

Default priors
for Bayesian and frequentist inference

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

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

G.Y. Yi
University of Waterloo, Canada

April 22, 2009

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

Abstract

We investigate the choice of default prior for use with likelihood to facilitate Bayesian and frequentist inference. Such a prior is a density or relative density that weights an observed likelihood function leading to the elimination of parameters not of interest and then to providing a density type assessment for a parameter of interest. For regular models with independent coordinates we develop a prior for the full parameter that has second order accuracy]; it is closely linked to the original Bayes approach and provides an extension of the familiar right invariant measure to general contexts. We then develop a modified prior that is targetted on a component parameter of interest and by targetting avoids the marginalization paradoxes of Dawid, Stone and Zidek (1973). In particular this modifies the Jeffreys' prior and provides extensions to Welch-Peers theory. 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 applied to an observed likelihood function. The choice of prior is based directly on continuity in the model and an absence of information as to how the parameter value was generated.

One Bayesian role for a default prior is to provide a reference allowing subsequent modification by an objective, subjective, personal, elicited or expedient prior. From a frequentist viewpoint a default prior can be viewed as a device to substitute integration on the parameter space for integration on the sample space

and thus to directly use the likelihood function. From either view it offers a flexible and easily implemented exploratory approach to statistical inference.

We construct default priors directly from the continuity that is present in most regular statistical models, in the coordinate distribution or quantile functions. For any general point in the parameter space we can examine how parameter change causes change in the model near an observed data point. For a particular independent coordinate at the data point the resulting change in the quantile as a function of the parameter is typically easy to calculate. And then for the full data vector the corresponding volume change is available as a function of the parameter. This reflects the sensitivity of the parameter at the data point. It then corresponds closely to replacing sample space integration by analogous parameter space integrations. This is developed in Section 2, leading to the default prior given below by (6) or (8) and then illustrated on some familiar models.

Then in Sections 3 and 4 we consider examples of exact and approximate priors based on model continuity. As part of this we show that the default prior needs in general to be targetted on the parameter of interest, when there is a type of nonlinearity in that parameter; this is an aspect of the marginalization paradoxes of Dawid, Stone and Zidek (1973). In Section 5 we use available third order approximations to p -values and posterior probabilities to derive a suitably targetted prior defined on the profile curve of the parameter of interest, and we then extend this to the full parameter space, leading to a full default targetted prior, given below by (19); this information based approach is however limited to the case of scalar interest and scalar nuisance parameters.

In Section 6 we use the continuity based approach to develop targetted priors for the general vector interest vector nuisance parameter case. Section 7 records a brief discussion. Our goal is to examine the structure of priors for which stated levels for posterior inference are realized, at least approximately. One conclusion is that to attain this beyond a first order of approximation, the prior in general needs to depend on the data.

The term objective prior is frequently used to refer to what we call a default

prior: in our view the term objective prior is appropriate just in contexts where it is known that θ arose from some density $g(\theta)$. There is a large literature on the construction of various types of default priors, and a complete review is beyond the scope of this paper. Our approach is closely related to the original Bayes proposal and has connections to the search for so-called “matching priors”, which require that posterior bounds on a scalar parameter θ have the reproducible confidence property, to some order of approximation. A different approach to developing default priors is based on information processing, as in Zellner (1988) and Bernardo (1979): in particular the reference priors of Bernardo are those that maximize the Kullback-Leibler divergence between the posterior and the prior. Yuan and Clarke. Kass and Wasserman. Mukerjee et al. latest Diccio and Young. George and McCulloch. Check Richardson and Green (1997) Wasserman (2000). Some further comments on various choices of default prior are provided in the Discussion.

Suppose¹ we have a single observation on a scalar parameter θ from an assumed model with density $f(y; \theta)$ and distribution function $F(y; \theta)$. For an observed value y^0 , the p -value as a function of θ is $F(y^0; \theta)$. The posterior probability function of θ is $\int_{\theta} f(y^0; \vartheta)\pi(\vartheta)d\vartheta$. If these two inference functions are equal for all θ giving equivalence of posterior and frequentist inference,

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

then differentiation of both sides with respect to θ gives

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

or equivalently

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

where the subscript notation indicates differentiation with respect to the relevant argument. In Section 2 we show that this generalizes constant priors for location parameters to give general approximate priors for non-location models. The

¹additional material added by NR which may be out of place at the moment but might be able to replace parts of the Introduction

prior can then be interpreted as $\pi(\theta) \propto dy/d\theta$, where the derivative is computed with $F(y; \theta)$ held fixed. This also enables the extension of flat priors to vector parameters.

However, with vector parameters the simple correspondence between p -values and posterior probabilities outlined here cannot be attained for each component parameter, partly because the exact frequentist p -value requires the elimination of the nuisance parameter. To extend this correspondence to more complex models we use third order approximations for marginal posterior probabilities and for frequentist p -value functions in Section 5, and use the continuity arguments in Section 6.

2 Default priors from continuity

Consider a sample $y = (y_1, \dots, y_n)$ from a parametric model with density $f(y; \theta)$ and distribution function $F(y; \theta)$ where θ is a vector of length p and f is continuous in y and θ . An intrinsic property of this is that a small change in the parameter θ causes a change in the distribution of any particular component y_i , and that this in turn causes a change in the distribution of the vector y . Continuity in the model then establishes an approximate location relationship between y and θ in a neighbourhood of the observed data point, and this relationship determines our proposed default prior. More generally we can have independent coordinates with parametric models as just described.

