Ann. Inst. Statist. Math.
Vol. 45, No. 3, 567–578 (1993)

INFERENTIAL DISTRIBUTIONS FOR NON-BAYESIAN PREDICTIVE FIT

HISATAKA KUBOKI

Department of Communications and Systems, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182, Japan

(Received June 28, 1991; revised April 7, 1992)

Abstract. This article proposes a non-Bayesian procedure for constructing inferential distributions which can be used for producing predictive distributions. The concepts of bootstrap and of predictive likelihood are employed for developing the method. A result is obtained for exponential families, and the Bayesian prediction based on Jeffreys' prior is newly justified.

Key words and phrases: Bootstrap, estimative fit, exponential family, inferential distribution, Jeffreys' prior, predictive distribution, predictive fit, predictive likelihood.

1. Introduction

Let $f(y)$ be the distribution of a future observation y. Suppose that we can use n independent observations $x^n = (x_1, \ldots, x_n)$ from the distribution $g(x)$. We desire some probability statement about y using the observations x^n. Here we call any distribution employed for this purpose a predictive distribution for y.

Assume that y and x^n are independent but that x^n provides information on y through the same indexing parameter. When parametric families $\{f(y \mid \theta) : \theta \in \Theta\}$ and $\{g(x \mid \theta) : \theta \in \Theta\}$, where $\Theta \subseteq \boldsymbol{R}^k$, are given, a predictive distribution $\hat{f}(y \mid x^n)$ is obtained by the method of estimative fit

$$\hat{f}(y \mid x^n) = f(y \mid \hat{\theta}_n), \tag{1.1}$$

using the maximum likelihood estimate $\hat{\theta}_n = \hat{\theta}_n(x^n)$ of θ, or by the method of predictive fit

$$\hat{f}(y \mid x^n) = \int f(y \mid \theta) p(\theta \mid x^n) d\theta, \tag{1.2}$$

using a standardized weighing function $p(\theta \mid x^n)$ on θ based on the observations x^n or using a posterior distribution $p(\theta \mid x^n)$ of θ given x^n if a prior distribution $p(\theta)$ for θ is available. An 'inferential' distribution for θ is any standardized weighing

function on θ based on x^n, and is employed for producing a predictive distribution by (1.2) (Akaike, 1978).

Since the estimative fit implicitly assumes that θ is known to be $\hat{\theta}_n$, it can lead to serious underestimation of the dispersion of y. In fact, if we evaluate the badness of a predictive distribution $\hat{f}(y \mid x^n)$ by the expected neg-entropy

$$
\begin{aligned}
(1.3) \qquad & E_\theta[I\{f(\cdot \mid \theta), \hat{f}(\cdot \mid x^n)\}] \\
& = \int I\{f(\cdot \mid \theta), \hat{f}(\cdot \mid x^n)\} g(x^n \mid \theta) dx^n \\
& = \int \left[\int f(y \mid \theta) \log \left\{ \frac{f(y \mid \theta)}{\hat{f}(y \mid x^n)} \right\} dy \right] g(x^n \mid \theta) dx^n,
\end{aligned}
$$

then the estimative fit (1.1) often gives higher values of (1.3) for all θ than the Bayesian predictive fit (1.2) with an inferential distribution based on a vague prior on θ (Aitchison (1975), Murray (1977), Ng (1980)). This fact demonstrates that the device of using inferential distributions is effective in producing predictive distributions. The problem then is to choose an inferential distribution. Although an inferential distribution can be formally constructed by way of a prior distribution $p(\theta)$ of θ, this procedure leads to the still debated question of specifying $p(\theta)$. Thus, we are interested on how to construct inferential distributions without assuming any prior knowledge about θ.

Recently, two approaches were introduced. The first is the bootstrap method (Harris (1989)). Let $f(y \mid \hat{\theta}_n)$ be an estimative fit and $p(\cdot \mid \theta)$ the sampling distribution of $\hat{\theta}_n$. Evaluating $p(\cdot \mid \theta)$ at the estimated value $\hat{\theta}_n$ of θ, we have an inferential distribution $p^H(\cdot \mid x^n) = p(\cdot \mid \hat{\theta}_n)$. This in turn gives a predictive distribution

$$
\hat{f}^H(y \mid x^n) = \int f(y \mid \tau) p^H(\tau \mid x^n) d\tau.
$$

Here τ is a generic symbol for possible values of $\hat{\theta}_n(x^n)$. The second method by El-Sayyad *et al.* (1989) is applicable when a sufficient reduction t_n of x^n exists. Let $f(y \mid \hat{\theta}_{n-1})$ be an estimative fit based on $n-1$ observations x^{n-1}. Then the conditional distribution of $\hat{\theta}_{n-1}$ given t_n can be used as an inferential distribution because it is independent of θ, and therefore it gives a predictive distribution

$$
\hat{f}^E(y \mid x^n) = \int f(y \mid \tau) p^E(\tau \mid t_n) d\tau,
$$

where $p^E(\cdot \mid t_n)$ denotes the conditional distribution of $\hat{\theta}_{n-1}$ given t_n.

