Default priors

for Bayesian and frequentist inference

D.A.S. Fraser*, N. Reid

University of Toronto, Canada

E. Marras

Centre for Advanced Studies and Development, Sardinia

University of Rome "La Sapienza", Rome

G.Y. Yi

University of Waterloo, Canada

July 2, 2009

*Address for correspondence: Department of Statistics, University of Toronto, 100 St. George Street, Toronto, Canada M5S 3G3

Abstract

We investigate the choice of default prior for use with likelihood to facilitate Bayesian and frequentist inference. Such a prior is a density or relative density that weights an observed likelihood function leading to the elimination of parameters not of interest and then to providing a density type assessment for a parameter of interest. For independent responses from a continuous model, we develop a prior for the full parameter that is closely linked to the original Bayes approach and provides an extension of the right invariant measure to general contexts. We then develop a modified prior that is targetted on a component parameter of interest and by targetting avoids the marginalization paradoxes of Dawid, Stone and Zidek (1973). In particular this modifies the Jeffreys' prior and provides extensions to Welch-Peers theory. These two approaches are combined to develop a prior that targets a vector interest parameter in the presence of a vector nuisance parameter. Examples are given to clarify the computation of the priors.

Keywords: Haar measure; Invariance; Jeffreys prior; Likelihood asymptotics; Noninformative prior; Nuisance parameter; Objective prior; Subjective prior.

1 Introduction

We develop default priors for Bayesian and frequentist inference in the context of a statistical model $f(y;\theta)$ and observed data y^0. A default prior is a density or relative density used as a weight function applied to an observed likelihood function. The choice of prior is based directly on assumed smoothness in the model and an absence of information as to how the parameter value was generated.

One Bayesian role for a default prior is to provide a reference allowing subsequent modification by an objective, subjective, personal, elicited or expedient prior. From a frequentist viewpoint a default prior can be viewed as a device to replace integration on the sample space by integration on the parameter space and thus to use the likelihood function directly. From either view it offers a flexible and easily implemented exploratory approach to statistical inference.

There is a large literature on the construction of many types of default priors, variously called non-informative, non-subjective, or objective; a complete review is beyond the scope of this paper. The term objective prior has obvious scientific interpretation and perhaps should be reserved for contexts where it is known that θ arose from some density $g(\theta)$. A very helpful survey of various methods for constructing defaulty priors is given in Kass & Wasserman (1996). In most discussions there is at least some emphasis on ensuring correct calibration of posterior probability limits, in the sense that these limits represent probability under the model, at least approximately. A recent discussion of this appears in Berger (2006); Goldstein (2006) gives a contrary view. Our own view is that such calibration is necessary to ensure that posterior inference does not give misleading results. Calibration of Bayes procedures is reviewed in Little (2006).

Broadly speaking, approaches to default priors in the literature include those based on notions of invariance and generalized invariance, on information or divergence measures, and on the goal of matching posterior and frequentist inferences to some order of approximation. For a scalar parameter model, all of these approaches lead to Jeffreys' prior $\pi_J(\theta) \propto i^{1/2}(\theta)$, where $i(\theta)$ is the expected Fisher information in the model. Jeffreys (1961) derived this default prior based on invariance arguments, and this was further pursued by Box & Tiao (1973) as data-translated likelihoods; see Kass (1990). George & McCulloch (1993) derived a class of invariant priors and developed a link between this approach and that based on divergence methods. Divergence methods can be framed in the context of the information processing that takes a prior distribution to a posterior distribution, as in Zellner (1988). The reference prior approach of Bernardo (1979) and Berger & Bernardo (1992) seeks to maximize the Kullback-Liebler divergence between the posterior and prior distribution: Clarke & Barron (1994) related this to least favorable distributions. This approach has been extended to families of divergence measures; a recent treatment is Ghosh et al. (2009). A more direct construction of reference priors for scalar parameters is given in Berger et al. (2009). Welch & Peers (1963) derived Jeffreys' prior by a probability matching argument

based on Edgeworth exapnsions.

In extending these results to problems with nuisance parameters several difficulties arise. The Welch-Peers approach was developed in Peers (1965), Tibshirani (1989), and in several papers by Mukerjee and colleagues. A review of this literature is provided by Datta & Mukerjee (2004); see also Reid et al. (2003). These extensions addressed the construction of matching priors using asymptotic arguments based on Edgeworth expansions, and the construction turns out to be difficult, and sometimes not possible. The reference prior approach to the construction of priors in the presence of nuisance parameters involves difficulties both in the ordering of the parameters and in the construction of compact subsets of the parameter space, which are still unresolved. Clarke & Yuan (2004) give a survey of information based priors for problems with nuisance parameters. Jeffreys (1961) recognized that his arguments based on invariance led to unsuitable priors in regression-scale problems, and recommended a modified approach treating location and scale parameters as independent: see Kass & Wasserman (1996).

We construct default priors directly by examining how parameter change determines change in the model near an observed data point. The corresponding volume change as a function of the parameter reflects the sensitivity of the parameter at the data point and is the link to replacing sample space integration by parameter space integration. This is developed in Section 2, leading to the default prior given below by (7) or (9). In Section 3 we consider examples of exact and approximate priors using this construction. As part of this we show that the default prior needs in general to be targetted on the parameter of interest, when there is a type of nonlinearity in that parameter; this is an aspect of the marginalization paradoxes of Dawid, Stone and Zidek (1973). In Section 4 we use third order approximations for p-values and posterior probabilities to derive a suitably targetted prior defined on the profile curve of the parameter of interest, and we then extend this to the full parameter space, leading to a full default targetted prior, given below by (25).

The information based approach however seems to be limited to the case of

scalar interest and scalar nuisance parameters, and in order to extend this to vector sub-parameters we return to the approach of Section 2. This is described in Section 5, and Section 6 records a brief discussion.

Our goal throughout is to examine the structure of priors for which stated levels for posterior inference are realized, at least approximately. Our development is not rigorous, but we require the model to be smoothly differentiable in both y and θ, and assume the log-likelihood function can be expanded in Taylor series to at least third order, in the usual manner of asymptotic expansions. Our method of construction of default priors entails dependence on the data. Data-dependent priors have been discussed in the literature in various contexts, such as Box & Cox (1964) and Wasserman (2000). Pierce & Peters (1994) noted that to obtain agreement of Bayesian and frequentist higher order approximations data-dependent priors would be required in general. This work responded to a question raised by in the discussion of Pierce & Peters (1992). Clarke (2007) discusses the role of data-dependent priors in the context of priors constructed by information processing arguments.

2 Default priors from model properties

Suppose we have a single observation on a scalar parameter θ from a model with density $f(y;\theta)$ and distribution function $F(y;\theta)$, and that F is continuously differentiable in both y and θ. For an observed value y^0, the p-value as a function of θ is $F(y^0;\theta)$. The posterior survivor function for θ is $s(\theta) = \int_\theta f(y^0;\vartheta)\pi(\vartheta)d\vartheta$. If these two inference functions are to be equal for all θ, giving equivalence of posterior and frequentist inference, then

$$F(y^0;\theta) = \int_\theta f(y^0;\vartheta)\pi(\vartheta)d\vartheta.$$

Differentiation of both sides with respect to θ gives

$$\frac{\partial}{\partial\theta}F(y^0;\theta) = -f(y^0;\theta)\pi(\theta)$$

and thus determines the prior as

$$\pi(\theta) = \left| \frac{F_\theta(y^0;\theta)}{F_y(y^0;\theta)} \right|, \tag{1}$$

where the subscript notation indicates differentiation with respect to the relevant argument. The derivation of this default prior shows that the parameter space integration provides a duplicate of the sample space integration; in other words the posterior survivor function $s(\theta)$ is exactly equal to the frequentist p-value function, which records the percentile position of the data with respect to possible θ values.

In the special case of a location model, $F(y;\theta) = F(y-\theta)$, (1) gives a constant prior for θ; otherwise (1) it gives the precise generalization, in terms of a re-expression of θ. The prior can also be interpreted as $\pi(\theta) \propto |dy/d\theta|$, where the derivative is computed with $F(y;\theta)$ held fixed.

Another way of describing this is to note that in a location model the quantile at any observed value y^0 shifts to $y^0 + d\theta$ when θ changes to $\theta + d\theta$, i.e. $F(y^0;\theta) = F(y^0 + d\theta; \theta + d\theta)$. For a non-location model this generalizes by requiring the total differential of $F(y;\theta)$ be equal to zero at y^0:

$$dF(y;\theta)|_{y=y^0} = \frac{\partial F(y^0;\theta)}{\partial \theta} d\theta + \frac{\partial F(y^0;\theta)}{\partial y} dy = 0;$$

thus the effect at y^0 of parameter change at θ is

$$\left. \frac{dy}{d\theta} \right|_{y^0} = -\frac{F_\theta(y^0;\theta)}{F_y(y^0;\theta)}. \tag{2}$$

The same calculation applies when θ is a vector of dimension p:

$$\left. \frac{dy}{d\theta'} \right|_{y^0} = -\frac{F_{\theta'}(y^0;\theta)}{F_y(y^0;\theta)} \tag{3}$$

which is a $1 \times p$ row vector. This generalizes translation invariance to local translation invariance (Fraser, 1964). Equation (2) can also be written in terms of the quantile function by setting $u = F(y;\theta)$, solving for the u-quantile $y = y(u;\theta)$ and then differentiating directly:

$$\left. \frac{dy}{d\theta} \right|_{y^0} = y_\theta(u,\theta)|_{u=F(y^0;\theta)}\,. \tag{4}$$

In (2), (3) and (4) differentiation with respect to θ is calculated with the p-value $F(y^0;\theta)$ held fixed. Any pivotal quantity that is a one-to-one function of $F(y^0;\theta)$ gives the same definition of $dy/d\theta$.

Now for a sample of independent observations, $y = (y_1, \ldots, y_n)$, each y_i has a corresponding row vector $V_i(\theta)$ defined by (3) using its distribution function. The change then in the vector variable y at y^0 under differential change in θ is now

$$\left.\frac{dy}{d\theta'}\right|_{y^0} = \left\{ \begin{array}{c} V_1(\theta) \\ \vdots \\ V_n(\theta) \end{array} \right\} = V(\theta), \tag{5}$$

where $V(\theta)$ is an $n \times p$ matrix that we call the sensitivity of θ at y^0. We denote the columns of $V(\theta)$ by $\{v_1(\theta) \cdots v_p(\theta)\}$ where the $n \times 1$ vector $v_j(\theta)$ gives the data displacement when the jth coordinate of θ is changed by $d\theta_j$.

