Statistics & Probability Letters 5 (1987) 113–118
North-Holland

March 1987

A NEW DEFINITION OF THE PREDICTIVE LIKELIHOOD

Siddhartha CHIB
University of Missouri, Columbia, MO, USA

S. Rao JAMMALAMADAKA
University of California, Santa Barbara, CA, USA

Ram C. TIWARI
University of North Carolina, Charlotte, NC, USA

Received February 1986
Revised August 1986

Abstract: In this paper we present a widely applicable definition of the predictive likelihood based on estimators that are either sufficient or approximately sufficient. Under regularity conditions, this predictive likelihood is shown to equal the Bayes prediction density up to terms of order $O_p(n^{-1})$. For the cases where this predictive likelihood is difficult to compute, an approximate predictive likelihood is provided which differs from the proposed predictive likelihood by $O_p(n^{-1})$. To illustrate the ideas, the approximate predictive likelihood and the Bayes prediction density are obtained for a pth order non-circular autoregression.

Keywords: approximate sufficiency, Bayes prediction density, predictive likelihood, pth order non-circular autoregression.

1. Introduction

Recently there has been much interest in prediction problems for parameteric statistical models using non-Bayesian methods. The main idea is to derive prediction functions, called predictive likelihoods, that map the future observations onto the real line, and which do not depend on the unknown parameter. The related definitions of Lauritzen (1974), Hinkley (1979) and Butler (1985), which involve conditioning on the value of a minimal sufficient statistic are, however, difficult to use especially in models with dependent random variables where sufficient statistics, when available, have intractable exact sampling distributions. To overcome this problem, Davison (1986) using the Bayes set-up, derives an approximate predictive likelihood by expanding the posterior prediction density. Some other definitions, based on maximum likelihood (ML) estimators possessing the best-asymptotically-normal property are contained in Cooley and Parke (1985) and Chib (1985).

In this paper we present a new difinition of the predictive likelihood based on estimators that are either sufficient or approximately sufficient. Of course, these estimators are not necessarily ML estimators. We show that under fairly weak regularity conditions, our predictive likelihood equals the Bayes prediction density upto terms of order $O_p(n^{-1})$, where n is the sample size. Using the results of Durbin (1980) and Abril (1985) we also provide an $O_p(n^{-1})$ approximation to the predictive likelihood, called the approximate predictive likelihood, that is easily computable. Interestingly, under some conditions, our approximate predictive likelihood, which is derived from entirely frequentist principles is shown to be related to the

predictive likelihood obtained by Davison (1986) from the Bayes set-up. As an example, we derive the approximate predictive likelihood for a pth order non-circular autoregressive process with unknown variance. The corresponding Bayes prediction density, and Davison's predictive likelihood, is also derived under the assumption that substantive prior information on the parameters is available.

2. Exact predictive likelihood

Let $\{Y_n,\ n \geqslant 1\}$ be a stationary sequence of real-valued randomvariables defined on a common probability space (Ω, B, P). For each $n \geqslant 1$, let the random-vector $Y_{(n)} = (Y_1, \ldots, Y_n)'$ have a joint pdf $f(y_{(n)};\ \theta)$ where $\theta = (\theta_1, \ldots, \theta_k) \in \theta$, an open subset of R^k, $k \geqslant 1$. Let θ_0 be the true value of the unknown parameter θ. Assume that, for each $n \geqslant 1$, the support $\{y_{(n)}\colon f(y_{(n)};\ \theta) > 0\}$ is independent of θ.

Let $\hat{\theta}_n = \hat{\theta}_n(y_{(n)})$ be an estimator of the parameter θ_0. We assume that $\hat{\theta}_n$ is a unique unbiased estimator of θ_0, and if not unbiased that the bias term is of order $O_p(n^{-1})$, i.e.

$$E_{\theta_0}\,\hat{\theta}_n = \theta_0 + O_p(n^{-1}) \quad \text{as } n \to \infty.$$

We further assume that $\hat{\theta}_n$ is a sufficient or approximately sufficient *estimator* of θ_0. (See definition 1 below.) These assumptions are not unduly restrictive and cover many of the interesting cases.

Definition 1 (Abril, 1985). Let $\hat{\theta}_n = \hat{\theta}(y_{(n)})$ be an estimator of the parameter θ_0. We say $\hat{\theta}_n$ is a (proper) *approximately sufficient estimator* for θ_0, if the joint pdf of $Y_{(n)}$ can be factorized as

$$f(y_{(n)};\ \theta) = g^*(\hat{\theta}_n;\ \theta)h^*(y_{(n)}) \times \left[1 + k^*(y_{(n)};\ \theta,\ n)\right] \qquad (1^*)$$

or as

$$f(y_{(n)};\ \theta) = g(\hat{\theta}_n;\ \theta)h(y_{(n)}) \times \left[1 + k(y_{(n)};\ \theta,\ n)\right] \qquad (1)$$

where the function k^* and k are $O_p(n^{-v/2})$ for some positive integer v, uniformly in $y_{(n)}$ and θ in a neighborhood of θ_0, $g(\hat{\theta}_n;\ \theta)$ is the marginal pdf of $\hat{\theta}_n$ while g^* need not be. Clearly, $\hat{\theta}_n$ is sufficient if the joint pdf $f(y_{(n)};\ \theta)$ can be expressed as (1*) or (1) without the $O_p(n^{-v/2})$ terms.

Note that although the function $g^*(\hat{\theta}_n;\ \theta)$ in (1*) can be deduced immediately from an inspection of $f(y_{(n)};\ \theta)$, the second representation is usually difficult to write down.

