the subjective prior in the second entry.

As an example of a poor fit, choose $w(\theta) = .1$, the uniform prior on [1, 11]. Now, minimization as in (8) shows that the best approximation to the uniform prior is a posterior with $n = 1$ and $x_1 = 7$. The minimal value of the relative entropy is .1823156. This gives a poor match, because the posterior $w_N(\cdot|x_1 = 7)$ is unimodal with mode near 6.5. But it is the best match one can find. Consideration of the graphs of the uniform and of the closest posterior suggests

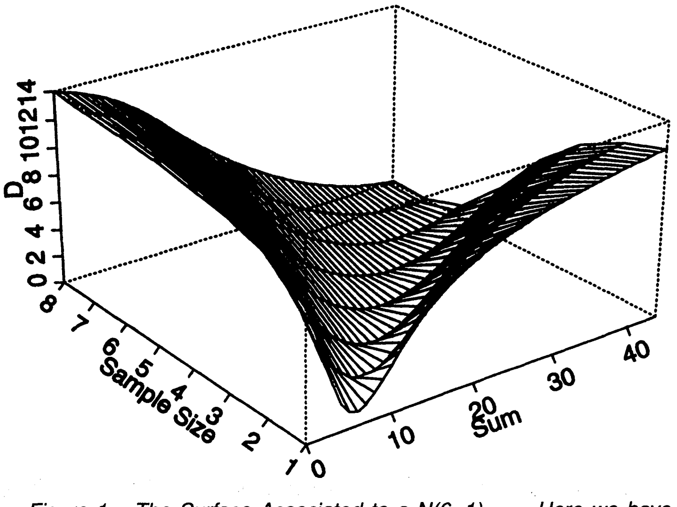

Figure 1. The Surface Associated to a N(6, 1)$_{[1,11]}$. Here we have plotted D(w$_N$(·|Σx_i)‖ N(6, 1)$_{[1,11]}$) as a function of the sample size n and the value of the sufficient statistic, Σx_i. The axis to the left gives n; the axis to the right gives the value of Σx_i; the vertical axis gives the relative entropy. The minimal relative entropy, .01260359, is achieved for n = 6 and Σx_i = 37. Beyond the portion of the plane shown, the surface continues to increase.

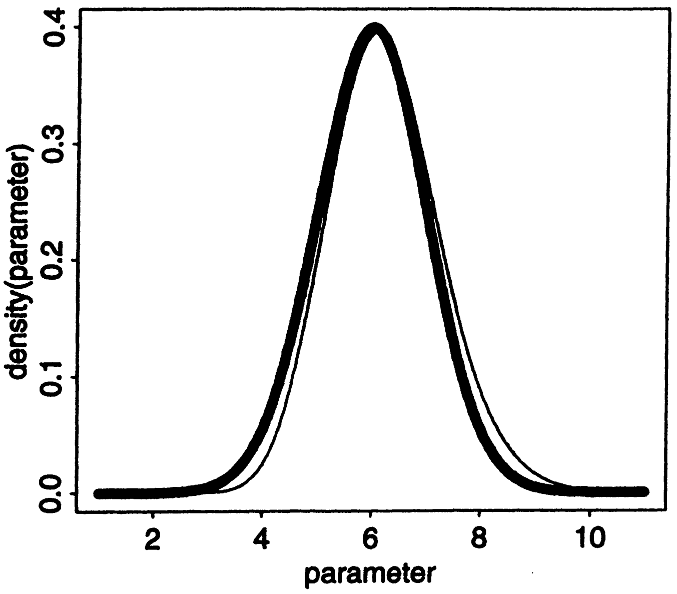

Figure 2. The Truncated Normal and Its Closest Approximation. The truncated normal of Figure 1 (heavy line) is seen to be close to the approximation given by exp(−6θ)$\theta^{36.5}$ divided by its integral over [1, 11] (thin line). The slight shift between the two densities reflects the tail of the posterior.

that the uniform prior represents information equivalent to one outcome, even if uniform inferences cannot reasonably be derived form a Poisson(θ) with the Jeffreys prior.

In general, graphs such as those in Figure 1 are qualitatively similar for smooth priors. In the limit, as the pair $(n, \sum x_i)$ increases in any Euclidean norm, the surface height over the $D = 0$ plane increases. So the minimum of (7) is attained within a finite region in the positive quadrant of the $D = 0$ plane. Restricting to $[0, n_{\max}] \times [0, (\sum x_i)_{\max}]$, for instance, the surface can exhibit various shapes. Usually there is a valley that decreases to the minimal value and then increases as asymptotic normality dominates. In practice, one can use ever larger sample sizes and values of the sum until the minimum occurs inside a rectangular block.

We comment that if one evaluates the information in priors that are not smooth, then the surface usually becomes irregular. Priors that are continuous but not differentiable are hard to match with a posterior based on the Jeffreys prior. Indeed, priors that are not continuous can generate a surface that only increases so that the void data set achieves the minimum. The lack of smoothness is hard to match in relative entropy; by the time asymptotic normality has become dominant, the data sets that would be representative have already been bypassed. Consequently, the minima for nonsmooth priors may occur closer to the origin than for smooth priors. Thus a smooth noninformative prior is already almost as close as possible to nonsmooth inferences.

As an example where matching is successful despite the lack of smoothness, choose w to be piecewise linear, assigning most of its mass near 6. Figure 3 shows the surface generated for w chosen to be the piecewise linear function graphed in Figure 4. Figure 4 also shows the closest approximating posterior. Although w is not smooth the closest approximation is good; the minimal relative entropy distance is .03839288, achieved for $n = 4$ and $\sum x_i = 24$.

