Submitted to the Brazilian Journal of Probability and Statistics
arXiv: `math.PR/0000000`

On default priors and approximate location models

D.A.S. Fraser[a,1] **and N. Reid**[a,1]

[a] *Department of Statistics, University of Toronto*

Abstract. A prior for statistical inference can be one of three basic types: a mathematical prior originally proposed in Bayes (1763), a subjective prior presenting an opinion, or a truly objective prior based on an identified frequency reference. In this note we consider a method for deriving a mathematical prior based on approximate location models. This produces a mathematical posterior, and any practical interpretation of such a posterior is in terms of exact or approximate confidence under the postulated model. We describe how a proposed prior can be simply checked for consistency with confidence methods, using expansions about the maximum likelihood estimator.

1 Introduction

Priors have been used for more than two centuries (Bayes, 1763) to weight an observed likelihood function, thus providing a prominent procedure for statistical inference. As such it is arguably the oldest formalized methodology in statistics, and long predates the formal introduction of the likelihood function (Fisher, 1922). The procedure provides a rich methodology for examining a statistical model with observed data and has had profound effects in liberalizing statistical analyses. It does give rise to statements presented as probabilities, in contrast to the confidence procedure of Fisher (1930) and Fraser (2003) that gives rise to statements labeled as confidence. Fraser (2010a) showed that the two procedures lead to the same result for scalar parameters only in location models; he then argued that this meant that confidence was not entitled to be labeled as probability, although he could equally have argued that his analysis showed that Bayesian inference was not entitled to be labeled as confidence.

As an example of Lindley's (1958) analysis, if we have an observed datum y^0 from a Normal $(\theta, 1)$ distribution, the confidence 95% lower bound is $y^0 - 1.64$, and the posterior 95% lower bound is $y^0 - 1.64$ under the prior $\pi(\theta)d\theta \propto d\theta$; this prior represents the translation invariance of the location

[1] Supported by NSERC

AMS 2000 subject classifications. Primary 62F15; secondary 62E20

Keywords and phrases. Conditioning, Confidence, Default prior, Jeffreys prior, Noninformative prior, Objective prior, Reference prior, Subjective prior

model. Similarly, the observed p-value and the posterior survivor value are the same for all θ and given by $\Phi(y^0 - \theta)$, where Φ is the standard normal distribution function. Of course the prior is not proper and thus cannot represent probabilities.

The probability lemma for calculating conditional probability takes two probability inputs and produces one probability output. If one input however is absent and a convenient mathematical object is substituted, then the conditions of the lemma are not fulfilled and the output is not a probability on the basis of the lemma. It may have attractive properties by other arguments; indeed it is our view that translation invariance as represented in a prior does widely give confidence, of first or sometimes higher order accuracy. These arguments are expanded on in Fraser (2010a,b,c). One conclusion is that the Bayes initiative was prescient, a very early and profound initial step towards confidence (Fisher, 1930), leaving fine tuning over some two hundred years to clarify the concept.

In Section 2 we describe the location invariance, which was in fact used by Bayes (1763) in an augmented analogy, and then describe approximate translation invariance based on model continuity (Fraser et al., 2010c). Priors with this invariance were called default priors in Fraser et al. (2010c), and lead to second order confidence for linear parameters. Section 3 discusses necessary and sufficient conditions for a prior to have approximate default properties. In Section 4 we show that even with default priors, confidence statements for curved parameters will be accurate only to first order, and describe some related work.

2 Location models and the location relationship

A location model $f(y - \theta)$, where we initially suppose the variable y and parameter θ are scalar, has the property that a displacement a to y and a parallel displacement a to θ leaves the statistical model unchanged. Conversely, if a transformation as just described produces a new variable and new parameter with an unchanged probability distribution, then the model is a location model. Let $z = y - \theta$ and let z_β be the β-quantile of the $f(z)$ distribution. Then the confidence inversion of the interval $(-\infty, z_\beta)$ produces the β-confidence lower bound $y - z_\beta$. And in a related way the prior $\pi(\theta)d\theta \propto d\theta$ having location invariance gives the posterior lower bound $y - z_\beta$ with a claimed valuation β; again these are equal. The two approaches are of course familiar but the unifying theme perhaps less so; for further discussion, see Fraser & Reid (2002).

A similar result holds for a p-dimensional variable y and p-dimensional

parameter θ having a location model $f(y-\theta)$. If the parameter of interest ψ is θ_1, then as above the β-confidence lower bound is obtained by inverting the pivot $y_1 - \psi$ giving the β-confidence lower bound $y_1 - z_\beta$. And the marginal posterior lower bound obtained by using the location invariant prior $\pi(\theta)d\theta \propto d\theta$ gives $y_1^0 - z_\beta$, and fully agrees with the β-confidence lower bound. This property does not extend however to parameters of interest ψ that are not linear in θ; an example is given in the §4.

