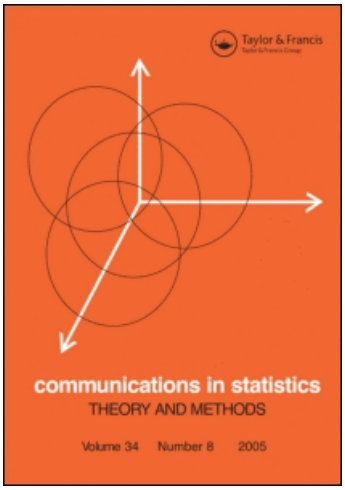

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

Assessing Sensitivity to Priors Using Higher Order Approximations

N. Reid [a];Y. Sun [a]
[a] Department of Statistics, University of Toronto, Toronto, Ontario, Canada

Online publication date: 28 April 2010

To cite this Article Reid, N. andSun, Y.(2010) 'Assessing Sensitivity to Priors Using Higher Order Approximations', Communications in Statistics - Theory and Methods, 39: 8, 1373 — 1386
To link to this Article: DOI: 10.1080/03610920802401138
URL: http://dx.doi.org/10.1080/03610920802401138

Communications in Statistics—Theory and Methods, 39: 1373–1386, 2010

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920802401138

Taylor & Francis
Taylor & Francis Group

Assessing Sensitivity to Priors Using Higher Order Approximations

N. REID AND Y. SUN

Department of Statistics, University of Toronto,
Toronto, Ontario, Canada

Higher order likelihood methods lead to an easily implemented and highly accurate approximation to both joint and marginal posterior distributions. This makes it quite straightforward to assess the influence of the prior, and to assess the effect of changing priors, on the posterior quantiles. We discuss this in the light of some simple examples that illustrate in concrete form the potential for marginal posterior densities from seemingly uninformative priors to be poorly calibrated.

Keywords Bayesian inference; Laplace approximation; Matching priors; Posterior quantiles; *p*-Values.

Mathematics Subject Classification Primary 62F15; Secondary 62E20.

1. Introduction

We consider inference in parametric models based on the likelihood function. While inference in a scalar parameter model is relatively straightforward, the incorporation of nuisance parameters can make non-Bayesian inference more complicated. In particular, it can be difficult to find a marginal or conditional likelihood function that depends only on the parameter of interest and yet retains all the information in the data. Bayesian inference is, in principle, much simpler, as the posterior marginal distribution of the parameter of interest is taken to provide a complete summary of the evidence in the data (and prior) about this parameter. Although high-dimensional integration may be needed to compute this marginal posterior distribution, this is a computational problem that has largely been solved by the development of Markov chain Monte Carlo (MCMC) methods, which are now very widely used in applications of Bayesian inference. Very accurate approximations, based on the Laplace expansion of the posterior, can also be used to bypass needed high-dimensional integrals.

Received May 4, 2008; Accepted August 11, 2008
With best wishes to Professor Akahira.
Address correspondence to N. Reid, Department of Statistics, University of Toronto, Toronto, Ontario M5S 3G3, Canada; E-mail: reid@utstat.utoronto.ca

Thus, given a model, the main difficulty is the choice of a prior density, and this is arguably increasingly difficult as the dimension of the parameter increases. One approach to constructing priors is to use posterior distributions from earlier related experiments or data sets: this may be quite useful in repeated applications of similar inference, such as might arise in on-line quality control experiments. A quite different approach is the elicitation of subjective priors, from an expert or group of experts, using systematic methods to construct priors that are consistent and capture the main features of prior opinions in the form of a probability distribution. This might be quite useful in certain decision-making contexts, particularly if the subjectivity involved is not an issue.

In many scientific contexts, however, these two approaches may be either not available or satisfactory, and some more *ad hoc* approach is needed. Development of classes of priors variously called *objective*, *noninformative*, *vague*, and so on has been considered. The general idea is to construct a somewhat arbitrary prior using a rule that can be expected to give good posterior inference. A good posterior density will be, first, a proper posterior, i.e., a density that integrates to 1, and proper posteriors can often result from improper priors. Beyond that requirement there does not seem to be a consensus on the meaning of *good*. One approach that is rather well developed is the approach of reference posteriors developed by Berger and Bernardo (1992). Very roughly speaking, a reference posterior is constructed from a prior that maximizes the Kullback–Liebler distance from the prior to the posterior; the prior that achieves this is called a *reference prior*. Another approach, developed by Welch and Peers (1963), is to find a prior that leads to posterior quantiles that are also confidence bounds under the model, either exactly or to a high order of approximation. Such priors are usually called *matching priors* in the literature. A survey of matching priors is given in Datta and Mukerjee (2004). Approaches based on information distances are discussed in Clarke and Wasserman (1995).

All these approaches have the distinct drawback that in a multi-parameter model, the prior needs to be targeted on the parameter of interest, and different priors for the full parameter are needed, depending on which parameter is of interest. This targeting seems to be unavoidable and is related to the marginalization paradox of Dawid et al. (1973). An exception to this requirement of targeting arises in the location-scale model, where, as Peers (1965) pointed out, the conditions for matching are satisfied simultaneously for both location and scale parameters. Peers (1965) provided a condition under which this may be expected to hold. The practical consequence of this is that a single prior for a vector parameter may not provide well-calibrated inference for a particular parameter of interest. Alternatively, if the inference is well calibrated for one parameter, it may not be for another. (We use the term *well-calibrated* interchangeably with probability matching, to mean that the probability associated with the posterior marginal distribution function has an interpretation as a p-value in repeated sampling from the model.) This is well known in the literature on the development of noninformative or objective priors but often seems to be overlooked or ignored in applied work, where it is very common to assign a flat prior to a vector parameter and proceed to marginalize to various parameters of interest. A recent example is Gelman et al. (2007).

