Sankhyā: The Indian Journal of Statistics
1997, Volume 59, Series A, Pt. 2, pp. 215-231

REFERENCE PRIORS UNDER THE CHI-SQUARED DISTANCE*

By BERTRAND CLARKE
University of British Columbia, Vancouver
and
DONGCHU SUN
University of Missouri-Columbia, Columbia

SUMMARY. For smooth parametric families in exponential form equipped with a smooth prior density on a real parameter θ, reference priors for use with independent, identically distributed data are obtained by maximising the expected Chi-squared distance between a prior density and its corresponding posterior density. We identify an asymptotic expansion for this Chi-squared distance and use it to define a functional which can be optimised so as to obtain a reference prior. Performing this optimisation for a unidimensional parameter leads to a prior which is proportional to $1/\sqrt{detI(\theta)}$, where θ is a d-dimensional real parameter and $I(\theta)$ is the Fisher information matrix. We argue the relevance of the Chi-squared distance by noting its relationship with the Chi-squared goodness-of-fit statistic and present computational results to show how this Chi-squared reference prior performs for $d = 1$. In addition, we briefly consider the case where nuisance parameters are present. Finally, we discuss the importance of the choice of measure of distance on prior selection and evaluation.

1. Introduction

In a parametric Bayesian model, one typically assumes that the likelihood function is known so that only the prior density on the parameter remains to be specified. Numerous prior selection techniques have been proposed. Those aimed at obtaining noninformativity are derived, roughly, from one of three principles: invariance, frequentist-matching, and functional optimisation.

Paper received. April 1997.
AMS (1991) *subject classification.* 62A15, 62C10.
Key words and phrases. Reference priors; distance measures; prior selection.
* Clarke's research was partially supported by the Faculty of Science, University of Columbia. Sun's work was partially supported by a Research Board Grant from the University of Missouri-System, a Summer Research Fellowship and a Research Council grant from the University of Missouri-Columbia.

For invariance, important contributions have been made by Jeffreys (1961), George and McCulloch (1992), Eaves (1985), Chang and Eaves (1990), Datta and Ghosh (1996). While invariance is important it does not typically specify priors uniquely by itself. Indeed, this was recognized by Jeffreys (1961) when he proposed priors proportional to $(\det I(\theta))^{1/2}$ where θ is a d-dimensional real parameter, and $I(\theta)$ is the Fisher information matrix from the parametric family $p(\cdot|\theta)$. A problem with invariance is that the prior is invariant under transformations which alter the behaviour of the estimand. For instance, the prior one would want for estimating a variance is not the same as the prior one would want for estimating the inverse of a variance.

Matching frequentist coverage probabilities has been used by Peers (1965), Tibshirani (1989), Mukerjee and Dey (1993), Datta and Ghosh (1995) and Sun and Ye (1996). Jeffreys prior emerges in this context also, see Hartigan (1983). This class of prior selection principles is useful for comparison purposes because it translates frequentist procedures into Bayesian terms. On the other hand, if a prior merely duplicates frequentist behavior one can argue that one might just as well use a frequentist analysis.

The earliest example of functional optimisation for prior selection is due to Hartigan (1965). More recently, functionals based on information-theoretic quantities have been explored by numerous authors. Chief amongst these is the idea of the reference prior, introduced by Bernardo (1979). This approach has been developed in Berger and Bernardo (1989, 1992a,b) and among others. A key feature of this method is that it can be used in the presence of nuisance parameters. A recent review and annotated bibliography on reference priors is given by Kass and Wasserman (1996).

The central idea of this work, see Bernardo (1979), is to maximize the expected relative entropy or Kullback-Leibler distance between a prior and its corresponding posterior. The priors thereby obtained can be interpreted as noninformative in the sense that they change the most on average upon receipt of the data. Thus, they can be used directly in the absence of information or used as Bernardo (1979) originally intended, namely as a benchmark for evaluating other, subjective, priors. Relative entropy reference priors also admit interpretations in data compression and channel capacity, see Clarke and Barron (1994).

Alternatively, one might seek reference priors which have an interpretation in terms of goodness-of-fit rather than in information. This would lead one to use a different measure of distance. In particular, motivated by the Chi-squared goodness-of-fit statistic, one can use the Chi-squared distance defined by $\chi^2(p,q) = \int (p-q)^2/q$ for densities p, q. Thus, for priors $w(\theta)$ and posteriors $w(\theta|X^n)$ we examine $E_m\chi^2(w(\cdot|X^n), w(\cdot))$ and maximize it over choices of $w(\theta)$ in an asymptotic sense. Here, $X^n = (X_1, ..., X_n)$ is i.i.d. from the parametric family $p(\cdot|\theta)$ and E_m denotes expectation with respect to the mixture density $m(x^n) = \int w(\theta)p(x^n|\theta)d\theta$. We note that the Chi-squared distance and the relative entropy are not metrics despite having metric-like properties.

The Chi-squared distance is much stronger than the relative entropy. Indeed, writing the relative entropy as $D(p||q) = \int p\ell n(p/q)$ we have $\chi^2(p,q) \geq D(p||q)$. So, asymptotic expressions for the expected Chi-squared distance are different also. Note that because the posterior can be written as $w(\theta|x^n) = w(\theta)p(x^n|\theta)/m(x^n)$, we have that

$$E_m\chi^2(w(\cdot|X^n), w(\cdot)) = \int_{\mathbb{R}}\int_{\mathbb{R}^\kappa} w(\theta|X^n)p(X^n|\theta)dX^n d\theta - 1. \qquad \ldots(1.1)$$

So, it is seen that the inner integral is the expected value of the posterior density with respect to $p(\cdot|\theta)$. Now, one can derive a Chi-squared reference prior using an asymptotic expression for $E_\theta w(\theta|X^n)$ for parametric families in exponential form. Specifically, we use

$$\begin{aligned} Ew(\theta|X^n) \quad = \quad & \frac{n^{d/2}|I(\theta)|^{1/2}}{2^{d/2}(2\pi)^{d/2}}\left\{1 + (\frac{1}{n})\left(\frac{\nabla w(\theta)^t I^{-1}(\theta)\nabla w(\theta)}{2w^2(\theta)} - \frac{3tr\ \nabla^2 w(\theta)I^{-1}(\theta)}{4w(\theta)}\right.\right. \\ & \left.\left. -2^{d/2}Ep_{2,\theta}(Z)e^{-Z^tZ/2} - \frac{2^{d/2}}{w(\theta)}\nabla w^t(\theta)I^{-1/2}(\theta)EZp_{1,\theta}(Z)e^{-Z^tZ/2}\right)\right. \\ & \left. +o(\frac{1}{n})\right\} \end{aligned} \qquad \ldots(1.2)$$

which is valid pointwise in θ, where p_1 and p_2 are polynomials arising from an Edgeworth expansion and Z is a d-dimensional mean zero normal random variable with the $d \times d$ identity as its variance matrix, see Clarke and Sun (1993).