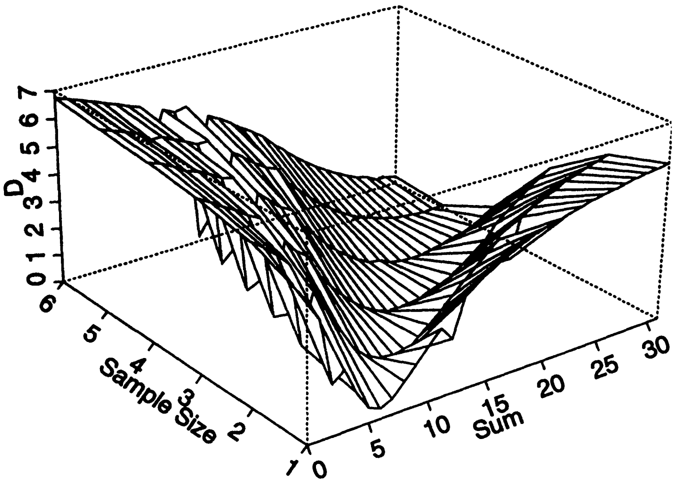

Figure 3. The Surface Associated to a Piecewise Linear Density. Here w is given by (f(θ) + .01)/3.6, where f is piecewise linear. The function f is zero for θ < 3 or θ > 9. on [5.5, 6.5], f(θ) = 1; on [3, 5.5], f(θ) = .4θ− 1.2; and, on [6.5, 9], f(θ) = −.4x + 3.6. The surface is seen to be irregular but decreases to a minimum of .03839288 at n = 4 and Σx_i = 24, after which it increases.

Several observations can be made from the cases we have examined. First, the minimizing value of the sufficient statistic tracks the location of the subjective density for the parameter. Also, as the concentration of the subjective density for the parameter increases, the value of $D_{\min}$ decreases and the sample size at the minimum increases. We get a better fit for more highly concentrated w's, and the concentration of w represents sample size. This is used in the next section to give a design criterion. Third, it takes a larger sample size to get the same concentration as the true value shifts to regions where the Jeffreys prior is smaller

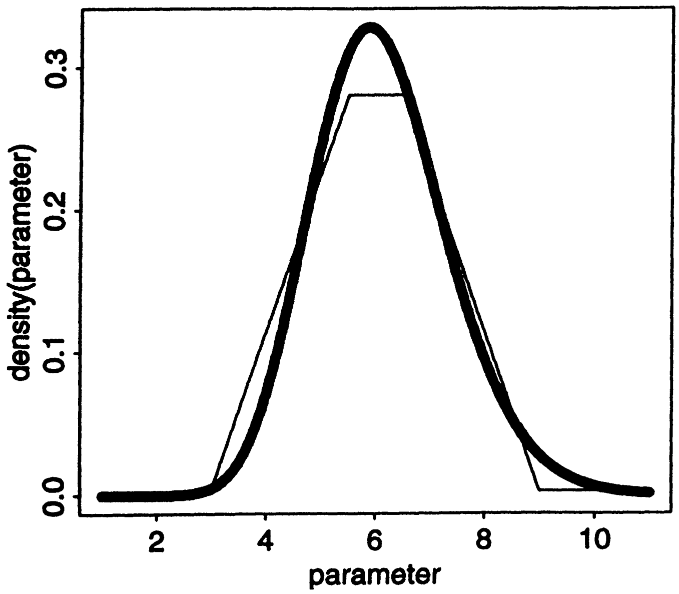

Figure 4. The Piecewise Linear Density and Its Closest Approximation. The piecewise linear density of Figure 3 is plotted and seen to be close to its closest approximation given by exp(−4θ)$\theta^{23.5}$ divided by its integral over [1, 11]. Because the tails nearly match, the discrepancy occurs mostly at the mode.

and symmetric priors tend to reflect a smaller sample size than priors that are not symmetric, the decrease tracking the loss of symmetry. This reflects asymptotic normality. Fourth, the best matches tend to be in the midrange of the parameter space rather than near the boundaries. Finally, the sample size of the best approximation can be unreasonably low, particularly for nonsmooth w's, and the mean of the approximation can be shifted from the mean of w. This occurs when the tails match poorly, and may indicate the unattainability of the inferences represented by w.

4. SAMPLE SIZE IMPLICATIONS

Now, we regard w as an inferential density for the parameter. That is, if we obtain w, then the inferences it would give would meet our estimation goal. The technique here for sample size selection ensures that for every member w of a class of inferential densities Γ there will exist a data set that would yield a posterior as close as possible to w. In fact, an experiment will yield only one data set. Consideration of alternative data sets ensures that we have collected enough data to achieve our estimation goal.

Let w be a member of the class Γ. By the technique of Section 3, we match each $w \in \Gamma$ to a data set x^n that would best permit the inferences represented by w. Then we scan the data sets obtained from all the w's and choose the largest of their sample sizes. We denote this maximum sample size by n^*.

For this to be successful, the minimum of $D(w_N(\cdot|x^n) \| w(\cdot))$ must be small enough so that approximating each w by a posterior will be meaningful. This is partially assured by the asymptotic normality of the posterior. Indeed, $D(w_N(\cdot|x^n)\|w_n(\cdot))$ tends to zero in P_θ probability and in L^1 when θ is the true value and w_n is the limiting normal sequence. Thus $D(w_N(\cdot|x^n)\|w(\cdot))$ will be small if Γ is a collection of normal densities with small variance. More generally, if w is sufficiently concentrated and smooth, then there will typically exist a data set for which $D(w_N(\cdot|x^n)\|w(\cdot))$ is meaningfully small. On the other hand, if this does not occur, then it may indicate that the inferences represented by w cannot be justified from the experiment. Use of such w's is inappropriate.