The purpose of the present paper is to develop a non-Bayesian procedure for obtaining inferential distributions. Section 2 describes the method. Here the concepts of bootstrap and of predictive likelihood are employed for the construction of inferential distributions. Section 3 deals with the case where $g(x \mid \theta)$ belongs to an exponential family. Here the Bayesian predictive fit based on Jeffreys' ((1961), Section 3.10) prior is looked at from the view presented in Section 2. Section 4 illustrates our method with gamma model.

2. A non-Bayesian procedure for constructing inferential distributions

Let $z^m = (z_1, \ldots, z_m)$ be a random sample of size m from $g(x \mid \theta)$. We assume that it is unobservable. Consider an estimative fit $f(y \mid \hat{\vartheta}_m)$, where $\hat{\vartheta}_m = \hat{\vartheta}_m(z^m)$ is the maximum likelihood estimate of θ based on z^m. Although the value of $\hat{\vartheta}_m$ is unobserved, the distribution $p(\cdot \mid \theta)$ of $\hat{\vartheta}_m$ can be specified except for the true value of θ. Thus adjusting uncertainty in $\hat{\vartheta}_m$ by integrating with respect to this distribution evaluated at the estimated value $\hat{\theta}_n$ of θ, we have a predictive distribution

$$\hat{f}(y \mid x^n) = \int f(y \mid \tau) p(\tau \mid \hat{\theta}_n) d\tau.$$

Here τ is a generic symbol for possible values of the unobserved $\hat{\vartheta}_m(z^m)$. If we take $m = n$, this is identical to the bootstrap predictive distribution $\hat{f}^H(y \mid x^n)$ of Harris (1989). If $\hat{\vartheta}_m$ is consistent, then $\hat{f}(y \mid x^n)$ works like the estimative fit (1.1) when m is sufficiently large.

This observation suggests the use of functions which assess the relative credibility of the possible outcomes of the unobserved $\hat{\vartheta}_m$ in the light of the observed x^n as inferential distributions. The fundamental point in this view is that we are dealing with functions which express evidence of z^m taking into account x^n. In the case stated above, the distribution of z^m evaluated at $\theta = \hat{\theta}_n$ was used. On the other hand, since all information about z^m is contained in the joint distribution of (x^n, z^m), a function obtained by elimination of θ from it will be an alternative choice for our purpose. Such a function is called a predictive likelihood of z^m given x^n. Of course, different ways of eliminating θ produce different predictive likelihoods (Bjørnstad (1990)). Let $\mathrm{plik}(z^m \mid x^n)$ denote a predictive likelihood for z^m given x^n. We now consider the problem of making some predictive statement about $\hat{\vartheta}_m$ using $\mathrm{plik}(z^m \mid x^n)$.

Suppose for the time being that we can use a function $\pi(z^m)$ with which the knowledge about the occurrence of z^m is a priori assessed. By analogy with the role of the likelihood function in Bayes' formula, we can view the predictive likelihood $\mathrm{plik}(z^m \mid x^n)$ as the function through which the observation x^n modifies our knowledge about the values of z^m. Then the evidence of each z^m is expressed as

$$\pi(z^m)\mathrm{plik}(z^m \mid x^n) \tag{2.1}$$

using the observation x^n. Let S_τ be the set of all unobserved samples z^m satisfying $\hat{\vartheta}_m(z^m) = \tau$ for a particular value τ of $\hat{\vartheta}_m$. Then the a posteriori plausibility of τ is obtained by accumulating (2.1) with respect to $z^m \in S_\tau$.

Here, it should be noted that each value of z^m is not necessarily uniform in its possibility of occurrence. Thus, it will be inadequate to take always $\pi(z^m) = c$. (In fact, see Section 4.) However, we now know the probability law of z^m. Thus, we can use this knowledge for the choice of $\pi(z^m)$. If the value of the unobserved sample is z^m, then the probability with which the value z^m will be observed can be estimated by the parametric maximum likelihood estimate

$$\pi(z^m) = \prod_{i=1}^{m} g\{z_i \mid \hat{\vartheta}_m(z^m)\}. \tag{2.2}$$

This procedure is the parametric bootstrap (Efron (1982), Section 5.2). The same idea is also used by Harris (1989) as mentioned in Section 1.

