Noninformative Priors and Nuisance Parameters

Bertrand CLARKE and Larry WASSERMAN*

We study the conflict between priors that are noninformative for a parameter of interest versus priors that are noninformative for the whole parameter. Our investigation leads us to maximize a functional that has two terms: an asymptotic approximation to a standardized expected Kullback-Leibler distance between the marginal prior and marginal posterior for a parameter of interest, and a penalty term measuring the distance of the prior from the Jeffreys prior. A positive constant multiplying the second terms determines the tradeoff between noninformativity for the parameter of interest and noninformativity for the entire parameter. As the constant increases, the prior tends to the Jeffreys prior. When the constant tends to 0, the prior becomes degenerate except in special cases. This prior does not have a closed-form solution, but we present a simple, numerical algorithm for finding the prior. We compare this prior to the Berger–Bernardo prior.

KEY WORDS: Asymptotic information; Reference prior; Tradeoff prior.

1. INTRODUCTION

Consider the distance between a prior density $\pi(\theta)$ for a k-dimensional real parameter θ and the posterior density corresponding to it. A *reference prior* may be defined to be

$$\arg\sup_{\pi\in\Gamma} E(d(\pi(\cdot), \pi(\cdot \mid Y^n))),$$

where d is a measure of distance, Γ is a set of priors, and $y^n = (y_1, \ldots, y_n)$ is an outcome of $Y^n = (Y_1, \ldots, Y_n)$. Here E refers to the expectation over the marginal distribution for Y^n induced by the prior and the model. Choices for d such as Hellinger distance (Jeffreys 1961, sec. 3.10) have been considered. Here we choose a distance that has an information theoretic motivation, the Shannon mutual information. For this choice we write

$$\begin{aligned}\gamma_n(\pi) &= I(\Theta; Y^n)\\ &= \int m(y^n) \int \pi(\theta \mid y^n) \log \frac{\pi(\theta \mid y^n)}{\pi(\theta)}\, d\theta\, dy^n\\ &= E_m K(\pi(\cdot \mid Y^n), \pi(\cdot)),\end{aligned}$$

where K is the Kullback–Leibler distance defined by $K(p, q) = \int p \log(p/q)$ and m is the marginal density for the data defined by $m(y^n) = \int \pi(\theta) f(y^n \mid \theta)\, d\theta$. We assume $f(y^n \mid \theta) = \prod_{i=1}^n f(y_i \mid \theta)$, where $f(\cdot \mid \theta)$ is a parametric family.

It can be proved that asymptotically maximizing $\gamma_n(\pi)$ over prior densities results in Jeffreys's prior (see Clarke and Barron 1990b). But this assumes that all parameters are of interest. More generally, consider $\theta = (\omega, \lambda)$ where $\omega \in \mathbb{R}^{k_1}$ is a parameter of interest and $\lambda \in \mathbb{R}^{k_2}$ is a nuisance parameter. In this case Jeffreys's prior (Jeffreys 1961, sec. 3.10) has been criticized for leading to unacceptable results. An example of this $Y_i \sim N(\theta_i, 1)$ where $Y_1, \ldots, Y_n$ are independent and $\omega = \sum_i \theta_i^2$. The Jeffreys prior for $\theta_1, \ldots, \theta_n$ is a flat prior, but conventional wisdom holds that better inferences result from priors that shrink toward some point. Another example where Jeffreys's prior does not work well is in estimating the product of two normal means (see Berger and Bernardo 1989). When we refer to Jeffreys's prior, we refer to the formal application of Jeffreys's rule. We remind the reader that Jeffreys did not advocate using this rule in all problems.

One possible explanation for this poor behavior of Jeffreys's prior is that is achieves optimal noninformativity for θ at the cost of being partially informative for ω. In Bernardo's (1979) terminology, Jeffreys's prior maximizes the missing information for θ but not for ω. This led Bernardo (1979) and Berger and Bernardo (1989, 1992a, 1992b) to define a reference prior that is essentially a stepwise Jeffreys's prior. Their method seems to give reasonable priors, but the motivation for the method is unclear.

In this article we propose an alternative method of constructing priors. First, we find an asymptotic expression for the Kullback–Leibler distance between the marginal prior and marginal posterior for ω accurate to $o(1)$. We then find the prior that maximizes this distance subject to a penalty term that measures distance from the Jeffreys prior so as to obtain a prior that gives up some of its noninformativity for λ to become more noninformative for ω. The prior depends on a scalar α that reflects the trade-off between being noninformative for λ and being noninformative for ω. So, we obtain a continuum of priors varying between the Jeffreys prior and the marginally noninformative prior. Although these priors cannot be expressed in closed form, we describe a simple algorithm for finding them numerically. We compare our solution to the Berger–Bernardo prior in two examples: the bivariate binomial and the multinomial.

We should say at the outset that by a "noninformative prior" we mean a prior that has, asymptotically, large expected distance from the posterior in a given experiment.

