

Entropies and Metrics That Lead to Jeffreys and Reference Priors

J.K. Ghosh, Arijit Chakrabarti and Tapas Samanta
Purdue University and Indian Statistical Institute

1. Introduction

The most popular objective priors are the reference priors of Bernardo and Berger, with scalar parameters arranged according to order of importance and a step by step algorithm that constructs conditional prior given the parameters appearing in the earlier order. It is perhaps not unfair to say they are descended from the Jeffreys prior with some of the standard objections to that prior well taken care of. Many different orderings are possible but usually only a small number of them turn out to be important for one reason or other. See for example Bernardo (1979) Berger and Bernardo (Proc 4th Valencia Conference, 1991), Berger and Sun (AS, 2008, Proc. of Valencia Conference, 2006), Berger, Bernardo and Sun ((much expected) monograph on Reference priors). Moreover, with reference priors as originally defined, the Jeffrey's prior is just a reference prior with all parameters given equal importance. Our aim is to explore the simplest form of construction as given in Bernardo (1979) to see if this throws some further light on these priors, Jeffrey's and (one-at-a-time) reference priors.

In this write-up we consider only the Jeffrey's prior for simplicity.

In Section 2 we briefly survey the three different methods of construction of an objective prior that lead to Jeffreys prior. This begins with Bernardo's method based on maximizing a measure of difference between prior and posterior which arises from Shannon's notion of missing entropy. A second method is based on probability matching, i.e., matching of frequentist and posterior probability distributions of the approximate pivotal quantity. A third method, due to Jeffreys and rediscovered by Ghosal, Ghosh and Ramamoorthi (1997), is to obtain the Jeffreys prior as a weak limit of discrete uniforms on points obtained by packing Hellinger spheres of smaller and smaller diameters. The Berger-Bernardo type ordering and step-by-step algorithm can be applied to the third case and leads to priors resembling reference priors but involving arithmetic mean rather than geometric means of information measure as in the Berger-Bernardo algorithm. The same sort of thing can also be done with probability matching, but only for two parameters, and then instead of geometric or arithmetic means, one gets harmonic means of the information numbers (see e.g. Ghosh and Ramamoorthi (2003, pp. 224-229).)

In general the reference priors based on geometric means seem more tractable and lead to proper posteriors more frequently. We conjecture, if conjecture is the right word, that this is because geometric means lead

to products (rather than sums or inverse of sums) which blend better with likelihood.

In Section 3 we explore the main aim of this paper, which is to see which kind of measures of divergence between posterior and prior would lead to the Jeffrey’s prior after maximization and if such measures are relatively easy to characterize. In fact our first conjecture was that the only other such measure is likely to be the Hellinger distance and that most likely this property will not hold at least for the L_1 -norm. It turns out that Hellinger distance also leads to the Jeffreys prior, when maximized, and contrary to our intuition, the same is also true for the L_1 -norm. However, the property fails for L_2 -norm. We still do not understand why this is so, but some possibilities are discussed. We also observe an interesting decomposition of the divergence, which is similar in the three divergence measures leading to Jeffreys prior. It seems to us a similar decomposition for reference priors may help us understand how inference for different parameters is affected by choice of different reference priors for different sample size. That may help in choosing our objective prior from among different reference priors.

In section 4 we explore an application of these ideas to the problem of choosing an objective prior for a “real life” example involving random effects studied in Gelman (2006). We compare the different priors, old and new, proposed by Gelman and compare them through the numerical values of the three different measures which lead to the Jeffreys prior.

2. Methods of Construction of an Objective Prior Leading to Jeffreys Prior

Jeffreys prior is one of the widely used objective priors which is defined as

$$\pi(\theta) = [\det(I_{ij}(\theta))]^{1/2},$$

where $(I_{ij}(\theta))$ denotes the Fisher information matrix. We briefly describe below some methods of construction of an objective prior that lead to Jeffreys prior.

