Objective Priors: An Introduction for Frequentists[1]

Malay Ghosh

Abstract.   Bayesian methods are increasingly applied in these days in the theory and practice of statistics. Any Bayesian inference depends on a likelihood and a prior. Ideally one would like to elicit a prior from related sources of information or past data. However, in its absence, Bayesian methods need to rely on some "objective" or "default" priors, and the resulting posterior inference can still be quite valuable.

Not surprisingly, over the years, the catalog of objective priors also has become prohibitively large, and one has to set some specific criteria for the selection of such priors. Our aim is to review some of these criteria, compare their performance, and illustrate them with some simple examples. While for very large sample sizes, it does not possibly matter what objective prior one uses, the selection of such a prior does influence inference for small or moderate samples. For regular models where asymptotic normality holds, Jeffreys' general rule prior, the positive square root of the determinant of the Fisher information matrix, enjoys many optimality properties in the absence of nuisance parameters. In the presence of nuisance parameters, however, there are many other priors which emerge as optimal depending on the criterion selected. One new feature in this article is that a prior different from Jeffreys' is shown to be optimal under the chi-square divergence criterion even in the absence of nuisance parameters. The latter is also invariant under one-to-one reparameterization.

Key words and phrases: Asymptotic expansion, divergence criterion, first-order probability matching, Jeffreys' prior, left Haar priors, location family, location–scale family, multiparameter, orthogonality, reference priors, right Haar priors, scale family, second-order probability matching, shrinkage argument.

1. INTRODUCTION

Bayesian methods are increasingly used in recent years in the theory and practice of statistics. Their implementation requires specification of both a likelihood and a prior. With enough historical data, it is possible to elicit a prior distribution fairly accurately. However, even in its absence, Bayesian methods, if judiciously

Malay Ghosh is Distinguished Professor, University of Florida, 223 Griffin–Floyd Hall, Gainesville, Florida 32611-8545, USA (e-mail: ghoshm@stat.ufl.edu).

[1]Discussed in 10.1214/11-STS338A and 10.1214/11-STS338B; rejoinder at 10.1214/11-STS338REJ.

used, can produce meaningful inferences based on the so-called "objective" or "default" priors.

The main focus of this article is to introduce certain objective priors which could be potentially useful even for frequentist inference. One such example where frequentists are yet to reach a consensus about an "optimal" approach is the construction of confidence intervals for the ratio of two normal means, the celebrated Fieller–Creasy problem. It is shown in Section 4 of this paper how an "objective" prior produces a credible interval in this case which meets the target coverage probability of a frequentist confidence interval even for small or moderate sample sizes. Another situation, which has often become a real challenge for frequentists, is to find a suitable method for elimination

of nuisance parameters when the dimension of the parameter grows in direct proportion to the sample size. This is what is usually referred to as the Neyman–Scott phenomenon. We will illustrate in Section 3 with an example of how an objective prior can sometimes overcome this problem.

Before getting into the main theme of this paper, we recount briefly the early history of objective priors. One of the earliest uses is usually attributed to Bayes (1763) and Laplace (1812) who recommended using a uniform prior for the binomial proportion p in the absence of any other information. While intuitively quite appealing, this prior has often been criticized due to its lack of invariance under one-to-one reparameterization. For example, a uniform prior for p in the binomial case does not result in a uniform prior for p^2. A more compelling example is that a uniform prior for σ, the population standard deviation, does not result in a uniform prior for σ^2, and the converse is also true. In a situation like this, it is not at all clear whether there can be any preference to assign a uniform prior to either σ or σ^2.

In contrast, Jeffreys' (1961) general rule prior, namely, the positive square root of the determinant of the Fisher information matrix, is invariant under one-to-one reparameterization of parameters. We will motivate this prior from several asymptotic considerations. In particular, for regular models where asymptotic normality holds, Jeffreys' prior enjoys many optimality properties in the absence of nuisance parameters. In the presence of nuisance parameters, this prior suffers from many problems—marginalization paradox, the Neyman–Scott problem, just to name a few. Indeed, for the location–scale models, Jeffreys himself recommended alternate priors.

There are several criteria for the construction of objective priors. The present article primarily reviews two of these criteria in some detail, namely, "divergence priors" and "probability matching priors," and finds optimal priors under these criteria. The class of divergence priors includes "reference priors" introduced by Bernardo (1979). The "probablity matching priors" were introduced by Welch and Peers (1963). There are many generalizations of the same in the past two decades. The development of both these priors rely on asymptotic considerations. Somewhat more briefly, I have discussed also a few other priors including the "right" and "left" Haar priors.

The paper does not claim the extensive thorough and comprehensive review of Kass and Wasserman (1996), nor does it aspire to the somewhat narrowly focused,

but a very comprehensive review of probability matching priors as given in Ghosh and Mukerjee (1998), Datta and Mukerjee (2004) and Datta and Sweeting (2005). A very comprehensive review of reference priors is now available in Bernardo (2005), and a unified approach is given in the recent article of Berger, Bernardo and Sun (2009).

While primarily a review, the present article has been able to unify as well as generalize some of the previously considered criteria, for example, viewing the reference priors as members of a bigger class of divergence priors. Interestingly, with some of these criteria as presented here, it is possible to construct some alternatives to Jeffreys' prior even in the absence of nuisance parameters.

The outline of the remaining sections is as follows. In Section 2 we introduce two basic tools to be used repeatedly in the subsequent sections. One such tool involving asymptotic expansion of the posterior density is due to Johnson (1970), and Ghosh, Sinha and Joshi (1982), and is discussed quite extensively in Ghosh, Delampady and Samanta (2006) and Datta and Mukerjee (2004). The second tool involves a shrinkage argument suggested by Dawid and used extensively by J. K. Ghosh and his co-authors. It is shown in Section 3 that this shrinkage argument can also be used in deriving priors with the criterion of maximizing the distance between the prior and the posterior. The distance measure used includes, but is not limited to, the Kullback–Leibler (K–L) distance considered in Bernardo (1979) for constructing two-group "reference priors." Also, in this section we have considered a new prior different from Jeffreys even in the one-parameter case which is also invariant under one-to-one reparameterization. Section 4 addresses construction of priors under probability matching criteria. Certain other priors are introduced in Section 5, and it is pointed out that some of these priors can often provide exact and not just asymptotic matching. Some final remarks are made in Section 6.

Throughout this paper the results are presented more or less in a heuristic fashion, that is, without paying much attention to the regularity conditions needed to justify these results. More emphasis is placed on the application of these results in the construction of objective priors.

2. TWO BASIC TOOLS

An asymptotic expansion of the posterior density began with Johnson (1970), followed up later by Ghosh,

Sinha and Joshi (1982), and many others. The result goes beyond that of the theorem of Bernstein and Von Mises which provides asymptotic normality of the posterior density. Typically, such an expansion is centered around the MLE (and occasionally the posterior mode), and requires only derivatives of the log-likelihood with respect to the parameters, and evaluated at their MLE's. These expansions are available even for heavy-tailed densities such as Cauchy because finiteness of moments of the distribution is not needed. The result goes a long way in finding asymptotic expansion for the posterior moments of parameters of interest as well as in finding asymptotic posterior predictive distributions.

The asymptotic expansion of the posterior resembles that of an Edgeworth expansion, but, unlike the latter, this approach does not need use of cumulants of the distribution. Finding cumulants, though conceptually easy, can become quite formidable, especially in the presence of multiple parameters, demanding evaluation of mixed cumulants.

We have used this expansion as a first step in the derivation of objective priors under different criteria. Together with the shrinkage argument as mentioned earlier in the Introduction, and to be discussed later in this section, one can easily unify and extend many of the known results on prior selection. In particular, we will see later in this section how some of the reference priors of Bernardo (1979) can be found via application of these two tools. The approach also leads to a somewhat surprising result involving asymptotic expansion of the distribution function of the MLE in a fairly general setup, and is not restricted to any particular family of distributions, for example, the exponential family, or the location–scale family. A detailed exposition is available in Datta and Mukerjee (2004, pages 5–8).

For simplicity of exposition, we consider primarily the one-parameter case. Results needed for the multiparameter case will occasionally be mentioned, and, in most cases, these are straightforward, albeit often cumbersome, extensions of one-parameter results. Moreover, as stated in the Introduction, the results will be given without full rigor, that is, without any specific mention of the needed regularity conditions.

We begin with $X_1, \ldots, X_n | \theta$ i.i.d. with common p.d.f. $f(X|\theta)$. Let $\hat{\theta}_n$ denote the MLE of θ. The likelihood function is denoted by $L_n(\theta) = \prod_1^n f(X_i|\theta)$ and let $\ell_n(\theta) = \log L_n(\theta)$. Let $a_i = n^{-1}[d^i \ell_n(\theta)/d\theta^i]_{\theta=\hat{\theta}_n}$, $i = 1, 2, \ldots$, and let $\hat{I}_n = -a_2$, the observed per unit Fisher information number. Consider a twice differentiable prior π. Let $T_n = \sqrt{n}(\theta - \hat{\theta}_n)\hat{I}_n^{1/2}$, and let $\pi_n^*(t)$ denote the posterior p.d.f. of T_n given $X_1, \ldots, X_n$.

Then, under certain regularity conditions, we have the following result.

THEOREM 1. $\pi_n^*(t) = \phi(t)[1 + n^{-1/2}\gamma_1(t; X_1, \ldots, X_n) + n^{-1}\gamma_2(t; X_1, \ldots, X_n)] + O_p(n^{-3/2})$, where $\phi(t)$ is the standard normal p.d.f., $\gamma_1(t; X_1, \ldots, X_n) = a_3 t^3/(6\hat{I}_n^{3/2}) + (t/\hat{I}_n^{1/2})\pi'(\hat{\theta}_n)/\pi(\hat{\theta}_n)$ and

$$\gamma_2(t; X_1, \ldots, X_n)$$

$$= \frac{1}{24\hat{I}_n^2}a_4 t^4 + \frac{1}{72\hat{I}_n^3}a_3^2 t^6 + \frac{1}{2\hat{I}_n}t^2\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}$$

$$+ \frac{1}{6\hat{I}_n^2}a_3 t^4\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} - \frac{a_4}{8\hat{I}_n^2}$$

$$- \frac{15a_3^2}{72\hat{I}_n^3} - \frac{1}{2\hat{I}_n}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} - \frac{a_3}{2\hat{I}_n^2}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}.$$

The proof is given in Ghosh, Delampdy and Samanta (2006, pages 107–108). The statement involves a few minor typos which can be corrected easily. We outline here only a few key steps needed in the proof.

We begin with the posterior p.d.f.,

$$\pi(\theta|X_1, \ldots, X_n)$$

(2.1)
$$= \exp[\ell_n(\theta)]\pi(\theta)\Big/\int \exp[\ell_n(\theta)]\pi(\theta)\,d\theta.$$

Substituting $t = \sqrt{n}(\theta - \hat{\theta}_n)\hat{I}_n^{1/2}$, the posterior p.d.f. of T_n is given by

$$\pi_n^*(t) = C_n^{-1}\exp[\ell_n\{\hat{\theta}_n + t(n\hat{I}_n)^{-1/2}\} - \ell_n(\hat{\theta}_n)]$$

$$\cdot \pi\{\hat{\theta}_n + t(n\hat{I}_n)^{-1/2}\},$$

(2.2)
$$\text{where } C_n = \int \exp[\ell_n\{\hat{\theta}_n + t(n\hat{I}_n)^{-1/2}\}$$

$$- \ell_n(\hat{\theta}_n)]$$

$$\cdot \pi\{\hat{\theta}_n + t(n\hat{I}_n)^{-1/2}\}\,dt.$$

The rest of the proof involves a Taylor expansion of $\exp[\ell_n\{\hat{\theta}_n + t(n\hat{I}_n)^{-1/2}\}$ and $\pi\{\hat{\theta}_n + t(n\hat{I}_n)^{-1/2}\}$ around $\hat{\theta}_n$ up to a desired order, and collecting the coefficients of $n^{-1/2}$, n^{-1}, etc. The other component is evaluation of C_n via momets of the N(0, 1) distribution.

