Journal of the Royal Statistical Society
SERIES B Statistical Methodology

*J. R. Statist. Soc.* B (2010)
**72**, *Part* 5, *pp.* 631–654

# Default priors for Bayesian and frequentist inference

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

*University of Toronto, Canada*

E. Marras

*Centre for Advanced Studies, Research and Development in Sardinia, Pula, Italy*

and G. Y. Yi

*University of Waterloo, Canada*

[Received July 2009. Revised March 2010]

**Summary.** We investigate the choice of default priors for use with likelihood for 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 that are not of interest and then 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 targeted on a component parameter of interest and by targeting avoids the marginalization paradoxes of Dawid and co-workers. This modifies Jeffreys's prior and provides extensions to the development of Welch and Peers. These two approaches are combined to explore priors for a vector parameter of interest in the presence of a vector nuisance parameter. Examples are given to illustrate the computation of the priors.

*Keywords*: Invariant prior; Jeffreys prior; Likelihood asymptotics; Marginalization paradox; Non-informative 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 that is 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 about 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 viewpoint a default prior offers a flexible exploratory approach to statistical inference.

There is a large literature on the construction of many types of default prior, 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 $\theta$ arose from some density $g(\theta)$. A very helpful survey of

*Address for correspondence*: N. Reid, Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario, M5S 3G3, Canada.
E-mail: reid@utstat.toronto.edu

 1369–7412/10/72631

various methods for constructing default priors is given in Kass and Wasserman (1996). In most discussions there is some emphasis on ensuring correct calibration of posterior probabilities, in the sense that these represent probability under the model, at least approximately. A recent discussion of this appears in Berger (2006); Goldstein (2006) gives a contrary view. Our 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 these approaches lead to Jeffreys's 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 on the basis of invariance arguments, and this was further pursued by Box and Tiao (1973) through the concept of data-translated likelihoods; see Kass (1990). George and McCulloch (1993) derived a class of invariant priors and developed a link to 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 and Bernardo (1992) seeks to maximize the Kullback–Leibler divergence from the prior to the posterior distribution: Clarke and Barron (1994) related this to least favourable 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 and Peers (1963) derived Jeffreys's prior by a probability matching argument based on Edgeworth expansions.

In extending these results to problems with nuisance parameters several difficulties arise. The Welch–Peers approach was used in Peers (1965), Tibshirani (1989) and in several papers by Mukerjee and colleagues; a review of this literature is provided by Datta and Mukerjee (2004). These extensions considered the construction of matching priors by 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 and Yuan (2004) have given 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 and 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 expression (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 targeted on the parameter of interest when there is a type of non-linearity in that parameter; this is an aspect of the marginalization paradox of Dawid *et al.* (1973). In Section 4 we use third-order approximations for $p$-values and posterior probabilities to derive a suitably targeted 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 targeted prior, given in Section 4 by equation (28).

The information-based approach, however, seems to be limited to the case of scalar interest and scalar nuisance parameters, and to extend this to vector subparameters we return to

the approach of Section 2. This is described in Section 5, and in Section 6 we record 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 $\theta$, and assume that 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 and Cox (1964) and Wasserman (2000). Pierce and Peters (1994) noted that agreement of Bayesian and frequentist higher order approximations would in general require data-dependent priors. Clarke (2007) discussed the role of data-dependent priors in the context of priors constructed by information processing arguments.

## 2. Default priors from model properties

Suppose that we have a scalar parameter $\theta$ and a single observation from a model with density function $f(y;\theta)$ and distribution function $F(y;\theta)$, where $F$ is stochastically increasing and continuously differentiable in both $y$ and $\theta$. For an observed value $y^0$, the $p$-value as a function of $\theta$ is $F(y^0;\theta)$ and the posterior right-hand tail distribution function is

$$s(\theta) = c(y^0) \int_\theta f(y^0;\vartheta)\,\pi(\vartheta)\,\mathrm{d}\vartheta.$$

With location models these two functions are equal, giving support to the Bayes (1763) proposal for inference. If these two inference functions are to be equal more generally, thus providing equivalence of posterior and frequentist inference, then

$$F(y^0;\theta) = c \int_\theta f(y^0;\vartheta)\,\pi(\vartheta)\,\mathrm{d}\vartheta.$$

Differentiating both sides with respect to $\theta$ gives

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

which determines the data-dependent prior as

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

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 $\theta$-values.

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

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

$$\begin{aligned} 0 &= \mathrm{d}F(y;\theta)|_{y=y^0} \\ &= F_\theta(y^0;\theta)\,\mathrm{d}\theta + F_y(y^0;\theta)\,\mathrm{d}y; \end{aligned}$$

the effect at $y^0$ of parameter change at $\theta$ is

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

The same calculation when $\theta$ is a vector of dimension $p$, with $y$ still a scalar, gives

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

translation invariance becomes local translation invariance (Fraser, 1964). For a vector $a$ we write $a'$ for $a^{\mathrm{T}}$ when there is no risk of confusion with differentiation. 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:

$$\begin{aligned} \left.\frac{\mathrm{d}y}{\mathrm{d}\theta}\right|_{y^0} &= \left.\frac{\partial}{\partial\theta}y(u;\theta)\right|_{u=F(y^0;\theta)} \\ &= y_\theta(u;\theta)|_{u=F(y^0;\theta)}. \end{aligned} \tag{4}$$

In equations (2)–(4) differentiation with respect to $\theta$ is calculated with the $p$-value $F(y;\theta)$ held fixed, and then evaluated at the data point. Any pivotal quantity that is a one-to-one function of $F(y;\theta)$ gives the same definition of $\mathrm{d}y/\mathrm{d}\theta$.

With a sample of independent observations, $y = (y_1, \ldots, y_n)$, each $y_i$ has a corresponding row vector, $V_i(\theta)$ say, which is obtained from equation (3) by using its distribution function. The change in the vector variable $y$ at $y^0$ relative to change in $\theta$ is

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

where $V(\theta)$ is an $n \times p$ matrix that we call the sensitivity of $\theta$ at $y^0$. We denote the columns of $V(\theta)$ by $\{v_1(\theta) \ldots v_p(\theta)\}$ where the $n \times 1$ vector $v_j(\theta)\,\mathrm{d}\theta_j$ gives the data displacement when the $j$th co-ordinate of $\theta$ is changed by $\mathrm{d}\theta_j$.

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 $\theta$ then $V(\theta) = 1$, and as noted above with a flat prior for $\theta$ posterior probabilities are equal to observed $p$-values. Indeed Bayes (1763) used an analogy (Stigler, 1982) and invariance argument to recommend the flat prior $\pi(\theta) = c$ for the parameter $\theta$, 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 $\theta$, which gives posterior probabilities, since equation (5) can be written $\mathrm{d}y = V(\theta)\,\mathrm{d}\theta$. By analogy with the location model, a natural default prior is $\pi_V(\theta) \propto |V(\theta)| = |V(\theta)'\,V(\theta)|^{1/2}$, the volume element determined by $V(\theta)$. As $\mathrm{d}y = V(\theta)\,\mathrm{d}\theta$ is $p$ dimensional for an increment $\mathrm{d}\theta$ at $\theta$, it is convenient to use maximum likelihood co-ordinates $\hat{\theta}(y)$ in place of $y$.

Writing $l(\theta;y) = \log\{f(y;\theta)\}$ for the log-likelihood function, and $\hat{\theta} = \hat{\theta}(y)$ for the maximum likelihood statistic that is obtained by solving the score equation $l_\theta(\theta;y) = 0$, we find

the connection between $y$ and $\hat{\theta}(y)$ by calculating the total derivative of the score equation $l_\theta(\theta; y) = 0$. At $(\hat{\theta}^0; y^0)$ we have

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

where the differentials $\mathrm{d}\hat{\theta}$ and $\mathrm{d}y$ are respectively $p \times 1$ and $n \times 1$ vectors, and

$$l_{\theta;y'}(\hat{\theta}^0; y^0) = (\partial/\partial\theta)(\partial/\partial y')\, l(\theta; y) = H'$$

is the gradient of the score function at the data point. Solving for $\mathrm{d}\hat{\theta}$ gives

$$\mathrm{d}\hat{\theta} = \hat{\jmath}^{-1} H'\,\mathrm{d}y,$$

where $\hat{\jmath} = j(\hat{\theta}^0; y^0) = -\partial^2 l(\hat{\theta}^0; y^0)/\partial\theta\,\partial\theta'$ is the observed Fisher information. Substituting $\mathrm{d}y = V(\theta)\,\mathrm{d}\theta$ from equation (5) gives

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

which presents the sample space change $\mathrm{d}\hat{\theta}$ at $\hat{\theta^0}$ in terms of parameter space change $\mathrm{d}\theta$ at arbitrary $\theta$. Expressing this as a volume element determined by $W$ gives our default prior

$$\pi(\theta)\,\mathrm{d}\theta \propto |W(\theta)|\,\mathrm{d}\theta = |\hat{\jmath}^{-1} H' V(\theta)|\,\mathrm{d}\theta. \tag{7}$$

As at expression (1) this is a data-dependent prior, although we have suppressed the dependence of $\pi(\cdot)$ on $y^0$ in expression (7).

For calculations with component parameters it may be natural to standardize with respect to observed information. Let $\hat{\jmath}^{1/2}$ be a right square root of the observed information matrix $\hat{\jmath}$ and define the information-standardized vector differential

$$\hat{\jmath}^{1/2}\,\mathrm{d}\hat{\theta} = \hat{\jmath}^{1/2}\, W(\theta)\,\mathrm{d}\theta = (\hat{\jmath}^{-1/2})'\, H'\, V(\theta) = \tilde{W}(\theta)\,\mathrm{d}\theta. \tag{8}$$

The rescaled default prior is then

$$\pi(\theta)\,\mathrm{d}\theta \propto |\tilde{W}(\theta)|\,\mathrm{d}\theta, \tag{9}$$

which is equivalent to expression (7); the matrix $\tilde{W}(\theta)$ is used in example 9 in Section 5.

As we shall see these priors can lead to posterior survivor values that duplicate to second order the frequentist $p$-values that are available from asymptotic theory for linear component parameters; otherwise the marginalization paradox of Dawid *et al.* (1973) provides a limiting factor in the presence of parameter curvature. Marginalization issues are examined in Sections 4 and 5.

