

Default priors
for Bayesian and frequentist inference

D.A.S. Fraser and N. Reid*
University of Toronto, Canada

E. Marras
Centre for Advanced Studies and Development, Sardinia
University of Rome “La Sapienza”, Rome

G.Y. Yi
University of Waterloo, Canada

June 2, 2008

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

Abstract

We investigate the choice of default prior for use with likelihood to facilitate Bayesian and frequentist inference. Such a prior is a density or relative density that weights an observed likelihood function leading to the elimination of parameters not of interest and accordingly providing a density type assessment for a parameter of interest. For regular models with independent coordinates we develop a second-order prior for the full parameter based on an approximate location relation from near a parameter value to near the observed data point; this derives directly from the coordinate distribution functions and is closely linked to the original Bayes approach. We then develop a modified prior that is targetted on a component parameter of interest and avoids the marginalization paradoxes of Dawid, Stone and Zidek (1973); this uses some extensions of Welch-Peers theory that modify the Jeffreys prior and builds more generally on the approximate location property. A third type of prior is then developed that targets a vector interest parameter in the presence of a vector nuisance parameter and is based more directly on the original Jeffreys approach. Examples are given to clarify the computation of the priors and the flexibility of the approach.

Key words: Default prior; Interest parameter; Jeffreys prior; Nuisance parameter; Objective prior; Subjective prior.

1 Introduction

We develop default priors for Bayesian and frequentist inference in the context of a statistical model $f(y; \theta)$ and observed data y^0 . A default prior is a density or relative density used as a weight function to be applied to the observed likelihood function. The choice of prior is based on characteristics of the model together with an absence of information as to how the parameter value was generated. For this we assume the response y has dimension n and the parameter θ dimension p and restrict attention to regular models with continuous responses.

One Bayesian role for a default prior is to provide a reference allowing subse-

quent modification by an objective, subjective, personal or expedient prior. From a frequentist viewpoint a default prior can be viewed as a device to directly use Fisher's (1922) likelihood function. From either view it offers a flexible and easily implemented exploratory approach to statistical inference.

We focus on a parameter to data location-type relationship that is implicit in the coordinate distribution functions. The emphasis is on continuity and is closely related to the original approach of Bayes and Laplace.

As a simple example suppose the model is location $f(y-\theta)$. Translation and invariance properties (Bayes, 1763) suggest the flat prior $\pi(\theta) = c$ for the parameter θ , where c is a constant; it follows immediately that the frequentist p -value $p(\theta) = \int^{y^0} f(y-\theta)dy$ is equal to the Bayesian survivor value $s(\theta) = \int_{\theta} f(y^0-\alpha)\pi(\alpha)d\alpha$. A similar argument is available with the location model $f\{y-\beta(\theta)\}$ using the flat prior $\pi(\theta)d\theta = c\beta'(\theta)d\theta$ relative to the location parameterization $\beta(\theta)$. Recent likelihood and large sample results show that such translation structure is widely available as an approximation.

In Section 2 we give some background on the approximate translation property. In Sections 3 and 4 we develop corresponding improved priors for default Bayesian and frequentist inference. In particular in Section 3 we use directly the continuity implicit in many models and obtain a location based default prior (8) that is appropriate for parameters that have an intrinsic linearity. Linearity of parameters is discussed in Section 4. In Section 5 we discuss the exponential model approximation that provides a basis for recent third order inference concerning scalar parameters; crucial to this is a canonical exponential-type reparameterization $\varphi(\theta)$ that allows second order comparison of differential change in θ in moderate deviations; this relates closely to the approximate location relationship initiated by Welch & Peers (1963). In Section 6 we develop targetted priors for a parameter $\psi(\theta)$ of primary interest, in the presence of a nuisance parameter $\lambda(\theta)$, but restrict attention to scalar component parameters. An alternative view is presented in Section 7 and then priors for components that can be vector valued are examined in Section 8.

The term objective prior is frequently used to describe what we call a default prior. In our view the term objective prior is appropriate to contexts where it is known that θ arose from some density $g(\theta)$. A different approach to developing default priors is based on information processing, as in Zellner (1988) and Bernardo (1979). Some further comments on various choices of default prior are provided in the Discussion.

2 Continuity: How the parameter affects the data variable

With minimum continuity in the model, a small change in the parameter θ causes a change in the distribution of any particular coordinate; this is an intrinsic property of the model. This in turn causes a change in the full distribution on the composite sample space, and thereby provides an approximate location relationship from a general parameter value to a neighbourhood of the observed data point. This property gives a default prior, as described in Section 3 below.

First consider a scalar coordinate with a location model $f(y - \theta)$. The model is invariant under location change: if θ is changed to $\theta + d\theta$, the distribution of y is shifted by the amount $d\theta$, and the quantile at an observed y^0 shifts to $y^0 + d\theta$. This generalizes to continuous models $f(y; \theta)$. The relation between change in θ and the corresponding change at y^0 can be calculated directly by using the quantile function $y = y(u, \theta)$:

$$dy = \left. \frac{\partial}{\partial \theta'} y(u, \theta) \right|_{y^0} d\theta. \quad (1)$$

We can also obtain the change by taking the total differential of $u = F(y; \theta)$ and solving for dy :

$$dy = -\frac{F_{;\theta}(y^0; \theta)}{F_y(y^0; \theta)} d\theta,$$

where the subscripts denote differentiation with respect to the corresponding argument. If θ is a vector of dimension p we have analogously that the change at a

scalar data point y^0 is

$$V(\theta)d\theta = -\frac{1}{F_y(y^0; \theta)} F_{y^0; \theta'}(y^0; \theta) d\theta \quad (2)$$

where $V(\theta)$ is a $1 \times p$ row vector. This generalizes translation invariance to local translation invariance.