REMARK 1. The above result is useful in finding certain expansions for the posterior moments as well. In particular, noting $\theta = \hat{\theta}_n + (n\hat{I}_n)^{-1/2}t_n$, it follows that the asymptotic expansion of the posterior mean of θ is given by

$$E(\theta|X_1, \ldots, X_n)$$

(2.3)
$$= \hat{\theta}_n + n^{-1}\left\{\frac{a_3}{2\hat{I}_n^2} + t\frac{\pi'(\hat{\theta}_n)}{\hat{I}_n\pi(\hat{\theta}_n)}\right\} + O_p(n^{-3/2}).$$

Also, $V(\theta|X_1, \ldots, X_n) = (n\hat{I}_n)^{-1} + O_p(n^{-3/2})$.

A multiparameter extension of Theorem 1 is as follows. Suppose that $\theta = (\theta_1, \ldots, \theta_p)^T$ is the parameter vector and $\hat{\theta}_n$ is the MLE of θ. Let

$$a_{jr} = -\hat{I}_{njr} = n^{-1} \frac{\partial^2 \ell_n(\theta)}{\partial \theta_j \, \partial \theta_r}\bigg|_{\theta=\hat{\theta}_n},$$

$$a_{jrs} = n^{-1} \frac{\partial^3 \ell_n(\theta)}{\partial \theta_j \, \partial \theta_r \, \partial \theta_s}\bigg|_{\theta=\hat{\theta}_n}$$

and $\hat{I}_n = ((\hat{I}_{njr}))$. Then retaining only up to the $O(n^{-1/2})$ term, the posterior of $W_n = \sqrt{n}(\theta - \hat{\theta}_n)$ is given by

(2.4)
$$\begin{aligned}
\pi_n^*(w) &= (2\pi)^{-1/2} \exp[-(1/2)w^T \hat{I}_n w] \\
&\quad \cdot \bigg[1 + n^{-1/2}\bigg\{\sum_{j=1}^p w_j \bigg(\frac{\partial \log \pi}{\partial \theta_j}\bigg)\bigg|_{\theta=\hat{\theta}_n} \\
&\quad + \frac{1}{6}\sum_{j,r,s} w_j w_r w_s a_{jrs}\bigg\} \\
&\quad + O_p(n^{-1})\bigg].
\end{aligned}$$

Next we present the basic shrinkage argument of J. K. Ghosh discussed in detail in Datta and Mukherjee (2004). The prime objective here is evaluation of $E[q(X,\theta)|\theta] = \lambda(\theta)$, say, where X and θ can be real- or vector-valued. The idea is to find first $\int \lambda(\theta)\bar{\pi}_m(\theta)\,d\theta$ through a sequence of priors $\{\bar{\pi}_m(\theta)\}$ defined on a compact set, and then shrinking the prior to degeneracy at some interior point, say, θ of the compact set. The interesting point is that one never needs explicit specification of $\bar{\pi}_m(\theta)$ in carrying out this evaluation. We will see several illustrations of this in this article.

First, we present the shrinkage argument in a nutshell. Consider a proper prior $\bar{\pi}(\cdot)$ with a compact rectangle as its support in the parameter space, and $\bar{\pi}(\cdot)$ vanishes on the boundary of support, while remaining positive in the interior. The support of $\bar{\pi}(\cdot)$ is the closure of the set. Consider the posterior of θ under $\bar{\pi}(\cdot)$ and, hence, obtain $E^{\bar{\pi}}[q(X,\theta)|X]$. Then find $E[\{E^{\bar{\pi}}(q(X,\theta)|X)\}|\theta] = \lambda(\theta)$ for θ in the interior of the support of $\bar{\pi}(\cdot)$. Finally, integrate $\lambda(\cdot)$ with respect to $\bar{\pi}(\cdot)$, and then allow $\bar{\pi}(\cdot)$ to converge to the degenerate prior at the true value of θ at an interior point of the support of $\pi(\theta)$. This yields $E[q(X,\theta)|\theta]$. The calculation assumes integrability of $q(X,\theta)$ over the joint distribution of X and θ. Such integrability allows change in the order of integration.

When executed up to the desired order of approximation, under suitable assumptions, these steps can lead to significant reduction in the algebra underlying higher order frequentist asymptotics. The simplification arises from two counts. First, although the Bayesian approach to frequentist asymptotics requires Edgeworth type assumptions, it avoids an explicit Edgeworth expansion involving calculation of approximate cumulants. Second, as we will see, it helps establish the results in an easily interpretable compact form. The following two sections will demonstrate multiple usage of these two basic tools.

3. OBJECTIVE PRIORS VIA MAXIMIZATION OF THE DISTANCE BETWEEN THE PRIOR AND THE POSTERIOR

3.1 Reference Priors

We begin with an alternate derivation of the reference prior of Bernardo. Following Lindley (1956), Bernardo (1979) suggested a Kullback–Leibler (K–L) divergence between the prior and the posterior, namely, $E[\log \frac{\pi(\theta|X)}{\pi(\theta)}]$, where expectation is taken over the joint distribution of X and θ. The target is to find a prior π which maximizes the above distance. It is shown in Berger and Bernardo (1989) that if one does this maximization for a fixed n, this may lead to a discrete prior with finitely many jumps, a far cry from a diffuse prior. Hence, one needs an asymptotic maximization.

First write $E[\log \frac{\pi(\theta|X)}{\pi(\theta)}]$ as

(3.1)
$$\begin{aligned}
&E\bigg[\log \frac{\pi(\theta|X)}{\pi(\theta)}\bigg] \\
&= \iint \log \frac{\pi(\theta|X)}{\pi(\theta)} \pi(\theta|X) m^\pi(X)\,d\theta\,dX \\
&= \iint \log \frac{\pi(\theta|X)}{\pi(\theta)} L_n(\theta)\pi(\theta)\,dX\,d\theta \\
&= \int \pi(\theta) E\bigg[\log \frac{\pi(\theta|X)}{\pi(\theta)}\bigg|\theta\bigg]\,d\theta,
\end{aligned}$$

where $X = (X_1, \ldots, X_n)$, $L_n(\theta) = \prod_1^n f(X_i|\theta)$, the likelihood function, and $m^\pi(X)$ denotes the marginal of X after integrating out θ. The integrations are carried out with respect to a prior π having a compact support, and subsequently passing on to the limit as and when necessary.

Without any nuisance parameters, Bernardo (1979) showed somewhat heuristically that Jeffreys' prior

achieves the necessary maximization. A more rigorous proof was supplied later by Clarke and Barron (1990, 1994). We demonstrate heuristically how the shrinkage argument can also lead to the reference priors derived in Bernardo (1979). To this end, we first consider the one-parameter case for a regular family of distributions. We rewrite

$$
\begin{aligned}
E\left[\log \frac{\pi(\theta|X)}{\pi(\theta)}\right] = \int \pi(\theta) E[\log \pi(\theta|X)|\theta]\, d\theta \\
- \int \pi(\theta) \log \pi(\theta)\, d\theta.
\end{aligned}
$$
(3.2)

Next we write

$$
E^{\bar\pi}[\log \pi(\theta|X)|X] = \int \log \pi(\theta|X)\bar\pi(\theta|X)\, d\theta.
$$

From the asymptotic expansion of the posterior, one gets

$$
\begin{aligned}
\log \pi(\theta|X) = (1/2)\log(n) - (1/2)\log(2\pi) \\
- n\frac{(\theta - \hat\theta_n)^2}{2}\hat I_n + (1/2)\log(\hat I_n) \\
+ O_p(n^{-1/2}).
\end{aligned}
$$

Since $\frac{n(\theta-\hat\theta_n)^2}{2}\hat I_n$ converges a posteriori to a χ_1^2 distribution as $n \to \infty$, irrespective of a prior π, by the Bernstein–Von Mises and Slutsky's theorems, one gets

$$
\begin{aligned}
E^{\bar\pi}[\log \pi(\theta|X)] \\
(3.3) \quad = (1/2)\log(n) - (1/2)\log(2\pi e) \\
+ (1/2)\log(\hat I_n) + O_p(n^{-1/2}).
\end{aligned}
$$

Since the leading term in the right-hand side of (3.3) does not involve the prior $\bar\pi$, and $\hat I_n$ converges almost surely (P_θ) to $I(\theta)$, applying the shrinkage argument, one gets from (3.3)

$$
\begin{aligned}
E[\log \pi(\theta|X)|\theta] \\
(3.4) \quad = (1/2)\log(n) - (1/2)\log(2\pi e) \\
+ \log(I^{1/2}(\theta)) + O(n^{-1/2}).
\end{aligned}
$$

In view of (3.2), considering only the leading terms in (3.4), one needs to find a prior π which maximizes $\int \log\{\frac{I^{1/2}(\theta)}{\pi(\theta)}\}\pi(\theta)\, d\theta$. The integral being nonpositive due to the property of the Kullback–Leibler information number, its maximum value is zero, which is attained for $\pi(\theta) = I^{1/2}(\theta)$, leading once again to Jeffreys' prior.

The multiparameter generalization of the above result without any nuisance parameters is based on the asymptotic expansion

$$
\begin{aligned}
E[\log \pi(\theta|X)|\theta] \\
= (p/2)\log(n) - (p/2)\log(2\pi e) \\
+ \int \log\{|I(\theta)|^{1/2}/\pi(\theta)\}\pi(\theta)\, d\theta \\
+ O(n^{-1/2}),
\end{aligned}
$$

and maximization of the leading term yields once again Jeffreys' general rule prior $\pi(\theta) = |I(\theta)|^{1/2}$.

In the presence of nuisance parameters, however, Jeffreys' general rule prior is no longer the distance maximizer. We will demonstrate this in the case when the parameter vector is split into two groups, one group consisting of the parameters of interest, and the other involving the nuisance parameters. In particular, Bernardo's (1979) two-group reference prior will be included as a special case.