## 3. Examples of default priors

### 3.1. Example 1: normal theory linear regression

Suppose that $y_i$ follows a normal distribution with mean $X_i\beta$ and variance $\sigma^2$, where $X_i$ is the $i$th row of an $n \times p$ design matrix $X$, $\beta$ is $p \times 1$ 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 distribution, gives the quantile functions as the usual expressions $y_1 = X_1\beta + \sigma z_1, \ldots, y_n = X_n\beta + \sigma z_n$ for the model. We compute $V(\theta)$ for fixed $u$, as described at equation (4), or equivalently for fixed $z$, obtaining

$$V(\theta) = \left.\frac{\partial(X\beta + \sigma z)}{\partial(\beta_1 \ldots \beta_p \sigma^2)}\right|_{y^0} = \left\{X, \frac{z^0(\theta)}{2\sigma}\right\},$$

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 $\theta$. The sample space derivative of the log-likelihood $l_{;y} = (\partial/\partial y)l = (X\beta - y)/\sigma^2$ and $l_{\theta';y}(\theta; y) = \{\sigma^{-2}X, \sigma^{-4}(y - X\beta)\}$ give

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

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

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

leading to

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

or equivalently to

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

in the modified parameterization $(\beta', \sigma)$.

The default prior given by expression (7) is

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

which is the familiar right invariant prior. This example illustrates how expression (7) modifies the invariance argument of Jeffreys to adapt to the local location form of the distribution function for each co-ordinate. Jeffreys's prior is the square root of the determinant of the expected Fisher information matrix, which leads to the left invariant prior for $(\beta, \sigma^2)$, which is proportional to $\mathrm{d}\beta\,\mathrm{d}\sigma^2/\sigma^{p+2}$. 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 $\beta$, whereas the right invariant prior does and agrees with Jeffreys's (1961) proposal for a modified rule for location–scale settings (Kass and Wasserman, 1996).

### 3.2. Example 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 location model example, or from direct calculation, that the default prior for the location parameter $\mu$ is $c\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2$ and the associated posterior for $(\mu_1, \mu_2)$ is $N\{(y_1^0, y_2^0); I/n\}$. For any component parameter that is linear in $(\mu_1, \mu_2)$ we then have exact agreement between frequentist $p$-values and Bayesian survivor probabilities.

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

$$\mathrm{d}\mu = \begin{pmatrix} \cos(\alpha) & -\rho\sin(\alpha) \\ \sin(\alpha) & -\rho\cos(\alpha) \end{pmatrix} \begin{pmatrix} \mathrm{d}\rho \\ \mathrm{d}\alpha \end{pmatrix}$$

and a similar formula for $\mathrm{d}\hat{\mu}$ in terms of $(\mathrm{d}\hat{\rho}, \mathrm{d}\hat{\alpha})'$. Then from $\mathrm{d}\hat{\mu} = \mathrm{d}\mu$ for the initial co-ordinates we obtain

$$\begin{pmatrix} \mathrm{d}\hat{\rho} \\ \mathrm{d}\hat{\alpha} \end{pmatrix} = \begin{pmatrix} \cos(\hat{\alpha}) & -\hat{\rho}\sin(\hat{\alpha}) \\ \sin(\hat{\alpha}) & -\hat{\rho}\cos(\hat{\alpha}) \end{pmatrix}^{-1} \begin{pmatrix} \cos(\alpha) & -\rho\sin(\alpha) \\ \sin(\alpha) & -\rho\cos(\alpha) \end{pmatrix} \begin{pmatrix} \mathrm{d}\rho \\ \mathrm{d}\alpha \end{pmatrix}$$

giving

$$W(\theta) = \begin{Bmatrix} \cos(\hat{\alpha}-\alpha) & \rho\sin(\hat{\alpha}-\alpha) \\ -\hat{\rho}^{-1}\sin(\hat{\alpha}-\alpha) & \rho\hat{\rho}^{-1}\cos(\hat{\alpha}-\alpha) \end{Bmatrix}, \tag{11}$$

with respect to the new co-ordinates; and then from expression (7) or (9) we obtain the default prior $\pi(\theta)\,\mathrm{d}\theta \propto \rho\,\mathrm{d}\rho\,\mathrm{d}\alpha$ for the full parameter. This is equivalent to the default flat prior $\mathrm{d}\mu_1\,\mathrm{d}\mu_2$ calculated directly from the location parameter $(\mu_1, \mu_2)$ and then transformed by using the Jacobian $|\partial(\mu_1, \mu_2)/\partial(\alpha, \rho)|$.

However, this prior is not appropriate for marginal inference when the parameter of interest is the radial distance $\rho$, which is a non-linear function of the mean vector $\mu$. The marginal distribution of $y_1^2 + y_2^2$ depends only on $\rho$, and the $p$-value function from this marginal distribution is

$$p(\rho) = \Pr\{\chi_2^2(\rho^2) \leqslant n(y_1^2 + y_2^2)\},$$

where $\chi_2^2(\delta^2)$ is a non-central $\chi^2$-distribution with 2 degrees of freedom and non-centrality parameter $\delta^2$ and the $y$s are fixed at their observed values. In contrast the posterior survivor function for $\rho$ under the flat prior $\mathrm{d}\mu_1\,\mathrm{d}\mu_2$ is

$$s(\rho) = \Pr[\rho^2 \leqslant \chi_2^2\{n(y_1^2 + y_2^2)\}].$$

Numerical calculation confirms that there can be substantial undercoverage for right-hand tail intervals based on this marginal posterior (Fraser and Reid, 2002). 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

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

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 $\mu_i$ and variance 1.

This discrepancy does not appear in the first order of asymptotic theory, where both the Bayesian and the frequentist approximation to $(\hat{\rho} - \rho)\sqrt{n}$ is the standard normal distribution, so to this order of approximation $p$-values and marginal survivor probabilities are identical. This is simply reflecting the fact that any prior that does not depend on $n$ is in the limit swamped by the data and has no effect on the posterior inference. Thus to study the agreement between Bayesian and frequentist inference it is necessary to consider exact distributions or at least higher order approximations.

The inappropriateness of the point estimator of $\rho$ developed from the prior $\pi(\mu)\,\mathrm{d}\mu \propto \mathrm{d}\mu$ was pointed out in Stein (1959) and was discussed in detail in Cox and Hinkley (1974), pages 46 and 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 a locally defined location parameter. This is a version of the marginalization

paradox of Dawid *et al.* (1973), where assigning a prior to a full parameter and then marginalizing the resulting posterior necessarily conflict with the approach of reducing to a one-parameter model and assigning a marginalized prior to that parameter of interest. Curvature and local location parameters are described in more detail in Appendix A.3. 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.

### 3.3. Example 3: transformation models

The preceding examples are special cases of transformation models. In Appendix A.1 we briefly record some background on transformation models and show that the locally defined prior (7) reproduces the right invariant prior for that model type; thus expression (7) can be written

$$\pi(\theta)\,\mathrm{d}\theta \propto |W(\theta)|\,\mathrm{d}\theta = c\,\mathrm{d}\nu(\theta)$$

where $\mathrm{d}\nu(\theta)$ is the right invariant measure on the transformation group. Transformation model theory shows that this prior is fully accurate for reproducing frequentist $p$-values, provided that the parameter of interest has a form of linearity, 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 that is described by the sensitivity matrix $V(\theta)$.

### 3.4. Example 4: Welch–Peers approximation

As noted above, the construction of the default prior by using the sensitivity matrix $V(\theta)$, or the modification $W(\theta)$, gives a flat prior when $\theta$ is a location parameter. If we have a scalar parameter model with location parameter $\beta(\theta)$, then this construction gives the flat prior $\pi(\theta)\,\mathrm{d}\theta \propto \mathrm{d}\beta(\theta)$. More generally for a scalar parameter model, an approximate location parameter prior is proportional to $i^{1/2}(\theta)$, where $i(\theta)$ is the expected Fisher information $E_\theta\{-l''(\theta)\}$. This was established in Welch and Peers (1963), by showing that this choice led to the equality, to $O(n^{-1})$, of approximations to confidence and posterior bounds. This in turn implies that

$$z = \int^{\hat\theta} i^{1/2}(t)\,\mathrm{d}t - \int^{\theta} i^{1/2}(t)\,\mathrm{d}t \tag{12}$$

has a limiting standard normal distribution, and to second order has a distribution that is free of $\theta$. In quantile form this can be written $\hat\beta = \beta + z$, where $\beta(\theta) = \int^\theta i^{1/2}(t)\,\mathrm{d}t$ is the constant information reparameterization and $z$ is a quantile of the $\theta$-free distribution. Then, for fixed $z$, $\mathrm{d}\hat\beta = \mathrm{d}\beta$ gives the Jeffreys prior $\mathrm{d}\beta \propto i^{1/2}(\theta)\,\mathrm{d}\theta$. The interpretation of this prior in terms of an approximate location parameter was discussed in Kass (1990).

As a special case of a general scalar parameter model, suppose that

$$f(s;\varphi) = \exp\{\varphi s - k(\varphi)\}h(s),$$

with a sample point data $s^0 = 0$; we assume that the parameter is centred so that $\hat\varphi^0 = 0$. Then observed information is the same as expected information and equation (12) gives

$$z = \int^{\hat\varphi} j^{1/2}_{\varphi\varphi}(t)\,\mathrm{d}t - \int^{\varphi} j^{1/2}_{\varphi\varphi}(t)\,\mathrm{d}t.$$

Then using $j_{\varphi\varphi}(\hat\varphi) = k''(\hat\varphi)$ and the score equation $s - k'(\hat\varphi) = 0$ gives $\mathrm{d}s = k''(\hat\varphi)\,\mathrm{d}\hat\varphi$ and

$$\mathrm{d}z = j^{-1/2}(\hat\varphi)\,\mathrm{d}\hat\varphi - j^{1/2}(\varphi)\,\mathrm{d}\varphi.$$

Thus for fixed quantile $z$ we have

$$\mathrm{d}s = j_{\varphi\varphi}^{1/2}(\hat{\varphi}^0)\; j_{\varphi\varphi}^{1/2}(\varphi)\,\mathrm{d}\varphi; \tag{13}$$

this gives an expression that is analogous to expression (6), but in terms of the score variable instead of the maximum likelihood estimator. This links the location parameter approach for constructing priors to that based on Fisher information. In Section 4 we extend this linking to develop targeted priors from exponential family approximations.