Now consider the case with independent scalar coordinates y_i each with a row vector $V_i(\theta)d\theta$ recording the shift when θ is changed to $\theta + d\theta$. The change in the vector variable y at y^0 is then

$$V(\theta)d\theta = \begin{pmatrix} V_1(\theta) \\ \vdots \\ V_n(\theta) \end{pmatrix} d\theta \quad (3)$$

where $V(\theta)$ is an $n \times p$ matrix that we call the sensitivity of θ relative to the data y^0 . We can also write $V(\theta) = \{v_1(\theta) \cdots v_p(\theta)\}$ where an $n \times 1$ component vector $v_j(\theta)$ gives the data displacement $v_j(\theta)d\theta_j$ when the j th coordinate of θ is changed by $d\theta_j$.

In the next sections $V(\theta)$ is used to develop default priors based directly on approximate translation properties. When evaluated at the maximum likelihood estimate $\hat{\theta}^0$, the array $V = V(\hat{\theta}^0)$ gives tangents to an approximately ancillary statistic (Fraser & Reid, 2001). In Sections 5 and 6 we describe how V enables the construction of an approximate exponential family model, which then leads to a Jeffreys' type default prior $\pi(\theta)$ for inference concerning a scalar parameter of interest in the presence of a scalar nuisance parameter. Then in Section 8 we describe the modification needed when the component parameters can be vector valued.

3 Default priors from continuity

In this section we use the sensitivity matrix $V(\theta)$ at (3) to determine the parameter-to-variable relationship implied by continuity. From this we obtain default priors

(8) that are a general model extension of the familiar right invariant priors common for regression and related models.

From (2) and (3), we have, at y^0 ,

$$dy = V(\theta)d\theta. \quad (4)$$

The corresponding volume $|V(\theta)| = |V'(\theta)V(\theta)|^{1/2}$ gives a second order default prior, but there are advantages to making some refinements. For this we derive the connection between the coordinates of y and the coordinates of the maximum likelihood estimator $\hat{\theta}$ by computing the total differential of the score equation $l_\theta(\theta; y) = 0$ at $(y^0, \hat{\theta}^0)$. This gives

$$d\hat{\theta} = \hat{j}^{-1}Hdy$$

where $H = \ell_{\theta; y'}(\hat{\theta}^0; y^0)$ is the gradient of the score function at the data point and $\hat{j} = j(\hat{\theta}^0; y^0) = -\partial^2 \ell(\hat{\theta}^0; y^0) / \partial \theta \partial \theta'$ is the observed information. We then combine these and obtain

$$d\hat{\theta} = \hat{j}^{-1}HV(\theta)d\theta = W(\theta)d\theta, \quad (5)$$

where $W(\theta) = \hat{j}^{-1}HV(\theta)$. This records how a change $d\hat{\theta}$ at $\hat{\theta}^0$ is related to a corresponding parameter increment $d\theta$ at θ .

There are some notational advantages to working with information standardized departures $\hat{j}^{1/2}d\hat{\theta}$ at the observed data, where $\hat{j}^{1/2}$ is a right square root of the observed information \hat{j} . We then obtain

$$\hat{j}^{1/2}d\hat{\theta} = \hat{j}^{1/2}W(\theta)d\theta = \tilde{W}(\theta)d\theta, \quad (6)$$

where $\tilde{W}(\theta) = (\hat{j}^{-1/2})'HV(\theta)$ provides a $p \times p$ matrix multiple of $d\theta$ that directly corresponds under distribution displacement to an increment at the observed data. In terms of volume this gives

$$|\hat{j}^{1/2}d\hat{\theta}| = |(\hat{j}^{-1/2})'HV(\theta)|d\theta = J_\theta(\theta)d\theta, \quad (7)$$

and the default prior is

$$\pi(\theta)d\theta = |\tilde{W}(\theta)|d\theta = J_\theta(\theta)d\theta; \quad (8)$$

we use the notation $J_\theta(\theta)$, as it is analogous to a root observed information determinant.

Example 3.1. Consider a statistical model with scalar variable, scalar parameter, and asymptotic properties. Welch & Peers (1963) show in effect that

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

has a parameter-free distribution to second order where $i(\theta)$ is expected Fisher information: they however formulate this in terms of the equality of confidence and posterior bounds. We rewrite this as

$$z = \hat{\beta} - \beta$$

where $\beta(\theta) = \int^\theta i^{1/2}(t)dt$ is the constant information reparametrization, and z is $N(0, 1)$ to first order and free of θ to second order. The default prior follows trivially from $\hat{\beta} = \beta + z$ and we obtain $d\hat{\beta} = d\beta$ giving the flat prior $\pi(\theta)d\theta = cd\beta$. This is of course the Jeffreys (1939) prior for this scalar case. For some discussion of transformations to location and of the Jeffreys prior see Kass (1990).

Example 3.2. Normal regression. Consider the normal linear model $y = X\beta + \sigma z$ where $z = (z_1, \dots, z_n)'$ is a sample from the standard normal and $\theta' = (\beta', \sigma^2)$. The quantile functions are $y_1 = X_1\beta + \sigma z_1, \dots, y_n = X_n\beta + \sigma z_n$, and

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

where $z^0(\theta) = z(y^0, \theta) = (y^0 - X\beta)/\sigma$ is the standardized residual corresponding to data y^0 and parameter value θ .