In this article we illustrate aspects of this lack of calibration on some relatively simple and well-studied examples. We also show that higher order approximations are a very quick and accurate method for computing posterior distribution functions and for checking the sensitivity of the posterior to the choice of prior.

In Sec. 2 we present the higher order approximation to the marginal posterior distribution function and briefly describe its derivation. In Sec. 3 we study in detail the simple, but instructive, example of inference for the length of the mean vector of a multivariate normal. In Sec. 4 we present two logistic regression examples. We find that the marginal posterior distribution of the ED_{50} in simple logistic regression is not well calibrated when flat priors are used for the regression coefficients. We conclude with a brief discussion.

2. Approximate Posterior Distributions

Let $Y = (Y_1, \ldots, Y_n)$ be a random vector of independent, identically distributed observations from a density $f(y \mid \theta)$, where $\theta = (\theta_1, \ldots, \theta_k)'$ is a vector of unknown parameters, and Y_i is a scalar random variable. We write $\ell(\theta; y) = \log f(y; \theta)$ for the log-likelihood function based on the sample $y = (y_1, \ldots, y_n)$ and denote by $\hat{\theta}$ the maximum likelihood estimate, assumed to be obtained by solving the score equation

$$\partial \ell(\theta; y)/\partial \theta = 0.$$

The *observed Fisher information* function is denoted by $j(\theta) = -\partial^2 \ell(\theta; y)/\partial\theta\, \partial\theta'$; it is a $k \times k$ matrix, and the dependence on y is suppressed in the notation. The *expected Fisher information* is $i(\theta) = E\{j(\theta)\}$, where the expectation is under the model $f(y; \theta)$. The matrix $i(\theta)$ is $O(n)$, and $j(\theta)$ is $O_p(n)$, in i.i.d. sampling. Finally, we denote the prior by $\pi(\theta)$, and the posterior by $\pi(\theta \mid y)$:

$$\pi(\theta \mid y) = \frac{\exp\{\ell(\theta; y)\}\pi(\theta)}{\int \exp\{\ell(\theta; y)\}\pi(\theta) d\theta}. \tag{1}$$

Often the normalizing constant for the posterior need not be computed explicitly, as the inference depends on ratios of the posterior at different values of θ.

For inference about component parameters we use the general notation $\theta = (\psi, \lambda)$, where ψ is the parameter of interest, usually scalar, and λ is a nuisance parameter. It is convenient to consider ψ as a component of θ, although this is not strictly necessary and may instead represent a constraint on θ. Inference about ψ is based on the marginal posterior density

$$\pi_m(\psi \mid y) = \frac{\int \exp\{\ell(\theta; y)\}\pi(\theta) d\lambda}{\int \exp\{\ell(\theta; y)\}\pi(\theta) d\theta}. \tag{2}$$

Exact integration of the numerator is rarely possible and difficult to compute numerically if the dimension of λ is greater than about 3. Markov chain Monte Carlo methods bypass this integration by obtaining samples from $\pi_m(\psi \mid y)$ using simulation techniques.

When ψ is a scalar, Laplace approximation of integrals in (1) and (2) leads to the following approximations (Tierney and Kadane, 1986):

$$\pi(\theta \mid y) \doteq \frac{1}{\sqrt{(2\pi)^{k/2}}} \exp\{\ell(\theta) - \ell(\hat{\theta})\} |j(\hat{\theta})|^{1/2} \frac{\pi(\theta)}{\pi(\hat{\theta})} \tag{3}$$

and

$$\pi_m(\psi \mid y) \doteq \frac{1}{\sqrt{(2\pi)}} \exp\{\ell(\hat{\theta}_\psi) - \ell(\hat{\theta})\} \frac{|j(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}} \frac{\pi(\hat{\theta}_\psi)}{\pi(\hat{\theta})}. \tag{4}$$

In (4) we have used the notation $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$, where $\hat{\lambda}_\psi$ is the constrained maximum likelihood estimate of λ for fixed ψ, assumed to satisfy the equation $\partial \ell(\theta)/\partial \lambda = 0$. The function $\ell_p(\psi) = \ell(\hat{\theta}_\psi)$ is usually called the *profile log-likelihood* and sometimes the *concentrated log-likelihood*. The Fisher information function is partitioned according to the partition of θ as

$$j(\theta) = \begin{pmatrix} j_{\psi\psi}(\theta) & j_{\psi\lambda}(\theta) \\ j_{\lambda\psi}(\theta) & j_{\lambda\lambda}(\theta) \end{pmatrix}. \tag{5}$$

It is shown in Tierney and Kadane (1986) that the relative error in (4) is $O(n^{-3/2})$ even though the relative errors in the approximations of the numerator and the denominator are just $O(n^{-1})$, because the constrained maximum likelihood estimator $\hat{\lambda}_\psi$ is within $O_p(n^{-1/2})$ of the full maximum likelihood estimator $\hat{\lambda}$ for ψ within a $\sqrt{n}$ neighborhood of ψ.