We first consider the simplest case of a single observation from a scalar parameter location model $f(y; \theta) = f(y - \theta)$. This model is invariant under location change in the parameter: the quantile at any observed value y^0 shifts to $y^0 + d\theta$ when θ changes to $\theta + d\theta$, i.e. $F(y^0; \theta) = F(y^0 + d\theta; \theta + d\theta)$. For a non-location model this generalizes by having the total differential of $F(y; \theta)$ be equal to zero at y^0 :

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

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

$$\left. \frac{dy}{d\theta} \right|_{y^0} = -\frac{F_\theta(y^0; \theta)}{F_y(y^0; \theta)}. \quad (1)$$

We can obtain an alternate expression by setting $u = F(y; \theta)$, solving for the u -quantile $y = y(u; \theta)$ and then differentiating directly:

$$\left. \frac{dy}{d\theta} \right|_{y^0} = y_\theta(u, \theta)|_{u=F(y^0; \theta)}, \quad (2)$$

where the partial differentiation with respect to θ is calculated for fixed quantile level $u = u^0 = F(y^0; \theta)$. In (1) and in (2) the derivative on the left hand side is calculated with the probability $F(y^0; \theta)$ held fixed.

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

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

which is a $1 \times p$ row vector. This generalizes translation invariance to local translation invariance (Fraser, 1964).

Now for the full response vector $y = (y_1, \dots, y_n)$, each y_i has a corresponding row vector $V_i(\theta)$ defined by (3) using the coordinate distribution functions; the change then in the vector variable y at y^0 under differential change in θ is

$$\left. \frac{dy}{d\theta'} \right|_{y^0} = \left\{ \begin{array}{c} V_1(\theta) \\ \vdots \\ V_n(\theta) \end{array} \right\} = V(\theta), \quad (4)$$

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

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

a translation or invariance argument to recommend the flat prior $\pi(\theta) = c$ for the parameter θ .

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

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

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

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

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

where $H' = \ell_{\theta; y'}(\hat{\theta}^0; y^0)$ is the gradient of the score function at the data point and $\hat{j} = j(\hat{\theta}^0; y^0) = -\partial^2\ell(\hat{\theta}^0; y^0)/\partial\theta\partial\theta'$ is the observed information. Combining this with (4) we obtain

$$d\hat{\theta} = \hat{j}^{-1}H'V(\theta)d\theta = W(\theta)d\theta \tag{5}$$

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

$$\pi(\theta)d\theta = |\hat{j}^{-1}HV(\theta)|d\theta = |W(\theta)|d\theta. \tag{6}$$

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

$$\hat{j}^{1/2}d\hat{\theta} = \hat{j}^{1/2}W(\theta)d\theta = \tilde{W}(\theta)d\theta. \quad (7)$$

The rescaled default prior is then

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

3 Examples: exact default priors from continuity

We illustrate the default prior in settings similar to Bayes' original location model, where posterior quantiles agree with frequentist inference.

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

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

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

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

where $\hat{\sigma}^0$ is abbreviated as $\hat{\sigma}$ to simplify notation. The observed information is $\hat{j} = \text{diag}\{X'X/\hat{\sigma}^2, n/(2\hat{\sigma}^4)\}$; combining these using (6) and least squares projection

properties gives

$$\begin{aligned}
W(\theta) &= \left\{ \begin{array}{cc} \hat{\sigma}^2(X'X)^{-1} & 0 \\ 0 & 2\hat{\sigma}^4/n \end{array} \right\} \left(\begin{array}{c} X'/\hat{\sigma}^2 \\ \hat{z}^{0'}/\hat{\sigma}^3 \end{array} \right) \{ X \quad z^0(\theta)/(2\sigma) \} \\
&= \left\{ \begin{array}{cc} I & (X'X)^{-1}X'z^0(\theta)/(2\sigma) \\ 2\hat{z}^{0'}\hat{\sigma}X/n & \hat{z}^{0'}z^0(\theta)\hat{\sigma}/(n\sigma) \end{array} \right\} \\
&= \left\{ \begin{array}{cc} I & (\hat{\beta}^0 - \beta)/2\sigma^2 \\ 0 & \hat{\sigma}^2/\sigma^2 \end{array} \right\}.
\end{aligned}$$

leading to

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

It thus follows from (6) that the default prior is

$$\pi(\theta)d\theta \propto |W(\theta)|d\theta \propto d\beta d\sigma^2/\sigma^2, \tag{9}$$

which is the familiar right invariant prior. This example illustrates how (6) modifies the invariance argument of Jeffreys to adapt to the underlying location parameterization. Jeffreys' argument based on the use of the determinant of the expected Fisher information matrix to construct default priors leads to the left invariant prior $d\beta d\sigma^2/\sigma^4$, which is usually regarded as incorrect for this problem; for example the associated posterior does not reproduce the t -distribution with the usual degrees of freedom for inference about components of β , whereas the right invariant prior does.

Example 3.2. Normal circle. As a special case of the normal theory linear model let (y_1, y_2) be distributed as $N\{(\mu_1, \mu_2), I/n\}$. It follows either from the preceding example, or from direct calculation, that the default prior for the location parameter μ is $cd\mu_1d\mu_2$, which gives the $N\{(y_1^0, y_2^0); I/n\}$ posterior for (μ_1, μ_2) . For any component parameter linear in (μ_1, μ_2) we then have exact agreement between frequentist p -values and Bayesian survivor probabilities.

