INFORMATION TRADEOFF

BERTRAND CLARKE
Department of Statistics, University of British Columbia
2021 West Mall, Vancouver, B. C., Canada V6T 1Z2
LARRY WASSERMAN
Department of Statistics, Carnegie Mellon University, Pittsburgh,
Pennsylvania 15213, U.S.A.

A prior may be noninformative for one parameter at the cost of being informative for another parameter. This leads to the idea of tradeoff priors: priors that give up noninformativity for some parameters to achieve noninformativity for others. We propose a general framework where priors are selected by optimizing a functional with two components. The first component formalizes the requirement that the optimal prior be noninformative for the parameter of interest. The second component is a penalty term that forces the optimizing prior to be close to some target prior. Optimizing such a functional results in a parameterized family of priors from which a specific prior may be selected as the tradeoff prior. An important particular example of such functionals is provided by choosing the first term to be the marginal missing information for the parameter of interest (generalizing Bernardo's notion of missing information) and the second term to be the relative entropy between the unknown prior and the Jeffreys prior. In this case we find a closed form expression for the tradeoff prior and we make explicit connections with the Berger-Bernardo prior. In particular, we show that under certain conditions, the Berger-Bernardo prior and the Jeffreys prior are special cases of the tradeoff prior. We consider several examples.

Some key words: Asymptotic information; Noninformative priors; Nuisance parameters.

1. INTRODUCTION

The most common method for constructing noninformative priors is due to Jeffreys (1961). Although this method works well in the absence of nuisance parameters, some authors have argued that Jeffreys prior leads to unacceptable inferences if nuisance parameters are present. There are at least three general methods for constructing noninformative priors that explicitly take account of nuisance parameters. The first is the Berger-Bernardo method (Bernardo 1979 and Berger and Bernardo 1989) which uses $J_\omega(\lambda)$ as a prior for λ given ω, where $J_\omega(\lambda)$ is the Jeffreys prior on the nuisance parameter λ for fixed values of the parameter of interest ω. Then a marginal model for the data that depends only on ω is found by integrating out λ from the model. The prior they advocate is the product of the reference prior from this marginal model and $J_\omega(\lambda)$. The second method is due to Tibshirani (1989) who found priors that match the probability of posterior credible regions to coverage probability, asymptotically. This method has been refined by Mukerjee and Dey (1992). Recently, Ghosh and Mukerjee (1992) and Clarke and Wasserman (1993) suggested a third method based on finding priors that optimize certain objective functions. In this paper we continue the investigation begun in Clarke and Wasserman (1993).

We define a functional that has two terms. The first is a generalization of Bernardo's idea of the missing information. The generalization involves defining the notion of missing information for one parameter. The second term is a penalty that ensures the optimal prior is not too far from a target prior. In particular, the penalty term stops the optimizing prior from being degenerate. For instance, in the case developed below we use the Jeffreys prior. The result is a prior that is noninformative for a parameter of interest but that is not too far from the Jeffreys prior which is noninformative for the whole parameter. However, any target prior could be used. A subjectively chosen prior would work as well. A scalar factor α in the second term controls the tradeoff between noninformativity for the parameter of interest (as formalized in the first term) and noninformativity for the nuisance parameter (as formalized in the second term).

Other functionals which may be considered to be special cases of the general formulation described here include those considered by Ghosh and Mukerjee (1992) and Clarke and Wasserman (1993). A particular instance of the general form of functional studied here uses a marginal form of Bernardo's missing information as the first term and uses the relative entropy between the unknown prior and Jeffreys prior in the second. This differs from the functional studied in Clarke and Wasserman (1993) only in the reversal of the arguments of the relative entropy in the second term. This alters the interpretation of the functional and permits us to obtain closed form solutions and to derive useful analytic results.

The outline of the present paper is as follows. In Section 2 we begin by describing a general framework for tradeoff priors. Then we specialize to a particular functional based on the relative entropy. For this case we find the optimal prior as a function of the tradeoff parameter α and examine the dependence of the tradeoff prior on α in the context of a many normal means example. Then we give general conditions under which the tradeoff prior reduces to the Berger-Bernardo prior for certain values of the tradeoff parameter. In Section 3 we find the new priors and compare them to the Berger-Bernardo priors in three more cases: the univariate normal, the nested binomial, and the multinomial. In Section 4 we consider the choice of the scalar α and in Section 5 we discuss the results and briefly remark on another special case based on the Chi-squared distance.

2. TRADEOFF PRIORS

2.1. The General Framework

Given a prior $p(\theta)$ and a model $f(y|\theta)$, let $p(\theta|y_1^n)$ be the posterior density for the parameter θ given data $y_1^n = (y_1, \ldots, y_n)$ where $\theta = (\omega, \lambda)$, ω is the parameter of interest and λ is the nuisance parameter. Assuming independence, the model and prior induce a marginal density $m(y_1^n) = \int \prod_i f(y_i|\theta)p(\theta)d\theta$ for y_1^n. Let d be a measure of distance on probability densities defined on the parameter space. We do not require d to be a metric; in particular we wish to consider choices of d that are asymmetric. Let $d_n^{(1)}$ be the expected value of $d(p_\omega(\cdot|y_1^n), p_\omega(\cdot))$ where $p_\omega(\omega) = \int p(\omega, \lambda)d\lambda$ and $p_\omega(\omega|y_1^n) = \int p(\omega, \lambda|y_1^n)d\lambda$ are the marginals for the parameter of interest. The expectation is with respect to the marginal $m(y_1^n) = \int p(y_1^n|\theta)p(\theta)d\theta$. Let $d^{(1)} = \lim_n d_n^{(1)} - c(n)$ where $c(n)$

