Ann. Inst. Statist. Math.
Vol. 55, No. 1, 1–6 (2003)

PROBABILITY MATCHING PRIORS FOR PREDICTING A DEPENDENT VARIABLE WITH APPLICATION TO REGRESSION MODELS

GAURI SANKAR DATTA[1] AND RAHUL MUKERJEE[2]

[1]*Department of Statistics, University of Georgia, Athens, GA 30602-1952, U.S.A.*
[2]*Indian Institute of Management, Post Box No. 16757, Calcutta 700 027, India*

(Received April 11, 2001; revised November 21, 2001)

Abstract. In a Bayesian setup, we consider the problem of predicting a dependent variable given an independent variable and past observations on the two variables. An asymptotic formula for the relevant posterior predictive density is worked out. Considering posterior quantiles and highest predictive density regions, we then characterize priors that ensure approximate frequentist validity of Bayesian prediction in the above setting. Application to regression models is also discussed.

Key words and phrases: Bayesian prediction, frequentist validity, highest predictive density region, noninformative prior, posterior quantile, regression, shrinkage argument.

1. Introduction

The problem of characterizing priors that ensure approximate frequentist validity of Bayesian credible sets has been of substantial recent interest; see, e.g., Ghosh and Mukerjee (1998) and Mukerjee and Reid (1999) for review and references. These priors, known as probability matching priors, are in a sense noninformative and yield accurate frequentist confidence sets; see Tibshirani (1989). In a related development, recently Datta *et al.* (2000) addressed the corresponding problem where interest lies in predicting a future observation rather than estimating a parameter. They characterized priors ensuring approximate frequentist validity of Bayesian predictive regions obtained via consideration of (a) posterior quantiles and (b) highest predictive density (HPD).

Notwithstanding the previous work in this area including that of Datta *et al.* (2000), an important issue in prediction yet remains to be addressed. This relates to the case where each observation involves a dependent variable and an independent variable, both possibly vector-valued. Quite commonly, then one has knowledge of both the variables in past observations and also of the independent variable in a new observation. Given these, the prediction problem involves the dependent variable in the new observation. The work of Datta *et al.* (2000) takes no cognizance of the distinction between such dependent and independent variables and hence their results cannot be applied to the present problem that can often arise in practice especially in regression settings.

We consider the above problem of predicting a dependent variable from a Bayesian viewpoint. Section 2 gives an asymptotic formula for the relevant posterior predictive density. Then Section 3 characterizes priors ensuring approximate frequentist validity of Bayesian prediction as considered here in the senses (a) and (b) above. While a shrinkage argument employed in Section 3 has been considered previously in other contexts, the

final matching conditions in Section 3 are new. Furthermore, they allow applications to the practically important regression models that cannot be handled by the existing results. These applications covering, in particular, the exponential regression model and several varieties of the normal regression model are discussed in Section 4.

2. Posterior predictive density

Consider an independent variable X and a dependent variable Y, both possibly vector-valued, having a joing density $f(x,y;\theta)$, where the parametric vector $\theta = (\theta_1,\ldots,\theta_p)'$ lies in an open subset of R^p. Let (X_i,Y_i), $i \geq 1$, be a sequence of independent and identically distributed observations on (X,Y). The first n of these are the past observations where both X and Y are known while the $(n+1)$-th one is a new observation where only X is known. We consider Bayesian prediction of Y_{n+1}, based on $d = \{(x_i,y_i), 1 \leq i \leq n\}$ and x_{n+1}, using a prior density $\pi(\cdot)$ which is positive and thrice continuously differentiable. Here x_i and y_i are the realized values of X_i and Y_i respectively. Along the line 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). The per observation Fisher information matrix $I \equiv I(\theta)$ is supposed to be positive definite for all θ. 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; cf. Bickel and Ghosh (1990).

