The Annals of Statistics
2000, Vol. 28, No. 5, 1414–1426

BAYESIAN PREDICTION WITH APPROXIMATE FREQUENTIST VALIDITY[1]

BY GAURI SANKAR DATTA, RAHUL MUKERJEE, MALAY GHOSH AND TREVOR J. SWEETING

University of Georgia, Indian Institute of Management, University of Florida and University of Surrey

We characterize priors which asymptotically match the posterior coverage probability of a Bayesian prediction region with the corresponding frequentist coverage probability. This is done considering both posterior quantiles and highest predictive density regions with reference to a future observation. The resulting priors are shown to be invariant under reparameterization. The role of Jeffreys' prior in this regard is also investigated. It is further shown that, for any given prior, it may be possible to choose an interval whose Bayesian predictive and frequentist coverage probabilities are asymptotically matched.

1. Introduction. Bayesian analyses are often based on noninformative priors. One important approach for the development of such priors is based on the *probability matching criterion* which requires matching the posterior coverage probability of a Bayesian credible set for a parameter of interest with the corresponding frequentist coverage probability asymptotically up to a certain order. Such priors, originally developed by Welch and Peers (1963), have received considerable attention in recent years. Among others, one may refer to Peers (1965), Stein (1985), Tibshirani (1989), Severini (1991, 1993), Ghosh and Mukerjee (1992), Mukerjee and Jey (1993), Nicolaou (1993), DiCiccio and Stern (1994), Sweeting (1995), Datta and Ghosh (1995), Datta (1996), Sun and Ye (1996), Rousseau (1997), Mukerjee and Ghosh (1997), and many other references contained in these papers.

In the absence of nuisance parameters, the matching criterion based on posterior quantiles leads to Jeffreys' prior (Jeffreys, 1961) for a scalar parameter. However, this need not necessarily be so in the presence of nuisance parameters. Furthermore, different matching priors may emerge depending on which parameter is viewed as the parameter of interest.

A natural alternative approach is to match asymptotically the coverage probability of a Bayesian credible set of a future observation with the corresponding frequentist probability. This is particularly attractive when the main problem is prediction, and not estimation, and thus there is no particular reason to treat a certain parameter as the parameter of interest in preference to

---

[1]Supported in part by the ASA/NSF/BLS/Census Fellowship Program, the Indian Institute of Management and the National Science Foundation.

AMS 1991 *subject classifications*. 62C10, 62F15.

Key words and phrases. Highest predictive density region, Jeffreys' prior, noninformative prior, posterior quantile, prediction interval.

others. The focus in this paper is therefore on prediction. Our study provides some theoretical insight into the relationship between Bayesian and frequentist approaches to predictive inference. Furthermore, the results of this paper may contribute to the development of objective Bayesian methodology, which is an area of increasing interest.

Two popular ways of coverage matching are through (a) posterior quantiles, and (b) highest posterior density regions. One of the aims of the present article is to characterize priors that accomplish this goal both via (a) and (b) for the prediction problem. This has been carried out in Sections 3 and 5 respectively after presenting the preliminaries in the next section. Since the leading term in the asymptotic expansion for the posterior predictive density is usually nonnormal, our results differ from those arising in the context of probability matching priors for parameters. For example, under prediction based on quantiles, even in the scalar parameter case, Jeffreys' prior does not automatically emerge as a solution of a differential equation but, as Theorem 1 below reveals, a new approach is needed in examining its role. Also, Bartlett adjustability cannot be invoked in handling the highest posterior density region for prediction (contrast DiCiccio and Stern (1994)). The examples presented in these sections indicate that the predictive matching approach is a promising tool for the development of sensible objective priors.

In Section 4 we consider the construction of prediction intervals in the scalar parameter case which have approximately equal Bayesian predictive and frequentist coverage probabilities. This provides a predictive analogue of the construction of posterior intervals having approximately equal Bayesian and frequentist coverage. The latter problem is considered by Severini (1993) and Sweeting (1999). The technique used in Section 4 for the construction of matching prediction regions, however, is quite different from that used for the construction of matching posterior regions.

The general discussion on Bayesian prediction is admirably presented in the books of Aitchison and Dunsmore (1975) and Geisser (1993); see also Kuboki (1998) for a study of reference priors for prediction using an information theoretic approach. To our knowledge, however, this matching idea as described earlier has not been addressed before. Recently, primarily in the context of frequentist inference under a curved exponential model, Komaki (1996) gave an asymptotic expression for a posterior predictive density and his overall recommendation seems to be in favour of Jeffreys' prior. Our explicit characterizations help in understanding how far Jeffreys' prior can yield asymptotically valid frequentist inference for the problem of prediction and demonstrate that it works only in some but not all situations.

Our method should be of appeal also to frequentists because we are providing asymptotically valid frequentist procedures which have valid Bayesian interpretations as well. Our analysis also provides an insight into the theoretical problem of defining a predictive distribution from a frequentist standpoint. In this connection, we refer to Barndorff-Nielsen and Cox (1996) and Vidoni (1998) for recent results on frequentist prediction and further references.

