On the implementation of local probability matching priors for interest parameters

By TREVOR J. SWEETING

Department of Statistical Science, University College London, WC1E 6BT, UK

trevor@stats.ucl.ac.uk

SUMMARY

Probability matching priors are priors for which the posterior probabilities of certain specified sets are exactly or approximately equal to their coverage probabilities. These priors arise as solutions of partial differential equations that may be difficult to solve, either analytically or numerically. Recently Levine & Casella (2003) presented an algorithm for the implementation of probability matching priors for an interest parameter in the presence of a single nuisance parameter. In this paper we develop a local implementation that is very much more easily computed. A local probability matching prior is a data-dependent approximation to a probability matching prior and is such that the asymptotic order of approximation of the frequentist coverage probability is not degraded. We illustrate the theory with a number of examples, including three discussed in Levine & Casella (2003).

Some key words: Coverage probability bias; Data-dependent prior; Higher-order asymptotics; Metropolis-Hastings algorithm; Nuisance parameter; Probability matching prior.

1. INTRODUCTION

In this paper we explore the practical implementation of probability matching priors for an interest parameter. Probability matching priors are prior distributions for which the posterior probabilities of certain specified sets are exactly or approximately equal to their coverage probabilities. There has been considerable development of probability matching priors since the paper of Welch & Peers (1963). In that paper it was shown that, in the case of n independent and identically distributed observations from a distribution known up to a single real parameter, Jeffreys' invariant prior achieves probability matching to $O_p(n^{-1})$. That is, posterior probabilities

1

coincide with coverage probabilities with an asymptotic error of order n^{-1}. In a development of this result, Peers (1965) obtained a partial differential equation satisfied by probability matching priors for a specified interest parameter in the presence of nuisance parameters. This work was further developed by Stein (1985), Tibshirani (1989), Datta & J. K. Ghosh (1995) and Datta & M. Ghosh (1995b). These results can be used to deduce frequentist properties of nonsubjective Bayesian priors, thereby providing some justification for their use in terms of repeated sampling performance. There is an extensive literature on nonsubjective Bayes methods; see for example Ghosh & Mukerjee (1992), Sweeting (1995), Kass & Wasserman (1996) and Bernardo & Ramón (1998). Sweeting (2001) specifically discusses aspects of coverage probability bias associated with Bayesian priors. From a frequentist point of view, the use of probability matching priors provides additional noninformative Bayesian interpretation of confidence intervals and supplies some conditional validity for confidence statements. Alternatively, probability matching priors can be viewed simply as a technical device for the elimination of nuisance parameters for the construction of frequentist confidence intervals. Two recent reviews of probability matching priors are Datta & Mukerjee (2004) and Datta & Sweeting (2004).

The partial differential equation satisfied by probability matching priors can only be solved in particular cases; some examples are given in Datta & M. Ghosh (1995b). If the problem is parameterised so that the nuisance parameter is orthogonal to the interest parameter then the solutions are of a very simple form. However, the construction of an orthogonal parameterisation itself involves the solution of a partial differential equation. Thus, as pointed out by Levine & Casella (2003), in many real problems it will be impossible to solve the equations analytically and often extremely difficult to solve them numerically. As a consequence, Levine & Casella (2003) proposed a general algorithm for the implementation of probability matching priors for a single interest parameter in the presence of a single real nuisance parameter and showed how this algorithm works in some examples. A major drawback to their method of implementation is the substantial amount of computing time required.

In the present article we investigate an implementation of probability matching priors for interest parameters based on local approximations that avoids the need to solve the partial differential equation or to obtain an orthogonal parameterisation. A

local probability matching prior is a data-dependent approximation to a particular probability matching prior such that the posterior distributions under the two priors are the same to $O_p(n^{-1})$. Credible intervals for an interest parameter of posterior probability α calculated under a local prior will also be confidence intervals with confidence level α to $O_p(n^{-1})$. This is the same order of approximation as achieved by using a probability matching prior and therefore to the asymptotic order considered a local matching prior will serve just as well for the construction of confidence intervals. Finally, the method described here can be implemented in parameter spaces of dimension greater than two.

2. PROBABILITY MATCHING PRIORS FOR AN INTEREST PARAMETER

We begin by reviewing the partial differential equation satisfied by probability matching priors for an interest parameter in the presence of nuisance parameters. We then obtain a general form of solution that can be approximated in a simple way, leading to a local probability matching prior.