### *3.5. Example 5: non-linear regression*

Suppose that $y_i$ are independently normally distributed with mean $x_i(\beta)$ and variance $\sigma^2$ for $i = 1, \ldots, n$, with $x_i(\beta)$ a known non-linear function of the $p \times 1$ vector $\beta$. As in example 1, the quantile functions are $y_i = x_i(\beta) + \sigma z_i$ and, with $\theta' = (\beta', \sigma^2)$, the sensitivity matrix $V(\theta)$ is

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

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

$$H = \frac{1}{\hat{\sigma}^2}\{X(\hat{\beta}^0) \quad \hat{z}^0/\hat{\sigma}\},$$
$$\hat{\jmath} = \begin{pmatrix} \hat{\jmath}_{11}/\hat{\sigma}^2 & 0 \\ 0 & n/2\hat{\sigma}^4 \end{pmatrix}$$

where

$$\hat{\jmath}_{11} = \sum_{i=1}^{n} X_i(\hat{\beta}) X_i'(\hat{\beta}) - \sum_{i=1}^{n} \{y_i - x_i(\hat{\beta})\}\, \partial^2 x_i(\hat{\beta})/\partial\beta\, \partial\beta'$$

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

$$W(\theta) = \begin{pmatrix} \hat{\sigma}^2 \hat{\jmath}_{11}^{-1} & 0 \\ 0 & 2\hat{\sigma}^4/n \end{pmatrix} \begin{Bmatrix} X'(\hat{\beta}^0)/\hat{\sigma}^2 \\ \hat{z}^{0\prime}/\hat{\sigma}^3 \end{Bmatrix} \{X(\beta) \quad z^0(\theta)/2\sigma\},$$
$$= \begin{Bmatrix} \hat{\jmath}_{11}^{-1} \sum X_i(\hat{\beta})\, X_i'(\beta) & \hat{\jmath}_{11}^{-1} \sum X_i'(\hat{\beta})\, z_i(\theta)/2\sigma \\ 2\hat{\sigma} \sum \hat{z}_i X_i(\beta)/n & (\hat{\sigma}/\sigma) \sum \hat{z}_i z_i(\theta)/n \end{Bmatrix},$$

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 non-linear function of $\beta$ 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 the linear space that is spanned by the columns of $X(\hat{\beta}^0)$, the default prior becomes $\mathrm{d}\tilde{\beta}\,\mathrm{d}\sigma^2/\sigma^2$ to $O(n^{-1})$, where $\mathrm{d}\tilde{\beta}$ designates a flat prior in co-ordinates of the tangent plane projection at the fitted data point. The two-group reference prior for this example is $|X(\beta)'\, X(\beta)|^{1/2}\,\mathrm{d}\beta\,\mathrm{d}\sigma^2/\sigma^2$ (Yang and Berger, 1996), which was also proposed on the grounds of invariance by Eaves (1983).

### 3.6. Example 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 model:

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

Jeffreys's 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 equation (2) to obtain

$$V_i(\theta) = \frac{\Gamma'(\theta)\ F(y_i^0;\theta) - \displaystyle\int_0^{y_i^0} z^{\theta-1} \log(z) \exp(-z)\, \mathrm{d}z}{\exp(-y_i^0)(y_i^0)^{\theta-1}}.$$

The $i$th entry of $H$ is $(1/y_i^0)$, so from expression (7) we have $\pi(\theta) \propto \Sigma\, V_i(\theta)/y_i^0$.

Fig. 1 shows Jeffreys's prior and $\pi(\theta)$ 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 and then have slightly different curvature as they respond differently to curvature in the model.

This example can be extended to the two-parameter gamma model, with shape $\theta$ and scale parameter $\lambda$:

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

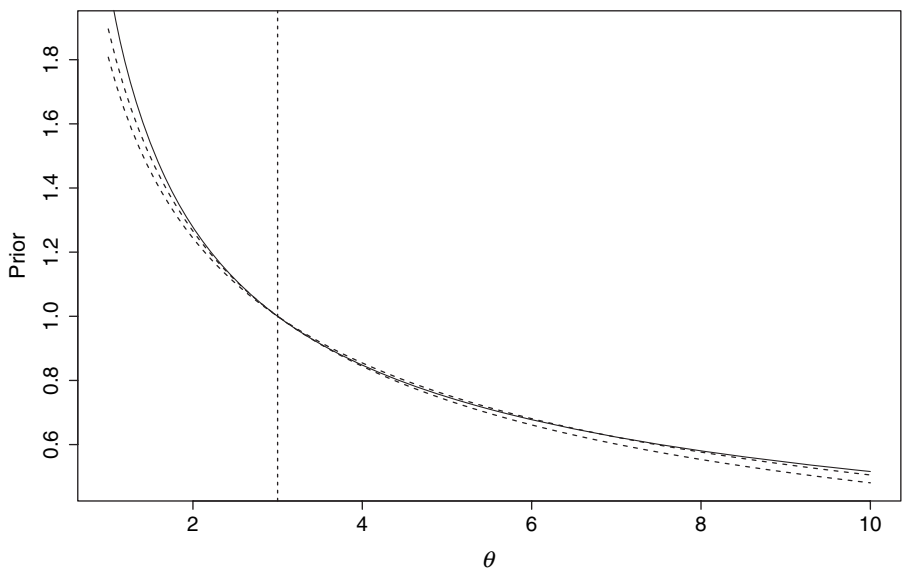


**Fig. 1.** Priors for the shape parameter of a gamma distribution: Jeffreys's prior $\pi_J(\theta)$ (———) is proportional to $\psi''(\theta)^{1/2}$ where $\psi''$ is the trigamma function; the default prior (-------) that is proposed here is based on expressions (2) and (7) and is presented here for two different samples $y^0$ of size 30 from a gamma distribution with true value $\theta = 3$; the priors are normalized to equal 1 at the true value

The sensitivity matrix $V(\lambda, \theta)$ has two columns: the $i$th element in the column corresponding to $\lambda$ is $V_{i1}(\lambda, \theta) = y_i^0/\lambda$, and the $i$th element in the column corresponding to $\theta$ is

$$V_{i2}(\lambda, \theta) = \frac{\Gamma'(\theta)\, F_1(\lambda y_i^0; \theta) - \int_0^{\lambda y_i^0} z^{\theta-1} \log(z) \exp(-z)\, \mathrm{d}z}{\lambda(\lambda y_i^0)^{\theta-1} \exp(-\lambda y_i^0)}$$

where $F_1(\cdot)$ is the distribution function for the one-parameter gamma model that was used above. The default prior (7) is

$$\pi(\theta, \lambda) \propto \frac{1}{\lambda}(\bar{V} - \hat{\lambda}\bar{W}), \tag{14}$$

where $\bar{V} = (1/n)\Sigma V_{i1}$ and $\bar{W} = (1/n)\Sigma V_{i2}/y_i^0$. Transforming expression (14) to the orthogonal parameterization $(\theta, \mu)$, where $\mu = \theta/\lambda$, gives

$$\pi(\theta, \mu) \propto \frac{1}{\mu}\left(\bar{V} - \frac{\hat{\theta}}{\hat{\mu}}\bar{W}\right). \tag{15}$$

Numerical work which is not shown indicates that the second factor in expression (15) depends very slightly on $\mu$. In contrast, the reference priors that were developed by Liseo (1993) and Sun and Ye (1996) take the form $(1/\mu)\, h(\theta)$, where $h(\theta) = [\{\theta\, \psi''(\theta) - 1\}/\theta]^{1/2}$ for the reference prior, $\{\theta\, \psi''(\theta) - 1\}/\theta^{1/2}$ for a $\theta$-matching prior that was developed in Sun and Ye (1996), and $\{\theta\, \psi''(\theta) - 1\}^{1/2}$ for Jeffreys's prior.

## 4. Targeted default priors: scalar components

The approach that was developed in the preceding sections gives a default prior for a vector parameter, but the resulting posterior is not appropriately targeted on component parameters unless the components are linear, in the sense discussed in Appendix A.3. To develop default priors that are targeted on parameters of interest, we use an approach that is motivated by higher order asymptotics and by the interpretation of the Welch–Peers prior as a location-model-based default prior as noted in example 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 targeting 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 thus may need recalibration for determining posterior probabilities for curved parameters.

The higher order approximations that are used to derive the information-based prior are accurate to $O(n^{-3/2})$ in continuous models, but the expression of the results in terms of information quantities only will be accurate just to $O(n^{-1})$. We write $\theta = (\psi, \lambda)$, where $\psi$ is a scalar parameter of interest and $\lambda$ 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, which is assumed unique, of $\partial l(\theta)/\partial\lambda = 0$.

The Laplace approximation to the marginal posterior survivor function for $\psi$ is

$$s(\psi) = \Phi(r_{\mathrm{B}}^*) = \Phi\{r + (1/r)\log(q_{\mathrm{B}}/r)\}, \tag{16}$$

where

$$r = \mathrm{sgn}(\hat{\psi} - \psi)[2\{l(\hat{\theta}) - l(\hat{\theta}_\psi)\}]^{1/2}, \tag{17}$$

$$q_{\mathrm{B}} = l'_{\mathrm{p}}(\psi) \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)}, \tag{18}$$

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

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

$$g(s; \theta) = \exp\{l(\theta) + \varphi(\theta)' s\}\, h(s) \tag{19}$$

with observed data $s^0 = 0$, and using the saddlepoint approximation. Some discussion of the use of this exponential family model $\{l(\theta), \varphi(\theta)\}$ as a full third-order surrogate for the original model is given in Davison *et al.* (2006) and Reid and Fraser (2010).