2. Preliminaries. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed possibly vector-valued random variables with a common density $f(x;\theta)$ where $\theta = (\theta_1, \ldots, \theta_p)'$ is a parameter vector lying in an open subset Ω of R^p. We consider Bayesian prediction of X_{n+1}, with approximate frequentist validity, based on $d = (X_1, \ldots, X_n)'$ using a prior density $\pi(\cdot)$ which is supposed to be positive and thrice continuously differentiable for all θ. Along the lines of Ghosh and Mukerjee (1993), we work essentially under the assumptions of Johnson (1970) and also need the Edgeworth assumptions of Bickel and Ghosh (1990). All formal expansions for the posterior, as used here, are valid for sample points in a set S with P_θ-probability $1 + o(n^{-1})$ uniformly over compact sets of θ. The set S may be defined following Section 2 of Bickel and Ghosh (1990).

Let $l(\theta) = n^{-1}\sum_{i=1}^n \log f(X_i;\theta)$ and $\hat{\theta}$ be the maximum likelihood estimator of θ based on d. With $D_j \equiv \partial/\partial\theta_j$, let

$$a_{jr} = \{D_j D_r l(\theta)\}_{\theta=\hat{\theta}},\ a_{jrs} = \{D_j D_r D_s l(\theta)\}_{\theta=\hat{\theta}},\ c_{jr} = -a_{jr},$$
$$\pi_j(\theta) = D_j\pi(\theta),\ f_j(x;\theta) = D_j f(x;\theta),\ f_{jr}(x;\theta) = D_j D_r f(x;\theta).$$

The matrix $C = ((c_{jr}))$ will be positive definite over S. Let $C^{-1} = ((c^{jr}))$ and $\tilde{\pi}(x_{n+1}|d)$ be the posterior predictive density of X_{n+1} given d under the prior $\pi(\cdot)$. Then algebra similar to that in Ghosh and Mukerjee (1991) [see also Komaki (1996)] shows that

$$\begin{aligned}\tilde{\pi}(x_{n+1}|d) &= f(x_{n+1};\hat{\theta}) \\ &\quad + \frac{1}{2n}\left[c^{st}\left\{c^{jr}a_{jrs} + \frac{2\pi_s(\hat{\theta})}{\pi(\hat{\theta})}\right\} f_t(x_{n+1};\hat{\theta})\right. \\ &\qquad \left. + c^{jr} f_{jr}(x_{n+1};\hat{\theta})\right] + o(n^{-1}).\end{aligned} \tag{2.1}$$

In (2.1) and elsewhere, unless otherwise specified, we follow the summation convention with sums ranging from 1 to p.

3. Frequentist validity of posterior quantiles. We first consider the case where the X_i, $i \geq 1$, are scalar-valued. Then the posterior quantiles of X_{n+1}, given $d = (X_1, \ldots, X_n)'$ are well-defined. For $0 < \alpha < 1$, let $q(\theta,\alpha)$ be such that

$$\int_{q(\theta,\alpha)}^{\infty} f(u;\theta)du = \alpha. \tag{3.1}$$

Denote the posterior probability measure under the prior $\pi(\cdot)$ by $P^\pi\{\cdot|d\}$. Define

$$\mu_j(\theta,\alpha) = \int_{q(\theta,\alpha)}^{\infty} f_j(u;\theta)du,\ \mu_{jr}(\theta,\alpha) = \int_{q(\theta,\alpha)}^{\infty} f_{jr}(u;\theta)du, \tag{3.2}$$

$$h(\pi,\alpha)=q(\hat\theta,\alpha)+\frac{1}{nf(q(\hat\theta,\alpha);\hat\theta)}\left[c^{st}\left\{\frac{1}{2}c^{jr}a_{jrs}+\frac{\pi_s(\hat\theta)}{\pi(\hat\theta)}\right\}\mu_t(\hat\theta,\alpha)+\frac{1}{2}c^{jr}\mu_{jr}(\hat\theta,\alpha)\right]. \tag{3.3}$$

Then by (2.1), $P^{\pi}\{X_{n+1}>h(\pi,\alpha)|d\}=\alpha+o(n^{-1})$. Thus $h(\pi,\alpha)$ provides an explicit representation for the $(1-\alpha)$th posterior quantile of X_{n+1} up to the order of approximation $o(n^{-1})$.

We now proceed to characterize priors ensuring approximate frequentist validity of the posterior quantiles of X_{n+1}. The approach of Ghosh and Mukerjee (1993) helps in computing $P_\theta\{X_{n+1}>h(\pi,\alpha)\}$. We take an auxiliary prior $\bar\pi(\cdot)$, satisfying the conditions of Bickel and Ghosh (1990), such that $\bar\pi(\cdot)$ and its first order partial derivatives vanish on the boundaries of a rectangle containing θ as an interior point. We find $P^{\bar\pi}\{X_{n+1}>h(\pi,\alpha)|d\}$, up to $o(n^{-1})$, using an approximation, analogous to (2.1), for the posterior density of X_{n+1} given d under $\bar\pi(\cdot)$. Then $E_\theta[P^{\bar\pi}\{X_{n+1}>h(\pi,\alpha)|d\}]$, as computed up to $o(n^{-1})$, is integrated with respect to $\bar\pi(\cdot)$ and finally $\bar\pi(\cdot)$ is allowed to converge weakly to the degenerate measure at θ to get an approximation for $P_\theta\{X_{n+1}>h(\pi,\alpha)\}$. By (3.3), the above steps yield