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

$$a_{jr} = \{D_jD_rl(\theta)\}_{\theta=\hat{\theta}}, \quad a_{jrs} = \{D_jD_rD_sl(\theta)\}_{\theta=\hat{\theta}}, \quad c_{jr} = -a_{jr},$$
$$\pi_j(\theta) = D_j\pi(\theta), \quad f_j(x,y;\theta) = D_jf(x,y;\theta), \quad f_{jr}(x,y;\theta) = D_jD_rf(x,y;\theta).$$

The matrix $C = ((c_{jr}))$ is positive definite over S. Let $C^{-1} = ((c^{jr}))$. Then, following Datta *et al.* (2000) (see also Komaki (1996)) and using the summation convention, the posterior joint density of (X_{n+1},Y_{n+1}), given $d = \{(x_i,y_i), 1 \leq i \leq n\}$, under the prior $\pi(\cdot)$, is seen to be

$$\begin{aligned}(2.1)\quad &\tilde{\pi}(x_{n+1},y_{n+1} \mid d)\\ &= f(x_{n+1},y_{n+1};\hat{\theta}) + \frac{1}{2n}\{A^t(\pi)f_t(x_{n+1},y_{n+1};\hat{\theta}) + c^{jr}f_{jr}(x_{n+1},y_{n+1};\hat{\theta})\}\\ &\quad + o(n^{-1}),\end{aligned}$$

where

$$(2.2)\qquad A^t(\pi) = c^{st}\left\{c^{jr}a_{jrs} + \frac{2\pi_s(\hat{\theta})}{\pi(\hat{\theta})}\right\}.$$

We now proceed to obtain the posterior predictive density of Y_{n+1}, given d and x_{n+1}. With reference to the density $f(x,y;\theta)$, let $g(x;\theta)$ be the marginal density of X and $h(y \mid x;\theta)$ be the conditional density of Y given X. Then $f_t(x,y;\theta) = g_t(x;\theta)h(y \mid x;\theta) + g(x;\theta)h_t(y \mid x;\theta)$, where $g_t(x;\theta) = D_tg(x;\theta)$ and $h_t(y \mid x;\theta) = D_th(y \mid x;\theta)$. Hence, $\int_{-\infty}^{\infty} f_t(x,y;\theta)dy = g_t(x;\theta)$. Similarly, $\int_{-\infty}^{\infty} f_{jr}(x,y;\theta)dy = g_{jr}(x;\theta)$, where $g_{jr}(x;\theta) = D_jD_rg(x;\theta)$. Then by (2.1),

$$(2.3)\qquad \int_{-\infty}^{\infty} \tilde{\pi}(x_{n+1},y_{n+1} \mid d)dy_{n+1}$$

$$= g(x_{n+1};\hat{\theta}) + \frac{1}{2n}\{A^t(\pi)g_t(x_{n+1};\hat{\theta}) + c^{jr}g_{jr}(x_{n+1};\hat{\theta})\} + o(n^{-1}).$$

By (2.1) and(2.3), after some simplification, the posterior predictive density of Y_{n+1}, given d and x_{n+1}, under the prior $\pi(\cdot)$ is found to be

$$\begin{aligned}(2.4)\quad \pi^*(y_{n+1} \mid d, x_{n+1}) &= \tilde{\pi}(x_{n+1}, y_{n+1} \mid d) \Big/ \int_{-\infty}^{\infty} \tilde{\pi}(x_{n+1}, y_{n+1} \mid d)dy_{n+1} \\ &= h(y_{n+1} \mid x_{n+1};\hat{\theta}) + \frac{1}{2n}\{A^t(\pi)h_t(y_{n+1} \mid x_{n+1};\hat{\theta}) + c^{jr}b_{jr}(y_{n+1} \mid x_{n+1};\hat{\theta})\} \\ &\quad + o(n^{-1}),\end{aligned}$$

where

$$b_{jr}(y_{n+1} \mid x_{n+1};\hat{\theta}) = \{f_{jr}(x_{n+1}, y_{n+1};\hat{\theta}) - g_{jr}(x_{n+1};\hat{\theta})h(y_{n+1} \mid x_{n+1};\hat{\theta})\}/g(x_{n+1};\hat{\theta}).$$