is an appropriate standardizing constant. (It is necessary to standardize $d_n^{(1)}$ since it tends to infinity.) This will be the first term in our functional. Now let q be another probability density on the parameter space, chosen subjectively or perhaps by a noninformativity principle that does not include the knowledge that λ is a nuisance parameter. Let $\tilde{d}$ be another measure of distance on densities on the entire parameter space. The tradeoff functional is

$$F(p, \alpha) = d^{(1)}(p(\omega|x^n), p(\omega)) - \alpha \tilde{d}(p, q)$$

We emphasize that d is being used to measure distance between ω densities while $\tilde{d}$ is being used to measure distance between θ densities. Choices for d and $\tilde{d}$ include the relative entropy, the Chi-squared measure of distance, Hellinger's measure of distance all of which are instances of Csiszar f-divergences (Csiszar 1967). Another class of choices for d and $\tilde{d}$ is provided by the power divergence family, see Read and Cressie (1988) for a definition and physical interpretations in certain situations.

Our task is to find the class of priors p_α which optimize the tradeoff functional and then to choose a specific value for α. The result will be called the tradeoff prior. In general, identifying $d^{(1)}$ is difficult. Often it will be a functional at least as complicated as one that has the form

$$\int G(\omega, \lambda, p(\omega, \lambda), \nabla p(\omega, \lambda), \nabla^2 p(\omega, \lambda)) d\omega d\lambda.$$

Functionals of this form have been studied extensively and in many cases the Euler-Lagrange equations have been derived. Typically these are partial differential equations that are difficult to solve; see Elsgolts (1977) for an elementary treatment and Courant and Hilbert (1965) more generally. Note that the difficulties involved have to do with mathematical tractability; they do not reflect on the usefulness of the priors from a statistical perspective.

2.2. Motivation for a Special Case

Here we restrict to the case that both d and $\tilde{d}$ are relative entropy, although in Section 5 we discuss obtaining a functional from use of the Chi-squared distance.

The relative entropy between two densities p and q is $D(p||q) = \int p \log(p/q)$. We denote the expected relative entropy between between the prior $p(\theta)$ and its corresponding posterior $p(\theta|y_1^n)$ by $K(\Theta; Y_1^n) = \int D(p(\cdot|y_1^n)||p(\cdot))m(y_1^n)dy_1^n$. We will assume that the parameter space is compact. This can be achieved by truncating the parameter space if necessary. Ibrigamov and H'asminsky (1973) and Clarke and Barron (1990) show that $K(\Theta; Y_1^n) - c(n) = -D(p||J) + \kappa(J) + o(1)$ where $c(n) = (r/2)\log(n/(2\pi e))$, $r = \text{dimension}(\theta)$, $\kappa(J)$ is a constant and J is Jeffreys' prior – the normalized square root of the determinant of the Fisher information matrix. Following Bernardo (1979) we define the (standardized) missing information to be $i(p) = \lim_{n\to\infty}\{K(\Theta; Y_1^n) - c(n) - \kappa(J)\} = -D(p||J)$. See also Polson (1992). Note that $i(p)$ is maximized over all priors by taking $p = J$, see Clarke and Barron (1994). It is in this sense that Jeffreys' prior is least informative for θ. This does not imply that J is least informative for the parameter of interest ω.

To construct priors that are noninformative for ω we proceed as follows. Define $K(\Omega; Y_1^n) = \int D(p_\omega(\cdot|y_1^n)||p_\omega(\cdot))m(y_1^n)dy_1^n$. Then, as in Ghosh and Mukerjee (1992) and Clarke and Wasserman (1993) we have that

$$K(\Omega; Y_1^n) = \int \int p(\omega, \lambda) \log \frac{S(\omega, \lambda)}{p(\omega)} d\omega d\lambda + d(n) + o(1)$$

where $S(\omega, \lambda) = \{|I| \, |I_{22}|^{-1}\}^{1/2}$, I is the Fisher information matrix for θ, I_{22} is the lower right hand r_2 by r_2 block of I, $d(n) = (r_2/2) \log(n/(2\pi e))$ and r_2 is the dimension of λ. Now define the marginal (standardized) missing information by $i^\omega(p) = \lim_{n \to \infty} \{K(\Omega; Y_1^n) - d(n)\} = \int \int p(\omega, \lambda) \log \frac{S(\omega,\lambda)}{p(\omega)} d\omega d\lambda$. In this case $d^{(1)}$ is the functional $i^\omega(p)$.

The quantity $S = S(\omega, \lambda) = \{|I| \, |I_{22}|^{-1}\}^{1/2}$ admits two interpretations. First, a Fisher information can be regarded as the asymptotic variance of an efficient estimator such as the MLE. Thus $I(\omega, \lambda)^{-1}$ can be regarded as the asymptotic variance of the MLE $\hat{\theta} = (\hat{\omega}, \hat{l}ambda)$ and $I_{2,2}(\omega, \lambda)^{-1}$ can be regarded as the asymptotic variance of the MLE $\hat{\lambda}$ when ω is presumed known. In fact, an efficient estimator (or pseudo-estimator) is used to establish the asymptotic expansions of $K(\Theta, Y^n)$ and $K(\Omega, Y^n)$ to obtain the tradeoff functional. This interpretation is unsatisfactory because it is unrelated to prior densities. A second interpretation is based on regarding S as the product of two Jeffreys priors., one for θ and one for λ assuming ω is known. When I is block diagonal, S is $I_{1,1}$. This interpreation is unsatisfactory because the conditional prior for λ from Jeffreys prior for θ is notin general the same as the Jeffreys prior for λ when ω is known.