More generally in transformation models, the default prior given by the right invariant measure of the transformation group ensures that posterior probability bounds are the same as confidence bounds for a wide range of parameters, with the caveat, again, that if the parameter of interest is not simply related to the transformation structure, that is, does not have a linearity property, then posterior inference can again disagree with confidence.

Fraser et al. (2010c) described how continuity in a general statistical model can lead to posterior inference that has valid confidence to $O(n^{-1})$, with n the size of an independent, identically distributed sample, or more generally an amount of information. The distribution function for each component of a vector response y can be used to give an approximate location model, and thence to an analogue of the location invariant prior described above; the distribution function provides a link between the parameter and the variable, in the following sense. In a location model for a single variable y with scalar parameter θ, $f(y;\theta) = f(y-\theta)$, we have $F(y+d\theta;\theta+d\theta) = F(y;\theta)$ by the location property, where F is the distribution function. We can also write this as

$$\frac{dy}{d\theta} = -\frac{F_{;\theta}(y;\theta)}{F_y(y;\theta)} = 1,$$

where $F_{;\theta}$ is the derivative of the distribution function with respect to θ. This shows that change in y is compensated by change in θ; the defining feature of a location model. In a non-location model, we can derive a similar result locally:

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

where now y^0 is a fixed point, usually the observed data point. Equivalently, in terms of the quantile function $F^{-1}(u;\theta) = y(u;\theta)$,

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

For a sample of independent observations $y_1,\ldots,y_n$ from the same model we can do this for each observation, giving $y = y(u;\theta) = \{y_1(u_1;\theta),\ldots y_n(u_n;\theta)\}'$,

and express its derivative with respect to θ as a vector of length n:

$$\{V_1(\theta), \ldots, V_n(\theta)\}' = \frac{\partial y}{\partial \theta}\Big|_{y^0} = \frac{\partial y(u;\theta)}{\partial \theta}\Big|_{u_i = F(y_i^0;\theta)}.$$

If θ is a vector of length p, then the same construction can be used, with each $V_i(\theta)$ a row vector of length p:

$$\frac{dy}{d\theta'}\Big|_{y^0} = \begin{pmatrix} V_1(\theta) \\ \vdots \\ V_n(\theta) \end{pmatrix} = V(\theta), \tag{2.1}$$

which is an $n \times p$ matrix that explicitly links change at the observed data with parameter change at various θ values. Fraser et al. (2010c) refer to $V(\theta)$ as the *sensitivity matrix*. In the location model $f(y_i - x_i'\beta)$, $V_i(\beta) = x_i'$, and $V(\beta)$ is simply the design matrix and thus records sample space directions indicated by possible changes in the regression parameter.

More generally the matrix $V(\theta)$ presents sample space directions intrinsic to an approximate location model at the observed data (Fraser et al., 2010b), and then the default prior $\pi(\theta)d\theta \propto |V(\theta)|d\theta$ ensures that posterior probabilities for linear components of θ are equivalent, to second order, to the p-values computed from frequentist methods. It is sometimes more convenient to re-express the relationship at (2.1) in the form,

$$dy = V(\theta)d\theta,$$

and then to derive the corresponding relationship between $\hat\theta$ and θ, where $\hat\theta$ is the maximum likelihood estimator. This is derived in Fraser et al. (2010c) to be

$$d\hat\theta = W(\theta)d\theta, \tag{2.2}$$

at $(y^0, \hat\theta^0)$, where $W(\theta) = j^{-1}(\hat\theta^0)H'(\hat\theta^0; y^0)V(\theta)d\theta$, $j(\theta) = -\partial^2\ell(\theta;y)/\partial\theta\partial\theta'$ is the observed Fisher information, and

$$H'(\hat\theta^0; y^0) = \ell_{\theta;y'}(\hat\theta^0; y^0) = \frac{\partial^2}{\partial\theta\partial y'}\ell(\theta;y)\Big|_{(y^0;\hat\theta^0)}.$$

This leads to the default prior

$$\pi(\theta)d\theta \propto |W(\theta)|d\theta. \tag{2.3}$$

Inference based on this default prior is typically accurate to $O(n^{-1})$, for linear parameters. In models having a right invariant prior it reproduces

that prior; in particular in the regression model $y_i = x_i'\beta + \sigma z_i$ it gives the $\pi(\theta) = d\beta d\sigma/\sigma$, for any fixed distribution for z. Additional examples are given in Fraser et al. (2010c).