The parametric framework is as follows. Let Y^n be a dataset from a distribution known apart from a parameter $\theta = (\theta_1, \ldots, \theta_d) \in \Omega$, an open subset of $\mathcal{R}^d$. Suppose that a prior density $\pi(\cdot)$ is available for θ. Then the posterior density of θ is

$$\pi(\theta|Y^n) \propto \pi(\theta) L_n(\theta), \tag{1}$$

where $L_n(\theta)$ is the likelihood function associated with Y^n. We will assume sufficient regularity conditions on the likelihood function and prior density for the validity of the asymptotic analysis leading to the partial differential equation below. Write $l_n(\theta) = \log L_n(\theta)$. Suppressing the dependence on n, let $J(\theta)$ be the negative Hessian matrix of $l_n(\theta)$ and $I(\theta) = E_\theta\{J(\theta)\}$, Fisher's information matrix. Suppose that θ_1 is considered to be the parameter of primary interest and that $(\theta_2, \ldots, \theta_d)$ is the vector of nuisance parameters. Let $z_\alpha(Y^n)$ be the α quantile of the marginal posterior distribution of θ_1 under π; that is $z_\alpha(Y^n)$ satisfies $\text{pr}_\pi\{\theta_1 \leq z_\alpha(Y^n)|Y^n\} = \alpha$. The prior π is said to be probability matching if it is also true that

$$\text{pr}_\theta\{\theta_1 \leq z_\alpha(Y^n)\} = \alpha + O_p(n^{-1}) \tag{2}$$

for every α, $0 < \alpha < 1$.

Partition the information matrix as

$$I(\theta) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix},$$

where I_{11} and I_{22} are the submatrices corresponding to θ_1 and $(\theta_2, \ldots, \theta_d)$ respectively, and define

$$I_{11.2} = I_{11} - I_{12}I_{22}^{-1}I_{21}.$$

Let $\kappa_{ij}(\theta)$ be the (i,j)th element of $n^{-1}I(\theta)$. In particular, $\kappa_{11} = n^{-1}I_{11}$. In the case where Y^n consists of n independent and identically distributed observations $Y_1, \ldots, Y_n$, Peers (1965) showed that π is a probability matching prior for θ_1 if and only if it is a solution of the partial differential equation

$$D_j\{(\kappa^{11})^{-1/2}\kappa^{1j}\pi\} = 0, \tag{3}$$

where $D_j = \partial/\partial\theta_j$, $\kappa^{ij}(\theta)$ is the (i,j)th element of $n\{I(\theta)\}^{-1}$ and we have used the summation convention. If θ_1 and $(\theta_2, \ldots, \theta_d)$ are orthogonal (Cox & Reid, 1987) then it follows immediately from (3) that

$$\pi(\theta) \propto \{\kappa^{11}(\theta)\}^{-1/2}h(\theta_2, \ldots, \theta_d) \propto \{I_{11}(\theta)\}^{1/2}h(\theta_2, \ldots, \theta_d),$$

where the function h is arbitrary, as shown by Tibshirani (1989). We will consider solutions of (3) in the general case where θ_1 and $(\theta_2, \ldots, \theta_d)$ are not necessarily orthogonal.

Let $\lambda = \log\{(\kappa^{11})^{1/2}\pi\}$. Then equation (3) may be re-expressed as

$$\alpha^j D_j(\lambda) + D_j(\alpha^j) = 0, \tag{4}$$

where $\alpha^j = (\kappa^{11})^{-1}\kappa^{1j}$. Since $\alpha^1 = 1$, the subsidiary equations of this Lagrange equation are of the form

$$\frac{d\theta_j}{d\theta_1} = \alpha^j$$

for $j = 2, \ldots, d$. We shall assume that a parameterisation has been chosen that satisfies

$$\alpha^j = \alpha^j(\theta_1, \theta_j) \tag{5}$$

for $j = 2, \ldots, d$. Clearly (5) will be true in any parameterisation when $d = 2$, which is the case treated in Levine & Casella (2003). When $d > 2$ there will exist many parameterisations that achieve (5). In particular, under an orthogonal parameterisation

4

we have $\alpha^j = 0$, $j = 2, \ldots, d$. The subsidiary equations have solutions of the form $\psi_j(\theta_1, \theta_j) = \text{constant}$. Define $\eta_1 = \theta_1$, $\eta_j = \psi_j(\theta_1, \theta_j)$ for $j = 2, \ldots, d$, with inverse transformation $\theta_1 = \eta_1$, $\theta_j = \gamma_j(\eta_1, \eta_j)$. Then

$$\frac{\partial \lambda}{\partial \eta_1} = \sum_{j=1}^{d} \frac{\partial \lambda}{\partial \theta_j} \frac{\partial \theta_j}{\partial \eta_1} = \alpha^j D_j(\lambda) = -D_j(\alpha^j)$$

from equation (4). Write $v(\eta) = D_j(\alpha^j)$. Since $(\kappa^{11})^{-1} = n^{-1} I_{11.2}$, it follows that solutions to (4) are of the form

$$\pi(\theta) \propto \{I_{11.2}(\theta)\}^{1/2} e^{-\phi(\eta)}, \tag{6}$$

where $\phi(\eta) = \int v(\eta) d\eta_1$ is any indefinite integral of $v(\eta)$ with respect to η_1. Thus an arbitrary multiplicative function of $(\eta_2, \ldots, \eta_d)$ is implicit in (6).

