Statistica Sinica **13**(2003), 1165-1178

PREDICTIVE DENSITIES: GENERAL ASYMPTOTIC RESULTS AND ADMISSIBILITY

Gang Wei and Ruhal Mukerjee

Hong Kong Baptist University and Indian Institute of Management

Abstract: For general regular parametric models, we compare predictive densities under the criterion of average Kullback-Leibler divergence. Asymptotic results are given via a Bayesian route without any assumption on curved exponentiality. We also address the issue of asymptotic admissibility of predictive densities and give a complete characterization when the underlying parameter is scalar-valued. Bayes predictive densities are considered in particular and the status of probability matching priors in this regard is examined. Finally, we indicate the consequences of working under more general α-divergences.

Key words and phrases: α-divergence, Bayes predictive density, estimative density, Jeffreys' prior, Kullback-Leibler divergence, probability matching prior.

1. Introduction

Predictive densities have been of considerable recent interest in the context of predicting a future observation from a parametric model on the basis of past ones. A predictive density may or may not belong to the parametric family under consideration. In the former case, it is referred to as an estimative density. Bayes and generalized Bayes predictive densities (Komaki (1996), Corcuera and Giummole (1999a), Datta, Mukerjee, Ghosh and Sweeting (2000)) are, on the other hand, notable examples of the latter case. In a pioneering work, Aitchison (1975) showed that Bayes predictive densities can dominate estimative densities with regard to closeness to the true density. This observation was further reinforced by Harris (1989) who introduced parametric bootstrap predictive densities. Vidoni (1995) gave a computationally simpler approximation to the proposal of Harris (1989). In recent years, significantly new ground was broken by Komaki (1996) and Corcuera and Giummole (1999a, b) for curved exponential models. Under the criterion of average Kullback-Leibler divergence, Komaki (1996) obtained asymptotic results on improving estimative densities. Corcuera and Giummole (1999b) extended these results to more general α-divergences. Further results on generalized Bayes predictive densities, associated with α-divergences, were reported by Corcuera and Giummole (1999a). More references on predictive densities are available in the last three papers; see also Komaki (2001) for interesting results on a shrinkage predictive distribution for multivariate normal observations. We refer to Barndorff-Nielsen and Cox (1996, Section 2) for a brief but very informative review of the developments in the general area of prediction.

The present article has two objectives. First, we aim to compare predictive densities for general regular parametric models, avoiding any assumption on curved exponentiality. This issue was left open by Corcuera and Giummole (1999a) in their concluding remarks. We work with a very general class of predictive densities and, without the curved exponentiality assumption, obtain an essentially complete subclass thereof in an asymptotic sense. In particular, for curved exponential models, this subclass includes the improved versions of estimative densities proposed in the literature. The above is done in Section 2 and the proofs, that follow via a transparent Bayesian route, are sketched in the appendix.

Our second objective is to make further comparisons within the essentially complete subclass mentioned above. This is done via consideration of asymptotic admissibility, an issue that has been hitherto unexplored in the context of predictive densities, even for curved exponential models. Borrowing ideas from the asymptotic theory of point estimation, rather than prediction, we give a complete characterization for such admissibility when the underlying model is indexed by a scalar parameter θ. This is done in Section 3. In Section 4, we give examples illustrating how the present general results can facilitate meaningful comparisons even for vector θ.

Throughout, we maintain an interest in Bayes predictive densities, especially those arising from improper priors that have been advocated from other considerations. In particular, our examples examine how probability matching priors for prediction (Datta et al (2000)) behave in the present setup. Quite counterintuitively it is seen that, even for scalar θ, Jeffreys' prior may be inadmissible in spite of having the probability matching property. It is also seen that for vector θ the present approach can help in narrowing down the class of probability matching priors.

For ease in presentation, in Sections 2-4, we work under the criterion of average Kullback-Leibler divergence. As noted in Komaki (1996), this is a natural measure of divergence having a concrete interpretation, for instance, in coding theory. Indeed, all our results have their counterparts for more general α-divergences. This point has been indicated briefly in Section 5.

2. Asymptotic Results on Predictive Densities

2.1. Preliminaries

Let $X_1, \ldots, X_n$ be possibly vector-valued observations that are independent and identically distributed with common density $f(x;\theta)$, where θ is an unknown parameter. On the basis of $Z = (X_1, X_2, \ldots, X_n)$, we intend to predict a future independent observation from the model $f(x;\theta)$. Standard regularity conditions are assumed. In particular the parameter space for $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$ is supposed to be R^p or an open set therein, and the support $\mathcal{X}$ of $f(x,\theta)$ is supposed to be free from θ. Furthermore, the per observation Fisher information matrix $I \equiv I(\theta)$ is supposed to be positive definite for all θ, and we assume the existence of a valid Edgeworth expansion for the distribution of $\sqrt{n}(\hat{\theta} - \theta)$, where $\hat{\theta}$ is the maximum likelihood estimator of θ based on Z. All stochastic expansions in this paper are over a set with $P_\theta-$probability $1 + o(n^{-2})$, uniformly on compact sets of θ (cf. Bickel and Ghosh (1990)).