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

$$p(\psi) = \Phi(r^*_{\mathrm{f}}) = \Phi\{r + (1/r)\log(q_{\mathrm{f}}/r)\}; \tag{20}$$

where $r$ is as given above, and two equivalent expressions for $q_{\mathrm{f}}$ are

$$\begin{aligned} q_{\mathrm{f}} &= \frac{|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi) \quad \varphi_\lambda(\hat{\theta}_\psi)|}{|\varphi_\theta(\hat{\theta})|} \frac{|j(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}, \\ &= \{\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)\} \left\{ \frac{|j_{\varphi\varphi}(\hat{\theta})|}{|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|} \right\}^{1/2}; \end{aligned} \tag{21}$$

see for example Fraser *et al.* (1999) and Fraser and Reid (2002). The second version of $q_{\mathrm{f}}$ 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. The scalar parameter $\chi(\theta)$ is a rotated co-ordinate 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 $\psi(\theta) = \psi$. Information quantities concerning $\psi$ or $\lambda$ are calculated within the approximating exponential model; in particular the matrix $j_{(\lambda\lambda)}(\hat{\theta}_\psi)$ is the nuisance information matrix at the constrained maximum, re-expressed in scaling provided by the canonical exponential parameter $\varphi(\theta)$:

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

where $\varphi_\lambda(\theta) = \partial\varphi(\theta)/\partial\lambda$. This calculation is straightforward because the second derivative is evaluated at the constrained maximum likelihood estimator; if it is evaluated at other points in the parameter space then there is an additional term from the first derivative of the log-likelihood function.

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

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

where $V_i(\theta)$ is the $i$th row of the sensitivity matrix (5). This canonical parameter $\varphi(\theta)$ becomes the primary reference parameterization for calculations of our default priors. A derivation of the $r_f^*$-approximation is beyond the scope of this paper but was described in Reid (2003) and Fraser *et al.* (1999); see also Brazzale *et al.* (2007), chapter 8. The role of $V(\hat{\theta}^0)$ 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 equations (16) and (20) 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 setting $q_B = q_f$. This was suggested in Casella *et al.* (1995) and was developed further in Fraser and Reid (2002), where it was called strong matching. Strictly speaking inference for $\psi$ can be obtained by using equation (20) alone, but the close parallel between equations (16) and (20) determines some aspects of the prior that are needed to ensure frequentist validity, at least to the present order of approximation. Equating $q_f$ to $q_B$ gives

$$\frac{\pi(\hat{\theta}_\psi)}{\pi(\hat{\theta})} \propto \frac{l'_p(\psi)|j_{\lambda\lambda}(\hat{\theta}_\psi)||\varphi_\theta(\hat{\theta})|}{|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi) \quad \varphi_\lambda(\hat{\theta}_\psi)||j(\hat{\theta})|} = \frac{l'_p(\psi)}{\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)} \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j(\hat{\theta})|^{1/2}} \frac{|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}}{|j_{\varphi\varphi}(\hat{\theta})|^{1/2}}. \tag{22}$$

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

$$\pi(\theta)\,\mathrm{d}\theta = c\frac{l'(\theta; y^0)}{\varphi(\hat{\theta}^0) - \varphi(\theta)}\,\mathrm{d}\theta = \mathrm{d}\beta(\theta), \tag{23}$$

where

$$\beta(\theta) = \int^{\theta} l'(\vartheta; y^0)/\{\varphi(\hat{\theta}^0) - \varphi(\vartheta)\}\,\mathrm{d}\vartheta$$

is the local location parameter at the observed data point. This prior leads to third-order equality of posterior probabilities and $p$-values and we refer to it as a third-order prior. In Appendix A.6 we show that the prior $\mathrm{d}\beta(\theta)$ is equivalent to second order to Jeffreys's prior $\mathrm{d}\beta$ of example 4.

With nuisance parameters expression (22) gives the data-dependent prior only on the profile curve $\mathcal{C}_\psi = \{\theta : \theta = (\psi, \hat{\lambda}_\psi)\}$ in the parameter space and does not directly give a prior over the full parameter space. To extend the prior for arbitrary values of $\lambda$ we shall use the Welch and Peers (1963) methods from example 4.

We first simplify expression (22) and write

$$\pi(\hat{\theta}_\psi)\,\mathrm{d}\psi\,\mathrm{d}\lambda = c\frac{l'_p(\psi)}{\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)}\,\mathrm{d}\psi\,|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}\,\mathrm{d}\lambda, \tag{24}$$

on the profile curve $\mathcal{C}_\psi$. An extension of the argument in Appendix A.6 allows us to express the first factor in terms of observed information, to second order. In the neighbourhood of a value $\psi$, the imputed parameter $\chi$ is the canonical parameter for the conditional distribution of the score variable on the line $L_\psi$, which is the translation of the profile curve $\mathcal{C}_\psi$ into the score space; Fig. 2. This conditional distribution is obtained from the approximating exponential family (19). The first factor in equation (24) can be rewritten in terms of an increment in this score variable, giving

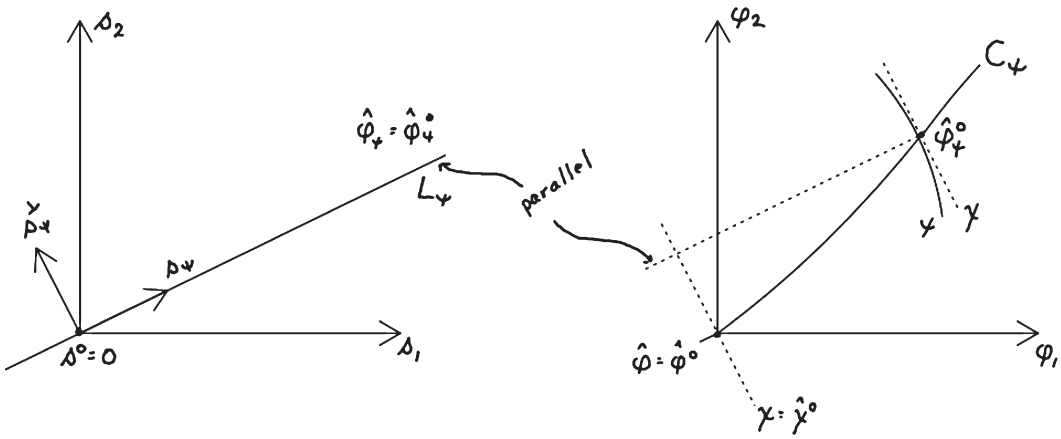


**Fig. 2.** Score and canonical parameter spaces with $p = 2$: the contour of $\psi(\varphi) = \psi$, a tested value; the observed data $s^0 = 0$; the observed maximum likelihood value $\hat{\varphi}^0 = 0$; the constrained maximum likelihood value $\hat{\varphi}^0_\psi$ lying on the profile curve $\mathcal{C}_\psi$; the observed nuisance score line $L_\psi$; the rotated $\varphi$-co-ordinate $\chi(\varphi)$ with contour $\chi = \hat{\chi}^0_\psi$ which is tangent to $\psi(\varphi) = \psi$ at $\hat{\varphi}^0_\psi$; the parameter $\chi$ tilts the profile likelihood and shifts the nominal profile distribution along the line $L_\psi$

$$\begin{aligned} \mathrm{d}s_\psi &= \frac{l'_{\mathrm{p}}(\psi)}{\chi(\hat{\varphi}) - \chi(\hat{\varphi}_\psi)}\,\mathrm{d}\psi \\ &= \frac{l'_{\mathrm{p}}(\psi)\ \psi_\chi(\hat{\theta}_\psi)}{\chi(\hat{\varphi}) - \chi(\hat{\varphi}_\psi)}\,\mathrm{d}\chi \end{aligned} \tag{25}$$

where $s_\psi$ is the score co-ordinate along the line $L_\psi$. Then extending equation (12) of example 4 we obtain the relationship between the score variable $s_\psi$ and the maximum likelihood variable $\chi(\hat{\varphi}_\psi)$ as

$$\mathrm{d}z_1 = j^{-1/2}_{\chi\chi\cdot\lambda}(\hat{\varphi}^0)\,\mathrm{d}s_\psi - j^{1/2}_{\chi\chi\cdot\lambda}(\hat{\varphi}_\psi)\,\mathrm{d}\chi,$$

where the $j_{\chi\chi\cdot\lambda}$ denotes information calculation for the $\varphi$-linear parameter $\chi$, with contours parallel to the tangent to the $\psi$-contour at $\hat{\varphi}_\psi$ but laterally displaced, and obtained from the exponential model (19). As at equation (13), with $\mathrm{d}z_1 = 0$ we have

$$\mathrm{d}s_\psi = j^{1/2}_{\chi\chi\cdot\lambda}(\hat{\varphi}^0)\, j^{1/2}_{\chi\chi\cdot\lambda}(\hat{\varphi}_\psi)\,\mathrm{d}\chi \tag{26}$$

indicating how a data change $\mathrm{d}s_\psi$ relates to parameter change $\mathrm{d}\chi$ at the value $\psi$, on the profile curve $\mathcal{C}_\psi$.

We now interpret the second factor in equation (24) by using a similar argument applied to the complementary curve in the parameter space determined by a fixed value of $\psi$. This gives

$$\mathrm{d}z_2 = j^{-1/2}_{(\lambda\lambda)}(\hat{\varphi}_\psi)\,\mathrm{d}s^\perp_\psi - j^{1/2}_{(\lambda\lambda)}(\varphi)\,\mathrm{d}(\lambda),$$

where $s^\perp_\psi$ is the score co-ordinate along the curve with fixed $\psi$. Thus using $\mathrm{d}z_2 = 0$ in the $\psi$-fixed model we have

$$\mathrm{d}s^\perp_\psi = j^{1/2}_{(\lambda\lambda)}(\hat{\varphi}_\psi)\ j^{1/2}_{(\lambda\lambda)}(\varphi)\,\mathrm{d}(\lambda),$$

giving a Welch–Peers complement to equation (26). The nuisance information away from the constrained maximum needs careful calculation; see Appendix A.2.