To this end, suppose $\theta = (\theta_1, \theta_2)$, where θ_1 ($p_1 \times 1$) is the parameter of interest and θ_2 ($p_2 \times 1$) is the nuisance parameter. We partition the Fisher information matrix $I(\theta)$ as

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

First begin with a general conditional prior $\pi(\theta_2|\theta_1) = \phi(\theta)$ (say). Bernardo (1979) considered $\phi(\theta) = |I_{22}(\theta)|^{1/2}$. The marginal prior $\pi(\theta_1)$ for θ_1 is then obtained by maximizing the distance $E[\log\frac{\pi(\theta_1|X)}{\pi(\theta_1)}]$. We begin by writing

$$
(3.5) \quad \log\frac{\pi(\theta_1|X)}{\pi(\theta_1)} = \log\frac{\pi(\theta|X)}{\pi(\theta)} - \log\frac{\pi(\theta_2|\theta_1,X)}{\pi(\theta_2|\theta_1)}.
$$

Writing $\pi(\theta) = \pi(\theta_1)\phi(\theta)$ and $|I(\theta)| = |I_{22}(\theta)| \cdot |I_{11.2}(\theta)|$, where $I_{11.2}(\theta) = I_{11}(\theta) - I_{12}(\theta)I_{22}^{-1}(\theta) \cdot I_{21}(\theta)$, the asymptotic expansion and the shrinkage argument together yield

$$
\begin{aligned}
E\left[\log\frac{\pi(\theta|X)}{\pi(\theta)}\right] \\
= (p/2)\log(n) - (p/2)\log(2\pi e) \\
+ \int \pi(\theta_1) \\
(3.6) \quad \cdot \left\{\int \phi(\theta) \right. \\
\left. \cdot \log\frac{|I_{22}(\theta)|^{1/2}|I_{11.2}(\theta)|^{1/2}}{\pi(\theta_1)\phi(\theta)}\, d\theta_2\right\}\, d\theta_1 \\
+ O(n^{-1/2})
\end{aligned}
$$

and

$$E\left[\log \frac{\pi(\theta_2|\theta_1, X)}{\pi(\theta_2|\theta_1)}\right]$$

$$(3.7) \quad = (p_2/2)\log(n) - (p_2/2)\log(2\pi e)$$

$$+ \int \pi(\theta_1)\left\{\int \phi(\theta) \log \frac{|I_{22}(\theta)|^{1/2}}{\phi(\theta)}\, d\theta_2\right\} d\theta_1$$

$$+ O(n^{-1/2}).$$

From (3.5)–(3.7), retaining only the leading term,

$$E\left[\log \frac{\pi(\theta_1|X)}{\pi(\theta_1)}\right]$$

$$(3.8) \quad \approx (p_1/2)\log(n) - (p_1/2)\log(2\pi e)$$

$$+ \int \pi(\theta_1)$$

$$\cdot \left\{\int \phi(\theta) \log \frac{|I_{11.2}(\theta)|^{1/2}}{\pi(\theta_1)}\, d\theta_2\right\} d\theta_1.$$

Writing $\log \psi(\theta_1) = \int \phi(\theta) \log |I_{11.2}(\theta)|^{1/2}\, d\theta_2$, once again by property of the Kullback–Leibler information number, it follows that the maximizing prior $\pi(\theta_1) = \psi(\theta_1)$.

We have purposely not set limits for these integrals. An important point to note [as pointed out in Berger and Bernardo (1989)] is that evaluation of all these integrals is carried out over an increasing sequence of compact sets K_i whose union is the entire parameter space. This is because most often we are working with improper priors, and direct evaluation of these integrals over the entire parameter space will simply give $+\infty$ which does not help finding any prior. As an illustration, if the parameter space is $\mathcal{R} \times \mathcal{R}^+$ as is typically the case for location–scale family of distributions, then one can take the increasing sequence of compact sets as $[-i, i] \times [i^{-1}, i]$, $i \geq 2$. All the proofs are usually carried out by taking a sequence of priors π_i with compact support K_i, and eventually making $i \to \infty$. This important point should be borne in mind in the actual derivation of reference priors. We will now illustrate this for the location–scale family of distributions when one of the two parameters is the parameter of interest, while the other one is the nuisance parameter.

EXAMPLE 1 (Location–scale models). Suppose $X_1, \ldots, X_n$ are i.i.d. with common p.d.f. $\sigma^{-1}f((x - \mu)/\sigma)$, where $\mu \in (-\infty, \infty)$ and $\sigma \in (0, \infty)$. Consider the sequence of priors π_i with support $[-i, i] \times [i^{-1}, i]$, $i = 2, 3, \ldots$. We may note that $I(\mu, \sigma) = \sigma^{-2}\begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix}$, where the constants c_1, c_2 and c_3 are functions of

f and do not involve either μ or σ. So, if μ is the parameter of interest, and σ is the nuisance parameter, following Bernardo's (1979) prescription, one begins with the sequence of priors $\pi_{i2}(\sigma|\mu) = k_{i2}\sigma^{-1}$ where, solving $1 = k_{i2}\int_{i-1}^{i}\sigma^{-1}\, d\sigma$, one gets $k_{i2} = (2\log i)^{-1}$. Next one finds the prior $\pi_{i1}(\mu) = k_{i1}\exp[\int_{-i}^{i} k_{i2}\sigma^{-1}\log(\sigma^{-1})\, d\sigma]$ which is a constant not depending on either μ or σ. Hence, the resulting joint prior $\pi_i(\mu, \sigma) = \pi_{i1}(\mu)\pi_{i2}(\sigma|\mu) \propto \sigma^{-1}$, which is the desired reference prior. Incidentally, this is Jeffreys' independence prior rather than Jeffreys' general rule prior, the latter being proportional to σ^{-2}. Conversely, when σ is the parameter of interest and μ is the nuisance parameter, one begins with $\pi_{i2}(\mu|\sigma) = (2i)^{-1}$ and then, following Bernardo (1979) again, one finds $\pi_{i1}(\sigma) = c_{i1}\exp[\int_{i-1}^{i}(2i)^{-1}\log(1/\sigma)]\, d\mu] \propto \sigma^{-1}$. Thus, once again one gets Jeffreys' independence prior. We will see in Section 5 that Jeffreys' independence prior is a right Haar prior, while Jeffreys' general rule prior is a left Haar prior for the location–scale family of distributions.

EXAMPLE 2 (Noncentrality parameter). Let $X_1, \ldots, X_n|\mu, \sigma$ be i.i.d. $N(\mu, \sigma^2)$, where μ real and $\sigma(> 0)$ are both unknown. Suppose the parameter of interest is $\theta = \mu/\sigma$, the noncentrality parameter. With the reparameterization (θ, σ) from (μ, σ), the likelihood is rewritten as $L(\theta, \sigma) \propto \sigma^{-n}\exp[-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n}(X_i - \theta\sigma)^2]$. Then the per observation Fisher information matrix is given by $I(\theta, \sigma) = \begin{pmatrix} 1 & \theta/\sigma \\ \theta/\sigma & (\theta^2+2)/\sigma^2 \end{pmatrix}$. Consider once again the sequence of priors π_i with support $[-i, i] \times [i^{-1}, i]$, $i = 2, 3, \ldots$. Again, following Bernardo, $\pi_{i2}(\sigma|\theta) = k_{i2}\sigma^{-1}$, where $k_{i2} = (2\log i)^{-1}$. Noting that $I_{11.2}(\theta, \sigma) = 1 - \theta^2/(\theta^2+2) = 2/(\theta^2 + 2)$, one gets $\pi_{i1}(\theta) = k_{i1}\exp[\int_{-i}^{i}\log(\sqrt{2}/(\theta^2 + 2)^{1/2})\, d\sigma] \propto (\theta^2 + 2)^{-1/2}$. Hence, the reference prior in this example is given by $\pi_R(\theta, \sigma) \propto (\theta^2+2)^{-1/2}\sigma^{-1}$. Due to its invariance property (Datta and Ghosh, 1996), in the original (μ, σ) parameterization, the two-group reference prior turns out to be $\pi_R(\mu, \sigma) \propto \sigma^{-1}(\mu^2 + 2\sigma^2)^{-1/2}$.

Things simplify considerably if θ_1 and θ_2 are orthogonal in the Fisherian sense, namely, $I_{12}(\theta) = 0$ (Huzurbazaar, 1950; Cox and Reid, 1987). Then if $I_{11}(\theta)$ and $I_{22}(\theta)$ factor respectively as $h_{11}(\theta_1)h_{12}(\theta_2)$ and $h_{21}(\theta_1)h_{22}(\theta_2)$, as a special case of a more general result of Datta and Ghosh (1995c), it follows that the two-group reference prior is given by $h_{11}^{1/2}(\theta_1)h_{22}^{1/2}(\theta_2)$.

EXAMPLE 3. As an illustration of the above, consider the celebrated Neyman–Scott problem (Berger

and Bernardo, 1992a, 1992b). Consider a fixed effects one-way balanced normal ANOVA model where the number of observations per cell is fixed, but the number of cells grows to infinity. In symbols, let $X_{i1}, \ldots, X_{ik}|\theta_i$ be mutually independent $N(\theta_i, \sigma^2)$, $k \geq 2$, $i = 1, \ldots, n$, all parameters being assumed unknown. Let $S = \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \bar{X}_i)^2/(n(k-1))$. Then the MLE of σ^2 is given by $(k-1)S/k$ which converges in probability [as $n \to \infty$ to $(k-1)\sigma^2/k$], and hence is inconsistent. Interestingly, Jeffreys' prior in this case also produces an inconsistent estimator of σ^2, but the Berger–Bernardo reference prior does not.

To see this, we begin with Fisher Information matrix $I(\theta_1, \ldots, \theta_n, \sigma) = k \operatorname{Diag}(\sigma^{-2}, \ldots, \sigma^{-2}, (1/2)n\sigma^{-4})$. Hence, Jeffreys' prior $\pi_J(\theta_1, \ldots, \theta_n, \sigma^2) \propto (\sigma^2)^{-n/2-1}$ which leads to the marginal posterior $\pi_J(\sigma^2|X) \propto (\sigma^2)^{-nk/2-1} \exp[-n(k-1)S/(2\sigma^2)]$ of σ^2, X denoting the entire data set. Then the posterior mean of σ^2 is given by $n(k-1)S/(nk-2)$, while the posterior mode is given by $n(k-1)S/(nk+2)$. Both are inconsistent estimators of σ^2, as these converge in probability to $(k-1)\sigma^2/k$ as $n \to \infty$.

In contrast, by the result of Datta and Ghosh (1995c), the two-group reference prior $\pi_R(\theta_1, \ldots, \theta_n, \sigma^2) \propto (\sigma^2)^{-1}$. This leads to the marginal posterior $\pi_R(\sigma^2| X) \propto (\sigma^2)^{-n(k-1)/2-1} \exp[-n(k-1)S/(2\sigma^2)]$ of σ^2. Now the posterior mean is given by $n(k-1)S/(n(k-1)-2)$, while the posterior mode is given by $n(k-1)S/(n(k-1)+2)$. Both are consistent estimators of σ^2.

EXAMPLE 4 (Ratio of normal means). Let X_1 and X_2 be two independent $N(\theta\mu, \mu)$ random variables, where the parameter of interest is θ. This is the celebrated Fieller–Creasy problem. The Fisher information matrix in this case is $I(\theta, \mu) = \begin{pmatrix} \mu^2 & \mu\theta \\ \mu\theta & 1+\theta^2 \end{pmatrix}$. With the transformation $\phi = \mu(1 + \theta^2)^{1/2}$, one obtains $I(\theta, \phi) = \operatorname{Diag}(\phi^2(1+\theta^2)^{-2}, 1)$. Again, by Datta and Ghosh (1995c), the two-group reference prior $\pi_R(\theta, \phi) \propto (1+\theta^2)^{-1}$.

EXAMPLE 5 (Random effects model). This example has been visited and revisited on several occasions. Berger and Bernardo (1992b) first found reference priors for variance components in this problem when the number of observations per cell is the same. Later, Ye (1994) and Datta and Ghosh (1995c, 1995d) also found reference priors for this problem. The case involving unequal number of observations per cell was considered by Chaloner (1987) and Datta, Ghosh and Kim (2002).