For inference about ψ it is more useful to have an expression for the cumulative distribution function. First, if θ is scalar, then

$$\Pr(\Theta \leq \theta \mid y) \doteq \Phi(r_B^*) = \Phi\left(r + \frac{1}{r} \log \frac{q_B}{r}\right) \tag{6}$$

$$r = \text{sign}(q_B)[2\{\ell(\hat{\theta}) - \ell(\theta)\}]^{1/2} \tag{7}$$

$$q_B = -\ell'(\theta) j^{-1/2}(\hat{\theta}) \frac{\pi(\hat{\theta})}{\pi(\theta)}.$$

For a scalar parameter of interest ψ with nuisance parameter λ we have

$$\Pr(\Psi \leq \psi \mid y) \doteq \Phi(r_B^*) \tag{8}$$

$$r = \text{sign}(q_B)[2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]^{1/2} \tag{9}$$

$$q_B = -\ell_p'(\psi) j_p^{-1/2}(\hat{\psi}) \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)}, \tag{10}$$

where r_B^* is defined by (6). Note that q_B takes the form of a score statistic, modified by the prior ratio, and that r is the signed square root of the log-likelihood ratio statistic.

Approximations (6) and (8) have relative error $O(n^{-3/2})$ for ψ in $\sqrt{n}$-neighborhoods of $\hat{\psi}$, and are often called r^* approximations. The form given here is originally due to Barndorff-Nielsen (1986, 1990) in a non-Bayesian context and can be shown to be asymptotically equivalent to the Lugannani and Rice (1980) version:

$$\Phi(r_B^*) \doteq \Phi(r) + \phi(r)\left(\frac{1}{r} - \frac{1}{q_B}\right); \tag{11}$$

neither version seems to dominate the other in terms of numerical properties, but r_B^* is a convenient expression of an asymptotically pivotal statistic.

The derivation of (6) from (3) involves changing the variable of integration from θ to r and then expressing the integrand as $\exp(-r_B^{*2}/2)$ by exponentiating the differential times the prior and completing the square. The derivation of the Bayesian marginal posterior in the form given by (11) is due to DiCiccio and Martin (1991). Several textbook treatments are now available; for example, in Barndorff-Nielsen and Cox (1994, chap. 6), Severini (2000, chap. 7), and Brazzale et al. (2007, chap. 8).

We can easily compare these approximate marginal posteriors under two different priors, π_1 and π_2, say, as the only effect is through the prior ratio in q_B. Thus, for example,

$$r^*_{B,\pi_1} - r^*_{B,\pi_2} = \frac{1}{r}\log\left(\frac{q_{B,\pi_1}}{q_{B,\pi_2}}\right) = \frac{1}{r}\left\{\log\frac{\pi_1(\hat\theta)}{\pi_1(\hat\theta_\psi)} - \log\frac{\pi_2(\hat\theta)}{\pi_2(\hat\theta_\psi)}\right\}. \tag{12}$$

We will use (12) to compare priors in Sec. 4.

3. Normal Circle

In this section we study in detail a simple model for which exact calculations are straightforward. This serves to illustrate several aspects of noninformative and objective priors. Assume that $y \sim N(\mu, I/n)$ follows a $k \times 1$ multivariate normal density, where I is the $k \times k$ identity matrix. The factor $1/n$ is used to distinguish the sample size from the dimension of the parameter space; in the asymptotic theory k is fixed and $n \to \infty$. We suppose that the scalar parameter of interest is the length of μ:

$$\psi = \|\mu\| = (\mu_1^2 + \cdots + \mu_k^2)^{1/2}.$$

If we use the (improper) joint prior $\pi(\mu)d\mu \propto d\mu$, a flat prior for the vector of means, the marginal posterior for ψ is proper and can be computed exactly from the noncentral χ^2 distribution:

$$\Pr_m(\Psi \geq \psi \mid y) = \Pr\{\chi^2_k(n\|y\|^2) \geq n\psi^2\}. \tag{13}$$

The r^* approximation to this is also easily computed using (8)–(10) and is given by $\Phi(r_B^*)$, where

$$r_B^* = \sqrt{n}(\hat\psi - \psi) + \frac{1}{\sqrt{n}(\hat\psi - \psi)}\log\left\{\left(\frac{\hat\psi}{\psi}\right)^{(k-1)/2}\right\} \tag{14}$$

and $\hat\psi = \|y\|$ is the maximum likelihood estimate of ψ.

This approximation is extremely accurate; in Fig. 1 the difference between the exact and approximate values is not visible. As a result we provide a small numerical comparison in Table 1. The accuracy of the approximation does not appear to degrade substantially with increasing numbers of nuisance parameters.

However, although the r_B^* approximation is very close to the exact noncentral χ^2, both probabilities give the wrong answer for this problem, in the sense that the posterior quantiles at level α given by these formulas are not α-level confidence bounds under the model. For this model there is an exact solution based

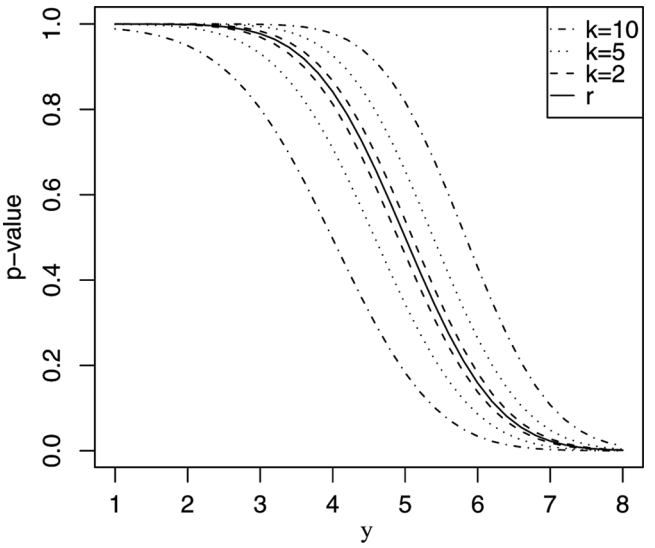

Figure 1. Exact and approximate p-value function and survivor function for ψ, for $n = 1$, $\hat{\psi} = 5$, and $k = 2, 5, 10$. The normal approximation to $r = \sqrt{n}(\hat{\psi} - \psi)$ is the solid line in the center. The three Bayesian marginal survivor functions are above this curve, and the three frequentist p-value functions are below this curve.