Suppose that the set $\Gamma = \Gamma(\varepsilon_1, \varepsilon_2)$ consists of densities for Θ that have variances between two small constants $\varepsilon_2 > \varepsilon_1 > 0$. We consider three ways to choose a sample size. First is the inference-matching method, which uses the largest sample size n^* of the data sets found by matching the inferences of all elements w of Γ. Second is the uniform method, which obtains the smallest sample size n_u that ensures all posteriors formed from at least n_u data points achieve the preassigned posterior variance. Third is the sequential method, in which one accumulates data until the posterior variance drops below a threshold value. We write n_s for this method and note that it is random.

We illustrate these methods for three likelihoods: the normal, the binomial, and the Poisson. They coincide for the normal but differ for the binomial and Poisson. For the normal and binomial, all three methods can be used. For the Poisson, the uniform method cannot be applied. For the binomial and Poisson, the sequential method has the problem that the posterior variance may increase when the sample size increases. In each case we use a conjugate prior in place of the noninformative prior w_N for the convenience of closed-form solutions. Using a noninformative prior will give sample sizes that are somewhat larger, because it contains less information.

First, for a normal likelihood with a normal prior all of the methods for selecting a sample size coincide, because the posterior variance is independent of the data. Indeed, let $X_i \sim N(\theta, \alpha^2)$ and for w_N use $\theta \sim N(\mu, \tau^2)$, where μ, τ^2, and α^2 are known. Then $(\Theta|X^n)$ is a normal random variable with mean $E(\Theta|X^n)$ and variance given by $\mathrm{var}(\Theta|X^n) = ((\alpha^2/n)\tau^2)/(\tau^2 + (\alpha^2/n))$. Now the posterior variance depends only on the sample size, and it is straightforward to verify that $\mathrm{var}(\Theta|X^n)$ is strictly decreasing as a function of n. This means that adding more data always reduces the posterior variance; that is, for all x^n for all x_{n+1}, $\mathrm{var}(\theta|x^n) > \mathrm{var}(\theta|x^{n+1})$.

Using the uniform method, one can find an n_u to force the posterior variance to be uniformly small in particular, less than, say, ε_1. If we choose Γ to be the set of normal densities with variance ε_1 then n^* equals n_u, because each member of Γ can be realized as the posterior from a data set of size n_u. The sequential method is seen to give the same result as the uniform and inference matching methods, because the posterior variance is independent of the data apart from the sample size.

Second, suppose that $X_i \sim \mathrm{Bernoulli}(p)$, so that $Y_n = \sum_{i=1}^n X_i$ is $\mathrm{Binomial}(n,p)$, and let $w(p)$ be a conjugate $\mathrm{Beta}(\alpha, \beta)$ prior. Now the density of $(P|Y_n), w(p|y_n)$ is $\mathrm{Beta}(y + \alpha, n - y + \beta)$, so the variance is not a decreasing function of sample size and depends on more than just n. In this case the three methods will not be equivalent. Because the X_i's are bounded, $\sup_{x^n} \mathrm{var}(\theta|X^n = x^n)$ decreases monotonically at rate $O(1/n)$. As a consequence, one can identify a value n_u that will guarantee a small posterior variance. But this value will typically be greater than the value of n_s because the posterior variance obtained may decrease much faster than the supremum over all possible posterior variances. The inference-matching method gives results between the sequential method and the uniform method.

As a final example let $X_i \sim \mathrm{Poisson}(\theta)$ and, for convenience, use the conjugate prior $\mathrm{Gamma}(\alpha, \beta)$. Given data x^n, the posterior density, $w(\theta|x^n)$, is $\mathrm{Gamma}(n\bar{x} + \alpha, n + \beta)$ with variance $(n\bar{x} + \alpha)/(1 + n)^2$. Because the supremal posterior variance is infinite, the uniform method will not work. The sequential and inference-matching methods will give distinct nontrivial results, with $n^* \geq n_s$ for appropriately chosen Γ's.

4.1 The Inference-Matching Method

Let Γ be a set of smooth densities representing the inferences that we wish to entertain. For instance, we might use $\Gamma(\varepsilon_1, \varepsilon_2)$ as earlier. Whatever the definition of Γ, Theorem 2.1 guarantees that for each $w \in \Gamma$, there will exist a data set $x^n(w)$ achieving

$$x^n(w) = \underset{x^n}{\operatorname{argmin}}\ D(w_N(\cdot|x^n)\|w(\cdot)),$$

with sample size denoted by n_w. For smooth, sufficiently concentrated w, this infimum should be meaningfully small, as noted earlier. Also, it is often the case that when a sufficient statistic such as $\bar{X}$ exists, $D(w_N(\cdot|\bar{x})\|w(\cdot))$ gives a convex surface as a function of n and $\bar{x}$, thereby guaranteeing the existence of a minimizing data set. Moreover, although it is conceivable that the infimum might not be attained by any data set, there will exist data sets for which it is attained to any specified degree of approximation.

