SOME ASPECTS OF MATCHING PRIORS

N. REID
University of Toronto

R. MUKERJEE
Indian Institute of Management

D.A.S. FRASER
University of Toronto

Priors for which Bayesian and frequentist inference agree, at least to some order of approximation, are called 'matching priors', and have been proposed as candidates for noninformative priors in Bayesian inference. We give an overview of the original work of Welch and Peers and some more recent developments.

1. Introduction

In the context of parametric inference, a matching prior is a prior for which posterior probability statements about the parameter also have an interpretation as confidence statements in the sampling model. The idea appears to have been proposed first by Lindley (1958). There have been several attempts to develop matching priors, starting with Welch and Peers (1963). Matching priors in principle hold the promise of providing a possible frequentist/Bayesian compromise and of providing default priors for routine use in Bayesian inference. This is attractive from some frequentist points of view because the Bayesian approach to inference provides a simple way to eliminate nuisance parameters, and typical frequentist approaches are rather more complicated. Default priors are attractive from some Bayesian points of view as they might be expected to be more widely accepted than subjective priors. In addition the inference from a default prior can be compared to that from priors developed otherwise, as a possible check on the robustness of the inference to the prior.

We consider here priors that lead to approximate matching, to some order of approximation in the sample size n. The terminology in the literature is not standardized, and we follow here the version used in Mukerjee and Reid (1999): we call a prior *first order matching* if it ensures approximate frequentist validity of a Bayesian posterior credible set with margin of error $O(n^{-1})$, and *second order matching* if it does so with a margin of error of $O(n^{-3/2})$. Since the posterior distribution is typically asymptotically normal for any choice of prior, it is only at the $O(n^{-1/2})$ term of the asymptotic expansion that matching leads to a class of priors. The relevant asymptotic expansions are typically in powers of $n^{-1/2}$, so first order matching actually involves the second order term in the expansion.

Welch and Peers (1963) showed that Jeffreys' prior is the unique first order matching prior in sampling from a model with a scalar parameter.

Peers (1965) showed that this result does not extend to models with vector parameters. There have been a number of results established that are at least partial analogues of Welch and Peers for the vector parameter case, but no single approach has emerged, and these more general matching priors can usually be constructed only for particular models. In Section 2 we review the results of Welch and Peers (1963) and Peers (1965), and also outline the Edgeworth expansion technique used in these and following papers. In Section 3 we review some extensions that have been established in the literature using this approach. There are a number of other approaches to default priors, different from frequentist matching, and some of these are also mentioned in Section 3.

In Section 4 we discuss a different approach, using recent results from the development of higher order asymptotic approximations using saddlepoint type expansions. This leads to a notion of strong matching, which requires priors that are dependent on the data. Data dependent priors were developed in the context of the transformed regression model by Box and Cox (1964), and Wasserman (2000) discusses the necessity of data dependent priors in mixture models. It seems likely that data dependent priors are needed in a general approach to matching priors.

For a review of matching priors from a different perspective, with more emphasis on the underlying partial differential equations, see Ghosh and Mukerjee (1998).

2. The Welch–Peers approach to matching priors

We assume throughout that $Y = (Y_1, \ldots, Y_n)$ is a sample of independent, identically distributed, observations with joint density $f(y \mid \theta)$, where $\theta \in \mathbb{R}^k$. We assume that we have a prior $\pi(\theta)$, and when $k = 1$ we denote by $\theta^{(1-\alpha)}(y)$ the $(1-\alpha)$ posterior quantile defined by

$$\mathrm{P}_{\theta|Y}\{\theta \le \theta^{(1-\alpha)}(y) \mid y\} = \int_{-\infty}^{\theta^{(1-\alpha)}(y)} \pi(\theta|y)\,\mathrm{d}\theta = 1-\alpha, \tag{2.1}$$

where $\pi(\theta \mid y)$ is the posterior density of θ, given y.