2.1 Jeffreys prior as Minimizer of Information

Let $\mathbf{X} = (X_1, X_2, \dots, X_n) \sim p(\mathbf{x}|\boldsymbol{\theta})$, $\mathbf{x} \in \mathcal{X}$, $\boldsymbol{\theta} \in \Theta \subset \mathcal{R}^d$, where $\boldsymbol{\theta}$ has prior density $p(\boldsymbol{\theta})$. Let $p(\boldsymbol{\theta}|\mathbf{x})$ and $m(\mathbf{x})$ denote respectively the posterior density of $\boldsymbol{\theta}$ given \mathbf{x} and the marginal density of \mathbf{x} . Using an idea of Lindley, Bernardo (1979) considers the average Kullback-Leibler divergence between prior and posterior, namely,

$$\begin{aligned} J(p(\cdot)) &= E \left\{ \log \frac{p(\boldsymbol{\theta}|\mathbf{X})}{p(\boldsymbol{\theta})} \right\} \\ &= \int_{\mathcal{X}} \left[\int_{\Theta} \log \left\{ \frac{p(\boldsymbol{\theta}|\mathbf{x})}{p(\boldsymbol{\theta})} \right\} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \right] m(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (2.1)$$

$$= \int_{\Theta} \left\{ \int_{\mathcal{X}} \left[\int_{\Theta} \log \left\{ \frac{p(\boldsymbol{\theta}'|\mathbf{x})}{p(\boldsymbol{\theta}')} \right\} p(\boldsymbol{\theta}'|\mathbf{x}) d\boldsymbol{\theta}' \right] p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \right\} p(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad (2.2)$$

and suggests that $-J$ is a measure of information in the prior. Thus a prior that maximizes J may be considered to be noninformative. It is also observed that the functional J is expected to be a nice function of the prior only asymptotically. We briefly describe below how the asymptotic maximization is done.

Fix an increasing sequence of compact rectangle K_i whose union is the whole parameter space Θ . Fix i and consider only priors p_i supported on K_i and let $n \rightarrow \infty$.

We assume X_1, \dots, X_n are i.i.d. and also assume conditions under which posterior distribution is asymptotically normal in appropriate sense (see Clarke and Barron (1990, 1994)). Then the functional J can be suitably approximated by

$$\begin{aligned} \hat{J}(p_i) &= \int_{K_i} \left\{ \int_{\mathcal{X}} \left[\int_{\mathcal{R}^d} \log \hat{p}_i(\boldsymbol{\theta}'|\mathbf{x}) \hat{p}_i(\boldsymbol{\theta}'|\mathbf{x}) d\boldsymbol{\theta}' \right] p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \right\} p_i(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\quad - \int_{K_i} (\log p_i(\boldsymbol{\theta})) p_i(\boldsymbol{\theta}) d\boldsymbol{\theta} \end{aligned} \quad (2.3)$$

where \hat{p}_i is the approximating d -dimensional normal distribution $N(\hat{\boldsymbol{\theta}}, I^{-1}(\boldsymbol{\theta})/n)$. We now use the well-known fact about the exponent of a multivariate normal density that

$$\int -\frac{1}{2}(\boldsymbol{\theta}' - \hat{\boldsymbol{\theta}})^t (I^{-1}(\boldsymbol{\theta})/n)^{-1} (\boldsymbol{\theta}' - \hat{\boldsymbol{\theta}}) \hat{p}_i(\boldsymbol{\theta}'|\mathbf{x}) d\boldsymbol{\theta}' = -\frac{d}{2}$$

and get

$$\hat{J}(p_i) = \left\{ -\frac{d}{2} \log(2\pi) - \frac{d}{2} + \frac{d}{2} \log n \right\} + \int_{K_i} \log \left\{ \frac{(\det(I(\boldsymbol{\theta})))^{\frac{1}{2}}}{p_i(\boldsymbol{\theta})} \right\} p_i(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

The prior that maximizes the above is given by

$$p_i(\boldsymbol{\theta}) = \begin{cases} c_i [\det(I(\boldsymbol{\theta}))]^{1/2} & \text{on } K_i; \\ 0 & \text{elsewhere} \end{cases} \quad (2.4)$$

where c_i is a normalizing constant such that (2.4) is a probability density on K_i .