Here we give a result that ensures the existence of a minimizing data set in a marginal probability sense. Heuristically, we can also give a nonprobabilistic argument based on a strong form of the asymptotic normality of the posterior in relative entropy. Let $\varepsilon > 0$ and suppose that there exists an N_ε so that for $n > N_\varepsilon$ the relative entropy distance between a posterior based on n observations and its limiting normal is less than ε. Next, choose ε so small that all the limiting normals, with variances $O(1/n)$, are further away from w than w_N is for $n \geq N_\varepsilon$. Now we can restrict to samples of size less than N_ε. For bounded random variables, we are minimizing a continuous function on a compact set, so the minimum must occur in that set. If the random variables are unbounded, then we use the fact that $p(x|\theta) \to 0$ as $x \to \infty$. Thus as any x_i increases without bound, the posterior must shift, thereby increasing the relative entropy between the posterior and w. So the x_i's must remain within a compact set on which a minimum must be attained. This argument can be formalized in some cases where a sufficient statistic for the parameter is present.

Once we have assigned a data set $x^n(w)$ to each member w of Γ, we may choose

$$\begin{aligned} n^* &= \max\{n(w)|x(w)^{n(w)} \\ &= \underset{x^n}{\operatorname{argmax}}\ D(w_N(\cdot|x^n)\|w(\cdot)), w \in \Gamma\} \quad (9) \end{aligned}$$

as a sample size. This n^* guarantees that every inferential density in Γ has a good approximation by a posterior based on a sample of size not more than n^*. It does not guarantee that every posterior based on a sample of size n^* is a member of Γ. We call the choice of n^* given in (9) the inference-matching method.

Clearly, the inference-matching method depends on the choice of Γ. If Γ consists of all densities for θ with variance less than some ε, then the method gives $n^* = \infty$. If Γ is too small, then it may exclude densities that concentrate at parameter values we want to consider and thus give an unrealistically small value for n^*. Indeed, if Γ is a singleton set $\{w\}$, then $n^* = n(w)$. Neither extreme is a realistic formulation of the sample size problem. We suggest that the best choice for Γ in general is $\Gamma(\varepsilon_1, \varepsilon_2)$ or its intersection with normal densities having variances in $(\varepsilon_1, \varepsilon_2)$.

The method is conservative because, provided that Γ is well chosen, we are assured that any element of Γ can be approximated by a posterior formed from a data set we might get, with the exception that for certain potential inferences we have obtained a bit more data than we require. Specifically, if some members of Γ are best approximated by using a sample of size $n < n^*$, then the extra $n^* - n$ data points that we have collected will result in a degree of precision greater than the minimum that we actually required. This form of conservatism can be reduced only by using a smaller class of inferential densities; that is, by having a better idea of what conclusions to anticipate in advance of experimentation.

Our first result is a theorem giving conditions under which n^* is well defined. For generality, we choose the set Γ to be the class of all continuous densities for Θ supported on the same set as w_N, and denote this by Γ_1. The result states that for n sufficiently large, the marginal probability that $D(w_N(\cdot|X^n)\|w(\cdot))$ is large can be made uniformly close to unity, where the uniformity is over w in Γ_1. Apart from the uniformity over Γ_1, this is the same statement as Theorem 3.1. Indeed, if Γ is a singleton set, then Theorem 4.1 reduces to Theorem 3.1.

The uniformity is possible because we are evaluating probabilities in the mixture distribution with respect to a fixed proper prior w_N. In effect, this means that we can reduce the domain of the mixture to a compact set in the parameter space with high W_N probability on which the Fisher information is bounded away from zero. As a consequence, certain convergences used in the proof are uniform in P_θ for θ in the compact set. Densities for the parameter that assign little mass to the compact set therefore do not matter.

In practice, one would choose a finite set of densities in Γ so that no member of Γ is too far from at least one element of the finite set. Then for each member w of the finite class, one can find $n(w)$ and thereby obtain a suitable value for n^*.

Theorem 4.1. Assume the Wald (1949) hypotheses for consistency of the maximum likelihood estimator $\hat{\theta}$ and let $K > 0$. For any preassigned $\varepsilon > 0$, there is an N_ε so that for all $n > N_\varepsilon$,

$$\inf_{w \in \Gamma_1} M_n(D(w_N(\cdot|X^n)\|w(\cdot)) > K) \geq 1 - \varepsilon. \quad (10)$$

Proof. See Appendix A.

This result shows that n^* is well defined in a sense similar to that used in Theorem 3.1. Uniformly over $w \in \Gamma_1$, the M_n probability that the minimal value of $D(w_N(\cdot|X^n)\|w)$ is achieved for $n < N(\varepsilon)$ is large. That is, if we have $N(\varepsilon)$ such that $\inf_w M_n(D(w_N(\cdot|X^n)\|w) > K) \geq 1 - \varepsilon$, then $n > N(\varepsilon)$ means that, uniformly in w and with high M_n probability, we have that $D(w_N(\cdot|X^n)\|w) > K$. Thus it is likely that for all $w, n(w)$ occurs for $n < N(\varepsilon)$, implying that $n^* \leq N(\varepsilon)$.

This result is anomalous, because n^* will be infinite for any choice of Γ that includes densities of arbitrarily high concentration. The anomaly is resolved by observing that the result uses the marginal density for the data that is the mixture of the likelihood with respect to w_N. Indeed, consider a density w in Γ_1 that is highly concentrated at a parameter value θ'. For choices of θ that are outside a small neighborhood of θ', the P_θ probability of the event $\{D(w_N(\cdot|X^n)\|w) > K\}$ will increase, roughly uniformly in θ as n increases. In contrast, the P_θ probability of that event will decrease to zero for values of θ in a shrinking neighborhood of θ'. But the prior probability of such neighborhoods of θ' will be arbitrarily small, so those θ's do not contribute much to the marginal probability. In particular, the P_θ probability can be very small for some θ's, provided that the rest of the θ's have W_N measure near 1 and the P_θ probability is large for them.