This sensitivity matrix $V(\theta)$ forms the basis for the construction of a default prior. If we are in a simple location model with scalar y and scalar θ then $V(\theta) = 1$, and we have $F(y^0 - \theta) = \int^{y^0} f(y - \theta)dy = \int_\theta f(y^0 - \alpha)d\alpha$, and the sample space integration is replaced by parameter space integration. Thus with a flat prior for θ posterior probabilities are equal to observed p-values. Indeed Bayes (1763) used a translation or invariance argument to recommend the flat prior $\pi(\theta) = c$ for the parameter θ; in effect proposing a confidence argument well before Fisher.

In non-location models the sensitivity matrix $V(\theta)$ enables integration with respect to y, which gives p-values, to be converted to integration with respect to θ, which gives posterior probabilities. Thus a natural default prior is the volume element determined by $V(\theta) : \pi(\theta)d\theta \propto |V(\theta)|d\theta = |V(\theta)'V(\theta)|^{1/2}d\theta$. To link this default prior to the development of priors from information matrices derived in the next sections, we make some refinements, using the coordinates given by the maximum likelihood estimator.

We write $\ell(\theta; y) = \log f(y; \theta)$ for the log-likelihood function, and $\hat{\theta} = \hat{\theta}(y)$ for the maximum likelihood statistic obtained by solving the score equation $\ell_\theta(\theta; y) = 0$. The connection between y and $\hat{\theta}(y)$ is obtained by evaluating the total derivative

of the score equation $\ell_\theta(\theta; y) = 0$; at $(\hat{\theta}^0; y^0)$ we have

$$\ell_{\theta\theta'}(\hat{\theta}^0; y^0)d\hat{\theta} + \ell_{\theta;y'}(\hat{\theta}^0; y^0)dy = 0,$$

where the differentials $d\hat{\theta}$ and dy are respectively $p \times 1$ and $n \times 1$ vectors, and $\hat{\theta}^0 = \hat{\theta}(y^0)$ is the maximum likelihood estimate with data y^0. Solving for $d\hat{\theta}$ gives

$$d\hat{\theta} = \hat{j}^{-1}H'dy$$

where $H' = \ell_{\theta;y'}(\hat{\theta^0}; y^0)$ is the gradient of the score function at the data point and $\hat{j} = j(\hat{\theta}^0; y^0) = -\partial^2\ell(\hat{\theta}^0; y^0)/\partial\theta\partial\theta'$ is the observed Fisher information. Combining this with (5) we obtain

$$d\hat{\theta} = \hat{j}^{-1}H'V(\theta)d\theta = W(\theta)d\theta \tag{6}$$

which presents the sample space change $d\hat{\theta}$ at $\hat{\theta^0}$ in terms of parameter space change $d\theta$ at arbitrary θ. In particular for any given volume increment at the data y^0 we have determined the direct equivalent volume increment at any parameter value θ of interest; this gives the default prior as the parameter space support volume that corresponds to a fixed data volume increment at y^0:

$$\pi(\theta)d\theta = |\hat{j}^{-1}HV(\theta)|d\theta = |W(\theta)|d\theta. \tag{7}$$

For calculations with component parameters, described in Section 5, there are advantages to standardizing with respect to observed information. For this let $\hat{j}^{1/2}$ be a right square root of the observed information matrix $\hat{j}$ and consider the standardized vector differential

$$\hat{j}^{1/2}d\hat{\theta} = \hat{j}^{1/2}W(\theta)d\theta = \tilde{W}(\theta)d\theta. \tag{8}$$

The rescaled default prior is then

$$\pi(\theta)d\theta = |\tilde{W}(\theta)|d\theta. \tag{9}$$

These priors can lead to posterior survivor values that duplicate to second and third order the frequentist p-values available from asymptotic theory, although care must be taken when constructing marginal posteriors, as the marginalization paradox of Dawid et al. (1973) is a limiting factor. These issues are taken up in Section 4.

3 Examples of default priors

The first three examples are similar to Bayes' original location model, and lead to posterior quantiles that agree with frequentist inference. In the normal linear model we recover Jeffreys' modified rule, the right invariant prior.

Example 3.1: Normal theory linear regression. Suppose y_i follows a normal distribution with mean $X_i\beta$ and variance σ^2, where X_i is the ith row of an $n \times p$ design matrix X, β is $r \times 1$, and $\theta' = (\beta', \sigma^2)$. Inverting $u_i = F(y_i;\theta) = \Phi\{(y_i - X_i\beta)/\sigma\} = \Phi(z_i)$, where $\Phi(\cdot)$ is the distribution function for the standard normal, gives the quantile functions $y_1 = X_1\beta + \sigma z_1, \ldots, y_n = X_n\beta + \sigma z_n$. We compute $V(\theta)$ for fixed u, as described at (4), or equivalently for fixed z, obtaining

$$V(\theta) = \left.\frac{dy}{d\theta}\right|_{y^0} = \{X, z^0(\theta)/2\sigma\},$$

where $z^0(\theta) = z(y^0, \theta) = (y^0 - X\beta)/\sigma$ is the standardized residual corresponding to data y^0 and parameter value θ. The likelihood gradient $\ell_{;y} = (X\beta - y)/\sigma^2$ then gives the score gradient $\ell_{\theta';y}(\theta; y) = \{\sigma^{-2}X, \sigma^{-4}(y - X\beta)\}$ and

$$H = \{X/\hat\sigma^2, (y^0 - X\hat\beta^0)/\hat\sigma^4\} = (X/\hat\sigma^2, \hat z^0/\hat\sigma^3),$$

where $\hat\sigma^0$ is abbreviated as $\hat\sigma$ to simplify notation. The observed information is $\hat j = \mathrm{diag}\{X'X/\hat\sigma^2, n/(2\hat\sigma^4)\}$; combining these using (7) and least squares projection properties gives

$$\begin{aligned} W(\theta) &= \begin{Bmatrix} \hat\sigma^2(X'X)^{-1} & 0 \\ 0 & 2\hat\sigma^4/n \end{Bmatrix} \begin{pmatrix} X'/\hat\sigma^2 \\ \hat z^{0\prime}/\hat\sigma^3 \end{pmatrix} \{\ X \quad z^0(\theta)/(2\sigma)\ \} \\ &= \begin{Bmatrix} I & (X'X)^{-1}X'z^0(\theta)/(2\sigma) \\ 2\hat z^{0\prime}\hat\sigma X/n & \hat z^{0\prime}z^0(\theta)\hat\sigma/(n\sigma) \end{Bmatrix} \\ &= \begin{Bmatrix} I & (\hat\beta^0 - \beta)/2\sigma^2 \\ 0 & \hat\sigma^2/\sigma^2 \end{Bmatrix}. \end{aligned}$$

leading to

$$\begin{aligned} d\hat\beta &= d\beta + (\hat\beta^0 - \beta)d\sigma^2/2\sigma^2 \\ d\hat\sigma^2 &= \hat\sigma^2 d\sigma^2/\sigma^2. \end{aligned}$$

It thus follows from (7) that the default prior is

$$\pi(\theta)d\theta \propto |W(\theta)|d\theta \propto d\beta d\sigma^2/\sigma^2, \tag{10}$$

which is the familiar right invariant prior. This example illustrates how (7) modifies the invariance argument of Jeffreys to adapt to the underlying location parameter form of the distribution function for each coordinate. Jeffreys' prior is the square root of the determinant of the expected Fisher information matrix, which leads to the left invariant prior $d\beta d\sigma^2/\sigma^4$. This is usually regarded as incorrect for this problem; for example the associated posterior does not reproduce the t-distribution with the usual degrees of freedom for inference about components of β, whereas the right invariant prior does, and agrees wutg Jeffreys' (1961) proposal for a modified rule for location-scale settings (Kass & Wasserman, 1996).

Example 3.2: Normal circle. As a special case of the normal theory linear model let (y_1, y_2) be distributed as $N\{(\mu_1, \mu_2); I/n\}$. It follows either from the preceding example, or from direct calculation, that the default prior for the location parameter μ is $cd\mu_1 d\mu_2$, which gives the $N\{(y_1^0, y_2^0); I/n\}$ posterior for (μ_1, μ_2). For any component parameter linear in (μ_1, μ_2) we then have exact agreement between frequentist p-values and Bayesian survivor probabilities.

Suppose now that we reparameterize the model as $\theta = (\psi, \alpha)$ where $\mu_1 = \psi\cos\alpha$ and $\mu_2 = \psi\sin\alpha$. The quantile functions are $y_1 = \psi\cos\alpha + z_1$ and $y_2 = \psi\sin\alpha + z_2$, where z_1, z_2 are independent standard normal variables. This gives

$$W(\theta) = \left\{ \begin{array}{cc} \cos(\hat{\alpha} - \alpha) & \psi\sin(\hat{\alpha} - \alpha) \\ -\hat{\psi}^{-1}\sin(\hat{\alpha} - \alpha) & \psi\hat{\psi}^{-1}\cos(\hat{\alpha} - \alpha) \end{array} \right\}, \tag{11}$$

and then from (7) or (9) we obtain the default prior $\pi(\theta)d\theta \propto \psi d\psi d\alpha$ for the full parameter. This is equivalent to the default flat prior $d\mu_1 d\mu_2$ calculated directly from the location parameter (μ_1, μ_2).

However, this prior is not appropriate for marginal inference when the parameter of interest is the radial distance ψ, which is a nonlinear function of the mean vector μ. To see this note that the marginal distribution of $y_1^2 + y_2^2$ depends only

on ψ, and the p-value function from this marginal distribution is

$$p(\psi) = \Pr\{\chi^2_2(\psi^2) \leq n(y_1^2 + y_2^2)\},$$

where $\chi^2_2(\delta^2)$ is a noncentral chi-square with 2 degrees of freedom and noncentrality parameter δ^2 and the y's are fixed at their observed values. In contrast the posterior survivor function for ψ under the flat prior $d\mu_1 d\mu_2$ is

$$s(\psi) = \Pr[\psi^2 \leq \chi^2_2\{n(y_1^2 + y_2^2)\}].$$

Numerical calculation confirms there can be substantial under-coverage for right tail intervals based on the marginal posterior. In the extension to k dimensions, with y_i distributed as $N(\mu_i, 1/n), i = 1, \ldots, k$, it can be shown that

$$s(\psi) - p(\psi) = \frac{k-1}{\psi\sqrt{n}} + O(\frac{1}{n})$$

so the discrepancy increases linearly with the number of dimensions. The scaling of the variances by $1/n$ enables this asymptotic analysis: we could equivalently model independent observations $y_{ij}, j = 1, \ldots n$ from normal distributions with mean μ_i and variance 1.