Let $\hat{\eta}$ and $\hat{\theta}$ be the maximum likelihood estimators of η and θ respectively, let θ_{10} be the true value of θ_1 and write $\phi_i = \partial \phi / \partial \eta_i$. Then, for η within $O(n^{-1/2})$ neighbourhoods of $\hat{\eta}$, by Taylor expansion we have

$$
\begin{aligned}
\phi(\eta) &= \phi(\hat{\eta}) + (\eta_1 - \hat{\eta}_1) v(\hat{\eta}) + \sum_{j=2}^{d} (\eta_j - \hat{\eta}_j) \phi_j(\hat{\eta}) + O_p(n^{-1}) \\
&= \phi(\hat{\eta}) + (\theta_1 - \hat{\theta}_1) v(\hat{\eta}) + \sum_{j=2}^{d} (\eta_j - \hat{\eta}_j) \phi_j(\theta_{10}, \hat{\eta}_2, \ldots, \hat{\eta}_d) + O_p(n^{-1})
\end{aligned}
$$

since $\hat{\theta}_1 - \theta_{10} = O_p(n^{-1/2})$. Now write $u(\theta) = v(\eta) = D_j(\alpha^j)$. Since $v(\hat{\eta}) = u(\hat{\theta})$ it follows that

$$\phi(\eta) = \phi(\hat{\eta}) + (\theta_1 - \hat{\theta}_1) u(\hat{\theta}) + \phi(\theta_{10}, \eta_2, \ldots, \eta_d) - \phi(\theta_{10}, \hat{\eta}_2, \ldots, \hat{\eta}_d) + O_p(n^{-1}). \tag{7}$$

We now define the data-dependent prior

$$\pi_l(\theta) \propto \{I_{11.2}(\theta)\}^{1/2} e^{-\theta_1 u(\hat{\theta})}. \tag{8}$$

Since $\int_{\theta_{10}}^{\theta_1} v(\eta) d\eta_1 = \phi(\theta_1, \eta_2, \ldots, \eta_d) - \phi(\theta_{10}, \eta_2, \ldots, \eta_d) = (\theta_1 - \theta_{10}) u(\hat{\theta}) + O_p(n^{-1})$ from (7), it follows that (8) is an $O_p(n^{-1})$ local approximation to the prior

$$\pi_0(\theta) \propto \{I_{11.2}(\theta)\}^{1/2} e^{-\int_{\theta_{10}}^{\theta_1} v(\eta) d\eta_1}.$$

Despite the dependence on the unknown true value θ_{10}, the prior π_0 is one version of the probability matching prior (6), so that (2) holds for this prior and hence for its local approximation (8). We refer to (8) as a local probability matching prior.

Although the probability matching prior (6) was derived by choosing a particular orthogonal parameterisation $\eta_2, \ldots, \eta_d$, the local probability matching prior (8) is expressed in terms of the original parameterisation θ, making it tractable for practical implementation. The local prior (8) is not invariant to reparameterisation of θ_1. However, alternative versions will all be equivalent to $O_p(n^{-1})$. Furthermore, choosing alternative parameterisations for $(\theta_2, \ldots, \theta_d)$ that satisfy (5) will lead to local approximations to other probability matching priors. Finally, although Peers (1965) obtained the partial differential equation (3) for the case of independent and identically distributed observations, (3) will hold more generally when Y^n is a dataset indexed by some parameter n under suitable regularity conditions on the likelihood function.

3. COMPUTATIONAL ASPECTS

In applications it may be straightforward to obtain the partial derivatives of α^j, and hence the function u, directly. Alternatively, these may be more conveniently obtained from the partial derivatives of the elements of the information matrix. In more complex applications the information matrix will not be available analytically. However, it turns out that we can usually obtain an $O_p(n^{-1})$ equivalent local probability matching prior to (8) by expressing the required derivatives in terms of observed quantities.

We begin by noting the identities

$$
\begin{aligned}
D_j\{(\kappa^{11})^{-1}\} &= (\kappa^{11})^{-2}\kappa^{1r}\kappa^{1s}D_j(\kappa_{rs}) \\
D_j(\kappa^{1j}) &= -\kappa^{1r}\kappa^{js}D_j(\kappa_{rs}).
\end{aligned}
$$

Using the product rule and substituting these identities gives

$$
\begin{aligned}
u = D_j(\alpha^j) &= (\kappa^{11})^{-2}\kappa^{1r}(\kappa^{1j}\kappa^{1s} - \kappa^{11}\kappa^{js})D_j(\kappa_{rs}) \\
&= \alpha^r(\kappa^{11}\alpha^j\alpha^s - \kappa^{js})D_j(\kappa_{rs}).
\end{aligned} \tag{9}
$$

We will use this formula in the example discussed in §5.2.