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

$\psi \sin \alpha + z_2$, where z_1, z_2 are independent standard normal variables. This gives

$$W(\theta) = \left\{ \begin{array}{cc} \cos(\hat{\alpha} - \alpha) & \psi \sin(\hat{\alpha} - \alpha) \\ -\hat{\psi}^{-1} \sin(\hat{\alpha} - \alpha) & \psi \hat{\psi}^{-1} \cos(\hat{\alpha} - \alpha) \end{array} \right\}, \quad (10)$$

and then from (6) or (8) we obtain the default prior $\pi(\theta)d\theta \propto \psi d\psi d\alpha$ for the full parameter. This is equivalent to the default flat prior $d\mu_1 d\mu_2$ for the location parameter (μ_1, μ_2) .

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

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

where $\chi_2^2(\delta^2)$ is a noncentral chi-square with 2 degrees of freedom and noncentrality parameter δ^2 and the y 's are fixed at their observed values. In contrast the posterior survivor function for ψ under the flat prior $d\mu_1 d\mu_2$ is

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

For example with $\sqrt{n(y_1^2 + y_2^2)}^{1/2} = 5$, 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.

Note that this discrepancy does not appear in the first order of asymptotic theory: the limiting distribution of $\sqrt{n}\hat{\psi} = \sqrt{n(y_1^2 + y_2^2)}^{1/2}$ is $N(\psi, 1)$, and the limiting distribution of the exact posterior for ψ is $N(\hat{\psi}, 1/n)$, so to this order of approximation p -values and marginal survivor probabilities are identical. This is simply reflecting the fact that any prior not depending on n is in the limit swamped

by the data and has no effect on the posterior inference. Thus to study the agreement between Bayesian and frequentist inference it is necessary to consider either the exact distributions, or higher order approximations.

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

This example illustrates in simple form the difficulty with the default prior (6) and any ‘flat’ prior for a vector parameter. It is not possible to achieve approximate equality of Bayesian and frequentist inferences beyond the simple asymptotic normal limit when the parameter of interest is curved in the local location parameter. This is another version of the marginalization paradox of Dawid et al. (1973) and it thus seems necessary to target the prior on the particular parameter component of interest. This is well-recognized in the literature on the construction of reference priors, but seems less well appreciated in other contexts. In Sections 6 and 7 we describe how to adapt a version of the default prior (6) or (8) to target a particular parameter of interest. Discussion of the local location parameter and associated non-linearity is deferred to the Appendix.

Example 3.3. Transformation models. The preceding examples and many more are special cases of transformation models. In the Appendix we briefly record the links to this general type of model and the result that our locally defined prior (6) reproduces the right invariant prior for that model type, thus (6) can be written

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

where $d\nu(\theta)$ is the right invariant measure on the transformation group. The details are given in the Appendix.

4 Examples: approximate default priors

Some other examples exhibit an approximate location relationship and thus inherit some of the reliability of Bayes original example. One way of quantifying this is by using asymptotic criteria.

Example 4.1. The Welch-Peers approximation. As noted above, the construction of the default prior using the sensitivity matrix $V(\theta)$, or the modification to $W(\theta)$, gives a flat prior when θ is a location parameter. If we have a scalar parameter model in which the location parameter is $\beta(\theta)$, then this construction gives the flat prior $\pi(\theta)d\theta \propto d\beta(\theta)$. For any scalar parameter model, an approximate location parameter is available, and is proportional to $i^{1/2}(\theta)$, where $i(\theta)$ is the expected Fisher information $E_{\theta}\{-\ell''(\theta)\}$. This was established in Welch & Peers (1963), by showing that

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

has a limiting standard normal distribution, and to second order has a distribution free of θ . In Welch & Peers (1963) this was formulated in terms of the equality of confidence and posterior bounds. The result can be expressed in quantile form as

$$\hat{\beta} = \beta + z_{\gamma},$$

where $\beta(\theta) = \int^{\theta} i^{1/2}(t)dt$ is the constant information reparametrization and z_{γ} is the γ quantile of the standard normal. Then $d\hat{\beta} = d\beta$ for fixed quantile, giving the flat prior $d\beta \propto i^{1/2}(\theta)d\theta$. This is simply Jeffreys' (1939) prior for this scalar case, and its interpretation in terms of an approximate location parameter is discussed in Kass (1990).

This example links the location-parameter approach for constructing priors to that based on Fisher information. In Sections 5 and 6 we extend this linking to develop targetted priors from exponential family approximations.

Example 4.2. Nonlinear regression. Suppose y_i are independently normally distributed with mean $x_i(\beta)$ and variance σ^2 for $i = 1, \dots, n$, with $x_i(\beta)$ a known nonlinear function of the $p \times 1$ vector β . As in example 3.1, the quantile functions are $y_i = x_i(\beta) + \sigma z_i$ where $z_i = \Phi^{-1}(u_i)$, and the sensitivity matrix $V(\theta)$ obtained by differentiating these for fixed z is

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

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

$$H = \frac{1}{\hat{\sigma}^2} \left\{ X(\hat{\beta}^0) \quad \hat{z}^0/\hat{\sigma} \right\}$$

$$\hat{j} = \left\{ \begin{array}{cc} X'(\hat{\beta}^0)X(\hat{\beta}^0)/\hat{\sigma}^2 & 0 \\ 0 & n/2\hat{\sigma}^4 \end{array} \right\}$$

where again for notational convenience we write $(\hat{\sigma}^0)^2 = \{y - x(\hat{\beta}^0)\}'\{y - x(\hat{\beta}^0)\}/n$ as just $\hat{\sigma}^2$.