This discrepancy does not appear in the first order of asymptotic theory: the limiting distribution of $\sqrt{n}\hat{\psi} = \sqrt{n}(y_1^2 + y_2^2)^{1/2}$ is $N(\psi, 1)$, and the limiting distribution of the exact posterior for ψ is $N(\hat{\psi}, 1/n)$, so to this order of approximation p-values and marginal survivor probabilities are identical. This is simply reflecting the fact that any prior not depending on n is in the limit swamped by the data and has no effect on the posterior inference. To study the agreement between Bayesian and frequentist inference it is necessary to consider either the exact distributions, or higher order approximations.

The inappropriateness of the point estimator developed from the prior $\pi(\mu)d\mu \propto d\mu$ was pointed out in Stein (1959) and is discussed in detail in Cox & Hinkley (1974, p. 46 and p. 383).

This example illustrates in simple form the difficulty with the default prior (7) and any 'flat' prior for a vector parameter. It is not possible to achieve approximate equality of Bayesian and frequentist inferences beyond the simple asymptotic

normal limit when the parameter of interest is curved in the local location parameter. This is a version of the marginalization paradox of Dawid et al. (1973): they described settings where assigning a prior to a full parameter and then marginalizing necessarily conflicts with the approach of reducing to a one-parameter model and marginalizing the prior to that parameter of interest. The need to target the prior on the particular parameter component of interest is well-recognized in the literature on the construction of reference priors, but seems less well appreciated in other contexts. In Section 4 we give a method to adapt the default prior (9) to target a particular parameter of interest.

Example 3.3: Transformation models. The preceding examples and many more are special cases of transformation models. In the Appendix we briefly record the links to this general type of model and the result that our locally defined prior (7) reproduces the right invariant prior for that model type, thus (7) can be written

$$\pi(\theta)d\theta \propto |W(\theta)|d\theta = cd\nu(\theta)$$

where $d\nu(\theta)$ is the right invariant measure on the transformation group. Transformation model theory shows that this prior is fully accurate fro reproducing frequentist p-values, provided that the parameter of interest is linear in the transformation parameter, thus avoiding the marginalization issues of Dawid et al. (1973).

In the next three examples the default prior is based on the approximate location relationship described by the sensitivity matrix $V(\theta)$.

Example 3.4: The Welch-Peers approximation. As noted above, the construction of the default prior using the sensitivity matrix $V(\theta)$, or the modification to $W(\theta)$, gives a flat prior when θ is a location parameter. If we have a scalar parameter model in which the location parameter is $\beta(\theta)$, then this construction gives the flat prior $\pi(\theta)d\theta \propto d\beta(\theta)$. For any scalar parameter model an approximate location parameter is proportional to $i^{1/2}(\theta)$, where $i(\theta)$ is the expected Fisher information $E_\theta\{-\ell''(\theta)\}$. This was established in Welch & Peers (1963), by showing that this choice led to the equality of confidence and posterior bounds. It can also

be expressed as the result that

$$z = \int^{\hat{\theta}} i^{1/2}(t)dt - \int^{\theta} i^{1/2}(t)dt$$

has a limiting standard normal distribution, and to second order has a distribution free of θ. In quantile form this is

$$\hat{\beta} = \beta + z,$$

where $\beta(\theta) = \int^{\theta} i^{1/2}(t)dt$ is the constant information reparametrization and z is a fixed quantile of the θ-free distribution. Then $d\hat{\beta} = d\beta$ for fixed quantile, giving Jeffreys' prior $d\beta \propto i^{1/2}(\theta)d\theta$. The interpretation of this prior in terms of an approximate location parameter is discussed in Kass (1990).

This example links the location-parameter approach for constructing priors to that based on Fisher information. In Section 4 we extend this linking to develop targetted priors from exponential family approximations.

Example 3.5. Nonlinear regression. Suppose y_i are independently normally distributed with mean $x_i(\beta)$ and variance σ^2 for $i = 1, \ldots, n$, with $x_i(\beta)$ a known nonlinear function of the $p \times 1$ vector β. As in Example 3.1, the quantile functions are $y_i = x_i(\beta) + \sigma z_i$ where $z_i = \Phi^{-1}(u_i)$, and with $\theta' = (\beta', \sigma^2)$, the sensitivity matrix $V(\theta)$ obtained by differentiating the quantile functions for fixed z is

$$V(\theta) = \begin{bmatrix} X_1(\beta) & \{y_1^0 - x_1(\beta)\}/2\sigma^2 \\ \vdots & \vdots \\ X_n(\beta) & \{y_n^0 - x_n(\beta)\}/2\sigma^2 \end{bmatrix} = \begin{bmatrix} X(\beta) & z^0(\theta)/2\sigma \end{bmatrix},$$

where $X_i(\beta) = \partial x_i(\beta)/\partial\beta'$. We also have

$$\begin{aligned} H' &= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} X(\hat{\beta}^0) & \hat{z}^0/\hat{\sigma} \end{bmatrix}, \\ \hat{j} &= \begin{Bmatrix} \hat{j}_{11}/\hat{\sigma}^2 & 0 \\ 0 & n/2\hat{\sigma}^4 \end{Bmatrix} \end{aligned}$$

where $\hat{j}_{11} = \sum_{i=1}^n \dot{x}_i'(\hat{\beta})\dot{x}_i(\hat{\beta}) - \sum_{i=1}^n \{y_i - x_i(\hat{\beta})\}\ddot{x}_i(\hat{\beta})$ and again for notational convenience we write $\hat{\sigma}^2 = (\hat{\sigma}^0)^2 = \{y - x(\hat{\beta}^0)\}'\{y - x(\hat{\beta}^0)\}/n$.

We then obtain

$$W(\theta) = \begin{bmatrix} \hat{\sigma}^2 \hat{j}_{11}^{-1} & 0 \\ 0 & 2\hat{\sigma}^4/n \end{bmatrix} \begin{Bmatrix} X(\hat{\beta}^0)'/\hat{\sigma}^2 \\ \hat{z}^{0\prime}/\hat{\sigma}^3 \end{Bmatrix} \begin{Bmatrix} X(\beta) & z^0(\theta)/2\sigma \end{Bmatrix}$$

$$= \begin{Bmatrix} \hat{j}_{11}^{-1} \sum X_i(\hat{\beta})' X_i(\beta) & \hat{j}_{11}^{-1} \sum X_i(\hat{\beta})' z_i(\theta)/(2\sigma) \\ 2\hat{\sigma} \sum \hat{z}_i X_i(\beta)/n & (\hat{\sigma}/\sigma) \sum \hat{z}_i z_i(\theta)/n \end{Bmatrix}$$

where $\hat{z}_i = \{y_i - x_i(\hat{\beta})\}/\hat{\sigma}$ and $z_i(\theta) = \{y_i - x_i(\beta)\}/\sigma$. The determinant of $W(\theta)$ has the form $h(\beta)/\sigma^2$, where $h(\beta)$ is a nonlinear function of β determined by the derivatives $X(\beta)$ of the mean function. Using the approximation

$$x(\beta) = x(\hat{\beta}^0) + X(\hat{\beta}^0)'(\beta - \hat{\beta}^0) + wn^{-1/2}$$

where w is orthogonal to $\mathcal{L}\{X(\hat{\beta}^0)\}$, the default prior becomes $d\tilde{\beta}d\sigma^2/\sigma^2$ to $O(n^{-1})$, where $d\tilde{\beta}$ designates a flat prior in coordinates of the tangent plane projection at the fitted data point. The two-group reference prior for this example is proportional to $|X(\beta)'X(\beta)|^{1/2}/\sigma$ (Yang & Berger,1996), which was also proposed on the grounds of invariance by Eaves (1983).

Example 3.6. Gamma distribution. As an example of a one-parameter model which is neither location nor scale, we consider default priors for the shape parameter of a gamma distribution:

$$f(y;\theta) = \frac{1}{\Gamma(\theta)} y^{\theta-1} e^{-y}.$$

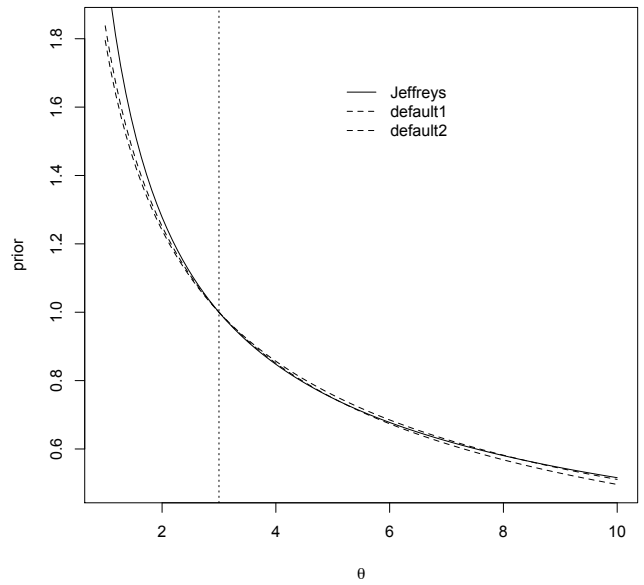
Jeffreys' prior is $\pi_J(\theta) \propto \psi''(\theta)^{1/2}$, where $\psi(\theta) = \log \Gamma(\theta)$. To construct a location-based prior for a sample of n, we use (2) with (7) to determine $V_i(\theta)$,

$$V_i(\theta) = \frac{\Gamma'(\theta)F(y_i^0;\theta) - \int_0^{y_i^0} z^{\theta-1} \log(z) e^{-z} dz}{e^{-y_i^0}(y_i^0)^{\theta-1}},$$

and then have $\pi_V(\theta) \propto \{\sum V_i^2(\theta)\}^{1/2}$.

Figure 1 shows Jeffreys' prior $i^{1/2}(\theta)$ and $\pi_V(\theta)$ for two different samples of size 30 from the gamma distribution with shape parameter $\theta = 3$. The priors are normalized to equal 1 at $\theta = 3$. The priors agree in the neighbourhood of the observed maximum likelihood value, then have slightly different curvature as they respond differently to curvature in the model.