Then from the likelihood gradient $\ell_{;y} = (X\beta - y)/\sigma^2$ we obtain the score gradient $\ell_{\theta';y}(\theta; y) = \{\sigma^{-2}X, \sigma^{-4}(y - X\beta)\}$, and thus have

$$\begin{aligned} H' &= \{X/\hat{\sigma}^2, (y^0 - X\hat{\beta}^0)/\hat{\sigma}^4\} \\ &= \{X/\hat{\sigma}^2, \hat{z}^0/\hat{\sigma}^3\}, \end{aligned} \tag{9}$$

where $\hat{\sigma}^0$ is abbreviated as $\hat{\sigma}$ to simplify notation.

Then combining this with $\hat{j} = \text{diag}(X'X/\hat{\sigma}^2, n/2\hat{\sigma}^4)$ and $V(\theta)$, and using (5) we obtain

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

giving

$$W(\theta) = \begin{pmatrix} I & (\hat{\beta}^0 - \beta)/2\sigma^2 \\ 0 & \hat{\sigma}^2/\sigma^2 \end{pmatrix},$$

and thus producing, via (8), the default prior

$$\pi(\theta)d\theta = |W(\theta)| = J_\theta(\theta) = cd\beta d\sigma^2/\sigma^2, \quad (10)$$

which is the familiar right invariant prior.

Example 3.3. Transformation models. The appearance of the right invariant measure for Example 3.2 suggests the examination of the more general transformation model. For such models the transformation group directly generates the ancillary variable and orbit, and gives the standard conditioning. Accordingly for illustration here we just examine the simplified model after the conditioning and thus have that the group $G = \{g\}$ is equivalent to the sample space $Y = \{y\}$ and to the parameter space $\Omega = \{\theta\}$, all of dimension p . Then with continuity and smoothness of the group we have

$$y = \theta z$$

where z has a fixed distribution and the y distribution is generated from the transformation θ applied to the z distribution; the multiplication is group multiplication usually noncommutative as for example with the regression model. Let $g(z)$ be the density for the error variable z relative to the usual left invariant measure $d\mu(z)$. The statistical model for y can then be written

$$f(y; \theta)dy = g(\theta^{-1}y)d\mu(y).$$

The notation is simplified if the group coordinates are centered so that the identity element is at the maximum density point of the error density, $g(z) \leq g(e)$

where e designates the identity element satisfying $ez = z$. The maximum likelihood group element $\hat{\theta}(y)$ is thus the solution of $\theta^{-1}y = e$ which gives $\hat{\theta}(y) = y$. We then have from (7) that

$$d\hat{\theta} = \left| \frac{\partial y}{\partial \theta} \right|_{y^0} d\theta = \left| \frac{\partial(\theta z)}{\partial \theta} \right|_{y^0} d\theta$$

where the differentiation is for fixed pivot z with the subsequent substitution $z = z^0(\theta) = z(y^0, \theta)$. The related transformation theory (e.g. Fraser, 1979) then shows that the right side of the equation is a constant times the right invariant measure $d\nu(\theta)$ on the group. We thus have that the general prior developed in this section simplifies in the group model case to the right invariant prior.

The default prior (8) for the full parameter is appropriate only for component parameters that have a linearity property. For curved components it does not resolve the marginalization issues of Dawid et al. (1973) and it thus seems necessary to target the prior on the particular parameter component of interest (Section 6).

Example 3.4. Normal circle. Consider a simple example discussed in Fraser & Reid (2002): let (y_1, y_2) be distributed as $N\{(\mu_1, \mu_2), I\}$. The natural default prior is $cd\mu_1d\mu_2$ and it leads to the $N\{(y_1^0, y_2^0); I\}$ posterior for (μ_1, μ_2) . For any component parameter linear in (μ_1, μ_2) we have clear agreement between frequentist p -values and Bayesian survivor probabilities. By contrast, if we reparameterize the model as $\theta = (\psi, \alpha)$ where $\mu_1 = \psi \cos \alpha$ and $\mu_2 = \psi \sin \alpha$, there can be substantial differences between p -values and Bayesian survival probabilities for ψ . The p -value function is

$$p(\psi) = \Pr\{\chi_2^2(\psi^2) \leq y_1^2 + y_2^2\}$$

and the posterior survivor function is

$$s(\psi) = \Pr\{\chi_2^2(y_1^2 + y_2^2) \leq \psi^2\},$$

where $\chi_2^2(\delta^2)$ is a noncentral chisquare with 2 degrees of freedom and noncentrality parameter δ^2 . For example with $r = (y_1^2 + y_2^2)^{1/2} = 5$, the difference can be as much as eight percentage points for ψ near the data value 5, and for $\psi = 6$ the difference is $s(6) - p(6) = 0.1818 - 0.1375 = 0.0443$, which indicates substantial undercoverage for right tail intervals based on the marginal posterior. In the

extension to k dimensions, with y_i distributed as $N(\mu_i, 1/n), i = 1, \dots, k$, it can be shown that

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

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

For $k = 2$ we have $z_1 = y_1 - \psi \cos \alpha$ and $z_2 = y_2 - \psi \sin \alpha$, leading to

$$W(\theta) = \begin{pmatrix} \cos(\hat{\alpha} - \alpha) & \psi \sin(\hat{\alpha} - \alpha) \\ -\hat{\psi}^{-1} \sin(\hat{\alpha} - \alpha) & \psi \hat{\psi}^{-1} \cos(\hat{\alpha} - \alpha) \end{pmatrix}, \quad (11)$$

and then from (8) we obtain the default prior $J_\theta(\theta)d\theta \propto \psi d\psi d\alpha$ for the full parameter; this is of course equivalent to the default flat prior $d\mu_1 d\mu_2$ for the location parameter (μ_1, μ_2) . The discussion in the preceding paragraph however clearly indicates that this prior is not appropriate when the radial distance ψ is the parameter of interest, and thus that ψ is curved in the manner next to be described.

