The Canadian Journal of Statistics
Vol. 36, No. 4, 2008, Pages 613–622
La revue canadienne de statistique

On probability matching priors

Ana-Maria STAICU and Nancy M. REID

Key words and phrases: Approximate Bayesian inference; Laplace approximation; orthogonal parameters; probability matching prior; tail probability approximation.

MSC 2000: Primary 62F15; secondary 62E99.

Abstract: First-order probability matching priors are priors for which Bayesian and frequentist inference, in the form of posterior quantiles, or confidence intervals, agree to a second order of approximation. The authors show that the matching priors developed by Peers (1965) and Tibshirani (1989) are readily and uniquely implemented in a third-order approximation to the posterior marginal density. The authors further show how strong orthogonality of parameters simplifies the arguments. Several examples illustrate their results.

À propos des lois a priori appariées en probabilité

Résumé : On dit que des lois a priori sont appariées en probabilité au premier ordre lorsque les quantiles a posteriori ou les intervalles de confiance des approches bayésienne et fréquentiste coïncident au second ordre. Les auteures montrent que les lois a priori appariées de Peers (1965) et Tibshirani (1989) conduisent à une approximation de troisième ordre simple et unique de la densité marginale a posteriori. Les auteures montrent aussi en quoi l'orthogonalité forte des paramètres simplifie les arguments. Plusieurs exemples illustrent leur propos.

1. INTRODUCTION

We consider parametric models for a response $Y = (Y_1, \ldots, Y_n)^\top$ with joint density $f(y;\theta)$, where the parameter $\theta^\top = (\psi, \lambda^\top)$ is assumed to be a d-dimensional vector with ψ the scalar component of interest. The log-likelihood function of the model is denoted by $\ell(\theta) = \log f(y;\theta)$. We write $j(\theta) = -n^{-1}\ell_{\theta^\top\theta}(\theta; y) = -n^{-1}\partial^2\ell(\theta;y)/\partial\theta^\top\partial\theta$ for the observed information matrix and $i(\theta) = n^{-1}\mathrm{E}\left\{-\ell_{\theta^\top\theta}(\theta;Y);\theta\right\}$ for the expected Fisher information matrix, per observation, and use subscript notation to indicate the partition of these matrices in accordance with the partition of the parameter; for example $i_{\psi\lambda}(\theta) = n^{-1}\mathrm{E}\left\{-\partial^2\ell(\theta)/\partial\psi\partial\lambda;\theta\right\}$.

In the absence of subjective prior information about the parameter θ, it may be natural to use a prior which leads to posterior probability limits that are also frequentist limits in that

$$\mathrm{pr}_\pi\{\psi \le \psi^{(1-\alpha)}(Y) \mid Y\} = \mathrm{pr}_\theta\{\psi^{(1-\alpha)}(Y) \ge \psi\} + O(n^{-1}),$$

where $\psi^{(1-\alpha)}(Y)$ is the upper $(1-\alpha)$ quantile of the marginal posterior distribution function $\Pi(\psi \mid Y)$, assumed to have density $\pi(\psi \mid Y)$. Following Datta & Mukerjee (2004), we call such priors first-order probability matching priors.

In a model with a scalar parameter, Welch & Peers (1963) showed that $\pi(\theta) \propto i^{1/2}(\theta)$ is the unique first-order probability matching prior. In models with nuisance parameters, Peers (1965) derived a class of first-order matching priors for ψ, as solutions to a partial differential equation. See also Mukerjee & Ghosh (1997), who provided a simpler derivation. In general this differential equation is not easy to solve, unless the components ψ and λ are orthogonal with respect to expected Fisher information, i.e., $i_{\psi\lambda}(\theta) = 0$. In this case Tibshirani (1989) and Nicolau (1993) showed that a family of solutions is

$$\pi(\psi, \lambda) \propto i_{\psi\psi}^{1/2}(\psi, \lambda) g(\lambda), \tag{1}$$

where $g(\lambda)$ is an arbitrary function. Sometimes consideration of higher order matching enables restriction of the class of functions $g(\lambda)$, occasionally enabling a unique matching prior to be defined; see Mukerjee & Dey (1993).

Levine & Casella (2003) proposed solving the partial differential equation numerically in models with a single nuisance parameter. Sweeting (2005) considered vector nuisance parameters and introduced data-dependent priors that locally approximate the matching priors. Both papers suggest implementing these priors using a Metropolis–Hastings algorithm, a rather computationally intensive procedure. Our work is closely connected to DiCiccio & Martin (1993), who use matching priors in approximate Bayesian inference as an alternative to more complicated frequentist formulas.

In this paper we argue that as long as one is satisfied with an approximation to the marginal posterior accurate to $O(n^{-3/2})$, the choice $g(\lambda) = 1$ in (1) is the simplest, and show that the marginal posterior approximation with this choice gives results that in simulations are verified to be very close to correct, from a frequentist point of view. The resulting marginal posterior is invariant to reparametrization, and is easily calculated with available software.

The paper is organized as follows. Section 2 presents the third-order approximation to the marginal posterior. Section 3 justifies the choice $g(\lambda) = 1$. Section 4 discusses models where the orthogonal components can be obtained without solving the differential equations. In Section 5 some examples illustrate the results. Section 6 provides our conclusions.