For contrast, suppose that one were to assess probability in P_{θ_0} for a fixed value θ_0 instead of in M_n. Now consider $P_{\theta_0}(D(w_N(\cdot|X^n)\|w(\cdot)) > K)$ for various choices of w. For any fixed w and $K > 0$, one can prove that this probability will go to unity uniformly over θ_0's in a compact set. But uniformity over w fails. Indeed, posterior normality guarantees that for each θ_0 there exists a sequence w_n of densities for the parameter that concentrates at θ_0, so that $D(w_N(\cdot|X^n)\|w_n(\cdot))$ converges to zero in P_{θ_0}. If w_n is not permitted to depend on the data, then $D(w_N(\cdot|X^n)\|w_n(\cdot))$ converges to a constant rather than to infinity. If probability is assessed in P_{θ_0}, then it is only possible to prove a result when the variances of the densities in Γ are bounded away from zero.

4.2 Properties of the Methods

First, we formally define the uniform and the sequential methods. The uniform method chooses sample size $n_u = n_u(\varepsilon)$ to be the least value of n satisfying

$$\sup_{x^{n'}} \operatorname*{var}_{w_N}(\Theta|x^{n'}) < \varepsilon \tag{11}$$

for $n' > n$. That is, as soon as n' exceeds n, $\mathrm{var}_{w_N}(\Theta|x^{n'}) < \varepsilon$ for all $x^{n'}$. As noted in the Poisson(θ) example, $n_u(\varepsilon)$ may not exist. The sequential method collects data until the posterior variance $\mathrm{var}_{w_N}(\theta|x^k)$ falls below ε. That is, we choose $n_s = n_s(\varepsilon)$ to be the first value of k that satisfies

$$\operatorname*{var}_{w_N}(\Theta|x^{k-1}) > \varepsilon, \qquad \operatorname*{var}_{w_N}(\Theta|x^k) < \varepsilon. \tag{12}$$

The expected value of this stopping time is $En_s = \sum_{k=1}^{\infty} kP$ (stop at k). The consistency of the posterior implies the existence of n_s.

Now we can provide bounds on n^* in terms of these two methods.

Proposition 4.1. Bounds on n^*: Let n^* be obtained from the class of densities $\Gamma = \Gamma(\varepsilon_1, \varepsilon_2)$ for Θ. Then

$$n_s(\varepsilon_2) \le n^* \le n_u(\varepsilon_1). \tag{13}$$

Proof. See Appendix B.

Proposition 4.2. Assume that $\int \theta^4 w(\theta)\,d\theta < \infty$ and for $a, b > 0$. Let μ be the number of upcrossings of $[a, b]$ by $\mathrm{var}(\Theta|X^n)$. If Wald's hypotheses for consistency hold (see Thm. 3.1), then

$$\overline{\lim_{n\to\infty}}\, E\mu \le 1 + \frac{b}{b-a} = 1 + \frac{1}{1-(a/b)}\ .$$

Proof. See Appendix B.

Together, these results show that n^* is bounded above by the uniform method and below by the sequential method, and that the lower bound is not too loose in that the posterior variance will not increase very often. Thus n^* is a compromise between a method that cannot always be applied, and one that may give unrealistically small sample sizes. In addition, these bounds may be used to approximate n^*, and they provide a comparison of three Bayesian techniques for sample size selection based only on the posterior.

5. DISCUSSION

The point of this article is to demonstrate that the optimality property satisfied by a reference prior permits one to associate to any other density for the parameter a data set that represents the information it contains. This information may be pre-experimental in the sense of Section 3, or in the sense of Section 4. In the latter case, consideration of a large class of inferential densities justifies a sample size procedure. We comment that the main quantity relates data to inferences and so can be used for calibration. For instance, one can convert a posterior based on dependent data to a posterior based on independent data that would give equivalent inferences, and one can combine the data sets from priors reflecting independent pre-experimental beliefs.

A potential paradox may arise from multidimensional parameters. If θ is multidimensional, then it may be possible to use the formulation given here to evaluate how hard it is to estimate parameters. Indeed, there may be in some examples a sort of "antagonism" between parameters. Consider $\theta = (\theta_1, \theta_2)$ and let

$$n_{1,2} = \operatorname{argmin} D(w_N(\theta_1, \theta_2|X^n)\|w(\theta_1)w(\theta_2)),$$

the version of (1) that one would examine for joint estimation of (θ_1, θ_2), where w is a joint density for the parameters. Now let

$$n_1 = \operatorname{argmin} D(w_N(\theta_1|X^n)\|w(\theta_1))$$

and

$$n_2 = \operatorname{argmin} D(w_N(\theta_2|X^n)\|w(\theta_2));$$

these are the quantities that one would examine for estimating θ_1 and θ_2 separately—and for determining how hard they were to estimate to a specified degree of accuracy. One would say that θ_1 and θ_2 are antagonistic if $n_{1,2} \ge n_1 + n_2$; that is, it takes more data to estimate the parameters together than to estimate them individually. Indeed, one might take this a step further and seek a sort of "uncertainty principle" in which the sort of data that permit good estimation of one parameter prevent good estimation of another.