For $1 \le i,\ j \le p$, let $D_i \equiv \partial/\partial\theta_i$, $f_i(x;\theta) = D_i f(x;\theta)$, $f_{ij}(x;\theta) = D_i D_j f(x;\theta)$, $m_i(x;\theta) = f_i(x;\theta)/f(x;\theta)$ and $m_{ij}(x;\theta) = f_{ij}(x;\theta)/f(x;\theta)$. On the basis of Z, we consider a very general class $\mathcal{F}$ of predictive densities such that each $f^*(x;Z) \in \mathcal{F}$ admits a stochastic expansion of the form

$$f^*(x;Z) = f(x;\hat{\theta}) + n^{-1} e_1(x;Z) + M_n, \tag{2.1}$$

where M_n is at most of order $O(n^{-2})$, $e_1(x;Z)$ is at most of order $O(1)$ and satisfies

$$\int_{\mathcal{X}} e_1(x;Z)dx = 0. \tag{2.2}$$

For each $f^*(x;Z) \in \mathcal{F}$, it is assumed that the integrals

$$\begin{aligned} W(Z) &= \int_{\mathcal{X}} \left\{ e_1(x;Z)/f(x;\hat{\theta}) \right\}^2 f(x;\hat{\theta})dx, \\ V_i(Z) &= \int_{\mathcal{X}} e_1(x;Z) m_i(x;\hat{\theta})dx, \quad V_{ij}(Z) = \int_{\mathcal{X}} e_1(x;Z) m_{ij}(x;\hat{\theta})dx, \end{aligned} \tag{2.3}$$

exist and satisfy

$$W(Z) = B(\theta) + o(1),\ \ V_i(Z) = G_i(\theta) + o(1),\ \ V_{ij}(Z) = G_{ij}(\theta) + o(1), \tag{2.4}$$

where $B(\cdot)$, $G_i(\cdot)$ $(1 \le i \le p)$ and $G_{ij}(\cdot)$ $(1 \le i,\ j \le p)$ are some smooth functions with functional form free from n.

The condition (2.2) is natural since both $f^*(x;\ Z)$ and $f(x;\ \hat{\theta})$ are densities. The existence of the integrals in (2.3) is necessary in order that stochastic expansions for the divergence measures considered here be available. The conditions in (2.4) are needed for computing the average of any such divergence up to the desired order of approximation. The class $\mathcal{F}$ is very rich and contains virtually all predictive densities that are amenable to an asymptotic analysis. This includes estimative densities associated with the maximum likelihood estimator or a perturbed version thereof, parametric bootstrap predictive densities of Harris (1989), as well as Bayes and generalized Bayes predictive densities.

In this and the next two sections, the predictive densities $f^*(x;Z)\in\mathcal{F}$ are evaluated on the basis of their average Kullback-Leibler divergence from the true density, as given by

$$H(\theta)=\mathrm{E}_\theta\left[\int_{\mathcal{X}} f(x;\theta)\log\{f(x;\theta)/f^*(x;Z)\}\,dx\right]. \tag{2.5}$$

The smaller is H as a function of θ, the better is the predictive fit of $f^*(x;Z)$. Let $H_0(\theta)$ be $H(\theta)$ when $f^*(x;Z)$ is taken as the simple estimative density $f(x;\hat\theta)$.

Some more notation will help. Let $m^{(1)}(x;\theta)$ and $m^{(2)}(x;\theta)$ be vectors, of orders $p\times 1$ and $p^2\times 1$, having elements $m_i(x;\theta)(1\le i\le p)$ and $m_{ij}(x;\theta)(1\le i,\ j\le p)$, respectively. Define $m(x;\theta)=\left[m^{(1)}(x;\theta)^T,\ m^{(2)}(x;\theta)^T\right]^T$ and

$$A(\theta)=\begin{bmatrix}A_{11}(\theta) & A_{12}(\theta)\\ A_{21}(\theta) & A_{22}(\theta)\end{bmatrix}=\int_{\mathcal{X}}\left\{m(x;\theta)m(x;\theta)^T\right\}f(x;\theta)dx, \tag{2.6}$$

where $A_{11}(\theta)$ is $p\times p$ and so on. Note that $A_{11}(\theta)=I$ where I is the per observation Fisher information matrix. Write $I^{-1}=((I^{ij}))$ and let $\Psi\equiv\Psi(\theta)$ be a $p^2\times 1$ vector with elements $I^{ij}(1\le i,\ j\le p)$. Also, let $L_{jrs}\equiv\mathrm{E}_\theta\{D_jD_rD_s\log f(X_1;\theta)\}$.

2.2. Asymptotic results

Consider $f^*(x;Z)\in\mathcal{F}$ having a stochastic expansion as in (2.1), the associated $W(Z)$, $V_i(Z)$, $V_{ij}(Z)$ and $B(\theta)$, $G_i(\theta)$, $G_{ij}(\theta)$ being as in (2.3) and (2.4), respectively. Let $V^{(1)}(Z)$, $V^{(2)}(Z)$, $G^{(1)}(\theta)$ and $G^{(2)}(\theta)$ be vectors, of orders $p\times 1$, $p^2\times 1$, $p\times 1$ and $p^2\times 1$, having elements $V_i(Z)$ $(1\le i\le p)$, $V_{ij}(Z)$ $(1\le i,\ j\le p)$, $G_i(\theta)$ $(1\le i\le p)$ and $G_{ij}(\theta)$ $(1\le i,\ j\le p)$, respectively. Then with reference to such an $f^*(x;Z)$, the following theorem holds. In part (a) of the theorem, and elsewhere, we follow the summation convention with sums ranging from 1 to p over repeated subscripts or superscripts.