2. APPROXIMATE BAYESIAN INFERENCE

The Laplace approximation to the marginal posterior density $\pi(\psi \mid y)$ is given by

$$\pi(\psi \mid y) \doteq c|j_p(\hat{\psi})|^{1/2} \exp\{\ell_p(\psi) - \ell_p(\hat{\psi})\} \left\{ \frac{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|}{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|} \right\}^{1/2} \frac{\pi(\psi, \hat{\lambda}_\psi)}{\pi(\hat{\psi}, \hat{\lambda})},$$

where $\hat{\lambda}_\psi$ is the constrained maximum likelihood estimate, $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ is the profile log-likelihood for ψ, $\hat{\theta}^\top = (\hat{\psi}, \hat{\lambda}^\top)$ is the full maximum likelihood estimate, and $j_p(\hat{\psi}) = -\ell_p''(\hat{\psi})$ is the observed information corresponding to the profile log-likelihood. In the independently and identically distributed sampling context Tierney & Kadane (1986) showed that this approximation has relative error $O(n^{-3/2})$.

The corresponding $O(n^{-3/2})$ approximation to the marginal posterior tail probability is

$$\mathrm{pr}_\pi(\Psi \geq \psi \mid Y) = 1 - \Pi(\psi \mid y) \doteq \Phi(r) + \left(\frac{1}{r} - \frac{1}{q_B} \right) \phi(r) \tag{2}$$

where ϕ and Φ are the standard normal density and standard normal distribution function respectively, and

$$\begin{aligned} r &= \operatorname{sign}(\hat{\psi} - \psi) \left[2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\} \right]^{1/2}, \\ q_B &= -\ell_p'(\psi)\{j_p(\hat{\psi})\}^{-1/2} \left\{ \frac{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|}{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|} \right\}^{1/2} \frac{\pi(\hat{\psi}, \hat{\lambda})}{\pi(\psi, \hat{\lambda}_\psi)}; \end{aligned}$$

this was derived in DiCiccio & Martin (1991). An asymptotically equivalent approximation to (2), called *Barndorff-Nielsen's approximation* after Barndorff-Nielsen (1986), provides approximate posterior quantiles for ψ

$$\mathrm{pr}_\pi(\Psi \geq \psi \mid Y) \doteq \Phi(r_B^*), \tag{3}$$

where $r_B^* = r + r^{-1} \log(q_B / r)$.

When the model is given in an orthogonal parameterization $\theta^\top = (\psi, \lambda^\top)$, another version of the Laplace approximation to the marginal posterior density for ψ can be obtained by using the adjusted profile log-likelihood function $\ell_a(\psi) = \ell_p(\psi) - \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|$ (Cox & Reid 1987),

$$\pi(\psi \mid y) \doteq c|j_a(\hat{\psi})|^{1/2} \exp\{\ell_a(\psi) - \ell_a(\hat{\psi})\} \frac{\pi(\psi, \hat{\lambda}_\psi)}{\pi(\hat{\psi}, \hat{\lambda})}, \tag{4}$$

where $j_a(\psi) = -\ell_a''(\psi)$. This approximation also has a relative error of $O(n^{-3/2})$, and was discussed in more detail in DiCiccio & Martin (1993).

3. FIRST-ORDER PROBABILITY MATCHING PRIORS

When the model is given in an orthogonal parameterization, the first-order matching prior for the parameter of interest ψ, given by (1), enters approximation (2) as a ratio, so the relevant quantity is

$$\frac{i_{\psi\psi}^{1/2}(\hat{\psi}, \hat{\lambda})g(\hat{\lambda})}{i_{\psi\psi}^{1/2}(\psi, \hat{\lambda}_\psi)g(\hat{\lambda}_\psi)}.$$

Although the function $g(\lambda)$ is an arbitrary factor in (1), for sufficiently smooth g the ratio $g(\hat{\lambda})/g(\hat{\lambda}_\psi) = 1 + O_p(n^{-1})$, as a consequence of the result that $\hat{\lambda}_\psi = \hat{\lambda} + O_p(n^{-1})$ under parameter orthogonality. It follows that the approximation to $\Pi(\psi \mid y)$ in (2) is unique to $O(n^{-1})$. An approximation to the marginal posterior probabilities to $O(n^{-1})$ leads to posterior quantiles for ψ to $O_p(n^{-3/2})$, as can be verified by inversion of the relevant asymptotic series, as outlined in the Appendix.

The first-order matching prior for ψ,

$$\pi_U(\psi, \lambda) \propto i_{\psi\psi}^{1/2}(\psi, \lambda) \tag{5}$$

has the simplest analytical form under the class of Tibshirani's matching priors, and gives the same marginal posterior distribution for the parameter of interest as if any other matching prior of form (1) were used instead. Accordingly, we call this prior "the unique matching prior for the component ψ" under the orthogonal parameterization ψ and λ. This uniqueness was noted in DiCiccio & Martin (1993) in a discussion of the relation between the Bayesian third-order approximation (2) and a frequentist version developed in Barndorff-Nielsen (1986).