We then obtain

$$W(\theta) = \left\{ \begin{array}{cc} \hat{\sigma}^2(X'(\hat{\beta}^0)X(\hat{\beta}^0))^{-1} & 0 \\ 0 & 2\hat{\sigma}^4/n \end{array} \right\} \left\{ \begin{array}{c} X'(\hat{\beta}^0)/\hat{\sigma}^2 \\ \hat{z}^0/\hat{\sigma}^3 \end{array} \right\} \{X(\beta) \quad z^0(\theta)/2\sigma\}$$

which is similar in form to that in the ordinary regression in Example 3.1. To examine the prior based on $W(\theta)$ we introduce temporary coordinates based on the projection to the tangent plane at the fitted data point. In terms of these new coordinates appropriately standardized we have

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

where w is orthogonal to $\mathcal{L}\{X(\hat{\beta}^0)\}$. It follows then as in Example 3.1 that the default prior is $d\beta d\sigma^2/\sigma^2$ to second order, where $d\beta$ now designates a flat prior in the tangent plane coordinates.

This gives

$$nW(\theta) = \left\{ \begin{array}{cc} \hat{j}_{11}^{-1} \sum \hat{x}'_i(\hat{\beta}) \hat{x}_i(\beta) & \hat{j}_{11}^{-1} \sum \hat{x}'_i(\hat{\beta}) z_i(\theta) \\ \sum \hat{z}_i \hat{x}_i(\beta) & \sum \hat{z}_i z_i(\theta) \end{array} \right\}$$

where $\hat{z}_i = \{y_i - x_i(\hat{\beta})\}/\hat{\sigma}$ and $z_i(\theta) = \{y_i - x_i(\beta)\}/\sigma$.

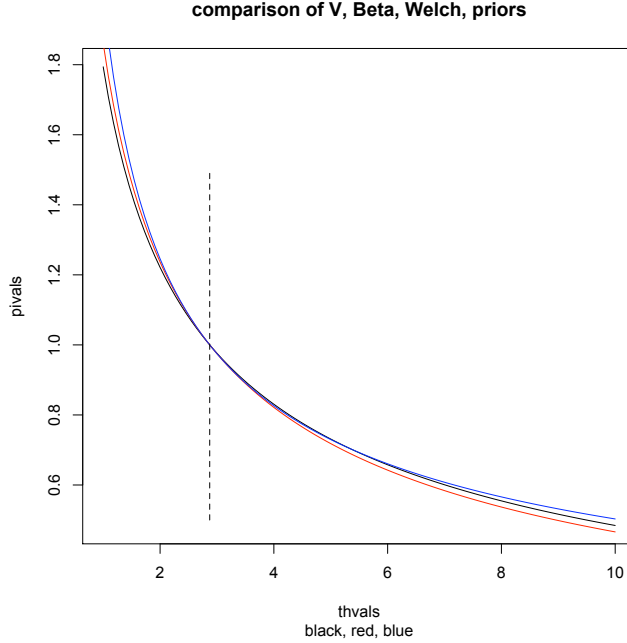
Example 4.3. Gamma distribution. As an example of a one-parameter model which is neither location nor scale, we consider default priors for the shape parameter of a gamma distribution:

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

Jeffreys' prior is $\pi_J(\theta) \propto \psi''(\theta)^{1/2}$, where $\psi(\theta) = \log \Gamma(\theta)$. To construct a location-based prior for a sample of n , we use (1) with (6) to determine $V_i(\theta)$,

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

Figure 1: Comparison of three default priors for the shape parameter of a Gamma distribution. Jeffreys' prior $\pi_J(\theta)$ is proportional to the trigamma function $\psi''(\theta)$; the default prior proposed here is based on (1) and (6), and $\pi_S(\theta)$ is a location model approximation to Jeffreys' prior. The three priors have been standardized to equal 1 at $\hat{\theta}^0$.



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

Figure 1 shows π_J and π_V for a sample of size 30 from the gamma distribution with shape parameter $\theta = 3$. The priors are normalized to equal 1 at $\theta = \hat{\theta} = 2.578$. Also shown is the strong matching prior given in Fraser & Reid (2002), $\pi_S(\theta) \propto \{\psi'(\theta) - \psi'(\hat{\theta}^0)\}/(\theta - \hat{\theta}^0)$, which on Taylor expansion recovers $\pi_J(\theta)$ as the leading term; for some details see Sections 8(iiii) and (v). We note that the priors agree in the neighbourhood of the observed maximum likelihood value, then have slightly different curvature as they respond differently to curvature in the model.

Example 4.4. Two parameter gamma model. We illustrate some of the issues with vector parameters by extending this example to the two-parameter gamma:

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

The sensitivity matrix $V(\mu, \theta)$ has two columns, and the column corresponding to

μ is simply $(y_1^0/\mu, \dots, y_n^0/\mu)'$, while the column corresponding to θ has elements

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

where the notation F_1 and f_1 refer to the gamma densities for $\mu = 1$ used above.

The associated default prior is given by

$$\pi(\theta, \mu) \propto |V(\mu, \theta)'V(\mu, \theta)|^{1/2} = \frac{|y^0|}{\mu} \left\{ \sum \tilde{V}_i^2 - (\sum y_i \tilde{V}_i)^2 / \sum y_i^2 \right\}^{1/2}.$$

[It is not clear if \tilde{V}_i depends on μ or not. The detailed expression for \tilde{V}_i is

$$\begin{aligned} \tilde{V}_i(\theta, \mu) &= -\frac{F_\theta(y_i^0; \theta, \mu)}{f(y_i^0; \theta, \mu)} \\ &= \frac{-(d/d\theta) \int_{-}^{y_i^0/\mu} \frac{1}{\Gamma(\theta)} z^{\theta-1} e^{-z} dz}{\frac{1}{\Gamma(\theta)} \frac{y_i^{0(\theta-1)}}{\mu^\theta} e^{-y_i^0/\mu}} \\ &= \frac{\Gamma'(\theta) F_1(y_i^0/\mu; \theta) - \int_0^{y_i^0/\mu} z^{\theta-1} \log z e^{-z} dz}{\frac{y_i^{0(\theta-1)}}{\mu^\theta} e^{-y_i^0/\mu}} \end{aligned}$$