Thus, we assume (1.2) is uniformly good on compact sets in the parameter space and we use the terms of order $O(n^{((d/2)-1)})$ to define a functional to be optimised. For unidimensional exponential families, the resulting Euler-Lagrange equation can be solved and the prior obtained can be regarded as the reference prior under the Chi-squared distance. This reference prior is proportional to reciprocal of the square root of the Fisher information. That is, the reference prior under the Chi-squared distance for a unidimensional parameter restricted to a compact set K is $1/c\sqrt{I(\theta)}$ where $c = \int_K 1/\sqrt{I(\theta)}d\theta$. Although (1.2) holds in the d dimensional case, we restrict to the unidimensional case for formal results because otherwise the optimisation is intractable. However, the unidimensional case suggests a prior proportional to $1/\sqrt{detI(\theta)}$ as the Chi-squared reference prior for a general d-dimensional parameter.

In addition to the goodness-of-fit interpretation already mentioned, this prior admits an estimation interpretation. Consider a value of θ for which $I(\theta)$ is large. Then, the Cramèr-Rao lower bound, $1/I(\theta)$, is small. So, an efficient estimator for θ will require comparatively little data to obtain credible sets of prescribed length and credibility. By contrast, for a value of θ with small $I(\theta)$ the Cramèr-Rao lower bound is large and even an efficient estimator will require relatively more data for good estimation. For such θ, a prior based on $1/\sqrt{I(\theta)}$

is large, increasing the value of $w(\theta|x^n)$, thereby making those values of θ more likely under the posterior. In particular, a prior proportional to $1/\sqrt{I(\theta)}$ puts much of its weight on those parameter values which are hardest to discriminate. Jeffreys prior is the result of normalizing $\sqrt{I(\theta)}$ and gives a reverse weighting; it puts much of its weight on parameter values with large $I(\theta)$ and are easier to discriminate.

This observation suggests that the Chi-squared reference prior is uninformative in a stronger sense than Jeffreys prior, for instance, and so is a better benchmark for comparison. Specifically, actual use of a prior proportional to $1/\sqrt{detI(\theta)}$ flattens the posterior density making discrimination amongst θ-values more difficult. This shows that $1/\sqrt{I(\theta)}$ does what one wants a reference prior to do: As Bernardo (1979) observed, one does not intend to use a reference prior for estimation necessarily, rather one uses it for comparison purposes so as to assess the strength of the assumptions implicit in, for instance, an informative or subjective prior. This may be useful in prior sensitivity analysis and experimental design.

Another point of obtaining a reference prior under a different measure of distance is to show that the distance matters. Most proposed noninformative priors are based on generalizing properties satisfied by Jeffreys prior and it has been the impression that non-informative priors would not depend too much on the distance used. That we obtain a prior proportional to $1/\sqrt{I(\theta)}$, which assigns most of its mass where Jeffreys prior does not is therefore striking.

The structure of this paper is as follows. In Section 2 we restrict to the unidimensional parameter case and identify a functional which we optimise so as to obtain a Chi-squared reference prior. In Section 3 we formalise a goodness-of-fit interpretation for the Chi-squared distance, discuss some properties of the priors we derive, and provide computational results. In Section 4 we discuss some of the issues raised by changing the distance.

2. Unidimensional Parameters in an Exponential Family

Assume that p_θ is a unidimensional parametric family in exponential form with the natural parameterization. That is, write

$$p(x|\theta) = e^{\theta \cdot x - \psi(\theta)}.$$

As a consequence of Theorem 2.1 and Proposition 2.2 in Clarke and Sun (1993) we have

$$\begin{aligned}\int w(\theta|x^n)p(x^n|\theta)dx^n &= \sqrt{\frac{nI(\theta)}{4\pi}}\left(1 - \frac{3w''(\theta)}{4nw(\theta)I(\theta)} + \frac{w'(\theta)^2}{2nI(\theta)w(\theta)^2}\right. \\ &\quad \left. - \frac{3}{n}\left(\frac{\mu_4-3}{24}\right) - \frac{3}{4}\left(\frac{\mu_3}{6n}\right)\frac{w'(\theta)}{w(\theta)I(\theta)^{1/2}}\right) + o\left(\frac{1}{\sqrt{n}}\right), \qquad \ldots(2.1)\end{aligned}$$

where $p_1(Z) = (\mu_3/6)(Z^3-3Z)$, $p_2(Z) = \frac{\mu_3}{72}(Z^3-3Z)+\frac{(\mu_4-3)}{24}(3-6Z^2+Z^4)$, and we have used $\mu_i = E(I(\theta)^{-1/2}\nabla \log\, p(X_1|\theta))^i$ and $\sigma^2 = \mathrm{Var}\,(I(\theta)^{-1/2}\nabla \log p(X_1|\theta)) = 1$.

2.1 *Solution of the Euler-Lagrange equation.* We assume, as is typically done in the reference prior literature, that (2.1) holds uniformly in θ on a compact interval $[a, b]$. While it is difficult to justify this assumption formally, we note that Egoroff's theorem guarantees the existence of a set with arbitrarily large probability on which the convergence can be taken as uniform. We have taken this to be an interval here because all the quantities that occur in the derivation of (1.2) are continuous in θ. We comment that it is well known that the limit of relative entropy reference priors which result from truncating to compact sets depends on the sequence of compact sets.