If an orthogonal parameterization is not explicitly available, the differential equations defining parameter orthogonality can be used in conjunction with (5) to give an expression for the prior in the original parameterization. We use the invariance argument presented in Mukerjee & Ghosh (1997) to express the matching prior in terms of the original parameterization.

More precisely, if our model is given in a parameterization $\phi^\top = (\psi, \eta^\top)$ not necessarily orthogonal, let $\theta^\top = (\psi, \lambda^\top)$ be an orthogonal reparameterization. Such an orthogonal reparameterization always exists when ψ is scalar; it is a solution of the partial differential equation

$$i_{\psi\eta}(\phi) = \frac{\partial\lambda(\phi)}{\partial\psi}\left\{\frac{\partial\lambda(\phi)}{\partial\eta^\top}\right\}^{-1} i_{\eta\eta}(\phi) \tag{6}$$

(Cox & Reid 1987). The unique first-order matching prior $\pi_U(\psi, \lambda)$ can be written in the original parameterization as

$$\pi_U(\psi, \eta) \propto i_{\psi\psi.\eta}^{1/2}(\psi, \eta)J(\psi, \eta), \tag{7}$$

where $i_{\psi\psi.\eta}(\psi, \eta) = i_{\psi\psi}(\psi, \eta) - i_{\psi\eta}(\psi, \eta)\{i_{\eta\eta}(\psi, \eta)\}^{-1}i_{\eta\psi}(\psi, \eta)$ is the (ψ, ψ) component of the expected Fisher information in the orthogonal parameterization, and $J(\psi, \eta) = |\partial\lambda/\partial\eta^\top|_+$ is the Jacobian of the transformation. In accordance with calling prior (5) a unique matching prior in the orthogonal parameterization (ψ, λ), the prior (7) shall be referred to as the unique matching prior in the (ψ, η) parameterization.

The analogy between (5) and (7) can also be justified by noting that in the orthogonal parameterization $\theta^\top = (\psi, \lambda^\top)$, the unique matching prior for ψ is proportional to the square root of the inverse of the asymptotic variance for $\hat{\psi}$. For a general parameterization $\phi^\top = (\psi, \eta^\top)$, the variance of $\hat{\psi}$ is the inverse of the partial information for ψ, i.e., $i^{\psi\psi}(\phi) = \{i_{\psi\psi.\eta}(\phi)\}^{-1}$

(Severini 2000, ch. 3.6), so the matching prior (7) in parameterization ϕ is a natural extension of the unique matching prior (5).

The unique matching prior (7) is similar to the local probability matching prior proposed by Sweeting (2005). The two priors share the term involving the partial information $i_{\psi\psi.\eta}(\theta)$; the extra factor in Sweeting's local prior is proportional to a local approximation of the Jacobian $J(\psi, \eta)$, based only on the parameter of interest and on the overall maximum likelihood estimate, see Sweeting (2005, equation (8)). An advantage of the unique matching prior is invariance to reparameterization.

4. STRONG ORTHOGONALITY

We examine in some detail models for which the orthogonal reparameterization has the property that $\hat{\lambda}_\psi = \hat{\lambda}$ holds for all ψ, which we call strong orthogonality. Barndorff-Nielsen & Cox (1994, ch. 3.6) pointed out that if $\hat{\lambda}_\psi = \hat{\lambda}$ holds for all ψ, then the components ψ and λ must be orthogonal. For models which admit strong orthogonality, the difficulty of obtaining matching priors can be reduced significantly, and the Bayesian posterior quantiles derived using (3) approximate the exact Bayesian posterior quantiles to $O(n^{-2})$.

For simplicity consider η to be a scalar nuisance parameter. If the score function corresponding to the nuisance parameter η has the form

$$\ell_\eta(\psi, \eta; y) \propto h\{\lambda(\psi, \eta); y\}, \tag{8}$$

for some functions $h(\,\cdot\,; y)$ and $\lambda(\,\cdot\,,\cdot\,)$ with $|\partial\lambda(\psi,\eta)/\partial\eta| \neq 0$, where the proportionality refers to non-zero functions which depend on the parameter only, then λ and ψ are strongly orthogonal. This follows from the equivariance of the constrained maximum likelihood estimator $\hat{\eta}_\psi$.

A simple form of (8) frequently encountered is $h\{\lambda(\psi,\eta); y\} = \lambda(\psi,\eta) - \tilde{p}(y)$, where we assume $|\partial\lambda(\psi,\eta)/\partial\eta| \neq 0$. Such is the case for the mean value reparameterization in the exponential family model. Another class of models giving strong orthogonality of parameters are those with likelihood orthogonality: i.e., $L(\psi,\eta) = L_1(\psi)L_2\{\lambda(\psi,\eta)\}$. The one-way random effects model in Section 5.3 belongs to this class.