3. Probability matching conditions

3.1 *Posterior quantiles*

First suppose the dependent variable Y is scalar valued so that posterior quantiles of Y_{n+1}, with reference to the predictive density (2.4) are well-defined. For $0 < \alpha < 1$, let $q(\theta, \alpha, x)$ be such that $\int_{q(\theta,\alpha,x)}^{\infty} h(y \mid x;\theta)dy = \alpha$. Then by (2.4), the $(1-\alpha)$-th posterior predictive quantile of Y_{n+1}, given d and x_{n+1}, is

$$(3.1)\qquad Q(\alpha, \pi) = q(\hat{\theta}, \alpha, x_{n+1}) + n^{-1}W(\pi),$$

where the explicit form of $W(\pi)$, which is at most of order $O(1)$ and can involve α, d and x_{n+1} in adition to $\pi(\cdot)$, will not be needed in the sequel.

We now characterize priors that ensure frequentist validity, with margin of error $o(n^{-1})$, of the posterior quantiles of Y_{n+1}. The shrinkage argument of Ghosh and Mukerjee (1993) helps in this regard; see Mukerjee and Reid (2000) for more details on why this argument works. 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. Let $P^{\bar{\pi}}\{\cdot \mid d, x_{n+1}\}$ denote the posterior probability measure under $\bar{\pi}(\cdot)$. Then by (3.1) and an approximation, analogous to (2.4), for posterior predictive density of Y_{n+1} given d and x_{n+1} under $\bar{\pi}(\cdot)$, one can check that

$$\begin{aligned}(3.2)\quad P^{\bar{\pi}}\{Y_{n+1} > Q(\alpha, \pi) \mid d, x_{n+1}\} &= \alpha + \frac{1}{2n}\{A^t(\bar{\pi}) - A^t(\pi)\}\int_{q(\hat{\theta},\alpha,x_{n+1})}^{\infty} h_t(y_{n+1} \mid x_{n+1};\hat{\theta})dy_{n+1} + o(n^{-1}).\end{aligned}$$

We next integrate by parts $E_\theta[P^{\bar{\pi}}\{Y_{n+1} > Q(\alpha, \pi) \mid (X_1, Y_1), \ldots, (X_n, Y_n), X_{n+1}\}]$, as computed up to $o(n^{-1})$ via (3.2) (recall the definition of d), with respect to $\bar{\pi}(\cdot)$ and finally allow $\bar{\pi}(\cdot)$ to converge weakly to the degenerate measure at θ. Using (2.2), these steps eventually yield

$$(3.3)\qquad P_\theta\{Y_{n+1} > Q(\alpha, \pi)\} = \alpha - \frac{1}{n\pi(\theta)}D_s\{I^{st}V_t(\theta, \alpha)\pi(\theta)\} + o(n^{-1}),$$

where $I^{-1} = ((I^{st}))$ is the inverse of the information matrix I, and

$$V_t(\theta, \alpha) = E_\theta \left\{ \int_{q(\theta,\alpha,X)}^{\infty} h_t(y \mid X; \theta) dy \right\}. \tag{3.4}$$

In view of (3.3), a prior $\pi(\cdot)$ ensures frequentist validity, up to $o(n^{-1})$, of the posterior quantiles of Y_{n+1} if and only if the condition

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

holds for every α.

3.2 *HPD regions*

HPD regions for Y_{n+1} are meaningful even when the dependent variable Y is possibly vector-valued. For $0 < \alpha < 1$, let $m(\theta, \alpha, x)$ be such that $\int h(y \mid x; \theta) dy = \alpha$, where the integral is over $H(\theta, \alpha, x) = \{y : h(y \mid x; \theta) \geq m(\theta, \alpha, x)\}$. Let $U_t(\theta, \alpha) = E_\theta\{\int h_t(y \mid X; \theta) dy\}$, the integral being over $H(\theta, \alpha, X)$. Then arguing as in Subsection 3.1, it can be shown that a prior $\pi(\cdot)$ ensures frequentist validity, up to $o(n^{-1})$, of HPD regions for Y_{n+1} if and only if the condition

$$D_s\{I^{st} U_t(\theta, \alpha)\pi(\theta)\} = 0 \tag{3.6}$$

holds for every α. The details are omitted here to save space.