If the following were also true

$$\mathrm{P}_{Y|\theta}\{\theta^{(1-\alpha)}(Y) \ge \theta\} = \int 1\{\theta^{(1-\alpha)}(y) \ge \theta\} f(y \mid \theta)\,\mathrm{d}y = 1-\alpha$$

then Bayesian and frequentist inference for θ, in the form of one-sided posterior limits or one-sided confidence limits, would be in perfect agreement. This is the case in a simple location model $f(y \mid \theta) = f_0(y-\theta)$ when $\pi(\theta)$ is constant, but of course we cannot expect the result to hold more generally. However, Welch and Peers (1963) proved that if we require instead

$$\mathrm{P}_{Y|\theta}\{\theta^{(1-\alpha)}(Y) \ge \theta\} = 1-\alpha+O(n^{-1}), \tag{2.2}$$

there is a unique prior for which (2.2) holds, given by $\pi(\theta) \propto \{i(\theta)\}^{1/2}$, where $i(\theta)$ is the expected Fisher information in a single observation

$$i(\theta) = n^{-1} E_{Y|\theta}\{-\partial^2 \log f(y \mid \theta\}/\partial\theta^2).$$

This prior is improper, as are all matching priors, but most Bayesian approaches permit improper priors as long as the posterior is proper.

It is useful to review the steps in the establishment of this result in order to see the limitations that arise in trying to extend it. We denote the log-likelihood function based on the sample y by

$$\ell(\theta) = \ell(\theta; y) = \log f(y \mid \theta).$$

The posterior cumulative distribution function, $\Pi(\theta)$, is given by

$$\Pi(\theta|y) = \int_{-\infty}^{\theta} \frac{\exp\{\ell(t)\}\pi(t)\,\mathrm{d}t}{\int \exp\{\ell(t)\}\pi(t)\,\mathrm{d}t}.$$

Using the results of Johnson (1970), we can obtain an Edgeworth expansion for the posterior distribution, although it is convenient to write this as an expansion for the posterior of $\sqrt{n}(\theta-\hat\theta)/\hat\sigma$, where $\hat\theta$ is the solution to $\ell'(\hat\theta) = 0$ and

$$\hat\sigma^{-2} = j(\hat\theta) = -\frac{1}{n}\frac{\partial^2\ell(\theta; y)}{\partial\theta^2}\bigg|_{\theta=\hat\theta}$$

is the observed per-observation Fisher information. A Cornish–Fisher inversion of the resulting expansion leads to an expansion for the posterior quantile of the form

$$\begin{aligned}(2.3)\quad &\theta^{(1-\alpha)}(y)\\ &= \hat\theta + \frac{\hat\sigma}{\sqrt{n}}\left[z_\alpha + \frac{1}{\sqrt{n}}\{(z_\alpha^2+2)A_3(y) + A_1(y)\} + \frac{1}{n}u(z_\alpha, \pi, y) + \cdots\right],\end{aligned}$$

where z_α is the $(1-\alpha)$ quantile of the standard normal distribution. Expressions for A_3, A_1 and u are given explicitly in Mukerjee and Reid (1999, (2.8)). The probability on the left-hand side of (2.2) is computed from (2.3) and turns out to have the expansion

$$(2.4)\qquad 1-\alpha + \frac{1}{\sqrt{n}}\phi(z_\alpha)T_1(\pi,\theta) + \frac{1}{n}z_\alpha\phi(z_\alpha)T_2(\pi,\theta) + \cdots$$

where $\phi(\cdot)$ is the standard normal density,

$$(2.5)\qquad T_1(\pi,\theta) = \frac{1}{\pi(\theta)}\frac{\mathrm{d}}{\mathrm{d}\theta}[\{i(\theta)\}^{-1/2}\pi(\theta)],$$

and the form of T_2 can be obtained from Mukerjee and Dey (1993).

From (2.5) we see that $T_1 = 0$, and the posterior quantile thus has frequentist coverage $1-\alpha+O(n^{-1})$, if and only if $\pi(\theta) \propto i^{1/2}(\theta)$. It is shown in Welch and Peers (1963) (see also Mukerjee and Ghosh, 1997) that for this choice of prior, $T_2 = 0$ if and only if

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{E_{Y|\theta}\{\ell'(\theta)\}^3}{\{i(\theta)\}^{3/2}} = 0 \tag{2.6}$$

which is a condition on the model. Thus matching to the next order of approximation can be achieved only in special cases.

The derivation in Welch and Peers (1963) is slightly different from that outlined above, but they employed the same basic techniques. They required the exact posterior quantile (not our approximation given in (2.2)) to have frequentist coverage $1 - \alpha + O(n^{-1})$, by seeking a prior for which the posterior cumulative distribution function $\Pi(\theta|y)$ is distributed as a uniform $(0,1)$ random variable under the sampling model $f(y \mid \theta)$. They established an asymptotic expansion for the moment generating function of $\Phi\{\Pi(\theta|y)\}$ (where Φ is the normal cumulative distribution function) and compared the expansion to the moment generating function for a standard normal random variable.