For simplicity, we consider here only the case with equal number of observations per cell. Let $Y_{ij} = m + \alpha_i + e_{ij}$, $j = 1, \ldots, n, i = 1, \ldots, k$. Here m is an unknown parameter, while α_i's and e_{ij} are mutually independent with α_i's i.i.d. $N(0, \sigma_\alpha^2)$ and e_{ij} i.i.d. $N(0, \sigma^2)$. The parameters m, σ_α^2 and σ^2 are all unknown. We write $\bar{Y}_i = \sum_{j=1}^{n} Y_{ij}/n$, $i = 1, \ldots, k$, and $\bar{Y} = \sum_{i=1}^{k} \bar{Y}_i/k$. The minimal sufficient statistic is $(\bar{Y}, T, S)$, where $T = n \sum_{i=1}^{k} (\bar{Y}_i - \bar{Y})^2$ and $S = \sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2$.

The different parameters of interest that we consider are m, σ_α^2/σ^2 and σ^2. The common mean m is of great relevance in meta analysis (cf. Morris and Normand, 1992). Ye (1994) pointed out that the variance ratio σ_α^2/σ^2 is of considerable interest in genetic studies. The parameter is also of importance to animal breeders, psychologists and others. Datta and Ghosh (1995d) have discussed the importance of σ^2, the error variance. In order to find reference priors for each one of these parameters, we first make the one-to-one transformation from $(m, \sigma_\alpha^2, \sigma^2)$ to (m, r, u), where $r = \sigma^{-2}$ and $u = \sigma^2/(n\sigma_\alpha^2 + \sigma^2)$. Thus, $\sigma_\alpha^2/\sigma^2 = (1-u)/(nu)$, and the likelihood $L(m, r, u)$ can be expressed as

$$L(m, r, u)$$
$$= r^{nk/2}u^{k/2} \exp[-(r/2)\{nku(\bar{Y} - m)^2 + uT + S\}].$$

Then the Fisher information matrix simplifies to $I(m, r, u) = k \operatorname{Diag}(nru, n/(2r^2), 1/(2u^2))$. From Theorem 1 of Datta and Ghosh (1995c), it follows now that when m, r and u are the respective parameters of interest, while the other two are nuisance parameters, the reference priors are given respectively by $\pi_{1R}(m, r, u) = 1$, $\pi_{2R}(m, r, u) = r^{-1}$ and $\pi_{3R}(m, r, u) = u^{-1}$.

3.2 General Divergence Priors

Next, back to the one-parameter case, we consider the more general distance (Amari, 1982; Cressie and Read, 1984)

$$(3.9) \qquad D^\pi = \left[1 - E\left\{\frac{\pi(\theta|X)}{\pi(\theta)}\right\}^{-\beta}\right] \Big/ \{\beta(1-\beta)\},$$
$$\beta < 1,$$

which is to be interpreted as its limit when $\beta \to 0$. This limit is the K–L distance as considered in Bernardo (1979). Also, $\beta = 1/2$ gives the Bhattacharyya–Hellinger (Bhattacharyya, 1943; Hellinger, 1909) distance, and $\beta = -1$ leads to the chi-square distance (Clarke

and Sun, 1997, 1999). In order to maximize D^π with respect to a prior π, one re-expresses (3.9) as

$$
\begin{aligned}
D^\pi &= \left[1 - \iint \pi^{\beta+1}(\theta)\pi^{-\beta}(\theta|X)L_n(\theta)\,dX\,d\theta\right] \\
&\quad /\{\beta(1-\beta)\} \\
&= \left[1 - \int \pi^{\beta+1}(\theta)E[\{\pi^{-\beta}(\theta|X)\}|\theta]\,d\theta\right] \\
&\quad /\{\beta(1-\beta)\}.
\end{aligned}
\tag{3.10}
$$

Hence, from (3.10), maximization of D^π amounts to minimization (maximization) of

$$
\int \pi^{\beta+1}(\theta)E[\{\pi^{-\beta}(\theta|X)\}|\theta]\,d\theta
\tag{3.11}
$$

for $0 < \beta < 1$ ($\beta < 0$). First consider the case $0 < |\beta| < 1$. From Theorem 1, the posterior of θ is

$$
\begin{aligned}
\pi(\theta|X) &= \frac{\sqrt{n}\hat{I}_n^{1/2}}{(2\pi)^{1/2}}\exp\left[-\frac{n}{2}(\theta-\hat{\theta}_n)^2\hat{I}_n\right] \\
&\quad \cdot [1 + O_p(n^{-1/2})].
\end{aligned}
\tag{3.12}
$$

Thus,

$$
\begin{aligned}
\pi^{-\beta}&(\theta|X) \\
&= n^{-\beta/2}(2\pi)^{\beta/2}\hat{I}_n^{-\beta/2} \\
&\quad \cdot \exp\left[\frac{n\beta}{2}(\theta-\hat{\theta}_n)^2\hat{I}_n\right][1 + O_p(n^{-1/2})].
\end{aligned}
\tag{3.13}
$$

Following the shrinkage argument, and noting that conditional on θ, $\hat{I}_n \overset{p}{\to} I(\theta)$, while $n(\theta-\hat{\theta}_n)^2\hat{I}_n \overset{d}{\to} \chi_1^2$, it follows heuristically from (3.13)

$$
\begin{aligned}
E[\pi^{-\beta}&(\theta|X)] \\
&= n^{-\beta/2}(2\pi)^{\beta/2}[I(\theta)]^{-\beta/2}(1-\beta)^{-1/2} \\
&\quad \cdot [1 + O_p(n^{-1/2})].
\end{aligned}
\tag{3.14}
$$

Hence, from (3.14), considering only the leading term, for $0 < \beta < 1$, minimization of (3.11) with respect to π amounts to minimization of $\int[\pi(\theta)/I^{1/2}(\theta)]^\beta\pi(\theta)\,d\theta$ with respect to π subject to $\int \pi(\theta)\,d\theta = 1$. A simple application of Holder's inequality shows that this minimization takes place when $\pi(\theta) \propto I^{1/2}(\theta)$. Similarly, for $-1 < \beta < 0$, $\pi(\theta) \propto I^{1/2}(\theta)$ provides the desired maximization of the expected distance between the prior and the posterior. The K–L distance, that is, when $\beta \to 0$, has already been considered earlier.

REMARK 2. Equation (3.14) also holds for $\beta < -1$. However, in this case, it is shown in Ghosh, Mergel and Liu (2011) that the integral $\int\{\pi(\theta)/I^{1/2}(\theta)\}^{-\beta} \cdot \pi(\theta)\,d\theta$ is uniquely minimized with respect to $\pi(\theta) \propto I^{1/2}(\theta)$, and there exists no maximizer of this integral when $\int \pi(\theta)\,d\theta = 1$. Thus, in this case, there does not exist any prior which maximizes the posterior-prior distance.

REMARK 3. Surprisingly, Jeffreys' prior is not necessarily the solution when $\beta = -1$ (the chi-square divergence). In this case, the first-order asymptotics does not work since $\pi^{\beta+1}(\theta) = 1$ for all θ. However, retaining also the $O_p(n^{-1})$ term as given in Theorem 1, Ghosh, Mergel and Liu (2011) have found in this case the solution $\pi(\theta) \propto \exp[\int^\theta \frac{2g_3(t)-I'(t)}{4I(t)}\,dt]$, where $g_3(t) = E[-\frac{d^3\log p(X_1|t)}{dt^3}|t]$. We shall refer to this prior as $\pi_{\mathrm{GML}}(\theta)$. We will show by examples that this prior may differ from Jeffreys' prior. But first we will establish a hitherto unknown invariance property of this prior under one-to-one reparameterization.

THEOREM 2. *Suppose that ϕ is a one-to-one twice differentiable function of θ. Then $\pi_{\mathrm{GML}}(\phi) = C\pi_{\mathrm{GML}}(\theta)|\frac{d\theta}{d\phi}|$, where $C(>0)$, the constant of proportionality, does not involve any parameters.*

PROOF. Without loss of generality, assume that ϕ is a nondecreasing function of θ. By the identity

$$
g_3(\phi) = I'(\phi) + E\left[\left(\frac{d^2\log f}{d\phi^2}\right)\left(\frac{d\log f}{d\phi}\right)\right],
$$

$\pi'_{\mathrm{GML}}(\phi)/\pi_{\mathrm{GML}}(\phi)$ reduces to

$$
\begin{aligned}
&\pi'_{\mathrm{GML}}(\phi)/\pi_{\mathrm{GML}}(\phi) \\
&= \frac{I'(\phi) + 2E[(d^2\log f/d\phi^2)(d\log f/d\phi)]}{4I(\phi)}.
\end{aligned}
\tag{3.15}
$$

Next, from the relation $I(\phi) = I(\theta)(d\theta/d\phi)^2$, one gets the identities

$$
\begin{aligned}
I'(\phi) &= I'(\theta)\left(\frac{d\theta}{d\phi}\right)^3 \\
&\quad + 2I(\theta)(d\theta/d\phi)(d^2\theta/d\phi^2);
\end{aligned}
\tag{3.16}
$$

$$
\begin{aligned}
&\left(\frac{d^2\log f}{d\phi^2}\right)\left(\frac{d\log f}{d\phi}\right) \\
&= \left\{\frac{d^2\log f}{d\theta^2}\left(\frac{d\theta}{d\phi}\right)^2 + \frac{d\log f}{d\theta}\cdot\frac{d^2\theta}{d\phi^2}\right\} \\
&\quad \cdot \left(\frac{d\log f}{d\theta}\cdot\frac{d\theta}{d\phi}\right).
\end{aligned}
\tag{3.17}
$$

From (3.15)–(3.17), one gets, after simplification,

$$\pi'_{\mathrm{GML}}(\phi)/\pi_{\mathrm{GML}}(\phi)$$

$$(3.18) \qquad = \frac{\pi'_{\mathrm{GML}}(\theta)}{\pi_{\mathrm{GML}}(\theta)}\frac{d\theta}{d\phi} + \frac{d^2\theta/d\phi^2}{d\theta/d\phi}.$$

Now, on integration, it follows from (3.18) $\pi_{\mathrm{GML}}(\phi) = C\pi_{\mathrm{GML}}(\phi)(d\theta/d\phi)$, which proves the theorem. □

EXAMPLE 6. Consider the one-parameter exponential family of distributions with $p(X|\theta) = \exp[\theta X - \psi(\theta) + h(X)]$. Then $g_3(\theta) = I'(\theta)$ so that $\pi(\theta) \propto \exp[\frac{1}{4} \int \frac{I'(\theta)}{I(\theta)} d\theta] = I^{1/4}(\theta)$, which is different from Jeffreys' $I^{1/2}(\theta)$ prior. Because of the invariance result proved in Theorem 2, in particular, for the Binomial(n, p) problem, noting that $p = \exp(\theta)/[1 + \exp(\theta)]$, one gets $\pi_{\mathrm{GML}}(p) \propto p^{-3/4}(1 - p)^{-3/4}$, which is a Beta$(\frac{1}{4}, \frac{1}{4})$ prior, different from Jeffreys' Beta$(\frac{1}{2}, \frac{1}{2})$ prior, Laplace's Beta$(1, 1)$ prior or Haldane's improper Beta$(0, 0)$ prior. Similarly, for the Poisson (λ) case, one gets $\pi_{\mathrm{GML}}(\lambda) \propto \lambda^{-1/4}$, again different from Jeffreys' $\pi_J(\lambda) \propto \lambda^{-1/2}$ prior. However, for the N$(\theta, 1)$ distribution, since $I(\theta) = 1$ and $g_3(\theta) = I'(\theta) = 0, \pi_{\mathrm{GML}}(\theta) = c(> 0)$, a constant, which is the same as Jeffreys' prior. It may be pointed out also that for the one-parameter exponential family, for the chi-square divergence, π_{GML} differs from Hartigan's (1998) maximum likelihood prior $\pi_H(\theta) = I(\theta)$.