on the marginal likelihood of $\hat{\psi} = \|y\|$, which depends only on ψ and is also given by the noncentral χ^2 distribution:

$$\Pr_m(\widehat{\Psi} \geq \hat{\psi}) = \Pr\{\chi^2_k(n\psi^2) \geq n\|Y\|^2\}. \tag{15}$$

This can also be approximated by a frequentist version of the r^* approximation $\Phi(r_F^*)$, where

$$r_F^* = \sqrt{n}(\hat{\psi} - \psi) - \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log\left\{\left(\frac{\hat{\psi}}{\psi}\right)^{(k-1)/2}\right\}. \tag{16}$$

The exact and approximate p-value functions given by these approximations are also illustrated in Fig. 1. General formulas for r_F^* are given in Brazzale et al. (2007, chaps. 2 and 8). This suggests that the Bayesian confidence limits at the usual α-levels will systematically uncover the true value; i.e., the true confidence level of a $1 - \alpha$ posterior confidence limit will be less than $1 - \alpha$, which can be verified by evaluating (15) or (16) using quantiles obtained using (13) or (14).

Table 1
Comparison of the approximate survivor function for ψ, based on (8) with the exact value based on the noncentral χ^2 distribution; $n = 1$, $\hat{\psi} = 5$

k	Exact	0.99	0.95	0.75	0.5	0.25	0.05	0.01
5		0.9898	0.9495	0.7491	0.4991	0.2494	0.0499	0.00997
10	$\Phi(r_B^*)$	0.9897	0.9493	0.7486	0.4987	0.2494	0.0498	0.00997
20		0.9899	0.9500	0.7506	0.5012	0.2511	0.0504	0.01012

The poor behavior of the flat prior $\pi(\mu)d\mu \propto d\mu$ for inference about $\|\mu\|$ was pointed out by Stein (1959) in the context of point estimation and is discussed in Cox and Hinkley (1974, chap. 2), also in this context. It is also an example of the marginalization paradox of Dawid et al. (1973). If we noted at the outset that $\|y\|$ has a distribution only depending on ψ, then we could consider constructing a posterior density based on $\|y\|$ and using the marginal prior $\int \pi(\mu)d\mu$, where the integral is over the set of μ with fixed length. This prior,

$$\pi(\mu) \propto \|\mu\|^{-(k-1)}, \tag{17}$$

is also the *matching prior*, in the sense of Peers (1965) and Tibshirani (1989), and the *reference prior*, as discussed in Datta and Ghosh (1995). The marginalization paradox arises because the marginal posterior obtained from the flat prior for the full parameter is not consistent with the marginal posterior obtained this way. A general discussion of reference priors for noncentrality parameters is given in Berger et al. (1998).

We can see from (16) and (14) that, to order $O(n^{-3/2})$, using the prior (17) will recover the formula for r_F^* from that of r_B^*. We can also use these expressions to note that

$$r_B^* - r_F^* = \frac{k-1}{\sqrt{n}(\hat{\psi} - \psi)} \log\left(\frac{\hat{\psi}}{\psi}\right) = \frac{k-1}{\psi\sqrt{n}} + O(n^{-1}) \tag{18}$$

in $\sqrt{n}$ neighborhoods of $\hat{\psi}$, from which we can see that the discrepancy is increasing in k and decreasing with n, as we would expect.

Using the flat prior for μ is clearly informative for the parameter $\psi = \|\mu\|$, and this example is simple enough that the choice $\pi(\mu)d\mu \propto d\mu$ is obviously poor. In more complex models this is not always so clear. For example, Cox and Hinkley (1974, chap. 10) discussed the following exponential regression example due to Mitchell (1967): $Y_i \sim N(\mu_i, \sigma^2)$ where

$$\mu_i = \beta_0 + \beta_1 \rho^{x_0 + ja}$$

where x_0 and a are known. The prior $\pi(\beta_0, \beta_1\rho, \sigma) \propto d\beta_0 d\beta_1 d\log\sigma d\rho$ for $0 \le \rho \le 1$ leads to a marginal posterior for ρ that is improper, with unbounded spikes of mass at the two points 0 and 1. Improper posteriors are usually taken in Bayesian work to invalidate the prior, so from this point of view one could say that this prior would never be used. However, one could readily construct a very close approximation to this behavior through a slightly modified prior leading to a proper posterior.

It might be argued that in the context of the normal circle problem a flat prior would not normally be used. In the next section we consider for the normal circle example how to use (12) to check the sensitivity of the marginal posterior to the choice of prior.