The relationship between variable and parameter in (2.3) is rather powerful. It shows that locally there is a location model $f(\hat{\beta} - \beta)$ that agrees with the given model to first derivative at the data value in the sample space, and agrees with the given model in the parameter space, for parameter values in the moderate deviation range, $O(n^{-1/2})$ about the maximum likelihood value. The default prior as defined by this local location model can be viewed as 'flat' in the location parameter coordinates. This suggests that for an arbitrary prior, we can examine its effect on the likelihood by seeing whether or not it agrees with this flat prior, to some order of approximation. If it does not, then the posterior survival probability will typically differ from the p-value, and the posterior intervals will thus not have the claimed probability content under the model. We consider some aspects of this, following Fraser & Reid (1995), Fraser & Sun (2010), Fraser et al. (2010a) and Fraser (2010d). We restrict attention to the scalar parameter case, as the vector parameter case raises some technical difficulties.

In the scalar case we have $d\hat{\theta} = w(\theta)d\theta$ where $w(\theta)$ is a scalar function of θ. We can sometimes integrate the equation directly and obtain the location parameter:

$$\beta \;=\; \int_{\theta_0}^{\theta} w(\theta)d\theta. \tag{2.4}$$

As a simple example consider a sample $y = (y_1, \ldots, y_n)'$ of independent observations from the scale family with density $f(y_i;\sigma) = (1/\sigma)f(y_i/\sigma)$ where the form of f is known. Since $F(y_i;\sigma) = F(y_i/\sigma)$, we have

$$V_i(\sigma) = -F_{;\sigma}(y_i;\sigma)/F_y(y_i;\sigma)|_{y_i^0} = y_i^0/\sigma,$$

where $y^0 = (y_1^0, \ldots, y_n^0)'$ is the observed sample. In quantile form the variable-parameter relationship $y_i = \sigma u_i$ gives the same result. Thus we have the default prior $\pi(\sigma) \propto |V(\sigma)|$, where

$$|V(\sigma)| = \{V(\sigma)'V(\sigma)\}^{1/2} = \frac{1}{\sigma}(\sum y_i^{02})^{1/2},$$

proportional to the usual prior for a scale parameter. Alternatively if we work with $W(\sigma)$ defined at (2.3), we obtain

$$|w(\sigma)| = \frac{\hat{\sigma}^0}{\sigma},$$

where σ_0 is the maximum likelihood estimate. The two forms of the prior are equivalent, as constants of proportionality do not matter.

Now consider the direct integration route to obtain the location relationship. From (4.1) we obtain $\beta = \sigma_0(\log\sigma - \log\sigma_0)$; this produces the logarithmic reparameterization as one might expect.

To gain more insight into the equation $d\hat{\theta} = w(\theta)d\theta$, we consider the following expansion. Let $w = w(\theta_0)$ and $c = w'(\theta_0)$; a second order approximation for the location parameter is given by

$$\beta = \int_{\theta_0}^{\theta} \{w + c(\theta - \theta_0)\}d\theta = \theta_0 + w(\theta - \theta_0) + c(\theta - \theta_0)^2/2. \tag{2.5}$$

where often $w = 1$ which would simplify the final expression. This exact or approximate location reparameterization gives an approximating second order quantile presentation of the model as $\hat{\beta} = \beta + z$ where z has the fixed distribution describing the location residual $\hat{\beta} - \beta$.

3 Is a prior second-order default?

For the scalar case we now discuss a simple test criterion for whether a proposed prior has the default properties found with the flat prior for the location model context.

For this, the basic input is of course the log-likelihood function $\ell^0(\theta) = \log f(y^0;\theta)$. To work easily with this likelihood we choose convenient coordinates that are centered at the observed maximum likelihood value and scaled with respect to observed information; specifically we take the new parameter coordinate to be $\tilde{\theta} = \hat{\jmath}_{\theta\theta}^{1/2}(\theta - \hat{\theta}^0)$ where $\hat{\jmath}_{\theta\theta}$ is the observed information. In these new coordinates the likelihood function has the simplified form

$$\ell(\tilde{\theta}) = -\tilde{\theta}^2/2 - \alpha_3\tilde{\theta}^3/6n^{1/2} + O(n^{-1}),$$

where $\alpha_3/n^{1/2}$ is the negative third derivative of likelihood at the maximum with respect to the standardized coordinates and the dependence on sample size has been made explicit.