If the information matrix is not available analytically, or if we simply wish to avoid calculating $I(\theta)$ and its derivatives, it will usually be possible to estimate the required

quantities from the data. Write $w_j(\theta) = \frac{1}{2}D_j(\log \kappa^{11})$. Performing a Taylor expansion of $\log\{I_{11.2}(\theta)\}$ about $\theta = \hat{\theta}$, we obtain an $O_p(n^{-1})$ equivalent local probability matching prior to (8),

$$\pi_l(\theta) \propto \exp\{-\theta_1 u(\hat{\theta}) - \sum_{j=1}^{d} \theta_j w_j(\hat{\theta})\} . \tag{10}$$

Now note that

$$D_j(\log \kappa^{11}) = -\kappa^{11}\alpha^r\alpha^s D_j(\kappa_{rs}) . \tag{11}$$

It follows from (10), (9) and (11) that it will be enough to use $O_p(n^{-1/2})$ estimators of $\kappa_{rs}(\theta)$ and $D_j\{\kappa_{rs}(\theta)\}$, all evaluated at $\theta = \hat{\theta}$. For $n\kappa_{rs}(\hat{\theta})$ we can simply use the (r, s)th element of the observed information matrix $J = J(\hat{\theta})$, while for $D_j\{\kappa_{rs}(\theta)\}$ note that

$$D_j(\kappa_{rs}) = n^{-1}E_\theta(l_{jr}l_s + l_{js}l_r + l_j l_r l_s) . \tag{12}$$

If the data consist of n independent and identically distributed observations then, to $O_p(n^{-1/2})$, each of the three terms on the right-hand side of (12) may be estimated by using the averages of the corresponding observed quantities evaluated at $\hat{\theta}$. Otherwise we may be able to estimate these terms using suitable averages from the data, as we illustrate in §5.2. We emphasise that the above modifications still give rise to a valid local probability matching prior.

Once we have obtained a suitable local probability matching prior, any of the usual Bayesian computational strategies can be adopted to obtain the posterior distribution of θ_1. In simple cases analytical solutions may be available. Otherwise a Markov chain Monte Carlo scheme may be applied in the usual way, as illustrated in the examples in §5. Since probability matching is only to $O_p(n^{-1})$, an alternative strategy would be to use an $O_p(n^{-1})$ Laplace approximation to the posterior distribution of θ_1; see for example Tierney & Kadane (1986) or Sweeting (1996). One version of this takes the form

$$\pi_l\{\theta_1|Y^n\} \propto \pi_l\{\theta_1, \hat{\theta}_2(\theta_1)\}L_n\{\theta_1, \hat{\theta}_2(\theta_1)\}|J_{22}\{\theta_1, \hat{\theta}_2(\theta_1)\}|^{-1/2} \tag{13}$$

uniformly to $O_p(n^{-1})$, where $\hat{\theta}_2(\theta_1)$ is the conditional maximum likelihood estimator of θ_2 for fixed θ_1 and the matrix $J(\theta)$ is partitioned in exactly the same way as the information matrix $I(\theta)$. Alternatively, and to the same order of approximation, we

can approximate $\hat{\theta}_2(\theta_1)$ by $\bar{\theta}_2(\theta_1) = \hat{\theta}_2 - (\theta_1 - \hat{\theta}_1)J_{22}^{-1}J_{21}$, which may be computation-ally more convenient. The above Laplace approximation under the local probability matching prior (10) then takes the form

$$\pi_l(\theta_1|Y^n) \propto \exp[-\theta_1\{u(\hat{\theta}) + w_1(\hat{\theta}) - w(\hat{\theta})J_{22}^{-1}J_{21}\}]L_n\{\theta_1, \bar{\theta}_2(\theta_1)\}|J_{22}\{\theta_1, \bar{\theta}_2(\theta_1)\}|^{-1/2},\tag{14}$$

where $w(\theta) = (w_2(\theta), \ldots, w_d(\theta))$. Approximate posterior quantiles of θ_1, and hence approximate confidence intervals for θ_1, can now be obtained from (13) or (14) using one-dimensional numerical integration, a Monte Carlo scheme or further asymptotic analysis. The latter case is discussed in Sweeting (1996), where a standardised signed-root log likelihood ratio statistic $\tilde{R}_1$ associated with θ_1 is defined. The α posterior quantile of θ_1 is computed as $\tilde{R}_1^{-1}(z_\alpha)$ to $O_p(n^{-1})$, where z_α is the α quantile of the standard normal distribution. This scheme is implemented in §5.2.

Finally in this section we note that the use of any $n^{1/2}$-consistent estimator of θ in place of the maximum likelihood estimator $\hat{\theta}$ will not change the order of approximation in (2). Alternatively, any consistent estimator of θ would produce (2) with $O_p(n^{-1})$ replaced by $o_p(n^{-1/2})$. It might be advantageous to use such alternative estimators if the likelihood function is not easily maximised or is poorly behaved in finite samples.

4. PRELIMINARY EXAMPLES

In this section we obtain local probability matching priors in three examples for which analytic solutions are available. These examples are discussed in Datta & M. Ghosh (1995b).