4 Linearity and marginalization paradoxes

A posterior distribution for a vector parameter can often lead to a marginal posterior for a component parameter of interest that has anomalous properties: for example, a marginal posterior can contradict the posterior based on the model from the component variable that appears in the marginal posterior. This was highlighted by Dawid, Stone & Zidek (1973) but is often under-appreciated in applications of Bayesian inference.

We will call a parameter contour $\psi(\theta) = \psi_0$ linear if a change $d\lambda$ in the nuisance parameter λ for fixed ψ generates through (5) a direction at the data point that is confined to a subspace free of λ and with dimension equal to $\dim(\lambda)$. For the normal circle example we note that the radius ρ is curved but the angle α is linear.

The linearity condition defines a location relationship between the nuisance parameter λ for fixed ψ and change at the data point. As such it provides an

invariant or flat prior for the constrained model, and thereby leads to a marginal model with the nuisance parameter eliminated. This avoids the marginalization paradoxes and parallels the elimination of a linear parameter in the standard location model.

We now examine the case with scalar θ_1 and scalar θ_2 , with parameter of interest $\psi(\theta)$ and develop the linear parameter that coincides with $\psi(\theta)$ in a small neighbourhood of the observed maximum likelihood value $\hat{\theta}^0$ and otherwise is defined by the linearity. As a preliminary step we use (5) with $W(\theta)$ to link sample space change at $\hat{\theta}^0$ with parameter space change at a parameter value θ ,

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

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

First we examine the parameter $\psi(\theta)$ near $\hat{\theta}^0$ on the parameter space and find that an increment $(d\theta_1, d\theta_2)$ with no effect on $\psi(\theta)$ must satisfy $d\psi(\theta) = \hat{\psi}_1^0 d\theta_1 + \hat{\psi}_2^0 d\theta_2 = 0$ where $\psi_i(\theta) = \partial\psi(\theta)/\partial\theta_i$; i.e. $d\theta_1 = -(\hat{\psi}_2^0/\hat{\psi}_1^0)d\theta_2$. Next we use (12) to determine the corresponding sample space increment at $\hat{\theta}^0$, and obtain

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

thus (c_1, c_2) so defined gives a direction $(c_1, c_2)dt$ on the sample space that corresponds to no ψ -change. Finally we use the inverse of (12) to determine the parameter space increment at a general point θ that corresponds to the preceding sample space increment, giving

$$d\theta = \begin{pmatrix} w^{11}(\theta)c_1 + w^{12}(\theta)c_2 \\ w^{21}(\theta)c_1 + w^{22}(\theta)c_2 \end{pmatrix} dt, \tag{13}$$

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

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

or the implicit equation $\theta_2 = \theta_2(\theta_1)$ as the solution of the differential equation

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

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

Example 4.1: We reconsider the regression Example 3.2, but for notational ease we consider the simple location-scale version with design matrix $X = 1$, and we construct the linear parameter that agrees near $\hat{\theta}^0$ with the quantile parameter $\mu + k\sigma$ for some fixed value k . From $W(\theta)$ in that example we obtain

$$\begin{aligned} d\hat{\mu} &= d\mu + (\hat{\mu}^0 - \mu)d\sigma/\sigma \\ d\hat{\sigma} &= \hat{\sigma}^0 d\sigma/\sigma. \end{aligned} \tag{14}$$

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

$$\begin{aligned} d\hat{\mu} &= d\mu - \mu d\sigma/\sigma \\ d\hat{\sigma} &= d\sigma/\sigma. \end{aligned} \tag{15}$$

Inverting this gives

$$\begin{aligned} d\mu &= d\hat{\mu} + \mu d\hat{\sigma} \\ d\sigma &= \sigma d\hat{\sigma}. \end{aligned} \tag{16}$$

First we examine $\mu + k\sigma$ in the neighbourhood of $\hat{\theta}^0$ on the parameter space and have that an increment $(d\mu, d\sigma)$ must satisfy $d(\mu + k\sigma) = 0$ at $\hat{\theta}^0 = (\hat{\mu}^0, \hat{\sigma}^0) = (0, 1)$; this gives $d\mu = -kd\sigma$ at $\hat{\theta}^0$. Next we determine the corresponding increment at $\hat{\theta}^0$ on the sample space $\{(\hat{\mu}, \hat{\sigma})\}$; from (15) we have $d\hat{\mu} = d\mu$ and $d\hat{\sigma} = d\sigma$ at this point. It follows then that $d\hat{\mu} = -kd\hat{\sigma}$. Finally we determine what restriction $d\hat{\mu} = -kd\hat{\sigma}$ on the sample space implies for $(d\mu, d\sigma)$ at a general point on the parameter space; from (16) this is

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

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

Example 4.2 In the normal circle case, Example 3.4, with parameter of interest $\psi = (\theta_1^2 + \theta_2^2)^{1/2}$, the increment on the parameter space at $\hat{\theta}^0$ with fixed ψ satisfies

$d\theta_1 = -\tan \hat{\alpha}^0 d\theta_2 = -(y_2^0/y_1^0)d\theta_2$. This then translates to the sample space at (y_1^0, y_2^0) using the specialized version of (12) to give satisfying $dy_2 = -(y_2^0/y_1^0)dy_1$ and the preceding then translates back to a general point on the parameter space using the specialized version of (13) to give a line through $\hat{\theta}^0$ described by $d\theta_2 = -(y_2^0/y_1^0)d\theta_1$ perpendicular to the radius and tangent to the circle through the data point, as the linear parameter equivalent to ψ near $\hat{\theta}_0$.