2. BACKGROUND

2.1 The Berger–Bernardo Method

We begin by reviewing the method of Bernardo (1979) and Berger and Bernardo (1992a). Recall that the Jeffreys

* Bertrand Clarke is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907. Larry Wasserman is Associate Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. Wasserman's research was supported by National Science Foundation Grant DMS-9005858 and National Institutes of Health Grant R01-CA54852-01. Part of this research was conducted at the Cornell Workshop on Conditional Inference sponsored by the U.S. Army Mathematical Sciences Institute and the Statistics Research Center (June 3–14, 1991). The authors thank Nick Polson for several helpful discussions and an associate editor and a referee for useful suggestions.

prior is defined by

$$J(\omega, \lambda) = \frac{|\mathbf{I}(\theta)|^{1/2}}{\int |\mathbf{I}(\theta)|^{1/2}\, d\theta},$$

where $\mathbf{I}$ is the Fisher information matrix and $|\cdot|$ indicates the determinant. The Berger–Bernardo method has two steps. First, use the Jeffreys prior for the nuisance parameter conditional on the parameter of interest; that is, set $\pi(\lambda|\omega) \propto |\mathbf{I}_{22}(\omega, \lambda)|^{1/2}/c(\lambda)$, where $\mathbf{I}_{22}$ is the lower right $k_2 \times k_2$ block of $\mathbf{I}$ and $c(\lambda)$ normalizes the density over a compact set. (To use this method, one must truncate the parameter space to a compact set so that the conditional Jeffreys prior is integrable. We also truncate to compact sets as needed). To obtain a joint reference prior, note that when $\pi(\lambda|\omega)$ is as just described, then

$$\mathbf{I}(\Theta; Y^n) = \frac{k_1}{2}\log\frac{n}{2\pi e} + \int \pi(\omega)\log\exp\left\{\frac{1}{2}\int \pi(\lambda|\omega)\log|I|\,|I_{22}|^{-1}\, d\lambda\right\} d\omega - \int \pi(\omega)\log\{\pi(\omega)\}\, d\omega + o(1)$$

under various regularity conditions; see Lemma 3.1. Asymptotically maximizing the standardized distance $I(\Theta; Y^n) - (k_1/2)\log(n/2\pi e)$ over all marginal prior densities for ω gives

$$\pi(\omega) = \exp\left\{\frac{1}{2}\int \pi(\lambda|\omega)\log|I|\,|I_{22}|^{-1}\, d\lambda\right\}\Big/ c,$$

where c is a normalizing constant. The Berger–Bernardo prior is the product $\pi(\omega)\pi(\lambda|\omega)$.

This method achieves noninformativity over the nuisance parameter first and, subject to this, is least informative for the parameter of interest. But the Jeffreys prior achieves maximal noninformativity in an information-theoretic sense. So, using the Jeffreys prior on the nuisance parameter in the first step sacrifices noninformativity where it is needed most—on the parameter of interest.

Examination of the preceding expression for $\mathbf{I}(\Theta; Y^n)$ led Ghosh and Mukerjee (1992) to consider $\int \pi(\omega)\log\{\pi(\omega)\} \times d\omega$ as a "penalty term" effectively forcing shrinkage towards a uniform prior. In Section 4 we use the information-theoretic formulation to motivate a single maximization of a two-term functional instead of a two-step maximization. Note that the Berger–Bernardo method, the Ghosh–Mukerjee method and our method all reduce to the Jeffreys prior when there are no nuisance parameters.

2.2 Noninformativity in the Absence of Nuisance Parameters

Bernardo's method is based on the idea that in the limit for large n, $\gamma_n(\pi)$ represents the missing information for someone using prior π. Thus the prior that maximizes this "missing information" is the one that will be most changed by experimentation. It is in this sense that the prior is least informative.

But Bernardo (1979) noted that $\lim_{n\to\infty}\gamma_n(\pi)$ is typically infinite (see also Davisson 1973; Ibrigamov and H'asminsky 1973). Berger and Bernardo sidestepped this by finding the prior π_n^* that maximizes γ_n and then taking the limit of π_n^* as $n \to \infty$. In contrast, we use a standardized version of $\gamma_n(\pi)$ that has a finite limit. Polson (1988) and Ghosh and Mukerjee (1992) argued in a similar way.