Figure 1: Priors for the shape parameter of a Gamma distribution. Jeffreys' prior $\pi_J(\theta)$ (solid line) is proportional to $\{\psi''(\theta)\}^{1/2}$ where ψ'' is the trigamma function; the default prior (dashed line) proposed here is based on (2) and (7), and presented here for two different samples y^0 of size 30 from a gamma distribution with true value $\theta = 3$. The priors are standardized to equal 1 at the true value of θ.



This example can be extended to the two-parameter gamma model, with shape θ and mean μ:

$$f(y;\mu,\theta) = \frac{1}{\Gamma(\theta)}\frac{y^{\theta-1}\theta^{\theta}}{\mu^{\theta}}\exp(-\theta y/\mu).$$

The sensitivity matrix $V(\mu,\theta)$ has two columns: the column corresponding to μ is simply $(y_1^0/\mu,\ldots,y_n^0/\mu)'$, while the column corresponding to θ has elements

$$\tilde{V}_i(\theta,\mu;y_i^0) = -\frac{F_{1\theta}(\theta y_i^0/\mu;\theta)}{\frac{1}{\mu}f_1(\theta y_i^0/\mu;\theta)}$$

where the notation F_1 and f_1 refer to the gamma densities for $\mu = \theta$ used above. The associated default prior is given by

$$\pi(\theta,\mu) \propto |V'(\mu,\theta)V(\mu,\theta)|^{1/2} = \frac{1}{\mu}\{\sum \tilde{V}_i^2 - (\sum y_i\tilde{V}_i)^2/\sum y_i^2\}^{1/2},$$

which is proportional to $(1/\mu)$ times a function of θ (and y^0) only.

4 Information based priors

The approach developed in the preceding sections gives a default prior for a vector parameter, but the resulting posterior is not appropriately targetted on component parameters unless the components are linear, in the sense discussed in the Appendix at (iii). To develop default priors that are targetted on parameters of interest, we use an approach motivated by higher order asymptotics and by the interpretation of the Welch-Peers prior as a location-model based default prior noted in Example 3.4. In that example, the Fisher information function defines locally a location parameter, and the resulting 'flat' prior is given by the Fisher information metric. To generalize this to the vector case, we can either generalize the location model approximation, which we did in the previous section, or the information approach, which we now consider. For targetting the prior on the parameter of interest, the information approach seems more directly accessible. In Section 5 we combine the two approaches to develop default priors for vector parameters of interest in the presence of nuisance parameters, although the resulting posterior is still subject to the marginalization paradox and may not give well-calibrated marginal posterior inference for curved parameters.

To examine the constraints needed to ensure that a marginal posterior is well-calibrated, we use higher order approximations for p-values and marginal posteriors available in the literature. We write $\theta = (\psi, \lambda)$, where ψ is a scalar parameter of interest and λ is a nuisance parameter, and let $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$ be the constrained maximum likelihood estimator, where $\hat{\lambda}_\psi$ is the solution (assumed unique) of $\partial \ell(\theta)/\partial \lambda = 0$.

The Laplace approximation to the marginal posterior survivor function for ψ is given by

$$s(\psi) = \Phi(r_B^*) = \Phi\{r + (1/r)\log(q_B/r)\}, \tag{12}$$

where

$$r = \text{sign}(\hat{\psi} - \psi)[2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]^{1/2}, \tag{13}$$

$$q_B = \ell'_{\text{p}}(\psi) j_{\text{p}}^{-1/2}(\hat{\psi}) \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)}, \tag{14}$$

$j_{\lambda\lambda}$ is the submatrix of the observed Fisher information matrix corresponding to the nuisance parameter, and $\ell_{\rm p}(\psi) = \ell(\hat\theta_\psi)$ is the profile log-likelihood function. This can be derived from Laplace approximation to the marginal posterior density for ψ: see, for example, Tierney & Kadane (1986), DiCiccio & Martin (1991) and Bédard et al (2007). The approximation has relative error $O(n^{-3/2})$ for ψ in $n^{-1/2}$-neighbourhoods of $\hat\psi$.

There is a parallel $O(n^{-3/2})$ p-value function for scalar $\psi(\theta)$, developed in Barndorff-Nielsen (1986) when there is an explicit ancillary function, and extended to general asymptotic models in Fraser & Reid (1993); see also Fraser et al. (1999) and Reid (2003). The analysis makes use of the observed likelihood function $\ell(\theta) = \ell(\theta; y^0)$ and the observed likelihood gradient $\varphi(\theta) = \ell_{;V}(\theta; y^0) = (\partial/\partial V)\ell(\theta; y)|_{y^0}$ in sample space directions V described below. Third order inference for any scalar parameter $\psi(\theta)$ is then available by replacing the model by an approximating exponential family:

$$g(s;\theta) = \exp\{\ell(\theta) + \varphi(\theta)'s\}h(s) \tag{15}$$

with observed data $s^0 = 0$, and using available approximation formulas for such (p,p) exponential families. This model is based on an expansion that ignores terms of second order, $O(n^{-1})$, but its construction ensures that p-values based on (15) have relative error of third order, $O(n^{-3/2})$. Some discussion of this use of the exponential family model $\{\ell(\theta), \varphi(\theta)\}$ as a full third order surrogate for the original model is given in Davison et al. (2006).

The p-value for testing $\psi(\theta) = \psi$ is

$$p(\psi) = \Phi(r_f^*) = \Phi\{r + (1/r)\log(q_f/r)\} \tag{16}$$

where r is as given above, and two equivalent expression for q_f are:

$$\begin{aligned} q_f &= \frac{|\varphi(\hat\theta) - \varphi(\hat\theta_\psi) \quad \varphi_\lambda(\hat\theta_\psi)|}{|\varphi_\theta(\hat\theta)|} \frac{|j(\hat\theta)|^{1/2}}{|j_{\lambda\lambda}(\hat\theta_\psi)|^{1/2}}, \\ &= \{\chi(\hat\theta) - \chi(\hat\theta_\psi)\} \left\{ \frac{|j_{\varphi\varphi}(\hat\theta)|}{|j_{(\lambda\lambda)}(\hat\theta_\psi)|} \right\}^{1/2}. \end{aligned} \tag{17}$$

The second version of q_f indicates that it can be presented as a parameter departure divided by its estimated standard error, and the first version gives a form

that is useful for computation.

In the expression for q_f, the canonical parameter $\varphi(\theta)$ is defined using sample space directional derivatives of the log-likelihood function:

$$\varphi(\theta) = \ell_{;V}(\theta; y^0) = \sum \ell_{y_i}(\theta; y^0) V_i(\hat{\theta}^0)$$

where $V_i(\theta)$ is the ith row of the sensitivity matrix (5). A derivation of the r_f^* approximation is beyond the scope of this paper, but is described in Reid (2003) and Fraser et al. (1999); see also Ch. 8 of Brazzale et al. (2007). The role of $V(\hat{\theta}^0)$ in the development of the approximation is to implement conditioning on an approximate ancillary statistic derived from a local location model, which is why the same matrix arises here as in the discussion of default priors.

Because the only difference between (12) and (16) is in the use of q_B or q_f, and only q_B involves the prior, we obtain equality of posterior and frequentist inference to $O(n^{-3/2})$ by deriving a prior so $q_B = q_f$. This was suggested in Casella et al. (1995), and was developed further in Fraser & Reid (2002), where it was called strong matching. Strictly speaking inference for ψ can be obtained using (16) alone, but the close parallel between (12) and (16) determines some aspects of the prior needed to ensure frequentist validity, at least to the present order of approximation. Thus we have

$$\frac{\pi(\hat{\theta}_\psi)}{\pi(\hat{\theta})} \propto \frac{\ell'_{\mathrm{p}}(\psi)|j_{\lambda\lambda}(\hat{\theta}_\psi)||\varphi_\theta(\hat{\theta})|}{|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi) \quad \varphi_\lambda(\hat{\theta}_\psi)||j(\hat{\theta})|}. \tag{18}$$

If the model is free of nuisance parameters, we obtain the strong matching prior described in Fraser & Reid (2002), which is given explicitly as

$$\pi(\theta)d\theta \propto \frac{\ell'(\theta; y^0)}{\varphi(\theta)} d\theta; \tag{19}$$

equivalently $\beta(\theta) = \int^\theta \{\ell'(\vartheta; y^0)\}/\varphi(\vartheta)d\vartheta$ is a local location parameter at the observed data point. By construction this prior leads to third order equality of posterior probabilities and p-values; accordingly we refer to it as a third order prior. The asymptotic equivalence of (19) and Jeffreys' prior is outlined in the Appendix at (iv).

Equation (18) gives a data-dependent prior along the profile curve $\mathcal{C}_\psi = \{\theta : \theta = (\psi, \hat{\lambda}_\psi)\}$ in the parameter space, but does not immediately give a prior over the full parameter space. To extend the prior for arbitrary values of λ beyond the profile curve we use the Welch & Peers (1963) construction based on Fisher information, as described below at (25).

Using the second expression for q_f in (17), we obtain

$$\pi(\hat{\theta}_\psi)d\psi d\lambda = c\frac{\ell'_{\rm p}(\psi)}{\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)}d\psi \cdot |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}d\lambda \tag{20}$$

on the profile curve C_ψ, ignoring terms of $O(n^{-3/2})$. The scalar parameter $\chi(\theta)$ is a rotated coordinate of the canonical parameter $\varphi(\theta)$ that is first derivative equivalent to $\psi(\theta) = \psi$ at $\hat{\theta}_\psi$; it is the unique locally defined scalar canonical parameter for assessing $\psi(\theta) = \psi$ (see for example, Fraser et al., 1999). The matrix $j_{(\lambda\lambda)}(\hat{\theta}_\psi)$ is the nuisance information matrix calculated with λ change re-expressed to the metric provided by $\varphi(\theta)$ for fixed ψ:

$$|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2} = |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}|\varphi_\lambda(\hat{\theta}_\psi)|^{-1},$$

where $\varphi_\lambda(\theta) = \partial\varphi(\theta)/\partial\lambda$.

Since (15) is a second order approximation in moderate deviations at the data point, we find it convenient to approximate the initial factor using the asymptotic relation between the score function and the maximum likelihood estimator, described in the Appendix at (iv),

$$\frac{\ell_\psi(\hat{\theta}_\psi)d\psi}{\hat{\chi} - \hat{\chi}_\psi} = |j_{(\psi\psi)\cdot\lambda}(\hat{\theta}_\psi)|^{1/2}d(\psi). \tag{21}$$

The notation $d(\psi) = d\chi_\psi$ denotes differential change in the parameter ψ, but expressed in terms of the φ metric.