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

5 The exponential model approximation

A statistical model appropriately conditioned can be approximated at a data point by an exponential model that is fully defined and determined by the observed log-likelihood $\ell(\theta) = \log f(y^0; \theta)$ together with a canonical reparameterization

$$\varphi(\theta) = \frac{d}{dV} \ell(\theta; y)|_{y^0}, \quad (17)$$

which is the gradient of the log-likelihood function at the observed data point and is calculated in sensitivity directions $V = V^0 = V(\hat{\theta}^0)$ recorded at (3). In more detail we have that the j th component of the row vector $\varphi(\theta)$ is given by the gradient of the log-likelihood in the direction v_j which is the j th column vector in the array V :

$$\varphi_j(\theta) = (d/dt)\ell(\theta; y^0 + tv_j)|_{t=0}. \quad (18)$$

For some background and an extension to the discrete context, see Davison et al (2006).

The exponential model that approximates the conditional model is

$$g(s; \theta) = \exp\{\ell(\theta) + \varphi(\theta)s\}h(s), \quad (19)$$

where s is a p -dimensional score variable and has observed value $s^0 = 0$. The

saddlepoint approximation to this exponential model is

$$g(s; \theta) = \frac{e^{k/n}}{(2\pi)^{p/2}} \exp\left\{-\frac{r^2(\theta; s)}{2}\right\} |j_{\varphi\varphi}(\hat{\theta})|^{-1/2}, \quad (20)$$

where $r^2(s; \theta) = 2[\ell(\hat{\theta}) - \ell(\theta) + \{\hat{\varphi} - \varphi(\theta)\}s]$ is the log likelihood ratio in the approximate model, $j_{\varphi\varphi}(\hat{\theta}) = j_{\varphi\varphi}\{\hat{\theta}(s)\}$ is the observed information from a score value s in the approximating model and k is constant.

The exponential model (19) uses only the observed likelihood $\ell(\theta)$ and its gradient $\varphi(\theta)$, and yet provides third order inference for scalar parameters; see, for example, Fraser, Reid & Wu (1999) and Andrews et al. (2005). If the original model is a full exponential model, then the preceding calculation just reproduces that model to third order as its saddlepoint approximation.

In the exponential model approximation (19), the observed and expected information functions are identical and Jeffreys' prior can be written

$$\pi(\theta)d\theta = |j_{\varphi\varphi}(\theta)|^{1/2}d\varphi = |j_{[\theta\theta]}(\theta)|^{1/2}d\theta \quad (21)$$

where $j_{\varphi\varphi}(\theta) = -(\partial^2/\partial\varphi\partial\varphi')\ell(\theta)$ is the information function in the exponential model and $j_{[\theta\theta]}(\theta) = \varphi'_\theta(\theta)j_{\varphi\varphi}(\theta)\varphi_\theta(\theta)$ is the same information but presented relative to the θ parameterization for integration purposes. The canonical parameterization $\varphi(\theta)$ provides a calibration not otherwise available concerning θ ; $j_{[\theta\theta]}(\theta)$ gives second-derivative likelihood information that is calculated in the φ scaling and then rescaled to the θ parameterization; for some details on recalibrating informations see the Appendix.

For a scalar parameter model as noted in Example 3.1, the Welch & Peers (1963) default prior $\pi(\theta) \propto |i(\theta)|^{1/2}$ has what is now called second order matching, in the sense that β -level posterior probability limits based on this prior provide confidence coverage β to second order; for this $i(\theta)$ is the expected Fisher information. The same result can be obtained by direct Taylor series approximation of the asymptotic model (Cakmak et al, 1998) and shows that

$$\int^{\hat{\theta}} j_{\varphi\varphi}^{1/2}(\theta)d\varphi(\theta) - \int^{\theta} j_{\varphi\varphi}^{1/2}(\theta)d\varphi(\theta) = z$$

is a pivotal variable z to the second order, thus closely paralleling Example 3.1. Thus Welch & Peers (1963) in the scalar case gives a location model to second order, using expected or observed information in the φ parameterization; from recent likelihood theory we prefer the observed information for most calculations. It is essential however that the observed information function be calculated in the canonical φ parameterization and then rescaled as needed or appropriate.

6 Targetted default priors: scalar components

We now consider using the exponential model approximation along with Welch-Peers to develop a default prior when interest centers on a scalar parameter $\psi(\theta)$ of interest with $\lambda(\theta)$ as a scalar nuisance parameter. In accord with theory mentioned in the preceding section there is a reduced model obtained by conditioning that has the same effective sample space dimension as the parameter; for some recent discussion see Fraser & Rousseau (2008).

The analysis of component parameters (Fraser & Reid, 1993) shows that the just described reduced model can then be written to third order as a conditional model relevant to the nuisance parameter λ times a marginal model relative to the interest parameter ψ . The marginal for ψ can be examined on the curve on the sample space defined by the constrained maximum likelihood estimate of the nuisance parameter: $S_\psi = \{s : \hat{\theta}_\psi = \hat{\theta}_\psi^0\}$. This profile curve at the data s^0 can however vary in orientation as the parameter ψ varies, a complicating phenomenon for inference that arises even in the $N(\mu, \sigma^2)$ case, as we will see in Example 6.1.