Theorem 1.

$$\begin{aligned}(a)\ \ H(\theta)-H_0(\theta) &= n^{-2}\Delta(\theta)+o(n^{-2}),\quad \textit{where}\\ \Delta(\theta) &= D_j\left\{I^{ij}G_i(\theta)\right\}-\tfrac{1}{2}I^{is}I^{jr}L_{jrs}G_i(\theta)\\ &\qquad +\tfrac{1}{2}\left\{B(\theta)-\Psi^TG^{(2)}(\theta)\right\}.\end{aligned} \tag{2.7}$$

$$\begin{aligned}(b)\ \ B(\theta)-\Psi^TG^{(2)}(\theta) &\ge q(\theta)^TIq(\theta)-\tfrac{1}{4}\Psi^TA_{22}(\theta)\Psi,\\ \textit{where}\quad q(\theta) &= I^{-1}\left\{G^{(1)}(\theta)-\tfrac{1}{2}A_{12}(\theta)\Psi\right\}.\end{aligned}$$

Theorem 1, proved in the appendix, helps to identify an essentially complete subclass of $\mathcal{F}$ under the criterion considered here. Let $\mathcal{F}_0$ be a subclass of $\mathcal{F}$ consisting of predictive densities for which

$$e_1(x;Z)=g_i(Z)f_i(x;\hat\theta)+g_{ij}(Z)f_{ij}(x;\hat\theta), \tag{2.8}$$

where the $g_i(Z)$ and $g_{ij}(Z)$ are at most of order $O(1)$ and satisfy

$$g_i(Z) = d_i(\theta) + o(1), \qquad g_{ij}(Z) = \frac{1}{2}I^{ij} + o(1), \tag{2.9}$$

the $d_i(\cdot)$ being some smooth functions with functional form free from n. By (2.3), (2.4) and (2.6), for such a predictive density in $\mathcal{F}_0$, we have

$$\begin{aligned} B(\theta) &= d(\theta)^T A_{11}(\theta) d(\theta) + \Psi^T A_{21}(\theta) d(\theta) + \frac{1}{4}\Psi^T A_{22}(\theta)\Psi, \\ G^{(i)}(\theta) &= A_{i1}(\theta) d(\theta) + \frac{1}{2}A_{i2}(\theta)\Psi, \qquad (i = 1,\ 2), \end{aligned} \tag{2.10}$$

where $d(\theta)$ is a $p \times 1$ vector with elements $d_i(\theta)(1 \le i \le p)$.

Theorem 2. *Given any predictive density in $\mathcal{F}$, there exists one in $\mathcal{F}_0$ such that $\Delta(\theta)$ for the latter does not exceed that for the former, for any θ.*

In view of Theorem 2, proved in the appendix, hereafter we consider only the class $\mathcal{F}_0$. For any member of $\mathcal{F}_0$, with the associated vector $d(\theta)$ defined via (2.9), it follows from (2.7) and (2.10) that $\Delta(\theta) = \Delta_0(\theta) + \Delta_1(\theta)$, where $\Delta_1(\theta)$ is the same for all members of $\mathcal{F}_0$, and

$$\Delta_0(\theta) = D_i\left\{d_i(\theta)\right\} - \frac{1}{2}I^{jr}L_{jrs}d_s(\theta) + \frac{1}{2}d(\theta)^T I d(\theta), \tag{2.11}$$

recalling that $A_{11}(\theta) = I$. By (2.9) the class $\mathcal{F}_0$ excludes estimative densities but includes the predictive densities constructed by Komaki (1996) for curved exponential models. Following Komaki (1996) or Datta et al (2000), it also includes the Bayes predictive density corresponding to any smooth and positive prior $\pi(\cdot)$, the associated $d_i(\cdot)$ being given by

$$d_i(\theta) = I^{ij}\frac{\pi_j(\theta)}{\pi(\theta)} + \frac{1}{2}I^{is}I^{jr}L_{jrs}, \qquad (1 \le i \le p), \tag{2.12}$$

where $\pi_j(\theta) = D_j\pi(\theta)$. Substituting (2.12) in (2.11), the part of $\Delta_0(\theta)$(or equivalently of $\Delta(\theta)$) that involves $\pi(\cdot)$ is

$$\Delta_\pi(\theta) = D_i\left\{I^{ij}\frac{\pi_j(\theta)}{\pi(\theta)}\right\} + \frac{1}{2}I^{ij}\left\{\frac{\pi_i(\theta)}{\pi(\theta)}\right\}\left\{\frac{\pi_j(\theta)}{\pi(\theta)}\right\}. \tag{2.13}$$

Since $D_i I^{ij} = I^{ir}I^{js}(L_{i,rs} + L_{irs})$ (see Ghosh and Mukerjee (1991)), where

$$L_{i,rs} = \mathrm{E}_\theta\left[\{D_i \log f(X_1;\ \theta)\}\{D_r D_s \log f(X_1;\ \theta)\}\right], \tag{2.14}$$

use of Bartlett conditions shows that (2.13) is in agreement with equation (17) of Corcuera and Giummole (1999a). They considered curved exponential models, but left open the issue of extension beyond such models. Incidentally, apart from settling this open issue, (2.13) looks somewhat simpler as well.

Before concluding this section, we indicate a connection between Theorem 1 and the findings in Corcuera and Giummole (2000). Let $\mathcal{F}^*$ be a subclass of $\mathcal{F}$ consisting of predictive densities for which $G^{(1)}(\theta) = 0$, identically in θ. Since $A_{11}(\theta) = I$, it is clear from (2.10) that any member of $\mathcal{F}_0$ with

$$d(\theta) = -\frac{1}{2} I^{-1} A_{12}(\theta) \Psi \tag{2.15}$$

belongs to $\mathcal{F}^*$. Furthermore, by Theorem 1(b) and (2.10), it is not hard to see that such a member of $\mathcal{F}_0$ minimizes $\Delta(\theta)$ over $\mathcal{F}^*$. This is in agreement with Proposition 3 of Corcuera and Giummole (2000) when the latter is interpreted in our setup with reference to Kullback-Leibler divergence. A member of $\mathcal{F}_0$ satisfying (2.15), however, is not guaranteed to minimize $\Delta(\theta)$, or equivalently $\Delta_0(\theta)$, over the class $\mathcal{F}_0$. This will be evident from Example 1 in the next section.