Since J can be derived by maximizing the missing information for θ it seems reasonable to find a non-informative prior for ω by maximizing the marginal missing information for ω. As noted in Ghosh and Mukerjee (1992) and Clarke and Wasserman (1993, Lemma 3.2) this leads to degenerate priors. The degeneracy occurs because maximizing $i^{\omega}(p)$ results in a singular prior implying exact knowledge of the nuisance parameter λ, that is the optimal prior concentrates on certain fixed values of λ, see also Theorem 4 below.. As an alternative, maximization of $i^{\omega}(p)$ less a penalty term gives a prior that retains some noninformativity for the nuisance parameters also. Formally, the functional we optimize is $F(p, \alpha) = i^{\omega}(p) - \alpha D(p||J)$ where α controls the tradeoff of noninformativity for ω versus of noninformativity for λ, and the relative entropy between p and J is the term $\tilde{d}$. We call the prior p_{α} that maximizes this $F(p, \alpha)$, the *(relative entropy) tradeoff prior.* The tradeoff parameter α controls the degree to which p_{α} is jointly noninformative for $\theta = (\omega, \lambda)$ (α large)versus the degree to which p_{α} is noninformative for ω only (α small).

One can regard the second term in the tradeoff functional as a penalty which will ensure the existence of reasonable solutions. Other penalty terms have been considered. Ghosh and Mukerjee (1992) used the entropy of p. This gives $p(\omega, \lambda) \propto S(\omega, \lambda)$ when $\alpha = 1$. Clarke and Wasserman (1993) used $D(J||p)$ as a penalty term. In this case no closed form solution is available but an algorithm was given for finding the solution. The change in the penalty here – using $D(p||J)$ of $D(J||p)$ – leads to closed form solutions and permits determination of a precise relationship with the Berger-Bernardo prior.

2.3. The Relative Entropy Tradeoff Prior

We begin by deriving the form of the (relative entropy) tradeoff prior p_α.

THEOREM 1. *For α not equal to zero or one, the prior p_α that maximizes $F(p, \alpha)$ over all priors is given by*

$$p_\alpha(\omega, \lambda) = \frac{S(\omega, \lambda)^{\frac{1}{\alpha}} J(\omega, \lambda)}{Z_\alpha(\omega)^{\frac{1}{\alpha+1}} \int Z_\alpha(\omega)^{\frac{\alpha}{\alpha+1}} d\omega}$$

where

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

REMARK 1: All proofs are in the appendix. The proof of this result does not hold for $\alpha = 0, 1$; this fact which is examined in the results below.

REMARK 2: The tradeoff prior is invariant under transformations of either the parameter of interest or the nuisance parameter. This follows from a similar result in Clarke and Wasserman (1993). Also, it has been shown in Datta and Ghosh (1993, Theorem 3.4).

Example: Many Normal Means

Now we give an example that shows how the tradeoff prior depends on α. Suppose $Y_1, \ldots, Y_n$ are independent and that $Y_i \sim N(\mu_i, 1)$. Let $\omega = r^2 = \sum \mu_i^2$ and set $\lambda = (\lambda_1, \ldots, \lambda_{n-1})$ where

$0 < r < R$, $0 < \lambda_1 \leq \pi$, $0 < \lambda_2 \leq \pi$, …, $0 < \lambda_{n-1} \leq 2\pi$ and

$$\begin{aligned}
\mu_1 &= r \sin\lambda_1 \sin\lambda_2 \cdots \sin\lambda_{n-2} \sin\lambda_{n-1} \\
\mu_2 &= r \sin\lambda_1 \sin\lambda_2 \cdots \sin\lambda_{n-2} \cos\lambda_{n-1} \\
\mu_3 &= r \sin\lambda_1 \sin\lambda_2 \cdots \cos\lambda_{n-2} \\
\mu_{n-1} &= r \sin\lambda_1 \cos\lambda_2 \\
\mu_n &= r \cos\lambda_1 .
\end{aligned}$$

Then, calculations show that

$$J(r, \lambda) \propto r^{p-1} \sin^{p-2}\lambda_1 \sin^{p-3}\lambda_2 \cdots \sin\lambda_{n-2}$$

and $S = 1$. Thus,

$$p_B(r, \lambda) \propto \sin^{p-2}\lambda_1 \sin^{p-3}\lambda_2 \cdots \sin\lambda_{n-2}$$

so that $p_B(\omega) \propto \omega^{-1/2}$.

The trade-off prior turns out to be

$$p_\alpha(r, \lambda) \propto r^{\alpha(n-1)/(\alpha+1)} \sin^{p-2}\lambda_1 \sin^{p-3}\lambda_2 \cdots \sin\lambda_{n-2}$$

so that

$$p_\alpha(\omega) \propto \omega^{(\alpha(n-2)-1)/(2(\alpha+1))}.$$

The two are equal if $\alpha = 0$. Figure 1 shows the tradoff prior for $\alpha = 0, 2, 4$ when $n = 2$ and $R = 1$. The first column of plots shows the marginal for ω while the second column shows the corresponding joint prior for μ_1 and μ_2. For $\alpha = 0$, the prior is noninformative for ω but is clearly far from the joint Jeffrey's prior (which is flat). By the time $\alpha = 4$, the joint prior is not quite close to being flat over the whole space. Note that when α is small, the prior shrinks towards the origin as is generally considered desirable.

2.4. Relation to the Berger-Bernardo Prior

In this section we review the Berger-Bernardo prior and we characterize the relationship between their prior and the tradeoff prior. Berger and Bernardo (1992a) show that their prior p_B is given by the formula

$$p_B = c_B J_\omega(\lambda) \exp\left\{ \int J_\omega(\lambda) \log S(\omega, \lambda) d\lambda \right\}$$

where $c_B > 0$.