This result is readily extended to the case where the nuisance parameter is a vector and h is then a vector of functions. More specifically, for $\eta^\top = (\eta_1, \ldots, \eta_{d-1})$ if the score function for the parametric model $f(y; \psi, \eta)$ has the form

$$\begin{aligned} \ell_{\eta_1}(\psi, \eta; y) &\propto h_1\{\lambda_1(\psi, \eta_1); y\}, \\ \ell_{\eta_k}(\phi; y) &\propto h_k\{\lambda_k(\psi, \eta_1, \ldots \eta_k), h_1(\,\cdot\,)f_{k,1}(\,\cdot\,), \ldots, h_{k-1}(\,\cdot\,)f_{k,k-1}(\,\cdot\,); y\}, \\ & \quad\; k = 2, \ldots, d-1 \end{aligned}$$

then $\hat{\lambda}_\psi = \hat{\lambda}$ and strong orthogonality holds. In these expressions we assume for each $1 \leq k \leq d-1$ that $h_k(\lambda_k, 0, \ldots, 0; y) = 0$ has a unique solution. For details of the proof, we refer the reader to Staicu (2007). We use strong orthogonality in the example of Section 5.5.

5. EXAMPLES

5.1. Linear exponential family.

Consider a sample of independently and identically distributed observations $Y = (Y_1, \ldots, Y_n)^\top$ from the model

$$f(y_i; \phi) = \exp\{\,\psi s(y_i) + \eta^\top t(y_i) - c(\phi) - d(y_i)\},$$

where $\phi^\top = (\psi, \eta^\top)$ is the full parameter and ψ the component of interest. An orthogonal reparameterization is given by $\theta^\top = (\psi, \lambda^\top)$ with $\lambda = E_\theta\{t_+(y)\}$, where $t_+(y) = \sum_{i=1}^n t(y_i)$. This can be obtained from the orthogonality equation (6), but more directly by noting that the arguments of the previous section ensure that $\hat{\lambda}_\psi = \hat{\lambda}$.

The unique first-order matching prior is

$$\pi_U(\phi) \propto i^{1/2}_{\psi\psi.\eta}(\phi)|c_{\eta\eta}(\phi)|_+.$$

It provides a unique marginal posterior distribution function for ψ to $O(n^{-3/2})$, as approximated by either (2) or (3). In these approximations, the expression for q_B simplifies to

$$q_B = -\ell_\psi(\tilde{\phi})i^{-1/2}_{\psi\psi.\eta}(\tilde{\phi})\left\{\frac{|i_{\eta\eta}(\hat{\phi})|}{|i_{\eta\eta}(\tilde{\phi})|}\right\}^{1/2},$$

where $i_{\psi\psi.\eta}(\phi) = c_{\psi\psi}(\phi) - c_{\psi\eta}(\phi)\{c_{\eta\eta}(\phi)\}^{-1}c_{\eta\psi}(\phi)$, $\tilde{\phi} = (\psi, \hat{\eta}_\psi)$ and $\hat{\phi} = (\hat{\psi}, \hat{\eta})$. The example is considered in DiCiccio & Martin (1993) as well.

5.2. Logistic regression.

We analyze the urine data of Davison & Hinkley (1997, Example 7.8). The presence or absence of calcium oxalate crystals in urine as well as specific gravity, pH, osmolarity, conductivity, urea concentration and calcium concentration are measured for 77 complete cases. The relationship between calcium oxalate crystals and the 6 explanatory variables is investigated under the logistic regression model. Matching priors for logistic regression are obtained numerically in Levine & Casella (2003) and Sweeting (2005); here we give a simple analytical solution.

The logistic regression model for a vector of independent random variables $Y_i \sim$ Binomial(m_i, p_i) has log-likelihood function

$$\ell(\beta) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}) - \sum_{i=1}^{n} m_i \log\{1 + e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pi}}\}.$$

Assume the parameter of interest is $\psi = \beta_p$, and take $\eta = (\beta_0, \ldots, \beta_{p-1})^\top$ to be the nuisance parameter. Since the model is in the exponential family, $\lambda = E_\beta\{t(y)\} = E_\beta(\sum_{i=1}^n y_i, \ldots, \sum_{i=1}^n y_i x_{p-1,i})$ is orthogonal to ψ. Therefore, the unique matching prior has the form

$$\pi_U(\beta) \propto i^{1/2}_{\psi\psi.\eta}(\beta)|i_{\eta\eta}(\beta)|_+, \tag{9}$$

with $i_{\psi\psi.\eta}(\beta) = i_{\psi\psi}(\beta) - i_{\psi\eta}(\beta)\{i_{\eta\eta}(\beta)\}^{-1}i_{\eta\psi}(\beta)$.

The block matrices which partition the expected Fisher information function have a simple form: $i_{\psi\psi}(\beta) = x_p^\top V(\beta)x_p$, $i_{\psi\eta}(\beta) = x_p^\top V(\beta)X_{-p}$ and $i_{\eta\eta}(\beta) = X_{-p}^\top V(\beta)X_{-p}$, where $V(\beta) = \mathrm{diag}\,\{m_i p_i(1-p_i)\}$, X is the $n \times (p+1)$ model matrix and $X_{-p} = X - \{x_p\}$ is the $n \times p$ matrix obtained by removing the column vector x_p. For $p = 2$ this example is discussed in Sweeting (2005), where comparison with his equation (18) shows that the factor $|i_{\eta\eta}(\beta)|_+$ in (9) is approximated by $\exp\{-2\beta_p T(\hat{\beta})\}$, a function that is log-linear in $\psi = \beta_p$; the function T depends only on x and the fitted probabilities.