The right factor in (20) can be rewritten giving

$$|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}d\lambda_\psi = |j_{(\lambda\lambda)}(\hat{\theta}_\psi)|d(\lambda_\psi)$$

where λ_ψ is the λ parametrization as used on the contour with ψ fixed and (λ_ψ) is that parametrization presented in the φ scaling. Thus $d(\lambda_\psi) = |\varphi_\lambda(\hat{\theta}_\psi)|d\lambda_\psi$ at the point where the ψ contour intersects $\mathcal{C}_\psi$.

Combining these modifications gives the following default prior as an adjusted Jeffreys' prior along the profile contour $\mathcal{C}_\psi$:

$$\begin{aligned} \pi(\hat\theta_\psi)d\psi d\lambda &= |j_{(\psi\psi)\cdot\lambda}(\hat\theta_\psi)|^{1/2}|j_{(\lambda\lambda)}(\hat\theta_\psi)|d(\psi)d(\lambda_\psi) \qquad (22)\\ &= |j_{\varphi\varphi}(\hat\theta_\psi)|^{1/2}|j_{(\lambda\lambda)}(\hat\theta_\psi)|^{1/2}d(\psi)d(\lambda_\psi). \qquad (23)\end{aligned}$$

To extend this off the contour $\mathcal{C}_\psi$ the calculations for nuisance information for various λ given ψ need to be carefully ordered: they are made within the exponential model defined by $\ell(\theta)$ and $\varphi(\theta)$, and are thus with respect to the φ-rescaled version of λ designated as (λ) and are then rescaled to other parametrizations as needed. The notation $d(\psi)d(\lambda_\psi)$ is used to emphasize this: some further explanation is given in the Appendix at (ii). We extend (23) to the full parameter space by replacing the information metric $|j_{(\lambda\lambda)}(\hat\theta_\psi)|^{1/2}d(\lambda_\psi)$ at the intersection with $\mathcal{C}_\psi$ by the extension off the profile contour that accomplishes Laplace integration performed at the profile contour $\mathcal{C}_\psi$; this gives the following general expression for the default prior,

$$\begin{aligned} \pi_\psi(\theta)d\psi d\lambda &= |j_{(\psi\psi)\cdot\lambda}(\hat\theta_\psi)|^{1/2}|j_{(\lambda\lambda)}(\hat\theta_\psi)|^{1/2}|j_{(\lambda\lambda)}(\theta)|^{1/2}d(\psi)d(\lambda_\psi) \qquad (24)\\ &= |j_{\varphi\varphi}(\hat\theta_\psi)|^{1/2}|j_{(\lambda\lambda)}(\theta)|^{1/2}d(\psi)d(\lambda_\psi), \qquad (25)\end{aligned}$$

which is Jeffreys' prior with a supplemental factor $|j_{(\lambda\lambda)}(\theta)|^{1/2}$ that accomplishes the targetting on the component ψ. The use of approximation (21) reduces the prior to second order accuracy.

Example 4.1: Normal (μ, σ^2). In Example 3.1 we obtained the default prior for the full parameter $\theta = (\mu, \sigma^2)$; we now illustrate the preceding targetted default prior for the components μ and σ^2. The canonical parameter is $(\varphi_1, \varphi_2) = (\mu/\sigma^2, 1/\sigma^2)$ which has information function

$$\begin{aligned} j_{\varphi\varphi}(\theta) &= n\begin{pmatrix} \varphi_2^{-1} & -\varphi_1\varphi_2^{-2} \\ -\varphi_1\varphi_2^{-2} & (1/2)\varphi_2^{-2} + \varphi_1^2/\varphi_2^3 \end{pmatrix} \\ &= n\begin{pmatrix} \sigma^2 & -\mu\sigma^2 \\ -\mu\sigma^2 & \sigma^4/2 + \mu^2\sigma^2 \end{pmatrix},\end{aligned}$$

and thus Jeffreys prior $(n\sigma^3/\sqrt{2})d\varphi_1 d\varphi_2 = (n/\sqrt{2}\sigma^3)d\mu d\sigma^2$. Without loss of generality we take the data point to be $(\hat{\mu}, \hat{\sigma}^2) = (0, 1)$. The re-standardized canonical parameter $(\tilde{\varphi}_1, \tilde{\varphi}_2)$ is $(n^{1/2}\mu/\sigma^2, n^{1/2}/\sqrt{2}\sigma^2)$ and has $j^0_{\hat{\varphi}\hat{\varphi}} = I$ with information function

$$j_{\tilde{\varphi}\tilde{\varphi}} = \begin{pmatrix} \sigma^2 & -\sqrt{2}\mu\sigma^2 \\ -\sqrt{2}\mu\sigma^2 & \sigma^4 + 2\mu^2\sigma^2 \end{pmatrix}$$

and Jeffreys prior

$$\sigma^3 d\tilde{\varphi}_1 d\tilde{\varphi}_2 = (n/\sqrt{2}\sigma^3)d\mu d\sigma^2.$$

With μ as interest parameter using the particular data choice we have $\mathcal{C}_\mu = \{(\mu, \hat{\sigma}_\mu)\} = \{(\mu, 1)\}$ and in moderate deviations $O(n^{-1})$ have $\hat{\sigma}^2_\mu = \hat{\sigma}^2 + \mu^2 = 1 + \delta^2/n = 1$ where $\mu = \delta/\sqrt{n}$ relative to $\hat{\mu} = 0$. For the nuisance information we have $j_{\sigma^2\sigma^2} = n\sigma^4/2$ and the recalibrated information using the $\tilde{\varphi}$ scaling is

$$\begin{aligned} j_{(\sigma^2\sigma^2)}(\hat{\theta}_\mu) &= \frac{n}{2\hat{\sigma}^4_\mu} \left| \frac{\partial\sigma^2}{\partial(\sigma)^2} \right|^{-2}_{(\mu,\hat{\sigma}^2_\mu)} \\ &= \frac{1}{2}(\mu^2 + \frac{1}{2})^{-1}\hat{\sigma}^4_\mu = 1. \end{aligned}$$

Thus on $\mathcal{C}_\mu$ with $\sigma^2 = \hat{\sigma}^2_\mu = 1$ we have the Jeffreys $|j_{\varphi\varphi}(\hat{\theta}_\mu)|^{1/2} d(\mu) d(\sigma^2) = c d\mu d\sigma^2$ and then the adjusted Jeffreys

$$|j_{\varphi\varphi}(\hat{\theta}_\mu)|^{1/2} |j_{(\sigma^2\sigma^2)}(\hat{\theta}_\mu)|^{1/2} d(\mu) d(\sigma^2) = c d\mu d\sigma^2.$$

Extending this off $\mathcal{C}_\mu$ by the constant information metric for σ^2 gives

$$\pi_\mu(\theta) d\mu d\sigma^2 = \frac{c}{\sigma^2} d\mu d\sigma^2;$$

this is the familiar right invariant measure.

With σ^2 as parameter of interest we have the profile $\mathcal{C}_{\sigma^2} = \{(\hat{\mu}_{\sigma^2}, \sigma^2)\} = \{(0, \sigma^2)\}$. For the nuisance information we have

$$j_{\mu\mu} = \frac{n}{\sigma^2}, \quad j_{(\mu\mu)} = \frac{n}{\sigma^2} \left(\frac{\partial\hat{\varphi}}{\partial\mu} \right)^{-2}_{(\hat{\mu}_{\sigma^2}, \sigma^2)} = \sigma^2,$$

using $\hat{\mu}_{\sigma^2} = 0$ on $\mathcal{C}_{\sigma^2}$; this gives the Jeffreys

$$|j_{\varphi\varphi}(\hat{\theta})|^{1/2} d(\mu) d(\sigma^2) = \sigma^{-3} d\mu d\sigma^2$$

and then the adjusted Jeffreys

$$|j_{\varphi\varphi}(\hat{\theta}_{\sigma^2})|^{1/2}|j_{(\mu\mu)}(\hat{\theta}_{\sigma^2})|^{1/2}d(\mu)d(\sigma^2) = \sigma^{-3}\sigma d\mu d\sigma^2.$$

Extending this using the constant information metric for μ gives the same expression, which again is the right invariant prior.

Example 4.2: Normal circle (continued). We saw in Example 3.2 that the default prior for the vector $\varphi = (\psi\cos\alpha, \psi\sin\alpha)$ did not correctly target the component parameter ψ. The following components of the targetted prior (23) are

$$\begin{aligned} \hat{\chi} - \hat{\chi}_\psi &= r - \psi, \quad \ell_\psi(\hat{\theta}_\psi) = r - \psi, \quad d\chi = d(\psi), \\ j_{\alpha\alpha}(\hat{\theta}_\psi) &= r\psi, \quad j_{(\alpha\alpha)}(\hat{\theta}_\psi) = r/\psi = j_{(\alpha\alpha)}(\theta). \end{aligned}$$

We then obtain the prior

$$\pi_\psi(\theta)d\psi d\lambda = \frac{r-\psi}{r-\psi}(r\psi)^{1/2}(\frac{r}{\psi})^{1/2}d\psi d\alpha = cd\psi d\alpha,$$

which is uniform in the radius ψ and the angle α. This agrees with several derivations of default priors, including Fraser & Reid (2002), who obtained default priors on the constrained maximum likelihood surface, and with Datta & Ghosh (1995) who obtained this as a reference prior, while noting that it was in the family of matching priors derived in Tibshirani (1989).

As another way of explaining (25), suppose that the full likelihood is first integrated with respect to the Jeffreys prior for the nuisance parameter,

$$|j_{(\lambda\lambda)}(\psi, \lambda_\psi)|^{1/2}d(\lambda_\psi) = |j_{[\lambda\lambda]}(\psi, \lambda_\psi)|^{1/2}d\lambda_\psi,$$

where the exponential parameter change $d(\lambda)$ is recalibrated to the change $d\lambda$ and the subscript is to indicate that this is done for fixed ψ. This integration on the parameter space has a Welch & Peers (1963) inversion to the sample space that uses the corresponding score variable s_2 at y^0 with differential

$$|j_{(\lambda\lambda)}(\psi, \hat{\lambda}_\psi)|^{-1/2}ds_2.$$

By contrast the ordinary sample space integration to obtain the marginal density relative to ψ uses just the score differential ds_2 for integration, which is $|j_{(\lambda\lambda)}(\psi,\hat{\lambda}_\psi)|^{1/2}$ times larger. Thus to directly duplicate the marginal density for ψ requires the rescaled Jeffreys

$$|j_{(\lambda\lambda)}(\psi,\hat{\lambda}_\psi)|^{1/2}|j_{[\lambda\lambda]}(\psi,\lambda)|^{1/2}d\lambda; \tag{26}$$

the additional factor is in fact the marginal likelihood adjustment to the ψ profile as developed differently in Fraser (2003).