Following Ibrigamov and H'asminsky (1973) and Clarke and Barron (1991), we may write (assuming certain regularity conditions)

$$\gamma_n(\pi) = \frac{k}{2}\log\frac{n}{2\pi e} + \int \pi(\theta)\log\frac{|\mathbf{I}(\theta)|^{1/2}}{\pi(\theta)}\, d\theta + o(1),$$

where k is the dimension of θ. We define the *standardized distance* $\tilde{\gamma}_n(\pi)$ by $\tilde{\gamma}_n(\pi) = \gamma_n(\pi) - (k/2)\log(n/2\pi e)$ and define the *asymptotic missing information* by

$$\tilde{\gamma}_\infty(\pi) \equiv \lim_{n\to\infty}\tilde{\gamma}_n(\pi) = \int \pi(\theta)\log\frac{|\mathbf{I}(\theta)|^{1/2}}{\pi(\theta)}\, d\theta.$$

Provided that $|\mathbf{I}(\theta|^{1/2}$ is integrable, a variational argument shows that $\tilde{\gamma}_\infty(\pi)$ is maximized by choosing $\pi(\theta) \propto |\mathbf{I}(\theta)|^{1/2}$, which, when normalized, is the Jeffreys prior. In a decision-theoretic formulation, Clarke and Barron (1990b) showed that Jeffreys's prior is least favorable under an entropy loss criterion.

3. NUISANCE PARAMETERS

Write $\theta = (\omega, \lambda)$, where ω is the parameter of interest and λ is a nuisance parameter. For a given joint prior $\pi(\omega, \lambda)$, we compute the expected Kullback–Leibler distance between the prior and posterior for the parameter of interest. Let $\gamma_n^\omega(\pi) = E(K(\pi(\omega|y^n), \pi(\omega)))$, where $\pi(\omega|y^n) = \int \pi(\omega, \lambda|y^n)\, d\lambda$ and $\pi(\omega) = \int \pi(\omega, \lambda)\, d\lambda$. Then we have the following lemma.

Lemma 3.1.

$$\gamma_n^\omega(\pi) = \frac{k_1}{2}\log\frac{n}{2\pi e} + \iint \pi(\omega, \lambda)\log\left\{\frac{S}{\pi(\omega)}\right\} d\omega\, d\lambda + o(1),$$

where k_1 is the dimension of ω and $S = \{|\mathbf{I}(\omega, \lambda)| \times |\mathbf{I}_{22}(\omega, \lambda)|^{-1}\}^{1/2}$.

(All proofs are contained in the Appendix.) We point out that the same formula appears, without a complete proof, in Ghosh and Mukerjee (1992).

Now we define the *asymptotic marginal missing information for* ω by $\tilde{\gamma}_\infty^\omega(\pi) = \lim_{n\to\infty}\{\gamma_n^\omega(\pi) - (k_1/2) \times \log(n/2\pi e)\} = \iint \pi(\omega, \lambda)\log\{S/\pi(\omega)\}\, d\omega\, d\lambda$. Because ω is the parameter of interest, we begin by maximizing $\tilde{\gamma}_\infty(\pi)$. This idea was suggested by Ghosh and Mukerjee (1992), who observed that the entropy of $\pi(\omega)$ appears as a penalty term in $\gamma_n^\omega(\pi) - (k_1/2)\log(n/2\pi e)$ that does not involve the nuisance parameter. They suggested that as a result, the maximizing priors would prove unsatisfactory. We verify their intuition in the next lemma.

Lemma 3.2. Suppose that ω and λ are orthogonal; that is, $\mathbf{I}_{12} = \mathbf{I}_{21} = 0$, where $\mathbf{I}_{12} = \mathbf{I}_{21}$ is the upper right $k_1 \times k_2$ block by the Fisher information matrix:

a. If $\mathbf{I}_{11}(\omega, \lambda)$ is a function of ω only, then any prior that satisfies $\pi(\omega) \propto \sqrt{|\mathbf{I}_{11}(\omega)|}$ maximizes $\tilde{\gamma}^{\omega}_{\infty}(\pi)$.
b. If $|\mathbf{I}_{11}| = f(\omega)g(\lambda)$ for some f and g with g not constant, g is continuous, and the parameter space is compact, then any prior that has ω marginal with density proportional to $\sqrt{f(\omega)}$ and has conditional distribution for λ given ω concentrating on the set $g^{-1}(\sup g)$ maximizes $\tilde{\gamma}^{\omega}_{\infty}(\pi)$.

Remark 1. Note that any statistical model may be reparameterized so that the parameters are orthogonal (see Cox and Reid 1987).

Part a of Lemma 3.2 shows that when inferences about the parameters can be separated, Jeffreys's prior maximizes $\tilde{\gamma}^{\omega}_{\infty}$. More generally, there is a class of priors that maximizes this quantity. In particular, Tibshirani's (1989) prior maximizes $\tilde{\gamma}^{\omega}_{\infty}(\pi)$. In Example 5.1 we will see that unlike the method we will present, the Berger–Bernardo prior does not reproduce the Jeffreys prior in this situation.

Part b of Lemma 3.2 shows that maximizing $\tilde{\gamma}^{\omega}_{\infty}(\pi)$ can lead to degenerate priors that effectively assume that the nuisance parameter is known. The reason for this behavior is that the criterion achieves noninformativity for the parameter of interest at the cost of exact knowledge for the nuisance parameter.