3. Admissibility Results for Scalar Parameter

We now turn to the problem of comparing predictive densities in $\mathcal{F}_0$ on the basis of $\Delta(\theta)$ or equivalently $\Delta_0(\theta)$. While minimization of $\Delta_0(\theta)$ uniformly in θ is not possible, the problem can be addressed via admissibility considerations. A predictive density in $\mathcal{F}_0$ will be called *second-order admissible* if there is no other member of $\mathcal{F}_0$ such that $\Delta_0(\theta)$ for the latter is less than or equal to that of the former for all θ, with strict inequality for some θ. A prior will be called second-order admissible if the corresponding Bayes predictive density is so in $\mathcal{F}_0$. Clearly, by Theorem 2, second-order admissibility in $\mathcal{F}_0$ is equivalent to that in the larger class $\mathcal{F}$.

We now consider the case of scalar θ and present a complete characterization of second-order admissibility in the above sense. Let (θ_-, θ_+) be the parameter space for θ, where $\theta_- < \theta_+$ and $\theta_- = -\infty$ or $\theta_+ = +\infty$ are possible. The $p \times 1$ vector $d(\theta)$, associated with any member of $\mathcal{F}_0$, now reduces to the scalar $d_1(\theta)$. Similarly, $I (\equiv I(\theta))$ becomes a scalar and we write $L (\equiv L(\theta))$ for L_{111}. Formula (2.11) for $\Delta_0(\theta)$ then becomes

$$\Delta_0(\theta) = \frac{1}{2} I \{d_1(\theta)\}^2 + d_1'(\theta) - \frac{1}{2} I^{-1} L d_1(\theta), \tag{3.1}$$

where the prime denotes differentiation with respect to θ.

Consider now two members $\bar{f}(x; Z)$ and $\tilde{f}(x; Z)$ of $\mathcal{F}_0$. Let $\bar{d}_1(\theta)$ and $\bar{\Delta}_0(\theta)$ be the expressions for $d_1(\theta)$ and $\Delta_0(\theta)$ associated with $\bar{f}(x; Z)$. Similarly, define $\tilde{d}_1(\theta)$ and $\tilde{\Delta}_0(\theta)$ with reference to $\tilde{f}(x; Z)$. Then by (3.1), $\tilde{\Delta}_0(\theta) - \bar{\Delta}_0(\theta) = (1/2) I \left[\{\lambda(\theta)\}^2 + 2I^{-1}\lambda'(\theta) + 2\lambda(\theta) b(\theta)\right]$, where $\lambda(\theta) = \tilde{d}_1(\theta) - \bar{d}_1(\theta)$ and $b(\theta) = \bar{d}_1(\theta) - (1/2) I^{-2} L$. Hence, following a result of Ghosh and Sinha (1981) in the context of point estimation, $\bar{f}(x;Z)$ is second-order admissible in $\mathcal{F}_0$ if and only if

$$\begin{aligned} &\int_{\theta_-}^{\theta_0} I(\theta)\exp\left\{\int_{\theta}^{\theta_0} b(u)I(u)du\right\}d\theta=\infty \quad \text{and} \\ &\int_{\theta_0}^{\theta_+} I(\theta)\exp\left\{\int_{\theta}^{\theta_0} b(u)I(u)du\right\}d\theta=\infty, \end{aligned} \tag{3.2}$$

for some $\theta_0 \in (\theta_-, \theta_+)$.

Further reduction of (3.2) is possible for Bayes predictive densities. By (2.12), for scalar θ and with reference to the predictive density corresponding to a prior $\pi(\cdot)$, we get

$$d_1(\theta) = I^{-1}\left\{\pi'(\theta)/\pi(\theta)\right\} + \frac{1}{2}I^{-2}L, \tag{3.3}$$

so that $b(\theta)$ becomes $I^{-1}\{\pi'(\theta)/\pi(\theta)\}$. Therefore, by (3.2), $\pi(\cdot)$ is second-order admissible if and only if

$$\int_{\theta_-}^{\theta_0} \frac{I(\theta)}{\pi(\theta)}d\theta = \infty \ \text{ and } \ \int_{\theta_0}^{\theta_+} \frac{I(\theta)}{\pi(\theta)}d\theta = \infty, \tag{3.4}$$

for some $\theta_0 \in (\theta_-,\ \theta_+)$. The conditions in (3.4) substantially simplify the study of second-order admissibility of priors under the criterion of average Kullback-Leibler divergence. In particular, Jeffreys' prior $\pi(\theta) \propto \{I(\theta)\}^{1/2}$ is second-order admissible if and only if

$$\int_{\theta_-}^{\theta_0} \{I(\theta)\}^{\frac{1}{2}}d\theta = \infty, \ \text{and} \ \int_{\theta_0}^{\theta_+} \{I(\theta)\}^{\frac{1}{2}}d\theta = \infty, \tag{3.5}$$

for some $\theta_0 \in (\theta_-, \theta_+)$.

Example 1. For the one-parameter location or scale models, it is immediate from (3.5) that Jeffreys' prior is second-order admissible when one works with the usual parameter spaces for these models. In particular, for the simple exponential model with scale parameter θ, one can readily check from (3.1) and (3.3) that the Bayes predictive density given by Jeffreys' prior has a smaller $\Delta_0(\theta)$, for every θ, than any member of $\mathcal{F}_0$ with $d(\theta)$ as in (2.15).

Example 2. Even beyond the one-parameter location or scale models, Jeffreys' prior can be second-order admissible. Consider the univariate normal model with both mean and variance equal to θ, where $\theta > 0$. Then $I(\theta) = (2\theta+1)/(2\theta^2)$ and, by (3.5), this conclusion about Jeffreys' prior follows.