$$P^{\bar\pi}\{X_{n+1}>h(\pi,\alpha)|d\}=\alpha+\frac{1}{n}c^{st}\left\{\frac{\bar\pi_s(\hat\theta)}{\bar\pi(\hat\theta)}-\frac{\pi_s(\hat\theta)}{\pi(\hat\theta)}\right\}\mu_t(\hat\theta,\alpha)+o(n^{-1}),$$

$$P_\theta\{X_{n+1}>h(\pi,\alpha)\}=\alpha-\frac{1}{n\pi(\theta)}D_s\{I^{st}\mu_t(\theta,\alpha)\pi(\theta)\}+o(n^{-1}), \tag{3.4}$$

where $I\equiv I(\theta)$ is the per observation (expected) Fisher information matrix at θ and $I^{-1}=((I^{st}))$.

The right hand side of (3.4) equals $\alpha+o(n^{-1})$ if and only if

$$D_s\{I^{st}\mu_t(\theta,\alpha)\pi(\theta)\}=0. \tag{3.5}$$

A prior $\pi(\cdot)$, satisfying (3.5) for every α, will ensure frequentist validity, up to $o(n^{-1})$, of the posterior quantiles of X_{n+1}.

THEOREM 1. *For scalar θ, if there exists a prior satisfying* (3.5) *for every α then it must be Jeffreys' prior*.

PROOF. Differentiation of both sides of (3.1) with respect to α yields

$$-f(q(\theta,\alpha);\theta)\frac{\partial q(\theta,\alpha)}{\partial\alpha}=1. \tag{3.6}$$

Suppose there exists a prior, say $\pi_0(\theta)$, satisfying (3.5) for every α. Then

$$\mu_1(\theta,\alpha)=\psi(\alpha)I(\theta)/\pi_0(\theta), \tag{3.7}$$

where $\psi(\alpha)$ does not involve the scalar-valued parameter θ. Also, following (3.2),

$$\mu_1(\theta, \alpha) = \int_{q(\theta,\alpha)}^{\infty} f_\theta(x; \theta) dx, \tag{3.8}$$

with $f_\theta(x; \theta) = \frac{\partial}{\partial \theta} f(x; \theta)$. Transforming $x = q(\theta, \beta)$ in the right hand side of (3.8) and using (3.6), one gets

$$\mu_1(\theta, \alpha) = \int_0^{\alpha} \frac{f_\theta(q(\theta, \beta); \theta)}{f(q(\theta, \beta); \theta)} d\beta. \tag{3.9}$$

Hence differentiation of both sides of (3.7) with respect to α yields

$$\frac{f_\theta(q(\theta, \alpha); \theta)}{f(q(\theta, \alpha); \theta)} = \frac{d}{d\alpha}\psi(\alpha)\frac{I(\theta)}{\pi_0(\theta)}. \tag{3.10}$$

Now observe that, analogous to (3.9),

$$I(\theta) = \int_{-\infty}^{\infty} \frac{\{f_\theta(x; \theta)\}^2}{f(x; \theta)} dx = \int_0^1 \left\{ \frac{f_\theta(q(\theta, \beta); \theta)}{f(q(\theta, \beta); \theta)} \right\}^2 d\beta. \tag{3.11}$$

Use of (3.10) in (3.11) shows that $I(\theta) \propto \{I(\theta)/\pi_0(\theta)\}^2$, that is, $\pi_0(\theta) \propto \{I(\theta)\}^{1/2}$, which proves the result. □

In particular, with the one-parameter location model or the one-parameter scale model, Jeffreys' prior satisfies (3.5) for every α. This can be verified along the lines of Example 2 below. Even outside one-parameter location or scale models, Jeffreys' prior may satisfy (3.5) for every α. This happens, for example, with the model specified by

$$f(x; \theta) = \theta(1 + \theta)(x + \theta)^{-2}, \ 0 < x < 1,$$

where $\theta > 0$. Here the Jeffreys' prior is proportional to $\{\theta(1 + \theta)\}^{-1}$. At the same time, even with scalar θ, there can be models where Jeffreys' prior does not satisfy (3.5) and hence, by Theorem 1, no solution to (3.5), valid for every α, is available. The following example serves as an illustration.

EXAMPLE 1. Let $f(x; \theta)$ represent the univariate normal model with both mean and variance equal to $\theta(> 0)$. Then $I(\theta) = (2\theta + 1)/(2\theta^2)$, and by (3.1), (3.2),

$$q(\theta, \alpha) = \theta + z_\alpha \theta^{1/2}, \ \mu_1(\theta, \alpha) = \phi(z_\alpha)\left(\theta^{1/2} + \tfrac{1}{2} z_\alpha\right)/\theta,$$

where $\phi(\cdot)$ is the standard univariate normal density and z_α is the corresponding $(1 - \alpha)$th quantile. Hence it can be seen that for no α Jeffreys' prior satisfies (3.5).