Example 1. Suppose that $Y_i \sim N(\mu, \sigma^2)$, $Y^n = (Y_1, \ldots, Y_n)$ and that $\theta_1 = \mu + \frac{1}{2}\sigma^2 = \log E\{\exp(Y_i)\}$ is the parameter of interest. If we take $\theta_2 = \sigma^2$, the information matrix is

$$I(\theta) = n\begin{pmatrix} \sigma^{-2} & -\frac{1}{2}\sigma^{-2} \\ -\frac{1}{2}\sigma^{-2} & \frac{1}{2}\sigma^{-4}(1 + \frac{1}{2}\sigma^2) \end{pmatrix},$$

from which we obtain $(\kappa^{11})^{-1} = n^{-1}I_{11.2} = \sigma^{-2}(1 + \frac{1}{2}\sigma^2)^{-1}$ and $\alpha^2 = (\kappa^{11})^{-1}\kappa^{12} = \sigma^2(1 + \frac{1}{2}\sigma^2)^{-1}$. Therefore $u(\theta) = D_2(\alpha^2) = (1 + \frac{1}{2}\sigma^2)^{-2}$, and from equation (8) a local probability matching prior is given by

$$\pi_l(\theta) \propto \sigma^{-1}(1 + \frac{1}{2}\sigma^2)^{-1/2}e^{-\theta_1(1 + \frac{1}{2}s^2)^{-2}},\tag{15}$$

8

where $s^2 = n^{-1}\sum_i(Y_i - \bar{Y})^2$ and $\bar{Y} = n^{-1}\sum_i Y_i$ are the maximum likelihood estimators of σ^2 and μ respectively.

In this problem $\phi = \mu - \log\sigma^2$ is orthogonal to θ_1 and probability matching priors for θ_1 are of the form

$$\pi(\theta) \propto \sigma^{-3}(1 + \tfrac{1}{2}\sigma^2)^{1/2}h(\theta_1 - \tfrac{1}{2}\sigma^2 - \log\sigma^2)\,, \tag{16}$$

where h is an arbitrary function (Datta & M. Ghosh, 1995b). Although not immediately obvious from (15) and (16), if we express both formulae in terms of (θ_1, ϕ) and carry out a first-order Taylor expansion, it can be seen directly that (15) is indeed locally equivalent to a prior of the form (16) to $O_p(n^{-1})$.

Example 2. Suppose again that $Y_i \sim N(\mu, \sigma^2)$ but that now $\theta_1 = \sigma^{-1}\mu$ is the parameter of interest. If we take $\theta_2 = \sigma$, the information matrix is

$$I(\theta) = \begin{pmatrix} 1 & \sigma^{-1}\theta_1 \\ \sigma^{-1}\theta_1 & 2\sigma^{-2}(1 + \tfrac{1}{2}\theta_1^2) \end{pmatrix}\,,$$

from which we obtain $(\kappa^{11})^{-1} = n^{-1}I_{11.2} = (1 + \tfrac{1}{2}\theta_1^2)^{-1}$ and $\alpha^2 = (\kappa^{11})^{-1}\kappa^{12} = -\tfrac{1}{2}\sigma\theta_1(1 + \tfrac{1}{2}\theta_1^2)^{-1}$. Therefore $u(\theta) = D_2(\alpha^2) = -\tfrac{1}{2}\theta_1(1 + \tfrac{1}{2}\theta_1^2)^{-1}$, so from (8) a local probability matching prior is given by

$$\pi_l(\theta) \propto (1 + \tfrac{1}{2}\theta_1^2)^{-1/2}e^{\frac{1}{2}\theta_1 s\bar{Y}(s^2 + \frac{1}{2}\bar{Y}^2)^{-1}}\,,$$

where $\bar{Y}$ and s^2 are defined in Example 1.

In this example it turns out that all priors of the form

$$\pi(\theta) \propto h(\phi) \tag{17}$$

are probability matching, where $\phi = \sigma(1 + \tfrac{1}{2}\theta_1^2)^{1/2}$ and h is an arbitrary function. Again, by Taylor expansion, it can be shown that π_l is locally of this form. Also, as previously noted we will obtain different local matching priors depending on the parameterisation used for θ_2. For example, if we had worked with the parameterisation (θ_1, ϕ) here, then we would have obtained a uniform prior in the (θ_1, σ) parameterisation, which is also a probability matching prior. From a Bayesian point of view there may be additional reasons for preferring one form of matching prior to another. For example, the reference prior (Bernardo, 1979) for which $h(\phi) = \phi$ in (17) avoids the marginalisation paradox; see for example the discussion in Datta & M. Ghosh

(1995b). However, in the more complex applications we envisage, it is likely to be extremely difficult, if not impossible, to verify such additional properties.