EXAMPLE 7. For the one-parameter location family of distributions with $p(X|\theta) = f(X - \theta)$, where f is a p.d.f., both $g_3(\theta)$ and $I(\theta)$ are constants implying $I'(\theta) = 0$. Hence, $\pi_{\mathrm{GML}}(\theta)$ is of the form $\pi_{\mathrm{GML}}(\theta) = \exp(k\theta)$ for some constant k. However, for the special case of a symmetric f, that is, $f(X) = f(-X)$ for all X, $g_3(\theta) = 0$, and then $\pi_{\mathrm{GML}}(\theta)$ reduces once again to $\pi(\theta) = c$, which is the same as Jeffreys' prior.

EXAMPLE 8. For the general scale family of distributions with $p(X|\theta) = \theta^{-1} f(\frac{X}{\theta}), \theta > 0$, where f is a p.d.f., $I(\theta) = \frac{c_1}{\theta^2}$ for some constant $c_1(> 0)$, where $g_3(\theta) = \frac{c_2}{\theta^3}$ for some constant c_2. Then $\pi_{\mathrm{GML}}(\theta) \propto \exp(c \log \theta) = \theta^c$ for some constant c. In particular, when $p(X|\theta) = \theta^{-1} \exp(-\frac{X}{\theta}), \pi_{\mathrm{GML}}(\theta) \propto \theta^{-3/2}$, different from Jeffreys' $\pi_J(\theta) \propto \theta^{-1}$ for the general scale family of distributions.

The multiparameter extension of the general divergence prior has been explored in the Ph.D. dissertation of Liu (2009). Among other things, he has shown that in the absence of any nuisance parameters, for $|\beta| < 1$, the divergence prior is Jeffreys' prior. However, on the boundary, namely, $\beta = -1$, priors other than Jeffreys' prior emerge.

4. PROBABILITY MATCHING PRIORS

4.1 Motivation and First-Order Matching

As mentioned in the Introduction, probability matching priors are intended to achieve Bayes-frequentist synthesis. Specifically, these priors are required to provide asymptotically the same coverage probability of the Bayesian credible intervals with the corresponding frequentist counterparts. Over the years, there have been several versions of such priors-quantile matching priors, matching priors for distribution functions, HPD matching priors and matching priors associated with likelihood ratio statistics. Datta and Mukerjee provided a detailed account of all these priors. In this article I will be concerned only with quantile matching priors.

A general definition of quantile matching priors is as follows: Suppose $X_1, \ldots, X_n|\theta$ i.i.d. with common p.d.f. $f(X|\theta)$, where θ is a real-valued parameter. Assume all the needed regularity conditions for the asymptotic expansion of the posterior around $\hat{\theta}_n$, the MLE of θ. We continue with the notation of the previous section. For $0 < \alpha < 1$, let $\theta^{\pi}_{1-\alpha}(X_1, \ldots, X_n) \equiv \theta^{\pi}_{1-\alpha}$ denote the $(1 - \alpha)$th asymptotic posterior quantile of θ based on the prior π, that is,

$$(4.1) \qquad \begin{aligned} &P^{\pi}[\theta \le \theta^{\pi}_{1-\alpha}|X_1, \ldots, X_n] \\ &= 1 - \alpha + O_p(n^{-r}) \end{aligned}$$

for some $r > 0$. If now $P[\theta \le \theta^{\pi}_{1-\alpha}|\theta] = 1 - \alpha + O_p(n^{-r})$, then some order of probability matching is achieved. If $r = 1$, we call π a first-order probability matching prior. If $r = 3/2$, we call π a second-order probability matching prior.

We first provide an intuitive argument for why Jeffreys' prior is a first-order probability matching prior in the absence of nuisance parameters. If $X_1, \ldots, X_n|\theta$ i.i.d. N$(\theta, 1)$ and $\pi(\theta) = 1, -\infty < \theta < \infty$, then the posterior $\pi(\theta|X_1, \ldots, X_n)$ is N$(\bar{X}_n, n^{-1})$. Now writing $z_{1-\alpha}$ as the $100(1 - \alpha)\%$ quantile of the N$(0, 1)$ distribution, one gets

$$(4.2) \qquad \begin{aligned} &P[\sqrt{n}(\theta - \bar{X}_n) \le z_{1-\alpha}|X_1, \ldots, X_n) \\ &= 1 - \alpha = P[\sqrt{n}(\bar{X}_n - \theta) \ge -z_{1-\alpha}|\theta], \end{aligned}$$

so that the one-sided credible interval $\bar{X}_n + z_{1-\alpha}/\sqrt{n}$ for θ has exact frequentist coverage probability $1 - \alpha$.

The above exact matching does not always hold. However, if $X_1, \ldots, X_n|\theta$ are i.i.d., then $\hat{\theta}_n|\theta$ is asymptotically N$(\theta, (nI(\theta))^{-1})$. Then, by the delta method, $g(\hat{\theta}_n)|\theta \sim \mathrm{N}[g(\theta), (g'(\theta))^2(nI(\theta))^{-1}]$. So if $g'(\theta) = I^{1/2}(\theta)$ so that $g(\theta) = \int^{\theta} I^{1/2}(t) \, dt, \sqrt{n}[g(\hat{\theta}_n) -$

$g(\theta)]|\theta$ is asymptotically N(0, 1). Hence, from (4.2), with the uniform prior $\pi(\phi) = 1$ for $\phi = g(\theta)$, coverage matching is asymptotically achieved for ϕ. This leads to the prior $\pi(\theta) = \frac{d\phi}{d\theta} = g'(\theta) = I^{1/2}(\theta)$ for θ.

Datta and Mukerjee (2004, pages 14–21) proved the result in a formal manner. They used the two basic tools of Section 3. In the absence of nuisance parameters, they showed that a first-order matching prior for θ is a solution of the differential equation

$$(4.3) \qquad \frac{d}{d\theta}(\pi(\theta)I^{-1/2}(\theta)) = 0,$$

so that Jeffreys' prior is the unique first-order matching prior. However, it does not always satisfy the second-order matching property.

4.2 Second-Order Matching

In order that the matching is accomplished up to $O(n^{-3/2})$ (second-order matching), one needs an asymptotic expansion of the posterior distribution function up to the $O(n^{-1})$ term, and to set up a second differential equation in addition to (4.3). This equation is given by (cf. Mukerjee and Dey, 1993; Mukerjee and Ghosh, 1997)

$$(4.4) \quad \begin{aligned} &\frac{1}{3}\frac{d}{d\theta}[\pi(\theta)I^{-2}(\theta)g_3(\theta)] + \frac{d^2}{d\theta^2}[\pi(\theta)I^{-1}(\theta)] \\ &\quad = 0, \end{aligned}$$

where, as before, $g_3(\theta) = -E[\frac{d^3 \log f(X|\theta)}{d\theta^3}|\theta]$. If Jeffrey's prior satisfies (4.4), then it is the unique second-order matching prior. While for the location and scale family of distributions, this is indeed the case, this is not true in general. Of course, in such an instance, there does not exist any second-order matching prior.

To see this, for $\pi_J(\theta) = I^{1/2}(\theta)$, (4.4) reduces to

$$\frac{1}{3}\frac{d}{d\theta}[I^{-3/2}(\theta)g_3(\theta)] + \frac{d^2}{d\theta^2}[I^{-1/2}(\theta)] = 0,$$

which requires $\frac{1}{3}I^{-3/2}(\theta)g_3(\theta) + \frac{d}{d\theta}(I^{-1/2}(\theta))$ to be a constant free from θ. After some algebra, the above expression simplifies to $(1/6)E[(\frac{d \log f}{d\theta})^3|\theta]/I^{3/2}(\theta)$. It is easy to check now that for the one-parameter location and scale family of distributions, the above expression does not depend on θ. However, for the one-parameter exponential family of distributions with canonical parameter θ, the same holds if and only if $I'(\theta)/I^{3/2}(\theta)$ does not depend on θ, or, in other words, $I(\theta) = \exp(c\theta)$ for some constant c. Another interesting example is given below.

EXAMPLE 9. $(X_1, X_2)^T \sim N_2[\binom{0}{0}, \binom{1 \ \rho}{\rho \ 1}]$. One can verify that $I(\rho) = (1 + \rho^2)/(1 - \rho^2)^2$ and $L_{1,1,1} = -\frac{2\rho(3+\rho^2)}{(1-\rho^2)^3}$ so that $L_{1,1,1}/I^{3/2}(\rho)$ is not a constant. Hence, π_J is not a second-order matching prior, and there does not exist any second-order matching prior in this example.

4.3 First-Order Quantile Matching Priors in the Presence of Nuisance Parameters

The parameter of interest is still real-valued, but there may be one or more nuisance parameters. To fix ideas, suppose $\theta = (\theta_1, \ldots, \theta_p)$, where θ_1 is the parameter of interest, while $\theta_2, \ldots, \theta_p$ are the nuisance parameters. As shown by Welch and Peers (1963) and later more rigorously by Datta and Ghosh (1995a) and Datta (1996), writing $I^{-1} = ((I^{jk}))$, the probability matching equation is given by

$$(4.5) \qquad \sum_{j=1}^{p} \frac{\partial}{\partial \theta_j}\{\pi(\theta)I^{j1}(I^{11})^{-1/2}\} = 0.$$

EXAMPLE 1 (Continued). First consider μ as the parameter of interest, and σ the nuisance parameter. Since each element of the inverse of the Fisher information matrix is a constant multiple of σ^2, any prior $\pi(\mu, \sigma) \propto g(\sigma)$, g arbitrary, satisfies (4.5). Conversely, when σ is the parameter of interest, and μ is the nuisance parameter, any prior $\pi(\mu, \sigma) \propto \sigma^{-1}g(\mu)$ satisfies (4.5).

A special case considered in Tibshirani (1989) is of interest. Here θ_1 is orthogonal to $(\theta_2, \ldots, \theta_p)$ in the Fisherian sense, that is, $I^{j1} = 0$ for $j = 2, 3, \ldots, p$.

With orthogonality, (4.5) simplifies to

$$\frac{\partial}{\partial \theta_1}\{\pi(\theta)I_{11}^{-1/2}\} = 0$$

(since $I^{11} = I_{11}^{-1}$). This leads to $\pi(\theta) = I_{11}^{1/2}h(\theta_2, \ldots, \theta_p)$, where h is arbitrary. Often a second-order matching prior removes the arbitrariness of h. We will see an example later in this section. However, this need not always be the case, and, indeed, as seen earlier in the one parameter case, second-order matching priors may not always exist. We will address this issue later in this section.