From the arguments above, we define the cumulative function $\Phi_m(\tau \mid x^n)$ of (2.1) with (2.2) by

$$\Phi_m(\tau \mid x^n) = \int_{\{z^m : \hat{\vartheta}_m(z^m) \le \tau\}} \left[\prod_{i=1}^{m} g\{z_i \mid \hat{\vartheta}_m(z^m)\}\right] \mathrm{plik}(z^m \mid x^n) dz^m. \tag{2.3}$$

In discrete case, the right-hand side is replaced by

$$\sum_{z^m \in \{z^m : \hat{\vartheta}_m(z^m) \le \tau\}} \left[\prod_{i=1}^{m} g\{z_i \mid \hat{\vartheta}_m(z^m)\}\right] \mathrm{plik}(z^m \mid x^n). \tag{2.4}$$

Let $\phi_m(\tau \mid x^n)$ denote the density of $\Phi_m(\tau \mid x^n)$ with respect to the Lebesgue measure on $\boldsymbol{R}^k$ or the counting measure on $\boldsymbol{Z}^k$, where $\boldsymbol{Z}$ denotes the set of integers. Then, we can regard $\phi_m(\hat{\vartheta}_m \mid x^n)$ as the function which orders the possible outcomes of the unobserved $\hat{\vartheta}_m$ in the light of x^n. Normalizing it to be a probability distribution, we obtain an inferential distribution

$$p^{B_m}(\tau \mid x^n) = \gamma_m(x^n)\phi_m(\tau \mid x^n),$$

where $\gamma_m(x^n)$ is a normalizing constant. Thus, from an estimative fit $f(y \mid \hat{\vartheta}_m)$, we can produce a predictive distribution $\hat{f}^{B_m}(y \mid x^n)$, adjusting uncertainty in $\hat{\vartheta}_m$ by integrating with respect to $p^{B_m}(\cdot \mid x^n)$:

$$\hat{f}^{B_m}(y \mid x^n) = \int f(y \mid \tau) p^{B_m}(\tau \mid x^n) d\tau.$$

When applying this procedure, we must determine the size m of the unobserved sampling. The parameter θ can be regarded as summarizing the infinite unobserved $(z_1, z_2, \ldots)$. This suggests the use of asymptotic approximations for $p^{B_m}(\tau \mid x^n)$. To illustrate this idea consider a normal model with

$$f(x \mid \theta) = g(x \mid \theta) = (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(x-\theta)^2\right\}, \qquad -\infty < \theta < \infty.$$

Let us compute $p^{B_m}(\tau \mid x^n)$ by using $\phi_m(\tau \mid x^n)$ defined by (3.1) of the next section. Then we obtain

$$p^{B_m}(\tau \mid x^n) = (2\pi\sigma_{m,n}^2)^{-1/2} \exp\left\{-\frac{1}{2}(\tau - \hat{\theta}_n)^2/\sigma_{m,n}^2\right\}$$

and

$$\hat{f}^{B_m}(y \mid x^n) = \{2\pi(1+\sigma_{m,n}^2)\}^{-1/2} \exp\left\{-\frac{1}{2}(y - \hat{\theta}_n)^2/(1+\sigma_{m,n}^2)\right\},$$

where $\sigma_{m,n}^2 = (m+n)/(mn)$. Since $\sigma_{m,n}^2 \to \sigma_n^2 = 1/n$ as $m \to \infty$,

$$p^{B_m}(\tau \mid x^n) \to p^{B}(\tau \mid x^n) = (2\pi\sigma_n^2)^{-1/2} \exp\left\{-\frac{1}{2}(\tau - \hat{\theta}_n)^2/\sigma_n^2\right\}$$

as $m \to \infty$. Let $\hat{f}^B(y \mid x^n)$ denote the predictive distribution corresponding to $p^B(\tau \mid x^n)$. Then the former predictive distribution is inferior to the latter because

$$E_\theta[I\{f(\cdot \mid \theta), \hat{f}^B(\cdot \mid x^n)\}] < E_\theta[I\{f(\cdot \mid \theta), \hat{f}^{B_m}(\cdot \mid x^n)\}]$$

for all $\theta \in \Theta$; this inequality follows immediately from

$$E_\theta[I\{f(\cdot \mid \theta), \hat{f}^{B_m}(\cdot \mid x^n)\}] = \frac{1}{2}\left\{\log(1+\sigma^2_{m,n}) + \frac{1+1/n}{1+\sigma^2_{m,n}} - 1\right\}$$

by differentiating it with respect to $\sigma^2_{m,n}$.

This result supports the idea of approximating $p^{B_m}(\tau \mid x^n)$ asymptotically in m. Let $p^B(\tau \mid x^n)$ denote such an approximation. Then for large m, $p^B(\hat{\vartheta}_m \mid x^n)$ expresses the relative credibility of the unobserved $\hat{\vartheta}_m$. If $\hat{\vartheta}_m$ is a consistent estimate of θ, then replacing $\hat{\vartheta}_m$ by θ, we can view $p^B(\theta \mid x^n)$ as the function which assesses the plausibility of the possible true values of θ in the light of x^n. Hence if there exists a function $p(\theta)$ such that

$$p^B(\theta \mid x^n) \propto p(\theta)\mathrm{lik}(\theta \mid x^n),$$

then $p^B(\theta \mid x^n)$ can be regarded as incorporating prior knowledge $p(\theta)$, and so using $p^B(\theta \mid x^n)$ will mean that (1.2) will be the Bayesian predictive distribution for y. Here $\mathrm{lik}(\theta \mid x^n)$ denotes the likelihood function for θ given x^n:

$$\mathrm{lik}(\theta \mid x^n) = \prod_{i=1}^n g(x_i \mid \theta).$$

3. A sampling justification of the Bayesian predictive fit based on Jeffreys' prior

Suppose that the $x_1, \ldots, x_n$ are independent and identically distributed with exponential family density with a canonical parametrization of the form

$$g(x \mid \theta) = a(\theta)b(x)\exp\{\theta' T(x)\}, \qquad \theta \in \Theta \subseteq \boldsymbol{R}^k,$$

relative to either Lebesgue measure on $\boldsymbol{R}^k$ or counting measure on $\boldsymbol{Z}^k$. This family is assumed to be regular. Then $t_n = T(x_1) + \cdots + T(x_n)$ is the minimal sufficient reduction of x^n. The statistics t_n has an exponential family density of the form

$$g^{n*}(t \mid \theta) = a(\theta)^n b^{n*}(t)\exp(\theta' t),$$

for some function $b^{n*}(t)$. Let $h(x^n \mid t_n)$ denote the conditional distribution of x^n given t_n.

Now set $\mu(\theta) = E_\theta(T)$ and $\Sigma(\theta) = E_\theta[\{T - \mu(\theta)\}\{T - \mu(\theta)\}']$. Let $\kappa(\theta)$ denote the cumulant generating function of $T(x)$: that is, $\kappa(\theta) = -\log a(\theta)$, and let $i(\theta)$ denote the Fisher information that x contains about the parameter θ. The maximum likelihood estimate $\hat{\vartheta}_m$ based on the unobserved z^m is a function of the

minimal sufficient reduction $s_m = T(z_1) + \cdots + T(z_m)$ of z^m: it is the unique solution of

$$s_m = m\mu(\theta) = m\frac{\partial\kappa(\theta)}{\partial\theta}.$$

Hence, we have only to assess the plausibility of s_m in the light of x^n, because

$$\pi(z^m) = \prod_{i=1}^{m} g\{z_i \mid \hat{\vartheta}_m(s_m)\} = h(z^m \mid s_m)g^{m*}\{s_m \mid \hat{\vartheta}_m(s_m)\}.$$

Furthermore, note that

$$\frac{\partial^2\kappa(\theta)}{\partial\theta\partial\theta'} = \Sigma(\theta) = i(\theta).$$

Here we apply the concept of predictive likelihood by Lauritzen (1974) and Hinkley (1979) to this problem. Hinkley ((1979), Definition 1) specifies the predictive likelihood for s_m given t_n as

$$\mathrm{lik}^*(s_m \mid t_n) = \frac{b^{m*}(s_m)b^{n*}(t_n)}{b^{(n+m)*}(t_n + s_m)}.$$

This is the conditional distribution of t_n given $t_n + s_m$. The predictive likelihood for s_m given x^n is then defined by

$$\mathrm{plik}(s_m \mid x^n) = h(x^n \mid t_n)\mathrm{lik}^*(s_m \mid t_n).$$

For this, see Bjørnstad (1990).

First, consider the continuous case. Since

$$\int_{\{z^m:\hat{\vartheta}_m(s_m)\le\tau\}} h(z^m \mid s_m)\,dz^m = \int_{\{s_m:\hat{\vartheta}_m(s_m)\le\tau\}} 1\,ds_m,$$

the cumulative function (2.3) becomes

$$\begin{aligned}
&\Phi_m(\tau \mid x^n) \\
&\quad = \int_{\{z^m:\hat{\vartheta}_m(s_m)\le\tau\}} h(z^m \mid s_m)h(x^n \mid t_n)g^{m*}\{s_m \mid \hat{\vartheta}_m(s_m)\}\mathrm{lik}^*(s_m \mid t_n)dz^m \\
&\quad = h(x^n \mid t_n)\int_{\{s_m:\hat{\vartheta}_m(s_m)\le\tau\}} g^{m*}\{s_m \mid \hat{\vartheta}_m(s_m)\}\mathrm{lik}^*(s_m \mid t_n)ds_m \\
&\quad = m^k h(x^n \mid t_n)\int_{\hat{\vartheta}_m\le\tau} g^{m*}\{m\mu(\hat{\vartheta}_m) \mid \hat{\vartheta}_m\}\mathrm{lik}^*\{m\mu(\hat{\vartheta}_m) \mid t_n\}|i(\hat{\vartheta}_m)|d\hat{\vartheta}_m,
\end{aligned}$$

where $s_m = m\mu(\hat{\vartheta}_m)$. (See Pitman ((1979), Appendix) and Kuboki (1984) for the operation employed in the integrals above.) Hence the density $\phi_m(\tau \mid x^n)$ of $\Phi_m(\tau \mid x^n)$ is given by

$$\phi_m(\tau \mid x^n) = m^k h(x^n \mid t_n)g^{m*}\{m\mu(\tau) \mid \tau\}\mathrm{lik}^*\{m\mu(\tau) \mid t_n\}|i(\tau)|. \tag{3.1}$$