Example 3. We now give an example where Jeffreys' prior is not second-order admissible but another improper prior, satisfying this admissibility property, is available. Let $f(x;\theta) = \theta(1+\theta)(x+\theta)^{-2}$, $0 < x < 1$, where $\theta > 0$. Then $I(\theta) = (1/3)\{\theta(1+\theta)\}^{-2}$ so that the second condition in (3.5) cannot hold for any $\theta_0 > 0$. Therefore, Jeffreys' prior is second-order inadmissible. On the other hand, by (3.4), the prior $\pi^*(\theta) \propto \{\theta(1+\theta)^2\}^{-1}$ is second-order admissible. Here $L = (1+2\theta)/\{\theta(1+\theta)\}^3$ and, from (3.1) and (3.3), one can check that $\pi^*(\theta)$ dominates Jeffreys' prior for every θ. At this stage, one may be concerned that $\pi^*(\theta)$ does not share the well-known invariance property of Jeffreys' prior. In the present context, however, this poses no serious problem since a Bayes predictive density based on any prior is invariant of the parameterization. Thus the prior $\pi^*(\theta)$ under the $\theta-$parameterization produces the same Bayes predictive density as a transformed version of $\pi^*(\theta)$ would under a one-to-one reparameterization.

Remark 1. As noted in Datta et al (2000), Jeffreys' prior is uniquely probability matching in Examples 1 and 3, in the sense of ensuring approximate frequentist validity of posterior quantiles of a future observation, whereas in Example 2 it is not so. On the other hand, our findings show that Jeffreys' prior is second-order admissible in Examples 1 and 2 but is not in Example 3. The last example is particularly revealing since it demonstrates that, even for scalar θ, Jeffreys' prior may be second-order inadmissible in spite of enjoying the probability matching property. On the whole, these examples suggest the absence of any general relationship between probability matching and second order admissibility properties of a prior.

Example 4. We now give an example that allows an exact analysis and enables us to examine how close the present asymptotic results are to the exact ones. With reference to the univariate normal model with mean θ $(-\infty < \theta < \infty)$ and variance unity, consider the class of improper priors of the form $\pi(\theta) = \exp(w\theta)$, where w is any real number. The average Kullback-Leibler divergence (2.5) can be obtained exactly for the resulting Bayes predictive densities and it is not hard to check that (2.5) is minimized, with respect to w, when $w = 0$, which corresponds to Jeffreys' prior (cf. Example 1). On the other hand, here $I = 1$, $L = 0$ and, from (3.1) and (3.3), it is evident that $\Delta_0(\theta)$ is also minimized over w at $w = 0$. Furthermore the conditions in (3.4) are met if and only if $w = 0$. Therefore, the asymptotic results are in perfect agreement with the exact one. In this example, if instead one considers a normal prior on θ, with specified mean and variance, then again (3.4) is met, i.e., such a normal prior becomes second order admissible. This too is in agreement with what one expects under an exact analysis with a proper prior.

4. Examples for Vector Parameter

General admissibility results, like those in Ghosh and Sinha (1981), are unknown for vector θ even in the context of point estimation. Hence analogues of conditions (3.2) and (3.4) are no longer available. However, formulae

(2.11)−(2.13) allow us to make meaningful comparisons in reasonable subclasses of $\mathcal{F}_0$, considering even models that do not belong to the curved exponential family. Illustrative examples follow.

Example 5. Consider the symmetric location-scale model $f(x;\theta) = \theta_2^{-1} f_0((x - \theta_1)/\theta_2), -\infty < x < \infty$, where $-\infty < \theta_1 < \infty$, $\theta_2 > 0$ and $f_0(\cdot)$ is a density on the real line that is symmetric about zero. Then

$$\begin{gathered} I_{11} = a_{11}/\theta_2^2, \;\; I_{22} = a_{22}/\theta_2^2, \;\; L_{112} = \xi_{112}/\theta_2^3, \\ L_{222} = \xi_{222}/\theta_2^3, \;\; I_{12} = L_{111} = L_{122} = 0, \end{gathered} \tag{4.1}$$

where a_{11}, a_{22}, ξ_{112} and ξ_{222} are constants that do not involve θ. Let $\mathcal{F}_1$ be a subclass of $\mathcal{F}_0$ consisting of predictive densities for which $d_1(\theta) = 0$ and $d_2(\theta)$ [$= d_2(\theta_2)$, say] is free from θ_1. By (2.12) and (4.1), $\mathcal{F}_1$ includes Bayes predictive densities corresponding to the natural class of priors $\pi(\cdot)$ that do not involve θ_1; furthermore, for any such prior,

$$d_2(\theta_2) = a_{22}^{-1}\theta_2\{\theta_2\rho(\theta_2) + k\}, \tag{4.2}$$

where $k = (1/2)(a_{11}^{-1}\xi_{112} + a_{22}^{-1}\xi_{222})$ and $\rho(\theta_2) = \pi_2(\theta)/\pi(\theta)$. For any member of $\mathcal{F}_1$, by (2.11) and (4.1),

$$\Delta_0(\theta) = \frac{1}{2} a_{22}\theta_2^{-2}\{d_2(\theta_2)\}^2 + d_2'(\theta_2) - k\theta_2^{-1} d_2(\theta_2), \tag{4.3}$$

where the prime denotes differentiation with respect to θ_2. Hence following Ghosh and Sinha (1981) as in the previous section, a predictive density in $\mathcal{F}_1$ is second-order admissible in $\mathcal{F}_1$ if and only if the corresponding $d_2(\theta_2)$ satisfies the conditions

$$\begin{gathered} \int_0^{\theta_0} \frac{1}{\theta_2^2} \exp\left\{\int_{\theta_2}^{\theta_0} \frac{b(u)a_{22}}{u^2} du\right\} d\theta_2 = \infty \quad \text{and} \\ \int_{\theta_0}^{\infty} \frac{1}{\theta_2^2} \exp\left\{\int_{\theta_2}^{\theta_0} \frac{b(u)a_{22}}{u^2} du\right\} d\theta_2 = \infty, \end{gathered} \tag{4.4}$$

for some $\theta_0 > 0$, where $b(u) = d_2(u) - a_{22}^{-1}ku$.