We combine the expressions for $\mathrm{d}s_\psi$ and $\mathrm{d}s^\perp_\psi$ and obtain

$$\begin{aligned}\pi(\theta)\,\mathrm{d}\theta &= \mathrm{d}s_\psi\,\mathrm{d}s^\perp_\psi \\ &= c\, j^{1/2}_{\chi\chi\cdot\lambda}(\hat\varphi_\psi)\,\mathrm{d}(\psi)\, j^{1/2}_{(\lambda\lambda)}(\hat\varphi_\psi)\, j^{1/2}_{(\lambda\lambda)}(\varphi)\,\mathrm{d}(\lambda).\end{aligned} \tag{27}$$

As described in Appendix A.2, we use the notation $(\theta)$ interchangeably with $\varphi$, and similarly use the notation $(\psi)$ and $(\lambda)$ for the parameter of interest and the nuisance parameter respectively, recalibrated through the canonical parameter: $\mathrm{d}(\psi) = \mathrm{d}\chi$, for example. In equation (27) we have a double Welch–Peers connection between sample space increment and various parameter change increments using score co-ordinates instead of maximum likelihood co-ordinates.

For further interpretation of the targeted default prior (27), let $j_{(\lambda_\chi\lambda_\chi)}(\hat\varphi_\psi)$ be the nuisance information relative to the contours with $\chi$ held fixed. Then $j^{1/2}_{\chi\chi\cdot\lambda}(\hat\varphi_\psi)$ and $j^{1/2}_{(\lambda_\chi\lambda_\chi)}(\hat\varphi_\psi)$ can be combined to give $|j_{\varphi\varphi}(\hat\varphi_\psi)|^{1/2}$ and we obtain

$$\pi(\theta)\,\mathrm{d}\theta = c|j_{\varphi\varphi}(\hat\varphi_\psi)|^{1/2}\,\frac{j^{1/2}_{(\lambda_\psi\lambda_\psi)}(\hat\varphi_\psi)}{j^{1/2}_{(\lambda_\chi\lambda_\chi)}(\hat\varphi_\psi)}\, j^{1/2}_{(\lambda_\psi\lambda_\psi)}(\varphi)\,\mathrm{d}(\psi)\,\mathrm{d}(\lambda_\psi). \tag{28}$$

The first factor is Jeffreys's prior for the full parameter; the second factor is an adjustment for curvature of the parameter of interest $\psi$ relative to its exponential linear version $\chi$; the third factor is a marginalization adjustment that is obtained initially as Jeffreys's prior for the nuisance parameter on the $\psi$-fixed curve, and then extended off the profile curve. The notation $(\lambda_\psi) = (\lambda)$ emphasizes the dependence of the nuisance parameterization on the parameter of interest. Some computational details for $j_{(\lambda_\chi\lambda_\chi)}(\hat\varphi_\psi)$ are recorded in Appendix A.2.

### *4.1. Example 7: normal $(\mu, \sigma^2)$*

In example 1 we obtained the default prior for the full parameter $\theta = (\mu, \sigma^2)$ as the right invariant prior $c\,\mathrm{d}\mu\,\mathrm{d}\sigma^2/\sigma^2$. We now examine how the preceding targeted default prior works for the components $\mu$ and $\sigma^2$. Without loss of generality because of model invariance we can take the data point to be $(\hat\mu, \hat\sigma^2) = (0, 1)$. The canonical parameter is $(\varphi_1, \varphi_2) = (\mu/\sigma^2, 1/\sigma^2)$, which has the information function

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

and thus the Jeffreys prior $|j_{\varphi\varphi}(\theta)|^{1/2}\,\mathrm{d}\varphi = (n\sigma^3/\sqrt{2})\,\mathrm{d}\varphi_1\,\mathrm{d}\varphi_2 = (n/\sigma^3\sqrt{2})\,\mathrm{d}\mu\,\mathrm{d}\sigma^2$ by using $|\partial\varphi/\partial\theta| = \sigma^{-6}$.

With $\mu$ as the parameter of interest and with the particular data choice we have in moderate deviations the profile curve $\mathcal{C}_\mu = \{(\mu, \hat\sigma_\mu)\} = \{(\mu, 1)\}$ where $\hat\sigma^2_\mu = \hat\sigma^2 + \mu^2 = 1 + \delta^2/n$, using centred co-ordinates for $\mu = \delta/n^{1/2}$ to $O(n^{-1})$ relative to the observed $\hat\mu = 0$. First we have, to second order, Jeffreys's prior along the profile curve where $\theta = \hat\theta_\mu = (\mu, 1)'$:

$$|j_{\varphi\varphi}(\theta)|^{1/2}\,\mathrm{d}\varphi_1\,\mathrm{d}\varphi_2 = \frac{n}{\sqrt{2}}\,\mathrm{d}\varphi_1\,\mathrm{d}\varphi_2 = \frac{n}{\sqrt{2}}\,\mathrm{d}\mu\,\mathrm{d}\sigma^2.$$

Next we calculate the $\varphi$-based nuisance information, which will give Jeffreys's prior for the constrained model:

$$j_{(\sigma^2\sigma^2)}(\theta) = j_{\sigma^2\sigma^2}\left|\frac{\partial(\varphi_1, \varphi_2)}{\partial\sigma^2}\right|^{-2} = \frac{n}{2\sigma^4}\left|\left(-\frac{\mu}{\sigma^4}, -\frac{1}{\sigma^4}\right)'\right|^{-2} = \frac{n}{2\sigma^4}\,\frac{\sigma^8}{1+\mu^2} = \frac{n\sigma^4}{2},$$

to second order. Next we note that a $\mu$-fixed contour is linear in the exponential parameterization so the curvature adjustment is 1. And next the marginalization adjustment is $(n\hat{\sigma}_\mu^4/2)^{1/2} = (n/2)^{1/2}$, ignoring terms of $O(n^{-1})$. Finally $\mathrm{d}\sigma^2$ on the profile where $\hat{\sigma}_\mu^2 = 1$ extends off the profile as $\mathrm{d}\sigma^2/\sigma^2$ using invariance for $\sigma^2$. This gives the $\mu$-targeted default prior

$$\frac{n}{\sqrt{2}} \times 1 \times \left(\frac{n}{2}\right)^{1/2} \frac{\mathrm{d}\mu\,\mathrm{d}\sigma^2}{\sigma^2} = \frac{n^{3/2}}{2\sigma^2}\,\mathrm{d}\mu\,\mathrm{d}\sigma^2$$

in moderate deviations. This is the right invariant prior as in example 1.

With $\sigma^2$ as the parameter of interest and with the special data choice we have the profile curve $\mathcal{C}_{\sigma^2} = \{(\hat{\mu}_{\sigma^2}, \sigma^2)\} = \{(0, \sigma^2)\}$. First we have Jeffreys's prior along the profile curve:

$$|j_{\varphi\varphi}(\hat{\theta}_{\sigma^2})|^{1/2}\,\mathrm{d}\varphi_1\,\mathrm{d}\varphi_2 = \frac{n\sigma^3}{\sqrt{2}}\,\mathrm{d}\varphi_1\,\mathrm{d}\varphi_2 = \frac{n}{\sigma^3\sqrt{2}}\,\mathrm{d}\mu\,\mathrm{d}\sigma^2.$$

Next we calculate the $\varphi$-based nuisance information, which is the second derivative in the $\varphi$-parameterization for fixed $\sigma^2$:

$$j_{(\mu\mu)}(\theta) = j_{\mu\mu}(\theta)\left|\frac{\partial\varphi}{\partial\mu}\right|^{-2} = \frac{n}{\sigma^2}\left|\left(\frac{1}{\sigma^2}, 0\right)'\right|^{-2} = n\sigma^2.$$

The curvature adjustment is 1 because the $\sigma^2$-fixed contour is linear in $\varphi$, and then the marginalization adjustment is $(n\sigma^2)^{1/2}$. Finally $\mathrm{d}\mu$ on the profile curve extends off the profile curve as $\mathrm{d}\mu$ by invariance. This gives the $\sigma^2$-targeted prior as

$$\pi_{\sigma^2}(\theta)\,\mathrm{d}\theta = \frac{n}{\sigma^3\sqrt{2}} \times 1 \times \sigma\sqrt{n}\,\mathrm{d}\mu\,\mathrm{d}\sigma^2 = \frac{n^{3/2}}{\sqrt{2}}\,\frac{\mathrm{d}\mu\,\mathrm{d}\sigma^2}{\sigma^2};$$

again this is the right invariant prior as in example 1.

### 4.2. Example 8: normal circle (continued)

We saw in example 2 that the default prior for the canonical parameter $(\mu_1, \mu_2) = (\rho\cos(\alpha), \rho\sin(\alpha))$ did not correctly target the component $\rho = (\mu_1^2 + \mu_2^2)^{1/2}$. Now consider the present targeted prior (28) with $\psi = \rho$. In terms of polar co-ordinates for the parameter we have that the profile curve for $\rho$ is $\mathcal{C}_\rho = \{(\rho, \hat{\alpha}^0)\}$. First we have the Jeffreys prior on the profile curve,

$$|j_{\varphi\varphi}(\rho, \hat{\alpha}^0)|^{1/2}\,\mathrm{d}(\alpha)\,\mathrm{d}\rho = 1 \times \rho\,\mathrm{d}\alpha\,\mathrm{d}\rho = \rho\,\mathrm{d}\alpha\,\mathrm{d}\rho$$

where $\mathrm{d}(\alpha) = \rho\,\mathrm{d}\alpha$ gives the $\varphi$-standardized measure for $\alpha$ at the profile $\mathcal{C}_\rho$. Next we calculate the $\varphi$-standardized nuisance information

$$j_{(\alpha\alpha)}(\theta) = j_{\alpha\alpha}(\theta)\left|\frac{\partial\varphi}{\partial\alpha}\right|^{-2} = \rho r|(-\rho\sin(\alpha), \rho\cos(\alpha))'|^{-2} = \frac{r}{\rho}.$$

The corresponding linear parameterization $(\lambda_\psi)$ has information 1 as derived from the standard normal error distribution; this gives the curvature adjustment $(r/\rho)^{1/2}$; and then finally there is the root nuisance information adjustment. We can then assemble the pieces for expression (28) and obtain

$$\pi_\rho(\theta)\,\mathrm{d}\theta = \rho\left(\frac{r}{\rho}\right)^{1/2}\left(\frac{r}{\rho}\right)^{1/2}\mathrm{d}\alpha\,\mathrm{d}\rho = c\,\mathrm{d}\alpha\,\mathrm{d}\rho$$

which is a flat prior in $\alpha$ and $\rho$. This agrees with several derivations of default priors, including Fraser and Reid (2002), who obtained default priors on the constrained maximum likelihood

surface, and with Datta and Ghosh (1995), who obtained this as a reference prior while noting that it was in the family of matching priors that was derived in Tibshirani (1989).