Under this assumption, we can integrate with respect to θ so as to get

$$\begin{aligned} E_m\chi^2(w(\cdot), w(\cdot|X^n)) &= \int_{\mathbb{R}}\int_{\mathbb{R}^K} w(\theta|X^n)p(X^n|\theta)dX^n d\theta - 1 \\ &= \int_{\mathbb{R}} \sqrt{\tfrac{nI(\theta)}{4\pi}}\left(1 - \tfrac{3w''(\theta)}{4nw(\theta)I(\theta)} + \tfrac{w'(\theta)^2}{2nI(\theta)w(\theta)^2} - \tfrac{3}{n}\left(\tfrac{\mu_4-3}{24}\right)\right. \\ &\quad \left. + \tfrac{3}{4}\left(\tfrac{\mu_3}{6n}\right)\tfrac{w'(\theta)}{w(\theta)I(\theta)^{1/2}}\right)d\theta - 1 + o\left(\tfrac{1}{\sqrt{n}}\right). \end{aligned} \quad \ldots(2.2)$$

Now the reference prior under the Chi-squared distance is the one that maximizes

$$\int_a^b \left\{ \frac{w'(\theta)^2}{2I^{1/2}(\theta)w(\theta)^2} - \frac{3w''(\theta)}{4w(\theta)I^{1/2}(\theta)} + \frac{\mu_3 w'(\theta)}{8w(\theta)} \right\} d\theta. \quad \ldots(2.3)$$

Note that the optimisation will not be affected by the first and fourth terms in the integrand of (2.2), since $\mu_4 = \mu_4(\theta)$ is constant as a function of w. For the unidimensional exponential family above, we have that $\mu_3 = \psi^{(3)}(\theta)/[\psi''(\theta)]^{3/2}$. Therefore (2.3) is equivalent to

$$\int_a^b \left\{ \frac{w'(\theta)^2}{2[\psi''(\theta)]^{1/2}w(\theta)^2} - \frac{3w''(\theta)}{4[\psi''(\theta)]^{1/2}w(\theta)} + \frac{\psi^{(3)}(\theta)w'(\theta)}{8[\psi''(\theta)]^{3/2}w(\theta)} \right\} d\theta. \quad \ldots(2.4)$$

Direct calculation shows that the Euler-Lagrange equation from (2.4) can be written as

$$4\frac{d}{d\theta}\left[\frac{w'(\theta)}{w(\theta)}\right] + 2\frac{d}{d\theta}\left[\frac{\psi^{(3)}(\theta)}{\psi''(\theta)}\right] - 2\frac{w'(\theta)\psi^{(3)}(\theta)}{w(\theta)\psi''(\theta)} - \left[\frac{\psi^{(3)}(\theta)}{\psi''(\theta)}\right]^2 = 0. \quad \ldots(2.5)$$

If we write $y(\theta) = \frac{w'(\theta)}{w(\theta)}$ and $\eta(\theta) = \frac{\psi^{(3)}(\theta)}{\psi''(\theta)}$, then (2.5) is equivalent to

$$2(2y' + \eta') - (2y + \eta)\eta = 0. \quad \ldots(2.6)$$

We can solve (2.6) explicitly. There are two cases.

Case I: If $2y + \eta = 0$, then from the definition of y and η, we have after integrating that $\log w^2(\theta) = -\log(\psi''(\theta)) + C$, for some constant C. So $w(\theta)$ is proportional to $1/\sqrt{\psi''(\theta)}$.

Case II: If $2y + \eta \neq 0$, then dividing (2.6) by $2y + \eta$, and integrating gives $\log(2y(\theta) + \eta(\theta)) = \log[C_2\sqrt{\psi''(\theta)}]$ for some constant C_2. Equivalently, we have $2y(\theta) = -\eta(\theta) + C_2\sqrt{\psi''(\theta)}$. Substituting the definitions of y and η, and integrating gives

$$w(\theta) = \frac{C_1}{\sqrt{\psi''(\theta)}} \exp\Big\{\frac{C_2}{2}\int_a^\theta \sqrt{\psi''(s)}ds\Big\}, \qquad \ldots(2.7)$$

for some constant $C_1 > 0$.

Note that the functional (2.4) is homogeneous so constant multiples of w do not change its value. Thus, substituting (2.7) into (2.4) gives a value independent of C_1. So, it is enough to optimise over C_2. Substituting (2.7) with $C_1 = 1$ into (2.3) one obtains

$$\int_a^b \frac{-2c_2^2\psi''^3 - 4(\psi^{(3)})^2 + 3\psi''\psi^{(4)}}{8(\psi'')^{5/2}}d\theta,$$

and it is seen that the choice $C_2 = 0$ maximizes the functional (and gives $1/\sqrt{\psi''(\theta)}$ as the solution achieving that maximum). Thus, within the class of solutions, the result of case I, $1/\sqrt{\psi''(\theta)}$, is the unique limit of a sequence of solutions which maximizes (2.3). Moreover, in both case I and case II, we obtain the standardized inverse square root of the Fisher information as the reference prior under the Chi-squared distance.

2.2. *Reference priors within smaller classes of priors.* Note that we have not verified that (2.3) has been maximized over the class of all twice continuously differentiable prior densities, the class used in deriving the Euler-Lagrange equation. In fact, it can be shown that within this large class the solution $1/\sqrt{\psi''(\theta)}$ is an inflection point. However, over a different collection of densities one can verify that priors proportional to $1/\sqrt{\psi''(\theta)}$ maximise the functional.

Suppose $\psi''(\theta)$ is decreasing and consider the class of priors w that are proportional to $f(\theta)/\sqrt{\psi''(\theta)}$ where the function f is positive and has a continuous derivative. For any such prior we have that

$$\frac{w'(\theta)}{w(\theta)} = \frac{f'(\theta)}{f(\theta)} - \frac{\psi^{(3)}(\theta)}{2\psi''(\theta)}.$$

An equivalent form for (2.3) is

$$\begin{aligned} G(w) \;=\; & -\tfrac{3}{4}\frac{w'(\theta)}{w(\theta)\sqrt{\psi''(\theta)}}\Big|_a^b - \tfrac{1}{4}\int_a^b \Big(\frac{w'\theta}{w(\theta)}\Big)^2 \frac{1}{\sqrt{\psi''(\theta)}}d\theta \\ & -\tfrac{1}{4}\int_a^b \Big(\frac{w'(\theta)}{w(\theta)}\Big)\frac{\psi^{(3)}(\theta)}{(\psi''(\theta))^{3/2}}d\theta. \end{aligned} \qquad \ldots(2.8)$$