4. AN ALTERNATIVE METHOD

The Jeffreys prior and the degenerate priors presented in the preceding section are two extreme cases. The first does not distinguish between nuisance parameters and parameters of interest; the second assumes that the nuisance parameter is known and hence not a nuisance. To interpolate between these two extremes, we define a functional that measures distance between the marginal prior and the marginal posterior subject to a penalty term that measures the distance from the Jeffreys prior. We then find the prior that maximizes this functional. Our method assumes that it is generally impossible to achieve simultaneously maximal noninformativity for the whole parameter and the parameter of interest.

Definition. The tradeoff prior for ω is the prior π_α that maximizes $\tilde{\gamma}^{\omega}_{\infty} - \alpha K(J, \pi_\alpha)$, where $K(\cdot, \cdot)$ is Kullback–Leibler distance, $J = J(\omega, \lambda)$ is Jeffreys's prior, and $\alpha > 0$.

Remark 2. If the conditions of Lemma 3.2.a hold and $\mathbf{I}_{22}$ is only a function of λ, then it is straightforward to show that π_α is Jeffreys's prior for every α.

The parameter α reflects the relative importance of the nuisance parameters. The penalty term forces shrinkage towards the Jeffreys prior. When $\alpha = 0$, this prior is the degenerate prior in Lemma 3.2.b. As α increases, the penalty term dominates, and the prior becomes noninformative for the whole parameter. In practice, α must be chosen to reflect the trade-off between these two extremes. We note that Ghosh and Mukerjee (1992) suggested maximizing $\tilde{\gamma}^{\omega}_{\infty} + \alpha H(\pi)$, where $H(\pi)$ is the entropy of π. But this prior shrinks towards the uniform prior instead of the Jeffreys prior as $\alpha \to \infty$.

Theorem 4.1. Assume that π_α is bounded away from 0. Then the prior π_α satisfies

$$\pi_\alpha(\omega, \lambda) = -\frac{\alpha J(\omega, \lambda)}{\log(S/\pi_\alpha(\omega)) + C_\alpha},$$

where C_α is the unique constant determined by the fact that π_α is positive and integrates to 1.

Corollary. $\lim_{\alpha \to \infty} \pi_\alpha = J$ as long as the limit exists.

Theorem 4.2. The trade-off prior is invariant under smooth monotone transformations of ω and λ.

Remark 3. In general, the trade-off prior, like the Berger–Bernardo prior, is not invariant under transformations that involve both ω and λ. This is to be expected, because ω is being singled out as a parameter of interest.

Remark 4. In current work we have been able to show that if instead the penalty $K(\pi_\alpha, J)$ is used, then the maximizing prior is

$$\pi_\alpha(\omega, \lambda) = \frac{S(\omega, \lambda)^{1/\alpha} J(\omega, \lambda)}{W_\alpha(\omega)^{1/(\alpha+1)} \int W_\alpha(\omega)^{\alpha/(\alpha+1)}\, d\omega},$$

where

$$W_\alpha(\omega) = \int S(\omega, \lambda)^{1/\alpha} J(\omega, \lambda)\, d\lambda.$$

If S is a function of ω only, then this prior reduces to the Berger–Bernardo prior when α tends to 0. Otherwise, it tends to a degenerate prior with density equal to $S(\omega, \lambda_\omega)/\int S(\omega, \lambda_\omega)\, d\omega$ on the manifold $\{(\omega, \lambda_\omega); \omega \in \mathbb{R}^{k_1}\}$. Here λ_ω is the point where $S(\omega, \lambda)$ is maximized for fixed ω.

The form of the solution in Theorem 4.1 suggests the following algorithm for finding π_α.

Algorithm.

Step 0. Choose $\pi^0(\omega, \lambda)$. Set $\pi^0(\omega) = \int \pi^0(\omega, \lambda)\, d\lambda$. Let $i = 1$. Repeat the next three steps until convergence.

Step 1. Find C such that $\int Z^i_C = 1$, where

$$Z^i_C = -\frac{\alpha J(\omega, \lambda)}{\log(S/\pi^{i-1}_\alpha(\omega)) + C}.$$

Step 2. Set $\pi^i_\alpha(\omega, \lambda) = Z^i_C$.

Step 3. Set $\pi^i(\omega) = \int \pi^i(\omega, \lambda)\, d\lambda$. Let $i = i + 1$.

The algorithm is used for the examples in the next section.

5. EXAMPLES

5.1 The Bivariate Binomial

Consider the following model for the germination of spores. Each of m spores has a probability p of germinating. Of the r spores that germinate, each has a probability q of bending in a particular direction. Let s be the number that bend in the specified direction. The probability model is

$$f(r, s|p, q, m) = \binom{m}{r} p^r (1-p)^{m-r} \binom{r}{s} q^s (1-q)^{r-s}$$

for $r = 1, \ldots, m$ and $s = 1, \ldots, r$. This is called the bivariate binomial model. Crowder and Sweeting (1989) and Polson and Wasserman (1990) have proposed priors for this model.