To illustrate, we take $\psi = \beta_6$, the coefficient of the effect of calcium concentration on the presence of calcium oxalate crystals in urine. The 95% posterior probability intervals using the Bayesian approach with matching prior (9) are given in Table 1. The frequentist calculations were carried out using the COND package in the HOA library bundle for R (Brazzale 2000). Although this package does not provide the Bayesian solution explicitly, the components needed are readily derived from the workspace. Also shown is the standard output from COND: two (first-order) normal approximations, and the frequentist version of the r_B^* approximation. The second normal approximation is based on the adjusted log-likelihood function $\ell_a(\psi)$, described above (4). While both approximations have relative error $O(n^{-1/2})$, the normal approximation based on $\ell_a(\psi)$ often seems to provide more accurate inferences in the presence of nuisance parameters, although this is not the case here. The third-order frequentist approximation is the saddlepoint approximation to the conditional distribution of $\sum_{i=1}^n x_{6i}y_i$ given $t = \hat{\lambda}$. The matching prior version is indeed equivalent to the frequentist solution, giving essentially the same confidence limits and P-values. A plot of the P-value function for β_6 (not shown) confirms that the survivor function for β_6 based on the matching prior accurately approximates the frequentist P-value function for all relevant values of β_6. We might expect that with 7 covariates and a sample size as large as 77, the data might swamp the prior, and first-order asymptotics would suffice. However, this is not the case; there is actually much less information in binary data than in continuous data. This point is expanded on in Brazzale, Davison & Reid (2007, pp. 58–9). Chapter 2 of the same book gives an overview of higher order frequentist approximations of the type presented here.

TABLE 1: Comparison of the 95% confidence intervals for β_6 and of the P-values for testing $\mathcal{H}_0$: $\beta_6 = 0$.

	95% CI for β_6		P-value
Normal approximation to MLE $\hat{\beta}_6$	(0.3169	1.250)	4.9887e-004
Normal approximation to conditional MLE $\hat{\beta}_6^c$	(0.2631	1.160)	9.3724e-004
Third-order frequentist approximation	(0.3224	1.208)	6.6893e-006
Laplace approximation with prior (9)	(0.3213	1.211)	5.3555e-006

5.3. Random effects model.

Consider the one-way random effects model $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, where τ_i and ε_{ij} are mutually independent with $\tau_i \sim \mathsf{N}(0, \sigma_\tau^2)$ and $\varepsilon_{ij} \sim \mathsf{N}(0, \sigma^2)$. For each i, the log-likelihood component is

$$\begin{aligned}\ell(\mu, \sigma_\tau^2, \sigma^2; y_i) &= -\frac{1}{2}(n_i - 1)\log\sigma^2 - \frac{1}{2}\log(\sigma^2 + n_i\sigma_\tau^2) - \frac{1}{2}n_i\mu^2(\sigma^2 + n_i\sigma_\tau^2)^{-1} \\ &\quad - \frac{1}{2}\frac{s_i^2}{\sigma^2} - \frac{1}{2}n_i\bar{y}_{i\cdot}^2(\sigma^2 + n_i\sigma_\tau^2)^{-1} + n_i\bar{y}_{i\cdot}\mu(\sigma^2 + n_i\sigma_\tau^2)^{-1},\end{aligned}$$

where $\bar{y}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$ and $s_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$. Note this has the form of an exponential family log-likelihood, with some canonical parameters depending on the sample size.

If $\psi = \mu$ is the parameter of interest, $\eta^\top = (\sigma_\tau^2, \sigma^2)$ is orthogonal to μ and a unique matching prior is obtained from (5). However the (ψ, ψ) component of the expected Fisher information matrix is a function only of the nuisance parameter

$$i_{\psi\psi}(\psi, \sigma_\tau^2, \sigma^2) \propto \sum_{i=1}^{k} n_i(\sigma_\tau^2 + n_i\sigma^2)^{-1},$$

so we can further simplify the unique matching prior for $\psi = \mu$ to the flat prior

$$\pi_U(\psi, \eta) \propto 1.$$

When $\psi = \sigma^2$ is the parameter of interest with $\eta = (\sigma_\tau^2, \mu)$ being the nuisance component, we take $\lambda_2 = \mu$, since $\hat{\lambda}_{2,\psi} = \bar{y}_{\cdot\cdot}$, where $\bar{y}_{\cdot\cdot} = N^{-1}\sum_{i=1}^{k} n_i y_i$ and $N = \sum_{i=1}^{k} n_i$. The differential equation (6) can then be used to obtain λ_1. In the balanced design with $n_1 = \cdots = n_k = n$, the score functions corresponding to the nuisance parameter η have the form

$$\begin{aligned}\ell_{\eta_1}(\psi, \eta) &= (\psi + n\eta_1)^{-2}\left\{-\frac{nk}{2}(\psi + n\eta_1) + \frac{n^2}{2}\sum_{i=1}^{k}(\bar{y}_{i\cdot} - \eta_2)^2\right\}, \\ \ell_{\eta_2}(\psi, \eta) &= nk(\psi + n\eta_1)^{-1}\{\bar{y}_{\cdot\cdot} - \eta_2\},\end{aligned}$$

and we can use the result described in Section 4 to identify $\lambda_1 = \psi + n\eta_1$ and $\lambda_2 = \eta_2$ as being orthogonal to the interest parameter ψ. Moreover for this reparameterization we have strong orthogonality:

$$\hat{\lambda}_{1,\psi} = \hat{\lambda}_1 = \frac{n}{k}\sum_{i=1}^{k}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \quad \text{and} \quad \hat{\lambda}_{2,\psi} = \hat{\lambda}_2 = \bar{y}_{\cdot\cdot} \quad \text{for all } \psi.$$