4. Checking Sensitivity to Priors

Suppose in the normal circle model we use instead a conjugate prior, say an independence prior with

$$\mu_i \sim N(0, \tau^2), \quad i = 1, \ldots, k$$

and τ^2 assumed known. To assess the effect on the posterior, at least to $O(n^{-3/2})$, we need only compute the prior ratio $\pi(\hat{\theta})/\pi(\hat{\theta}_\psi)$, which reduces to

$$\frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} = \exp\left\{-\frac{1}{2\tau^2}(\hat{\psi}^2 - \psi^2)\right\} \tag{19}$$

where $\hat{\psi} = (\sum y_i^2)^{1/2}$. The posterior based on this prior is plotted in Fig. 2(a) for several choices of τ^2; as $\tau^2 \to \infty$, the posterior approaches that given by (13) as expected. However, to duplicate the matching or reference prior we would need to take the rather odd choice $\tau^2 = (\hat{\psi}^2 - \psi^2)/\{2(k-1)\log(\hat{\psi}/\psi)\}$.

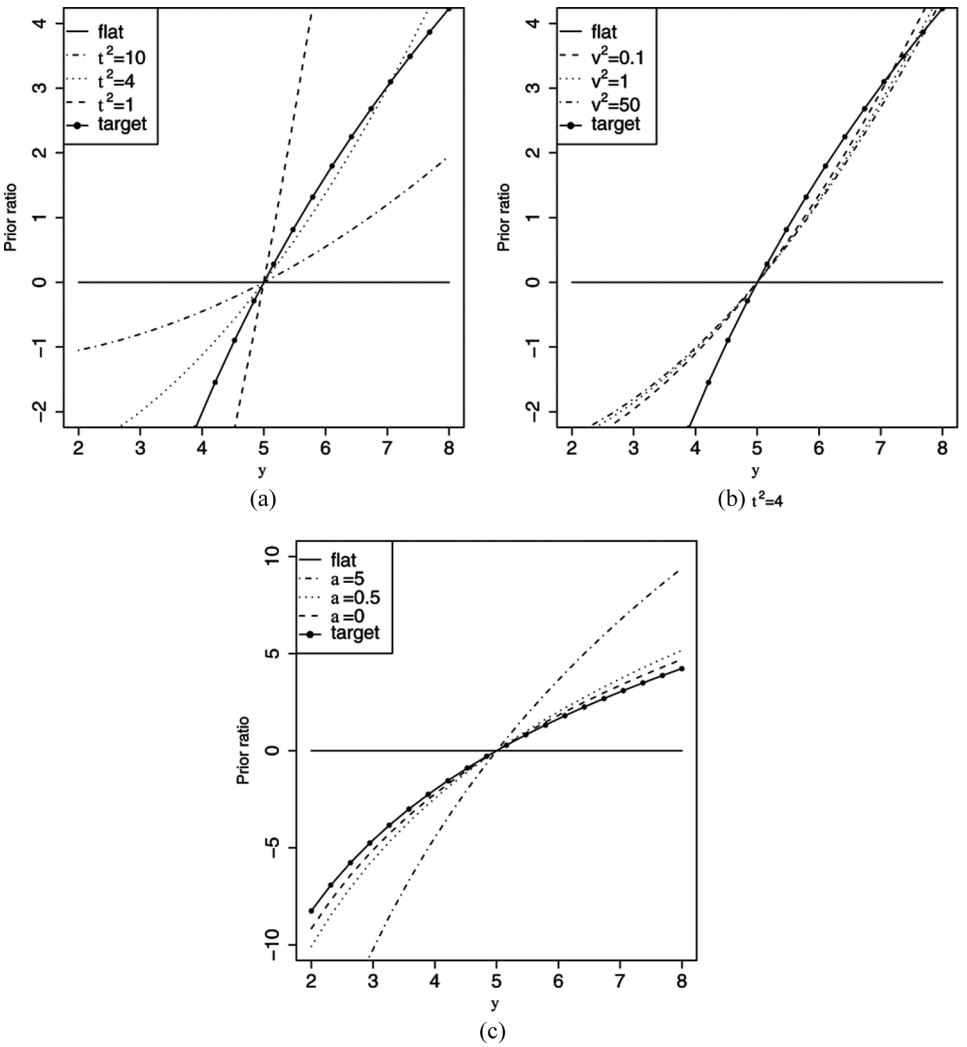

Figure 2. Comparison of the log-prior ratio $\log\{\pi(\hat{\theta})/\pi(\hat{\theta}_\psi)\}$ with that based on the reference, or matching, prior $\pi(\mu) \propto \psi^{-(k-1)}$, for $\hat{\psi} = 5$ and $k = 10$. (a) Independent $N(0, \tau^2)$ prior for μ_i; (b) prior (20) with $\tau^2 = 4$; (c) inverse gamma prior for σ^2, as at (21).

It seems plausible, and is often argued, that the influence of the prior can be decreased by using hierarchical priors. In this context we might, for example, use the hierarchy

$$\begin{aligned} \mu_i &\sim N(a, \tau^2), \quad i = 1, \ldots, k \\ a &\sim N(0, v^2); \end{aligned} \tag{20}$$

this leads to the prior ratio

$$\frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} = \exp\left\{-\frac{1}{2\tau^2}(\hat{\psi}^2 - \psi^2)\frac{k-1}{k}\right\},$$

if we take $v \to \infty$, which gives a very minor difference from the simple prior above. In computing this ratio we have assumed that the observed data is $(y_1, 0, \ldots, 0)$ to eliminate the term $\sum y_i$ from the prior ratio; this is without loss of generality since we can always rotate the data to achieve this.