APPENDIX A: PROOF OF THEOREM 4.1

Fix any $w \in \Gamma$ and $K \ge 0$, and write

$$\phi_{\tilde\theta}(\theta) = \frac{n^{d/2}|\mathbf{I}(\tilde\theta)|^{1/2}}{(2\pi)^{d/2}}\, e^{-(n/2)(\theta-\tilde\theta)\mathbf{I}(\tilde\theta)(\theta-\tilde\theta)}, \tag{A.1}$$

where $\tilde{\theta}$ is the posterior mean using the prior w_N and $\mathbf{I}(\theta)$ is the Fisher information matrix for $p_\theta(\cdot)$. Now we have

$$\begin{aligned} &M_n(D(w_N(\cdot|X^n)\|w(\cdot)) > K) \\ &= M_n\left(\int w_N(\theta|X^n) \ln \frac{w_N(\theta|X^n)}{\phi_{\tilde{\theta}}(\theta)}\, d\theta \right. \\ &\qquad + \int w_N(\theta|X^n) \ln \phi_{\tilde{\theta}}(\theta)\, d\theta \\ &\qquad \left. - \int w_N(\theta|X^n) \ln w(\theta)\, d\theta > K\right) \\ &\geq M_n\left(\int w_N(\theta|X^n) \ln \phi_{\tilde{\theta}}(\theta)\, d\theta \right. \\ &\qquad \left. - \int w_N(\theta|X^n) \ln w(\theta)\, d\theta > K\right), \qquad \text{(A.2)} \end{aligned}$$

because Kullback–Leibler numbers are nonnegative. The last bound is

$$\begin{aligned} M_n\bigg(&\frac{d}{2} \ln \frac{n}{(2\pi)} + \frac{1}{2} \ln |\mathbf{I}(\tilde{\theta})| - \frac{n}{2} \operatorname{tr}(\mathbf{I}(\tilde{\theta}) \underset{w_N}{\operatorname{var}}(\Theta|X^n)) \\ &- \int w_N(\theta|X^n) \ln w(\theta)\, d\theta > K\bigg). \qquad \text{(A.3)} \end{aligned}$$

Note that $\tilde{\theta}$ converges to θ_0 and $n \operatorname{tr}(\mathbf{I}(\tilde{\theta})\operatorname{var}(\theta|X^n))$ converges to d in P_{θ_0}, uniformly for θ_0 in any compact set in the parameter space. (These are standard results and straightforward to prove under the stated hypotheses; see, for instance, Clarke 1989.) Choose a compact set C in the parameter space that has W_N probability at least $1 - \varepsilon/2$. Now we can write M_n explicitly as a mixture and get a lower bound given by

$$\begin{aligned} \int_C w_N(\theta_0) P_{\theta_0}\bigg(&\int w_N(\theta|X^n) \ln w(\theta)\, d\theta \\ &< \frac{d}{2} \ln n - K - \varepsilon\bigg)\, d\theta_0. \qquad \text{(A.4)} \end{aligned}$$

For given $w \in \Gamma$, let $\Delta = \Delta_w = \{w > B\}$, where B is a large positive constant. Now the Lebesgue measure of Δ is $\lambda(\Delta) \leq 1/B$, independently of w, so we can bound (A.4) from below. First, bound the integral in the event in (A.4) by using the upper bound B on Δ^c and the inequality $\ln w(\theta) \leq w(\theta)$ on Δ. Second, restrict the domain of integration in the mixture to Δ^c (which has large Lebesgue measure). The result is that

$$\begin{aligned} \int_{\Delta^c \cap C} w_N(\theta_0) P_{\theta_0}\bigg(&\int_\Delta w_N(\theta|X^n) w(\theta)\, d\theta \\ &\leq \frac{d}{2} \ln n - K - \varepsilon - B\bigg)\, d\theta_0 \qquad \text{(A.5)} \end{aligned}$$

is a lower bound for (A.4). Note that in (A.5) the inner integral is over Δ, which is a small set, but that the parameter value indexing the probability is an element of the complement Δ^c, a large set. We lower bound (A.5) by taking a supremum of the posterior over Δ. This gives

$$\begin{aligned} \int_{\Delta^c \cap C} w_N(\theta_0) P_{\theta_0}\bigg(&\int_\Delta w(\theta)\, d\theta \leq \frac{d}{2} \ln n - K - \varepsilon - B \text{ and} \\ &\sup_{\theta \in \Delta} w_N(\theta|X^n) \leq 1\bigg)\, d\theta_0. \qquad \text{(A.6)} \end{aligned}$$

Because the first condition in the event is satisfied for n large enough, and B can be allowed to increase (at a rate $o(\ln n)$), the result will be demonstrated provided that

$$P_{\theta_0}\left(\sup_{\theta \in \Delta} w_N(\theta|X^n) \leq 1\right) \to 1 \qquad \text{(A.7)}$$

for all values of θ_0 in a large subset of Δ^c. We choose a subset of Δ^c so that all elements of it are at least $\eta > 0$ away from every element of Δ. Because B may be allowed to increase, we can choose a subset of Δ^c with w_N probability arbitrarily close to unity. This follows from the boundedness of w_N and the fact that Δ has small Lebesgue measure. Now suppose that a value θ_0 in this subset is true. Then there is an open set containing θ_0 which does not intersect Δ, the set over which the supremum in (A.7) is taken. Using this fact and the boundedness of w_N, it is sufficient to show that for each θ_0 and any $\eta > 0$,

$$P_{\theta_0}\left(\sup_{\theta \in B(\theta_0, \eta)^c} \frac{p(x^n|\theta)}{m(x^n)} > \alpha\right) \to 0. \qquad \text{(A.8)}$$

From Clarke and Barron (1990, eqn. 6.5) we have that for any $r' > 0$,

$$P_{\theta_0}\left(\int_{B(\theta_0, \delta)} w(\theta) p(x^n|\theta)\, d\theta > e^{-nr'} p(x^n|\theta_0)\right) = 1 - O\left(\frac{1}{n}\right).$$