Regardless of the method used, we find the partial information for ψ, $i_{\psi\psi.\eta}(\psi,\eta) \propto \psi^{-2}$, and the Jacobian of the transformation $|\partial\lambda/\partial\eta^{\top}| = n$. Then by using (7) we obtain the prior

$$\pi_U(\psi,\eta) \propto \psi^{-1},$$

which gives unique approximate matching inference based on matching priors in the orthogonal parameterization $(\psi, \lambda^{\top})$.

Computing the posterior using (2) or (3) involves very elementary calculations on the model; the main computational work involved is evaluating the constrained maximum likelihood estimator for a grid of 200 values for ψ.

We performed the same simulation study as Levine & Casella (2003). We randomly generated 100,000 data sets from the random effects model with $n = 10$ and $k = 3$, for $\mu = 10$ and $\sigma_\tau = \sigma = 1$, and calculated the 95% posterior intervals for the parameter of interest $\psi = \sigma^2$. The posterior interval was easily obtained by spline smoothing. The simulated coverage of the 95% posterior intervals was 94.991%; the coverage obtained by Levine & Casella using a Metropolis Hastings algorithm with the prior $\pi(\psi,\eta) \propto \psi^{-1}(\psi + n\eta_1)^{-1}$ was 92.3%.

5.4. Inverse gaussian model.

Suppose that $Y_i \sim \text{IG}(\mu, \sigma^2)$ with probability density function

$$f(y;\mu,\sigma^2) = \frac{y^{-3/2}}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2\mu^2 y}\right\}, \quad y > 0,$$

where $\mu > 0$ and $i = 1, \ldots, n$. This parameterization is orthogonal and the expected information matrix is $i(\mu,\sigma^2) = \text{diag}\,(\mu^{-3}\sigma^{-2}, \sigma^{-4}/2)$. When $\psi = \sigma^2$ and $\lambda = \mu$ we have strong orthogonality: $\hat{\lambda}_\psi = \hat{\lambda} = \bar{y}$, where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$. Hence all the first-order matching priors lead to unique approximation to the marginal posterior distribution as given by (2); the unique first-order matching prior is $\pi_U(\sigma^2,\mu) \propto \sigma^{-2}$. When the interest parameter is $\psi = \mu$, we no longer have strong orthogonality: $\hat{\lambda}_\psi = n^{-1}\sum_{i=1}^{n} y_i^{-1} + \bar{y}\psi^{-2} - 2\psi^{-1}$. The unique matching prior (5) is $\pi_U(\mu,\sigma^2) \propto \mu^{-3/2}\sigma^{-1}$. Datta & Ghosh (1995) propose the reverse reference prior $\pi_{\text{RR}}(\mu,\sigma^2) \propto \mu^{-3/2}\sigma^{-2}$, as it is a matching prior for each parameter in turn. This prior is of the form (1) with $g(\lambda) = \lambda^{-1/2}$, so both priors $\pi_{\text{RR}}(\mu,\sigma^2)$ and $\pi_U(\mu,\sigma^2)$ result in the same approximate Bayesian inference to order $O(n^{-1})$.

5.5. Multivariate normal mean.

Suppose that $Y_i \sim \mathsf{N}(\mu_i, 1)$ with $\mu_i \in \mathbf{R}$ for $i = 1, \ldots, p$, and take the parameter of interest to be $\psi = (\mu_1^2 + \cdots + \mu_p^2)^{1/2}$. Datta & Ghosh (1995) use the reparameterization $(\psi, \lambda_1, \ldots, \lambda_{p-1})$ with $\mu_1 = \psi\cos\lambda_1$, $\mu_2 = \psi\sin\lambda_1\cos\lambda_2$, ..., $\mu_{p-1} = \psi\prod_{i=1}^{p-2}\sin\lambda_i\cos\lambda_{p-1}$, and lastly $\mu_p = \psi\prod_{i=1}^{p-2}\sin\lambda_i\sin\lambda_{p-1}$; the information in this reparameterization is

$$i(\psi,\lambda) = \text{diag}\left(1, \psi^2, \psi^2\sin^2\lambda_1, \ldots, \psi^2\prod_{i=1}^{p-2}\sin^2\lambda_i\right).$$

This reparameterization also gives strong orthogonality, as we now show. The constrained maximum likelihood estimate $\hat{\lambda}_{p-1,\psi}$ is the solution of $\ell_{\lambda_{p-1}}(\psi,\lambda) = 0$, where

$$\ell_{\lambda_{p-1}}(\psi,\lambda) \propto y_{p-1}\sin\lambda_{p-1} - y_p\cos\lambda_{p-1},$$

yielding $\hat{\lambda}_{p-1,\psi} = \hat{\lambda} = \arctan(y_p/y_{p-1})$.