Towards another way of explaining expression (28) from a somewhat different Welch–Peers viewpoint, suppose that the full likelihood is first integrated with respect to the Jeffreys prior for the nuisance parameter $\lambda$,

$$|j_{(\lambda\lambda)}(\psi, \lambda_\psi)|^{1/2}\, \mathrm{d}(\lambda_\psi);$$

this uses the $\varphi$-based reparameterization $(\lambda_\psi)$ for fixed $\psi$. This integration on the parameter space has a Welch and Peers (1963) equivalent on the sample space that uses the corresponding score variable $s_\psi^\perp$ at $y^0$ with differential

$$|j_{(\lambda\lambda)}(\psi, \hat{\lambda}_\psi)|^{-1/2}\, \mathrm{d}s_\psi^\perp.$$

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

$$|j_{(\lambda\lambda)}(\psi, \hat{\lambda}_\psi)|^{1/2}\, |j_{(\lambda\lambda)}(\psi, \lambda_\psi)|^{1/2}\, \mathrm{d}(\lambda_\psi); \tag{29}$$

the additional factor is in fact the marginal likelihood adjustment to the $\psi$-profile as developed differently in Fraser (2003).

The adjusted nuisance Jeffreys prior (29) leads to marginal likelihood for $\psi$, which then appears as an appropriately adjusted profile likelihood for that parameter of interest. This can then be integrated following the Welch–Peers pattern by using root profile information obtained from the exponential parameterization. This in turn gives the Jeffreys-type adjustment

$$|j^{(\psi\psi)}(\hat{\theta}_\psi)|^{-1/2}\, \mathrm{d}(\psi)$$

for the profile concerning $\psi$. The combined targeted prior for $\psi$ is then

$$\pi_\psi(\theta) = |j^{(\psi\psi)}(\hat{\theta}_\psi)|^{-1/2} |j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2} |j_{(\lambda\lambda)}(\theta)|^{1/2}\, \mathrm{d}(\psi)\, \mathrm{d}(\lambda_\psi), \tag{30}$$

for use with the full likelihood $L(\psi, \lambda)$ where $\mathrm{d}(\psi)$ is calculated on the profile from $\psi$-values on the profile and $\mathrm{d}(\lambda_\psi)$ is calculated conditionally for given $\psi$.

The rescaled Jeffreys integration for $\lambda$ on the parameter space produces marginal probability concerning $\psi$ with support $\mathrm{d}s_\psi$. For different $\psi$-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).

## 5. Vector component parameters

The information approach that was outlined in the preceding section requires the nuisance parameter to be scalar, so that the Welch–Peers analysis can be used to extend the default prior beyond the profile contour. In this section we return to the continuity approach to extend the 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 $\theta' = (\psi', \lambda')$.

We work with the parameter effects matrix $W(\theta)$ at expression (7) but the arguments are the same using information-adjusted matrix $\tilde{W}(\theta)$ defined at expression (8). The matrix $W(\theta)$ can be partitioned in accord with the components $\psi$ and $\lambda$ giving $W(\theta) = \{W_\psi(\theta), W_\lambda(\theta)\}$ so that

$$\mathrm{d}\hat{\theta} = \{W_\psi(\theta), W_\lambda(\theta)\} \begin{pmatrix} \mathrm{d}\psi \\ \mathrm{d}\lambda \end{pmatrix} = W_\psi(\theta)\, \mathrm{d}\psi + W_\lambda(\theta)\, \mathrm{d}\lambda.$$

To target the parameter on $\psi$, we separate the effects of $\psi$ and $\lambda$ by orthogonalization in the information-standardized co-ordinates, as at equation (30), and construct the targeted prior as

$$\pi_\psi(\theta)\,\mathrm{d}\theta \propto |W_{\psi\cdot\lambda}(\hat\theta_\psi)|\,\mathrm{d}\psi\,|W_\lambda(\theta)|\,\mathrm{d}\lambda,$$

where

$$W_{\psi\cdot\lambda}(\hat\theta_\psi) = W_\psi - W_\lambda(W_\lambda'W_\lambda)^{-1}W_\lambda'W_\psi$$

records the residual vectors for $W_\psi(\hat\theta_\psi)$ orthogonalized to $W_\lambda(\hat\theta_\psi)$. This is similar to expression (27), but without the middle factor, which is an adjustment for curvature. For a parameter value $\hat\theta_\psi = (\psi, \hat\lambda_\psi)$ on the profile curve $\mathcal{C}_\psi$ formed by the constrained maximum likelihood value, a change $\mathrm{d}\lambda$ in $\lambda$ with $\psi$ fixed generates a $(p-d)$-dimensional tangent plane $\mathcal{T}_\psi = \mathcal{L}\{W_\lambda(\hat\theta_\psi)\}$ at the observed $\hat\theta^0$ on the data space for $\hat\theta$. The term $W_{\psi\cdot\lambda}\,\mathrm{d}\psi$ thus presents the effect of a change $\mathrm{d}\psi$ in $\psi$, perpendicular to $\mathcal{T}_\psi$. When $\psi$ is a vector we would, however, expect this prior for $\psi$ to give second-order calibration only for linear parameter components of $\psi$ in accord with Dawid *et al.* (1973); otherwise we would have curvature effects in these components to take account of, as in the normal circle example; these will be examined elsewhere.

### 5.1. Example 9: linear regression (continued from example 1)

Suppose that $r = 3$, the parameter of interest is $\psi = (\beta_1, \beta_2)'$ and the nuisance parameter is $\lambda = (\beta_3, \sigma^2)$. To simplify the expressions we assume that the regression variables have been standardized so that $X'X/n = I$. We have

$$W(\theta) = \begin{pmatrix} 1 & 0 & 0 & (\hat\beta_1^0 - \beta_1)/2\sigma^2 \\ 0 & 1 & 0 & (\hat\beta_2^0 - \beta_2)/2\sigma^2 \\ 0 & 0 & 1 & (\hat\beta_3^0 - \beta_3)/2\sigma^2 \\ 0 & 0 & 0 & \hat\sigma^2/\sigma^2 \end{pmatrix},$$

giving

$$W_\psi(\hat\theta_\psi) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad W_\lambda(\theta) = \begin{pmatrix} 0 & t_1(\beta_1)\hat\sigma/2\sigma^2\sqrt{n} \\ 0 & t_2(\beta_2)\hat\sigma/2\sigma^2\sqrt{n} \\ 1 & t_3(\beta_3)\hat\sigma/2\sigma^2\sqrt{n} \\ 0 & \hat\sigma^2/\sigma^2 \end{pmatrix},$$

where $t_j(\beta_j) = (\hat\beta_j^0 - \beta_j)\sqrt{n}/\hat\sigma$ is the usual $t$-statistic. As $\hat\sigma_\psi = \hat\sigma$ to $O(n^{-1})$, we have $|W_{\psi\cdot\lambda}(\hat\theta_\psi)| = 1 + O(n^{-1})$ and $|W_\lambda(\theta)| = \sqrt{n}/(2\hat\sigma\sigma^2)\{1 + O(n^{-1})\}$, giving the prior $\pi(\theta) \propto \mathrm{d}\beta\,\mathrm{d}\sigma^2/2\sigma^2$, as would be expected from example 1. The same result is obtained by using the information-standardized version (9); the observed Fisher information matrix is $\mathrm{diag}(nI/\hat\sigma^2, n/2\hat\sigma^4)$ which gives

$$\tilde W_\psi(\hat\theta_\psi) = \sqrt{n}\begin{pmatrix} 1/\hat\sigma & 0 \\ 0 & 1/\hat\sigma \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \tilde W_\lambda(\theta) = \sqrt{n}\begin{pmatrix} 0 & t_1(\beta_1)/2\sigma^2\hat\sigma\sqrt{n} \\ 0 & t_2(\beta_2)/2\sigma^2\hat\sigma\sqrt{n} \\ 1/\hat\sigma & t_3(\beta_3)/2\sigma^2\hat\sigma\sqrt{n} \\ 0 & 1/\hat\sigma\sigma^2\sqrt{2} \end{pmatrix},$$

and thus, to order $O(n^{-1})$, $\pi(\theta) \propto \mathrm{d}\beta\,\mathrm{d}\sigma^2/\sigma^2$ in moderate deviation about the observed maximum likelihood value.

## 6. Discussion

We have described two approaches to default priors: one based on extending a location approximation, and one based on matching higher order approximations. There is a natural progression in complexity, following the model type.

For a scalar parameter model that is not location, asymptotic arguments lead to Jeffreys's prior $\pi_J(\theta)\,d\theta \propto i^{1/2}(\theta)\,d\theta$, which agrees with the Welch–Peers approach and gives matching probabilities to second order. The refined version (23) gives matching to third order, and is data dependent and incorporates conditioning on the approximate ancillary statistic.

The default prior that is based on the sensitivity matrix, which was derived in Section 2, extends this local location property to vector parameters. Underlying the construction of expression (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 by using Taylor series expansions, and the location parameter is given by the expression for $\beta$ following expression (23). 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 independent scalar coordinates $y_i$ does give a location model approximation, 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 expression (7) provides a rescaling on the parameter space so that in the new parameterization each parameter value has the same influence at the data point. However, in the case of non-linear parameters, as discussed in Section 4, more is needed to target the parameter of interest properly.