Consider now a prior of the form $\pi(\theta) \propto \theta_2^{-w}$, where w is any real number. By (4.2), for such a prior $d_2(\theta_2) = a_{22}^{-1}(k - w)\theta_2$. Then by (4.3), $\Delta_0(\theta) = a_{22}^{-1}(k-w)\{(1/2)(k-w)-k+1\}$, which is minimized over w at $w = 1$. From (4.4), it is easily seen that the resulting prior $\pi(\theta) \propto \theta_2^{-1}$ is second-order admissible in the entire class $\mathcal{F}_1$. As noted in Datta et al (2000), this is also the unique prior having the probability matching property mentioned in the last section.

Example 6. This example demonstrates how, for vector θ, the criterion of average Kullback-Leibler divergence can help in narrowing down the class of probability matching priors. Consider the bivariate Cauchy model

$$f(x;\theta)=\left[2\pi\theta_1\theta_2\left\{1+\left(\frac{x^{(1)}}{\theta_1}\right)^2+\left(\frac{x^{(2)}-\theta_3x^{(1)}}{\theta_2}\right)^2\right\}^{\frac{3}{2}}\right]^{-1},\ -\infty<x^{(1)},x^{(2)}<\infty,$$

where $x=(x^{(1)},x^{(2)})^T$, $\theta=(\theta_1,\ \theta_2,\ \theta_3)^T$, $\theta_1>0$, $\theta_2>0$ and $-\infty<\theta_3<\infty$. It can be seen here that

$$I^{11}=\frac{5}{3}\theta_1^2,\ I^{22}=\frac{5}{3}\theta_2^2,\ I^{33}=\frac{5}{3}\frac{\theta_2^2}{\theta_1^2},\ I^{12}=\frac{5}{6}\theta_1\theta_2,\ I^{13}=I^{23}=0. \tag{4.5}$$

Let $\mathcal{F}_1$ be a subclass of $\mathcal{F}_0$ consisting of Bayes predictive densities associated with priors of the form

$$\pi(\theta)\propto(\theta_1^{w_1}\theta_2^{w_2})^{-1}, \tag{4.6}$$

where w_1 and w_2 are any real numbers. By (2.13) and (4.5), we have $\Delta_\pi(\theta)=(5/24)\left\{3(w_1+w_2-2)^2+(w_1-w_2)^2-12\right\}$, which reaches minimum when $w_1=w_2=1$.

Following Datta et al (2000), a prior of the form (4.6) ensures approximate frequentist validity of highest posterior predictive density regions if and only if $w_1+w_2=2$. The optimal prior obtained in the last paragraph satisfies this condition. Thus the criterion of Kullback-Leibler divergence helps in this example to identify a unique prior among all those satisfying the matching condition.

5. Results under α-divergence

Before concluding, we briefly indicate the consequences of working with more general α-divergence measures that cover the Kullback-Leibler divergence as the special case $\alpha=-1$, and correspond to the Hellinger or chi-squared distances when $\alpha=0$ or 3, respectively. Following Corcuera and Giummole (1999a), the α-divergence of $f^*(x;\ Z)\in\mathcal{F}$ from the true density is defined as

$$\int_{\mathcal{X}}f(x;\ \theta)K_\alpha\left(\frac{f^*(x;\ Z)}{f(x;\ \theta)}\right)dx,$$

$$K_\alpha(t)=\begin{cases}t\log t, & \text{if } \alpha=1,\\ -\log t, & \text{if } \alpha=-1,\\ \dfrac{4}{1-\alpha^2}(1-t^{(1+\alpha)/2}), & \text{if } \alpha\neq 1,\ -1.\end{cases} \tag{5.1}$$

For $1 \leq i,\ j \leq p$, let $m^*_{ij}(x;\theta) = m_i(x;\theta)m_j(x;\theta) - I_{ij}(\theta)$, where $I_{ij}(\theta)$ is the (i,j)th element of the Fisher information matrix I. In order to study the predictive densities in $\mathcal{F}$ under the criterion of average α-divergence, we need a mild assumption that for every member of $\mathcal{F}$, as given by (2.1), the integrals $V^*_{ij}(Z) = \int_{\mathcal{X}} e_1(x;Z)m^*_{ij}(x;\hat{\theta})dx$ exist and satisfy $V^*_{ij}(Z) = G^*_{ij}(\theta) + o(1)$, the $G^*_{ij}(\cdot)$ $(1 \leq i,\ j \leq p)$ being some smooth functions with functional form free from n.

Then, under the criterion of average α-divergence, one can check that a counterpart of Theorem 1(a) holds with $\Delta(\theta)$ in the statement of Theorem 1(a) replaced by

$$\Delta^{(\alpha)}(\theta) = \Delta(\theta) + \frac{1}{4}(1+\alpha)I^{ij}G^*_{ij}(\theta). \tag{5.2}$$

Similarly, a counterpart of Theorem 2 holds with $\Delta(\theta)$ and $\mathcal{F}_0$ there replaced by $\Delta^{(\alpha)}(\theta)$ and $\mathcal{F}_0^{(\alpha)}$, respectively. Here $\mathcal{F}_0^{(\alpha)}$ is a subclass of $\mathcal{F}$ consisting of predictive densities for which

$$e_1(x;Z) = g_i(Z)f_i(x;\hat{\theta}) + g_{ij}(Z)f_{ij}(x;\hat{\theta}) + g^*_{ij}(Z)m^*_{ij}(x;\hat{\theta})f(x;\hat{\theta}), \tag{5.3}$$

where the $g_i(Z),\ g_{ij}(Z)$ and $g^*_{ij}(Z)$ are at most of order $O(1)$ and satisfy $g^*_{ij}(Z) = -(1/4)(1+\alpha)I^{ij} + o(1)$ $(1 \leq i,\ j \leq p)$, in addition to the conditions in (2.9). These results follow along the lines of the appendix with slightly heavier algebra, but no assumption of curved exponentiality is needed. In proving the counterpart of Theorem 2, one also requires a counterpart of Theorem 1(b) which is omitted here.