Next, we note that the score function corresponding to coordinate λ_{p-2}, $\ell_{\lambda_{p-2}}(\psi, \lambda)$ has the form

$$\ell_{\lambda_{p-2}}(\psi, \lambda) \propto y_{p-2} \sin \lambda_{p-2} - (y_{p-1} \cos \lambda_{p-1} + y_p \sin \lambda_{p-1}) \cos \lambda_{p-2},$$

and therefore the solution $\hat{\lambda}_{p-2,\psi}$ of the score equation $\ell_{\lambda_{p-2}}(\psi, \lambda) = 0$ is $\hat{\lambda}_{p-2,\psi} = \hat{\lambda}_{p-2} = \arctan(y_{p-1}/y_{p-2}) \cos \hat{\lambda}_{p-1} + (y_p/y_{p-2}) \sin \hat{\lambda}_{p-1}$; we continue with this backward procedure to obtain $\hat{\lambda}_\psi = \hat{\lambda}$. Having strong orthogonality, the unique matching prior is $\pi_U(\psi, \lambda) \propto 1$. Datta & Ghosh (1995) and Tibshirani (1989) obtained $\pi_R(\psi, \lambda) \propto \prod_{k=1}^{p-1} \sin^{p-1-k} \lambda_k$ as a first-order matching prior for ψ; this prior is also a reference prior. Both priors give the same posterior quantiles to third-order.

6. DISCUSSION

We have illustrated the use of $\pi_U(\theta)$, a particular choice of the Tibshirani–Peers matching prior, in two practical and two theoretical examples. Several further examples are discussed in Staicu (2007). The use of this prior in third-order approximations is quite straightforward, and avoids any simulation or numerical integration. There is no need to choose among the family of matching priors, in particular to search for a matching prior to higher order: in fact when $\hat{\lambda}_\psi = \hat{\lambda}$, the unique first-order matching prior is second-order matching if and only if the model has the following property (R. Mukerjee, personal communication):

$$\frac{\partial}{\partial \psi} \left[i_{\psi\psi}^{-3/2}(\theta) \mathrm{E}\left\{\ell_\psi^3(\theta)\right\} \right] = 0.$$

A reviewer has pointed out that with improper priors there is no guarantee that the posterior is proper, and this needs to be checked on a case-by-case basis. For the examples of Section 5, the posterior is indeed proper, and it seems possible that the matching argument could be used in conjunction with the third-order approximation to show that the unique prior will, under regularity conditions on the model, lead to proper posteriors.

The reference approach to noninformative priors, based on maximizing the Kullback–Liebler distance between the prior and the posterior, often gives posterior inferences which are frequentist matching, although they are not derived from this point of view. Kass & Wasserman (1996) provide an introduction to this literature, and show that under parameter orthogonality, and subject to rather strong regularity conditions, the reference prior is proportional to $|i_{\lambda\lambda}(\theta)|^{1/2} g(\psi)$, exactly opposite to (1).

The approximation used for the logistic regression example is to a conditional distribution, given a sufficient statistic for the nuisance parameter. The normal approximation to r_B^* has frequentist matching conditionally on this statistic, and hence unconditionally. It is also an approximation to a discrete distribution, whereas the theory of quantile matching implicitly assumes underlying continuity. The approximation is best viewed as matching a continuous (smoothed) version of the discrete distribution, as in Davison & Wang (2002), but the theoretical details to verify this have not yet been established. Rousseau (2000) provides the most detailed results on this aspect.

Severini (2007) has considered the construction of conditional priors $\pi(\lambda \mid \psi)$, using a notion of parameter orthogonality which he calls strongly unrelated. Although this work was not directly focussed on frequentist matching, a referee has suggested that it may be possible to use Severini's approach to extend the notion of matching priors.

APPENDIX

LEMMA 1. *For* $0 < \alpha < 1$, *denote by* $\hat{\psi}^{(1-\alpha)}(\pi, y)$ *the posterior quantile corresponding to prior* $\pi(\theta) \propto i_{\psi\psi}^{1/2}(\theta)g(\lambda)$, *which is defined by* $\Pi\{\hat{\psi}^{(1-\alpha)}(\pi, y) \mid y\} = 1 - \alpha$. *We assume that* $g(\lambda) \neq 0$ *has a continuous first derivative for all* λ. *Then*

$$\hat{\psi}^{(1-\alpha)}(\pi, y) = \hat{\psi}^{(1-\alpha)}(\pi_U, y) + O_p(n^{-3/2});$$

that is, the posterior quantile is unique to $O_p(n^{-3/2})$ *under the class of matching priors* $\pi(\theta)$.