So, expression (2.8) gives

$$G(\frac{f}{\sqrt{\psi''}}) = -\frac{3}{4}\Big(\frac{f'}{\sqrt{\psi''}f} - \frac{\psi^{(3)}}{2\psi''^{3/2}}\Big)\Big|_a^b - \frac{1}{4}\int_a^b (\frac{f'}{f})^2 \frac{1}{\psi''^{1/2}} d\theta + \frac{1}{16}\int_a^b \frac{(\psi^{(3)})^2}{\psi''^{5/2}} d\theta.$$

It is therefore enough to minimise

$$3\frac{f'}{f\psi''^{1/2}}\Big|_a^b + \int_a^b (\frac{f'}{f})^2 \frac{1}{\psi''^{1/2}} d\theta.$$

Writing $g(\theta) = f'(\theta)/f(\theta)$ and $h(\theta) = 1/\psi''(\theta)^{1/2}$ it is equivalent to minimise

$$3g(b)h(b) - 3g(a)h(a) + \int_{a+\epsilon}^{b-\epsilon} g^2(\theta)h(\theta)d\theta + \int_a^{a+\epsilon} g^2(\theta)h(\theta)d\theta + \int_{b-\epsilon}^b g^2(\theta)h(\theta)d\theta, \quad \ldots(2.9)$$

for any $\epsilon > 0$. Now it is seen that for any fixed but otherwise arbitrary set of boundary conditions, the choice of g identically zero is the unique limit of any sequence of functions for which (2.9) decreases to its minimal value of zero. In this sense one gets $g = 0$ and so $f'(\theta) = 0$, implying $f(\theta)$ is a constant at the minimum. Thus, the inverse square root of $\psi''(\theta)$ is the unique maximiser of (2.3) within the class of priors proportional to expressions of the form $f(\theta)/\psi''(\theta)^{1/2}$.

It may be desirable to maximize (2.3) over a smaller collection of densities. We can consider a class of priors which includes Jeffreys prior and $1/\sqrt{I(\theta)}$ a special cases. Assume $\psi^{(3)}(\theta)$ is not zero on (a, b) so that $\psi''(\theta)$ is not a constant. Let $w_\alpha(\theta) = I(\theta)^\alpha = (\psi''(\theta))^\alpha$, where normalizing constants have been neglected since the functional (2.3) is homogeneous.

Evaluating (2.8) at $w_\alpha(\theta)$ gives

$$G(w_\alpha) = -\frac{3\alpha\psi^{(3)}(\theta)}{4(\psi''(\theta))^{3/2}}\Big|_a^b - \frac{1}{4}\int_a^b (\alpha^2 + \alpha)\frac{\psi^{(3)}(\theta)^2}{\psi''(\theta)^{5/2}} d\theta. \quad \ldots(2.10)$$

Differentiating with respect to α, setting the derivative equal to zero and solving for the optimal α^* gives

$$\alpha^* = \frac{1}{2}\frac{-3\frac{\psi^{(3)}(\theta)}{\psi''(\theta)^{3/2}}\Big|_a^b}{\int_a^b \frac{\psi^{(3)}(\theta)^2}{\psi''(\theta)^{5/2}} d\theta} - \frac{1}{2}. \quad \ldots(2.11)$$

Differentiating (2.10) twice with respect to α gives a negative second derivative so α^* maximizes (2.10).

If we consider an exponential distribution with the natural parameterization for instance $p(x|\theta) = \exp(\theta x + \log(-\theta))$ where $\theta < 0$, then we have that $\psi''(\theta) = 1/\theta^2$ so that the numerator of the first term on the right hand side of (2.11) is seen to be zero ($\psi^{(3)}(\theta)/\psi''(\theta)^{3/2}$ is the constant -2) then $\alpha^* = -1/2$. That is,

a prior of the form $c/\sqrt{I(\theta)}$ maximizes G over the class $\{w_\alpha|\alpha \in \mathbb{R}\}$. It can be verified that the same result is obtained if p_θ is the gamma family with known shape parameter. Consequently, $1/\sqrt{I(\theta)}$ is less informative in the Chi-squared sense for these parametric families than Jeffreys prior is.

If we consider other distributions such as the Poisson, Binomial, or Geometric in their natural parametrizations, we generally find that α^* depends on the endpoints a and b and takes values $\pm\infty$ as $b = -a$ goes to infinity. This shows that over the small class of functions which are powers of the Fisher information we cannot separate the boundary effects from the behaviour on the interior (a, b) and that, moreover, Jeffreys prior does not minimise (2.3). That is, Jeffreys prior is not minimally informative even within a relatively small class of alternative priors.

We note that maximising over the class of priors proportional to $I(\theta)^\alpha$ gives a different result than maximising over the class of priors proportional to $f(\theta)/\sqrt{I(\theta)}$. This occurs because the larger class permits one to isolate the effect of the optimisation at the boundary points a, b from the optimisation on the interior (a, b). One can modify the functional obtained from the asymptotic behavior of $E\chi^2(w(\cdot), w(\cdot|X^n))$ by subtracting a penalty term. This may permit one to obtain a well defined maximum. However, even a seemingly tractable penalty term such as $\alpha \int ((w'/w)')^2 d\theta$ gives an equation which does not appear to have useful solutions.

3. Interpretations of the Chi-squared Reference Prior

In this section we motivate use of the Chi-squared distance by noting its relationship with the Chi-squared goodness-of-fit test. By contrast, the relative entropy from which one can derive the Jeffreys prior has an analogous motivation from information theoretic considerations. Next, we turn briefly to the cases when $d \geq 2$ or nuisance parameters are present. Then, we give some computational results comparing the Jeffreys prior and the prior based on $1/\sqrt{I(\theta)}$ in two examples, the *Binomial*(n, θ) and the *Exponential*(θ).