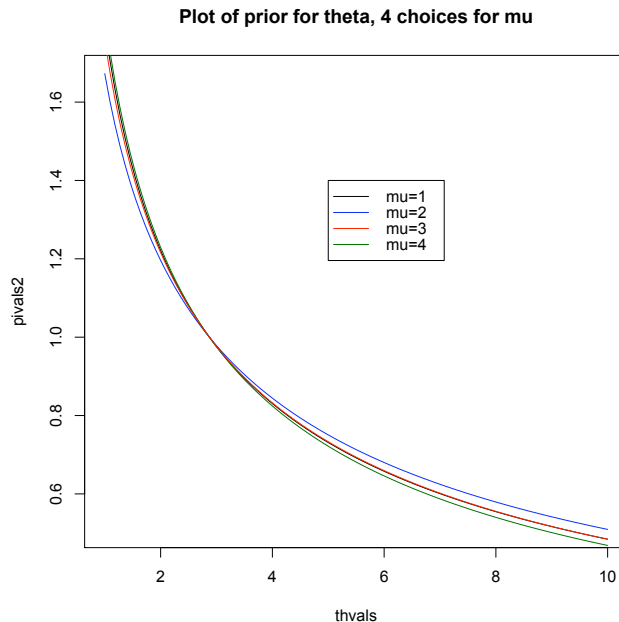
and when programmed in R seems to change with μ : see Figure 2.

Figure 2 plots the second factor in this joint prior for several values of μ . We can see that it is nearly constant in μ . The change with μ could be computational, although I don't think so, but it could also be a reflection of the fact that the parameter θ is not linear, so we don't exactly get what we want for the marginal prior of μ , which would be $1/\mu$, if we use the joint prior.]

5 Information based priors

The approach developed in the preceding sections gives a default prior for a vector parameter, but the resulting posterior is not appropriately targetted on component parameters unless the parametrization of the model is linear, in the sense discussed in the Appendix Section 8(iv). To develop default priors that are targetted on parameters of interest, we use initially a different approach, motivated by higher order asymptotics and by the similarity between the Welch-Peers prior and the default prior noted in Example 4.1. In that example, the Fisher information

Figure 2: Comparison of the default prior based on $V(\theta)$ for the shape parameter of a Gamma distribution, plotted here for four different values of μ . The four priors have been standardized to equal 1 at $\hat{\theta}^0$.



function defines locally a location parameter, and the resulting ‘flat’ prior is given by the Fisher information metric. To generalize this to the vector case, we can either generalize the location model approximation, which we did in the previous section, or the information approach, which we now consider. For targetting the prior on the parameter of interest, the information approach seems more directly accessible. For developing default priors for vector parameters, either approach can be used, as long as it is remembered that the resulting posterior is still subject to marginalization issues; that is, may not give well-calibrated marginal posterior inference.

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

solution (assumed unique) of $\partial\ell(\theta)/\partial\lambda = 0$.

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

$$s(\psi) = \Phi(r_B^*) = \Phi\{r + (1/r)\log(q_B/r)\} \quad (11)$$

where

$$r = \text{sign}(\hat{\psi} - \psi)[2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]^{1/2}, \quad (12)$$

$$q_B = \ell'_p(\psi)j_p^{-1/2}(\hat{\psi}) \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)}, \quad (13)$$

$j_{\lambda\lambda}$ is the submatrix of the observed Fisher information matrix corresponding to the nuisance parameter, and $\ell_p(\psi) = \ell(\hat{\theta}_\psi)$ is the profile log-likelihood function. This can be derived from the Laplace approximation to the marginal posterior density for ψ : see, for example, Tierney & Kadane (1986), DiCiccio & Martin (1993) and Bédard et al (2007). The approximation has relative error $O(n^{-3/2})$ for ψ in $n^{-1/2}$ -neighbourhoods of $\hat{\psi}$.

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

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

with observed data $s^0 = 0$, and using available approximation formulas for such (p, p) exponential families. Some discussion of this use of the exponential family model $\{\ell(\theta), \varphi(\theta)\}$ as a full third order surrogate for the original model is given in Davison et al (2006).

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

$$p(\psi) = \Phi(r_f^*) = \Phi\{r + (1/r)\log(q_f/r)\} \quad (14)$$

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

$$\begin{aligned} q_f &= \frac{|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi)|}{|\varphi_\theta(\hat{\theta})|} \frac{\varphi_\lambda(\hat{\theta}_\psi)}{|j(\hat{\theta})|^{1/2}} \frac{|j(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}, \\ &= \{\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)\} \left\{ \frac{|j_{\varphi\varphi}(\hat{\theta})|}{|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|} \right\}^{1/2}. \end{aligned} \quad (15)$$

The second version of q_f indicates that it can be presented as a parameter departure divided by its estimated standard error, and the first version gives a form that is useful for computation. We now describe the use of these two r^* approximations in developing priors.