THEOREM 2. (Ghosh and Mukerjee, 1992). *The prior p_α that maximizes $i^\omega(p)$ over all priors of the form $p(\omega, \lambda) = p(\omega) J_\omega(\lambda)$ is given by p_B.*

We thus see that the Berger-Bernardo prior maximizes the marginal missing information for ω subject to the condition that missing information is maximized for λ conditional on ω. Note that the joint tradeoff prior emerges from a single optimization whereas the Berger-Bernardo prior results from a two stage process in which one first must choose Jeffreys prior for the nuisance parameter. This implicitly assumes optimal transmission for the nuisance parameter, see Clarke and Barron (1994), which may or may not be valid. Furthermore, use of the penalty term replaces the use of Jeffreys prior on the nuisance parameter.

THEOREM 3. *If $S(\omega, \lambda)$ is a function of ω only, then $p_0 = \lim_{\alpha \downarrow 0} p_\alpha$ exists and is given by $p_0 \propto S(\omega) J_\omega(\lambda)$. Furthermore, $p_0 = p_B$.*

REMARK: This result confirms, in this special case, the intuition that trade-off priors interpolate between the Berger-Bernardo prior at $\alpha = 0$ and the Jeffreys prior at $\alpha = \infty$. We show below in Theorem 4 that this intuition does not fully generalize.

The next theorem shows that if S does depend on λ, then p_0 is degenerate. Let μ be Lebesgue measure, $\hat{s}(\omega) = \sup_\lambda S(\omega, \lambda)$ and $A_\omega = \{\lambda; S(\omega, \lambda) = \hat{s}(\omega)\}$. Usually, A_λ is a singleton set but for the sake of completeness, we also consider the case where A_λ has positive Lebesgue measure.

THEOREM 4. *Assume that S is continuous and bounded.*

Case 1: Suppose $A_\omega = \{\lambda_\omega\}$. Then as $\alpha \downarrow 0$, $p_\alpha(\lambda|\omega)$ converges to a point mass at λ_ω. The marginal converges to a distribution with density $S(\omega, \lambda_\omega)/\int S(\omega, \lambda_\omega)d\omega$.

Case 2: Suppose $\mu(A_\omega) > 0$. Let λ_ω be any point in A_ω. As $\alpha \downarrow 0$ we have that $\int_A p_\alpha(\lambda|\omega)d\lambda \to R_\omega(A)$ for every measurable A, where

$$R_\omega(A) = \frac{\int_{A \cap A_\omega} J(\omega, \lambda)d\lambda}{\int_{A_\omega} J(\omega, \lambda)d\lambda}.$$

The marginal for ω converges to a distribution with density given by

$$\frac{\int_{A_\omega} J(\omega, \lambda)d\lambda}{\int \int_{A_\omega} J(\omega, \lambda)d\lambda d\omega}.$$

In the next Theorem we show that agreement with the Berger-Bernardo prior may hold when $\alpha = -1$.

THEOREM 5. *If $S(\omega, \lambda)$ is a nontrivial function of λ and I has the following* *form:*

$$I = \begin{bmatrix} f_1(\omega)f_2(\lambda) & 0 \\ 0 & g(\lambda) \end{bmatrix}$$

then $p_\alpha = p_B$ *with* $\alpha = -1$.

Our development requires that α be non-negative. Thus, the correspondence with the Berger-Bernardo prior for $\alpha = -1$ is only a formal correspondence. It might be possible to shed some light on this curiosity by considering the tradeoff functional from an information theoretic perspective. This is discussed in Section 5.

3. OTHER EXAMPLES.

Here we consider a few other examples. In each case we consider the Jeffreys prior J, the Berger-Bernardo prior p_B and the tradeoff prior p_α.

3.1. The Univariate Normal

Consider a $N(\mu, \sigma^2)$ model with $\omega = \mu$ and $\lambda = \sigma$. Then $J(\mu, \sigma) \propto \sigma^{-2}$, $p_B(\mu, \sigma) \propto \sigma^{-1}$ and $p_\alpha(\mu, \sigma) \propto \sigma^{-\frac{1}{\alpha}-2}$. Usually, p_B, which is the right Haar measure, is preferred. We get this from $p_\alpha(\mu, \sigma)$ if we take $\alpha = -1$. Also, the more noninformative we wish to be about μ the smaller we should choose α and the further we get from the right Haar measure. Note that p_B seems to get the "right" answer by injecting $J_\omega(\lambda)$ in place of $J(\lambda|\omega)$. If right Haar measure is preferred then one can shrink towards the right Haar measure instead of Jeffreys' prior i.e. we can define $F(p, \alpha) = i^\omega(p) - \alpha D(p||R)$ where R is right Haar measure.

3.2. The Nested Binomial

Suppose that X is binomial (m, μ) and, given $X = x$, Y is binomial (x, ν). For example, X may be the number of survivors of a disease after one year and Y may be the number of survivors after the second year. Then $J(\mu, \nu) \propto \{(1 - \mu)\nu(1-\nu)\}^{-1/2}$ – see Crowder and Sweeting (1989) and Polson and Wasserman (1990). First suppose that $\omega = \mu$ and $\lambda = \nu$. Then $p_B(\mu, \nu) \propto \{\mu(1-\mu)\nu(1-\nu)\}^{-1/2}$ and $p_\alpha(\mu, \nu) \propto \{\mu^{1/(\alpha+1)}(1-\mu)\nu(1-\nu)\}^{-1/2}$. Note that $p_\alpha = p_B$ if $\alpha = 0$.