Even in situations such as this, where no solution to (3.5) valid for every α is available, the explicit formula (3.4) can be useful in comparing priors. A smaller absolute value of the term of order $O(n^{-1})$ in the right-hand side of (3.4) is indicative of a closer proximity to the correct frequentist coverage. To illustrate this point, one may consider Jeffreys' prior and the prior $\pi(\theta) \propto \theta^2$.

Following Ghosh and Mukerjee (1993), the latter prior ensures frequentist validity, up to $o(n^{-1})$, of the highest posterior density regions for θ. Then it can be checked that Jeffreys' prior entails a smaller absolute value of the term of order $O(n^{-1})$ in the right-hand side of (3.4) than the other prior if and only if either

$$z_\alpha < -(2/3)\{\theta^{1/2}(4+5\theta)\}/(1+\theta) \text{ or } z_\alpha > -2\theta^{1/2}. \tag{3.12}$$

For each θ, the condition (3.12) holds for an overwhelming part of the range (0,1) for α. Thus on the whole, Jeffreys' prior behaves much better than the other one with regard to the frequentist coverage of the posterior quantiles of a future observation.

Theorem 1 does not hold for vector-valued θ. As illustrated by the next example, there it is possible that Jeffreys' prior does not satisfy (3.5) but another solution to (3.5), valid for every α, is available.

EXAMPLE 2. Consider the location-scale model given by $f(x;\theta) = \theta_2^{-1} f^*((x-\theta_1)/\theta_2)$, where $-\infty < \theta_1 < \infty$ and $\theta_2 > 0$. Let k_α be such that $\int_{k_\alpha}^{\infty} f^*(u)du = \alpha$. Then by (3.1), (3.2), $q(\theta,\alpha) = \theta_1 + k_\alpha\theta_2$, $\mu_1(\theta,\alpha) = \theta_2^{-1} f^*(k_\alpha)$, $\mu_2(\theta,\alpha) = \theta_2^{-1} k_\alpha f^*(k_\alpha)$. Also, $I^{st} = b^{st}\theta_2^2$ for each s, t, where b^{st} is free from θ. It can be seen that (3.5) reduces to

$$r^j(\alpha) D_j(\theta_2 \pi(\theta)) = 0 \qquad \text{for all } \alpha,$$

where $r^j(\alpha) = b^{j1} + b^{j2}k_\alpha$. Hence $\pi(\theta) \propto \theta_2^{-1}$ satisfies (3.5) for every α and it is the *unique* prior satisfying this condition. From Mukerjee and Ghosh (1997), this is also the unique prior ensuring frequentist validity, up to $o(n^{-1})$, of the posterior quantiles of both θ_1 and θ_2.

As a specific illustration, consider the two-parameter Weibull model given by $f(x;\theta) = (\theta_2/\theta_1)(x/\theta_1)^{\theta_2 - 1}\exp\{-(x/\theta_1)^{\theta_2}\}$, where $\theta = (\theta_1, \theta_2)'$, $x > 0$, $\theta_1 > 0$ and $\theta_2 > 0$. Then, since this model can be written via the transformation $y = \log x$ in the form of a location-scale model with location and scale parameters $\phi_1 = \log\theta_1$, $\phi_2 = \theta_2^{-1}$ respectively, it follows that the unique prior satisfying (3.5) for every α is $\pi(\theta) \propto (\theta_1\theta_2)^{-1}$.

Finally, we indicate a satisfying invariance property of any solution to (3.5). Consider any reparameterization of the model given by a one-to-one transformation of θ. Then it can be shown that a prior $\pi(.)$ satisfies (3.5) for every α if and only if, under the reparameterization, the transformed version of $\pi(.)$ satisfies the transformed version of (3.5) for every α. This invariance property can be verified in a straightforward manner for $p = 1$, while for general p one has to proceed as in Theorem 3.1 of Datta and Ghosh (1996). This is in agreement with what happens under the corresponding estimation problem; cf. Datta and Ghosh (1996).

4. Prediction intervals. For the case of scalar θ, we will now consider prediction intervals for a scalar-valued future observation X_{n+1}. It turns out that, for any given prior, it is often possible to choose an interval which

has both Bayesian predictive probability and frequentist coverage probability equal to α, to the order $o(n^{-1})$.

Fix α and let $\gamma_\alpha(\theta)$ be any function, with functional form free from n, satisfying $0 < \alpha < \gamma_\alpha(\theta) < 1$. Write $\gamma_\alpha = \gamma_\alpha(\hat{\theta})$. Then as in the last section,

$$P^\pi\{h(\pi, 1-\gamma_\alpha+\alpha) < X_{n+1} \le h(\pi, 1-\gamma_\alpha)|d\} = \alpha + o(n^{-1}),$$

and as in (3.4),

$$P_\theta\{h(\pi, 1-\gamma_\alpha+\alpha) < X_{n+1} \le h(\pi, 1-\gamma_\alpha)\} = \alpha - \frac{1}{n\pi(\theta)}\frac{d}{d\theta}\left\{\frac{\pi(\theta)\xi_\alpha(\theta)}{I(\theta)}\right\} + o(n^{-1}),$$

where $\xi_\alpha(\theta) = \psi(\gamma_\alpha(\theta), \alpha, \theta)$ and

$$\psi(\gamma, \alpha, \theta) = \int_{q(\theta,1-\gamma+\alpha)}^{q(\theta,1-\gamma)} f_\theta(x;\theta)dx.$$

If we can find a function $\gamma_\alpha(\theta)$ for which $\psi(\gamma_\alpha(\theta), \alpha, \theta) = 0$, then for any prior the Bayesian predictive and the frequentist coverage probabilities will agree to $o(n^{-1})$. Sufficient conditions for the existence and uniqueness of such a function are given in the next lemma. Here $F(x;\theta)$ is the common distribution function of the observations $X_1, X_2, \ldots$.