Because the only difference between (11) and (14) is in the use of q_B or q_f , and only q_B involves the prior, we obtain equality of posterior and frequentist inference to $O(n^{-3/2})$ by deriving a prior so $q_B = q_f$. This was suggested in DiCiccio et al. (1995), and was developed further in Fraser & Reid (2002), where it was called strong matching. Strictly speaking inference for ψ can be obtained using (14) alone, but the close parallel between (11) and (14) does determine some aspects of the prior needed to ensure frequentist validity, at least to the present order of approximation. Thus we have

$$\frac{\pi(\hat{\theta}_\psi)}{\pi(\hat{\theta})} \propto \frac{\ell'_p(\psi) |j_{\lambda\lambda}(\hat{\theta}_\psi)| |\varphi_\theta(\hat{\theta})|}{|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi)| \varphi_\lambda(\hat{\theta}_\psi) |j(\hat{\theta})|}. \quad (16)$$

This gives a data-dependent prior along the profile curve $\mathcal{C}_\psi = \{\theta : \theta = (\psi, \hat{\lambda}_\psi)\}$ in the parameter space, but does not immediately give a prior over the full parameter space. To extend the prior for arbitrary values of λ beyond the profile curve we use the Welch & Peers (1963) construction based on Fisher information, as described below at (19).

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

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

where $V_i(\theta)$ is the i th row of the sensitivity matrix (4). A derivation of this r_f^* approximation is beyond the scope of this paper, but some notes are provided in

supplementary material. The role of $V(\hat{\theta}^0)$ in the development of the approximation is to implement conditioning on an approximate ancillary statistic derived from a local location model, which is why the same matrix arises here as in the discussion of default priors.

Using the second expression for q_f in (15), we obtain

$$\pi(\hat{\theta}_\psi)d\psi d\lambda = c \frac{\ell'_p(\psi)}{\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)} d\psi \cdot |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2} |j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2} d\lambda \quad (17)$$

to third order on the profile curve C_ψ . 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 matrix $j_{(\lambda\lambda)}(\hat{\theta}_\psi)$ is the nuisance information matrix $j_{\lambda\lambda}(\hat{\theta}_\psi) = -\ell_{\lambda\lambda}(\hat{\theta}_\psi)$ for λ at the constrained maximum, but calculated with λ change re-expressed to the metric provided by $\varphi(\theta)$ along the contour for given ψ :

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

where $\varphi_\lambda(\theta) = \partial\varphi(\theta)/\partial\lambda$ is $p \times (p - 1)$ and we are using the notation $|X| = |X'X|^{1/2}$ for an $n \times p$ matrix, here with $p = 2$. 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 the φ -rescaled version of λ designated as (λ) ; they are then rescaled to other parametrizations as needed; some details are given in the Appendix Section 8(ii).

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 factor²

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

²see the Appendix. (If we started with (16) perhaps this could be simplified because there is a $j_{\varphi\varphi}$ already in the formula at $\hat{\theta}$, so we just need to stick in a $j_{\varphi\varphi}(\hat{\theta}_\psi)$? Also there is an implicit $j_{(\lambda\lambda)}$ in the φ_λ term.)

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

The right factor in (17) can be rewritten giving

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

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

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

$$\begin{aligned}\pi(\hat{\theta}_\psi)d\psi d\lambda &= |j_{(\psi\psi)\cdot\lambda}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|d(\psi)d(\lambda) \\ &= |j_{\varphi\varphi}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}d(\psi)d(\lambda_\psi).\end{aligned}\quad (18)$$

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 equivalently the Laplace integration using $|j_{(\lambda\lambda)}(\psi, \hat{\lambda}_\psi)|^{1/2}d(\lambda)$ at the intersection with the profile curve \mathcal{C}_ψ just produces $(2\pi)^{1/2}$ times the profile value at ψ . Accordingly we replace the information metric $|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}d(\lambda_\psi)$ at the intersection with \mathcal{C}_ψ by the extended version applicable on the full contour as an equivalent of Laplace at \mathcal{C}_ψ ; this gives the following general expression for the default prior:

$$\begin{aligned}\pi_\psi(\theta)d\psi d\lambda &= c \frac{\ell_\psi(\hat{\theta}_\psi)}{\hat{\chi} - \hat{\chi}_\psi} d\psi \cdot |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\theta)|^{1/2}d\lambda \\ &= |j_{\varphi\varphi}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\theta)|^{1/2}d(\psi)d(\lambda_\psi).\end{aligned}\quad (19)$$

Example 5.1. Normal (μ, σ^2) . In Example 3.2 we obtained the default prior for the full parameter $\theta = (\mu, \sigma^2)$; we now illustrate the preceding targetted default prior by examining the components μ and σ^2 . The canonical parameter is $(\varphi_1, \varphi_2) = (\mu/\sigma^2, 1/\sigma^2)$ which has information function

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

$$= n \begin{pmatrix} \sigma^2 & -\mu\sigma^2 \\ -\mu\sigma^2 & \sigma^4/2 + \mu^2\sigma^2 \end{pmatrix},$$

and thus 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 $j_{\tilde{\varphi}\tilde{\varphi}}^0 = I$ with information function

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

and Jeffreys prior

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

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

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

Thus on \mathcal{C}_μ 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 off \mathcal{C}_μ by the constant information metric for σ^2 gives

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

this is the familiar right invariant measure.

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

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

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

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

and then the adjusted Jeffreys

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

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

Example 5.2. Normal circle (continued). We saw in Example 3.2 that the default prior for the vector $\varphi = (\psi \cos \alpha, \psi \sin \alpha)$ did not correctly target the component parameter ψ . Now, for inference about ψ , we have the following components of the targetted prior (18):

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

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).

Suppose³ 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

³I think the material that follows might be the additional explanation for footnote 5. But if we can get away without the square brackets that would be good.

parameter space has a Welch & Peers (1963) inversion to the sample space that uses the corresponding score variable s_2 at y^0 with differential

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

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

$$|j_{(\lambda\lambda)}(\psi, \hat{\lambda}_\psi)|^{1/2} |j_{[\lambda\lambda]}(\psi, \lambda)|^{1/2} d\lambda; \quad (20)$$

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

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

The adjusted nuisance Jeffreys prior (20) produces marginal likelihood for ψ , which then appears as an appropriately adjusted profile likelihood for that parameter of interest. This can then be integrated following the Welch-Peers pattern using root profile information obtained from the exponential parametrization. This gives the Jeffreys type adjustment

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

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

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

for use with the full likelihood $L(\psi, \lambda)$; this is in agreement with (19).