When m is sufficiently large, it follows from the saddle-point expansion of $g^{m*}(s_m \mid \theta)$ (Barndorff-Nielsen and Cox (1979)) that

$$g^{m*}\{m\mu(\tau) \mid \tau\} = (2\pi m)^{-k/2}|\Sigma(\tau)|^{-1/2}\{1 + O(m^{-1})\} \tag{3.2}$$

uniformly in s_m, provided $\tau = \hat{\vartheta}_m(s_m)$ belongs to a given, but arbitrary, compact subset K of Θ. In addition, the following consistency property of $\mathrm{lik}^*(s_m \mid t_n)$ is discussed by Hinkley (1979) and Mathiasen (1979):

$$\mathrm{lik}^*\{m\mu(\tau) \mid t_n\} = g^{n*}(t_n \mid \tau)\{1 + O(m^{-1})\} \tag{3.3}$$

uniformly in s_m such that $\tau = \hat{\vartheta}_m(s_m) \in K$, and in t_n on every bounded subset of the support of g^{n*}. Thus, under some regularity conditions which ensure the validity of the saddle-point expansion and the consistency, we can give an approximation to the right-hand side of (3.1) by

$$\left(\frac{m}{2\pi}\right)^{k/2} h(x^n \mid t_n) g^{n*}(t_n \mid \tau)|i(\tau)|^{1/2}. \tag{3.4}$$

Next, let us see that when the family is discrete we can also regard (3.4) as an approximation to the density of $\Phi_m(\tau \mid x^n)$. It follows from (2.4) that

$$\begin{aligned}
\Phi_m(\tau \mid x^n) &= h(x^n \mid t_n) \sum_{s_m \in \{s_m : \hat{\vartheta}_m(s_m) \le \tau\}} g^{m*}\{s_m \mid \hat{\vartheta}_m(s_m)\}\mathrm{lik}^*(s_m \mid t_n) \\
&= m^k h(x^n \mid t_n) \sum_{\bar{s}_m \in \{\bar{s}_m : \hat{\vartheta}_m(m\bar{s}_m) \le \tau\}} g^{m*}\{m\bar{s}_m \mid \hat{\vartheta}_m(m\bar{s}_m)\} \\
&\quad \cdot \mathrm{lik}^*(m\bar{s}_m \mid t_n)\frac{1}{m^k},
\end{aligned}$$

where $\bar{s}_m = s_m/m$. Under some regularity conditions, (3.2) and (3.3) is also true for discrete case. Then for sufficiently large m,

$$\begin{aligned}
&\Phi_m(\tau + \delta \mid x^n) - \Phi_m(\tau \mid x^n) \\
&\quad \approx \left(\frac{m}{2\pi}\right)^{k/2} h(x^n \mid t_n) \sum_{\bar{s}_m \in \{\bar{s}_m : \tau < \hat{\vartheta}_m(m\bar{s}_m) \le \tau + \delta\}} g^{n*}\{t_n \mid \hat{\vartheta}_m(m\bar{s}_m)\} \\
&\qquad \cdot |i\{\hat{\vartheta}_m(m\bar{s}_m)\}|^{-1/2}\frac{1}{m^k}.
\end{aligned}$$

Since $g^{n*}\{t_n \mid \hat{\vartheta}_m(m\bar{s}_m)\}|i\{\hat{\vartheta}_m(m\bar{s}_m)\}|^{-1/2}$ is continuous in $\bar{s}_m$, we have a further approximation

$$\begin{aligned}
&\Phi_m(\tau + \delta \mid x^n) - \Phi_m(\tau \mid x^n) \\
&\quad \approx \left(\frac{m}{2\pi}\right)^{k/2} h(x^n \mid t_n) \int_{\{s : \tau < \hat{\vartheta}_m(ms) \le \tau + \delta\}} g^{n*}\{t_n \mid \hat{\vartheta}_m(ms)\} \\
&\qquad \cdot |i\{\hat{\vartheta}_m(ms)\}|^{-1/2} ds \\
&\quad = \left(\frac{m}{2\pi}\right)^{k/2} h(x^n \mid t_n) \int_{\tau < \hat{\vartheta}_m \le \tau + \delta} g^{n*}(t_n \mid \hat{\vartheta}_m)|i(\hat{\vartheta}_m)|^{1/2} d\hat{\vartheta}_m.
\end{aligned}$$

Thus the approximation (3.4) follows.