LEMMA 1. *Suppose that, for each $\theta \in \Omega$, the equation $f_\theta(x;\theta) = 0$ has a unique solution $x(\theta)$, and $f_{\theta x}(x(\theta);\theta) \neq 0$, where $f_{\theta x}(x;\theta) = (\partial^2/\partial x\partial\theta)f(x;\theta)$. Let $\alpha \ge \alpha_0$, where*

$$\alpha_0 = \sup_{\theta\in\Omega}[\max\{F(x(\theta);\theta), 1 - F(x(\theta);\theta)\}]. \tag{4.1}$$

Then there exists a unique solution $\gamma = \gamma_\alpha(\theta)$ in $(\alpha, 1)$ to the equation

$$\psi(\gamma, \alpha, \theta) = 0.$$

PROOF. Without loss of generality, assume that $f_{\theta x}(x(\theta);\theta) > 0$. Then $f_\theta(x;\theta) < f_\theta(x(\theta);\theta) = 0$ for $x < x(\theta)$ and $f_\theta(x;\theta) > 0$ for $x > x(\theta)$. Since

$$\int_{-\infty}^{\infty} f_\theta(x;\theta)dx = 0,$$

we have

$$\psi(\alpha, \alpha, \theta) = \int_{-\infty}^{q(\theta,1-\alpha)} f_\theta(x;\theta)dx = -\int_{q(\theta,1-\alpha)}^{\infty} f_\theta(x;\theta)dx < 0 \tag{4.2}$$

provided that $q(\theta, 1-\alpha) \ge x(\theta)$, and

$$\psi(1, \alpha, \theta) = \int_{q(\theta,\alpha)}^{\infty} f_\theta(x;\theta)dx = -\int_{-\infty}^{q(\theta,\alpha)} f_\theta(x;\theta)dx > 0 \tag{4.3}$$

provided that $q(\theta, \alpha) \le x(\theta)$. Also, for $\alpha < \gamma < \gamma' < 1$,

$$\begin{aligned}\psi(\gamma', \alpha, \theta) - \psi(\gamma, \alpha, \theta) &= \int_{q(\theta,1-\gamma'+\alpha)}^{q(\theta,1-\gamma')} f_\theta(x;\theta)dx - \int_{q(\theta,1-\gamma+\alpha)}^{q(\theta,1-\gamma)} f_\theta(x;\theta)dx \\ &= \int_{q(\theta,1-\gamma)}^{q(\theta,1-\gamma')} f_\theta(x;\theta)dx - \int_{q(\theta,1-\gamma+\alpha)}^{q(\theta,1-\gamma'+\alpha)} f_\theta(x;\theta)dx > 0.\end{aligned}$$

The last inequality is justified by the fact that for x in $(q(\theta, 1-\gamma), q(\theta, 1-\gamma'))$, one has $x > q(\theta, 1-\gamma) > q(\theta, 1-\alpha) \geq x(\theta)$, implying thereby $f_\theta(x;\theta) > 0$. Similarly, for x in $(q(\theta, 1-\gamma+\alpha), q(\theta, 1-\gamma'+\alpha))$, one has $x < q(\theta, 1-\gamma'+\alpha) < q(\theta, \alpha) \leq x(\theta)$, implying thereby $f_\theta(x;\theta) < 0$. Hence the function $\psi(\gamma, \alpha, \theta)$ is increasing in γ; so (4.2) and (4.3) imply that there is a unique function $\gamma_\alpha(\theta)$ with $\alpha < \gamma_\alpha(\theta) < 1$ for which $\psi(\gamma_\alpha(\theta), \alpha, \theta) = 0$. Finally, the conditions required for (4.2) and (4.3) are readily seen to be equivalent to (4.1), which completes the proof. □

Computation of γ_α for any given value of α can easily be achieved using a Newton iteration. Even though the conditions in Lemma 4.1 appear to be restrictive (for example, one needs the coverage probability $\alpha > 1/2$), these are satisfied for all practical purposes. Also, as pointed out by a referee, the conditions of the lemma hold for distributions belonging to the one-parameter exponential family. Although it is not possible to get closed form expressions for the prediction intervals for an arbitrary member of the exponential family, Example 3 below indicates that the prediction intervals based on this lemma are of a natural form.