The Fisher information matrix is

$$\mathbf{I}(p, q) = m\begin{bmatrix} \{p(1-p)\}^{-1} & 0 \\ 0 & p\{q(1-q)\}^{-1} \end{bmatrix}.$$

The Jeffreys prior is $\pi_J(p, q) \propto (1-p)^{-1/2}q^{-1/2} \times (1-q)^{-1/2}$.

Polson and Wasserman (1990) showed that when q is the parameter of interest and p is the nuisance parameter, the Berger–Bernardo prior is $\pi_{BB}(p, q) \propto p^{-1/2}(1-p)^{-1/2} \times q^{-1/2}(1-q)^{-1/2}$. We used the algorithm in Section 4 to find π_α. The result is shown in Figure 1 for several values of α. The degenerate form of π_α when α is near 0 illustrates the effect described in Lemma 3.2.b. The convergence to the overall Jeffreys prior as α increases illustrates the corollary.

5.2 The Trinomial

Suppose that $y = (y_1, y_2, y_3)$, where the y_i's are nonnegative integers, and that

$$p(y|\theta, n) = \frac{n!}{y_1!y_2!y_3!}\theta_1^{y_1}\theta_2^{y_2}\theta_3^{y_3},$$

where $n = y_1 + y_2 + y_3$, $\theta = (\theta_1, \theta_2, \theta_3)$, $\theta_i \geq 0$, $i = 1, 2, 3$, and $\theta_3 = 1 - \theta_1 - \theta_2$. Thus y has a trinomial distribution. Let $\omega = \theta_1$ and $\lambda = \theta_2$. The Jeffreys prior is $J(\theta_1, \theta_2) \propto \theta_1^{-1/2}\theta_2^{-1/2}(1-\theta_1-\theta_2)^{-1/2}$; the Berger–Bernardo prior is $\pi_{BB} \propto \theta_1^{-1/2}(1-\theta_1)^{-1/2}\theta_2^{-1/2}(1-\theta_1-\theta_2)^{-1/2}$ (Berger and Bernardo 1991b). Figure 2 shows the ω marginal of the trade-off prior for $\alpha = 3$ and 5. Note that when $\alpha = 5$, the trade-off prior appears to be similar to the Berger–Bernardo.

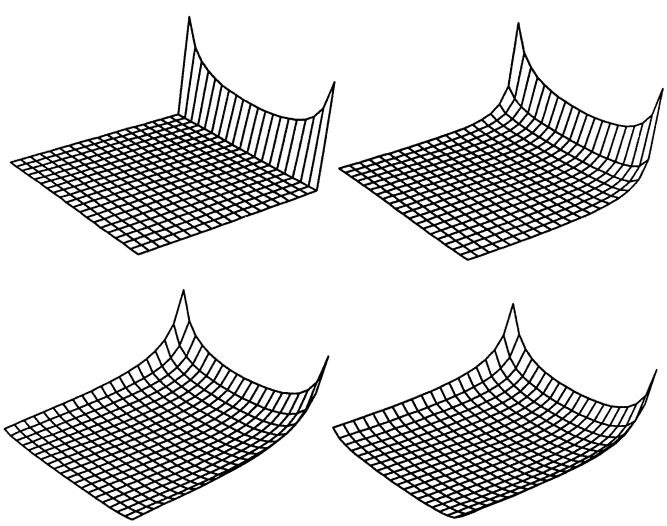

Figure 1. A Sequence of Plots Illustrating the Tradeoff Prior for the Bivariate Binomial Problem When q is the Parameter of Interest. The values of α are .001, .1, 1.0. The last plot shows the Jeffreys prior. When α = .001, the prior concentrates on the line p = 1. When α = 1, the prior is nearly equal to the Jeffreys prior.

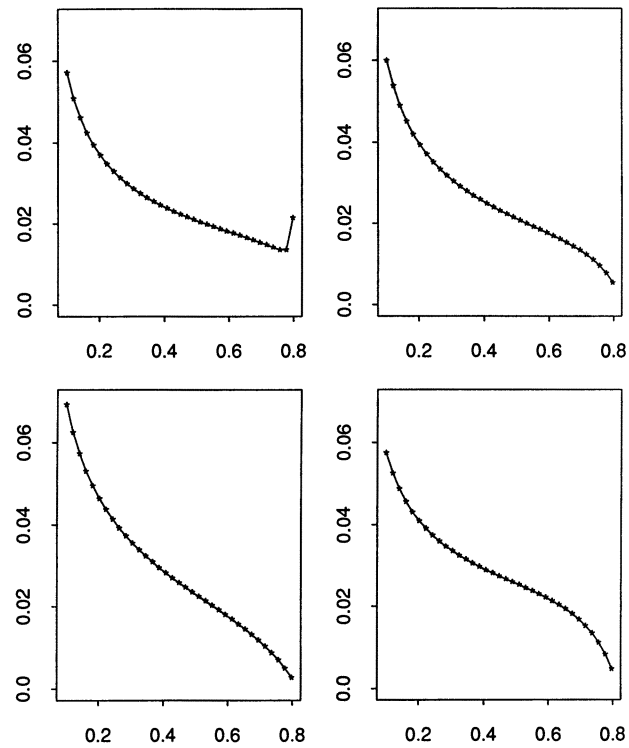

Figure 2. The Marginal of the Tradeoff Prior for the Trinomial Problem When p_1 is the Parameter of Interest. The values of α are 3 and 5. The penultimate plot is the Jeffreys prior. The last plot is the Berger–Bernardo prior.