We have not discussed whether or not posteriors based on $W(\theta)$ are proper. The development is local to the data point, and several approximations are made that assume that $\theta$ is within $O(n^{-1/2})$ of the maximum likelihood estimate $\hat{\theta}$. This suggests that posteriors would need to be checked case by case. 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

$$f(y;\theta) = \frac{\beta(y-\psi)^{\beta-1}}{\eta^{\beta}} \exp\left\{-\left(\frac{y-\psi}{\eta}\right)^{\beta}\right\}, \qquad y > \psi. \tag{31}$$

This model has a discontinuity related to the end point parameter, so the derivations here are just suggestive. 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 compute $V(\theta)$ formally by using equation (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) proposed a combination reference or right Haar prior for this model which is proportional to $1/\eta\beta$ and noted that the posterior is improper unless the range of $\psi$ is restricted. The factor $h(y^0, \psi)$ seems to enforce a restriction on the range of $\psi$, 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 is not possible to address non-linear parameters of interest unless the nuisance parameter is a scalar. Other approaches to deriving targeted priors for the full parameter space have analogous difficulties. The Welch–Peers approach leads to a family of priors $\pi(\theta)\,\mathrm{d}\theta \propto i_{\psi\psi}^{1/2}(\theta)\, g(\lambda)$ 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 targeted 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 case by case.

## Acknowledgements

We are grateful to the Joint Editor and referees for very careful and constructive comments. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada. We are indebted to Kexin Ji for her assistance with the manuscript and with Fig. 2.

## Appendix A

### A.1. Example 3: background on transformation models

The parameter $\theta$ 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 with density $g(z)$. 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 $\tilde{f}(y;\theta)\,\mathrm{d}y = \tilde{g}(\theta^{-1}y)\,\mathrm{d}\mu(y)$ where $\mathrm{d}\mu(\cdot)$ is the left invariant measure.

The notation is simplified if the group co-ordinates are centred so that the identity element is at the maximum density point of the conditional error density; thus $\tilde{g}(z) \leqslant \tilde{g}(e)$ 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 expression (7) that the default prior is

$$|W(\theta)|\,\mathrm{d}\theta = \left|\frac{\mathrm{d}y}{\mathrm{d}\theta}\right|_{y^0}\mathrm{d}\theta = \left|\frac{\mathrm{d}(\theta z)}{\mathrm{d}\theta}\right|_{\theta z = y^0}\mathrm{d}\theta \tag{32}$$

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 by using notation from Fraser (1979), page 144; let $J^*(h;g) = |\partial gh/\partial h|$ with variable $g$ and also $J^*(g) = J^*(g;e)$; this gives $\mathrm{d}\nu(g) = \mathrm{d}g/J^*(g)$ where $\mathrm{d}\nu(g)$ is the right invariant measure. We then have $\mathrm{d}\theta z = J^*(\theta z)\,\mathrm{d}\nu(\theta z)$ with $\theta$ as the variable. Then with $\theta z$ set equal to $y^0$ we obtain

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

using the right invariance of $\nu$; this is a constant times the standardized right invariant measure $\mathrm{d}\nu(\theta)$ on the group. We thus have that the default prior (32) is $\pi(\theta)\,\mathrm{d}\theta = c\,\mathrm{d}\nu(\theta)$.

### A.2. Section 4: rescaling the parameterization of the approximating exponential model

The exponential model approximation (19) to a general model depends on $\theta$ only through the observed log-likelihood function $l(\theta)$ and the observed log-likelihood gradient function $\varphi(\theta)$. The $r_{\mathrm{f}}^*$-approximation (20) is computed entirely within this model, with log-likelihood function

$$l(\theta;s) = l(\theta) + \varphi'(\theta)s \tag{33}$$

and observed data $s^0 = 0$.

For scalar $\theta$ and $\varphi$ we have $l_{(\theta)}(\theta) = l_{\varphi}(\theta) = l_{\theta}(\theta)\, \varphi_{\theta}^{-1}(\theta)$ where the subscripts as usual denote differentiation. Then differentiating again we obtain

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

An analogous formula is available for the vector case by 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 $\psi$ fixed. This model can have curved ancillary contours on the initial score space $\{s\}$ if for example $\psi$ is not linear in $\varphi(\theta)$. Correspondingly the differentiation with respect to $(\lambda)$ requires the use of the $\varphi$-metric for $\lambda$ given $\psi$ and the results indicate the use of the standardization $j^{0}_{\varphi\varphi} = I$. From the preceding scalar derivative expression we obtain

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

where as usual $|\varphi_{\lambda}|^2 = |\varphi'_{\lambda}\varphi_{\lambda}|$.

Now consider the nuisance information $j_{(\lambda_\chi\lambda_\chi)}(\hat{\varphi}_{\psi})$ calculated at a point $\hat{\varphi}_{\psi}$ on the profile curve $\mathcal{C}_{\psi}$ from observed data: see Fig. 2. The observed log-likelihood function has a maximum at $\hat{\varphi}_{\psi}$ when examined along the contour with $\psi$ fixed or when examined along the tangent contours with $\chi$ fixed. But the negative Hessian with respect to $\lambda$ or $(\lambda)$ will typically differ on two contours unless $\hat{\varphi}_{\psi} = \hat{\varphi}$, i.e. unless $\psi$ is at its maximum likelihood value $\hat{\psi}$. We seek an expression for the curvature $j_{(\lambda_\chi\lambda_\chi)}(\hat{\varphi}_{\psi})$ along the contour of the linear parameter $\chi$. The tangent plane to the likelihood at $\hat{\varphi}_{\psi}$ is $l_{\varphi}(\hat{\varphi}_{\psi})(\varphi - \hat{\varphi}_{\psi}) = 0$; and the tilted likelihood

$$\tilde{l}(\varphi) = l(\varphi) - l_{\varphi}(\hat{\varphi}_{\psi})(\varphi - \hat{\varphi}_{\psi})$$

has maximum at $\hat{\varphi}_{\psi}$; and accordingly the negative Hessians along $\psi$- or $\chi$-contours are connected as

$$j_{(\lambda_\chi\lambda_\chi)}(\hat{\varphi}_{\psi}) = j_{(\lambda\lambda)}(\hat{\varphi}_{\psi}) - l_{\varphi}(\hat{\varphi}_{\psi})\, \varphi_{\lambda\lambda}(\hat{\varphi}_{\psi}).$$

### *A.3. 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 that is obtained by marginalizing the posterior distribution for the full parameter is consistent with that obtained by using a prior distribution on the parameter of interest and applying it to the likelihood function from the marginal model. The normal circle problem of example 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 $\psi$. 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 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 $\mathrm{d}\lambda$ in the nuisance parameter $\lambda$ for fixed $\psi = \psi_0$ generates through expression (6) a direction at the data point that is confined to a subspace free of $\lambda$ 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 $\theta_2$ applied to $y_1 = \theta_1 + z_1$ and $y_2 = \theta_2 + z_2$ gives the $y_2$-direction which corresponds to fixed $y_1$. For the normal circle example we note that the radius $\psi$ is curved but the angle $\alpha$ is linear.

The linearity condition defines a location relationship between the nuisance parameter $\lambda$ for fixed $\psi$ 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 expression (6) we have

$$\begin{aligned} \mathrm{d}\hat{\theta}_1 &= w_{11}(\theta)\, \mathrm{d}\theta_1 + w_{12}(\theta)\, \mathrm{d}\theta_2, \\ \mathrm{d}\hat{\theta}_2 &= w_{21}(\theta)\, \mathrm{d}\theta_1 + w_{22}(\theta)\, \mathrm{d}\theta_2, \end{aligned} \tag{34}$$

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

First we examine the parameter $\psi(\theta)$ near $\hat{\theta}^0$ on the parameter space and find that an increment $(\mathrm{d}\theta_1, \mathrm{d}\theta_2)$ with no effect on $\psi(\theta)$ must satisfy $\mathrm{d}\psi(\theta) = \hat{\psi}_1^0\, \mathrm{d}\theta_1 + \hat{\psi}_2^0\, \mathrm{d}\theta_2 = 0$ where $\psi_i(\theta) = \partial\psi(\theta)/\partial\theta_i$, i.e. $\mathrm{d}\theta_1 = -(\hat{\psi}_2^0/\hat{\psi}_1^0)\, \mathrm{d}\theta_2$. Next we use expression (34) to determine the corresponding sample space increment

at $\hat{\theta}^0$, and obtain

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

thus $(c_1, c_2)$ so defined gives a direction $(c_1, c_2)'\,\mathrm{d}t$ on the sample space that corresponds to no $\psi$-change. Finally we use the inverse of expression (34) to determine the parameter space increment at a general point $\theta$ that corresponds to the preceding sample space increment, giving

$$\mathrm{d}\theta = \begin{pmatrix} w^{11}(\theta)c_1 + w^{12}(\theta)c_2 \\ w^{21}(\theta)c_1 + w^{22}(\theta)c_2 \end{pmatrix} \mathrm{d}t, \tag{35}$$

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

$$\theta = \hat{\theta}^0 + \int_0^1 \begin{pmatrix} w^{11}\{\hat{\theta}^0 + (c_1, c_2)'t\}c_1 + w^{12}\{\hat{\theta}^0 + (c_1, c_2)'t\}c_2 \\ w^{21}\{\hat{\theta}^0 + (c_1, c_2)'t\}c_1 + w^{22}\{\hat{\theta}^0 + (c_1, c_2)'t\}c_2 \end{pmatrix} \mathrm{d}t,$$

which describes the radial solution of the differential equation (35), or an implicit equation $\theta_2 = \theta_2(\theta_1)$ as a direct solution of the differential equation

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

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

### A.4. Example 10: linearity with the regression model

We reconsider the regression example 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

$$\begin{aligned} \mathrm{d}\hat{\mu} &= \mathrm{d}\mu + (\hat{\mu}^0 - \mu)\,\mathrm{d}\sigma/\sigma, \\ \mathrm{d}\hat{\sigma} &= \hat{\sigma}^0\,\mathrm{d}\sigma/\sigma. \end{aligned} \tag{36}$$

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

$$\begin{aligned} \mathrm{d}\hat{\mu} &= \mathrm{d}\mu - \mu\,\mathrm{d}\sigma/\sigma, \\ \mathrm{d}\hat{\sigma} &= \mathrm{d}\sigma/\sigma. \end{aligned} \tag{37}$$

Inverting this gives

$$\begin{aligned} \mathrm{d}\mu &= \mathrm{d}\hat{\mu} + \mu\,\mathrm{d}\hat{\sigma}, \\ \mathrm{d}\sigma &= \sigma\,\mathrm{d}\hat{\sigma}. \end{aligned} \tag{38}$$

First we examine $\mu + k\sigma$ in the neighbourhood of $\hat{\theta}^0$ on the parameter space and have that an increment $(\mathrm{d}\mu, \mathrm{d}\sigma)$ must satisfy $\mathrm{d}(\mu + k\sigma) = 0$ at $\hat{\theta}^0 = (\hat{\mu}^0, \hat{\sigma}^0) = (0, 1)$; this gives $\mathrm{d}\mu = -k\,\mathrm{d}\sigma$ at $\hat{\theta}^0$. Next we determine the corresponding increment at $\hat{\theta}^0$ on the sample space $\{(\hat{\mu}, \hat{\sigma})\}$; from expression (37) we have $\mathrm{d}\hat{\mu} = \mathrm{d}\mu$ and $\mathrm{d}\hat{\sigma} = \mathrm{d}\sigma$ at this point, which gives $\mathrm{d}\hat{\mu} = -k\,\mathrm{d}\hat{\sigma}$. Finally we determine what the restriction $\mathrm{d}\hat{\mu} = -k\,\mathrm{d}\hat{\sigma}$ on the sample space implies for $(\mathrm{d}\mu, \mathrm{d}\sigma)$ at a general point on the parameter space; from expression (38) this is

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

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 second order linear.