For every K_i , the (normalized) jeffreys prior $p_i(\boldsymbol{\theta})$, given by (2.4) maximizes $J(p)$ where K_i increases to the whole of Θ . Thus, in a sense, Jeffreys prior is obtained as a maximizer of Bernardo's measure J or Minimizer of information ($-J$) in a prior.

2.2 Other Methods Leading to Jeffreys Prior

A noninformative prior may be expected to provide inference that is similar to frequentist inference. Consider, for example the case when observations are i.i.d. $N(\theta, 1)$ and the improper uniform prior on \mathcal{R} is used. Then the frequentist distribution of the pivotal quantity $\theta - \bar{X}$ given θ is identical with the posterior distribution of $\theta - \bar{X}$ given \mathbf{X} . In the general case one may obtain priors by matching posterior and frequentist probability distribution of the approximate pivotal quantity upto a certain order of approximation. A precise definition of a probability matching prior for a single parameter is given below.

Let X_1, X_2, \dots, X_n be i.i.d. $p(x|\theta)$, $\theta \in \Theta \subset \mathcal{R}$. Assume regularity conditions needed for expansion of the posterior with the normal $N(\hat{\theta}, (nI(\hat{\theta}))^{-1})$ as the leading term. For $0 < \alpha < 1$, choose $\theta_\alpha(\mathbf{X})$ depending on the prior $p(\theta)$ such that

$$P\{\theta \leq \theta_\alpha(\mathbf{X})|\mathbf{X}\} = 1 - \alpha + O_p(n^{-1}). \quad (2.5)$$

It can be verified that $\theta_\alpha(\mathbf{X}) = \hat{\theta} + O_p(1/\sqrt{n})$. We say $p(\theta)$ is probability matching (to first order), if

$$P_\theta\{\theta \leq \theta_\alpha(\mathbf{X})\} = 1 - \alpha + O(n^{-1}) \quad (2.6)$$

(uniformly on compact sets of θ). In the normal case with $p(\theta) = \text{constant}$,

$$\theta_\alpha(X) = \bar{X} + z_\alpha/\sqrt{n}$$

where $P\{Z > z_\alpha\} = \alpha$, $Z \sim N(0, 1)$.

We have matched posterior probability and frequentist probability up to $O_p(n^{-1})$. For any prior $p(\theta)$ satisfying some regularity conditions, the two probabilities mentioned above agree up to $O(n^{-\frac{1}{2}})$, so $O(n^{-\frac{1}{2}})$ is too weak. On the other hand, if we strengthen $O(n^{-1})$ to, say, $O(n^{-\frac{3}{2}})$, in general no prior would be probability matching. So $O(n^{-1})$ is the right order.

We view probability matching in a slightly different but equivalent way. Instead of working with $\hat{\theta}_\alpha(\mathbf{X})$ one may choose to work with the approximate quantile $\hat{\theta} + z_\alpha/\sqrt{n}$ and require

$$P\{\theta < \hat{\theta} + z_\alpha/\sqrt{n}|\mathbf{X}\} = P_\theta\{\theta < \hat{\theta} + z_\alpha/\sqrt{n}\} + O_p(n^{-1}) \quad (2.7)$$

under θ (uniformly on compact sets of θ).

Each of these two probabilities has an expansion starting with $(1 - \alpha)$ and having terms decreasing in powers of $n^{-\frac{1}{2}}$. So for probability matching, we must have the same next term in the expansion.

In principle one would have to expand the probabilities and set the two second terms equal, leading to

$$-\frac{d(I(\theta))^{-1/2}}{d\theta} = (I(\theta))^{-1/2} \frac{1}{p(\theta)} \frac{dp(\theta)}{d\theta}. \quad (2.8)$$

The left-hand side comes from the frequentist probability, the right-hand side from the posterior probability (taking into account the limits of random quantities under θ). For details see Ghosh (1994, Chapter 9) or Ghosh and Mukerjee (1992) or Datta and Mukerjee (2004).