EXAMPLE 3. Suppose $f(x;\theta) = \theta^{-1}\exp(-x/\theta)$. Then $x(\theta) = \theta$ and $F(x(\theta);\theta) = 1 - e^{-1}$. Thus $\alpha_0 = 1 - e^{-1}$. For $\alpha = 0.9$ or 0.95, condition (4.1) is satisfied. In fact, considering $\psi(\gamma, \alpha, \theta)$ explicitly, it can be seen that in this example a unique $\gamma_\alpha(\theta)$, as envisaged above, exists for every α in (0,1). This $\gamma_\alpha(\theta)$ does not actually involve θ and is given by the unique solution in $(\alpha, 1)$ for γ of

$$(1-\gamma)\log(1-\gamma) - (1-\gamma+\alpha)\log(1-\gamma+\alpha) = 0.$$

Consider now any prior of the form $\pi(\theta) \propto 1/\theta^r$. Then, either from (3.3) or from the exact predictive distribution, it can be seen that the prediction interval considered in this section takes the natural form $(k_1\bar{X}, k_2\bar{X})$, where $\bar{X}$ is the arithmetic mean of $X_1, \cdots, X_n$, and k_1 and k_2 are constants which involve α and r.

5. Highest predictive density region. We now turn to the general case where the X_i, $i \geq 1$, are possibly vector-valued. While the posterior quantiles of X_{n+1} are well-defined for scalar X_i, they do not remain so with vector X_i. Even in the latter situation, however, one may consider a highest predictive density region for predicting X_{n+1}. We now explore the conditions under which such prediction has approximate frequentist validity.

For $0 < \alpha < 1$, let $m(\theta, \alpha)$ be such that

$$\int_A f(u;\theta)du = \alpha, \tag{5.1}$$

where $A \equiv A(\theta, \alpha) = \{u : f(u;\theta) \geq m(\theta, \alpha)\}$. Define $\xi_j(\theta, \alpha)$ by

$$\xi_j(\theta, \alpha) = \int_A f_j(u;\theta)du. \tag{5.2}$$

From (2.1), observe that

$$\tilde{\pi}(x_{n+1}|d) = f\left(x_{n+1}; \hat{\theta} + \frac{1}{n}g(\pi)\right) + \frac{1}{2n}c^{jr}f_{jr}(x_{n+1}; \hat{\theta}) + o(n^{-1}), \tag{5.3}$$

where $g(\pi) = (g_1(\pi), \ldots, g_p(\pi))'$ with

$$g_t(\pi) = c^{st}\left\{\frac{1}{2}c^{jr}a_{jrs} + \frac{\pi_s(\hat{\theta})}{\pi(\hat{\theta})}\right\}.$$

Noting that the second term in the right hand side of (5.3) is free from $\pi(\cdot)$, it follows from (5.1) and (5.3) that the highest predictive density region under the prior $\pi(\cdot)$ for predicting X_{n+1}, with coverage probability $\alpha + o(n^{-1})$ has the form

$$H(\pi, d) = \left\{x_{n+1} : f\left(x_{n+1}; \hat{\theta} + \frac{1}{n}g(\pi)\right) + \frac{1}{2n}c^{jr}f_{jr}(x_{n+1}; \hat{\theta}) \geq m\left(\hat{\theta} + \frac{1}{n}g(\pi), \alpha\right) + \frac{1}{n}\rho(d)\right\},$$

where $\rho(d)$ is at most of order $O(1)$ and is free from $\pi(\cdot)$. The explicit form of $\rho(d)$ is not needed in the sequel for studying approximate frequentist validity of $H(\pi, d)$.

After considerable algebra, arguing as in Section 3 we get

$$P_\theta\{X_{n+1} \in H(\pi, d)\} = \alpha - \frac{1}{n\pi(\theta)}D_s\{I^{st}\xi_t(\theta, \alpha)\pi(\theta)\} + o(n^{-1}). \tag{5.4}$$

The right hand side of (5.4) equals $\alpha + o(n^{-1})$ if and only if

$$D_s\{I^{st}\xi_t(\theta, \alpha)\pi(\theta)\} = 0. \tag{5.5}$$

A prior $\pi(\cdot)$, satisfying (5.5) for every α, will ensure frequentist validity, up to $o(n^{-1})$, of the highest predictive density prediction of X_{n+1}.

EXAMPLE 4. We will consider the highest predictive density region in the bivariate normal model, with zero means for simplicity, and unknown variances σ_1^2, σ_2^2 and correlation coefficient ρ. Since the matching prior from (5.5) remains invariant under reparameterization (cf. last paragraph of this section), for computational simplicity, we choose to work with an orthogonal reparameterization given by $\theta_1 = \rho\sigma_2/\sigma_1$, $\theta_2 = \sigma_2^2(1 - \rho^2)$ and $\theta_3 = \sigma_1^2$. The inverse of the information matrix under this reparameterization is given by $I^{-1}(\theta) = \text{Diag}(\theta_2/\theta_3,\ 2\theta_2^2,\ 2\theta_3^2)$. It can be checked that

$$m(\theta, \alpha) = \frac{(1 - \alpha)}{2\pi\sqrt{(\theta_2\theta_3)}},\ \ \xi_1(\theta, \alpha) = 0,$$

$$\theta_2\xi_2(\theta, \alpha) = \theta_3\xi_3(\theta, \alpha) = \frac{(1 - \alpha)\log(1 - \alpha)}{2},$$

and that a class of priors exists as solutions to (5.5) for *all* α. A subclass of solutions to (5.5) is given by $\pi(\theta) \propto \theta_2^{-r}\theta_3^{r-2}$ for an arbitrary r, or in the original parameterization the subclass of solutions is given by $\pi(\sigma_1, \sigma_2, \rho) \propto (1-\rho^2)^{-r}\sigma_1^{2r-4}\sigma_2^{2-2r}$. Taking $r = 3/2$, one gets the prior $\pi(\sigma_1, \sigma_2, \rho) \propto (1-\rho^2)^{-3/2}\sigma_1^{-1}\sigma_2^{-1}$ as a solution to (5.5) for every α, which is Jeffreys' prior and is probability matching for θ_1 and $\rho\sigma_1/\sigma_2$.