Thus, rearranging and increasing the domain in the integral gives

$$P_{\theta_0}\left(\frac{p(x^n|\theta_0)}{m(x^n)} < e^{nr'}\right) = 1 - O(1/n). \qquad \text{(A.9)}$$

From Wolfowitz (1949), we have that there exists an $r > 0$ so that

$$P_{\theta_0}\left(\sup_{\theta \in B(\theta_0, \eta)^c} \frac{p(x^n|\theta)}{p(x^n|\theta_0)} > e^{-nr}\right) < \varepsilon. \qquad \text{(A.10)}$$

Using (A.9) and (A.10) in (A.8), we see that the latter is

$$\begin{aligned} &P_{\theta_0}\left(\sup_{\theta \in B(\theta_0, \eta)^c} \frac{p(x^n|\theta)}{p(x^n|\theta_0)} \frac{p(x^n|\theta_0)}{m(X^n)} > \alpha\right) \\ &\qquad \leq O\left(\frac{1}{n}\right) + P_{\theta_0}\left(\sup_{\theta \in B^c} \frac{p(x^n|\theta)}{p(x^n|\theta_0)}\, e^{nr'} > \alpha\right) \qquad \text{(A.11)} \end{aligned}$$

by intersecting with the event $\{p_{\theta_0}/m_n < e^{nr'}\}$ and its complement. Because the second term is less than ε for $r' \leq r$, (A.9) is established.

APPENDIX B: PROOFS OF PROPOSITIONS 4.1 AND 4.2

Next we turn to providing a proof that n^* can be bounded above and below by n_u and n_s.

Proof of Proposition 4.1

We start with the first inequality in (13). Suppose that in the sequential method, we stop after having gotten x^k. Then we know that $\operatorname{var}_{w_N}(\Theta|x^k) < \varepsilon_2$ so that $W_N(\theta|x^k) \in \Gamma(\varepsilon_1, \varepsilon_2)$. Thus $w_N(\theta|x^k)$ is one of the choices of w used to obtain n^*. Indeed,

$$D(w_N(\cdot|x^n)\|w_N(\cdot|x^k))|_{x^n = x^k} = 0$$

and

$$\operatorname{argmin} D(w_N(\cdot|x^n)\|w_N(\cdot|x^k)) = k.$$

Because n^* is the supremum over the argmins, we have that $n^* \geq n_s$ as in (13).

For the second inequality in (13), note that the sample size $n_u(\varepsilon_1)$ is the least value of n so that

$$\sup_{x^{n'}, n' \geq n} \operatorname*{var}_{w_N}(\Theta | x^{n'}) < \varepsilon_1$$

and

$$n^*_{\Gamma(\varepsilon_1 \varepsilon_2)} = \sup\{n_w | \operatorname*{argmin}_{w \in \Gamma(\varepsilon_1 \varepsilon_2)} D(w_N(\cdot|x^n) \| w(\cdot)) = n_w\}.$$

Now for $w \in \Gamma_{\varepsilon_1 \varepsilon_2}$ there is an x^k, so that $w(\theta) = w_N(\theta|x^k)$ and hence

$$n_w = \operatorname*{argmin}_{w \in \Gamma_{\varepsilon_1 \varepsilon_2}} D(w_N(\cdot|x^n) \| w(\cdot)) = k.$$

Because $\text{var}_w(\Theta) = \text{var}_{w_N}(\Theta|x^k) \in (\varepsilon_1, \varepsilon_2)$, we have that $n_w \leq n_u(\varepsilon_1)$, and thus the second inequality in (13) follows.

Finally, we prove the result giving bounds on the upcrossings of the posterior variance.

Proof of Proposition 4.2

First, note that the density ratio $P_\theta(x^n)/m(x^n)$ is an m-martingale. Indeed,

$$\begin{aligned} E\left(\frac{p_\theta(x^n)}{m(x^n)} \middle| X^{n-1} \right) &= \int \frac{p_\theta(x^n)}{m(x^n)} m(x^n | x^{n-1}) \, dx_n \\ &= p_\theta(x^{n-1})/m(x^{n-1}). \end{aligned}$$

Similarly, $\theta^k w(\theta) p(x^n|\theta)/m(X^n)$ is an m-martingale for any k, as is $E(\theta^k|X^n)$:

$$\begin{aligned} &E(E(\Theta^k|X^n)|X^{n-1}) \\ &\quad = \int_{\mathcal{X}} \int_\Omega \theta^k \frac{p_\theta(x^n)}{m(x^n)} w(\theta) \, d\theta m(x_n|X^{n-1}) \, dx_n \\ &\quad = \int_\Omega \int_{\mathcal{X}} \frac{p_\theta(x^n)}{m(x^n)} m(x_n|X^{n-1}) \, dx_n \theta^k w(\theta) \, d\theta \\ &\quad = E(\Theta^k|X^{n-1}). \end{aligned}$$

The posterior variance is $E(\Theta^2|X^n) - (E(\Theta|X^n))^2$, in which the first term is a martingale. Because the function x^2 is convex and

$$E_m(E(\Theta|x^n)^2) \leq E_m E(\Theta^2|x^n) = \int \theta^2 w(\theta) \, d\theta < \infty,$$

theorem 35.1 of Billingsley (1986, p. 487) implies that $E(\theta|X^n)^2$ is a submartingale. So, $-\text{var}(\Theta|X^n)$ is the difference between a submartingale and a martingale; that is, the posterior variance $-\text{var}(\Theta|X^n)$ is a submartingale also.