The rescaled Jeffreys integration for λ on the parameter space produces marginal probability concerning ψ with support ds_1. For different ψ values the support can be on different lines through y^0, which is the rotation complication that has affected the development of marginal likelihood adjustments (Fraser, 2003). The choice of the standardized $\tilde{\varphi}(\theta)$ gives a common information scaling on the different lines through y^0 that are used to assess ψ. This provides sample space invariance and leads to the third order adjustment for marginal likelihood.

The adjusted nuisance Jeffreys prior (26) produces marginal likelihood for ψ, which then appears as an appropriately adjusted profile likelihood for that parameter of interest. This can then be integrated following the Welch-Peers pattern using root profile information obtained from the exponential parametrization. This gives the Jeffreys type adjustment

$$|j^{(\psi\psi)}(\hat{\theta}_\psi)|^{-1/2}d(\psi) = |j^{[\psi\psi]}(\hat{\theta}_\psi)|^{-1/2}d\psi$$

for the profile concerning ψ. The combined targetted prior for ψ is then

$$\begin{aligned}
\pi_\psi(\theta) &= |j^{[\psi\psi]}(\hat{\theta}_\psi)|^{-1/2}|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}|j_{[\lambda\lambda]}(\theta)|^{1/2}d\psi d\lambda \\
&= |j_{\varphi\varphi}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\theta)|^{1/2}d(\psi)d(\lambda) \\
&= |j_{[\theta\theta]}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\theta)|^{1/2}d\psi d\lambda
\end{aligned}$$

for use with the full likelihood $L(\psi,\lambda)$, which is the same as (25).

5 Targetted default priors: vector components

The information approach outlined in the preceding section requires that the nuisance parameter be scalar, so that the Welch-Peers approach can be used to extend the default prior beyond the profile contour. In this section we use the continuity approach to extend the preceding information-based approach to the case where the parameter of interest $\psi(\theta)$ and nuisance parameter $\lambda(\theta)$ are vector valued, with dimensions say d and $p-d$ and with $\theta' = (\psi', \lambda')$.

The parameter effects matrix $\tilde{W}(\theta)$ at (8) can be partitioned in accord with the components ψ and λ giving $\tilde{W}(\theta) = \{\tilde{W}_\psi(\theta), W_\lambda(\theta)\}$ so that

$$\hat{j}^{1/2} d\hat{\theta} = \tilde{W}_\psi(\theta) d\psi + \tilde{W}_\lambda(\theta) d\lambda.$$

To target the parameter on ψ, we separate the effects of ψ and λ following the pattern used in Section 4. We construct the targetted prior as

$$\begin{aligned} \pi_\psi(\theta) d\theta &\propto |\tilde{W}(\hat{\theta}_\psi)| \frac{|\tilde{W}_\lambda(\theta)|}{|\tilde{W}_\lambda(\hat{\theta}_\psi)} \\ &= |\tilde{W}_{\psi\cdot\lambda}(\hat{\theta}_\psi)| \cdot |\tilde{W}_\lambda(\theta)| d\psi d\lambda; \end{aligned}$$

in comparison with (22) the rescaling mentioned at (26) is built into the $\tilde{W}$ matrices.

For a parameter value $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$ on the profile curve $\mathcal{C}_\psi$ formed by the constrained maximum likelihood values, a change $d\lambda$ in λ with ψ fixed generates a $p-d$-dimensional tangent plane $\mathcal{T}_\psi = \mathcal{L}\{\tilde{W}_\lambda(\hat{\theta}_\psi)\}$ at the observed $\hat{\theta}^0$. The term $\tilde{W}_{\psi\cdot\lambda} d\psi$ presents the effect of a change $d\psi$ in ψ with a corresponding change in λ so that the consequent change in $\{\hat{j}^{1/2} d\theta\}$ at the observed $\hat{\theta}^0$ is perpendicular to $\mathcal{T}_\psi$. Then

$$\hat{j}^{1/2} d\hat{\theta} = \tilde{W}_{\psi\cdot\lambda}(\theta) d\psi + \tilde{W}_\lambda(\theta) d\lambda;$$

with $d\psi$ and $d\lambda$ interpreted as just described. Parameter change in ψ is measured along the profile curve $\mathcal{C}_\psi$. When ψ however is a vector we would expect this prior for ψ to be third order only for linear parameters; otherwise we could have curvature effects as in the normal circle example.

Example 5.1: Linear regression. Suppose $r = 3$ and let $\psi = (\beta_1, \beta_2)'$ be the parameter of interest and $\lambda = (\beta_3, \sigma^2)$ be the nuisance parameter. Then we have

$$W_\psi(\theta) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad W_\lambda(\theta) = \begin{pmatrix} 0 & (\hat\beta_1^0 - \beta_1)/\sigma^2 \\ 0 & (\hat\beta_2^0 - \beta_2)/\sigma^2 \\ 1 & (\hat\beta_3^0 - \beta_3)/\sigma^2 \\ 0 & \hat\sigma^2/\sigma^2 \end{pmatrix}, \tag{27}$$

leading to

$$\begin{aligned} \pi_\psi(\theta)d\theta &= \frac{\hat\sigma^2}{\hat\sigma^2_\psi}\frac{|\tilde W_\lambda(\theta)|}{|\tilde W_\lambda(\hat\theta_\psi)|} \\ &= \frac{\hat\sigma^2}{\hat\sigma^2_\psi}\frac{\{(\hat\beta_1^0 - \beta_1)^2 + (\hat\beta_2^0 - \beta_2)^2 + \hat\sigma^4\}^{1/2}}{\hat\sigma^2/\hat\sigma^2_\psi} \\ &= \frac{1}{\sigma^2}\{(\hat\beta_1^0 - \beta_1)^2 + (\hat\beta_2^0 - \beta_2)^2 + \hat\sigma^4\}^{1/2}. \end{aligned}$$

As $\hat\beta_i^0 - \beta_i$ is of order $n^{-1/2}$ we have $(\hat\beta_i^0 - \beta_i)^2 = O(n^{-1})$ and thus $\pi_\psi(\theta)$ simplifies to c/σ^2 to second order. This gives the prior $d\beta d\sigma^2/\sigma^2$ as expected from Example 3.1.

6 Discussion

We have described two approaches to defining default priors: one based on extending a location approximation, and one based on matching higher order approximations. There is a natural progression in complexity, according to the model type. If we have a location model, then the uniform prior of Bayes, $\pi(\theta)d\theta \propto d\theta$ ensures exact matching of frequentist p-values and posterior probability limits. In a scalar parameter model that is not location, asymptotic arguments lead to Jeffreys' prior $\pi_J(\theta)d\theta \propto i^{1/2}(\theta)d\theta$, which gives matching probabilities to second order.

The location version (19) gives matching to third order, and by being data dependent automatically incorporates conditioning on approximate ancillary statistics. The default prior based on the sensitivity matrix, derived in Section 2, extends this local location property to vector parameters. Underlying the construction of (7) is an approximation to the model at the data point by a tangent location

model. This model can be explicitly derived in the scalar parameter case using Taylor series expansions, and the location parameter is given by the expression for β following (19). In the vector parameter setting, the existence of a location model approximation to the original model, to $O(n^{-1})$ can be established (Cakmak et al., 1994), but the form of the location parameter is typically not available explicitly. The array $V(\theta)$ based on pivotals for each coordinate y_i gives a transformation model approximation related to this location parameterization, and in that sense is an $O(n^{-1})$ default prior. A reviewer has suggested a simpler way to interpret the role of the sensitivity matrix $V(\theta)$ in the default prior (7): this prior gives more weight to parameter values that have more influence at the data point. Operationally (7) provides a rescaling on the parameter space so that in the new parametrization each parameter value has the same influence at the data point. However more is needed in the case of non-linear parameters, in order to properly target the parameter of interest, as discussed in Section 4.

We have not discussed the propriety of posteriors based on $V(\theta)$. The development is local to the data point, and several *ad hoc* approximations are made that assume θ is within $O(n^{-1/2})$ of the maximum likelihood estimate $\hat{\theta}$. This suggests that posteriors would need to be checked on a case by case basis. It is possible however that the data dependence is an advantage in this regard. As an example, consider the three parameter Weibull distribution with density function

$$f(y;\theta) = \frac{\beta(y-\psi)^{\beta-1}}{\eta^{\beta}} \exp\{-(\frac{y-\psi}{\eta})^{\beta}\}, \quad y > \psi. \tag{28}$$

This model has a discontinuity related to the endpoint parameter, so the derivations here do not apply. Although the prior (7) based on linking to the maximum likelihood estimate cannot be constructed as $\hat{\psi}$ is not obtained from the score equation, it is possible to formally compute $V(\theta)$ using (5), with fixed quantile $z = \{y-\psi)/\eta\}^{\beta}$. This gives the volume element form

$$|V(\theta)'V(\theta)|^{1/2} \propto \frac{1}{\eta\beta} h^{1/2}(y^0, \psi),$$

where $h(y^0,\psi) = \Sigma(y_i^0-\psi)^2\Sigma(y_i^0-\psi)^2\log^2(y_i^0-\psi) - \{\Sigma(y_i^0-\psi)^2\log(y_i^0-\psi)\}^2$. Lin et al (2009) propose a combination reference/right Haar prior for this model

which is proportional to $1/(\eta\beta)$, and note that the posterior is improper unless the range of ψ is restricted. The factor $h(y^0, \psi)$ enforces a restriction on the range of ψ, since it is undefined for $\psi > y^0_{(1)}$.

To summarize, the main conclusions that emerge from the developments in this paper are that priors that ensure calibration of the resultant posterior inferences need to depend on the data, and that a global prior ensuring this calibration does not seem to be possible for nonlinear parameters of interest unless the nuisance parameter is a scalar. Other approaches to deriving targetted priors for the full parameter space have analogous difficulties. The Welch-Peers approach leads to a family of priors $\pi(\theta)d\theta \propto i_{\psi\psi}(\theta)^{1/2}g(\lambda)$ and efforts to choose a unique form for $g(\cdot)$ have had limited success. The reference prior approach requires care as well in the construction of targetted priors with vector nuisance parameters: in particular the parameters need to be ordered and grouped, and the results depend on this choice.