### A.5. Example 11: linearity for the normal case on the plane

For the normal circle example 2 with parameter of interest $\psi = (\theta_1^2 + \theta_2^2)^{1/2}$, the increment on the parame-

ter space at $\hat{\theta}^0$ with fixed $\psi$ satisfies $\mathrm{d}\theta_1 = -\tan(\hat{\alpha}^0)\,\mathrm{d}\theta_2 = -(y_2^0/y_1^0)\,\mathrm{d}\theta_2$. This then translates to the sample space at $(y_1^0, y_2^0)$ by using the specialized version of expression (34) to give $\mathrm{d}y_2 = -(y_2^0/y_1^0)\,\mathrm{d}y_1$, and this then translates back to a general point on the parameter space by using the specialized version of expression (35) to give a line through $\hat{\theta}^0$ described by $\mathrm{d}\theta_2 = -(y_2^0/y_1^0)\,\mathrm{d}\theta_1$, which is perpendicular to the radius and thus tangent to the circle through the data point; this is the linear parameter equivalent to $\psi$ near $\hat{\theta}_0$ and is fully linear in the location parameter $\theta$.

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 (Fraser and Sun, 2010).

### *A.6. Strong matching and information approximation*

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

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

where $\mathrm{d}\beta(\theta)$ is a locally defined linear parameter (Fraser and Reid, 2002).

If the model is a full exponential family with log-likelihood function $l(\theta) = \theta t - k(\theta)$ then the location reparameterization satisfies to second order

$$\begin{aligned}\frac{\mathrm{d}\beta(\theta)}{\mathrm{d}\theta} &= -\frac{t^0 - k'(\theta)}{\theta - \hat{\theta}^0} = \frac{k'(\theta) - k'(\hat{\theta}^0)}{\theta - \hat{\theta}^0} \\ &= \frac{(\theta - \hat{\theta}^0)\,k''(\hat{\theta}^0) + \frac{1}{2}(\theta - \hat{\theta}^0)^2\,k'''(\hat{\theta}^0)}{\theta - \hat{\theta}^0} \\ &= k''(\hat{\theta}^0)\{1 + \tfrac{1}{2}(\theta - \hat{\theta}^0)\,k'''(\hat{\theta}^0)/k''(\hat{\theta}^0)\};\end{aligned}$$

this agrees to the same order with the usual Jeffreys prior

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

The same argument can then be applied to the approximating exponential model (19) as used along the profile curve and leads to the approximation that is used at expression (25).

## References


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

Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Phil. Trans. R. Soc. Lond.*, **53**, 370–418; **54**, 296–325.

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

Berger, J. (2006) The case for objective Bayesian analysis. *Baysn Anal.*, **1**, 385–402.

Berger, J. and Bernardo, J. (1992) On the development of the reference prior method. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 35–60. Oxford: Clarendon.

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

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

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.

Box, G. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*. Reading: Addison-Wesley.

Brazzale, A. R., Davison, A. and Reid, N. (2007) *Applied Asymptotics*. Cambridge: Cambridge University Press.

Cakmak, S., Fraser, D. and Reid, N. (1994) Multivariate asymptotic model: exponential and location approximations. *Util. Math.*, **46**, 21–31.

Casella, G., DiCiccio, T. and Wells, M. (1995) Inference based on estimating functions in the presence of nuisance parameters: comment, Alternative aspects of conditional inference. *Statist. Sci.*, **10**, 179–185.

Clarke, B. (2007) Information optimality and Bayesian modelling. *J. Econmetr.*, **38**, 405–429.

Clarke, B. and Barron, A. (1994) Jeffreys prior is asymptotically least favorable under entropy risk. *J. Statist. Planng Inf.*, **41**, 37–60.
Clarke, B. and Yuan, A. (2004) Partial information reference priors: derivation and interpretations. *J. Statist. Planng Inf.*, **123**, 313–345.
Cox, D. and Hinkley, D. (1974) *Theoretical Statistics*. London: Chapman and Hall.
Datta, G. and Ghosh, M. (1995) Some remarks on noninformative priors. *J. Am. Stastist. Ass.*, **90**, 1357–1363.
Datta, G. and Mukerjee, R. (2004) *Probability Matching Priors: Higher Order Asymptotics*. New York: Springer.
Davison, A. C., Fraser, D. A. S. and Reid, N. (2006) Improved likelihood inference for discrete data. *J. R. Statist. Soc.* B, **68**, 495–508.
Dawid, A. P., Stone, M. and Zidek, J. (1973) Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. R. Statist. Soc.* B, **35**, 189–233.
DiCiccio, T. and Martin, M. (1991) Approximations of marginal tail probabilities for a class of smooth functions with applications to Bayesian and conditional inference. *Biometrika*, **78**, 891–902.
Eaves, D. (1983) On Bayesian nonlinear regression with an enzyme example. *Biometrika*, **70**, 373–379.
Fraser, D. A. S. (1964) Local conditional sufficiency. *J. R. Statist. Soc.* B, **26**, 46–51.
Fraser, D. A. S. (1979) *Inference and Linear Models*. New York: McGraw-Hill.
Fraser, D. A. S. (2003) Likelihood for component parameters. *Biometrika*, **90**, 327–339.
Fraser, D. A. S. and Reid, N. (1993) Third order asymptotic models: likelihood functions leading to accurate approximations to distribution functions. *Statist. Sin.*, **3**, 67–82.
Fraser, D. A. S. and Reid, N. (2002) Strong matching of frequentist and Bayesian inference. *J. Statist. Planng Inf.*, **103**, 263–285.
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.
Fraser, D. A. S. and Sun, Y. (2010) Some corrections for Bayes curvature. *Pak. J. Statist.*, to be published.
George, E. and McCulloch, R. (1993) On obtaining invariant prior distributions. *J. Statist. Planng Inf.*, **37**, 169–179.
Ghosh, J., Chakrabarti, A. and Samanta, T. (2009) Entropies and metrics that lead to Jeffreys and reference priors. *Technical Report*. Indian Statistical Institute, Calcutta. (Available from `http://www-stat.wharton.upenn.edu/statweb/Conference/OBayes09/lectures.htm`.)
Goldstein, M. (2006) Subjective Bayesian analysis: principles and practice. *Baysn Anal.*, **1**, 403–420.
Jeffreys, H. (1961) *Theory of Probability*, 3rd edn. Oxford: Oxford University Press.
Kass, R. (1990) Data-translated likelihood and Jeffreys's rules. *Biometrika*, **77**, 107–114.
Kass, R. and Wasserman, L. (1996) Formal rules for selecting prior distributions: a review and annotated bibliography. *J. Am. Statist. Ass.*, **91**, 1343–1370.
Lin, X., Sun, D. and Berger, J. (2009) Objective Bayes analysis under semi-invariance structure. University of South Carolina, Columbia. (Available from `http://www-stat.wharton.upenn.edu/statweb/Conference/OBayes09/lectures.htm`.)
Liseo, B. (1993) Elimination of nuisance parameters with reference priors. *Biometrika*, **80**, 295–304.
Little, R. (2006) Calibrated Bayes: a Bayes/frequentist roadmap. *Am. Statistn*, **60**, 1–11.
Peers, H. W. (1965) On confidence points and Bayesian probability points in the case of several parameters. *J. R. Statist. Soc.* B, **27**, 9–16.
Pierce, D. and Peters, D. (1994) Higher-order asymptotics and the likelihood principle: one-parameter models. *Biometrika*, **81**, 1–10.
Reid, N. (2003) Asymptotics and the theory of inference. *Ann. Statist.*, **31**, 1695–1731.
Reid, N. and Fraser, D. (2010) Mean log-likelihood and higher order approximation. *Biometrika*, **97**, 159–170.
Stein, C. (1959) An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.*, **30**, 877–880.
Stigler, S. M. (1982) Thomas Bayes's Bayesian inference. *J. R. Statist. Soc.* A, **145**, 250–258.
Sun, D. and Ye, K. (1996) Frequentist validity of posterior quantiles for a two-parameter exponential family. *Biometrika*, **83**, 55–65.
Tibshirani, R. (1989) Noninformative priors for one parameter of many. *Biometrika*, **76**, 705–708.
Tierney, L. and Kadane, J. (1986) Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Ass.*, **81**, 82–87.
Wasserman, L. (2000) Asymptotic inference for mixture models using data-dependent priors. *J. R. Statist. Soc.* B, **62**, 159–180.
Welch, B. L. and Peers, H. W. (1963) On formulae for confidence points based in intervals of weighted likelihoods. *J. R. Statist. Soc.* B, **25**, 318–329.
Yang, R. and Berger, J. (1996) A catalog of noninformative priors. *Discussion Paper 97-42*. Institute of Statistics and Decision Sciences, Duke University, Durham. (Available from `citeseer.ist.psu.edu/old/401050.html`.)
Zellner, A. (1988) Optimal information processing and Bayes' theorem. *Am. Statistn*, **42**, 278–284.