Clearly, the unique solution to (2.8) is the Jeffreys prior

$$p(\theta) \propto \sqrt{I(\theta)}.$$

2.3 The Jeffreys as a Limit of Discrete Uniforms

Consider a parametric space K which is compact in its metric ρ and a compact subset K_1 (which may be K itself). Let $D(\epsilon, K)$ be the packing, i.e. the number of balls of radius ϵ that can be packed into K .

A sort of discrete uniform probability of K_1 is $P_\epsilon(K_1) = \frac{D(\epsilon, K_1)}{D(\epsilon, K)}$. It can be proved, vide Theorem 8.3.1 due to Dembski (J. Theoretical Prob. (1990, 611-626)) and Theorem 8.4.1 due to Ghosal, Ghosh and Ramamoorthi (1997), that a limit exists under regularity conditions of $K \subset R^k$ and ρ =Hellinger metric. A similar result holds for the Kullback-Leibler divergence measure.

A simple heuristic explanation for both result is that both those measures are locally like Rao's Riemannian metric $\rho(\theta, \theta + d\theta) = \sum \sum I_{ij}(\theta) \partial\theta_i \partial\theta_j (1 + o(1))$.

It is hard to believe that the Jeffreys prior can be limit of discrete uniforms in the above sense if the above local representation fails to hold. For example the L_1 -metric is not locally like Rao's metric.

3. Entropies and Metrics Leading to Jeffreys Prior

The Lindley-Bernardo measure $J(p)$, considered in Section 2.1, was based on the average Kullback-Leibler divergence between prior and posterior. It was shown in that section how Jeffreys prior is obtained as maximizer of this measure. The question that we try to address in this section is for which kind of divergence measures, the corresponding functionals $J(p)$, when maximized with respect to the prior, lead to Jeffreys prior. We only have partial answer to this question. We find that the Lindley-Bernardo measure is not unique in this respect. For simplicity in illustration we consider the case with a real parameter.

An important property of the (average) Kullback-Leibler divergence is that it is invariant under smooth one-to-one transformation of the parameter. This is because for a smooth one-to-one function $\eta = g(\theta)$,

$$\frac{p(\theta|\mathbf{x})}{p(\theta)} = \frac{p^\eta(\eta|\mathbf{x})}{p^\eta(\eta)}$$

and therefore,

$$J(p(\cdot)) = \int_{\mathcal{X}} \int \frac{p(\theta|\mathbf{x})}{p(\theta)} p(\theta|\mathbf{x}) d\theta dm(\mathbf{x})$$

$$= \int_{\mathcal{X}} \int \frac{p^\eta(\eta|\mathbf{x})}{p^\eta(\eta)} p^\eta(\eta|\mathbf{x}) d\eta dm(\mathbf{x})$$

Thus the maximizing prior must be invariant under smooth one-to-one transformations.

We now consider a divergence measure $D(p(\cdot), p(\cdot|\mathbf{x}))$ between the prior and the posterior and the corresponding functional

$$\begin{aligned} J(p(\cdot)) &= \int_{\mathcal{X}} D(p(\cdot), p(\cdot|\mathbf{x})) m(\mathbf{x}) d\mathbf{x}. \\ &= \int \int D(p(\cdot), p(\cdot|\mathbf{x})) p(\mathbf{x}|\theta) d\mathbf{x} p(\theta) d\theta \end{aligned} \quad (3.1)$$

Examples of other divergence measures D include the Hellinger distance

$$\left\{ \int |p^{\frac{1}{2}}(\theta) - p^{\frac{1}{2}}(\theta|\mathbf{x})|^2 d\theta \right\}^{\frac{1}{2}}$$

and the L_r -distance, $r > 0$,

$$\left\{ \int |p(\theta) - p(\theta|\mathbf{x})|^r d\theta \right\}^{\frac{1}{r}}.$$

We consider a divergence measure of the form

$$D(p(\cdot), p(\cdot|\mathbf{x})) = \left\{ \int_{\Theta} d\left(\frac{p(\theta)}{p(\theta|\mathbf{x})}, 1\right) p^r(\theta|\mathbf{x}) d\theta' \right\}^s \quad (3.2)$$

for some $d(\cdot, \cdot)$ and $r, s > 0$.