We next introduce an inverse gamma prior for the variance of μ_i: suppose we take

$$\begin{aligned} \mu_i &\sim N(0, \tau^2), \quad i = 1, \ldots, k \\ \tau^{-2} &\sim \Gamma(\alpha, \beta), \end{aligned} \tag{21}$$

where α is the shape parameter and β the scale parameter for the gamma distribution. The prior ratio then becomes

$$\log \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} = -\left(\alpha + \frac{k}{2}\right) \log\left(\frac{\hat{\psi}^2}{\psi^2}\right)$$

as $\beta \to \infty$, which is plotted in Fig. 2(c) for various choices of α. This prior ratio gets quite close to the reference, or matching prior (17) as $\alpha \to 0$. Datta and Ghosh (1995) showed that the choice $\alpha = -1/2$, $\beta \to \infty$ gives the reference prior exactly.

Finally, we combine the third and second prior as

$$\mu_i \sim N(a, \tau^2), \quad a \sim N(0, v^2), \quad 1/\tau^2 \sim \Gamma(\alpha, \beta) \tag{22}$$

leading to the prior ratio

$$\log \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} = -\left(\alpha + \frac{k}{2}\right) \log \frac{\hat{\psi}^2}{\psi^2},$$

as $v, \beta \to \infty$, which is the same as for (21).

It is clear from this analysis that the hierarchical priors will not lead to good coverage of posterior probability intervals except in very special cases and that introducing a hierarchy of priors is not guaranteed to lead to priors with better frequentist behavior. Of course, they are not designed with this goal in mind, but they are widely used in linear models. In the next section we consider the use of flat priors in logistic regression.

5. Logistic Regression

Our first example is a simple linear logistic regression: we assume that y_i, $i = 1, \ldots, n$ is a sample of independent Bernoulli observations and that

$$\log\{\Pr(Y_i = 1)/\Pr(Y_i = 0)\} = \alpha + \beta x_i$$

where the x_i are taken as fixed covariates. The parameter $\theta = (\alpha, \beta)$ is at least approximately a location parameter, so it might be argued that the flat prior $\pi(\alpha, \beta)d\alpha\, d\beta \propto d\alpha\, d\beta$ is a reasonable starting point, and indeed this is the prior given as the default in both R and Winbugs. Suppose our parameter of interest $\psi = -\alpha/\beta$, which is the ED_{50}, the value of x for which the probability of success is 1/2. In Tables 2 and 3 we show the behavior of selected quantiles of r_F^* and r_B^* in simulations. As with the normal circle problem, the posterior quantiles systematically uncover the true value of ψ, relative to the sampling distribution. The intervals based on r_F^* have good central coverage but some imbalance in the two tails, which we think is due to the discreteness and the small sample size. The normal approximation to r_F^* was derived using the method outlined in Davison et al. (2006). It is difficult to do a full set of simulations over the parameter space for (α, β) with small values of n, as many parameter pairs lead to large numbers of samples for which the maximum likelihood estimates cannot be obtained.

Our second example is multiple logistic regression and is based on the data set `urine` in R; there are 77 binary observations, representing the presence or absence of certain biological markers in the urine, and there are 6 covariates. The data were discussed in Davison and Hinkley (1997), and r_F^* formulas for inference about components of the regression vector were developed in Brazzale (2000) and implemented in the `cond` library of the `hoa` package.

Table 2

Empirical coverage of upper and lower 2.5% limits based on the posterior marginal survivor function using a flat prior, $\Phi(r_B^*)$; the p-value function using the frequentist solution, $\Phi(r_F^*)$; and the normal approximation to the log-likelihood root, $\Phi(r)$. This is based on 10,000 simulations of the simple logistic regression model, based on a sample of size 15 from simulated data. The x sample was fixed throughout at (0.60, 0.67, 0.74, 0.81, 0.88, 0.96, 1.03, 1.10, 1.17, 1.24, 1.31, 1.38, 1.46, 1.53, 1.60) for four choices of α and β, all giving $\psi = 1$. For each subtable m is the number of simulations where the estimates of α and β could not be obtained; these simulations were omitted from the estimation of the coverage

	Left	Center	Right	Left	Center	Right
	$\alpha = 2, \beta = -2, m = 65$			$\alpha = -0.8, \beta = 0.8, m = 14$		
$\Phi(r)$	0.0315	0.9254	0.0431	0.0210	0.9320	0.0470
$\Phi(r_F^*)$	0.0232	0.9461	0.0306	0.0168	0.9494	0.0337
$\Phi(r_B^*)$	0.0471	0.8992	0.0674	0.0405	0.8804	0.0790
	$\alpha = 1.5, \beta = -1.5, m = 35$			$\alpha = 3, \beta = -3, m = 165$		
$\Phi(r)$	0.0265	0.9294	0.0440	0.0341	0.9266	0.0393
$\Phi(r_F^*)$	0.0204	0.9484	0.0312	0.0232	0.9497	0.0271
$\Phi(r_B^*)$	0.0410	0.8892	0.0697	0.0440	0.8963	0.0598

Table 3
Empirical coverage based on simulations as described in Table 2 for $\psi \neq 1$; m is the number of simulations omitted from the estimation of the coverage. The first subtable has $n = 15$ and the second $n = 30$

	Left	Center	Right	Left	Center	Right
	$\alpha = -2, \psi = 0.8, m = 174$			$\alpha = -2, \psi = 1.2, m = 39$		
$\Phi(r)$	0.0094	0.9356	0.0551	0.0411	0.9348	0.0240
$\Phi(r_F^*)$	0.0063	0.9527	0.0410	0.0300	0.9530	0.0170
$\Phi(r_B^*)$	0.0206	0.8936	0.0857	0.0734	0.8969	0.0357
	$\alpha = -2, \psi = 0.8 = -1.5, m = 0$			$\alpha = -2, \psi = 1.2 = -3, m = 0$		
$\Phi(r)$	0.0158	0.9468	0.0374	0.0362	0.9399	0.0239
$\Phi(r_F^*)$	0.0146	0.9525	0.0329	0.0313	0.9468	0.0219
$\Phi(r_B^*)$	0.0187	0.9346	0.0467	0.0440	0.9291	0.0269