Now expanding the log-prior in the standardized parameter gives

$$\log\pi(\tilde{\theta}) = a_0 + a_1\tilde{\theta}/n^{1/2} + O(n^{-1});$$

and then combining likelihood and prior gives the posterior

$$\pi(\tilde{\theta};\text{data}) = k\exp\{-(\tilde{\theta} - a_1/n^{1/2})^2/2 - \alpha_3(\tilde{\theta} - a_1/n^{1/2})^3/6n^{1/2}\}$$

to $O(n^{-1})$, where terms of $O(n^{-1})$ have been introduced to simplify the resulting expression. This shows that the prior to this order causes a translation by the amount $a_1/n^{1/2}$ on the likelihood in standardized coordinates. Do we actually want a prior hopefully not informative to be displacing the location likelihood information provided by the data?

If the initial parameterization θ is a linear and if we want the related location model to have a flat prior with slope or log-slope equal to zero then we would want $a_1/n^{1/2} = 0$ to the order $O(n^{-1})$. And we see of course that such a prior does not shift the likelihood information.

Alternatively however suppose that θ is not linear. Then, following §3, we can calculate the curvature c from the relationship $d\hat{\theta} = w(\theta)d\theta$. In the standardized parameterization $\tilde{\theta}$ we have $w(\tilde{\theta}) = w + c\tilde{\theta}/n^{1/2}$ which integrates to give the linear parameterization $\beta = w\tilde{\theta} + c\tilde{\theta}^2/2n^{1/2}$.

To obtain the prior re-expressed in the linear parameterization β we first calculate

$$\frac{d\beta}{d\tilde{\theta}} = w + c\tilde{\theta}/n^{1/2} = w\exp\{c\tilde{\theta}/wn^{1/2}\}$$

to second order, and then use use its inverse to adjust the prior to be relative to $\tilde{\theta}$. This gives the reexpressed log-prior relative to the linear parameterization as

$$\tilde{a}_0 + a_1\tilde{\theta}/n^{1/2} - (c/w)\tilde{\theta}/n^{1/2},$$

and leads to the following posterior expressed in terms of the linear parameterization

$$k\exp\{-(\tilde{\theta} - a_1^*/n^{1/2})^2/2 - \alpha_3(\tilde{\theta} - a_1^*/n^{1/2})^3/6n^{1/2}\},$$

where $a_1^* = a_1 - c/w$. Thus to have a prior that is flat relative to the location parameterization, and therefore does not shift the likelihood, we need to have $a_1^* = 0$, that is, $a_1 = c/w$, which in turn means the prior must take account of the first derivative of $w(\theta)$, or more generally the first derivative array of the sensitivity matrix $V(\theta)$.

In the scalar parameter case Welch & Peers (1963) showed that posterior quantiles had correct frequentist coverage to $O(n^{-1})$ if the prior is proportional to $i^{1/2}(\theta)$, where $i(\theta)$ is the expected Fisher information function in a single observation from the model, so in a sense the calculations above might rarely be needed. However, the requirement that $a_1 = c/w$ is weaker, as all arguments here are carried out locally near the maximum likelihood point. They might also be relevant in a setting where nuisance parameters were handled by combining a profile or adjusted profile likelihood with a prior for the parameter of interest only, as in Fraser et al. (2003).

4 Discussion

In the preceding section we considered local properties of a prior that is flat in a locally defined location parameter. When θ is a vector, the location parameter $\beta(\theta)$ is not available explicitly, as at (2.4), but is available implicitly (Fraser & Yi, 2003). However, an expansion of β similar to (2.5) is available, where $W(\theta)$ is a $p \times p$ matrix, and the analogues of w and c are a $p \times p$ matrix and a $p \times p \times p$ array, respectively. The structure of these coefficients is described in Fraser et al. (2010a), and the use of this expansion for constraining priors is investigated in Fraser (2010d).