Hence leaving out an irrelevant constant, we may consider that $\phi_m(\tau \mid x^n)$ is asymptotically expressible in the form

$$h(x^n \mid t_n)g^{n*}(t_n \mid \tau)|i(\tau)|^{1/2} = \mathrm{lik}(\tau \mid x^n)|i(\tau)|^{1/2}.$$

From this, the evidence of the unobserved value $\hat{\vartheta}_m$ is asymptotically assessed by $|i(\hat{\vartheta}_m)|^{1/2}\mathrm{lik}(\hat{\vartheta}_m \mid x^n)$ by using the observations x^n. Since $\hat{\vartheta}_m$ is a consistent estimate of the parameter θ, replacing $\hat{\vartheta}_m$ by θ yields $|i(\theta)|^{1/2}\mathrm{lik}(\theta \mid x^n)$, and therefore, this can be viewed also as a function ordering possible true values of θ in the light of x^n. Normalizing it to be a probability distribution, we have an inferential distribution $p^B(\theta \mid x^n)$ of the form

$$p^B(\theta \mid x^n) = \gamma(x^n)|i(\theta)|^{1/2}\mathrm{lik}(\theta \mid x^n), \tag{3.5}$$

where $\gamma(x^n)$ is a normalizing constant.

From a Bayesian point of view, the inferential distribution (3.5) is identical to the posterior distribution corresponding to Jeffreys' ((1961), Section 3.10) prior. Since the present approach gives an objective way of constructing inferential distributions for non-Bayesian prediction fit, we can view (3.5) as a result which justifies the use of Jeffreys' prior in Bayesian prediction based on x^n coming from an exponential family. Alternative derivations of this prior have been discussed by many authors: Jeffreys ((1961), Section 3.10), Box and Tiao ((1973), Section 1.3), Akaike (1978), Bernardo (1979), and so on. Among them, Box and Tiao's approach uses the likelihood for θ given x^n to choose prior distributions. They justify Jeffreys' prior on the grounds that it is noninformative or locally uniform for the parameter ω such that the reparametrization $\omega = \omega(\theta)$ makes the likelihood curve 'approximately data translated'. Our procedure is also based on the concept of likelihood. However, the parametric likelihood for θ given z^m evaluated at $\theta = \hat{\vartheta}_m$ is treated as a bootstrap estimate of the probability of the outcome z^m, and employed, therefore, as a priori assessment of the plausibility of z^m.

If we use $\pi(z^m) = c$ instead of (2.2), then from the discussion similar to the above, we have an inferential distribution of the form

$$p^U(\theta \mid x^n) \propto |i(\theta)|\mathrm{lik}(\theta \mid x^n).$$

This can be regarded as the posterior distribution corresponding to the uniform prior to the mean parameter $\mu = \mu(\theta)$ because $d\mu = |i(\theta)|d\theta$.

Incidentally, we are interested also in the performance of the inferential distribution (3.5) when we compare it with other inferential distributions. In the following section, we illustrate a superiority of (3.5) for the gamma model.

4. An illustration

We suppose that the classes of models to be observed and fitted are $\mathrm{Ga}(\alpha, \theta)$ and $\mathrm{Ga}(\beta, \theta)$, respectively, with known shape parameters α and β:

$$g(x \mid \theta) = \frac{\theta^{\alpha} x^{\alpha-1} e^{-\theta x}}{\Gamma(\alpha)} I_{(0,\infty)}(x) \quad \text{and} \quad f(y \mid \theta) = \frac{\theta^{\beta} y^{\beta-1} e^{-\theta y}}{\Gamma(\beta)} I_{(0,\infty)}(y).$$

The density of $t_n = x_1 + \cdots + x_n$ is given by

$$g^{n*}(t \mid \theta) = \frac{\theta^{n\alpha} t^{n\alpha-1} e^{-\theta t}}{\Gamma(n\alpha)} I_{(0,\infty)}(t).$$

Thus the Lauritzen-Hinkley predictive likelihood for $s_m = z_1 + \cdots + z_m$ given t_n is

$$\mathrm{lik}^*(s_m \mid t_n) \propto \frac{s_m^{m\alpha-1} t_n^{n\alpha-1}}{(t_n + s_m)^{m\alpha+n\alpha-1}} I_{(0,\infty)}(s_m).$$

Since $s_m = m\alpha/\hat{\vartheta}_m$, the bootstrap estimate of the probability of the outcome z^m is

$$\pi(z^m) \propto \frac{z_1^{\alpha-1} \cdots z_m^{\alpha-1}}{(z_1 + \cdots + z_m)^{m\alpha}}. \tag{4.1}$$

In addition, $i(\hat{\vartheta}_m) = \alpha/\hat{\vartheta}_m^2$. Then it follows that

$$p^{B_m}(\theta \mid x^n) = \frac{\left(\frac{m}{n}\hat{\theta}_n\right)^{m\alpha-1}}{B(n\alpha, m\alpha - 1)} \frac{\theta^{n\alpha-1}}{\left(\frac{m}{n}\hat{\theta}_n + \theta\right)^{n\alpha+m\alpha-1}} I_{(0,\infty)}(\theta),$$

where $\hat{\theta}_n = n\alpha/t_n$. This is an inverse beta distribution, written $\mathrm{InBe}(n\alpha, m\alpha - 1, (m/n)\hat{\theta}_n)$. Further (3.5) is computed to be

$$p^{B}(\theta \mid x^n) = \frac{t_n^{n\alpha} \theta^{n\alpha-1} e^{-t_n\theta}}{\Gamma(n\alpha)} I_{(0,\infty)}(\theta).$$