For the L_r -distance, $d(u, v) = |u - v|^r$, $s = \frac{1}{r}$, for the Hellinger distance, $d(u, v) = (\sqrt{u} - \sqrt{v})^2$, $r = 1$, $s = \frac{1}{2}$ and for the Kullback-Leibler divergence, $d(u, v) = -\log(u/v)$, $r = 1$, $s = 1$.

Note that the average divergence measure $J(p(\cdot))$ is invariant under smooth one-to-one transformation for Kullback-Leibler divergence, the Hellinger distance and the L_1 -distance but not for the L_r -distance with $r \neq 1$.

As in Section 2.1 we consider a sequence of increasing compact sets whose union is the whole of Θ . We fix such a compact set $[a, b]$ and consider priors $p(\cdot)$ supported on $[a, b]$. Assuming that the posterior $p(\theta|\mathbf{x})$ can be approximated by $N(\hat{\theta}, (nI(\theta))^{-1})$ distribution in appropriate sense we approximate $J(p)$ as

$$\int \int \left\{ \int d\left(\frac{p\left(\hat{\theta} + t/\sqrt{nI(\theta)}\right)}{\sqrt{nI(\theta)}} / \phi(t), 1\right) (nI(\theta)/2\pi)^{(r-1)/2} \phi^r(t) \right\}^s p(\mathbf{x}|\theta) d\mathbf{x} p(\theta) d\theta$$

where $\phi(t)$ is the standard normal density. Here we use the change of variable $t = \sqrt{nI(\theta)}(\theta' - \hat{\theta})$.

We give a brief sketch of proof for the Hellinger distance and the L_1 -distance.

For the L_1 -distance, we use the standard result on posterior normality which states that the L_1 -distance between the posterior and the approximating normal distribution converges to zero with probability one. In view of the inequalities that compare the Hellinger distance and the L_1 -distance (see, for example, Corollary 1.2.1, Ch.1, p.14 of Ghosh and Ramamoorthi (2003)), a similar result also holds for the Hellinger distance.

Consider first the case with L_1 -distance. Using the normal approximation the measure $J(p)$, as given in (3.1), can be approximated by

$$\hat{J}(p) = \int_a^b \int \left| \frac{p\left(\hat{\theta} + t/\sqrt{nI(\theta)}\right)}{\sqrt{nI(\theta)}} - \phi(t) \right| dt p(\mathbf{x}|\theta) d\mathbf{x} p(\theta) d\theta.$$

Now note that for fixed θ ,

$$\begin{aligned} & \int \left| \frac{p\left(\hat{\theta} + t/\sqrt{nI(\theta)}\right)}{\sqrt{nI(\theta)}} - \phi(t) \right| dt \\ &= 2 \int_{t \in A} \left[\phi(t) - \frac{p\left(\hat{\theta} + t/\sqrt{nI(\theta)}\right)}{\sqrt{nI(\theta)}} \right] dt \end{aligned}$$

$$\begin{aligned} \text{where } \{t \in A\} &= \left\{ t : \phi(t) > \frac{p\left(\hat{\theta} + t/\sqrt{nI(\theta)}\right)}{\sqrt{nI(\theta)}} \right\} \\ &= \left\{ t : |t| < \left(\log n + \log(I(\theta)/2\pi) - 2 \log p\left(\hat{\theta} + t/\sqrt{nI(\theta)}\right) \right)^{\frac{1}{2}} \right\}. \end{aligned}$$

Assuming that both $I(\theta)$ and $p(\theta)$ are nice functions on the compact parameter set $[a, b]$, and using properties of $\phi(t)$, $J(p)$ can be approximated as

$$2 - c_n \int_a^b \frac{p(\theta)}{\sqrt{I(\theta)}} p(\theta) d\theta \tag{3.3}$$

for some constant c_n .