6. DISCUSSION

There is a conflict between being noninformative for a parameter of interest and being noninformative for the entire parameter. This article's main contribution is to make this conflict explicit. Finding priors to achieve either goal leads to extreme solutions. Our method provides a compromise between these two extremes.

The usual method for constructing priors to account for the role of parameters of interest, the Berger–Bernardo method, involves the construction of a stepwise prior. Our method defines the prior via the maximization of a functional that has a simple interpretation. In this sense, our prior has a clearer meaning than the Berger–Bernardo prior. The advantage of the Berger–Bernardo prior is that there is no need to choose the trade-off parameter α.

In principle, α should be chosen to represent the relative importance of the parameter of interest. In practice, one can compute the trade-off prior for several values of α and then choose a prior that is approximately midway between the degenerate prior and the Jeffreys prior. This entails some subjectivity, which may seem to defeat the process of constructing automatic priors. But the choice of prior has been reduced to a search through a one-parameter family. The associate editor suggested that α be treated as a hyperparameter and then a prior placed on α. This is an interesting idea, but the prior will have to be chosen carefully, because it will have a large effect on the analysis, especially in cases where π_α becomes degenerate at $\alpha = 0$. Perhaps a prior could be calibrated in terms of the Kullback–Leibler distance of π_α

from the Jeffreys prior. This is in the same spirit as Jeffreys (1961, p. 275), who used a similar idea to construct priors for the alternative hypothesis in Bayesian testing. If we followed his approach, then we would use a prior $\pi(\alpha)$ that induces a uniform prior on $\tan^{-1}W^{1/2}$, where W is $K(\pi_\alpha, J) + K(J, \pi_\alpha)$. Another possibility is to use a prior of the form $\pi(\alpha) \propto \beta \exp(-\beta K(\pi_\alpha, J))$. For that matter, we might choose α to make π_α a given distance δ from J, where δ depends on, say, dimension (θ) − dimension (ω). All of these ideas deserve further investigation. For now, we suggest using a range of values of α. Good Bayesian statistical analyses examine the sensitivity to the choice of prior by using several alternative priors. In our method, this amounts to using several values of α.

The main consequence of our results for real problems is that it is now apparent that any attempt to construct noninformative priors for a particular parameter entails an increase in informativity for other parameters. Thus there is a hidden cost to tailoring a prior to a specific parameter. When interest focuses on one parameter in a high-dimensional model, the trade-off is likely to be large. In real problems, there is rarely a single parameter of interest, and this means that tradeoff must be assessed for several parameters. Further experience with real problems is needed to determine when the effects of adjusting noninformative priors will have a major impact on the analysis. In future work we hope to develop a diagnostic, perhaps based on $\tilde{\gamma}_\infty^\omega$, to determine when modifications of priors are needed to reduce the influence of the prior. Of course, this will depend on the sample size and the dimension of the model.

APPENDIX: PROOFS OF THEOREMS

Proof of Lemma 3.1. Write $y = y^n$ and let $m^\omega(y) = \int f(y|\omega, \lambda) \times \pi(\lambda|\omega)\, d\lambda$. Now, using standard manipulations, $\gamma_n^\omega(\pi) = -\int\int \pi(\omega, \lambda) K(f(y|\theta), m^\omega(y))\, d\lambda\, d\omega + \int\int \pi(\omega, \lambda) \times K(f(y|\theta), m(y))\, d\lambda\, d\omega$. From Clarke and Barron (1991),

$$K(f(y|\theta), m^\omega(y)) = \frac{k_1}{2}\log(n/2\pi e) + (1/2)\log|\mathbf{I}_{22}(\omega, \lambda)| - \log(\pi(\lambda|\omega)) + o(1)$$

and

$$K(f(y|\theta), m(y)) = \frac{k_1 + k_2}{2}\log(n/2\pi e) + (1/2)\log|\mathbf{I}(\omega, \lambda)| - \log(\pi(\omega, \lambda)) + o(1).$$