Now let $\omega = \nu$ and $\lambda = \mu$. Then $p_B(\mu, \nu) \propto \{\mu(1-\mu)\nu(1-\nu)\}^{-1/2}$ and $p_\alpha(\mu, \nu) \propto \{\mu^{-1/\alpha}(1-\mu)\nu(1-\nu)\}^{-1/2}$. Here, $p_\alpha = p_B$ if $\alpha = -1$. In this case, $S = \sqrt{\mu/(\nu(1-\nu))}$ so the degenerate distribution as $\alpha \downarrow 0$ is singular with support on the line $\mu = 1$ and density, along this line, proportional to $\{\nu(1-\nu)\}^{-1/2}$. This is like acting as if μ were known to be 1 and the adopting a Jeffreys prior for ν.

3.3. The Multinomial

Let $y = (y_1, \ldots, y_r)$ be an observation from a multinomial $\theta = (\theta_1, \ldots, \theta_r)$ where $\theta_i \geq 0$ and $\theta_r = 1 - \sum_{i=1}^{r-1} \theta_i$. Let $\omega = \theta_1$ and $\lambda = (\theta_2, \ldots, \theta_r)$. A recent discussion of this problem is in Berger and Bernardo (1992b). Tedious arithmetic shows that $|I| = \{\prod_{i=1}^r \theta_i\}^{-1}$ and $|I_{22}| = (1-\theta_1)\{\prod_{i=2}^r \theta_i\}^{-1}$. Thus, $J(\theta) = \Gamma(r/2)\pi^{-r/2} \prod_{i=1}^r \theta_i^{-1/2}$. Now, $J_\omega(\lambda) = |I_{22}|^{1/2}/W$ where $W = \int |I_{22}|^{1/2} d\theta_2 \ldots d\theta_{r-1} = (1-\theta_1)^{(r-3)/2}\pi^{(r-1)/2}/\Gamma((r-1)/2)$ and $J_\omega(\lambda) \propto (1 -$

$\theta_1)^{-(r-3)/2}\{\prod_{i=2}^r \theta_i\}^{-1/2}$. Also, $S = \theta^{-1/2}(1-\theta)^{-1/2}$ so that, using theorem 2, $p_B(\theta) \propto J(\theta)(1-\theta_1)^{-(r-2)/2}$. The tradeoff prior is $p_\alpha(\theta) \propto J(\theta)(1-\theta_1)^{-(r-2)/(2(\alpha+1))}$. Hence they are equal if $\alpha = 0$. It is interesting to note that $E_J(\omega) = 1/r$, $E_{p_B}(\omega) = 1/2$ and $E_{p_\alpha}(\omega) = (\alpha+1)/(\alpha r + 2)$. Thus, $E_{p_\alpha}(\omega)$ is half way between the two when $\alpha = 2/r$.

4. CHOOSING α.

Here, we briefly consider the selection of α. We begin by pointing out that it is best to examine a set of priors obtained from a range of values of α. Thus, several values of α should be considered. It is useful, however, to have a default value of α.

Following McCulloch (1989), the distance $D(\cdot||\cdot)$ may be calibrated in the following way. Let $z(d) = (1 + (1-e^{-2d})^{1/2})/2$. Then the relative entropy between a fair coin and a biased coin with success probability $z(d)$ is precisely d. This puts the relative entropy on the interval $[0.5, 1.0]$. We may interpret a distance $D(\cdot||\cdot) = d$ to be the discrepancy between $1/2$ and $z(d)$. Of course, such a calibration can be criticized on many grounds but at least it provides some guidance. Suppose we choose α so that $z(D(p_\alpha||J))$ takes some intermediate value, say $3/4$. This implies α should be chosen so that $D(p_\alpha||J) = \log(2/\sqrt{3})$. The plots on the left in Figure 2 show $D(p_\alpha||J)$ as a function of α for the nested binomial example from Section 3.2. The first case corresponds to $\omega = \mu$ and the second case is $\omega = \nu$. We see from the plots that the calibration criterion gives approximately $\alpha = 0.2$ for the first case and $\alpha = 0.4$ for the second case.

We note that these two values lie between 0 and 1 which is the range in which a qualitative change in the prior is observed in Figure 1 for the multinomial problem and in Figure 2 for the many normal means problem. The plots on the right show the prior for this suggested value of α.

In the first case, S is a function of ω only so the prior does not degenerate at $\alpha = 0$. Thus, $D(p_\alpha || J)$ varies slowly and a small value of α is selected. In the second case, the prior is degenerate at $\alpha = 0$ and the rapid change in $D(p_\alpha || J)$ leads to a larger value of α.

The second prior is similar to the Jeffreys prior but is more peaked. In contrast, the first prior has a more symmetric shape than the Jeffreys prior. That p_α concentrates more sharply toward values of $\mu = 1$ when ν is the parameter of interest has an intuitive explanation. Suppose, as in Section 3.2, we interpret μ as the probability of surviving the first year and ν as the probability of surviving the second year given that one survived the first year. It would be impossible to learn about ν unless we expected survivors after the first year. Thus, to declare ν to be the parameter of interest suggests that μ is not expected to be small.

5. DISCUSSION

The Berger-Bernardo prior and tradeoff prior both correct the Jeffreys prior to account for the role of the parameter of interest and for certain cases we have uncovered some connections between the two. The agreement between the Berger-Bernardo prior and tradeoff prior for $\alpha = 0, -1$ stated in Theorems 3 and 5 is confirmed in our examples. However, this does not cover all possible cases and it remains an open problem as to whether there exist models for which the tradeoff prior and the Berger-Bernardo prior will not agree for any α.