Using the inequality between arithmetic mean and geometric mean we see that (3.3) is maximized by (normalized) Jeffrey prior restricted to $[a, b]$.

Similarly, for the Hellinger distance, using normal approximation to posterior and a change of variable $t = \sqrt{nI(\theta)}(\theta' - \hat{\theta})$, we approximate $J(p)$

as

$$\hat{J}(p) = \int \int \left[2 - 2 \int \frac{1}{(2\pi)^{1/4}} e^{-\frac{1}{4}t^2} \frac{p^{1/2}(\hat{\theta} + t/\sqrt{nI(\theta)})}{(nI(\theta))^{1/4}} dt \right]^{1/2} p(\mathbf{x}|\theta) d\mathbf{x} p(\theta) d\theta$$

which is then approximately equal to

$$\int_a^b \left[2 - c_n \frac{p^{1/2}(\theta)}{(I(\theta))^{1/4}} \right]^{1/2} p(\theta) d\theta,$$

where c_n is a constant. By Jensen's inequality this is maximized by the (normalized) Jeffreys prior restricted to $[a, b]$.

If, however, one considers the L_2 -distance, one can show that the corresponding measure $J(p)$ is not maximized by the Jeffreys prior.

4. Simulations

Prior	L_1 distance	Hellinger	KL Distance
Uniform Prior over (0, 1000) (8 Schools Data)	1.93	1.30	4.06
Inverse Gamma(1,1) (8 Schools Data)	0.052	0.1	0.012
Half-Cauchy (8 Schools Data)	1.056	0.755	1.697
Half-Cauchy (3 schools data)	0.736	0.520	1.040
Uniform Prior over (0, 1000) (3 schools data)	1.562	1.067	2.157

Table 1: Table 1

Note:- 3 Schools data means data based on the first three schools of Table 2.

The model is as follows : y_1, \dots, y_j are independent.

- 1) Y_j follows $N(\theta_j, \sigma_j^2)$, σ_j^2 are known (third column of Table 2)
- 2) θ_j are iid $N(\mu, \sigma^2)$
- 3) μ has noninformative prior $N(0, 10^6)$
- 4) σ^2 has prior as described in first column of table.

The table results give the value of $D(p(\cdot), p(\cdot|x))$ as defined in the write-up for the given data. Approximation to the integral done using simulations from the on posterior distributions.

The Data can be found in Page 140 of Bayesian Data Analysis, Second Edition by Gelman, Carlin, Stern and Rubin and it is described there and in the paper of Gelman (2006). We give below the data as copied from the book in Table 2.

School	Estimated treatment effect y_j	Standard Error of effect estimate σ_j
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Table 2: Table 2

5. Concluding Remarks

We have shown that the Jeffreys prior arises from at least two different measures of distance (Hellinger, L_1) and a measure of divergence (Bernardo-Lindley). However no simple local characterisation (vide section 2.3) seems possible. If one maximizes in a neighbourhood of normalized $\sqrt{I(\theta)}$, it appears that one requires local properties of the measure. This would still make the result for L_1 somewhat odd.

In all three cases, the measure has an asymptotic expansion of terms with diminishing magnitude, of which the first two are same for all priors and the third depends on the prior. This suggests that as n increases the difference of reference priors will tend to diminish and one would have robust O'Bayes inference. This phenomenon seems to be different from results showing the effect of prior being washed away for large n . This second well-known fact would normally require much larger n than what is needed for the difference between two reference priors to be small.

Finally, the stress on suitable invariance of the measure would seem to require that information in a prior can only be captured by such invariant measures along with simultaneous presentation of prior-posterior diagrams as in Gelman(2006). Information in a prior cannot be measured except in the context of a model.