For any member of $\mathcal{F}_0^{(\alpha)}$, from (5.2) and (5.3), it can be seen that $\Delta^{(\alpha)}(\theta) = \Delta_0(\theta) + \Delta_1^{(\alpha)}(\theta)$, where $\Delta_1^{(\alpha)}(\theta)$ is the same for all members of $\mathcal{F}_0^{(\alpha)}$ and $\Delta_0(\theta)$, free from α, is as given by (2.11). The class $\mathcal{F}_0^{(\alpha)}$ contains the predictive densities constructed by Corcuera and Giummole (1999b) for curved exponential models. Because of the last term in (5.3), this class does not in general include Bayes predictive densities unless $\alpha = -1$. However, it includes the generalized Bayes predictive density corresponding to any smooth and positive prior. For curved exponential models, this follows from Corcuera and Giummole (1999a). Even otherwise, this can be shown if one starts from their equation (4), considers an expansion for the posterior density of $h = \sqrt{n}(\theta - \hat{\theta})$ (Ghosh and Mukerjee (1991)) and incorporates a normalizing factor. For any such generalized Bayes predictive density, (2.12) and (2.13) continue to hold.

Thus the key equations (2.11)$-$(2.13) remain unaltered as we pass to α-divergence from the Kullback-Leibler divergence. The only point is that now (2.12) and (2.13) correspond to generalized Bayes rather than Bayes predictive densities. Consequently, the admissibility results presented earlier readily extend themselves to the case of general α-divergence. Thus for scalar θ, (3.2) gives a characterization for second-order admissibility in $\mathcal{F}_0^{(\alpha)}$ and hence in $\mathcal{F}$. Similarly (3.4) becomes a necessary and sufficient condition for the second-order admissibility of the generalized Bayes predictive density corresponding to a prior $\pi(\cdot)$. Since (3.4) does not involve α, this entails a robustness property for priors in the sense that a given prior either generates second-order admissible generalized Bayes predictive densities for all α or it fails to do so for any α, depending on whether (3.4) holds or not.

Corcuera and Giummole (1999b) considered an even wider class of divergence measures of the form

$$\int_{\mathcal{X}} f(x;\ \theta) F\left(\frac{f^*(x;\ Z)}{f(x;\ \theta)}\right) dx,$$

where $F(\cdot)$ is a smooth strictly convex function that vanishes at 1. As in their paper, without loss of generality, let $F''(1) = 1$. Along the lines of the appendix, one can again check that, under any such divergence, a counterpart of Theorem 1(a) holds with $\Delta(\theta)$ there replaced by $\Delta^{(\alpha)}(\theta)$, where $\alpha = 2F'''(1){+}3$ and $\Delta^{(\alpha)}(\theta)$ is as in (5.2). This is in agreement with the findings of Corcuera and Giummole (1999b) who worked under curved exponentiality. Consequently, even with these divergences, one can develop counterparts of Theorem 2 and the second order admissibility results just as indicated above for $\alpha-$divergences. The details are omitted here.

Acknowledgement

We thank the Co-Editor, an associate editor and the referees for their very constructive suggestions. This work was supported by a grant from the Center for Management and Development Studies, Indian Institute of Management, Calcutta and the grant HKBU 2059/00P of RGC, Hong Kong.

Appendix. Proofs

Proof of Theorem 1. (a) By (2.5),

$$H(\theta) - H_0(\theta) = -\mathrm{E}_\theta \left[\int_{\mathcal{X}} f(x;\theta) \log\{f^*(x;Z)/f(x;\hat{\theta})\} dx\right]. \qquad \text{(A.1)}$$

Let the remainder term M_n in (2.1) be denoted by $n^{-2}e_2(x;\ Z)$, where $e_2(x;\ Z)$ is at most of order $O(1)$. Then by (2.1) and (2.2), $\int_{\mathcal{X}} e_2(x;\ Z)dx = 0$, and $f^*(x;Z)/f(x;\hat{\theta}) = 1 + n^{-1}U_1 + n^{-2}U_2$, where $U_i = e_i(x;Z)/f(x;\hat{\theta})$ $(i = 1, 2)$. Write $h = (h_1, h_2, \ldots, h_p)^T = \sqrt{n}(\theta - \hat{\theta})$ and note that $f(x;\theta) = f(x;\hat{\theta})\{1 + n^{-1/2}h_i m_i(x;\hat{\theta}) + (1/2)n^{-1}h_i h_j m_{ij}(x;\hat{\theta})\} + o(n^{-1})$. Using the above together with (2.2) and (2.3),

$$\int_{\mathcal{X}} f(x;\theta) \log\left\{\frac{f^*(x;Z)}{f(x;\hat{\theta})}\right\} dx$$

$$= n^{-3/2}h^TV^{(1)}(Z) + \frac{1}{2}n^{-2}\left\{h_ih_jV_{ij}(Z) - W(Z)\right\} + o(n^{-2}). \qquad \text{(A.2)}$$

Clearly, by (2.4),

$$\mathrm{E}_\theta\{h_ih_jV_{ij}(Z) - W(Z)\} = \Psi^TG^{(2)}(\theta) - B(\theta) + o(1). \qquad \text{(A.3)}$$