These results suggest that a completely general calibration of Bayesian posterior inferences is not possible through the choice of the prior, and that calibration needs to be checked on a case by case basis.

Appendix

(i) Example 3.3: Background on transformation models

The parameter θ of a transformation model is an element of a transformation group that operates smoothly and exactly on the sample space of the model; for background details see Fraser (1979). The response y is then generated as $y = \theta z$ where z is an error or reference variable on the sample space. An observed value $y = y^0$ then determines that the antecedent realized error value, say z^r, such that $Gy^0 = Gz^r$ and this subset is an ancillary contour.

Conditioning on the identified subset gives $y = \theta z$ where the connection between any two elements is one-to-one when the remaining variable is held fixed. The conditional model has the form $\tilde{f}(y;\theta)dy = \tilde{g}(\theta^{-1}y)d\mu(y)$ where $d\mu(\cdot)$ is the

left invariant measure.

The notation is simplified if the group coordinates are centered so that the identity element is at the maximum density point of the conditional error density; thus $\tilde{g}(z) \leq \tilde{g}(e)$ where e designates the identity element satisfying $ez = z$. The maximum likelihood group element $\hat{\theta}(y)$ is then the solution of $\theta^{-1}y = e$ which gives $\hat{\theta}(y) = y$. We then have from (7) that the default prior is

$$|W(\theta)|d\theta = \left|\frac{dy}{d\theta}\right|_{y^0} d\theta = \left|\frac{d(\theta z)}{d\theta}\right|_{\theta z = y^0} d\theta \tag{29}$$

where the differentiation is for fixed reference value z with the subsequent substitution $\theta z = y^0$ or $z = z^0(\theta) = z(y^0, \theta)$. The Jacobian can be evaluated using notation from Fraser (1979, p.144): let $J^*(h; g) = |\partial gh/\partial h|$ with variable g and then $J^*(g) = J^*(g; e)$; this gives $d\nu(g) = dg/J^*(g)$ where $d\nu(g)$ is the right invariant measure. We then have $d\theta z = J^*(\theta z)d\nu(\theta z)$ with θ as the variable. Then with θz set equal to y^0 we obtain

$$\begin{aligned} d\theta z &= J^*(y^0)d\nu(\theta z) \\ &= J^*(y^0)d\nu(\theta) \end{aligned}$$

using the right invariance of ν, which is a constant times the right invariant measure $d\nu(\theta)$ on the group. We thus have that the default prior (29) is $\pi(\theta)d\theta = cd\nu(\theta)$.

(ii) Section 4: Rescaling the parametrization of the approximating exponential model

The exponential model approximation to a general model (15) depends on θ only through the observed log-likelihood function $\ell(\theta)$ and the observed log-likelihood gradient function $\varphi(\theta)$. The r_f^* approximation (16) is computed entirely within this model, with log-likelihood function

$$\ell(\theta; s) = \ell(\theta) + \varphi'(\theta)s \tag{30}$$

and observed data $s = 0$.

For scalar θ and φ we have $\ell_{(\theta)}(\theta) = \ell_{\varphi}(\theta) = \ell_{\theta}(\theta)\varphi_{\theta}^{-1}(\theta)$ where the subscript as usual denotes differentiation. Then differentiating again we obtain

$$\ell_{(\theta\theta)}(\theta) = \ell_{\varphi\varphi}(\theta) = \ell_{\theta\theta}(\theta)\varphi_{\theta}^{-2}(\theta) - \ell_{\theta}(\theta)\varphi_{\theta\theta}(\theta)\varphi_{\theta}^{-3}(\theta).$$

An analogous formula is available for the vector case using tensor notation.

Now consider a vector $\theta = (\psi, \lambda)$ with scalar components. The information $j_{(\lambda\lambda)}(\theta)$ concerns the scalar parameter model with ψ fixed. This model can have curved ancillary contours on the initial score space $\{s\}$ if for example ψ is not linear in $\varphi(\theta)$. Correspondingly the differentiation with respect to (λ) requires the use of the φ metric for λ given ψ and the results depend on the use of the standardization $\hat{j}^{0}_{\varphi\varphi} = I$. From the preceding scalar derivative expression we obtain

$$j_{(\lambda\lambda)}(\theta) = j_{\lambda\lambda}(\theta)|\varphi_{\lambda}(\theta)|^{-2} - \ell_{\lambda}(\theta)\varphi_{\lambda\lambda}(\theta)|\varphi_{\lambda}(\theta)|^{-3},$$

where as usual $|\varphi_{\lambda}|^2 = |\varphi'_{\lambda}\varphi_{\lambda}|$. To convert back to the to the initial λ scale we write

$$j_{[\lambda\lambda]}(\theta) = j_{\lambda\lambda}(\theta) + \ell_{\lambda}(\theta)\varphi_{\lambda\lambda}(\theta)|\varphi_{\lambda}(\theta)|^{-1}.$$

(iii) Linear parameters and marginalization paradoxes

Dawid et al. (1973) showed that in some cases it is not possible to construct a prior for which the inference obtained by marginalizing the posterior distribution for the full parameter is consistent with that obtained by marginalizing the prior distribution to the parameter of interest and using this on the likelihood function based on a reduced model. The normal circle problem of Example 3.2 is a simple example of this, with the reduced model being that for $r^2 = y_1^2 + y_2^2$, which has a distribution depending only on the parameter of interest ψ. One conclusion of Dawid et al. (1973) is that improper priors for vector parameters may lead to anomalous results for inference about component parameters. The default priors of Section 2 share this drawback, and are only appropriate for marginal inference on component parameters that are linear, in the sense that they are consistent with location-type models inherent in their construction. We call such component parameters linear.

Formally, we call a parameter contour $\psi(\theta) = \psi_0$ linear if a change $d\lambda$ in the nuisance parameter λ for fixed $\psi = \psi_0$ generates through (6) a direction at the data point that is confined to a subspace free of λ and with dimension equal to $\dim(\lambda)$. This is an extension of the result for $f(y_1 - \theta_1, y_2 - \theta_2)$ where a change in θ_2 applied to $y_1 = \theta_1 + z_1, y_2 = \theta_2 + z_2$ gives the y_2 direction which corresponds to fixed y_1. For the normal circle example we note that the radius ψ is curved but the angle α is linear.

The linearity condition defines a location relationship between the nuisance parameter λ for fixed ψ and change at the data point. As such it provides an invariant or flat prior for the constrained model, and thereby leads to a marginal model with the nuisance parameter eliminated. This avoids the marginalization paradoxes and parallels the elimination of a linear parameter in the standard location model.

We now consider a two parameter model parameterized by $\theta = (\theta_1, \theta_2)$, with parameter of interest $\psi(\theta)$, and develop the linear parameter that coincides with $\psi(\theta)$ in a neighbourhood of the observed maximum likelihood value $\hat{\theta}^0$. From (6) we have

$$\begin{aligned} d\hat{\theta}_1 &= w_{11}(\theta)d\theta_1 + w_{12}(\theta)d\theta_2 \\ d\hat{\theta}_2 &= w_{21}(\theta)d\theta_1 + w_{22}(\theta)d\theta_2, \end{aligned} \tag{31}$$

which can be inverted using coefficients $w^{ij}(\theta)$ to express $d\theta$ in terms of $d\hat{\theta}$.

First we examine the parameter $\psi(\theta)$ near $\hat{\theta}^0$ on the parameter space and find that an increment $(d\theta_1, d\theta_2)$ with no effect on $\psi(\theta)$ must satisfy $d\psi(\theta) = \hat{\psi}_1^0 d\theta_1 + \hat{\psi}_2^0 d\theta_2 = 0$ where $\psi_i(\theta) = \partial\psi(\theta)/\partial\theta_i$; i.e. $d\theta_1 = -(\hat{\psi}_2^0/\hat{\psi}_1^0)d\theta_2$. Next we use (31) to determine the corresponding sample space increment at $\hat{\theta}^0$, and obtain

$$\frac{d\hat{\theta}_1}{d\hat{\theta}_2} = \frac{-\hat{w}_{11}^0\hat{\psi}_2^0 + \hat{w}_{12}^0\hat{\psi}_1^0}{-\hat{w}_{21}^0\hat{\psi}_2^0 + \hat{w}_{22}^0\hat{\psi}_1^0} = \frac{c_1}{c_2};$$

thus (c_1, c_2) so defined gives a direction $(c_1, c_2)dt$ on the sample space that corresponds to no ψ-change. Finally we use the inverse of (31) to determine the parameter space increment at a general point θ that corresponds to the preceding

sample space increment, giving

$$d\theta = \begin{pmatrix} w^{11}(\theta)c_1 + w^{12}(\theta)c_2 \\ w^{21}(\theta)c_1 + w^{22}(\theta)c_2 \end{pmatrix} dt, \tag{32}$$

as a tangent to the linearized version of $\psi(\theta)$. We then have either the explicit contour integral solution

$$\theta(t) = \hat{\theta}^0 + \int_0^t \begin{pmatrix} w^{11}(\theta(t))c_1 + w^{12}(\theta(t))c_2 \\ w^{21}(\theta(t))c_1 + w^{22}(\theta(t))c_2 \end{pmatrix} dt,$$

which describes the iterative solution of the differential equation (32), or the implicit equation $\theta_2 = \theta_2(\theta_1)$ as the direct solution of the differential equation

$$\frac{d\theta_2}{d\theta_1} = \frac{w^{21}(\theta)c_1 + w^{22}(\theta)c_2}{w^{11}(\theta)c_1 + w^{12}(\theta)c_2}.$$

This defines to second order a linear parameter that is equivalent to $\psi(\theta)$ near $\hat{\theta}^0$.