Note here that $p^{B_m}(\theta \mid x^n)$ is also expressed as

$$p^{B_m}(\theta \mid x^n) = \frac{t_n^{n\alpha} \theta^{n\alpha-1} \left(1 + \frac{t_n\theta}{m\alpha}\right)^{-m\alpha}}{(m\alpha)^{n\alpha} B(n\alpha, m\alpha - 1) \left(1 + \frac{t_n\theta}{m\alpha}\right)^{n\alpha-1}} I_{(0,\infty)}(\theta),$$

and that $(m\alpha)^{n\alpha} B(n\alpha, m\alpha - 1) \to \Gamma(n\alpha)$ as $m \to \infty$. The latter follows from the asymptotic expansion

$$\Gamma(u) \sim e^{-u} u^{u-1/2} \sqrt{2\pi} \left(1 + \frac{1}{12u} + \cdots\right), \qquad u \to \infty.$$

Now combining these facts, we find that $p^{B_m}(\theta \mid x^n)$ certainly approaches to $p^B(\theta \mid x^n)$ as $m \to \infty$, as was claimed in the preceding section.

Furthermore, calculations similar to the above yield

$$p^E(\theta \mid x^n) = \frac{\left(\frac{n-1}{n}\hat{\theta}_n\right)^{(n-1)\alpha}}{B\{\alpha,(n-1)\alpha\}} \frac{\left(-\frac{n-1}{n}\hat{\theta}_n + \theta\right)^{\alpha-1}}{\theta^{\alpha+(n-1)\alpha}} I_{(0,\infty)}\left(-\frac{n-1}{n}\hat{\theta}_n + \theta\right).$$

On the other hand, by considering the sampling distribution of $\hat{\theta}_n$, we have

$$p^H(\theta \mid x^n) = \frac{(n\alpha\hat{\theta}_n)^{n\alpha}}{\Gamma(n\alpha)} \frac{e^{-n\alpha\hat{\theta}_n/\theta}}{\theta^{n\alpha+1}} I_{(0,\infty)}(\theta).$$

In addition to these, consider the inferential distributions of the form

$$\begin{aligned} p^{V_a}(\theta \mid x^n) &= \frac{t_n^{n\alpha+a}\theta^{n\alpha+a-1}e^{-t_n\theta}}{\Gamma(n\alpha+a)} I_{(0,\infty)}(\theta) \\ &\propto i(\theta)^{(1-a)/2}\mathrm{lik}(\theta \mid x^n), \qquad -n\alpha < a < \infty. \end{aligned}$$

Obviously, $p^{V_0}(\theta \mid x^n) = p^B(\theta \mid x^n)$ and $p^{V_{-1}}(\theta \mid x^n) = p^U(\theta \mid x^n)$. For $a > 0$, $p^{V_a}(\theta \mid x^n)$ is also the posterior distribution of θ corresponding to the vague conjugate prior $\mathrm{Ga}(a, b)$, for which $b \to 0$.

We show in Fig. 1 the graphs of $p^B(\theta \mid x^n)$, $p^U(\theta \mid x^n)$, $p^H(\theta \mid x^n)$ and $p^E(\theta \mid x^n)$ for $\alpha = 2$, $n = 5$ and $\hat{\theta}_n = 1$. The inferential distribution obtained from the procedure of El-Sayyad *et al.* (1989) is quite different from the other three inferential distributions, as it is much more concentrated around the maximum likelihood estimate. Thus $\hat{f}^E(y \mid x^n)$ will work like the estimative fit $f(y \mid \hat{\theta}_n)$. Although the other three densities are similar in shape, the knowledge of $\hat{\theta}_n$ appears to be more effectively used in $p^B(\theta \mid x^n)$ and $p^H(\theta \mid x^n)$ than in $p^U(\theta \mid x^n)$, because the mode of $p^U(\theta \mid x^n)$ deviates more from $\hat{\theta}_n$ than those of the other two densities. In fact, as demonstrated below, this affects the performance of each predictive distribution.

Unfortunately, the predictive distributions obtained from the above inferential distributions are not expressible in explicit forms except those based on $p^{V_a}(\theta \mid x^n)$, $-n\alpha < a < \infty$. The predictive distribution $\hat{f}^{V_a}(y \mid x^n)$ corresponding to the inferential distribution $p^{V_a}(\theta \mid x^n)$ is $\mathrm{InBe}(\beta, n\alpha + a, t_n)$:

$$\hat{f}^{V_a}(y \mid x^n) = \frac{t_n^{n\alpha+a}}{B(\beta, n\alpha+a)} \frac{y^{\beta-1}}{(t_n+y)^{\beta+n\alpha+a}} I_{(0,\infty)}(y), \qquad -n\alpha < a < \infty.$$