Example 3. As a final example in this section we consider the one-way random effects model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $j = 1, \ldots, n$, $i = 1, \ldots, k$, where τ_i and ϵ_{ij} are mutually independent random variables with $\tau_i \sim N(0, \sigma_\tau^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$. Let $\bar{Y}_i = n^{-1} \sum_j Y_{ij}$ and $\bar{Y} = k^{-1} \sum_i \bar{Y}_i$. Objective priors for this model are discussed in Berger & Bernardo (1992), Ye (1994) and Datta & M. Ghosh (1995a, b). For comparison with Levine & Casella (2003) we suppose that the parameter of interest is $\theta_1 = \sigma^2$ and take $\theta_2 = \tau^2$, $\theta_3 = \mu$. Then, from Datta & M. Ghosh (1995b), the nonzero elements of the information matrix are $\kappa_{11} = \frac{1}{2}\{(n-1)\theta_1^{-2} + (\theta_1 + n\theta_2)^{-2}\}$, $\kappa_{12} = \kappa_{21} = \frac{1}{2}kn(\theta_1 + n\theta_2)^{-2}$, $\kappa_{22} = \frac{1}{2}kn^2(\theta_1 + n\theta_2)^{-2}$ and $\kappa_{33} = kn(\theta_1 + n\theta_2)^{-1}$. It follows that $(\kappa^{11})^{-1} = n^{-1}I_{11.2} = \frac{1}{2}(k-1)\theta_1^{-2}$, $\alpha^2 = k^{-1}$ and $\alpha^3 = 0$, giving $u(\theta) = 0$ and hence the local matching prior

$$\pi_l(\theta) \propto \theta_1^{-1},$$

which is an exactly probability matching prior for θ_1 in this case. This example was discussed in Levine & Casella (2003), assuming μ to be known.

5. ANALYSIS OF TWO EXAMPLES

In this section we consider two examples discussed in Levine & Casella (2003). Since both these examples involve discrete observations, we note that probability matching is not actually to second order. However, this order is correct if we consider asymptotic coverage probabilities in a very weak sense (see for example the discussion in Section 2 of Datta & Sweeting, 2004). An alternative would be to include continuity corrections based on uniform perturbations (Rousseau, 2000).

5.1 Logistic regression example

The data for this example (Levine & Casella, 2003) are taken from Hosmer & Lemeshow (2000, Table 1.1). The presence or absence of coronary heart disease was recorded for one hundred subjects and the relationship between coronary heart disease status, y, and age, x, was investigated using a logistic regression model. As in Levine & Casella (2003), we suppose that the unknown effect of age, θ_1, is the parameter of interest and that the intercept, θ_2, is a nuisance parameter. As noted by Levine

10

& Casella, analytic solutions to the probability matching prior equations (3) are not available and numerical solution appears to be difficult.

The information matrix for grouped binary data under the logistic regression model is

$$I(\theta) = \sum_{i=1}^{k} h_i(\theta) \begin{pmatrix} x_i^2 & x_i \\ x_i & 1 \end{pmatrix} ,$$

where

$$h_i(\theta) = \frac{n_i e^{\theta_2 + \theta_1 x_i}}{(1 + e^{\theta_2 + \theta_1 x_i})^2} ,$$

k is the number of age groups and n_i is the number of subjects in the ith age group, $i = 1, \ldots, k$. After some algebra we obtain $I_{11.2}(\theta) = R_{xx}(\theta)$ and $u(\theta) = 2T_{xp}(\theta)$, where $T_{xp}(\theta) = \{\sum_i h_i(\theta)\}^{-1} R_{xp}(\theta)$,

$$\begin{aligned} R_{xx}(\theta) &= \sum_i h_i(\theta)\{x_i - \bar{x}(\theta)\}^2 \\ R_{xp}(\theta) &= \sum_i h_i(\theta)\{x_i - \bar{x}(\theta)\}\{p_i(\theta) - \bar{p}(\theta)\} , \end{aligned}$$

$\bar{x}(\theta) = \{\sum_i h_i(\theta)\}^{-1} \sum_i h_i(\theta) x_i$, $\bar{p}(\theta) = \{\sum_i h_i(\theta)\}^{-1} \sum_i h_i(\theta) p_i(\theta)$ and $p_i(\theta) = \{1 + \exp(-\theta_2 - \theta_1 x_i)\}^{-1}$ is the probability of disease in the ith age group. From (8) a local probability matching prior that gives correct coverage probabilities for θ_1 to $O_p(n^{-1})$, where $n = \sum_i n_i$, is therefore

$$\pi_l(\theta) \propto \{R_{xx}(\theta)\}^{1/2} e^{-2\theta_1 T_{xp}(\hat{\theta})} . \tag{18}$$

With this prior, the logistic regression model was fitted to the coronary heart disease data by performing a random walk Metropolis-Hastings routine (Robert & Casella, 1999), which was run for 25000 iterations following a burn-in period. This yielded an approximate equi-tailed 95% confidence interval for θ_1 of 0.064 to 0.160. For comparison, the posterior distribution of θ_1 was also obtained using the signed-root log likelihood ratio approximation based on the Laplace approximation (13), as described in §3. This yielded values of 0.065 and 0.159, which are close to the above figures. In order to see the effect of the prior distribution, the equi-tailed 95% Bayesian credible region under a uniform prior for θ was calculated using the signed-root approximation. This was found to be 0.069 to 0.166. Since the coverage probability bias of equi-tailed Bayesian regions is $O_p(n^{-1})$ for every smooth prior (Hartigan, 1966), a uniform prior will also produce a 95% confidence interval for θ_1 to $O_p(n^{-1})$. Indeed, using the signed-root approximation again, we obtained the posterior probability of the above interval

11

under the local matching prior as approximately 0.948. However, the directional coverage errors were approximately 3.8% and 1.4%.