A special choice is $h \equiv 1$. The resultant prior $\pi(\theta) = I_{11}^{1/2}$ bears some intuitive appeal. Since under orthogonality, $\sqrt{n}(\hat{\theta}_{1n} - \theta_1)|\theta \sim N(0, I_{11}^{-1}(\theta))$, one may expect $I_{11}^{1/2}(\theta)$ to be a first-order probability matching prior. This prior is only a member within the class of

priors $\pi(\theta) = I_{11}^{1/2} h(\theta_2, \ldots, \theta_p)$, as found by Tibshirani (1989), and admittedly need not be second-order matching even when the latter exists. A recent article by Staicu and Reid (2008) has proved some interesting properties of the prior $\pi(\theta) = I_{11}^{1/2}(\theta)$. This prior is also considered in Ghosh and Mukerjee (1992).

For a symmetric location–scale family of distributions, that is, when $f(X) = f(-X)$, $c_2 = 0$, that is, μ and σ are orthogonal. Now, when μ is the parameter of interest and σ is the nuisance parameter, the class of first-order matching priors $\pi_1(\mu, \sigma)$ is characterized by $h_1(\sigma)$, where h_1 is arbitrary. Similarly, when σ is the parameter of interest and μ is the nuisance parameter, the class of first-order matching priors is characterized by $\pi_2(\mu, \sigma) = \sigma^{-1} h_2(\mu)$, where g_2 is arbitrary. The intersection of the two classes leads again to the unique prior $\pi(\mu, \sigma) = \sigma^{-1}$.

EXAMPLE 2 (Continued). Let $X_1, \ldots, X_n | \mu, \sigma$ be i.i.d. N(μ, σ^2), and $\theta = \mu/\sigma$ is again the parameter of interest. In order to find a parameter ϕ which is orthogonal to θ, we rewrite the p.d.f. in the form

$$(4.6) \quad \begin{aligned} & f(X|\theta, \sigma) \\ & = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(X - \theta\sigma)^2\right]. \end{aligned}$$

Then the Fisher information matrix

$$I(\theta, \sigma) = \left[\begin{pmatrix} 1 & \theta/\sigma \\ \theta/\sigma & \sigma^{-2}(\theta^2 + 2) \end{pmatrix}\right].$$

It turns out now if we reparameterize from (θ, σ) to (θ, ϕ), where $\phi = \sigma(\theta^2 + 2)^{1/2}$, then θ and ϕ are orthogonal with the corresponding Fisher information matrix given by $I(\theta, \phi) = \text{Diag}[2(\theta^2 + 2)^{-1}, \phi^{-2}(\theta^2 + 2)]$. Hence, the class of first-order matching priors when θ is the parameter of interest is given by $\pi(\theta, \phi) = (\theta^2 + 2)^{-1/2} h(\phi)$, where h is arbitrary.

4.4 Second-Order Quantile Matching Priors in the Presence of Nuisance Parameters

When θ_1 is the parameter of interest, and $(\theta_2, \ldots, \theta_p)$ is the vector of nuisance parameters, the general class of second-order quantile matching priors is characterized in (2.4.11) and (2.4.12) of Datta and Mukerjee (2004, page 12). For simplicity, we consider only the case when θ_1 is orthogonal to $(\theta_2, \ldots, \theta_p)$. In this case a first-order quantile matching prior $\pi(\theta_1, \theta_2, \ldots, \theta_p) \propto I_{11}^{1/2}(\theta) h(\theta_2, \ldots, \theta_p)$ is also second-order matching if and only if h satisfies (cf. Datta and Mukerjee, 2004,

page 27) the differential equation

$$\begin{aligned} (4.7) \quad & \sum_{s=2}^{p} \sum_{u=2}^{p} \frac{\partial}{\partial \theta_u} \left\{ I_{11}^{-1/2} I^{su} E\left(\frac{\partial^3 \log f}{\partial \theta_1^2 \, \partial \theta_s}\right) h \Big| \theta \right\} \\ & + (h/6) \frac{\partial}{\partial \theta_1} \left\{ I_{11}^{-3/2} E\left(\left(\frac{\partial \log f}{\partial \theta_1}\right)^3 \Big| \theta\right)\right\} \\ & = 0. \end{aligned}$$

We revisit Examples 1–5 and provide complete, or at least partial, characterization of second-order quantile matching priors.

EXAMPLE 1 (Continued). Let f be symmetric so that μ and σ are orthogonal. First let μ be the parameter of interest and σ the nuisance parameter. Then since both the terms in (4.7) are zeroes, every first-order quantile matching prior of the form $\sigma^{-1} h(\sigma) = q(\sigma)$, say, is also second-order matching. This means that an arbitrary prior of the form $\pi(\mu, \sigma)$ is second-order matching as long as it is only a function of σ. On the other hand, if σ is the parameter of interest and μ is the nuisance parameter, since the second term in (4.7) is zero, a first-order quantile matching prior of the form $\sigma^{-1} h(\mu)$ is also second-order matching if and only if $h(\mu)$ is a constant. Thus, the unique second-order quantile matching prior in this case is proportional to σ^{-1}, which is Jeffreys' independence prior.

EXAMPLE 2 (Continued). Recall that in this case writing $\theta = \mu/\sigma$, and $\phi = \sigma(\theta^2 + 2)^{1/2}$, the Fisher information matrix $I(\theta, \phi) = \text{Diag}[2(\theta^2 + 2)^{-1}, \phi^{-2}(\theta^2 + 2)]$. Also, $E[(\frac{\partial \log f}{\partial \theta})^3 | \theta, \phi] = -\frac{\partial \theta(\theta^2 + 3)}{(\theta^2 + 2)^3}$ and $E(\frac{\partial^3 \log f}{\partial \theta^2 \partial \phi} | \theta, \phi) = (4/\phi)(\theta^2 + 2)^{-2}$. Hence, (4.7) holds if and only if $h(\phi) = \phi^{-1}$. This leads to the unique second-order quantile matching prior $\pi(\theta, \phi) \propto (\theta^2 + 2)^{-1/2}$. Back to the original (μ, σ) parameterization, this leads to the prior $\pi(\mu, \sigma) \propto \sigma^{-1}$, Jeffreys' independence prior.

EXAMPLE 3 (Continued). Consider once again the Neyman–Scott example. Since the Fisher information matrix $I(\theta_1, \ldots, \theta_n, \sigma^2) = k \, \text{Diag}(\sigma^{-2}, \ldots, \sigma^{-2}, n\sigma^{-2})$, σ^2 is orthogonal to $(\theta_1, \ldots, \theta_n)$. Now, the class of second-order matching priors is given by $\sigma^{-2} h(\mu_1, \ldots, \mu_n)$, where h is arbitrary. Simple algebra shows that in this case both the first and second terms in (4.7) are zeroes so that every first-order quantile matching prior is also second-order matching.

EXAMPLE 4 (Continued). From Tibshirani (1989), it follows that the class of first-order quantile matching

priors for θ is of the form $(1 + \theta^2)^{-1} h(\phi)$, where h is arbitrary. Once again, since both the first and second terms in (4.7) are zeroes, every first-order quantile matching prior is also second-order matching.

EXAMPLE 5 (Continued). Again from Tibshirani (1989), the class of second-order matching priors when m, r and u are the parameters of interest are given respectively by $h_1(r, u)$, $r^{-1} h_2(m, u)$ and $u^{-1} h_3(m, r)$, where h_1, h_2 and h_3 are arbitrary nonnegative functions. Also, the prior $\pi_S(r, u) \propto (ru)^{-3/2}$ is second-order matching when m is the parameter of interest. On the other hand, any first-order matching prior is also second-order matching when either r or u is the parameter of interest.

It may be of interest to find an example where a reference prior is not a second-order matching prior. Consider the gamma p.d.f. $f(x|\mu, \lambda) = (\lambda^\lambda / \Gamma(\lambda)) \cdot \exp[-\lambda y/\mu] y^{\lambda-1} \mu^{-\lambda}$, where the mean μ is the parameter of interest. The Fisher information matrix is given by $\mathrm{Diag}(\lambda \mu^{-2}, \frac{d^2 \log \Gamma(\lambda)}{d\lambda^2} - 1/\lambda)$. Then the two-group reference prior of Bernardo (1979) is given by $\mu^{-1}[\frac{d^2 \log \Gamma(\lambda)}{d\lambda^2} - (1/\lambda)]^{1/2}$, while the unique second-order quantile matching prior is given by $\lambda \mu^{-1} \cdot [\frac{d^2 \log \Gamma(\lambda)}{d\lambda^2} - (1/\lambda)]$.

In some of these examples, especially for the location and location–scale families, one gets exact rather than asymptotic matching. This is especially so when the matching prior is a right-invariant Haar prior. We will see some examples in the next section.

5. OTHER PRIORS

5.1 Invariant Priors

Very often objective priors are derived via some invariance criterion. We illustrate with the location–scale family of distributions.

Let X have p.d.f. $p(x|\mu, \sigma) = \sigma^{-1} f((x - \mu)/\sigma)$, $-\infty < \mu < \infty$, $0 < \sigma < \infty$, where f is a p.d.f. Then, as found in Section 4, the Fisher information matrix $I(\mu, \sigma)$ is of the form $I(\mu, \sigma) = \sigma^{-2} \binom{c_1 \ c_2}{c_2 \ c_3}$. Hence, Jeffreys' general rule prior $\pi_J(\mu, \sigma) \propto \sigma^{-2}$. This prior, as we will see in this section, corresponds to a left-invariant Haar prior. In contrast, Jeffreys' independence prior $\pi_I(\mu, \sigma) \propto \sigma^{-1}$ corresponds to a right-invariant Haar prior.

In order to demonstrate this, consider a group of linear transformations $G = \{g_{a,b} - \infty < a < \infty, b > 0\}$, where $g_{a,b}(x) = a + bx$. The induced group of transformations on the parameter space will be denoted

by $\bar{G}$, where $\bar{G} = \{\bar{g}_{a,b}\}$, where $\bar{g}_{a,b}(\mu, \sigma) = (a + b\mu, b\sigma)$. The general theory of locally compact groups states that there exist two measures η_1 and η_2 on $\bar{G}$ such that η_1 is left-invariant and η_2 is right-invariant. What this means is that for all $\bar{g} \in \bar{G}$ and A a subset of G, $\eta_1(\bar{g}A) = \eta_1(A)$ and $\eta_2(A\bar{g}) = \eta_2(A)$, where $\bar{g}A = \{\bar{g}\bar{g}_* : \bar{g}_* \in A\}$ and $A\bar{g} = \{\bar{g}_*\bar{g} : \bar{g}_* \in A\}$. The measures η_1 and η_2 are referred to respectively as left- and right-invariant Haar measures. For the location–scale family of distributions, the left- and right-invariant Haar priors turn out to be $\pi_L(\mu, \sigma) \propto \sigma^{-2}$ and $\pi_R(\mu, \sigma) \propto \sigma^{-1}$, respectively (cf. Berger, 1985, pages 406–407; Ghosh, Delampady and Samanta, 2006, pages 136–138).

The right-Haar prior usually enjoys more optimality properties than the left-Haar prior. Some optimality properties of left-Haar priors are given in Datta and Ghosh (1995b). In Example 1, for the location–scale family of distributions, the right-Haar prior is Bernardo's reference prior when either μ or σ is the parameter of interest, while the other parameter is the nuisance parameter. Also, it is shown in Datta, Ghosh and Mukerjee (2000) that for the location–scale family of distributions, the right-Haar prior yields *exact* matching of the coverage probabilities of Bayesian credible intervals and the corresponding frequentist confidence intervals when either μ or σ is the parameter of interest, while the other parameter is the nuisance parameter.