3.1. *Statistical interpretation of the chi-squared distance.* The Chi-squared distance is motivated by the Chi-squared goodness-of-fit statistic. Consider a discretized form of a continuous random variable. If we have a guess as to the true distribution F, we might test this guess by using a Chi-squared goodness-of-fit test. The simplest form of this test is

$$\chi_s^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the number of observations in cell i, $E_i = nP(X_1 \in \text{cell } i)$, and k is the number of cells. Now select $0 = a_o < \ldots < a_k = 1$ so that

$-\infty = F^{-1}(a_o), \ldots, F^{-1}(1) = \infty$, where $p_i = a_i - a_{i-1}$ is the probability of the i^{th} cell under F, the distribution function of P. Let $\hat{p}_i$ be the empirical probability of the i^{th} cell, that is the number of outcomes in the i^{th} cell divided by the sample size, n. Denote the empirical distribution by $F_n(t) = (1/n)\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}$ and its conditional expectation given the cell counts by $\hat{F}_n(t) = E(F_n(t) \mid \hat{p}_1, \ldots, \hat{p}_k)$. Now, if we choose $t \in (a_{i-1}, a_i]$ we can show that

$$\hat{F}_n(t) = \sum_{r < i} \hat{p}_r + \frac{\hat{p}_i}{p_i}(F(t) - F(a_{i-1})).$$

Note that $(1/n)\chi_s^2 = \sum_{i=1}^{k}(\hat{p}_i - p_i)^2/p_i$. Now, a typical summand in the right hand side of this equation is

$$\frac{(\hat{p}_i - p_i)^2}{p_i} = \int_{F^{-1}(a_{i-1})}^{F^{-1}(a_i)} \frac{(\hat{p}_i - p_i)^2}{p_i^2} f(t)dt$$

where f is the density of F. The integrand can be written as $((\hat{p}_i f(t)/p_i - f(t))^2 / f(t)$ which can be recognized as $(\hat{F}_n'(t) - F'(t))^2/F'(t)$. Summing over s gives

$$\frac{1}{n}\chi_s^2 = \int_{-\infty}^{\infty} \frac{(\hat{F}_n'(t) - F'(t))^2}{F'(t)} dt,$$

i.e., the Chi-squared statistic is a rescaled version of the Chi-squared distance between a smoothed histogram and the true density.

In a sense that can be made precise by taking limits as the cell diameter shrinks, the Chi-squared distance is a measure of goodness-of-fit. Since we are maximizing it here, we are imposing a sort of 'badness of fit' criterion.

We have maximized the distance between a prior and its corresponding posterior over a class of priors. This can be justified in a frequentist way; Jeffreys prior admits an analogous interpretation in terms of repeated use of a channel. It is important to note that the justification for both the Jeffreys prior and the Chi-squared prior is asymptotic.

Suppose we have k data vectors of length n, $\tilde{X}_1, \ldots, \tilde{X}_k$ which arise from k repetitions of the same experiment, in a Bayesian way. That is, each data vector $\tilde{X}_i$ is the second stage of a two stage experiment, in which the first stage was choosing θ_i from a distribution with density w. We can now estimate each θ_i by $\hat{\theta}_i$ based on $\tilde{X}_i$. From these k estimates we can construct an estimator for w. Our optimisation procedure is identifying the density w which is hardest to estimate with a restricted class of estimators – histograms and expectations of histograms. If we believe that goodness-of-fit is the appropriate way to assess the performance of a posterior then we may be led to the Chi-squared distance and hence to $1/\sqrt{I(\theta)}$ as a noninformative prior.

3.2 *The case $d \geq 2$ and nuisance parameters.* In Section 2, we maximised an asymptotic expression for $E_m\chi^2(w(\theta|X^n), w(\theta))$ to obtain a prior proportional to $1/\sqrt{I(\theta)}$ for unidimensional parameters in exponential families. The general d-dimensional form of (1.2) can be used to obtain an analog to (2.1) which can be optimised by similar techniques. However, this is difficult due to the complexity of the Euler-Lagrange equations. Our results from Section 2 suggest that the result of such an optimisation would be a prior proportional to $1/\sqrt{detI(\theta)}$; that is, the reference prior under the Chi-squared distance remains the inverse square of the Jeffreys prior even in the multi-dimensional parameter setting.

An important area of application for prior selection principles is the case of nuisance parameters. We develop our formal approach for that context as follows. Suppose θ remains of interest but a nuisance parameter ϕ is also present. With some abuse of notation, the quantity we now want to maximise asymptotically is

$$E_m\chi^2(w(\theta|X^n), w(\theta)) = \int w(\theta|x^n)m(x^n|\theta)d\theta dx^n - 1, \qquad \ldots(3.1)$$

in which $m(x^n) = \int w(\theta,\phi)p(x^n|\theta,\phi)d\theta d\phi$, for a joint prior $w(\theta,\phi)$, $w(\theta|X^n)$ is the marginal posterior and $m(x^n|\theta) = \int w(\phi|\theta)p(x^n|\theta,\phi)d\phi$. The integral on the right hand side of (3.1) can be written as

$$\begin{aligned}&\int\int\int\int w(\theta,\phi|x^n)w(\phi'|\theta)p(x^n|\theta,\phi')dx^n d\phi d\phi' d\theta \\ &\qquad \approx \int\int\int w(\theta,\phi|x^n)w(\phi|\theta)p(x^n|\theta,\phi)dx^n d\phi d\theta,\end{aligned} \qquad \ldots(3.2)$$

in which the approximation $\phi = \phi'$ holds in the limit of large n because the set on which $\phi \neq \phi'$ has probability tending to zero. The right hand side of (3.2) reduces to

$$\int\int w(\phi|\theta)E_{\theta,\phi}w(\theta,\phi|X^n)d\theta d\phi, \qquad \ldots(3.3)$$

in which (1.2) can be used on the expected posterior. This gives a functional to optimise in the nuisance parameter case. Analogous to the reference prior algorithm of Berger and Bernardo (1989), one can note the following. Solution of the maximisation problem requires that one specify a conditional prior for ϕ given θ, say $w(\phi|\theta)$. One possibility is to use $1/\sqrt{detI_{22}(\theta,\phi)}$, where $I_{22}(\theta,\phi)$ is the Fisher information of ϕ, when θ is fixed. However, the exact expansions for this case are complicated and it is difficult to solve the Euler-Lagrange equations. We leave this as an open question to be treated elsewhere.

For the present we note that Datta and Ghosh (1996), see also Datta and Ghosh (1995), examined several desirable properties of noninformative priors in a number of examples. Consider the $Normal(\mu,\sigma^2)$ distribution in which the parameter of interest is $\theta = \mu/\sigma$ and the nuisance parameter is $\phi = \sigma$. In this case, the Chi-squared reference prior $w(\phi,\sigma) \propto |I(\theta,\phi)|^{-1/2} \propto \sigma$ satisfies Stein's condition as can be seen by examination of the proof of Theorem 1 in

Datta and Ghosh (1996). However, the priors derived here are not invariant. The Chi-squared distance is a function of the density ratio which is invariant with respect to transformations of the parameter space, but the invariance is lost when one takes the expectation over the data. We note that George and McCulloch (1992) restricted attention to invariant measures of distance and obtained generalizations of Jeffreys prior. They were determinants of Hessian matrices obtained from the distance and the parametric family.