6. Appendix: Based on heuristics

Based on the three special cases we indicate how we may obtain a more general result on measures of divergence between prior and posterior which are maximized by the Jeffreys prior. We consider the one dimensional case only for simplicity of notation, generalization is straightforward. Let $\mathbf{x} = (x_1, \dots, x_n)$. We start with a general divergence $D_p(\mathbf{x}) = D(p(\cdot), p(\cdot|\mathbf{x}))$ between prior and posterior. Conditions will be added essentially on D as we go along. To fix ideas, we note that typically, but not necessarily, $D = \int_a^b p(\theta|\mathbf{x}) d(\frac{p(\theta|\mathbf{x})}{p(\theta)}) d\theta$. (A1), where $\mathbf{x} = (x_1, \dots, x_n)$, $[a, b]$ is a bounded rectangle in Θ and the function d has suitable convexity. The function D of \mathbf{x} is then integrated with respect to \mathbf{x} , $\int D_p(\mathbf{x}) dm_p(\mathbf{x}) = \int_a^b [D_p(\mathbf{x}) p(\mathbf{x}|\theta') d\mathbf{x}] p(\theta') d\theta'$ where $m_p(\mathbf{x})$ is the marginal of \mathbf{x} , θ' maybe thought of as the "true value" of

θ_1 . We first replace the exact posterior by its normal approximation, namely, $N(\hat{\theta}, \frac{1}{nI(\hat{\theta})})$ where $\hat{\theta}$ is the mle based on \mathbf{x} . We also make a change of variables from θ to t where $t = (\theta - \hat{\theta})\sqrt{nI(\hat{\theta})}$, $d\theta = \frac{dt}{\sqrt{nI(\hat{\theta})}}$. Then the integral (A1) becomes approximately:

$\int_a^b [f D(p(\hat{\theta}+t/\sqrt{nI(\hat{\theta})}); \frac{1}{\sqrt{2\pi}} \exp^{-\frac{t^2}{2}})p(\mathbf{x}|\theta')d\theta']p(\theta')d\theta'$ (A2). Suppose the integrand is of the form $D = \int_a^b \frac{1}{\sqrt{2\pi}} \exp^{-\frac{t^2}{2}} f(\frac{p(\hat{\theta}+t/\sqrt{nI(\hat{\theta})})}{\sqrt{nI(\hat{\theta})}} / \frac{1}{\sqrt{2\pi}} \exp^{-\frac{t^2}{2}})dt$ and can

be further simplified to $\int_a^b \frac{1}{\sqrt{2\pi}} \exp^{-\frac{t^2}{2}} \Psi(\frac{p(\hat{\theta})}{\sqrt{nI(\hat{\theta})}})f(\sqrt{2\pi} \exp^{-\frac{t^2}{2}})dt(1 + 0(1)) = \Psi(\frac{p(\theta')}{\sqrt{nI(\theta')}})(constant)(1 + 0(1))$, since $\hat{\theta} \xrightarrow{p} \theta'$. Then the double integral

(A2) reduces to the one-dimensional integral $const \int_a^b \Psi(\frac{p(\theta')}{\sqrt{nI(\theta')}})p(\theta')d\theta'(1 + 0(1))$ (A3). Let c be the normalizing constant for $\sqrt{I(\theta')}$, i.e., $\int_a^b c\sqrt{I(\theta')}d\theta' = 1$. Suppose we can pull out a constant again from the integral and can write

(A3) as $[A_n - B_n \int_a^b \Psi(\frac{p(\theta')}{c\sqrt{nI(\theta')}})p(\theta')d(\theta')](1+0(1))$ (A4), where $B_n > 0$. In our three cases both A_n and B_n are > 0 and Ψ is a convex function. Assuming

both $B_n > 0$ and convexity of Ψ , (A4) is maximized by $p(\theta' = c\sqrt{I(\theta')})$, $\theta' \in [a, b]$. The integral in (A4) with convex Ψ was proposed as a measure of divergence by Ali and Silvey (JRSS, Ser B, 1966). Interestingly, this family

include the Kullback-Liebler , Hellinger and L_1 , and the α -divergences of Amari(1985, Differential Geometric Methods in Statistics), but not L_2 (see

e.g. Minka(Microsoft TR, 2005)). These divergences have become very popular recently in the machine learning literature. The variational methods of Michael Jordan are based on minimizing a Kullback-Liebler divergence.