The methods for dealing with the complication are addressed for marginal likelihood in Fraser (2003) and lead to coordinates at the data point s^0 that are standardized with respect to observed information as mentioned briefly in Section 3. This standardization seems intrinsic to the separation of likelihood information, so we require that $\varphi(\theta)$ be rescaled so that $\hat{j}_{\varphi\varphi} = I$. If an initial φ does not satisfy this, then $\tilde{\varphi}(\theta) = \hat{j}_{\varphi\varphi}^{1/2} \varphi(\theta)$ does, where $j_{\varphi\varphi}(\hat{\theta}^0) = (\hat{j}_{\varphi\varphi}^{1/2})'(\hat{j}_{\varphi\varphi}^{1/2})$ with $\hat{j}_{\varphi\varphi}^{1/2}$ given as a right root of the observed information matrix. When working just to second

order this standardization is not needed, but it does provide a natural invariance to sample space choice of coordinates (Fraser, 2003).

Fraser & Reid (2002) use recent third order likelihood methods to determine conditions on a default prior. For this let \mathcal{C} be the profile contour on the parameter space for $\psi(\cdot)$:

$$\mathcal{C} = \{\theta : \theta = \hat{\theta}_\psi^0\} = \{\theta : \theta = (\psi, \hat{\lambda}_\psi^0)\}.$$

Along this contour the observed likelihood function records the profile likelihood for $\psi(\theta)$: $L^0\{(\psi, \hat{\lambda}_\psi^0)\} = L_p(\psi)$. Then from (8.4) in Fraser & Reid (2002) we have

$$\pi(\hat{\theta}_\psi)d\psi d\lambda = c \frac{\ell_\psi(\hat{\theta}_\psi)}{\hat{\chi} - \hat{\chi}_\psi} d\psi \cdot |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2} |j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2} d\lambda \quad (22)$$

to third order. In (22), $\ell_\psi(\hat{\theta}_\psi)$ is the profile score function calculated from the observed profile likelihood $L_p(\psi)$ and the scalar parameter $\chi(\theta)$ is a rotated coordinate of the canonical parameter $\varphi(\theta)$ that is first derivative equivalent to $\psi(\theta) = \psi$ at $\hat{\theta}_\psi$. It is the unique locally defined scalar canonical parameter for assessing $\psi(\theta) = \psi$; see for example, Fraser, Reid & Wu (1999). The nuisance information matrix $j_{\lambda\lambda}(\hat{\theta}_\psi) = -\ell_{\lambda\lambda}(\hat{\theta}_\psi)$ is the information for λ at the constrained maximum; and $j_{(\lambda\lambda)}(\hat{\theta}_\psi)$ is the same, but rescaled to the metric provided by $\varphi(\theta)$ along the contour for given λ ,

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

where $\varphi_\lambda(\theta) = \partial\varphi(\theta)/\partial\lambda$ is $p \times (p - 1)$ and we use the notation $|X| = |X'X|^{1/2}$ for an $n \times p$ matrix; here $p = 2$. When not at a maximum likelihood value the calculations for nuisance information for various λ given ψ need to be carefully ordered: they are made within the exponential model defined by φ and are thus with respect to a φ -rescaled version of λ designated (λ) ; they are then rescaled to the initial λ parameterization; some details are given in the Appendix.

As our use of the exponential approximation in moderate deviations at the data point is second order we find it convenient to simplify to second order the initial factor

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

where $d(\psi) = d\chi_\psi$ is differential change in the parameter ψ on the profile contour, but expressed along the contour in terms of the φ metric, that is, in terms of the χ parameterization at that point; the two versions correspond respectively to third order and second order parameter linearization.

The right factor in (22) can be rewritten giving

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

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

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

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

The contour integration of the likelihood function for fixed ψ using the constant information metric $|j_{(\lambda\lambda)}(\psi, \lambda)|^{1/2}d(\lambda)$, or the Laplace integration using $|j_{(\lambda\lambda)}(\psi, \hat{\lambda}_\psi)|^{1/2}d(\lambda)$ at the intersection with the profile curve \mathcal{C} both produce just $(2\pi)^{1/2}$ times the profile value at ψ . Accordingly we replace the factor $|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}$ in (22) by $|j_{(\lambda\lambda)}(\theta)|^{1/2}$ to obtain the following general expression for the default prior:

$$\pi_\psi(\theta)d\psi d\lambda = c \frac{\ell_\psi(\hat{\theta}_\psi)}{\hat{\chi} - \hat{\chi}_\psi} d\psi \cdot |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\theta)|^{1/2}d\lambda \quad (23)$$

$$= |j_{\varphi\varphi}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\theta)|^{1/2}d(\psi)d(\lambda_\psi). \quad (24)$$

Example 6.1. Normal (μ, σ) . In Example 3.2 we obtained the default prior for the full parameter $\theta = (\mu, \sigma^2)$; we now use the higher order theory to directly address priors targetted on the components μ and σ^2 . The canonical parameter is $(\varphi_1, \varphi_2) = (\mu/\sigma^2, 1/\sigma^2)$ which has information function

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

and has Jeffreys prior $(n\sigma^3/\sqrt{2})d\varphi_1d\varphi_2 = (n/\sqrt{2}\sigma^3)d\mu d\sigma^2$. Without loss of generality we take the data point to be $(\hat{\mu}, \hat{\sigma}^2) = (0, 1)$. The re-standardized canonical parameter $(\tilde{\varphi}_1, \tilde{\varphi}_2)$ is $(n^{1/2}\mu/\sigma^2, n^{1/2}/\sqrt{2}\sigma^2)$ and has $j_{\tilde{\varphi}\tilde{\varphi}}^0 = I_2$ with information function

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

and Jeffreys prior

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

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

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

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

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

Extending this by the constant information metric for σ^2 gives

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

this is the familiar right invariant measure.