3.3 *Computational results.* Although the most appropriate comparison would be of our prior to a subjectively chosen prior (since that is the point of a reference prior) we consider the special case that this subjectively chosen prior is Jeffreys prior. That is, we are supposing that the experimenter believes he has the extra knowledge that the experiment is achieving optimal data transmission (or compression) and so has been led to believe that use of Jeffreys prior is justified. As a general rule, in the absence of this extra knowledge we do not expect the noninformative prior to perform well, but rather to give larger credible regions.

We consider two examples, the Binomial(n, θ) and the exponential(θ) (with mean $1/\theta$). For both cases we use a sample size of $n = 5$. This is relatively small. However, it is for small sample sizes that the prior influence is greatest and a virtue of the Bayesian approach is that it permits the addition of extra information by way of the prior. In fact, behavior of the posterior for other sample sizes is qualitatively the same for the Binomial example although some differences emerge for the exponential family. In general, the credible sets under the Chi-squared reference prior are skewed toward ranges of the parameter which correspond to low Fisher information.

In the Binomial cases, Figure 1 shows the posterior only for $x = 0, 1, 2$. (By symmetry, these are the same as for $x = 3, 4, 5$.) It is seen that the Chi-squared reference prior gives a proper posterior in all three cases and, heuristically, that the spread of the posterior under $1/\sqrt{I(\theta)}$ varies less than the spread of the posterior under Jeffreys prior, as the data change. This is due to the fact that Jeffreys prior here is $1/\sqrt{p(1-p)}$ which is unbounded at 0,1 with a minimum at $p = 1/2$ whereas the Chi-squared reference prior is $\sqrt{p(1-p)}$ which is well defined on [0,1] with a maximum at $p = 1/2$.

More specifically, we can compare, for instance, 90 percent highest posterior density (HPD) sets. In Table 1 the length of 90 percent HPD sets are given for three values of x and both posteriors. It is seen that the 90 percent HPD regions for the Chi-squared reference prior are less variable in length than for the relative entropy reference prior, Jeffreys prior. For central values of x Jeffreys prior gives longer credibility sets than does the Chi-squared reference prior, and for extreme values of x Jeffreys prior gives shorter credibility sets. This reflects the fact that Jeffreys prior puts least weight on the center of the interior of the parameter space.

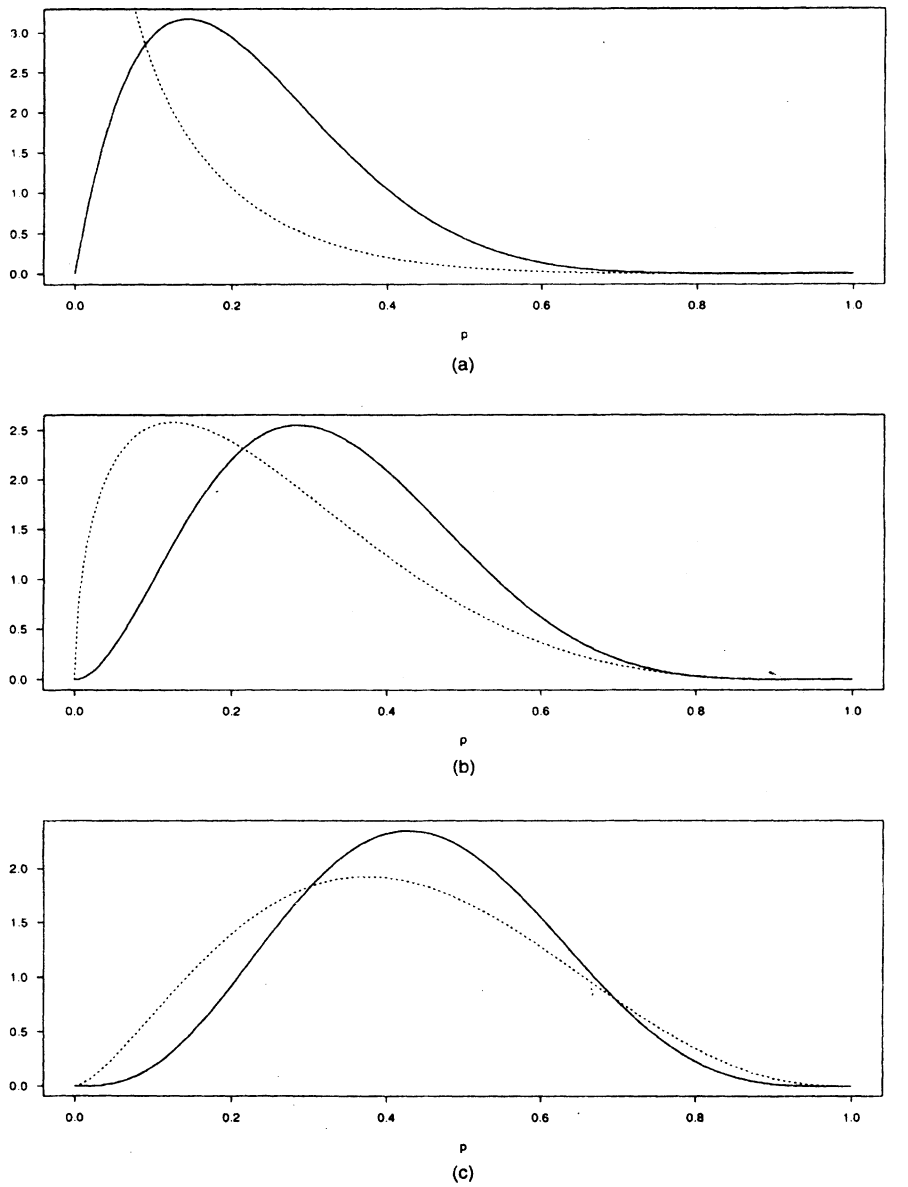

Figure 1: Posterior Densities of p for Given (a) $\sum_{i=1}^{5} X_i = 0$, (b) $\sum_{i=1}^{5} X_i = 1$ and (c) $\sum_{i=1}^{5} X_i = 2$, where $X_1, \cdots, X_5$ are i.i.d. from Bernoulli(p).