EXAMPLE 5. Consider the multivariate scale model with a density

$$f(x;\theta) = (\theta_1 \dots \theta_p)^{-1} f^*(x^{(1)}/\theta_1, \dots, x^{(p)}/\theta_p),$$

where $x = (x^{(1)}, \dots, x^{(p)})'$ and $\theta_1, \dots, \theta_p > 0$. Let m_α be such that

$$\int f^*(u_1, \dots, u_p)du = \alpha,$$

where the integral is over $\{u : f^*(u_1, \dots, u_p) \geq m_\alpha\}$. Then by (5.1), (5.2),

$$m(\theta, \alpha) = m_\alpha/(\theta_1 \dots \theta_p), \quad \xi_j(\theta, \alpha) = \gamma_j(\alpha)/\theta_j,$$

with $\gamma_j(\alpha)$ free from θ. Also, $I^{st} = b^{st}\theta_s\theta_t$ for each s, t, the quantities b^{st} being free from θ. Hence it can be seen that Jeffreys' prior $\pi(\theta) \propto (\theta_1 \dots \theta_p)^{-1}$ satisfies (5.5) for every α.

Similarly, it can be seen that for the multivariate location model $f(x;\theta) = f^*(x^{(1)} - \theta_1, \dots, x^{(p)} - \theta_p)$, Jeffreys' prior, given by $\pi(\theta) = constant$, satisfies (5.5) for every α. It can also be checked that for the multivariate location-scale model with a density with different scale parameters

$$f(x;\theta) = (\theta_{p+1} \dots \theta_{2p})^{-1} f^*((x^{(1)} - \theta_1)/\theta_{p+1}, \dots, (x^{(p)} - \theta_p)/\theta_{2p}),$$

where $x = (x^{(1)}, \dots, x^{(p)})'$, $\theta = (\theta_1, \cdots, \theta_{2p})'$ and $\theta_{p+1}, \dots, \theta_{2p} > 0$, or a density with the same scale parameter

$$f(x;\theta) = \theta_{p+1}^{-p} f^*((x^{(1)} - \theta_1)/\theta_{p+1}, \dots, (x^{(p)} - \theta_p)/\theta_{p+1}),$$

where $\theta = (\theta_1, \dots, \theta_{p+1})'$ and $\theta_{p+1} > 0$, there exist priors that satisfy (5.5) for every α. In the first case, such a prior is given by $\pi(\theta) \propto (\theta_{p+1} \cdots \theta_{2p})^{-1}$, while in the second case it is given by $\pi(\theta) \propto \theta_{p+1}^{-1}$. In both cases, however, Jeffreys' priors are not solutions to (5.5).

Following Ghosh and Mukerjee (1993), under both the multivariate location and scale models, Jeffreys' prior also ensures approximate frequentist validity, up to $o(n^{-1})$, of highest posterior density regions for θ. However, in general, there is no guarantee that Jeffreys' prior will always satisfy (5.5) for every α. In fact, as Example 6 below demonstrates, even with scalar θ and scalar X_i, $i \geq 1$, it is possible that Jeffreys' prior does not satisfy (5.5) but another solution to (5.5), valid for every α, is available.

EXAMPLE 6. We revisit Example 1. By (5.1) and (5.2), writing $z^* = z_{(1-\alpha)/2}$,

$$m(\theta, \alpha) = \theta^{-1/2}\phi(z^*), \quad \xi_1(\theta, \alpha) = -\theta^{-1}z^*\phi(z^*).$$

Hence the unique prior, satisfying (5.5) for every α, is $\pi(\theta) \propto (2\theta + 1)/\theta$. This is different from Jeffreys' prior and, following Ghosh and Mukerjee (1993), also from the priors ensuring frequentist validity, up to $o(n^{-1})$, of the highest posterior density regions for θ.

We now indicate some situations where no prior satisfying (5.5) for every α is available. Consider the case of scalar θ. Since both $I(\theta)$ and $\pi(\theta)$ are positive for all θ, a solution to (5.5), valid for every α, is available if and only if $\xi_1(\theta, \alpha)$ is of the form

$$\xi_1(\theta, \alpha) = Q(\theta)R(\alpha), \tag{5.6}$$

where $Q(\theta)$ does not involve α, $R(\alpha)$ does not involve θ, and $Q(\theta)$ is positive for all θ. For the truncated exponential model $f(x; \theta) = k(\theta)\exp(-x/\theta)$, $0 < x < 1$, where $k(\theta) = 1/[\theta\{1 - \exp(-1/\theta)\}]$ and $\theta > 0$, a factorization as in (5.6) is not possible. On the other hand, for the bivariate normal model with zero means, unit variances and correlation coefficient θ $(-1 < \theta < 1)$ such a factorization is possible but one cannot have $Q(\theta)$ positive over the entire parameter space. Hence in these two situations no solution to (5.5), valid for every α, exists. For the latter model, however, the condition (5.6) together with positiveness of $Q(\theta)$ is met if the parameter space for the correlation coefficient θ is changed to (0,1) or (-1,0).