Substituting above, we get

$$\begin{aligned}\gamma_n^\omega(\pi) &= \frac{k_1}{2}\log\frac{n}{2\pi e} + (1/2)\int\int \pi(\omega, \lambda) \times \log|\mathbf{I}(\omega, \lambda)|\,|\mathbf{I}_{22}(\omega, \lambda)|^{-1}\, d\omega\, d\lambda \\ &\quad + \int\int \pi(\omega, \lambda)\log\frac{\pi(\lambda|\omega)}{\pi(\lambda, \omega)}\, d\omega\, d\lambda + o(1) \\ &= \frac{k_1}{2}\log\frac{n}{2\pi e} + \int\int \pi(\omega, \lambda) \times \log\frac{\sqrt{|\mathbf{I}(\omega, \lambda)|\,|\mathbf{I}_{22}|^{-1}(\omega, \lambda)}}{\pi(\omega)}\, d\omega\, d\lambda + o(1).\end{aligned}$$

Proof of Lemma 3.2. Due to the orthogonality, $S = \sqrt{\mathbf{I}_{11}(\omega, \lambda)}$. To prove Part a, it is enough to note that

$$\tilde{\gamma}_\infty^\omega = \int \pi(\omega)\log\frac{S}{\pi(\omega)}\, d\omega.$$

For Part b, we have that

$$\begin{aligned}\tilde{\gamma}_\infty^\omega &= \int\int \pi(\omega)\log\frac{\sqrt{f(\omega)g(\lambda)}}{\pi(\omega)}\, d\omega\, d\lambda \\ &= \int \pi(\omega)\log\frac{\sqrt{f(\omega)}}{\pi(\omega)}\, d\omega + \int \pi(\lambda)\log\sqrt{g(\lambda)}\, d\lambda.\end{aligned}$$

Choosing the prior to have marginal density for ω proportional to $\sqrt{f(\omega)}$ maximizes the first term, and choosing the prior to have conditional distribution concentrated on $g^{-1}(\sup g)$ maximizes the second term.

Proof of Theorem 4.1. We apply a standard calculus of variations argument. Let $M = \{\delta : \Theta \to \mathbb{R};\ \sup|\delta| \le 1$ and $\int\int \delta(\omega, \lambda)\, d\omega\, d\lambda = 0\}$. For sufficiently small $\varepsilon > 0$, define

$$\begin{aligned}F(\pi, \varepsilon) &= \int\int (\pi(\omega, \lambda) + \varepsilon\delta(\omega, \lambda)) \\ &\quad \times \log\left\{\frac{S}{\int (\pi(\omega, \lambda') + \varepsilon\delta(\omega, \lambda'))\, d\lambda'}\right\} d\omega\, d\lambda \\ &\quad - \alpha\int\int J(\omega, \lambda)\log\frac{J(\omega, \lambda)}{(\pi(\omega, \lambda) + \varepsilon\delta(\omega, \lambda))}\, d\omega\, d\lambda.\end{aligned}$$

Then a necessary condition for π_α to maximize the criterion is that $(dF(\pi_\alpha, \varepsilon)/d\varepsilon)_{\varepsilon=0} = 0$ for every $\delta \in M$. Let $\pi_\alpha(\omega) = \int (\pi_\alpha(\omega, \lambda)\, d\lambda$ and $\delta(\omega) = \int \delta(\omega, \lambda)\, d\lambda$. Now,

$$\left(\frac{dF(\pi_\alpha, \varepsilon)}{d\varepsilon}\right)_{\varepsilon=0} = \int\int \delta(\omega, \lambda)\left(\log\frac{S}{\pi_\alpha(\omega)} + \alpha\frac{J(\omega, \lambda)}{\pi_\alpha(\omega, \lambda)}\right) d\omega\, d\lambda.$$

If this equals 0 for all $\delta \in M$, then for every constant C,

$$0 = \int\int \delta(\omega, \lambda)\left(\log\frac{S}{\pi_\alpha(\omega)} + \alpha\frac{J(\omega, \lambda)}{\pi_\alpha(\omega, \lambda)} + C\right) d\omega\, d\lambda.$$

Because $\delta \in M$ is arbitrary, we have that

$$\left(\log\frac{S}{\pi_\alpha(\omega)} + \alpha\frac{J(\omega, \lambda)}{\pi_\alpha(\omega, \lambda)} + C\right) = 0.$$

On rearrangement, this is the desired result. We see that π_α is a maximum by noting that

$$\left(\frac{d^2F(\pi_\alpha, \varepsilon)}{d\varepsilon^2}\right)_{\varepsilon=0} = -\int (\delta(\omega))^2 \Big/ \pi(\omega)\, d\omega - \alpha\int\int J(\omega, \lambda)\left(\frac{\delta(\omega, \lambda)}{\pi(\omega, \lambda)}\right)^2 d\omega\, d\lambda,$$

which is strictly negative. This defines a class of solutions to the optimization indexed by C. To identify the unique probability density function in this class we proceed as follows. Define

$$Z_{\omega,\lambda}(c) = \frac{J(\omega, \lambda)}{(1/\alpha)L(\omega, \lambda) + c},$$

where $L = \log(\pi_\alpha(\omega)/S)$. Now $Z_{\omega,\lambda}(c)$ is positive and is strictly decreasing for $c > -(1/\alpha)\inf L$. Furthermore, $\int Z_{\omega,\lambda}(c)$ increases without bound as c decreases to $-(1/\alpha)\inf L$ and tends to 0 as c increases. Consequently, there is a unique value for which it equals 1.