4. Application to regression models

Example 1. With scalar Y and possibly vector-valued X, let

$$g(x; \theta) = g(x; \psi), \qquad h(y \mid x; \theta) = \exp\{-\lambda' w(x)\} h^*[y \exp\{-\lambda' w(x)\}], \tag{4.1}$$

where $\theta = (\lambda', \psi')'$. Here $h^*(\cdot)$ is a density on the real line and each of λ, ψ and $w(x)$ is possibly vector-valued. Note that the independent variable X enters into the conditional distribution of Y via a scale parameter $\exp\{\lambda' w(x)\}$. The exponential regression model (Cox and Reid (1987)) is covered by (4.1) if one takes $h^*(v) = \exp(-v)$ for $v > 0$ and $= 0$ otherwise. Similarly, (4.1) covers the normal regression model with known coefficient of variation if one takes $h^*(v) = \phi(v - k)$, where $\phi(\cdot)$ is the standard univariate normal density and $k^{-1}(> 0)$ is the known coefficient of variation in the conditional distribution of Y.

In view of (3.4), after some algebra it can be checked that the following hold under (4.1): (a) I does not involve λ; (b) $I = \text{diag}(M_1, M_2)$, where M_1 and M_2 correspond to λ and ψ respectively; (c) $q(\theta, \alpha, x) = q_\alpha \exp\{\lambda' w(x)\}$, where q_α is the $(1 - \alpha)$-th quantile of the density $h^*(\cdot)$; (d) $V_t(\theta, \alpha)$ does not involve λ for any t; (e) $V_t(\theta, \alpha) = 0$ whenever t corresponds to some element of ψ. Hence one can verify that any prior $\pi(\cdot)$ that does not involve λ will satisfy the matching condition (3.5) for posterior quantiles. Furthermore, with additional algebra, the same conclusion is seen to hold also with respect to the matching condition (3.6) arising via HPD regions.

Example 2. Continuing with scalar Y and possibly vector-valued X now let

$$g(x; \theta) = g(x; \psi), \qquad h(y \mid x; \theta) = \delta^{-1} h^*[\{y - \lambda' w(x)\}/\delta], \tag{4.2}$$

where $\theta = (\delta, \lambda', \psi')'$. As before, $h^*(\cdot)$ is a density on the real line and each of λ, ψ and $w(x)$ is possibly vector-valued. The independent variable X enters into the conditional distribution of Y via a location parameter $\lambda' w(x)$ and $\delta(> 0)$ is a scale parameter underlying this conditional distribution. In particular, if X and Y are jointly normal (with no supplementary information available about the underlying parameters) then (4.2) arises and δ represents the conditional standard deviation of Y.

By (3.4) after some algebra one can check that the following hold under (4.2): (a) $I = \text{diag}(\delta^{-2}M_1, M_2)$, where M_1 and M_2 correspond to $(\delta, \lambda')'$ and ψ respectively; (b) neither M_1 nor M_2 involves δ or λ; (c) $q(\theta, \alpha, x) = \delta q_\alpha + \lambda' w(x)$, where q_α is the $(1-\alpha)$-th quantile of the density $h^*(\cdot)$; (d) $V_t(\theta, \alpha)$ is of the form $V_t(\theta, \alpha) = \delta^{-1} G_t(\psi, \alpha)$ whenever t corresponds to δ or some element of λ, $G_t(\psi, \alpha)$ being free from δ or λ; (e) $V_t(\theta, \alpha) = 0$ whenever t corresponds to some element of ψ. Hence it may be verified that any prior $\pi(\cdot)$ of the form $\pi(\theta) = \kappa(\psi)/\delta$, where $\kappa(\psi)(> 0)$ is a smooth function involving ψ alone, will satisfy the matching condition (3.5) for posterior quantiles. One can also check that the same conclusion holds for the matching condition (3.6) pertaining to HPD regions.