As an alternative to the relative entropy approach of Berger and Bernardo one can note that the expected Chi-squared distance between the marginal posterior and the marginal prior is

$$E_m\chi^2(p_\omega(\cdot|Y^n), p_\omega(\cdot)) = \int p(\omega,\lambda|y^n)p(y^n|\omega,\lambda)\frac{p(y^n|\omega,\lambda')}{p(y^n|\omega,\lambda)}p(\lambda'|\omega)d\omega d\lambda d\lambda' dy^n - 1$$

where $\chi^2(f,g) = \int(f-g)^2/g$. A suitable first term for a tradeoff functional can be identified by writing the density ratio as $exp(-n((1/n)\Sigma_{k=1}^n log(p(y_k|\lambda,\omega)/p(y_k|\omega\lambda'))$ so as to approximate it by $exp(-nD(P_{\omega,\lambda}||P_{\omega,\lambda'}))$, when λ is unidimensional. Taylor expanding this relative entropy results in $exp(-(n/2)I_{22}(\lambda-\lambda')^2))$. By using Fubini's theorem in the approximation, one can integrate over λ' first, y^n second and finally over ω,λ. The first integration is a mixture of normals; the second includes dependence on n through the expectation of a posterior. This latter quantity admits an asymptotic expansion in terms of the prior and its derivatives, see Clarke and Sun (1993). From this, a choice for $d^{(1)}$ can be identified. It has been argued that one over the determinant of the Fisher information (upon normalization) is the reference prior under the Chi-squared distance in the absence of nuisance parameters so we can identify a Chi-squared tradeoff functional by using $-\alpha\chi^2(p, 1/|I|)$ as a penalty term. In this case, $d^{(1)}$ involves first and second partial derivatives of the unknown prior and so likely will prove difficult to analyze, as anticipated in Section 2.1.

In addition to the choice of functional, the choice of α remains a problem, even for the functional $F(p,\alpha)$ examined here. In Section 4, we proposed one method for selecting α. This results, however, in α strictly positive in contrast to cases where the choice of α leads to agreement with the Berger-Bernardo prior. Clearly there is room for more work here. In the current scheme one could let α depend on the difference of dimension of the parameter of interest and the nuisance parameter. Also, the sensitivity of any method for choosing α to the truncation of the parameter space is a delicate issue.

We now give a brief, information theoretic interpretation of the relative entropy tradeoff functional. We can offer a physical interpretation of this functional as a sum of rates of transmission in an asymptotic information-theoretic sense. The first term of the functional is the first term in an asymptotic expansion for $K(\Omega, Y_1^n)$ which is the Shannon mutual information. By the chain rule for mutual information, this equals $K((\Omega,\Lambda), Y_1^n) - K(\Lambda, Y_1^n|\Omega)$, where the second term is the conditional Shannon mutual information. The first of these is an achievable rate of transmission for the channel defined by $p(y|\omega,\lambda)$. The second is an average achievable rate of transmission for the nuisance parameter, averaged over possible transmissions of the parameter of interest, see Cover and Thomas (1991, Chapter 14). Maximizing this term means that we want the difference between the rates of transmission for the full parameter and for the nuisance parameter to be as large as possible; i.e., we want as much of the information we get to be from the parameter of interest as possible.

The functional in the second term is the negative of the part of the constant term which depends on the prior in an expansion for $K((\Omega, \Lambda), Y_1^n)$, ignoring the fact that λ is a nuisance parameter. Thus, the functional we seek to maximize can be regarded as arising from the quantity $(K((\Omega, \Lambda), Y_1^n) - K(\Lambda, Y_1^n|\Omega)) + \alpha K((\Omega, \Lambda), Y_1^n)$ by examining the terms in its asymptotic expansion which depend on the prior. Since both terms are positive, we expect that a maximum, if it exists, will occur for negative α.

Two values of α are obviously of interest: If $\alpha = -1$ then two of the mutual informations cancel, leaving the negative conditional mutual information. In this case, maximization of the tradeoff functional reduces to minimization of the conditional information and this will in some cases lead to the Berger-Bernardo prior. If $\alpha = 0$ then the second term does not exist and outside of particular cases the tradeoff functional does not admit a unique maximum; see Theorem 3.

Introducing the factor α permits the maximization to result in a tradeoff amongst the rates of transmission the terms in the functional represent. In this sense, our functional can be regarded as a generalization of the notion of channel capacity, i.e., the supremal rate of communication permitted by a channel. It remains an open question whether a channel can be identified for which this maximization is an appropriate measure of performance.

It is curious that although the intuition based on Shannon information is verified, examining the terms of the tradeoff functional in isolation leads to a different intuition which is not verified. Specifically, the first term in the tradeoff functional appears to take positive and negative values. Indeed it can be written as $-D(P_{\lambda,\omega}||\tilde{J}) - H(\Lambda|\Omega) + logc$ where $\tilde{J}$ is the normalized form of S, with normalizing constant c, and $H(\cdot|\cdot)$ is the conditional entropy, which can be positive or negative. The relative entropy in the second term is always positive. Thus no statement about the sign of α is obvious.

An important problem which remains unresolved is that positive values of α appear to be more statistically useful even though the information-theoretic interpretation seems to suggest negative values should be expected.

Finally, we comment that it is not necessary to shrink towards the Jeffreys prior. Indeed, our methods could be used to modify any prior including a subjective prior. The formula for p_α is then modified in the obvious way. The resulting priors, indexed by α, could be used to perform sensitivity analysis around this prior by varying α.

ACKNOWLEDGMENT

The authors thank the associate editor and a referee for helpful comments. The second author's research was supported by grants from the National Science Foundation and the National Institutes of Health.