Let ν be the number of upcrossings of $[-b, -a]$ by $-\text{var}(\Theta|X^n)$; then, by the upcrossing inequality (see Hall and Heyde 1980, p. 15),

$$(-a + b)E\nu \leq E(-\text{var}(\Theta|X^n) + b)^+ - E(-\text{var}(\Theta|X_1) + b).$$

Note that ν is also the number of downcrossings of $[a, b]$ by $\text{var}(\Theta|X^n)$ and $\mu \leq \nu + 1$. Consequently,

$$\begin{aligned} E\mu &\leq 1 + \frac{E(-\text{var}(\Theta|X^n) + b)^+ - E(-\text{var}(\Theta|X_1) + b)^+}{b - a} \\ &\leq 1 + \frac{E(-\text{var}(\Theta|X^n) + b)^+}{(b - a)} \\ &\leq 1 + \frac{b}{b - a} + \frac{E_m \text{var}(\Theta|X^n)}{b - a}. \end{aligned}$$

To complete the proof, we show that as $n \to \infty$, the third term goes to zero. First, note that

$$\begin{aligned} &E_m \text{var}(\Theta|X^n)^2 \\ &\quad = E_m \left(\int \theta^2 w(\theta|X^n) \, d\theta - \left(\int \theta w(\theta|X^n) \, d\theta \right)^2 \right)^2 \\ &\quad \leq 2E_m \left(\int \theta^2 w(\theta|X^n) \, d\theta \right)^2 + 2E_m \left(\int \theta w(\theta|X^n) \, d\theta \right)^4 \\ &\quad \leq 4E_m \int \theta^4 w(\theta|X^n) \, d\theta = 4 \int \theta^4 w(\theta) \, d\theta, \end{aligned}$$

so that $\sup_n E_m \text{var}(\Theta|X^n)^2 < \infty$. Recall that for any p_θ, $n \text{ var}(\Theta|X^n) \xrightarrow{P_\theta} \mathbf{I}(\theta)^{-1}$. (This is standard and straightforward to prove under the stated hypotheses; see, for instance, Clarke 1989, p. 71). So, $\text{var}(\Theta|X^n) \xrightarrow{P_\theta} 0$. For any $\varepsilon > 0$, we have

$$M(\text{var}(\theta|X^n) > \varepsilon) = \int w(\theta) P_\theta(\text{var}(\theta|X^n) > \varepsilon) \, d\theta,$$

in which the probability in the integrand goes to zero pointwise as $n \to \infty$. By the bounded convergence theorem, the left side also goes to zero. Thus $\text{var}(\Theta|X^n) \to 0$ in M_n probability. So, by the corollary to theorem 25.12 of Billingsley (1986, p. 348), we have that $E \text{ var}(\Theta|X^n) \to 0$, and the proposition is proved.

[Received March 1994. Revised May 1995.]

REFERENCES

Berger, J. O., and Bernardo, J. M. (1989), "Estimating a Product of Means: Bayesian Analysis With Reference Priors," *Journal of the American Statistical Association*, 84, 200–207.

——— (1992), "On the Development of Reference Priors," in *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Clarendon Press.

Bernardo, J. M. (1979), "Reference Posterior Distributions for Bayesian Inference" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 41, 113–147.

Bickel, P. J., and Doksum, K. A. (1977), *Mathematical Statistics*, Oakland, CA: Holden-Day.

Billingsley, P. J. (1986), *Probability and Measure*, New York: John Wiley.

Clarke, B. (1989), "Asymptotic Cumulative Risk and Bayes Risk, Under Entropy Loss, With Applications," Ph.D. dissertation, University of Illinois, Dept. of Statistics.

Clarke, B., and Barron, A. R. (1990), "Information-Theoretic Asymptotics of Bayes Methods," *IEEE Transactions on Information Theory*, 36, 453–471.

——— (1994), "Jeffreys's Prior is Asymptotically Least Favorable Under Entropy Risk," *Journal of Statistical Planning and Inference*, 41, 37–60.

Csiszar, I. (1967), "Information-Type Measures of Difference of Probability Distributions and Indirect Observations," *Studia Scientiarum Mathematicarum Hungarica*, 2, 299–318.

Datta, G. S., and Ghosh, M. (1994a), "Some Remarks on Noninformative Priors," Technical Report 93-16, University of Georgia, Dept. of Statistics.

——— (1994b), "On the Invariance of Noninformative Priors," Technical Report 93-17, University of Georgia, Dept. of Statistics.

George, E. I., and McCulloch, R. (1992), "On Obtaining Invariant Prior Distributions," Technical Report 73, University of Chicago, Graduate School of Business.

Ghosh, J. K., and Mukerjee, R. (1992), "Noninformative Prior," in *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Clarendon Press.

Hall, P., and Heyde, C. C. (1980), *Martingale Limit Theory and Its Application*, San Diego: Academic Press.

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

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

Kass, R. E., and Wasserman, L. A. (1993), "Formal Rules for Selecting Prior Distributions: A Review and Annotated Bibliography," Technical Report 583, Carnegie-Mellon University, Dept. of Statistics.

Soofi, E. S. (1994), "Capturing the Intangible Concept of Information," *Journal of the American Statistical Association*, 89, 1243–1254.

Soofi, E. S., Ebrahimi, N., and Habibullah, M. (1995), "Information Distinguishability With Applications to Analysis of Failure Data," *Journal of the American Statistical Association*, 90, 657–668.

Wald, A. (1949), "Note on the Consistency of the Maximum Likelihood Estimate," *Annals of Mathematical Statistics*, 595–601.

Wolfowitz, J. (1949), "On Wald's Proof of the Consistency of the Maximum Likelihood Estimate," *Annals of Mathematical Statistics*, 601–602.