Exactly the same steps can be followed for a component of a vector parameter $\theta = (\theta_1, \ldots, \theta_k)$, using the marginal posterior. We define $\theta_1^{(1-\alpha)}(y)$ by

$$\int_{-\infty}^{\theta_1^{(1-\alpha)}(y)} \pi_m(t \mid y)\,\mathrm{d}t = 1 - \alpha$$

where $\pi_m(\theta_1 \mid y)$ is the posterior marginal density for a single component. Cornish–Fisher inversion of the Edgeworth expansion for the marginal posterior distribution function leads to an expression for the posterior quantile as

$$\theta_1^{(1-\alpha)}(y) = \hat{\theta}_1 + \frac{\hat{\sigma}_{11}}{\sqrt{n}}\left\{ z_\alpha + \frac{1}{\sqrt{n}}u_1(z_\alpha, \pi, y) + \frac{1}{n}u_2(z_\alpha, \pi, y) + \cdots \right\} \tag{2.7}$$

where $\hat{\sigma}_{11}^2 = j^{11}(\hat{\theta})$ is the $(1,1)$ component of the inverse of the observed per observation Fisher information matrix, and expressions for u_1 and u_2 are given in Mukerjee and Reid (1999, (2.8)). This leads to an expansion for $\mathrm{P}_{Y|\theta}\{\theta_1^{(1-\alpha)}(Y) \geq \theta_1\}$ of the same form as (2.4), where $T_1 = 0$ if and only if

$$\frac{\partial}{\partial\theta_a}\{i^{11}(\theta)^{-1/2}i^{a1}(\theta)\pi(\theta)\} = 0. \tag{2.8}$$

In (2.8) $\{i^{ab}(\theta)\}$ is the inverse of the per observation expected Fisher information matrix, and summation over a from 1 to k is implied. This result was obtained by Peers (1965), using the method of Welch and Peers (1963).

Expression (2.8) can be simplified if we assume that θ_1 is orthogonal to $(\theta_2, \ldots, \theta_k)$ with respect to expected Fisher information, i.e., $i_{1a}(\theta) = 0$ for $a = 2, \ldots, k$, in which case $T_1 = 0$ if and only if

$$\pi(\theta) \propto g(\theta_2, \ldots, \theta_k)\{i_{11}(\theta)\}^{1/2} \tag{2.9}$$

where $g(\cdot)$ is an arbitrary smooth function. These priors were derived from a slightly different approach in Tibshirani (1989), and a detailed discussion and derivation is given in Nicolau (1993).

One might hope that a unique solution to $\pi(\theta)$ could be established by applying (2.8) in turn to each component of θ, but the resulting set of k differential equations will not in general be mutually consistent, as shown by Peers (1965). This is apparent from (2.9) as well. While it is always possible to find a reparametrization that orthogonalizes one component of θ to the remaining components, it is not possible to find a single transformation that orthogonalizes all the components to each other (Cox and Reid, 1987).

One approach to choosing from among the set of priors satisfying (2.9) is to consider the next order term T_2 in (2.4). It is shown in Mukerjee and Ghosh (1997) that $T_2 = 0$ if and only if

$$\frac{1}{6} g(\theta_{(2)}) \frac{\partial}{\partial \theta_1}\{i_{11}(\theta)^{-3/2} i_{1,1,1}(\theta)\} + \frac{\partial}{\partial \theta_a}\frac{\partial}{\partial \theta_b}\{i_{11}(\theta)^{-1/2} i_{11a}(\theta) i^{ab}(\theta) g(\theta_{(2)})\} = 0$$

where summation over a and b from 2 to k is implied, we have used the notation $\theta_{(2)} = (\theta_2, \ldots, \theta_k)$ and the third order per observation information functions are

$$i_{11a}(\theta) = \frac{1}{n} E_{Y|\theta}\left\{ \frac{\partial^3 \ell(\theta)}{\partial \theta_1 \partial \theta_1 \partial \theta_a} \right\}$$

$$i_{1,1,1}(\theta) = \frac{1}{n} E_{Y|\theta}\left\{ \frac{\partial \ell(\theta)}{\partial \theta_1} \right\}^3 .$$

These conditions involve both the prior, through $g(\theta_{(2)})$, and the model. Mukerjee and Ghosh (1997) and Ghosh and Mukerjee (1998) give examples where a prior satisfying the first order matching condition (2.8) also satisfies $T_2 = 0$, where there is a first order matching prior but no second order matching prior, and where a class of first order matching priors is narrowed down by adding the second order matching condition. Peers (1965) also noted the impossibility, in general, of using the second order matching condition to choose among priors satisfying the first order matching condition.