APPENDIX: PROOFS

Proof of Theorem 1. Using a calculus of variations argument as in Clarke and Wasserman (1993) we see that p_α must satisfy

$$p_\alpha(\omega,\lambda) \propto \frac{S^{\frac{1}{\alpha}} J}{\{p_\alpha(\omega)\}^{\frac{1}{\alpha}}}.$$

Now integrate both sides with respect to λ and conclude that

$$p_\alpha(\omega) \propto \left\{ \int S^{\frac{1}{\alpha}} J d\lambda \right\}^{\frac{\alpha}{\alpha+1}}.$$

The conclusion follows since $\int\int p_\alpha(\omega,\lambda) d\omega d\lambda = 1$. □

Proof of Theorem 3. Since S depends only on ω we conclude, after some calculations, that

$$p_\alpha(\omega,\lambda) = \frac{Z_\alpha(\omega)^{\frac{\alpha}{\alpha+1}}}{\int Z_\alpha(\omega)^{\frac{\alpha}{\alpha+1}} d\omega}.$$

Now, $S = cJ/|I_{22}|^{1/2}$ for some $c > 0$. So, $J \propto S|I_{22}|^{1/2}$, and $J(\omega) \propto S(\omega) \int |I_{22}|^{1/2} d\lambda$. Now,

$$J(\lambda|\omega) = \frac{J(\omega,\lambda)}{J(\omega)} \propto \frac{S(\omega)|I_{22}|^{1/2}}{S(\omega)\int |I_{22}|^{1/2} d\lambda} \propto \frac{|I_{22}|^{1/2}}{\int |I_{22}|^{1/2} d\lambda} = J_\omega(\lambda).$$

As a result we obtain $\lim_{\alpha\downarrow 0} p_\alpha \propto S(\omega)J(\lambda|\omega) = S(\omega)J_\omega(\lambda) \propto p_B$. □

Proof of Theorem 4. Fix ω. Note that $p_\alpha(\lambda|\omega) = S^{1/\alpha}J/Z_\alpha$. Define $W(A) = \int_A J(\omega,\lambda)d\lambda$.

Case 1: Fix ω and let B be a closed sphere of radius ϵ around λ_ω. Choose ϵ sufficiently small so that $S(\omega,\lambda) > S(\omega,\lambda')$ for every $\lambda \in B, \lambda' \in B^c$. Let $s_0 = \inf_{\lambda\in B} S(\omega,\lambda)$ and let $v = S/s_0$. Then, by dominated convergence, $\int_{B^c} v^{1/\alpha} J d\lambda \to 0$. Hence,

$$\begin{aligned}
\int_B p_\alpha(\lambda|\omega)d\lambda &= \frac{\int_B S^{\frac{1}{\alpha}} J d\lambda}{\int S^{\frac{1}{\alpha}} J d\lambda} \\
&= \frac{\int_B v^{\frac{1}{\alpha}} J d\lambda}{\int_B v^{\frac{1}{\alpha}} J d\lambda + \int_{B^c} v^{\frac{1}{\alpha}} J d\lambda} \\
&\geq \frac{\int_B J d\lambda}{\int_B J d\lambda + \int_{B^c} v^{\frac{1}{\alpha}} J d\lambda} \\
&\to 1.
\end{aligned}$$

For the marginal we have

$$\begin{aligned}
p_\alpha(\omega) &= \frac{\left\{\int S^{1/\alpha} J d\lambda\right\}^{\alpha/(\alpha+1)}}{\int \left\{\int S^{1/\alpha} J d\lambda\right\}^{\alpha/(\alpha+1)} d\omega} \\
&= \frac{\left\{\int_B S^{1/\alpha} J d\lambda + \int_{B^c} S^{1/\alpha} J d\lambda\right\}^{\alpha/(\alpha+1)}}{\int \left\{\int_B S^{1/\alpha} J d\lambda + \int_{B^c} S^{1/\alpha} J d\lambda\right\}^{\alpha/(\alpha+1)} d\omega} \\
&= \frac{\left\{\int_B v^{1/\alpha} J d\lambda + \int_{B^c} v^{1/\alpha} J d\lambda\right\}^{\alpha/(\alpha+1)}}{\int \left\{\int_B v^{1/\alpha} J d\lambda + \int_{B^c} v^{1/\alpha} J d\lambda\right\}^{\alpha/(\alpha+1)} d\omega}.
\end{aligned}$$

Let $N(\alpha)$ denote the numerator of the last expression. Then $\log N(\alpha) = 1/(\alpha + 1) \log f(\alpha) g(\alpha)$ where

$$f(\alpha) = \left\{\int_B v^{1/\alpha} J d\lambda\right\}^\alpha$$

and

$$g(\alpha) = \left\{1 + \frac{\int_{B^c} v^{1/\alpha} J d\lambda}{\int_B v^{1/\alpha} J d\lambda}\right\}^\alpha.$$

Now $\int_{B^c} v^{1/\alpha} J d\lambda \to 0$ and $\int_B v^{1/\alpha} J d\lambda \geq \int_B J d\lambda > 0$ so $g(\alpha) \to 1$. And by the convergence of the L_p norm to the L_∞ norm we have that $f(\alpha) \to \sup v = \hat{s}/s_0$. This convergence is uniform in ω because the parameter space is compact. Thus $N(\alpha) \to \hat{s}/s_0$. For the denominator we have $\lim_\alpha \int N(\alpha) d\omega = \int \lim_\alpha N(\alpha) d\omega = \int \hat{s}/s_0 d\omega$ since the numerator is uniformly bounded.