Now we use (1.3) as a criterion for evaluating the badness of prediction fit. The superiority of $\hat{f}^{V_a}(y \mid x^n)$ over the estimative fit $f(y \mid \hat{\theta}_n)$ is discussed by Aitchison (1975). Following his discussion there, we have

$$E_\theta[I\{f(\cdot \mid \theta), \hat{f}^{V_a}(\cdot \mid x^n)\}] = \mathrm{const} - W(\beta, n\alpha + a, n\alpha). \tag{4.2}$$

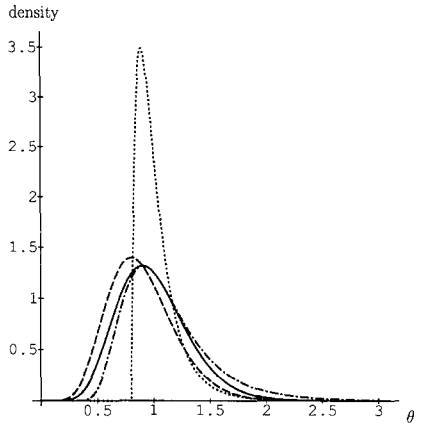

Fig. 1. Inferential distribution $p^B(\theta \mid x^n)$, shown by solid line, $p^U(\theta \mid x^n)$, dashed line, $p^H(\theta \mid x^n)$, dashed-and-dotted line, and $p^E(\theta \mid x^n)$, dotted line: for $\alpha = 2$, $n = 5$ and $\hat{\theta}_n = 1$.

However, the function $W(K, G, k)$ given by Aitchison should be here corrected as follows:

$$W(K,G,k) = \log\left\{\frac{\Gamma(K+G)}{\Gamma(G)}\right\} - K\log k + \frac{Kk}{k-1} - (K+G)\{\psi(K+k) - \psi(k)\},$$

where $\psi(u) = d\log\Gamma(u)/du$ is the digamma function.

We are interested in the value of a at which (4.2) is minimized. By differentiating (4.2) with respect to a, we have

$$\int_{n\alpha}^{\beta+n\alpha}\left\{\frac{d\psi(u)}{du} - \frac{d\psi(u+a)}{du}\right\}du,$$

which is negative for $-n\alpha < a < 0$ and positive for $a > 0$, because the trigamma function $d\psi(u)/du$ is strictly decreasing in $u > 0$; this fact follows from the expression

$$\frac{d\psi(u)}{du} = \sum_{j=0}^{\infty}\frac{1}{(u+j)^2}, \qquad u \neq 0, -1, -2, \ldots$$

(see, e.g., Olver (1974), p. 39). Accordingly, we find that (4.2) is minimized when $a = 0$. That is, $\hat{f}^B(y \mid x^n) = \hat{f}^{V_0}(y \mid x^n)$ is best in the class $\{\hat{f}^{V_a}(y \mid x^n) : -n\alpha < a < \infty\}$. This demonstrates the adequacy of assessing a priori the plausibility of z^m not by $\pi(z^m) = c$ but by (4.1).

It should also be discussed which of $\hat{f}^B(y \mid x^n)$ or $\hat{f}^H(y \mid x^n)$ is superior in terms of (1.3). However, as mentioned above, $\hat{f}^H(y \mid x^n)$ has not yet been computed in an explicit form. Hence, this problem remains to be settled.

Acknowledgements

The author would like to thank the two anonymous referees for their helpful comments.

References

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

Akaike, H. (1978). A new look at the Bayes procedure, *Biometrika*, **65**, 53–59.

Barndorff-Nielsen, O. and Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications (with discussion), *J. Roy. Statist. Soc. Ser. B*, **41**, 279–312.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian Inference (with discussion), *J. Roy. Statist. Soc. Ser. B*, **41**, 113–147.

Bjørnstad, J. F. (1990). Predictive likelihood: a review, *Statist. Sci.*, **5**, 242–265.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, Regional Conference Series in Applied Mathematics, No. 38, SIAM, Philadelphia.

El-Sayyad, G. M., Samiuddin, M. and Al-Harbey, A. A. (1989). On parametric density estimation, *Biometrika*, **76**, 343–348.

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

Hinkley, D. (1979). Predictive likelihood, *Ann. Statist.*, **7**, 718–728.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed., Oxford University Press, Oxford.

Kuboki, H. (1984). A generalization of the relative conditional expectation—further properties of Pitman's T^* and their applications to statistics, *Ann. Inst. Statist. Math.*, **36**, 181–197.

Lauritzen, S. L. (1974). Sufficiency, prediction and extreme models, *Scand. J. Statist.*, **2**, 23–32.

Mathiasen, P. E. (1979). Prediction Functions, *Scand. J. Statist.*, **6**, 1–21.

Murray, G. D. (1977). A note on the estimation of probability density functions, *Biometrika*, **64**, 150–152.

Ng, V. M. (1980). On the estimation of parametric density functions, *Biometrika*, **67**, 505–506.

Olver, F. W. J. (1974). *Asymptotics and Special Functions*, Academic Press, New York.

Pitman, E. J. G. (1979). *Some Basic Theory for Statistical Inference*, Chapman and Hall, London.