Example 3. The last two examples may give the impression that if $f(x, y; \theta)$ is separable as

$$f(x, y; \theta) = g(x, \psi) h(y \mid x, \lambda), \tag{4.3}$$

where $\theta = (\lambda', \psi')'$, then the matching priors in the present context are determined only through the conditional model $h(y \mid x, \lambda)$. We now demonstrate that this is not the case in general. Consider a model of the form (4.3) where X, Y, ψ and λ are all scalars and $h(y \mid x, \lambda)$ is the simple exponential density with mean $\exp(\frac{1}{2}\lambda^2 + \lambda x)$. Write $\lambda = \theta_1, \psi = \theta_2$. Then

$$I_{11} = E_\psi\{(\lambda + X)^2\}, \quad I_{12} = 0, \quad V_1(\theta, \alpha) = G(\alpha) E_\psi(\lambda + X), \; V_2(\theta, \alpha) = 0,$$

where the constant $G(\alpha)$ depends only on α and not on λ or ψ. Hence it can be seen that a prior satisfies the matching condition (3.5) if and only if it is of the form

$$\pi(\theta) = \kappa(\psi) E_\psi\{(\lambda + X)^2\} / E_\psi(\lambda + X),$$

where $\kappa(\psi)(> 0)$ is a smooth function involving ψ alone. Clearly, the above form of $\pi(\theta)$ is influenced by the marginal distribution of X as well.

Example 4. In the previous examples, the marginal density of X and the conditional density of Y given X involved no common parameter. Consider now a situation where this is not the case. Let the joint distribution of X and Y be bivariate normal with both means $\mu(\in R^1)$, both standard deviations $\sigma(> 0)$, and correlation coefficient $\rho(|\rho| < 1)$. Writing $\mu = \theta_1$, $\sigma = \theta_2$ and $\rho = \theta_3$, then

$$\begin{aligned}
&I_{11} = 2/\{\sigma^2(1 + \rho)\}, \quad I_{12} = I_{13} = 0, \quad I_{22} = 4/\sigma^2, \; I_{23} = -2\rho/\{\sigma(1 - \rho^2)\}, \\
&I_{33} = (1 + \rho^2)/(1 - \rho^2)^2, \\
&V_1(\theta, \alpha) = \{(1 - \rho)/(1 + \rho)\}^{1/2} \phi(z_\alpha)/\sigma, \quad V_2(\theta, \alpha) = z_\alpha \phi(z_\alpha)/\sigma, \\
&V_3(\theta, \alpha) = -\rho z_\alpha \phi(z_\alpha)/(1 - \rho^2),
\end{aligned}$$

where z_α is the $(1 - \alpha)$-th quantile of a standard normal variate and, as before, $\phi(\cdot)$ is the standard normal density. Hence considering a natural class of priors of the form

$\pi(\theta) = \{\sigma^r(1-\rho^2)^s\}^{-1}$, where r and s are any real numbers, one can verify that the unique prior in this class satisfying the matching condition (3.5) is given by $r = -1$, $s = 1$. With additional algebra, the same conclusion is seen to hold for the matching condition (3.6). Thus, either via posterior quantiles or via predictive HPD regions, one gets the unique probability matching prior $\pi(\theta) = \sigma/(1-\rho^2)$ within the natural class mentioned above. Interestingly, in contrast with what was seen in Example 2, this prior is not inversely proportional to the conditional standard deviation of Y.

Acknowledgements

We thank the refrees for very constructive suggestions. This work was supported by National Science Foundation, USA, and Center for Management and Development Studies, Indian Institute of Management, Calcutta.

References

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.

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

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

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

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

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

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

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

Mukerjee, R. and Reid, N. (2000). On the Bayesian approach for frequentist computations, Brazilian *Journal of Probability and Statistics* (to appear).

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