Example (Mukerjee and Ghosh, 1997). Suppose we are sampling from a bivariate normal distribution with mean (μ_1, μ_2), variances σ_1^2 and σ_2^2,

and correlation coefficient ρ. Define the parameter of interest θ_1 to be the regression parameter $\rho\sigma_2/\sigma_1$, and let $\theta_2 = \sigma_2^2(1-\rho^2)$, $\theta_3 = \sigma_1^2$, $\theta_4 = \mu_1$ and $\theta_5 = \mu_2$. The first order matching prior is of the form

$$\pi(\theta) \propto g(\theta_2, \theta_3, \theta_4, \theta_5)\left(\frac{\theta_3}{\theta_2}\right)^{1/2},$$

and invoking the second order matching criterion leads to the class of priors

$$\pi(\theta) \propto g(\theta_3, \theta_4, \theta_5)\theta_2^{-1}.$$

3. Other matching criteria

In view of the general non-existence of Welch–Peers type matching priors, several other approaches have been investigated. We give a very brief summary here. The review paper by Ghosh and Mukerjee (1998) provides more detail, along with a number of examples.

Instead of asking for the posterior quantile to have an interpretation as a confidence bound, we might work instead with the posterior distribution function directly. Consider

$$\Pi(w \mid Y) = \mathrm{P}_{\theta|Y}\{\sqrt{n}(\theta_1 - \hat{\theta}_1)/\hat{\sigma}_{11} \leq w \mid Y\}.$$

The frequentist counterpart of this distribution function is

$$F_{Y|\theta}(w) = \mathrm{P}_{Y|\theta}\{\sqrt{n}(\theta_1 - \hat{\theta}_1)/\hat{\sigma}_{11} \leq w\}.$$

The former quantity is random, but we might define matching by

$$E_{Y|\theta}\Pi(w \mid Y) = F_{Y|\theta}(w) + O(n^{-j})$$

for any fixed w. This approach was discussed in Mukerjee and Ghosh (1997), where they showed in particular that to first order ($j = 1$), this leads to the same condition as the Welch–Peers approach (2.8), but to the next order ($j = \frac{3}{2}$), this leads to a pair of equations $T_3(\pi, \theta) = 0$, $T_4(\pi, \theta) = 0$ as opposed to the single equation $T_2(\pi, \theta) = 0$ arising in the Welch-Peers approach (see Mukerjee and Ghosh, 1997, (2.13)). However, for the bivariate normal example discussed above, this leads to the same class of priors as quantile matching, even to second order.

Mukerjee and Reid (1999) considered matching under 'alternative' values, defined by requiring that

$$E_{Y|\theta}\,\mathrm{P}_{\theta|Y}\{\theta_1 + \delta(i^{11}/n)^{1/2} \leq \theta_1^{(1-\alpha)}(Y) \mid Y\}$$
$$= \mathrm{P}_{Y|\theta}\{\theta_1^{(1-\alpha)}(Y) \geq \theta_1 + \delta(i^{11}/n)^{1/2}\} + O(n^{-j})$$

and showed that to $O(n^{-1})$ this again requires $T_1 = 0$, as at (2.8), and to $O(n^{-3/2})$ requires a more stringent set of conditions.

In the bivariate normal example this leads to the second order matching priors

$$\pi(\theta) \propto g(\theta_4, \theta_5)\{(\theta_2\theta_3)\}^{-1}.$$

Mukerjee and Reid (2001) generalized the Welch-Peers results to quantiles of an arbitrary one-dimensional function $h(\theta)$, and found conditions under which the posterior quantile for $h(\theta)$ has the correct frequentist coverage to $O(n^{-1})$ and $O(n^{-3/2})$. By way of illustration, the generalization of (2.8) to this setting is (see also Datta and Ghosh (1995))

$$\frac{\partial}{\partial\theta_j}\left[\left\{i^{ab}(\theta)\frac{\partial h}{\partial\theta_a}\frac{\partial h}{\partial\theta_b}\right\}^{-1/2} i^{jc}(\theta)\frac{\partial h}{\partial\theta_c}\pi(\theta)\right] = 0,$$

where again there is an implied summation over a, b and c. In Mukerjee and Reid (2001) this was used to match Bayesian and frequentist tolerance limits, by choosing for $h(\theta)$ the $1-\beta$ quantile of the distribution function of Y_i. This matching criterion is more stringent than those derived above, and depends more strongly on the model.