In order to assess the accuracy of confidence intervals produced by the local matching prior (18), as in Levine & Casella (2003) we simulated smaller datasets having thirty subjects divided into three age groups. One thousand datasets were simulated and the 95% equi-tailed posterior intervals for θ_1 under the local matching prior (18) were calculated, each by using a random walk Metropolis-Hastings algorithm run for 25000 iterations following burn-in periods. The overall coverage rate was 94.8%, with 3.0% and 2.2% in the lower and upper tail respectively, indicating good directional coverage.

5.2 Beta-binomial example

As a final example we consider the beta-binomial example discussed in Levine & Casella (2003), which is a somewhat more demanding problem since the information matrix is not available analytically. We implement the modified version of the local matching prior obtained in §3.

The data, taken from Williams (1975), are from an experiment consisting of two groups of 16 pregnant female rats given control and test chemical diets respectively. Let n_{ij} and Y_{ij} be the number of rats from the jth litter in group i that were alive at four days and, respectively, that survived the 21-day lactation period. A beta-binomial distribution is adopted, where, for $j = 1, \ldots, n$ and $i = 1, 2$, $Y_{ij}|p_{ij} \sim \mathrm{Bi}(n_{ij}, p_{ij})$, $p_{ij} \sim \mathrm{Be}(\alpha_i, \beta_i)$ and α_i, β_i are unknown parameters. The expected proportion of survivals in group i is $\mu_i = \alpha_i/(\alpha_i + \beta_i)$. As in Levine & Casella (2003) we suppose that μ_1 and μ_2 are the two parameters of interest here. Since the likelihood contributions for the two groups are orthogonal we can consider the groups separately.

In this example it is not possible to obtain an analytical form for the information matrix. We therefore use the version of the local probability matching prior based on observed quantities described in §3. For each group separately each of the quantities on the right-hand side of (12) was estimated to $O_p(n^{-1/2})$ by the averages of the corresponding observed quantities for each of the 16 rats in the group; we omit the details. Furthermore, in this example it is simplest to evaluate the above quantities at moment estimates of θ, since the first two moments of the beta-binomial distribution

are available in closed form. In fact it made hardly any difference here whether the moment or maximum likelihood estimates were used.

We parameterise the problem by $\theta_{i1} = \log\{\mu_i/(1 - \mu_i)\}$ and $\theta_{i2} = \log(\alpha_i + \beta_i)$, $i = 1, 2$. Using the additional approximations to (8) described above for each group, we found that the resulting local prior for these data that gives probability matching to $O_p(n^{-1})$ for both μ_1 and μ_2 is

$$\pi_l(\theta) \propto e^{1.182\theta_{11} + 0.129\theta_{12} - 1.107\theta_{21} + 0.023\theta_{22}} . \tag{19}$$

Two random walk Metropolis-Hastings samplers, one for each group, were run for 50000 iterations following burn-in periods in order to achieve sufficient Monte Carlo precision. The posterior mean of μ_1 was 0.90 with five and ninety-five percentiles 0.86 and 0.94 respectively, while the posterior mean of μ_2 was 0.71 with five and ninety-five percentiles 0.57 and 0.82 respectively. The results for the control group are approximately the same as those obtained by Levine & Casella (2003), although for the treatment group the mean is slightly lower and the 90% confidence interval much wider. Similar results were obtained in repeat runs. In terms of the parameters $\alpha_i, \beta_i, i = 1, 2$, the local matching prior (19) becomes

$$\pi_l(\alpha_1, \beta_1, \alpha_2, \beta_2) \propto \alpha_1^{0.182}\beta_1^{-2.182}(\alpha_1 + \beta_1)^{0.129}\alpha_2^{-2.107}\beta_2^{0.107}(\alpha_2 + \beta_2)^{0.023} .$$

Note that there is no reason to expect the coefficients of α_i and β_i to be similar in the two groups since the two likelihood functions are concentrated in quite different regions of the parameter space. The two local priors will therefore be estimating different versions of the probability matching prior (6). In fact the maximum likelihood estimates of α_1 and β_1 here are 43.56 and 4.95 respectively, while those of α_2 and β_2 are 1.592 and 0.559 respectively.

Finally, as noted by Levine & Casella, the two posterior samples could also be used to obtain a credible interval for the log-odds $\log \Psi = \theta_{21} - \theta_{11}$. However, if an $O_p(n^{-1})$ confidence interval for $\log \Psi$ is required then this parameter needs to be declared as the parameter of interest and the associated local probability matching prior obtained for the four-parameter problem.

6. DISCUSSION

In this paper we have demonstrated a method for the implementation of probability matching priors for one parameter of many via local solutions to the probability matching prior partial differential equation. Although in particular applications some analytical work is needed to obtain the local prior (8), we have seen in §3 that the process can be fully automated in the sense that the local probability matching prior contains only observed quantities calculated from components of the likelihood function.