Finally, we return to the issue of invariance. In a manner similar to that described in Section 3, one can check that the same invariance property holds also with reference to (5.5). This property may be contrasted with the lack of such invariance under probability matching with highest posterior density regions for θ itself.

6. Conclusion. The problem of predicting the future value of a random variable is often viewed as similar to estimation of a suitable parametric function. Nonetheless, there are important differences between the two problems. One such difference is highlighted in this paper in the development of probability matching priors. For instance, while probability matching equations which provide frequentist justification based on a one-sided credible interval or a HPD set for a parametric function remain independent of the target coverage probability α, the corresponding equations for prediction usually depend on α [cf. (3.5) and (5.5)]. Despite this difficulty, there are examples where the probability matching approach to prediction problems leads to sensible noninformative priors (cf. Examples 2, 4 and 5). In particular, in Theorem 1, as in Welch and Peers (1963), we have established that in models depending on a scalar parameter, Jeffreys' prior is the only probability matching prior if one exists. This result lends further support to Jeffreys' prior as a default prior in the scalar case. Unfortunately, this result does not hold uniformly for all models such as a regular one-parameter exponential distribution (see Example 1).

Our overall assessment of the probability matching approach to prediction is that it promises to be a valuable tool for the development of sensible objective priors for Bayesian inference, and is worthy of further study.

Acknowledgments and Disclaimer. We thank the Editor, Associate Editor and referees for very constructive suggestions. This work was initiated when the first three authors met at the University of Nebraska–Lincoln in 1997.

This paper reports the results of research and analysis undertaken by the authors. Research results and conclusions expressed are those of the authors and have not been endorsed by Bureau of Labor Statistics and the Census Bureau.

REFERENCES

AITCHISON, J. and DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge Univ. Press.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1996). Prediction and asymptotics. *Bernoulli* **2** 319–340.

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.

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

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

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

DICICCIO, T. J. and STERN, S. E. (1994). Frequentist and Bayesian Bartlett correction of test statistics based on adjusted profile likelihoods. *J. Roy. Statist. Soc. Ser. B* **56** 397–408.

GEISSER, S. (1993). *Predictive Inference: An Introduction.* Chapman-Hall, New York.

GHOSH, J. K. and MUKERJEE, R. (1991). Characterization of priors under which Bayesian and frequentist Bartlett corrections are equivalent in the multiparameter case. *J. Multivariate Anal.* **38** 385–393.

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

GHOSH, J. K. and MUKERJEE, R. (1993). Frequentist validity of highest posterior density regions in multiparameter case. *Ann. Inst. Statist. Math.* **45** 293–302.

JEFFREYS, H. (1961). *Theory of Probability*. Oxford Univ. Press.

JOHNSON, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* **45** 851–864.

KOMAKI, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83** 299–314.

KUBOKI, H. (1998). Reference priors for prediction. *J. Statist. Plann. Inference* **69** 295–317.

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.

NICOLAOU, 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 sets and Bayesian probability points in the case of several parameters. *J. Roy. Statist. Soc. Ser. B* **27** 9–16.

ROUSSEAU, J. (1997). *Propriétés Asymptotiques des estimateurs de Bayes*. Ph.D. dissertation, Laboratoire de Statistiques, Théoriques et Appliquées, Paris VI.

SEVERINI, T. A. (1991). On the relationship between Bayesian and non-Bayesian interval estimates. *J. Roy. Statist. Soc. Ser. B* **53** 611–618.

SEVERINI, T. A. (1993). Bayesian interval estimates which are also confidence intervals. *J. Roy. Statist. Soc. Ser. B* **55** 533–540.

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

SUN, D. and YE, K. (1996). Frequentist validity of posterior quantiles for a two-parameter exponential family. *Biometrika* **83** 55–65.

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

SWEETING, T. J. (1999). On the construction of Bayes-confidence regions. *J. Roy. Statist. Soc. Ser. B* **61** 849–861.

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

VIDONI, P. A. (1998). A note on modified estimative prediction limits and distributions. *Biometrika* **85** 949–953.

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.

G. S. DATTA
DEPARTMENT OF STATISTICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602-1952
E-MAIL: gauri@stat.uga.edu

R. MUKERJEE
INDIAN INSTITUTE OF MANAGEMENT
JOKA, DIAMOND HARBOUR ROAD
P.O. BOX NO. 16757
ALIPORE POST OFFICE, CALCUTTA 700 027
INDIA
E-MAIL: rmuk1@hotmail.com

M. GHOSH
DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611-8545
E-MAIL: ghoshm@stat.ufl.edu

T. J. SWEETING
DEPARTMENT OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF SURREY
GUILDFORD, SURREY GU2 5XH
UNITED KINGDOM
E-MAIL: t.sweeting@eim.surrey.ac.uk