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

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

Thus on \mathcal{C} with $\hat{\mu}_\sigma^2 = 0$ we have the Jeffreys

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

and then the adjusted Jeffreys

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

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

Example 6.2. Normal circle (continued). We saw in Example 3.4 that the default prior for the vector $\varphi = (\psi \cos \alpha, \psi \sin \alpha)$ did not correctly target the component parameter ψ . Now, for inference about ψ , we have the following components of the targetted prior (23):

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

We then obtain the prior

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

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

7 Targetted default priors for scalar components: an alternative view

Frequentist likelihood theory provides third order accuracy for interest parameters using analysis that emphasizes continuity of the coordinate distribution function. The analysis obtains a marginal distribution on the profile curve $\mathcal{S}_\psi^0 = \{s : \hat{\theta}_\psi = \hat{\theta}_\psi^0\}$ and a conditional distribution concerning λ on a curve that is ancillary for λ when ψ is fixed. In particular if we work with the tangent exponential model from

Section 5 we have that \mathcal{S}_ψ^0 is a line on the score space and it is perpendicular to the contour $\psi(\theta) = \psi$ at $\hat{\varphi}_\psi = \varphi(\psi, \hat{\lambda}_\psi)$ on the canonical parameter space of $\varphi(\theta)$.

Now consider a neighbourhood of a value ψ and the corresponding profile line \mathcal{S}_ψ^0 . The probability differential on the sample space is $L(\psi, \lambda; s)ds_1ds_2$ where ds_1 is length on the line and ds_2 is distance from that line but for interpretation is used on the ancillary contour for λ ; see Fraser and Reid (1993, 2001). We now relate this to integration on the parameter space.

Suppose first that the full likelihood is integrated with respect to the Jeffreys prior for the nuisance parameter,

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

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

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

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

$$|j_{(\lambda\lambda)}(\psi, \hat{\lambda}_\psi)|^{1/2}|j_{[\lambda\lambda]}(\psi, \lambda)|^{1/2}d\lambda; \tag{25}$$

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

The rescaled Jeffreys integration for λ on the parameter space produces marginal probability concerning ψ with support ds_1 . For different ψ values the support can be on different lines through y^0 , which is the rotation complication that has complicated the development of marginal likelihood adjustments (Fraser, 2003). The choice of the standardized $\tilde{\varphi}(\theta)$ gives a common information scaling on the dif-

ferent lines through y^0 that are used to assess ψ . This provides sample space invariance and leads to the third order adjustment for marginal likelihood.

The adjusted nuisance Jeffreys prior (25) produces marginal likelihood for ψ , which then appears as an appropriately adjusted profile likelihood for that parameter of interest. This can then be integrated following the Welch-Peers pattern using root profile information obtained from the exponential parameterization. This gives the Jeffreys type adjustment

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

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

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

for use with the full likelihood $L(\psi, \lambda)$; this is in agreement with (22).

8 Targetted default priors: vector components

The information approach outlined above requires that the nuisance parameter be scalar, in order that the Welch-Peers approach can be used to extend the default prior beyond the profile contour. In this section we use the separation approach developed in these sections to extend the continuity approach to the case that 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 first partition the parameter effect matrix $\tilde{W}(\theta)$ at (6) in accord with the components ψ and λ , $\tilde{W}(\theta) = \{\tilde{W}_\psi(\theta), \tilde{W}_\lambda(\theta)\}$, and thus obtain

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

this separates the effects of ψ change and λ change at the observed data, and for example gives the prior $|\tilde{W}_\lambda(\theta)|d\lambda$ for λ , with ψ fixed, as specialized from (8).

Now to focus more closely on the parameter we examine how the variable, near the observed data y^0 , measures the parameter, in some physical sense. Following the pattern in the preceding section we have the Welch-Peers type effect of data change on λ given ψ as $|\tilde{W}_\lambda(\psi, \lambda)|d\lambda$ and the Welch-Peers type affect of data change on ψ on the profile curve $\mathcal{C} = \{(\psi, \hat{\lambda}_\psi)\}$ given as $|\tilde{W}_\psi(\psi, \hat{\lambda}_\psi)|d\psi$. This gives the composite targetted prior

$$\pi_\psi(\psi, \lambda)d\psi d\lambda = |\tilde{W}_\psi(\psi, \hat{\lambda}_\psi)||\tilde{W}_\lambda(\psi, \lambda)|d\psi d\lambda.$$

Example 8.1: Linear regression. Suppose $r = 3$ and let $\psi = (\beta_1, \beta_2)'$ be the parameter of interest and $\lambda = (\beta_3, \sigma^2)$ be the nuisance parameter. Then we have

$$W_\psi(\theta) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad W_\lambda(\theta) = \begin{pmatrix} 0 & (\hat{\beta}_1^0 - \beta_1)/\sigma^2 \\ 0 & (\hat{\beta}_2^0 - \beta_2)/\sigma^2 \\ 1 & (\hat{\beta}_3^0 - \beta_3)/\sigma^2 \\ 0 & \hat{\sigma}^2/\sigma^2 \end{pmatrix} \quad (26)$$

and thus deriving volumes as indicated after (4) in Section 2 we obtain

$$\begin{aligned} |W_\psi(\hat{\theta}_\psi)| &= 1, \\ |W_\lambda(\theta)| &= \left| \begin{array}{cc} 1 & (\hat{\beta}_3^0 - \beta_3)/\sigma^2 \\ (\hat{\beta}_3^0 - \beta_3)/\sigma^2 & \frac{\hat{\sigma}^4 + \sum_1^2 (\hat{\beta}_i^0 - \beta_i)^2/\sigma^4}{\sigma^4} \end{array} \right|^{1/2} \\ &= \left(\frac{\hat{\sigma}^4 + \sum_1^2 (\hat{\beta}_i^0 - \beta_i)^2/\sigma^4}{\sigma^4} \right)^{1/2}. \end{aligned}$$

As $\hat{\beta}_i^0 - \beta_i$ is of order $n^{-1/2}$ we have $(\hat{\beta}_i^0 - \beta_i)^2 = O(n^{-1})$ and thus $|W_\lambda(\theta)|$ simplifies to c/σ^2 to second order. This gives the prior $d\beta d\sigma^2/\sigma^2$ as expected from Example 3.2.