As discussed in the paper, alternative nuisance parameterisations will give rise to different versions of the local matching prior. Sometimes there will be a good reason for preferring a particular parameterisation, such as avoidance of a marginalisation paradox, or to achieve higher-order matching (Mukerjee & Ghosh, 1997). However, it will be difficult to identify such additional properties in more complex problems and an appropriate parameterisation should be guided by experience. For example, there may be a natural parameterisation for the model or a parameterisation similar to one known to have good properties in simpler models.

Peers (1965) showed that in general there will not exist a probability matching prior that matches posterior and coverage probabilities simultaneously to $O_p(n^{-1})$ when there is more than one parameter of interest. When such a prior does exist it is referred to as a simultaneous (marginal) probability matching prior (Datta, 1996). Initial results indicate that it may be possible to construct local priors with the simultaneous matching property even when no simultaneous probability matching prior exists. It would also be of interest to obtain local versions of other forms of probability matching priors, such as joint probability matching priors (Datta, 1996) and predictive matching priors (Datta et al., 2000).

ACKNOWLEDGEMENT

REFERENCES

BERGER, J. & BERNARDO, J. M. (1992). Reference priors in a variance components problem. In *Bayesian Analysis in Statistics and Econometrics*, Ed. P. K. Goel and N. S. Iyengar, pp. 177–94. New York: Springer-Verlag.

BERNARDO, J. M. (1979). Reference posterior distributions for Bayesian inference (with Discussion). *J. R. Statist. Soc.* B **41**, 113–47.

BERNARDO, J. M. & RAMÓN, J. M. (1998). An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *Statistician* **47**, 101–35.

COX, D. R. & REID, N. (1987). Parameter orthogonality and approximate conditional inference (with Discussion). *J. R. Statist. Soc.* B **53**, 79–109.

DATTA, G. S. (1996). On priors providing frequentist validity of Bayesian inference for multiple parametric functions. *Biometrika* **83**, 287–98.

DATTA, G. S. & GHOSH, J. K. (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika* **82**, 37–45.

DATTA, G. S. & GHOSH, M. (1995a). Hierarchical Bayes estimators for the error variance in one-way ANOVA models. *J. Statist. Plan. Infer.* **45**, 399–411.

DATTA, G. S. & GHOSH, M. (1995b). Some remarks on noninformative priors. *J. Am. Statist. Assoc.* **90**, 1357–63.

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

DATTA, G. S., MUKERJEE, R., GHOSH, M. & SWEETING, T. J. (2000). Bayesian prediction with approximate frequentist validity. *Ann. Statist.* **28**, 1414–26.

DATTA, G. S. & SWEETING, T. J. (2004). Probability matching priors. In *Bayesian Statistics: Modelling and Computation*, Ed. C. R. Rao and D. K. Dey. Elsevier Science. (To appear)

GHOSH, J. K. & MUKERJEE, R. (1992). Non-informative priors (with Discussion). In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 195–210. Oxford University Press.

HARTIGAN, J. A. (1966). Note on the confidence-prior of Welch and Peers. *J. R. Statist. Soc.* B **28**, 55–6.

HOSMER, D. W. & LEMESHOW, S. (2000). *Applied Logistic Regression*, 2nd ed. New York: Wiley.

KASS, R. E. & WASSERMAN, L. (1996). The selection of prior distributions by formal rules. *J. Am. Statist. Assoc.* **91**, 1343–70.

LEVINE, R. A. & CASELLA, G. (2003). Implementing matching priors for frequentist inference. *Biometrika* **90**, 127–37.

MUKERJEE, R. & GHOSH, M. (1997). Second-order probability matching priors. *Biometrika* **84**, 970–5.

PEERS, H. W. (1965). On confidence sets and Bayesian probability points in the case of several parameters. *J. R. Statist. Soc.* B **27**, 9–16.

ROBERT, C. P. & CASELLA, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

ROUSSEAU, J. (2000). Coverage properties of one-sided intervals in the discrete case and application to matching priors. *Ann. Inst. Statist. Math.*, **52**, 28–42.

STEIN, C. M. (1985). On the coverage probability of confidence sets based on a prior distribution. In *Sequential Methods in Statistics*, Banach Center Publications, **16**, Ed. R. Zieliński, pp. 485–514. Warsaw: PWN-Polish Scientific Publishers.

SWEETING, T. J. (1995). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **82**, 1–23.

SWEETING, T. J. (1996). Approximate Bayesian computation based on signed roots of log-density ratios (with Discussion). In *Bayesian Statistics 5*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 427–44. Oxford University Press.

SWEETING, T. J. (2001). Coverage probability bias, objective Bayes and the likelihood principle. *Biometrika* **88**, 657–75.

TIBSHIRANI, R.J. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604–8.

TIERNEY, L. & KADANE, J. (1986) Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Assoc.* **81**, 82–6.

WELCH, B. L. & PEERS, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. R. Statist. Soc.* B **35**, 318–29.

WILLIAMS, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**, 949–52.

YE, K. (1994). Bayesian reference prior analysis on the ratio of variances for balanced one-way random effect model. *J. Statist. Plan. Infer.* **41**, 267–80.