The model is

$$y_i \sim Bernoulli(p_i), \quad \text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_6 x_{6i}. \tag{23}$$

The parameter of interest is taken to be

$$\psi = \exp(\beta_0 + \beta_1 x_1^* + \cdots + \beta_6 x_6^*)/\{1 + \exp(\beta_0 + \beta_1 x_1^* + \cdots + \beta_6 x_6^*)\},$$

the probability of success at a fixed value x^* of the covariates. Figure 3 shows three p-value functions for ψ, for different values of x^* chosen so that $\hat{\psi}$ is approximately

Table 4
Empirical coverage of upper and lower 2.5% limits based on the posterior marginal survivor function using a flat prior, $\Phi(r_{B1}^*)$; a matching prior for ψ, $\Phi(r_{B2}^*)$; the p-value function using the frequentist solution, $\Phi(r_F^*)$; and the normal approximation to the log-likelihood root, $\Phi(r)$. This is based on 10,000 simulations of the logistic regression model fitted to the `urine` data set from Davison and Hinkley (1997). In the three subtables 273, 283, and 280 simulated data sets, respectively, led to nonconvergence; these simulations were omitted

	Left	Center	Right	Left	Center	Right
		$\hat{\psi} = 0.86$			$\hat{\psi} = 0.51$	
$\Phi(r)$	0.0228	0.9278	0.0494	0.0343	0.9363	0.0306
$\Phi(r_F^*)$	0.0296	0.9511	0.0192	0.0256	0.9552	0.0203
$\Phi(r_{B1}^*)$ (flat)	0.0177	0.8722	0.1101	0.0462	0.9087	0.0451
$\Phi(r_{B2}^*)$ (matching)	0.0274	0.9509	0.0217	0.0247	0.9532	0.0222
		$\hat{\psi} = 0.21$				
$\Phi(r)$	0.0469	0.9287	0.0244			
$\Phi(r_F^*)$	0.0198	0.9532	0.0269			
$\Phi(r_{B1}^*)$ (flat)	0.0940	0.8841	0.0218			
$\Phi(r_{B2}^*)$ (matching)	0.0260	0.9476	0.0263			

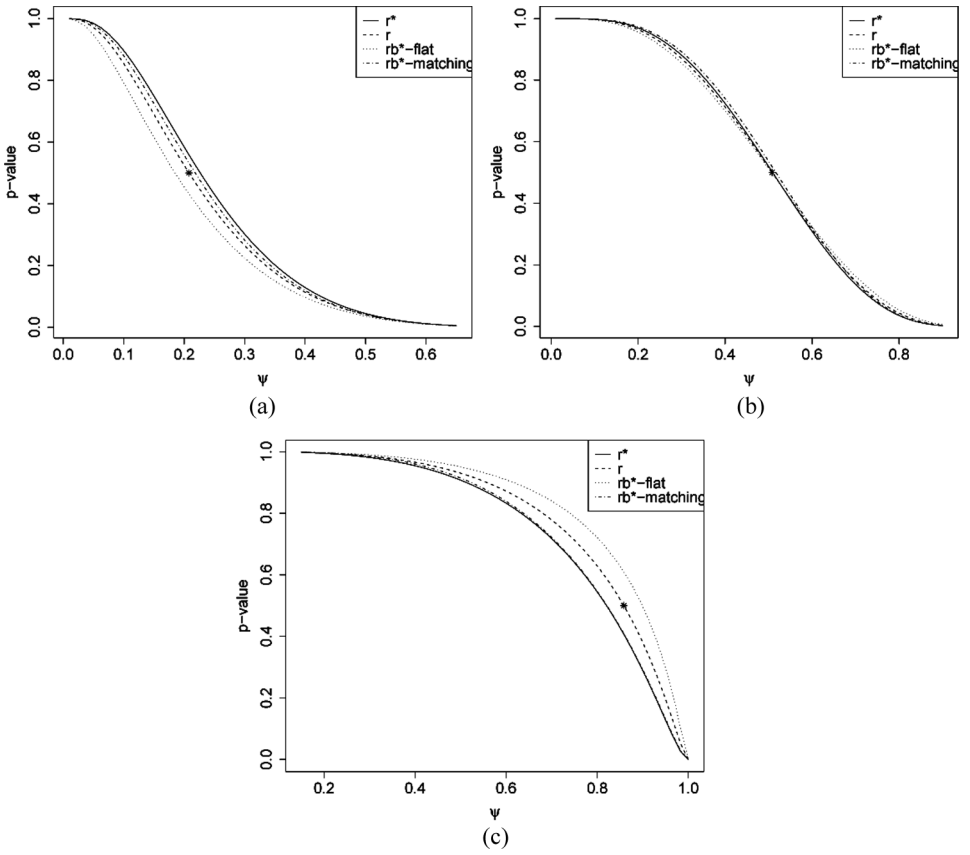

Figure 3. Bayesian posterior survivor function for $\psi = \exp(\beta^T x^*)/\{1 + \exp(\beta^T x^*)\}$ (dotted line), compared to the frequentist solution (solid line) and the normal approximation to the log-likelihood root (dashed line). Also shown is the Bayesian solution using the matching prior, as a dashed-dotted line; the differences between that and the frequentist solution are not very visible on the plot. The values of $\hat{\psi}$ from left to right are 0.21, 0.51, and 0.86.