The distribution-function matching approach described above used the standardized maximum likelihood estimate as the basis for the distribution, but it is also possible to consider matching the distribution function for the standardized score statistic or the generalized log-likelihood ratio statistic, or indeed other statistics that may be derived from quite different considerations. A survey of many of these matching criteria is given in Ghosh and Mukerjee (1998, Section 4).

Matching *prediction limits* leads to quite different expansions and criteria; details are provided in Datta et al. (2000). Sweeting (2001) discusses the role of matching priors for two-sided confidence intervals in scalar parameter models. For a comparison of Bayesian and frequentist approaches to prediction from another point of view, see Smith (1999).

4. Strong matching

The approach based on Edgeworth expansions does not generalize very well to problems with nuisance parameters, in the sense that there is no obvious criterion that works well in general models. Each has to be approached on a case by case basis. In retrospect this is perhaps not too surprising, since a generally acceptable approach would lead to frequentist inference for component parameters that was relatively straightforward, and to a Bayesian approach to prior selection in complex models that was also relatively straightforward, and both of these problems are the most difficult in their respective schools of inference.

Fraser and Reid (1996) developed a notion of *strong matching*, based on the approach to higher order asymptotic inference derived from the likelihood function developed by Barndorff-Nielsen, Fraser, and others, and most closely related to the saddlepoint or Laplace approximation rather than the Edgeworth approximation.

The basis of the frequentist approximation is the following approximation to the p-value for inference about θ_1:

$$\mathrm{P}_{Y|\theta}\{r(\theta_1, Y) \le r(\theta_1, y)\} \doteq \Phi(r) + \phi(r)\left(\frac{1}{r} - \frac{1}{q_{\mathrm{F}}}\right) \tag{4.1}$$

where

$$r = r(\theta_1, y) = \pm\sqrt{[2\{\ell_p(\hat{\theta}_1) - \ell_p(\theta_1)\}]}, \tag{4.2}$$

$\ell_p(\theta_1) = \ell\{\theta_1, \tilde{\theta}_{(2)}(\theta_1)\}$ is the profile log likelihood function, $\tilde{\theta}_{(2)} = \tilde{\theta}_{(2)}(\theta_1)$ is the restricted maximum likelihood estimate of $\theta_{(2)}$, and $q_{\mathrm{F}} = q_{\mathrm{F}}(\theta_1, y)$ is a maximum likelihood-type statistic with a relatively complicated definition which is derived in detail in Fraser, Reid and Wu (1999). There is a Bayesian version of approximation (4.1) as well:

$$\mathrm{P}_{\theta|Y}\{r(\Theta_1, y) \ge r(\theta_1, y) \mid y\} \doteq \Phi(r) + \phi(r)\left(\frac{1}{r} - \frac{1}{q_{\mathrm{B}}}\right) \tag{4.3}$$

where r is given by (4.2), $q_{\mathrm{B}} = q_{\mathrm{B}}(\theta_1, y)$ is a type of score statistic:

$$q_{\mathrm{B}} = \ell_1(\theta_1, \tilde{\theta}_{(2)})\left\{\frac{|\hat{j}_{\theta\theta}|}{|j_{22}(\theta_1, \tilde{\theta}_{(2)})|}\right\}^{-1/2} \frac{\pi(\hat{\theta})}{\pi(\theta_1, \tilde{\theta}_{(2)})} \tag{4.4}$$

where ℓ_1 is the score function for θ_1, j_{22} is the $\theta_{(2)}$ component of the observed information matrix.

For completeness we provide the expression for q_{F}, which relies on a reparametrization $\varphi = \varphi(\theta)$, discussed in Fraser, Reid and Wu (1999):

$$q_{\mathrm{F}} = \{\chi(\hat{\theta}) - \chi(\theta_1, \tilde{\theta}_{(2)})\}\left\{\frac{|\hat{j}_{\varphi\varphi}|}{|j_{(22)}(\theta_1, \tilde{\theta}_{(2)})|}\right\}^{1/2} \tag{4.5}$$

where the notation $j_{(22)}$ refers to the information submatrix corresponding to $\varphi_{(2)}$, and χ plays the role of the parameter of interest, defined by $\chi(\theta) = e_\psi^T \varphi(\theta)$, where $e_\psi = \psi_{\varphi'}(\hat{\theta}_\psi)/|\psi_{\varphi'}(\hat{\theta}_\psi)|$, and $\psi_{\varphi'}$ is the derivative of the first component of the inverse transformation $\theta(\varphi)$.