We next find the expectation of the first term on the right-hand side of (A.2) via a Bayesian route. To that effect, define $l(\theta) = n^{-1}\sum_{i=1}^n \log f(X_i;\theta)$, $c_{jr} = -\{D_jD_rl(\theta)\}_{\theta=\hat{\theta}}$, $a_{jrs} = \{D_jD_rD_sl(\theta)\}_{\theta=\hat{\theta}}$ and $((c^{jr})) = ((c_{jr}))^{-1}$. Consider now the posterior density of h, given Z, under an auxiliary prior $\bar{\pi}(\cdot)$ satisfying the conditions of Bickel and Ghosh(1990, p.1078). Let $\mathrm{E}^{\bar{\pi}}\{\cdot|Z\}$ denote expectation with respect to this posterior density. Using an expansion for this posterior density (see, for example, Ghosh and Mukerjee (1991)), we get

$$\mathrm{E}^{\bar{\pi}}\{h^TV^{(1)}(Z)|Z\} = n^{-\frac{1}{2}}\left\{\frac{\bar{\pi}_j(\hat{\theta})}{\bar{\pi}(\hat{\theta})}c^{ij} + \frac{1}{2}c^{is}c^{jr}a_{jrs}\right\}V_i(Z) + o(n^{-1/2}),$$

where $\bar{\pi}_j(\theta) = D_j\bar{\pi}(\theta)$. If one now calculates $\mathrm{E}_\theta\left[\mathrm{E}^{\bar{\pi}}\{h^TV^{(1)}(Z)|Z\}\right]$ and then employs a shrinkage argument, popular in Bayesian asymptotics (Mukerjee and Dey (1993)), then one gets

$$\mathrm{E}_\theta\left\{h^TV^{(1)}(Z)\right\} = n^{-1/2}\left[\frac{1}{2}I^{is}I^{jr}L_{jrs}G_i(\theta) - D_j\{I^{ij}G_i(\theta)\}\right] + o(n^{-1/2}), \quad \text{(A.4)}$$

recalling (2.4). Part (a) of the theorem is evident from (A.1)−(A.4).
(b) Let $\hat{q} = q(\hat{\theta})$ and $\hat{\Psi} = \Psi(\hat{\theta})$. By (2.3) and (2.6),

$$\int_{\mathcal{X}}\left[\{e_1(x;Z)/f(x;\hat{\theta})\} - \hat{q}^Tm^{(1)}(x;\hat{\theta}) - \frac{1}{2}\hat{\Psi}^Tm^{(2)}(x;\hat{\theta})\right]^2 f(x;\hat{\theta})dx$$

$$= W(Z) - 2\hat{q}^TV^{(1)}(Z) - \hat{\Psi}^TV^{(2)}(Z) + \hat{q}^TA_{11}(\hat{\theta})\hat{q} + \hat{q}^TA_{12}(\hat{\theta})\hat{\Psi} + \frac{1}{4}\hat{\Psi}^TA_{22}(\hat{\theta})\hat{\Psi}. \text{(A.5)}$$

By (2.4), the definition of $q(\theta)$ and the fact that $A_{11}(\theta) = I$, one can check that the right-hand side of (A.5) is $Q(\theta) + o(1)$ where $Q(\theta) = B(\theta) - \Psi^TG^{(2)}(\theta) - q(\theta)^TIq(\theta) + (1/4)\Psi^TA_{22}(\theta)\Psi$. Since the left-hand side of (A.5) is nonnegative, part (b) of Theorem 1 follows.

Proof of Theorem 2. Let $\hat{I} = I(\hat{\theta}) = ((\hat{I}^{ij}))$. By (2.10), equality holds in Theorem 1(b) for every member of $\mathcal{F}_0$. Hence by (2.7), it is enough to show that given any member of $\mathcal{F}$, there exists one of $\mathcal{F}_0$ such that the two have the same $G^{(1)}(\theta)$. In view of (2.8)−(2.10), for any $f^*(x;Z) \in \mathcal{F}$ with the associated $V_i(Z)$ given by (2.3), this follows considering $f_0(x;Z) \in \mathcal{F}_0$, where $f_0(x;Z) = f(x;\hat{\theta}) + n^{-1}\{g_i(Z)f_i(x;\hat{\theta}) + (1/2)\hat{I}^{ij}f_{ij}(x;\hat{\theta})\}$, with $(g_1(Z), g_2(Z), \ldots, g_p(Z))^T = \hat{I}^{-1}\{V^{(1)}(Z) - (1/2)A_{12}(\hat{\theta})\hat{\Psi}, V^{(1)}(Z)\}$ being defined as in the beginning of subsection 2.2.

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62**, 547-554.

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.

Corcuera, J. M. and Giummole, F. (1999a). A generalized Bayes rule for prediction. *Scand. J. Statist.* **26**, 265-279.

Corcuera, J. M. and Giummole, F. (1999b). On the relationship between α-connections and the asymptotic properties of predictive distributions. *Bernoulli* **5**, 163-176.

Corcuera, J. M. and Giummole, F. (2000). First order optimal predictive densities. In *Applications of Differential Geometry to Econometrics* (Edited by P. Marriott and M. Salmon), 214-229, Cambridge University Press, Cambridge.

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. (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 Sinha, B. K. (1981). A necessary and sufficient condition for second order admissibility with applications to Berkson's bioassay problem. *Ann. Statist.* **9**, 1334-1338.

Harris, I. R. (1989). Predictive fit for natural exponential families. *Biometrika* **76**, 675-684.

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83**, 299-313.

Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika* **88**, 859-864.

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.

Vidoni, P. (1995). A simple predictive density based on the p^* formula. *Biometrika* **82**, 855-863.

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China.

E-mail: gwei@hkbu.edu.hk

Indian Institute of Management, Post Box No. 16757, Calcutta 700 027, India.

E-mail: rmuk1@hotmail.com

(Received December 2002; accepted March 2003)