Proof of Corollary. Observe that $\pi_\alpha = J/[(1/\alpha)L + c_\alpha]$. Thus, $\lim_{\alpha\to\infty}\pi_\alpha = J/\lim_{\alpha\to\infty} c_\alpha$, and the denominator equals 1 because the left side integrates to 1.

Proof of Theorem 4.2. It suffices to show that the criterion functional is invariant. The second term is known to be invariant,

so we need only show that $\tilde{\gamma}^{\omega}_{\infty}(\pi)$ is invariant. We will prove this for ω and λ both scalar; the proof in higher dimensions is similar. Consider a transformation $(\omega, \lambda) \to (\beta, \sigma)$, where $\beta = g(\omega)$ and $\sigma = G(\lambda)$. Let $h = g^{-1}$ and $H = G^{-1}$. The Jacobian of the transformation is $\Lambda = \text{diag}(h'(\beta), H'(\sigma))$, so $p(\beta, \sigma) = \pi(h(\beta), H(\sigma))h'(\beta)H'(\sigma)$. Because $\mathbf{I}(\beta, \sigma) = \Lambda\mathbf{I}(\omega, \lambda)\Lambda^t$, we have $|\mathbf{I}(\beta, \sigma)| = \{h'(\beta)H'(\sigma)\}^2|\mathbf{I}(\omega, \lambda)|$ and $\mathbf{I}_{22}(\beta, \sigma) = \{H'(\sigma)\}^2\mathbf{I}_{22}(\omega, \lambda)$. Thus $S(\beta, \sigma) = h'(\beta)S(\omega, \lambda)$. Now

$$\begin{aligned}
\tilde{\gamma}^{\beta}_{\infty}(\pi) &= \int\int p(\beta, \sigma)\log\frac{S(\beta, \sigma)}{p(\beta)}\,d\beta\,d\sigma \\
&= \int\int \pi(h(\beta), H(\sigma))h'(\beta)H'(\sigma) \\
&\quad \times \log\frac{h'(\beta)S(h(\beta), H(\sigma))}{\pi(h(\beta))h'(\beta)}\,d\beta\,d\sigma \\
&= \int\int \pi(\omega, \lambda)h'(g(\omega))H'(G(\lambda)) \\
&\quad \times \log\frac{S(\omega, \lambda)}{\pi(\omega)}\,g'(\omega)G'(\lambda)\,d\omega\,d\lambda \\
&= \int\int \pi(\omega, \lambda)\log\frac{S(\omega, \lambda)}{\pi(\omega)}\,d\omega\,d\lambda = \tilde{\gamma}^{\omega}_{\infty}(\pi).
\end{aligned}$$

[Received October 1991. Revised July 1992.]

REFERENCES

Berger, J., and Bernardo, J. (1989), "Estimating the Product of Means: Bayesian Analysis With Reference Priors," *J. Amer. Statist. Assoc.,* 84, 200–207.

——— (1992a), "On the Development of the Reference Prior Method," in *Bayesian Statistics 4,* eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press, pp. 35–60.

——— (1992b), "Ordered Group Reference Priors With Application to the Multinomial Problem," *Biometrika,* 79, 25–38.

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

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

——— (1990b), "Bayes and Minimax Asymptotics of Entropy Risk," Technical Report 90-11, Purdue University, Dept. of Statistics.

——— (1991), "Entropy Risk and the Bayesian Central Limit Theorem," technical report, Purdue University, Dept. of Statistics.

Cox, D. R., and Reid, N. (1987), "Parameter Orthogonality and Approximate Conditional Inference," *Journal of the Royal Statistical Society,* Ser. B, 49, 1–18.

Crowder, M., and Sweeting, T. (1989), "Bayesian Inference for a Bivariate Binomial," *Biometrika,* 76, 599–604.

Ghosh, J. K., and Mukerjee, R. (1992), "Noninformative Priors," in *Bayesian Statistics 4,* ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press, pp. 195–210.

Ibrigamov, I. A., and H'asminsky, R. Z. (1973), "On the Information Contained in a Sample About a Parameter," *Second International Symposium on Information Theory,* pp. 295–309.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.). Oxford, U.K.: Clarendon Press.

Polson, N. (1988), "Bayesian Perspectives on Statistical Modeling," unpublished Ph.D. dissertation, University of Nottingham, Dept. of Mathematics.

Polson, N., and Wasserman, L. (1990), "Prior Distributions for the Bivariate Binomial," *Biometrika,* 77, 901–904.

Tibshirani, R. (1989), "Noninformative Priors for One Parameter of Many," *Biometrika,* 76, 604–608.