Setting $q_{\mathrm{B}} = q_{\mathrm{F}}$ ensures that to the order of approximations (4.1) and (4.3), Bayesian and frequentist inference limits agree. Since the Bayesian version depends on the prior, this determines in principle the form of the prior for the parameter of interest θ_1. In the bivariate normal example of

Section 2 this leads to a flat prior for θ_1. The strong matching prior for the full parameter is obtained by constructing an approximate location model, as described in Section 8 of Fraser and Reid (2001). Recent unpublished work of Fraser and Yi consider the construction of an approximate marginal likelihood for θ_1, thus avoiding the construction of any prior for the nuisance parameter.

Strong matching priors are dependent on the data, and in particular on the observed information, so do not lead to noninformative Bayesian inference in the conventional sense. They also have the disadvantage that the simplicity of Bayesian marginal inference is essentially lost in this approach. Approximations (4.1) and (4.3) have relative error $O(n^{-3/2})$ in $n^{-1/2}$ neighborhoods of the maximum likelihood estimate, so strong matching provides second order matching of confidence limits. Fraser and Reid (2001) discuss a simpler approach that provides matching to $O(n^{-1})$, and still uses a data dependent prior.

The emergence of a data dependent prior in attempts to match Bayesian and frequentist inference to second order seems to be inevitable; this was first discussed in the scalar parameter case in Pierce and Peters (1994), and is explored in more detail in Sweeting (2001). Sweeting (2001) also discusses, in the scalar parameter case, matching prediction limits, and matching two sided confidence limits. Both of these approaches lead to a different asymptotic result, for slightly different reasons. In considering two-sided confidence limits the forms of the expansions are such that the $O(1/\sqrt{n})$ terms in each tail cancel. The leading term in establishing even one-sided prediction limits is also $O(1/n)$.

In more complex models, data dependent priors have emerged in two quite different contexts. In Box and Cox (1964), Bayesian inference is considered for the model

$$y_i^{(\lambda)} = x_i'\beta + \sigma e_i \tag{4.6}$$

where the errors e_i are assumed to be independent standard normal random variables, and

$$y_i^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

Writing $\theta = (\beta, \sigma, \lambda)$, Box and Cox (1964) argue that the most natural noninformative prior is

$$\pi(\theta)\mathrm{d}\theta \propto \mathrm{d}\beta \frac{\mathrm{d}\sigma}{\sigma} \frac{\mathrm{d}\lambda}{(\dot{y}^{\lambda-1})^k}$$

where k is the dimension of β and $\dot{y}$ is the geometric mean of the observations. Essentially some information is needed on the scale of the observations, in order to sensibly assign a vague prior to the transformation parameter.

Wasserman (2000) considers matching priors in the mixture model

$$f(y;\theta) = \tfrac{1}{2}\phi(y) + \tfrac{1}{2}\phi(y-\theta)$$

where ϕ is the standard normal density. He shows that any fixed improper prior must lead to an improper posterior, and that no fixed prior can give matching of one-sided intervals to $O(n^{-1})$, but shows that the following data dependent prior solves both problems

$$\pi(\theta) \propto \{i(\theta)\}^{1/2} c(\theta; y)$$

where

$$c(\theta; y) = 1 - \prod \left\{ 1 + \frac{\phi(y_i - \theta)}{\phi(y_i)} \right\}$$

which in effect is a simple way to delete from the likelihood function the sample that comes entirely from the first part of the mixture distribution, and hence gives no information about θ, and use Jeffreys' prior for this pseudo-likelihood function. An analogous result is derived for a mixture of k normal distributions with differing means and variances. As discussed in Wasserman (2000), Diebolt and Robert (1994) used a similar idea with conjugate priors.

5. Conclusion

From several converging points of view, data dependent priors appear to be needed in a frequentist approach to Bayesian inference. They are possibly even needed in a Bayesian approach using default priors, as all default priors involve model averaging (through the calculation of expected information), and thus do not permit conditioning on any 'obvious' ancillaries.