For simplicity, we demonstrate this only for the normal example. Let $X_1, \ldots, X_n|\mu, \sigma^2$ be i.i.d. $N(\mu, \sigma^2)$, where $n \geq 2$. With the right-Haar prior $\pi_2(\mu, \sigma) \propto \sigma^{-1}$, the marginal posterior distribution of μ is Student's t with location parameter $\bar{X} = \sum_{i=1}^{n} X_i/n$, scale parameter $S/\sqrt{n}$, where $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$, and degrees of freedom $n - 1$. Hence, if $\mu_{1-\alpha}$ denotes the $100(1 - \alpha)$th percentile of this marginal posterior, then

$$
\begin{aligned}
1 - \alpha &= P(\mu \leq \mu_{1-\alpha}|X_1, \ldots, X_n) \\
&= P\big[\sqrt{n}(\mu - \bar{X})/S \\
&\quad \leq \sqrt{n}(\mu_{1-\alpha} - \bar{X})/S|X_1, \ldots, X_n\big] \\
&= P\big[t_{n-1} \leq \sqrt{n}(\mu_{1-\alpha} - \bar{X})/S\big],
\end{aligned}
$$

so that $\sqrt{n}(\mu_{1-\alpha} - \bar{X})/S = t_{n-1,1-\alpha}$, the $100(1-\alpha)$th percentile of t_{n-1}. Now

$$
\begin{aligned}
&P(\mu \leq \mu_{1-\alpha}|\mu, \sigma) \\
&= P\big[\sqrt{n}(\bar{X} - \mu)/S \geq -t_{n-1,1-\alpha}|\mu, \sigma\big] = 1 - \alpha \\
&= P(\mu \leq \mu_{1-\alpha}|X_1, \ldots, X_n).
\end{aligned}
$$

This provides the exact coverage matching probability for μ.

Next, with the same set up, when σ^2 is the parameter of interest, its marginal posterior is Inverse Gamma$((n-1)/2, (n-1)S^2/2)$. Now, if $\sigma^2_{1-\alpha}$ denotes the $100(1-\alpha)$th percentile of this marginal posterior, then $\sigma^2_{1-\alpha} = (n-1)S^2/\chi^2_{n-1;1-\alpha}$, where $\chi^2_{n-1;1-\alpha}$ is the $100(1-\alpha)$th percentile of the χ^2_{n-1} distribution. Now

$$P(\sigma^2 \le \sigma^2_{1-\alpha}|\mu, \sigma)$$
$$= P[(n-1)S^2/\sigma^2 \le \chi^2_{n-1;1-\alpha}|\mu, \sigma] = 1 - \alpha,$$

showing once again the exact coverage matching.

The general definition of a right-invariant Haar density on $\bar{\mathcal{G}}$ which we will denote by h_r must satisfy $\int_{A\bar{g}_0} h_r(x)\,dx = \int_A h_r(x)\,dx$, where $A\bar{g} = \{\bar{g}_*\bar{g} : \bar{g}_* \in A\}$. Similarly, a left invariant Haar density on $\bar{\mathcal{G}}$ which we will denote by h_l must satisfy $\int_{\bar{g}A} h_l(x)\,dx = \int_A h_l(x)\,dx$, where $\bar{g}A = \{\bar{g}\bar{g}_* : \bar{g}_* \in A\}$. An alternate representation of the right- and left-Haar densities are given by $P^{h_r}(A\bar{g}) = P^{h_r}(A)$ and $P^{h_l}(\bar{g}A) = P^{h_l}(A)$, respectively.

It is shown in Halmos (1950) and Nachbin (1965) that the right- and left-invariant Haar densities exist and are unique up to a multiplicative constant. Berger (1985) provides calculation of h_r and h_l in a very general framework. He points out that if $\bar{\mathcal{G}}$ is isomorphic to the parameter space Θ, then one can construct right- and left-invariant Haar priors on the parameter space Θ. A very substantial account of invariant Haar densities is available in Datta and Ghosh (1995b). Severini, Mukerjee and Ghosh (2002) have demonstrated the exact matching property of right invariant Haar densities in a prediction context under fairly general conditions.

5.2 Moment Matching Priors

Here we discuss a new matching criterion which we will refer to as the "moment matching criterion." For a regular family of distributions, the classic article of Bernstein and Von Mises (see, e.g., Ferguson, 1996, page 141; Ghosh, Delampady and Samanta, 2006, page 104) proved the asymptotic normality of the posterior of a parameter vector centered around the maximum likelihood estimator or the posterior mode and variance equal to the inverse of the observed Fisher information matrix evaluated at the maximum likelihood estimator or the posterior mode. We utilize the same asymptotic expansion to find priors which can provide high order matching of the moments of the posterior mean and the maximum likelihood estimator.

For simplicity of exposition, we shall primarily confine ourselves to priors which achieve the matching of the first moment, although it is easy to see how higher order moment matching is equally possible.

The motivation for moment matching priors stems from several considerations. First, these priors lead to posterior means which share the asymptotic optimality of the MLE's up to a high order. In particular, if one is interested in asymptotic bias or MSE reduction of the MLE's through some adjustment, the same adjustment applies directly to the posterior means. In this way, it is possible to achieve Bayes-frequentist synthesis of point estimates. The second important aspect of these priors is that they provide new viable alternatives to Jeffreys' prior even for real-valued parameters in the absence of nuisance parameters motivated from the proposed criterion. A third motivation, which will be made clear later in this section, is that with moment matching priors, it is possible to construct credible regions for parameters of interest based only on the posterior mean and the posterior variance, which match the maximum likelihood based confidence intervals to a high order of approximation. We will confine ourselves primarily to regular families of distributions.

Let $X_1, X_2, \ldots, X_n|\theta$ be independent and identically distributed with common density function $f(x|\theta)$, where $\theta \in \Theta$, some interval in the real line. Consider a general class of priors $\pi(\theta), \theta \in \Theta$ for θ. Throughout, it is assumed that both f and π satisfy all the needed regularity conditions as given in Johnson (1970) and Bickel and Ghosh (1990).

Let $\hat{\theta}_n$ denote the maximum likelihood estimator of θ. Under the prior π, we denote the posterior mean of θ by $\hat{\theta}_n^B$. The formal asymptotic expansion given in Section 2 now leads to $\hat{\theta}_n^B = \hat{\theta}_n + n^{-1}(\frac{a_3}{2\hat{I}_n^2} + \frac{1}{\hat{I}_n}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}) + O_p(n^{-3/2})$, where a_3 and $\hat{I}_n$ are defined in Theorem 1. The law of large numbers and consistency of the MLE now give $n(\hat{\theta}_n^B - \hat{\theta}_n) \xrightarrow{P} (\frac{-g_3(\theta)}{2I^2(\theta)} + \frac{1}{I(\theta)}\frac{\pi'(\theta)}{\pi(\theta)})$. With the choice $\pi(\theta) = \exp[-\frac{1}{2}\int^\theta \frac{g_3(t)}{I(t)}\,dt]$, one gets $\hat{\theta}_n^B - \hat{\theta}_n = O_p(n^{-3/2})$. We will denote this prior as $\pi_M(\theta)$.

Ghosh and Liu (2011) have shown that if ϕ is a one-to-one function of θ, then the moment matching prior $\pi_M(\phi)$ for ϕ is given by $\pi_M(\phi) = \pi_M(\theta)|\frac{d\theta}{d\phi}|^{3/2}$. We now see an application of this result.

EXAMPLE 6 (Continued). Consider the regular one-parameter exponential family of densities given by $f(x|\theta) = \exp[\theta x - \psi(\theta) + h(x)]$. For the canonical parameter θ, noting that $I(\theta) = \psi''(\theta)$ and $g_3(\theta) = \psi'''(\theta) = I'(\theta)$, $\pi_M(\theta) = \exp[\frac{1}{2}\int I'(\theta)/I(\theta)\,d\theta] =$

$I^{1/2}(\theta)$, which is Jeffreys' prior. On the other hand, for the population mean $\phi = \psi'(\theta)$ which is a strictly increasing function of θ [since $\psi''(\theta) = V(X|\theta) > 0$], the moment matching prior $\pi_M(\phi) = I(\phi)$. In particular, for the binomial proportion p, one gets the Haldane prior $\pi_H(p) \propto p^{-1}(1-p)^{-1}$, which is the same as Hartigan's (1964, 1998) maximum likelihood prior. However, for the canonical parameter $\theta = \text{logit}(p)$, whereas we get Jeffreys' prior, Hartigan (1964, 1998) gets the Laplace uniform$(0, 1)$ prior.

REMARK 4. It is now clear that a fundamental difference between priors obtained by matching probabilities and those obtained by matching moments is the lack of invariance of the latter under one-to-one reparameterization. It may be interesting to find conditions under which a moment matching prior agrees with Jeffreys' prior $I^{1/2}(\theta)$ or the uniform constant prior. The former holds if and only if $g_3(\theta) = I'(\theta)$, while the latter holds if and only if $g_3(\theta) = 0$.

The if part of the above results are immediate from the definition of $\pi_M(\theta)$. To prove the only if parts, note that if $\pi_M(\theta) = I^{1/2}(\theta)$, first taking logarithms, and then differentiating with respect to θ, one gets $\frac{I'(\theta)}{2I(\theta)} = \frac{g_3(\theta)}{2I(\theta)}$ so that $g_3(\theta) = I'(\theta)$. On the other hand, if $\pi(\theta) = c$, then taking logarithms, and then differentiating with respect to θ, one gets $g_3(\theta) = 0$.

The above approach can be extended to the matching of higher moments as well. Noting that $V_\pi(\theta|X_1, \ldots, X_n) = E_\pi[(\theta - \hat{\theta}_n)^2|X_1, \ldots, X_n)] - (\hat{\theta}_n^B - \hat{\theta}_n)^2$, it follows immediately that under the moment matching prior π_M, $V_\pi(\theta|X_1, \ldots, X_n) = (n\hat{I}_n)^{-1} + O_p(n^{-2})$. This fact helps construction of credible intervals for θ, the parameter of interest, centered at the posterior mean and scaled by the posterior standard deviation which enjoys the same asymptotic properties as the credible interval centered at the MLE and scaled by the square root of the reciprocal of the observed Fisher information number.

6. SUMMARY AND CONCLUSION

As mentioned in the Introduction, this article provides a selective review of objective priors reflecting my own interest and familiarity with the topics. I am well aware that many important contributions are left out. For instance, I have discussed only the two-group reference priors of Bernardo (1979). A more appealing later contribution by Berger and Bernardo (1992b) provided an algorithm for the construction of multigroup reference priors when these groups are arranged

in accordance to their order of importance. In particular, the one-at-a-time reference priors, as advocated by these authors, has proved to be quite useful in practice. Ghosal (1997, 1999) provided the construction of reference priors in nonregular cases, while a formal definition of reference priors encompassing both regular and nonregular cases has recently been proposed by Berger, Bernardo and Sun (2009).

Regarding probability matching priors, we have discussed only the quantile matching criterion. There are several others, possibly equally important probability matching criteria. Notable among these are the highest posterior density matching criterion as well as matching via inversion of test statistics, such as the likelihood ratio test statistic, Rao statistic or the Wald statistic. Extensive discussion of such matching priors is given in Datta and Mukerjee (2004). Datta et al. (2000) constructed matching priors via the prediction criterion, and related exact results in this context are available in Fraser and Reid (2002). The issue of matching priors in the context of conditional inference has been discussed quite extensively in Reid (1996).