9 Discussion

Default priors are widely used in Bayesian analysis and occasionally in frequentist analysis. A default prior is a density or relative density used to weight an observed likelihood to give a composite numerical assessment of possible parameter values,

commonly called a posterior density in Bayesian analysis. The choice of prior is based on model characteristics, a location relationship following Gauss, Laplace and Jeffreys or information processing properties as indicated by Zellner (1988) and Bernardo (1979); the resulting posterior is then available for subsequent objective, subjective, personal or expedient adjustments. A posterior interval may not have reproducibility and may not agree with intervals derived from component variables. Priors to avoid the first issue are called matching priors; and priors to avoid the second are called targetted priors.

The likelihood function provides an accessible easily implementable approach to obtaining first order inference from a statistical model with data; for this, the only data information typically used is the observed likelihood. Then without outside information concerning the source of the unknown parameter value, the model itself can be used to provide a neutral background determination of the weight function to be applied to the observed likelihood; this leads to second order inference for linear parameters. For nonlinear parameters the model again can be used to provide an adjustment for nonlinearity and leads to second order inference.

In principle such additional model information can include any relevant aspect of the model. Bayesian practice however has tended to emphasize model properties implicit in the process of converting a prior into a posterior; this is certainly an important aspect of model information but does not explicitly include such very direct model information as the continuity often present between variable and parameter, and available typically in coordinate distribution functions. By including such continuity information we have determined the conforming parameters and developed appropriate targetted priors for nonconforming parameters.

10 Appendix

For any given model with regularity we have an exponential model as a second order approximation; it uses observed log-likelihood $\ell(\theta)$ and observed log-likelihood

gradient $\varphi(\theta)$ and has log-likelihood form

$$\ell(\theta; s) = \ell(\theta) + \varphi(\theta)s$$

in moderate deviations with data $s = 0$. We want information calculations for this exponential model but the needed derivatives are often accessible just through the available parameterization θ .

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

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

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

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

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

where as usual $|\varphi_{\lambda}|^2 = |\varphi'_{\lambda}\varphi_{\lambda}|$. Then for Welch-Peers purposes we may want to integrate with respect to the available variable λ and would then correspondingly want to rescale the information to the initial λ scale:

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

References

- [1] Andrews, D.F., Fraser, D.A.S. and Wong, A. (2005). Computation of distribution functions from likelihood information near observed data. *J. Statist. Plann. Infer.* **134**, 180–193.

- [2] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc. London* **53**, 370–418; **54**, 296–325. Reprinted in *Biometrika* **45**(1958), 293–315.
- [3] Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion).
- [4] Cakmak, S., Fraser, D.A.S., McDunnough, P., Reid, N. and Yuan, X. (1998). Likelihood centered asymptotic model: exponential and location model versions. *J. Statist. Plann. Infer.* **66**, 211–222.
- [5] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.
- [6] Datta, G.S. and Ghosh, M. (1995). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90**, 1357–1363.
- [7] Davison, A.C., Fraser, D.A.S. and Reid, N. (2006). Likelihood inference for categorical data. *J. R. Statist. Soc B* **68**, 495–508.
- [8] Dawid, A.P., Stone, M. and Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. B* **35**, 189–233.
- [9] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Royal Soc. London A* **222**, 309–368.
- [10] Fraser, D.A.S. (1979). *Inference and Linear Models*. New York: McGraw-Hill.
- [11] Fraser D.A.S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327–339.
- [12] Fraser, D.A.S. and Reid, N. (1993). Third order asymptotic models: likelihood functions leading to accurate approximations for distribution functions. *Statist. Sinica* **3**, 67–82.
- [13] Fraser, D.A.S. and Reid, N. (2001). Ancillary information for statistical inference. In S.E. Ahmed and N. Reid (Eds), *Empirical Bayes and Likelihood Inference*, 185–207. New York: Springer-Verlag.

- [14] Fraser, D.A.S. and Reid, N. (2002). Strong matching of frequentist and Bayesian inference. *J. Statist. Plan. Infer.* **103**, 263–285.
- [15] Fraser, D.A.S. and Rousseau, J. (2008). Studentization and developing p -values. *Biometrika* **95**, 1–16.
- [16] 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.
- [17] Jeffreys, H. (1939). *Theory of Probability*. Oxford: Oxford University Press. Third Edition (1961).
- [18] Kass, R.E. (1990). Data-translated likelihood and Jeffreys’s rules. *Biometrika* **77**, 107–114.
- [19] Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.* **30**, 877–880.
- [20] Tibshirani, R. (1989). Noninformative priors fro one parameter of many. *Biometrika* **76**, 705–708.
- [21] Welch, B.L. and Peers H.W. (1963). On formulae for confidence points based in intervals of weighted likelihoods. *J. Roy. Statist. Soc. B* **25**, 318–329.
- [22] Zellner, A. (1988). Optimal information processing and Bayes’ theorem. *Amer. Statist.* **42**, 278–284.