The default prior, proportional to $|W(\theta)|$, can be constructed, as in Fraser et al. (2010c), but posterior marginal inference for component parameters will only be well-calibrated, that is agree with the p-value, to $O(n^{-1})$, if the component parameter is linear in the underlying (approximate) location parameter. To illustrate this in simple form, assume that (y_1, y_2) is a mean of n observations from the location normal distribution on the plane, and assume the parameter of interest is $\psi(\theta) = \theta_1 + k\theta_2^2/2n^{1/2}$. A contour of the parameter ψ crosses the θ_1 axis at right angles and with positive k bends to the left above and below this axis. At the observed data point $y^0 = (0, 0)'$, the constrained maximum likelihood estimate of θ_1 is given ψ is $\hat{\theta}^0_\psi = (\psi, 0)'$ neglecting terms in the likelihood equations of $O(n^{-1})$. Consider evaluating ψ with the statistic $t(y) = y_1 + ky_2^2/2n^{1/2}$, which has the same contours on the sample space as does ψ on the parameter space. This function is asymptotically normally distribution with mean ψ and variance 1. The p-value based on the marginal distribution of t is the probability left of the $t(y)$ contour through y^0, using the normal distribution with mean ψ and variance 1. The posterior survivor function is the probability to the right of the $\psi(\theta)$ contour, using the bivariate normal distribution with mean $y^0 = (0, 0)$ and identity covariance. These are equal when the curvature $k = 0$, but as k increases, the region of integration for the p-value decreases and the region of integration for the s-value increases. Thus the Bayes s-value will be mis-calibrated by a term that is $O(n^{-1/2})$, in contrast to the reproducibility inherent in the p-value. The example above is more general than it may appear, as it captures first order deviations from limiting normal distribution.

Some aspects of this discrepancy were discussed as part of the marginalization paradox (Dawid, 1973), but identifying curvature as an intrinsic cause is more recent (Fraser et al., 2010a; Fraser & Reid, 2002; Fraser & Sun, 2010).

Thus although Bayesian inference generally gives confidence statements

to a first order of approximation, and with the default prior leads to confidence statements to second order, as described above, more generally inference based on a prior comes with risks not often easily assessed. For some discussion including examples of the curvature discussed in Section 3 see Fraser (2010a,b,c). For some examples where Bayes and confidence methods routinely give identical results see Bedard (2008). And for three enigmatic examples where Bayesian inference has progressively greater difficulty in achieving an analysis, see Fraser et al. (2009).

References

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.

Bédard, M., Fraser, D.A.S. and Wong, A. (2008). Higher accuracy for Bayesian and frequentist inference: Large sample theory for small sample likelihood. *Statistical Science* **22**, 301-321.

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.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Royal Soc.* London A 222, 309-368.

Fisher, R. A. (1930). Inverse probability. *Proc. Camb. Phil. Soc.* **26**, 528-535.

Fraser, D.A.S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327-339.

Fraser, D.A.S.(2010a). Is Bayes posterior just quick and dirty confidence? *Statistical Science*, to appear.

Fraser, D.A.S. (2010b). Bayesian analysis or evidence based statistics. *International Encyclopedia of Statistical Science (Springer)*, to appear.

Fraser, D.A.S. (2010c). Bayesian inference: An approach to statistical Iinference. *Wiley Interdisciplinary Reviews: Computational Statistics*, to appear.

Fraser, D.A.S. (2010d). Default priors: Measuring Bayes bias for interest parameters. Manuscript.

Fraser, A.M. Fraser, D.A.S. and Fraser, M.J. (2010a). Parameter curvature revisited and the Bayesian frequentist divergence. *J. Statist. Research: Efron volume*, **44**, to appear.

Fraser, A.M., Fraser, D.A.S. and Staicu, A.-M. (2010b). Second order ancillary: A differential view with continuity. *Bernoulli*, **16**, 1208-1223.

Fraser, D.A.S., and Reid, N. (1995). Ancillaries and third order significance. *Utilitas Mathematica* **47**, 33-53.

Fraser, D.A.S., and Reid, N. (2002) Strong matching of frequentist and Bayesian parametric inference. *J. Statist. Plann. Infer.*. **103**, 263-285.

Fraser, D.A.S., Reid, N., Marras, E., and Yi, G.Y. (2010c). Default priors for Bayesian and frequentist inference. *J. Roy. Statist. Soc. B*, **75**, 631-654.

Fraser, D.A.S., Reid, N., Wong, A., and Yun Yi, G.(2003). Direct Bayes for interest parameters. *Valencia 7*, 529-533.

Fraser, D.A.S., Wong, A. and Sun, Ye (2009). Three enigmatic examples and inference from likelihood. *Canad. J. Statist.*, **37**, 161-181.

Fraser, D.A.S. and Sun, Y. (2010). Some corrections for Bayes curvature. *Pakistan J. Statist.*; **25**, 351-370.

Fraser, D.A.S. and Yi, G. Y. (2003). Location reparameterization and default priors for statistical analysis. *J. Iranian Statist. Soc.* **1**. 55-78.

Lindley, D.V. (1958). Fiducial distribution and Bayes theorem. *J. Roy. Statist. Soc.* B **20**, 102-107.

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

Department of Statistics,
University of Toronto,
100 St. George St.,
Toronto, Canada M5S 3G3
E-mail: dfraser@utstat.toronto.edu; reid@utstat.toronto.edu