6 Targetted default priors: vector components

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

The parameter effects matrix $\tilde{W}(\theta)$ at (7) can be partitioned in accord with the components ψ and λ giving $\tilde{W}(\theta) = \{\tilde{W}_\psi(\theta), W_\lambda(\theta)\}$ so that

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

We consider a parameter value $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$ on the profile curve \mathcal{C}_ψ formed by the constrained maximum likelihood values. A change $d\lambda$ in λ with ψ fixed generated a $p - d$ -dimensional tangent plane $\mathcal{T}_\psi = \mathcal{L}\{\tilde{W}_\lambda(\hat{\theta}_\psi)\}$ in $\{\hat{j}^{1/2}d\hat{\theta}\}$ at the observed $\hat{\theta}^0$. Now consider a change $d\psi$ in ψ with an appropriate corresponding change in λ so that the consequent change in $\{\hat{j}^{1/2}d\hat{\theta}\}$ at the observed $\hat{\theta}^0$ is perpendicular to \mathcal{T}_ψ ; let $\tilde{W}_{\psi,\lambda}d\psi$ designate this change with the column vectors in $\tilde{W}_{\psi,\lambda}$ being those of \tilde{W}_ψ after orthogonalization to \mathcal{T}_ψ . Then

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

with $d\psi$ and $d\lambda$ interpreted as just described.

We now follow the pattern in the preceding section and have the Welch-Peers effect of data change for λ with given ψ as $|\tilde{W}_\lambda(\psi, \lambda)|d\lambda$ and the Welch-Peers effect of data change in ψ along the profile curve $\mathcal{C}_\psi = \{\hat{\theta}_\psi\}$ as $|\tilde{W}_{\psi,\lambda}(\hat{\theta}_\psi)|d\psi$. This gives the composite targetted prior

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

where $|\psi|$ is measured along the profile curve \mathcal{C}_ψ . When ψ however is a vector this prior would target linear parameters and the Dawid Stone Zidek effect would be possible with nonlinearity.

Example 6.1: 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

$$W_\psi(\theta) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad W_\lambda(\theta) = \begin{pmatrix} 0 & (\hat{\beta}_1^0 - \beta_1)/\sigma^2 \\ 0 & (\hat{\beta}_2^0 - \beta_2)/\sigma^2 \\ 1 & (\hat{\beta}_3^0 - \beta_3)/\sigma^2 \\ 0 & \hat{\sigma}^2/\sigma^2 \end{pmatrix}. \quad (21)$$

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

$$\begin{aligned} |W_\psi(\hat{\theta}_\psi)| &= 1, \\ |W_\lambda(\theta)| &= \left| \begin{array}{cc} 1 & (\hat{\beta}_3^0 - \beta_3)/\sigma^2 \\ (\hat{\beta}_3^0 - \beta_3)/\sigma^2 & \{\hat{\sigma}^4 + \sum_1^3 (\hat{\beta}_i^0 - \beta_i)^2\} \sigma^{-4} \end{array} \right|^{1/2} \\ &= \left(\frac{\hat{\sigma}^4 + \sum_1^3 (\hat{\beta}_i^0 - \beta_i)^2}{\sigma^4} \right)^{1/2}. \end{aligned}$$

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 3.1.

7 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.

Appendix

(i) Transformation models

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

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

left invariant measure.

The notation is simplified if the group coordinates are centered so that the identity element is at the maximum density point of the conditional error density; thus $g(z) \leq \tilde{g}(e)$ where e designates the identity element satisfying $ez = z$. The maximum likelihood group element $\hat{\theta}(y)$ is then the solution of $\theta^{-1}y = e$ which gives $\hat{\theta}(y) = y$. We then have from (6) that the default prior is

$$|W(\theta)|d\theta = \left| \frac{dy}{d\theta} \right|_{y^0} d\theta = \left| \frac{d(\theta z)}{d\theta} \right|_{\theta z=y^0} d\theta \quad (22)$$

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

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

using the right invariance of ν . This 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 default prior (22) is $\pi(\theta)d\theta = cd\nu(\theta)$.

(ii) Information and the approximating exponential model

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 parametrization 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:

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

(iii) Strong matching and information approximation

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

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

where $d\beta(\theta)$ is a locally defined linear parameter.

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

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

whereas the usual Jeffreys' prior is

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

these are asymptotically equivalent.

(iv) 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 $\psi = \psi_0$ 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 ψ is curved but the angle α is linear.