Case 2: Let $v = S/\hat{s}$. Note that $v^{1/\alpha}$ converges to 0 for each fixed $\lambda \in A_\omega^c$ and that $0 \leq v^{1/\alpha} \leq 1$. Because of compactness, we conclude from the Lebesgue dominated convergence theorem that $\int_{A_\omega^c} v^{1/\alpha} J d\lambda \to 0$. Hence,

$$\begin{aligned}
\int_A p_\alpha(\lambda|\omega) d\lambda &= \frac{\int_{A \cap A_\omega} S^{\frac{1}{\alpha}} J d\lambda + \int_{A \cap A_\omega^c} S^{\frac{1}{\alpha}} J d\lambda}{\int_{A_\omega} S^{\frac{1}{\alpha}} J d\lambda + \int_{A_\omega^c} S^{\frac{1}{\alpha}} J d\lambda} \\
&= \frac{\hat{s}^{1/\alpha} W(A \cap A_\omega) + \int_{A \cap A_\omega^c} S^{\frac{1}{\alpha}} J d\lambda}{\hat{s}^{1/\alpha} W(A_\omega) + \int_{A_\omega^c} S^{\frac{1}{\alpha}} J d\lambda}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{W(A\cap A_\omega)+\int_{A\cap A_\omega^c} v^{\frac{1}{\alpha}} J d\lambda}{W(A_\omega)+\int_{A_\omega^c} v^{\frac{1}{\alpha}} J d\lambda} \\
&\to \frac{W(A\cap A_\omega)}{W(A_\omega)} = R_\omega(A).
\end{aligned}
$$

The proof of the convergence of the marginal density is omitted. □

Proof of Theorem 5. We get that

$$p_B(\omega,\lambda)=\frac{f_1(\omega)^{1/2} g(\lambda)^{1/2}}{\int f_1(\omega)^{1/2} d\omega \int g(\lambda)^{1/2} d\lambda}$$

and

$$p_\alpha(\omega,\lambda)=\frac{f_1(\omega)^{1/2} f_2(\lambda)^{(\alpha+1)/(2\alpha)} g(\lambda)^{1/2}}{\int f_1(\omega)^{1/2} d\omega \int f_2(\lambda)^{(\alpha+1)/(2\alpha)} g(\lambda)^{1/2} d\lambda}$$

and the result follows. □

REFERENCES.

Berger, J. and Bernardo, J. (1989). Estimating the product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84** 200-207.

Berger, J. and Bernardo, J. (1992a). On the development of the reference prior method. In *Bayesian Statistics 4: Proceedings of the Fourth International Meeting on Bayesian Statistics.* Clarendon Press: Oxford.

Berger, J. and Bernardo, J. (1992b). Ordered group reference priors with application to the multinomial problem. *Biometrika* **79** 25-37.

Bernardo, J. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc.* **41** 113-147.

Clarke, B. and Barron, A. (1990). Bayes and minimax asymptotics of entropy risk. Technical report #90-11, Dept. of Statistics, Purdue University.

Clarke, B. and Barron, A. (1994). Jeffreys prior is asymptotically least favorable under entropy risk. *J. Statist. Planning and Inference* **41** 37-60.

Clarke, B. and Sun, D. (1993) Reference priors under the Chi-squared distance. UBC Department of Statistics Technical Report 124.

Clarke, B. and Wasserman, L. (1993). Noninformative priors and nuisance parameters. *J. Amer. Statist. Assoc.* **88**, 1427-1432.

Courant, R. and Hilbert, D. (1965). *Methods of Mathematical Physics. Vol. 1.* Interscience, New York.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* John Wiley and Sons Inc. New York.

Crowder, M. & Sweeting, T. (1989). Bayesian inference for a bivariate binomial. *Biometrika* **76**, 599-604.

Datta, G. and Ghosh, M. (1993). On the invariance of noninformative priors. University of Georgia, Department of Statistics Technical Report 93-17.

Elsgolts, L. (1970). *Differential Equations and the Calculus of Variations.* Mir, Moscow.

Ghosh, J.K. and Mukerjee, R. (1992). Noninformative priors. In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting on Bayesian Statistics.* Clarendon Press: Oxford.

Ibrigamov, I.A. and H'asminsky, R.Z. (1973). On the information contained in a sample about a parameter. *2nd Int. Symp. on Info. Theory.* 295-309.

Jeffreys, H. (1961). *Theory of Probability.*(Third edition). Clarendon Press: Oxford.

McCulloch, R. (1989). Local Model Influence. *J. Amer. Statist. Assoc.* **84**, 473-478.

Mukerjee, R. and Dey, D.K. (1992). Frequentist validity of posterior quantiles in the presence of a nuisance parameter: higher order asymptotics. Technical report 92-07, Department of Statistics, The University of Connecticut.

Polson, N. (1992). On the expected amount of information from a non-linear model. *R. Roy. Statist. Soc.*, **54**, 889-895.

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.

Bertrand Clarke
Department of Statistics
University of British Columbia
2021 West Mall, Vancouver, B.C.
Canada V6T 1Z2

Larry Wasserman
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213

FIGURE CAPTIONS.

Figure 1. The many normal means problem for $n = 2$. The first column is $p_\alpha(\omega)$ the second is $p_\alpha(\mu_1, \mu_2)$.

Figure 2. The plots on the left show $D(p_\alpha||J)$ as a function of α for the nested binomial when the parameter of interest is μ and ν, respectively. The critical value of α corresponding to $D(p_\alpha||J) = \log(2/\sqrt{3})$ is indicated on the plots. The plots on the right show the tradeoff prior p_α when α is chosen so that $D(p_\alpha||J) = \log(2/\sqrt{3})$.