A different class of priors called "the maximum likelihood prior" was developed by Hartigan (1964, 1998). Roughly speaking, these priors are found by maximizing the expected distance between the prior and the posterior under a truncated Kullback–Leibler distance. Like the proposed moment matching priors, the maximum likelihood prior densities, when they exist, result in posterior means asymptotically negligible from the MLE's. I have alluded to some of these priors as a comparison with other priors as given in this paper.

With the exception of the right- and left-invariant Haar priors, the derivation of the remaining priors are based essentially on the asymptotic expansion of the posterior density as well as the shrinkage argument of J. K. Ghosh. This approach provides a nice unified tool for the development of objective priors. I believe very strongly that many new priors will be found in the future by either a direct application or slight modification of these tools.

The results of this article show that Jeffreys' prior is a clear winner in the absence of nuisance parameters for most situations. The only exception is the chisquare divergence where different priors may emerge. But that corresponds only to one special case, namely, the boundary of the class of divergence priors, while Jeffreys' prior continues its optimality in the interior. In the presence of nuisance parameters, my own recommendation is to find two- or multi-group reference priors following the algorithm of Berger and Bernardo

(1992a), and then narrow down this class of priors by finding their intersection with the class of probability matching priors. This approach can even lead to a unique objective prior in some situations. Some simple illustrations are given in this article. I also want to point out the versatility of reference priors. For example, for nonregular models, Jeffreys' general rule prior does not work. But as shown in Ghosal (1997) and Berger, Bernardo and Sun (2009), one can extend the definition of reference priors to cover these situations as well.

The examples given in this paper are purposely quite simplistic to aid understanding mainly of readers not familiar at all with the topic. Quite rightfully, they can be criticized as somewhat stylized. Both reference and probability matching priors, however, have been developed for more complex problems of practical importance. Among others, I may refer to Berger and Yang (1994), Berger, De Oliveira and Sanso (2001), Ghosh and Heo (2003), Ghosh, Carlin and Srivastava (1994) and Ghosh, Yin and Kim (2003). The topics of these papers include time series models, spatial models and inverse problems, such as linear calibration and problems in bioassay, in particular, slope ratio and parallel line assays. One can easily extend this list. A very useful source for all these papers is Bernardo (2005).

ACKNOWLEDGMENTS

REFERENCES

AMARI, S. (1982). Differential geometry of curved exponential families—Curvatures and information loss. *Ann. Statist.* **10** 357–387. MR0653513

BAYES, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **53** 370–418.

BERGER, J. O. (1985). *Statistical Decision Theory and Related Topics*, 2nd ed. Springer, New York. MR0804611

BERGER, J. O. and BERNARDO, J. M. (1989). Estimating a product of means. Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84** 200–207. MR0999679

BERGER, J. O. and BERNARDO, J. M. (1992a). On the development of reference priors (with discussion). In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 35–60. Oxford Univ. Press, New York. MR1380269

BERGER, J. O. and BERNARDO, J. M. (1992b). Reference priors in a variance components problem. In *Bayesian Analysis in Statistics and Econometrics* (P. K. Goel and N. S. Iyengar, eds.) 177–194. Springer, New York. MR1194392

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37** 905–938. MR2502655

BERGER, J. O., DE OLIVEIRA, V. and SANSO, B. (2001). Objective Bayesian analysis of spatially correlated data. *J. Amer. Statist. Assoc.* **96** 1361–1374. MR1946582

BERGER, J. O. and YANG, R. (1994). Noninformative priors and Bayesian testing for the AR(1) model. *Econometric Theory* **10** 461–482. MR1309107

BERNARDO, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **41** 113–147. MR0547240

BERNARDO, J. M. (2005). Reference analysis. In *Handbook of Bayesian Statisics* **25** (D. K. Dey and C. R. Rao, eds.). North-Holland, Amsterdam. MR2490522

BHATTACHARYYA, A. K. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **35** 99–109. MR0010358

BICKEL, P. J. and GHOSH, J. K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction—A Bayesian argument. *Ann. Statist.* **18** 1070–1090. MR1062699

CLARKE, B. and BARRON, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory* **36** 453–471. MR1053841

CLARKE, B. and BARRON, A. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Statist. Plann. Inference* **41** 37–60. MR1292146

CLARKE, B. and SUN, D. (1997). Reference priors under the chi-square distance. *Sankhyā A* **59** 215–231. MR1665703

CLARKE, B. and SUN, D. (1999). Asymptotics of the expected posterior. *Ann. Inst. Statist. Math.* **51** 163–185. MR1704652

CRESSIE, N. and READ, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46** 440–464. MR0790631

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy Statist. Soc. Ser. B* **53** 79–109. MR0893334

DATTA, G. S. (1996). On priors providing frequentist validity of Bayesian inference for multiple parametric functions. *Biometrika* **83** 287–298. MR1439784

DATTA, G. S. and GHOSH, J. K. (1995a). On priors providing frequentist validity for Bayesian inference. *Biometrika* **82** 37–45. MR1332838

DATTA, G. S. and GHOSH, J. K. (1995b). Noninformative priors for maximal invariant in group models. *Test* **4** 95–114. MR1365042

DATTA, G. S. and GHOSH, M. (1995c). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90** 1357–1363. MR1379478

DATTA, G. S. and GHOSH, M. (1995d). Hierarchical Bayes estimators of the error variance in one-way ANOVA models. *J. Statist. Plann. Inference* **45** 399–411. MR1341333

DATTA, G. S. and GHOSH, M. (1996). On the invariance of noninformative priors. *Ann. Statist.* **24** 141–159. MR1389884

DATTA, G. S., GHOSH, M. and MUKERJEE, R. (2000). Some new results on probability matching priors. *Calcutta Statist. Assoc. Bull.* **50** 179–192. MR1843620

DATTA, G. S., GHOSH, M. and KIM, Y. (2002). Probability matching priors for one-way unbalanced rendom effects models. *Statist. Decisions* **20** 29–51. MR1904422

DATTA, G. S. and MUKERJEE, R. (2004). *Probability Matching Priors: Higher Order Asymptotics.* Springer, New York. MR2053794

DATTA, G. S., MUKERJEE, R., GHOSH, M. and SWEETING, T. J. (2000). Bayesian prediction with approximate frequentist validity. *Ann. Statist.* **28** 1414–1426. MR1805790

DATTA, G. S. and SWEETING, T. J. (2005). Probability matching priors. In *Bayesian Thinking, Modeling and Computation. Handbook of Statistics* **25** (D. K. Dey and C. R. Rao, eds.). North-Holland, Amsterdam. MR2490523

FERGUSON, T. (1996). *A Course in Large Sample Theory.* Chapman & Hall/CRC Press, Boca Raton, FL. MR1699953

FRASER, D. A. S. and REID, N. (2002). Strong matching of frequentist and Bayesian parametric inference. *J. Statist. Plann. Inference* **103** 263–285. MR1896996

GHOSAL, S. (1997). Reference priors in multiparameter nonregular cases. *Test* **6** 159–186. MR1466439

GHOSAL, S. (1999). Probability matching priors for nonregular cases. *Biometrika* **86** 956–964. MR1741992

GHOSH, J. K., DELAMPADY, M. and SAMANTA, T. (2006). *An Introduction to Bayesian Analysis.* Springer, New York. MR2247439

GHOSH, M. and LIU, R. (2011). Moment matching priors. *Sankhyā A.* To appear.

GHOSH, J. K. and MUKERJEE, R. (1992). Non-informative priors (with discussion). In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 195–210. Oxford Univ. Press, New York. MR1380277

GHOSH, J. K., SINHA, B. K. and JOSHI, S. N. (1982). Expansion for posterior probability and integrated Bayes risk. In *Statistical Decision Theory and Related Topics III* **1** 403–456. Academic Press, New York. MR0705299

GHOSH, M., CARLIN, B. P. and SRIVASTAVA, M. S. (1994). Probability matching priors for linear calibration. *Test* **4** 333–357. MR1379796

GHOSH, M. and HEO, J. (2003). Default Bayesian priors for regression models with second-order autogressive residuals. *J. Time Ser. Anal.* **24** 269–282. MR1984597

GHOSH, M., MERGEL, V. and LIU, R. (2011). A general divergence criterion for prior selection. *Ann. Inst. Statist. Math.* **63** 43–58.

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.) 227–252. Nova Science Publishers, New York. MR1673669

GHOSH, M., YIN, M. and KIM, Y.-H. (2003). Objective Bayesian inference for ratios of regression coefficients in linear models. *Statist. Sinica* **13** 409–422. MR1977734

HALMOS, P. (1950). *Measure Theory.* Van Nostrand, New York. MR0033869

HARTIGAN, J. A. (1964). Invariant prior densities. *Ann. Math. Statist.* **35** 836–845. MR0161406

HARTIGAN, J. A. (1998). The maximum likelihood prior. *Ann. Statist.* **26** 2083–2103. MR1700222

HELLINGER, E. (1909). Neue Begründung der Theorie quadratischen Formen von unendlichen vielen Veränderlichen. *J. Reine Angew. Math.* **136** 210–271.

HUZURBAZAR, V. S. (1950). Probability distributions and orthogonal parameters. *Proc. Camb. Phil. Soc.* **46** 281–284. MR0034567

JEFFREYS, H. (1961). *Theory of Probability and Inference*, 3rd ed. Cambridge Univ. Press, London.

JOHNSON, R. A. (1970). Asymptotic expansions associated with posterior distribution. *Ann. Math. Statist.* **41** 851–864. MR0263198

KASS, R. E. and WASSERMAN, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91** 1343–1370. MR1478684

LAPLACE, P. S. (1812). *Theorie Analytique des Probabilities.* Courcier, Paris.

LINDLEY, D. V. (1956). On the measure of the information provided by an experiment. *Ann. Math. Statist.* **27** 986–1005. MR0083936

LIU, R. (2009). On some new contributions towards objective priors. Unpublished Ph.D. dissertation. Dept. Statistics, Univ. Florida, Gainesville, FL. MR2714091

MORRIS, C. N. and NORMAND, S. L. (1992). Hierarchical models for combining information and for meta-analysis. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 321–344. Oxford. Univ. Press, New York. MR1380284

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. MR1248017

MUKERJEE, R. and GHOSH, M. (1997). Second-order probability matching priors. *Biometrika* **84** 970–975. MR1625016

NACHBIN, L. (1965). *The Haar Integral.* Van Nostrand, New York. MR0175995

REID, N. (1996). Likelihood and Bayesian approximation methods (with discussion). In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 351–368. Oxford Univ. Press, New York. MR1425414

SEVERINI, T. A., MUKERJEE, R. and GHOSH, M. (2002). On an exact probability matching property of right-invariant priors. *Biometrika* **89** 952–957. MR1946524

STAICU, A.-M. and REID, N. (2008). On probability matching priors. *Canad. J. Statist.* **36** 613–622. MR2532255

TIBSHIRANI, R. J. (1989). Noninformative priors for one parameter of many. *Biometrika* **76** 604–608. MR1040654

WELCH, B. L. and PEERS, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy Statist. Soc. Ser. B* **35** 318–329. MR0173309

YE, K. (1994). Bayesian reference prior analysis on the ratio of variances for the balanced one-way random effect model. *J. Statist. Plann. Inference* **41** 267–280. MR1309613