0.8, 0.5, and 0.2, respectively. Again we see that the normal approximation to the distribution of r_B^* produces limits that uncover the true value, compared to the normal approximation to the distribution of r_F^*. The simulations summarized in Table 4 verify that the frequentist approximation is very accurate. There is in this example a matching prior for ψ, derived from the family of matching priors proposed in Tibshirani (1989) and discussed in the context of logistic regression in Staicu and Reid (2008), and Fig. 3 and Table 4 also show the results of using this prior, instead of the flat prior on β. This confirms that the problem with the coverage of intervals based on r_B^* is not due to the approximation of the exact marginal posterior distribution by $\Phi(r_B^*)$ but rather due to the prior.

6. Discussion

The examples in the previous sections show that using flat priors for vector parameters, and then marginalizing to scalar parameters that are nonlinear in the

original parameters, leads to Bayesian marginal posterior limits that do not have an interpretation as confidence limits. This observation has been made several times in the literature, and in particular the development of reference priors and matching priors takes due account of this. However, it is not widely emphasized and seems to be overlooked in a great deal of applied work. The construction of priors with some properties of calibration is important for settings in which there is an expectation that posterior limits are well calibrated, but unfortunately the construction of such targeted priors is difficult and needs to be undertaken separately for each parameter of interest. This means that it is not usually suitable to use a single prior to construct marginal posterior distributions for several different parameters of interest.

An example from the scientific literature of the failure of flat priors for several parameters is given in Heinrich (2006).

In some special situations it might not be required that the posterior marginal distribution have a frequency, or model-based, interpretation, in which case it would seem appropriate to use informative priors and acknowledge the dependence of the inference on the prior.

These are all small-sample results, in the sense that as n increases the effect of the prior goes to zero. However, the size of n needed depends on the number of nuisance parameters, as seen in the normal circle problem, where the discrepancy between the Bayesian and frequentist versions of r^* grows as $k/\sqrt{n}$. With discrete data and a large number of nuisance parameters, as in the second example of Sec. 5, the sample size needed to eliminate the effect of the prior may be rather large. In personal communication Thomas Richardson has suggested that it might be possible to check the coverage of Bayesian posterior intervals using a parametric bootstrap approach. The r_B^* approximation would be very useful for this as it is much faster to calculate than MCMC approximations.

Acknowledgments

This research was presented at the Gregynog Conference on April 20, 2008; NR would like to acknowledge the support of the organizers. We would like to thank the referees, as well as Don Fraser and Thomas Richardson, for helpful comments. The research was partially funded by the Natural Sciences and Engineering Council of Canada.

References

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

Barndorff-Nielsen, O. E. (1990). Approximate interval probabilities. *J. Roy. Stat. Soc. B* 52:485–496.

Barndorff-Nielsen, O. E., Cox, D. R. (1994). *Inference and Asymptotics*. London: Chapman and Hall.

Berger, J., Bernardo, J. (1992). On the development of reference priors. In: Berger, J., Bernardo, J., Dawid, A., Smith, A., eds. *Bayesian Statistics 4*. Oxford: Oxford University Press, pp. 34–49.

Berger, J. O., Philippe, A., Robert, C. (1998). Estimation of quadratic functions: noninformative priors for non-centrality parameters. *Statistica Sinica* 8:359–375.

Brazzale, A. R. (2000). Practical Small-Sample Parametric Inference. Ph.D. thesis, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne, Switzerland.

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

Clarke, B., Wasserman, L. (1995). Information trade-off. *TEST* 4:19–38.

Cox, D., Hinkley, D. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Datta, G., Ghosh, M. (1995). Some remarks on non-informative priors. *J. Am. Stat. Assoc.* 90:1357–1363.

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

Davison, A., Fraser, D. A. S., Reid, N. (2006). Likelihood inference for categorical data. *J. Roy. Stat. Soc. B* 68:495–508.

Davison, A. C., Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Dawid, A., Stone, M., Zidek, J. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Stat. Soc. B* 35:189–233.

DiCiccio, T. J., Martin, M. A. (1991). Approximations of marginal tail probabilities for a class of smooth functions with applications to Bayesian and conditional inference. *Biometrika* 78:891–902.

Gelman, A., Fagan, J., Kiss, A. (2007). An analysis of the New York City police department by "stop-and-frisk" policy in the context of claims of racial bias. *J. Am. Stat. Assoc.* 102:813–823.

Heinrich, J. (2006). The Bayesian approach to setting limits—what to avoid. In: Lyons, L., Unel, M. K., eds. *Statistical Problems in Particle Physics, Astrophysics and Cosmology*. London: Imperial College Press.

Lugannani, R., Rice, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. Appl. Probab.* 12:475–490.

Mitchell, A. (1967). Contribution to the discussion of I.J. Good. *J. Roy. Stat. Soc. B* 29:423–424.

Peers, H. (1965). On confidence points and Bayesian probability points in the case of several parameters. *J. Roy. Stat. Soc. B* 27:9–16.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

Staicu, A., Reid, N. (2008). On the uniqueness of matching priors. *Can. J. Stat.* 36:613–622.

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

Tibshirani, R. (1989). Non-informative priors for one parameter of many. *Biometrika* 76:604–608.

Tierney, L., Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.* 81:82–86.

Welch, B., Peers, H. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Stat. Soc. B* 25:318–329.