Proof. Let z_α denote the $100(1-\alpha)$ percentile of a standard normal variate and let $j^{\psi\psi}(\psi, \lambda)$ stand for the (ψ, ψ) component of the inverse of the observed information matrix. The Cornish–Fisher inversion of the Edgeworth expansion for the marginal posterior distribution of ψ leads to

$$\begin{aligned}\hat{\psi}^{(1-\alpha)}(\pi, y) &= \hat{\psi} + n^{-1/2}\{j^{\psi\psi}(\hat{\theta})\}^{1/2} z_\alpha \\ &\quad + n^{-1}\{j^{\psi\psi}(\hat{\theta})\}^{1/2} u_1(z_\alpha, \pi, y) + O_p(n^{-3/2}),\end{aligned}$$

where $u_1(z_\alpha, \pi, y) = A_{11}(\pi, y) + A_{12}(y) + (z_\alpha^2 + 2)A_3(y)$ with

$$A_{11}(\pi, y) = \{j^{\psi\psi}(\hat{\theta})\}^{-1/2}\left\{\frac{\pi_\psi(\hat{\theta})}{\pi(\hat{\theta})} j^{\psi\psi}(\hat{\theta}) + \frac{\pi_{\lambda^\top}(\hat{\theta})}{\pi(\hat{\theta})} j^{\lambda\psi}(\hat{\theta})\right\}$$

$\pi_\psi(\theta) = \partial\pi(\theta)/\partial\psi$, $\pi_\lambda(\theta) = \partial\pi(\theta)/\partial\lambda$ and the expressions for A_{12} and A_3 are given in Mukerjee & Reid (1999). It suffices to show $A_{11}(\pi, y)$ does not depend on $g(\lambda)$ to order $O_p(n^{-1/2})$. By the assumptions on g and the consistency of the maximum likelihood estimator we have $g_\lambda(\hat{\lambda})/g(\hat{\lambda}) = O_p(1)$ and the result follows.

ACKNOWLEDGEMENTS

The authors were partially supported by the Natural Sciences and Engineering Research Council of Canada. They would like to acknowledge helpful discussions with Anthony Davison, Don Fraser, Rahul Mukerjee and Malay Ghosh. The authors also wish to thank the Associate Editor and the reviewers for their suggestions.

REFERENCES

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

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

A. R. Brazzale (2000). *Practical Small-Sample Parametric Inference*. Doctoral dissertation, Département de mathématiques, École polytechnique fédérale de Lausanne, Lausanne, Switzerland.

A. R. Brazzale, A. C. Davison & N. M. Reid (2007). *Applied Asymptotics: Case Studies in Small Sample Statistics*. Cambridge University Press.

D. R. Cox & N. M. Reid (1987). Parameter orthogonality and approximate conditional inference (with Discussion). *Journal of the Royal Statistical Society Series B*, 49, 1–39.

G. S. Datta & M. Ghosh (1995). Some remarks on noninformative priors. *Journal of the American Statistical Association*, 90, 1357–1363.

G. S. Datta & R. Mukerjee (2004). Probability Matching Priors: Higher Order Asymptotics. *Lecture Notes in Statistics* 178, Springer, New York.

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

A. C. Davison & S. Wang (2002). Saddlepoint approximations as smoothers. *Biometrika*, 89, 933–938.

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

T. J. DiCiccio & M. A. Martin (1993). Simple modifications for signed roots of likelihood ratio statistics. *Journal of the Royal Statistical Society Series B*, 55, 305–316.

R. E. Kass & L. A. Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343–1370.

R. A. Levine & G. Casella (2003). Implementing matching priors for frequentist inference. *Biometrika*, 90, 127–137.

R. Mukerjee & D. K. Dey (1993). Frequentist validity of posterior quantiles in the presence of a nuisance parameter: higher order asymptotics. *Biometrika*, 80, 499–505.

R. Mukerjee & M. Ghosh (1997). Second-order probability matching priors. *Biometrika*, 84, 970–975.

R. Mukerjee & N. M. Reid (1999). On a property of probability matching priors: Matching the alternative coverage probabilities. *Biometrika*, 86, 333–340.

A. Nicolaou (1993). Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *Journal of the Royal Statistical Society Series B*, 55, 377–390.

H. W. Peers (1965). On confidence points and Bayesian probability points in the case of several parameters. *Journal of the Royal Statistical Society Series B*, 27, 9–16.

J. Rousseau (2000). Coverage properties of one-sided intervals in the discrete case and applications to matching priors. *Annals of the Institute of Statistical Mathematics*, 52, 28–42.

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

T. A. Severini (2007). Integrated likelihood functions for non-Bayesian inference. *Biometrika*, 94, 529–542.

A. M. Staicu (2007). *On some Aspects of Likelihood Methods with Applications in Biostatistics*. Doctoral dissertation, Department of Statistics, University of Toronto.

T. J. Sweeting (2005). On the implementation of local probability priors for interest parameters. *Biometrika*, 92, 47–57.

R. J. Tibshirani (1989). Noninformative priors for one parameter of many. *Biometrika*, 76, 604–608.

L. J. Tierney & J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82–86.

B. L. Welch & B. L. Peers (1963). On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society Series B*, 25, 318–329.

Received 14 September 2007
Accepted 16 June 2008

Ana-Maria STAICU: a.staicu@bristol.ac.uk
Department of Mathematics, University of Bristol
University Walk, Bristol BS8 1TW
United Kingdom

Nancy M. REID: reid@utstat.utoronto.ca
Department of Statistics, University of Toronto
100 Saint George Street, Toronto, Ontario
Canada, M5S 3G3