Example A.1. We reconsider the regression Example 3.1, but for notational ease restrict attention to the simple location-scale version with design matrix $X = 1$. We construct the linear parameter that agrees with the quantile parameter $\mu + k\sigma$ near $\hat{\theta}^0$ for some fixed value of k. From $W(\theta)$ in that example we obtain

$$\begin{aligned} d\hat{\mu} &= d\mu + (\hat{\mu}^0 - \mu)d\sigma/\sigma \\ d\hat{\sigma} &= \hat{\sigma}^0 d\sigma/\sigma. \end{aligned} \tag{33}$$

For simplicity here and without loss of generality due to location scale invariance, we work with observed data $(\hat{\mu}^0, \hat{\sigma}^0) = (0, 1)$ and have

$$\begin{aligned} d\hat{\mu} &= d\mu - \mu d\sigma/\sigma \\ d\hat{\sigma} &= d\sigma/\sigma. \end{aligned} \tag{34}$$

Inverting this gives

$$\begin{aligned} d\mu &= d\hat{\mu} + \mu d\hat{\sigma} \\ d\sigma &= \sigma d\hat{\sigma}. \end{aligned} \tag{35}$$

First we examine $\mu + k\sigma$ in the neighbourhood of $\hat{\theta}^0$ on the parameter space and have that an increment $(d\mu, d\sigma)$ must satisfy $d(\mu + k\sigma) = 0$ at $\hat{\theta}^0 = (\hat{\mu}^0, \hat{\sigma}^0) = (0, 1)$; this gives $d\mu = -kd\sigma$ at $\hat{\theta}^0$. Next we determine the corresponding increment at

$\hat{\theta}^0$ on the sample space $\{(\hat{\mu}, \hat{\sigma})\}$; from (34) we have $d\hat{\mu} = d\mu$ and $d\hat{\sigma} = d\sigma$ at this point, which gives $d\hat{\mu} = -kd\hat{\sigma}$. Finally we determine what the restriction $d\hat{\mu} = -kd\hat{\sigma}$ on the sample space implies for $(d\mu, d\sigma)$ at a general point on the parameter space; from (35) this is

$$\frac{d\mu}{d\sigma} = \frac{\mu - k}{\sigma}$$

with initial condition $(\mu, \sigma) = (0, 1)$. This gives $\mu = -k(\sigma - 1)$ or $\mu + k\sigma = k$, which shows that $\mu + k\sigma$ is linear.

Example A.2. For the normal circle Example 3.2 with parameter of interest $\psi = (\theta_1^2 + \theta_2^2)^{1/2}$, the increment on the parameter space at $\hat{\theta}^0$ with fixed ψ satisfies $d\theta_1 = -\tan\hat{\alpha}^0 d\theta_2 = -(y_2^0/y_1^0)d\theta_2$. This then translates to the sample space at (y_1^0, y_2^0) using the specialized version of (31) to give $dy_2 = -(y_2^0/y_1^0)dy_1$, and this then translates back to a general point on the parameter space using the specialized version of (32) to give a line through $\hat{\theta}^0$ described by $d\theta_2 = -(y_2^0/y_1^0)d\theta_1$, which is perpendicular to the radius and thus tangent to the circle through the data point; this is the linear parameter equivalent to ψ near $\hat{\theta}_0$.

An extension of this linearity leads to a locally defined curvature measure that calibrates the marginalization discrepancy and can be used to correct for such discrepancies to second order. We do not pursue this here.

(iv) Strong matching and information approximation

In the scalar case, strong matching of Bayesian and frequentist approximations gives the expression for the prior as

$$\frac{\pi(\theta)}{\pi(\hat{\theta}^0)} = \frac{d\beta(\theta)}{d\theta} = -\frac{\ell_\theta(\theta; y^0)}{\varphi(\theta) - \varphi(\hat{\theta}^0)}$$

where $d\beta(\theta)$ is a locally defined linear parameter (Fraser & Reid, 2002).

If the model is a full exponential family with log-likelihood function $\ell(\theta) = \theta t - k(\theta)$ then we obtain

$$\frac{d\beta(\theta)}{d\theta} = -\frac{t^0 - k'(\theta)}{\theta - \hat{\theta}^0} = \frac{k'(\theta) - k'(\hat{\theta}^0)}{\theta - \hat{\theta}^0}$$

$$
\begin{aligned}
&= \frac{k'(\hat\theta^0) + (\theta-\hat\theta^0)k''(\hat\theta^0) + (1/2)(\theta-\hat\theta^0)^2 k'''(\hat\theta^0)}{\theta-\hat\theta^0} \\
&= k''(\hat\theta^0)\{1 + (1/2)(\theta-\hat\theta^0)k'''(\hat\theta^0)/k''(\hat\theta^0)\}
\end{aligned}
$$

whereas the usual Jeffreys' prior is

$$
i^{1/2}(\theta) = k''(\theta)^{1/2} = \{k''(\hat\theta^0) + (\theta-\hat\theta^0)k'''(\hat\theta^0)\}^{1/2} = k''(\hat\theta^0)^{1/2}\{1 + (1/2)(\theta-\hat\theta^0)k'''(\hat\theta^0)/k''(\hat\theta^0)\};
$$

these are asymptotically equivalent.

The same argument can be applied to the approximating exponential model (15), and this leads to the approximation used at (21).

Acknowledgements

We are grateful to the editors and referees for very careful and constructive comments. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–322.

[2] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc. London* **53**, 370–418; **54**, 296–325. Reprinted in *Biometrika* **45**(1958), 293–315.

[3] Bédard, M., Fraser, D.A.S. and Wong, A.C.M. (2007). Higher accuracy for Bayesian and frequentist inference: large sample theory for small sample likelihood. *Statist. Sci.* **22**, 301-321.

[4] Berger, J.O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* **1**, 385–402, DOI:10.1214/06-BA115.

[5] Berger, J.O. and Bernardo, J.M. (1992). On the development of the reference prior method. in *Bayesian Statistics 4: Proceedings of the Fourth Valencia Inter-*

national Meeting, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 35-60. Clarendon Press, Oxford.

[6] Berger, J.O., Bernardo, J.M. and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37**, 905–938.

[7] Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc.* B **41**, 113–147 (with discussion).

[8] Box, G.E.P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, MA.

[9] Box, G. and Cox, D.R. (1964). An analysis of transformations. *J. Roy. Statist. Soc.* B **26**, 102–107.

[10] Casella, G., DiCiccio, T.J. and Wells, M.T. (1995). Inference based on estimating functions in the presence of nuisance parameters: Comment: Alternative aspects of conditional inference. *Statistical Science* **10**, 179–185.

[11] Clarke, B.S. (2007). Information optimality and Bayesian modelling. *Journal of Econometrics*, **138**, 405–429.

[12] Clarke, B.S. and Barron, A. (1994). Jeffreys prior is asymptotically least favorable under entropy risk. *J. Statist. Plann. Infer.* **41**, 37–60.

[13] Clarke, B.S. and Yuan, A. (2004). Partial information reference priors: derivation and interpretations. *J. Statist. Plann. Infer.* **123**, 313–345.

[14] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics.* Chapman & Hall, London.

[15] Datta, G.S. and Ghosh, M. (1995). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90**, 1357–1363.

[16] Datta, G.S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics.* Springer-Verlag, New York.

[17] Davison, A.C., Fraser, D.A.S. and Reid, N. (2006). Likelihood inference for categorical data. *J. R. Statist. Soc* B **68**, 495–508.

[18] Dawid, A.P., Stone, M. and Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc.* B **35**, 189–233.

[19] DiCiccio and Martin (1991). Approximations of marginal tail probabilities for a class of smooth functions with applications to Bayesian and conditional inference. *Biometrika* **78**, 891–902.

[20] Eaves, D. (1983). On Bayesian nonlinear regression with an enzyme example. *Biometrika* **70**, 373–379.

[21] Fraser, D.A.S. (1979). *Inference and Linear Models.* McGraw-Hill, New York.

[22] Fraser, D.A.S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327–339.

[23] Fraser, D.A.S. and Reid, N. (1993). Third order asymptotic models: likelihood functions leading to accurate approximations for distribution functions. *Statist. Sinica* **3**, 67–82.

[24] Fraser, D.A.S. and Reid, N. (2002). Strong matching of frequentist and Bayesian inference. *J. Statist. Plan. Infer.* **103**, 263–285.

[25] Fraser, D.A.S., Reid, N. and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–264.

[26] George, E.I. and McCulloch, R. (1993). On obtaining invariant prior distributions. *J. Statist. Plann. Infer.* **37**, 169–179.

[27] J.K. Ghosh, J.K., Chakrabarti, A. and Samanta, T. (2009). Entropies and metrics that lead to Jeffreys and reference priors. preprint.

[28] Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. *Bayesian Analysis* **1**, 403 – 420, DOI:10.1214/06-BA116.

[29] Jeffreys, H. (1961). *Theory of Probability.* 3rd Edition. Oxford University Press, Oxford.

[30] Kass, R.E. (1990). Data-translated likelihood and Jeffreys's rules. *Biometrika* **77**, 107–114.

[31] Kass, R.E and Wasserman, L. (1996). Formal rules for selecting prior distributions: a review and annotated bibliography. *J. Am. Statist. Assoc.* **91**, 1343–70.

[32] Lin, X., Sun, D. and Berger, J.O. (2009). Objective Bayesian analysis under semi-invariance structure. presented at OBayes 09, June, 2009. `http://www-stat.wharton.upenn.edu/statweb/Conference/OBayes09/lectures.htm`, accessed on June 20, 2009.

[33] Little, R. (2006). Calibrated Bayes: a Bayes/frequentist roadmap. *Amer. Statist.* **60**, 1–11.

[34] Peers, H.W. (1965). On confidence points and Bayesian probability points int he case of several parameters. *J. R. Statist. Soc.* B, **27**, 9–16.

[35] Pierce, D.A. and Peters, D.L. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. Roy. Statist. Soc.* B **54**, 701–737.

[36] Pierce, D.A. and Peters, D.L. (1994). Higher-order asymptotics and the likelihood principle: one-parameter models. *Biometrika* **81**, 1–10.

[37] Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.* **31**, 1695–1731.

[38] Reid, N., Mukerjee, R. and Fraser, D.A.S. (2003) Some aspects of matching priors. *Mathematical Statistics and Applications: Festschrift for C. VanEeden* (M. Moore, S. Froda, C. Léger, eds.) 31–44. Lecture notes Monograph Series 42, Institute of Mathematical Statistics, Hayward.

[39] Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.* **30**, 877–880.

[40] Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 705–708.

[41] Tierney, L.J. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Ass.* **81**, 82–87.

[42] Wasserman, L. (2000). Asymptotic inference for mixture models using data dependent priors. *J. Roy. Statist. Soc.* B **62**, 159–180.

[43] Welch, B.L. and Peers, H.W. (1963). On formulae for confidence points based in intervals of weighted likelihoods. *J. Roy. Statist. Soc.* B **25**, 318–329.

[44] Yang, R. and Berger, J.O. (1996). A catalog of noninformative priors. Duke University Technical Report 96-42.
`http://citeseer.ist.psu.edu/old/401050.html`, accessed June 16, 2009.

[45] Zellner, A. (1988). Optimal information processing and Bayes' theorem. *Amer. Statist.* **42**, 278–284.