..... Using the Jeffreys prior; ——— Using the χ^2- reference prior.

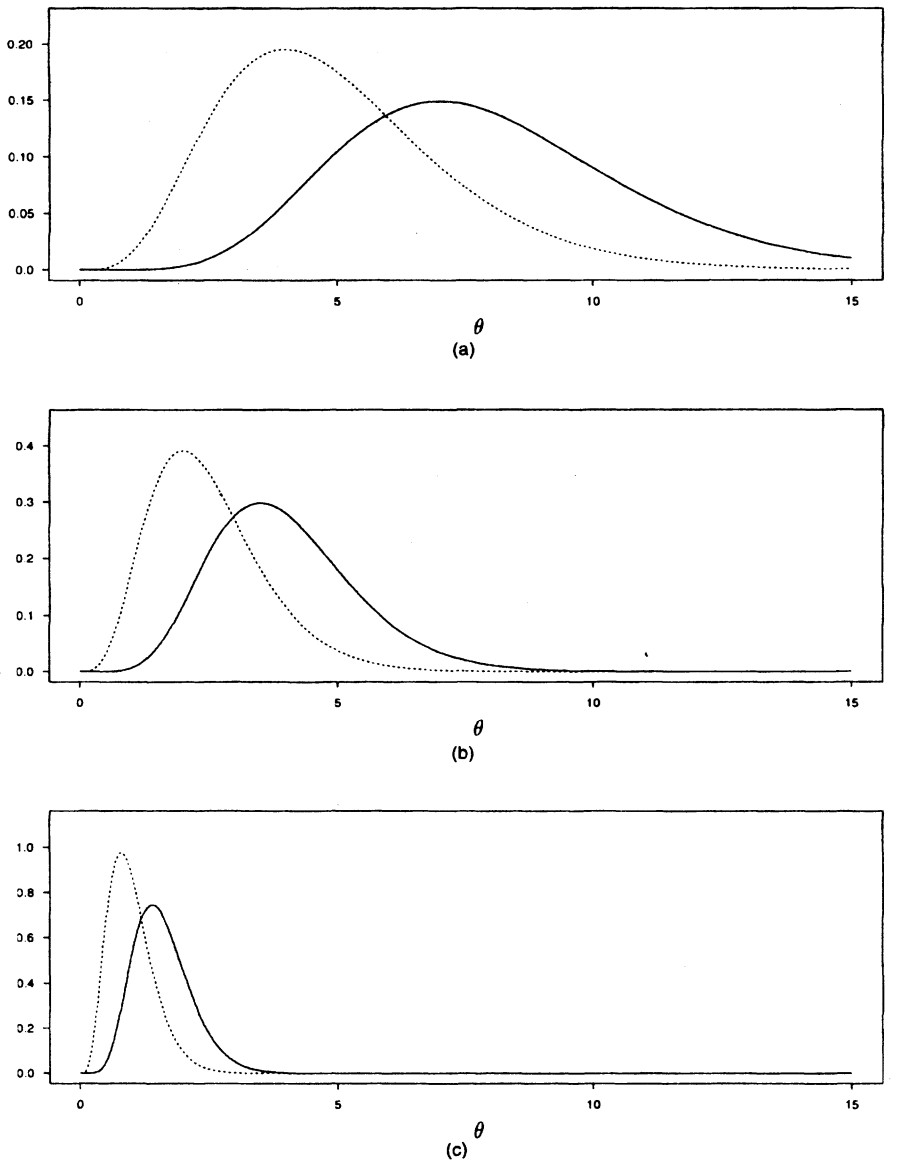

Figure 2: Posterior Densities of θ for Given (a) $\sum_{i=1}^{5} X_i = 1$, (b) $\sum_{i=1}^{5} X_i = 2$ and (c) $\sum_{i=1}^{5} X_i = 5$, where $X_1, \cdots, X_5$ are i.i.d. exponential random variables with common mean $1/\theta$.

..... Using the Jeffreys prior; ——— Using the χ^2- reference prior.

TABLE 1. LENGTHS OF POSTERIOR 90% HPD CREDIBLE SETS OF p FOR GIVEN $X = \sum_{i=1}^{5} X_i$, WHERE $X_1, \ldots, X_5$ ARE I.I.D. FROM BERNOULLI (p).

data	Jeffreys 90% HPD length	the χ^2 - 90% HPD length
$X = 0$	.2435	.3972
$X = 1$	.4824	.4798
$X = 2$	.6071	.5169

In the exponential example, we get some similar results and some different results. Here, Figure 2 shows that both posteriors are well defined and, as before, Jeffreys prior always gives tighter posteriors. Recall that we expect $1/\bar{X}$ to be near the mean of the posterior. Now, when $\sum_{i=1}^{5} x_i = 1$, we get $\bar{X} = .2$ suggesting the mean ought to be near 5. It is seen that for this case the posterior based on Jeffreys prior is centered closer to 5 than the posterior based on $1/\sqrt{I(\theta)}$ and in addition is more concentrated. (Other values of $\bar{X}$ give a similar observation.) However, the Chi-squared reference prior is more conservative in that it assigns more mass to the θ's which are hardest to discriminate, and these tend to be the values which are far from what one would expect if one believed Jeffreys prior.

The extra weight the Chi-squared reference prior gives to large parameter values protects an experimenter if these hard to estimate parameter values are important. This may be useful for estimating dispersion parameters. In this example, all the HPD regions for the Chi-squared reference prior are larger than those from Jeffreys prior and are shifted to the right. In Table 2 we see that as $\sum x_i$ varies the Jeffreys prior intervals stay at about 83% of the length of the $1/\sqrt{I(\theta)}$ intervals in all three cases.

TABLE 2. POSTERIOR 90% HPD CREDIBLE SETS OF θ FOR GIVEN $X = \sum_{i=1}^{5} X_i$, WHERE $X_1, \ldots, X_5$ ARE I.I.D. EXPONENTIAL RANDOM VARIABLES WITH COMMON MEAN $1/\theta$.