There are other approaches to default or noninformative priors that do not involve matching frequentist limits or confidence bounds. A good review is given in Kass and Wasserman (1996), and there is also a default Bayes web page at `www.stat.missouri.edu/bayes/`.

The most prominent alternative default prior is Berger and Bernardo's reference prior, which is noninformative in the sense of maximizing the Kullback–Liebler distance between the prior and the posterior, or equivalently minimizing the Kullback–Liebler distance between the likelihood function and the marginal distribution of the data. An accessible introduction to reference priors is given in Kass and Wasserman (1996). Reference priors are often first order matching in problems with nuisance parameters, but there are now a number of examples described where the reference prior is not second order matching (Garvan and Ghosh, 1997; Ghosh and Kim, 2001; Yin and Ghosh, 2001). A data-dependent type of reference prior is developed in Clarke and Yuan (2001).

REFERENCES

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* 26, 211–252.

Clarke, B. and Yuan, A. (2001). Partial information reference priors. Preprint available at `http://hajek.stat.ubc.ca/~riffraff/publ.html`.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* 49, 1–39.

Datta, G.S. and Ghosh, J.K. (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika* 82, 37–45.

Datta, G.S., Mukerjee, R., Ghosh, M. and Sweeting, T.J. (2000). Bayesian prediction with approximate frequentist validity. *Ann. Statist.* 28, 1414–1426.

Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* 56, 363–375.

Fraser, D.A.S., and Reid, N. (1996). Bayes posteriors for scalar interest parameters. In *Bayesian Statistics* V (J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith, eds.), pp. 581–585. Oxford Univ. Press, Oxford.

Fraser, D.A.S. and Reid, N. (2001). Strong matching of frequentist and Bayesian parametric inference. *J. Statist. Plann. Inference.* To appear.

Fraser, D.A.S., Reid, N. and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* 86, 249–264.

Garvan, C.W. and Ghosh, M. (1997). Noninformative priors for dispersion models. *Biometrika* 84, 976–982.

Ghosh, M. and Mukerjee, R. (1998). Recent developments on probability matching priors. In *Applied Statistical Science. III* (S.E. Ahmed, M. Ahsanullah and B.K. Sinha, eds.), pp. 227–252. Nova Science, New York.

Ghosh, M. and Kim, Y.-H. (2001). The Behrens-Fisher problem revisited: a Bayes-frequentist synthesis. *Canad. J. Statist.* 29, 5–18.

Johnson, R.A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* 41, 851–864.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* 91, 1343–1370.

Lindley, D.V. (1958). Fiducial distributions and Bayes' theorem. *J. Roy. Statist. Soc. Ser. B* 20, 102–107.

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

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

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

Mukerjee, R. and Reid, N. (2001). Second-order probability matching priors for a parametric function with application to Bayesian tolerance limits. *Biometrika* 88, 587–592.

Nicolau, A. (1993). Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *J. Roy. Statist. Soc. Ser. B* 55, 377–390.

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

Pierce, D.A. and Peters, D. (1994). Higher-order asymptotics and the likelihood principle: one parameter models. *Biometrika* 81, 1–10.

Smith, R.L. (1999). Bayesian and frequentist approaches to parametric predictive inference (with discussion). In *Bayesian Statistics* VI (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.), pp. 589–612. Oxford Univ. Press, Oxford.

Sweeting, T.J. (2001). Coverage probability bias, objective Bayes and the likelihood principle. *Biometrika* 88, 657–676.

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

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *J. Roy. Statist. Soc. Ser. B* 62, 159–180.

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

Yin, M. and Ghosh, M. (2001). Bayesian and likelihood inference for the generalized Fieller-Creasy problem. In *Empirical Bayes and Likelihood Inference* (S.E. Ahmed and N. Reid, eds.), pp. 121–137. Lecture Notes in Statist., vol 148. Springer-Verlag, New York.

N. Reid
Department of Statistics
University of Toronto
100 St. George St.
Toronto, ON M5S 3G3
Canada
reid@utstat.toronto.edu

R. Mukerjee
Indian Institute of Management
Joka, Diamond Harbour Road
Post Box No. 16757
Alipore Post Office
Calcutta 700 027
India
rmuk1@hotmail.com

D.A.S. Fraser
Department of Statistics
University of Toronto
100 St. George St.
Toronto, ON M5S 3G3
Canada
dfraser@utstat.toronto.edu