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

We now examine the case with scalar θ_1 and scalar θ_2 and 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 ,

$$\begin{aligned} d\hat{\theta}_1 &= w_{11}(\theta)d\theta_1 + w_{12}(\theta)d\theta_2 \\ d\hat{\theta}_2 &= w_{21}(\theta)d\theta_1 + w_{22}(\theta)d\theta_2; \end{aligned} \tag{23}$$

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 (23) 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 (23) to determine the parameter space increment at a general point θ that corresponds to the preceding sample space increment, giving

$$d\theta = \begin{pmatrix} w^{11}(\theta)c_1 + w^{12}(\theta)c_2 \\ w^{21}(\theta)c_1 + w^{22}(\theta)c_2 \end{pmatrix} dt, \quad (24)$$

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

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

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

$$\begin{aligned} d\hat{\mu} &= d\mu + (\hat{\mu}^0 - \mu)d\sigma/\sigma \\ d\hat{\sigma} &= \hat{\sigma}^0 d\sigma/\sigma. \end{aligned} \quad (25)$$

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

$$\begin{aligned} d\hat{\mu} &= d\mu - \mu d\sigma/\sigma \\ d\hat{\sigma} &= d\sigma/\sigma. \end{aligned} \quad (26)$$

Inverting this gives

$$\begin{aligned} d\mu &= d\hat{\mu} + \mu d\hat{\sigma} \\ d\sigma &= \sigma d\hat{\sigma}. \end{aligned} \tag{27}$$

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 (26) we have $d\hat{\mu} = d\mu$ and $d\hat{\sigma} = d\sigma$ at this point, which gives $d\hat{\mu} = -kd\hat{\sigma}$. Finally we determine what the restriction $d\hat{\mu} = -kd\hat{\sigma}$ on the sample space implies for $(d\mu, d\sigma)$ at a general point on the parameter space; from (27) 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 8.2. For the normal circle Example 3.2 with parameter of interest $\psi = (\theta_1^2 + \theta_2^2)^{1/2}$, the increment on the parameter space at $\hat{\theta}^0$ with fixed ψ satisfies $d\theta_1 = -\tan \hat{\alpha}^0 d\theta_2 = -(y_2^0/y_1^0)d\theta_2$. This then translates to the sample space at (y_1^0, y_2^0) using the specialized version of (23) to give $dy_2 = -(y_2^0/y_1^0)dy_1$, and this then translates back to a general point on the parameter space using the specialized version of (24) 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.

(v) Details for gamma example

The model is $f(y_i; \theta) = \Gamma^{-1}(\theta)y_i^{\theta-1}\exp(-y_i)$, and the three priors considered are the Jeffreys prior, $\pi_J(\theta) \propto i^{1/2}(\theta)$ where $i(\theta)$ is the expected Fisher information, $\pi_V(\theta)$, the default prior proposed at (6), with V_i defined using (1), and the strong

matching prior (Fraser & Reid, 2002) defined by $\pi_S(\theta) \propto \ell'(\theta)/\ell_{;V}(\theta)$. The calculations are as follows:

$$\begin{aligned}\ell(\theta; y) &= \theta \sum \log y_i - n \log \Gamma(\theta) = \theta t - n\psi(\theta) \\ i(\theta) &= -\ell''(\theta) = n\psi''(\theta) \\ \ell'(\theta; y^0) &= t^0 - n\psi'(\theta)\end{aligned}$$

and

$$\begin{aligned}V_i(\theta) &= -\frac{F_\theta(y_i^0; \theta)}{F_y(y_i^0; \theta)} \\ &= \frac{\int_0^{y_i^0} \Gamma^{-1}(\theta) z^{\theta-1} e^{-z} dz}{\Gamma^{-1}(\theta) (y_i^0)^{\theta-1} e^{-y_i^0}} \\ &= \frac{\Gamma'(\theta) F(y_i^0; \theta) - \int_0^{y_i^0} z^{\theta-1} \log(z) e^{-z} dz}{e^{-y_i^0} (y_i^0)^{\theta-1}}\end{aligned}$$

which was computed in R by numerical integration. The prior given by (6) is

$$\pi_V(\theta) \propto \left(\sum V_i^2(\theta) \right)^{1/2}.$$

Using $V = V(\hat{\theta}^0)$ to compute the strong matching prior gives

$$\begin{aligned}d\beta(\theta) &\propto -\frac{\ell'(\theta; y^0)}{\ell_{;V}(\theta; y^0) - \ell_{;V}(\hat{\theta}^0; y^0)} \\ &= \frac{t^0 - n\psi'(\theta)}{\theta \sum_{i=1}^n V_i(\hat{\theta}^0)/y_i^0} \\ &= \frac{1}{\theta - \hat{\theta}^0} \frac{n\{\psi'(\theta) - \psi'(\hat{\theta}^0)\}}{V_i(\hat{\theta}^0)/y_i^0}\end{aligned}$$

References

- [1] Barndorff-Nielsen, O.E. (1986).
- [2] Bédard, M., Fraser, D.A.S. and Wong, A.C.M. (2007). Higher accuracy for Bayesian and frequentist inference: large sample theory for small sample likelihood. *Biometrika* **22**, 300-321.
- [3] Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113-147 (with discussion).

- [4] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.
- [5] Datta, G.S. and Ghosh, M. (1995). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90**, 1357–1363.
- [6] Davison, A.C., Fraser, D.A.S. and Reid, N. (2006). Likelihood inference for categorical data. *J. R. Statist. Soc B* **68**, 495–508.
- [7] 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.
- [8] DiCiccio, T.J. et al (1995)
- [9] DiCiccio and Martin (1993)
- [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., Reid, N. and Wu, J. (1999).
- [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., Reid, N. and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–264.
- [16] George, E.I. and McCulloch.
- [17] Jeffreys, H. (1939). *Theory of Probability*. Oxford: Oxford University Press. Third Edition (1961).
- [18] Kass, R. and Wasserman, L. (1995).
- [19] Kass, R.E. (1990). Data-translated likelihood and Jeffreys’s rules. *Biometrika* **77**, 107–114.

- [20] Mukerjee, R. et al.
- [21] Richardson and Green
- [22] Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.* **30**, 877–880.
- [23] Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 705–708.
- [24] Tierney and Kadane (1986)
- [25] 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.
- [26] Zellner, A. (1988). Optimal information processing and Bayes' theorem. *Amer. Statist.* **42**, 278–284.