	Jeffreys 90% HPD		the χ^2 - 90% HPD	
data	credible set	length	credible set	length
$X = 0$	(1.508663, 8.355397)	6.846734	(3.467376, 12.371172)	8.903796
$X = 1$	(0.754332, 4.177699)	3.423367	(1.733688, 6.185586)	4.451898
$X = 2$	(1.508663, 8.355397)	1.369347	(0.693475, 2.474234)	1.780759

4. Issues Raised by Change of Noninformativity Principle

What do we want a prior to do? Even if we only want the prior to reflect pre-experimental beliefs we must have a benchmark for assessing the information in the prior. That is, we want to compare the posterior location and dispersion from our subjective prior to the results from the use of benchmark priors. These benchmark priors gain legitimacy from the noninformativity principle from which they are derived.

Consider the following example. Suppose we have a subjective prior and we want to assess the location and dispersion of the posterior it generates. For the location, we would want two priors, one which locates the posterior rapidly at the true value, and one which does so very slowly. Likewise for the dispersion, we want one prior which gives a posterior that concentrates rapidly, and one which gives a posterior that concentrates slowly. These would permit evaluations of the impact of subjective information on inferences. The computational results here and elsewhere show that Jeffreys prior locates the posterior rapidly. This is to be expected from its optimality property which represents the extra information of optimal transmission. So, Jeffreys prior can be used for comparisons of location estimates. On the other hand, the reference prior used here optimises a 'badness-of-fit' criterion and so gives a sort of worst case for comparisons of dispersion estimates, as the Chi-squared interpretation suggests.

In contrast to other measures of distance which have been used for prior selection, for instance the Kullback-Leibler distance, the Chi-squared distance is stronger. The stronger the measure of distance is, the more weight the resulting prior will have to assign to those values of the parameter amongst which it is hardest to discriminate. Consequently, the worse in practice the resulting prior will be. Because of this dependence on distance, it is difficult to have an absolute notion of noninformativity. Indeed, if one were to use a measure of distance weaker than the relative entropy then one would expect a prior which weighted easily discriminable parameters more than Jeffreys prior does.

For instance, consider the Hellinger metric which has a geometric interpretation in the context of packing spheres in a simplex. We see that

$$E_m \int (\sqrt{w(\theta)} - \sqrt{w(\theta|x)})^2 \, d\theta = 2(1 - \int \sqrt{w(\theta)} E_m \sqrt{w(\theta|x)} d\theta),$$

so techniques analogous to those used here may apply to the expected square root of the posterior. This may generate a differential equation whose solution is a prior that performs 'better' than Jeffreys prior does. We anticipate that it, too, would be a power of the determinant of the Fisher information matrix higher than 1/2. If true, this lends weight to the proposal that priors based on powers of the Fisher information be used for sensitivity analysis.

Finally, we note that for the Chi-squared distance the reference prior quantity has a decision-theoretic interpretation from $E_m \chi^2(w(\cdot|X^n), w(\cdot)) = \int w(\theta) \chi^2$

$(p^n(\cdot|\theta), m_n(\cdot))$. That is, the reference prior quantity is the Bayes risk of using the mixture distribution as an estimator for the true density p_θ^n. By contrast, if the Chi-squared distance is used as a loss function then the Bayes risk is the minimum of $\int w(\theta) E_\theta \chi^2(p_\theta, q) d\theta$ as q ranges over all estimators of p_θ. One can verify that the Bayes estimator under Chi-squared loss is the standardized square root of $\int p_\theta^2(X^n) w(\theta) dX^n \Big/ \int p_\theta^2(X^n) w(\theta) dX^n d\theta$. For the Hellinger distance, the Bayes estimator is $(\int w(\theta) \sqrt{p_\theta} d\theta)^2$ upon standardization.

ACKNOWLEDGEMENT. The authors thank Dr. Ian McKay for helpful conversations about goodness of fit and an anonymous referee for invaluable comments made after critically reading an earlier version of this manuscript.

References

BERGER, J.O. and BERNARDO, J.M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84** 200-207.

– – –– (1992a). Ordered group reference priors with applications to multinomial and components of variance problems. *Biometrika*, **79**, 25-38.

– – –– (1992b). On the development of reference priors. In *Bayesian Analysis IV*, J.M. Bernardo, *et. al.*, (Eds.). Oxford University Press, Oxford.

BERNARDO, J.M. (1979) Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **41** 113-147.

MUKERJEE, R. and DEY, K.D. (1993). Frequentist validity of posterior quantiles in the presence of a nuisance parameter: higher order asymptotics. *Biometrika* **80**, 499-505.

CHANG, T. and EAVES, D.M. (1990). Reference priors for the orbit in a group model. *Ann. Statist.* **18-4** 1595-1614.

CLARKE, B. and BARRON, A.R. (1994) Jeffreys prior is asymptotically least favorable under entropy risk. *J. Statistical Planning and Inference* **41** 37-60.

CLARKE, B. and SUN, D.C. (1993). Reference priors under the Chi-square distance. *Technical Report*, Department of Statistics, University of British Columbia.

DATTA, G.S. and GHOSH, M. (1996). On the invariance of noninformative priors. To appear in *Annals of Statistics.*

– – –– (1995). Some remarks on noninformative priors. To appear in *J. Amer. Statist. Assoc.*, December.

EAVES, D.M. (1985). On maximising missing information about a hypothesis. *J. R. Statist.* Soc. B, **47-2** 263-266.

GEORGE, E. and MCCULLOCH, R. (1989). On obtaining invariant prior distributions. *Technical Report*, Graduate School of Business, University of Chicago.

HARTIGAN, J.A. (1965). The asymptotically unbiased prior distribution. *Ann. Statist.* **36-4** 1137-1152.

– – –– (1983). *Bayes Theory*, Springer-Verlag, New York.

JEFFREYS, H. (1961). *Probability Theory*, Oxford University Press, New York.

KASS, R.E. and WASSERMAN, L. (1996). Formal rules for selecting prior distributions: a review and annotated bibliography, To appear *J. Amer. Statist. Assoc.*

PEERS, H.W. (1965). On confidence sets and Bayesian probability points in the case of several parameters. *J. Royal Statist. Soc.*, Ser. B, **27**, 9-16.

SUN, D. and YE, K. (1996). Frequentist validity of posterior quantiles for a two-parameter exponential family. *Biometrika*, **83**.

TIBSHIRANI, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76** 604-608.

DEPARTMENT OF STATISTICS
UNIVERSITY OF BRITISH COLUMBIA
6356 AGRICULTURAL ROAD, ROOM 333
VANCOUVER
CANADA
riffraff.stat.ubc.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF MISSOURI-COLUMBIA
COLUMBIA, MO 65211
USA