The prediction problem can be described as follows. Let the random-vector $Y(n+s) = (Y_{(n)}', Z')'$, where $Z = (Y_{n+1}, \ldots, Y_{n+s})'$ is unobserved, have a joint pdf $f(y_{(n)}, z;\ \theta_0)$. Given a realization $y_{(n)}$ of $Y_{(n)}$ we seek to estimate the conditional pdf $h(z;\ \theta_0 \mid y_{(n)})$, of z given $Y_{(n)} = y_{(n)}$. The idea of predictive likelihood provides a solution to this problem.

Consider now a definition of the predictive likelihood which is more general than the definitions proposed in Lauritzen (1974), Hinkley (1979) and Butler (1986) all of which are based on minimal sufficient statistics. The new definition requires estimators that may be only approximately sufficient, not necessarily sufficient.

Definition 2. Let $\hat{\theta}_n$ and $\hat{\theta}_{n+s}$ be either both sufficient or both approximately sufficient estimators of θ_0 satisfying (1). Then the predictive likelihood of Z given $Y_{(n)} = y_{(n)}$ is defined as

$$L(z \mid y_{(n)}) = \frac{p(y_{(n)},\ z \mid \hat{\theta}_{n+s})}{p(y_{(n)} \mid \hat{\theta}_n)}, \qquad (3)$$

where

$$p(y_{(i)} \mid \hat{\theta}_i) = \frac{f(y_{(i)};\ \theta_0)}{g(\hat{\theta}_i;\ \theta_0)}$$

is the conditional pdf of $Y_{(i)}$ given $\hat{\theta}_i$, $i = n,\ n+s$.

Remark 1. Since our definition is based on unique sufficient or approximately sufficient *estimator* of θ_0, satisfying (1), we do not need the kind of Jacobian term that is necessary in Butler's definition. (See equation 2.4, Butler, 1986). In his case, the Jacobian term is needed to ensure that the same predictive likelihood is obtained for different 1–1 specifications of the minimal sufficient statistic. Secondly, although it is clear that the denominator in (3), being free of z, is not needed in the above definition, its inclusion makes it match the Bayes prediction density (4) (See Theorem 1).

Remark 2. If $\hat{\theta}_i$ is sufficient, then (3) is free of θ, while if $\hat{\theta}_i$ is approximately sufficient then (3) is free of θ upto terms of $O_p(n^{-v/2})$ which for $y_{(i)}$ and θ in a neighborhood of θ_0 go to zero as $n \to \infty$. Thus, by appropriately relaxing the restrictive condition that the predictive likelihood be free of θ for all n (imposed in the definitions of Lauritzen, Hinkley and Butler), we are able to provide a more widely applicable definition of the predictive likelihood.

Definition 3. Let $\lambda(\theta)$ be a prior pdf on the parameter space θ, then the Bayes prediction density of Z, given $Y_{(n)} = y_{(n)}$, is defined as

$$f^B(z \mid y_{(n)}) = \frac{\int_\Theta f(y_{(n)}, z; \theta)\, \lambda(\theta)\, d\theta}{\int_\Theta f(y_{(n)}; \theta)\, \lambda(\theta)\, d\theta}, \tag{4}$$

provided the integrals exist.

The next result provides a justification for our definition of the predictive likelihood, from the Bayes point of view.

Theorem 1. *Let* $\zeta(\hat{\theta}_i) = \int g(\hat{\theta}_i; \theta)\, \lambda(\theta)\, d\theta$, $i = n$, $n+s$, *exist. Assume* (A1) $\hat{\theta}_i$, $i = n$, $n+s$, *are uniformly consistent for* θ_0 *in a neighborhood,* $N(\theta_0)$, *around* θ_0, (A2) $\zeta(\theta)$ *possesses a continuous derivative* $\zeta'(\theta)$ *for all* $\theta \in N(\theta_0)$, *and* (A3) $\hat{\theta}_{n+s} - \hat{\theta}_n = O_p(n^{-1})$ *as* $n \to \infty$, *for* $\hat{\theta}_i \in N(\theta_0)$. *Then*

(i) $f^B(z \mid y_{(n)}) = L(z \mid y_{(n)})\left[1 + O_p(n^{-1})\right]$,

if $\hat{\theta}_i$ *is sufficient, or approximately sufficient with* $v \geqslant 2$, *and*

(ii) $f^B(z \mid y_{(n)}) = L(z \mid y_{(n)})\left[1 + O_p(n^{-1/2})\right]$,

if $\hat{\theta}_i$ *is approximately sufficient with* $v = 1$.

Proof. Consider the case when $\hat{\theta}_i$, $i = n$, $n+s$ are sufficient. Notice that f^B in (4) can be written as

$$f^B(z \mid y_{(n)}) = \frac{\int p(y_{(n+s)} \mid \hat{\theta}_{n+s}) \cdot g(\hat{\theta}_{n+s}; \theta)\, \lambda(\theta)\, d\theta}{\int p(y_{(n)} \mid \hat{\theta}_n) \cdot g(\hat{\theta}_n; \theta)\, \lambda(\theta)\, d\theta}$$

$$= L(z \mid y_{(n)}) \frac{\zeta(\hat{\theta}_{n+s})}{\zeta(\hat{\theta}_n)}$$

$$= L(z \mid y_{(n)}) \cdot \left\{1 + \frac{\zeta'(\theta^*)}{\zeta(\hat{\theta}_n)}(\theta_{n+s} - \hat{\theta}_n)\right\},$$

where $\theta^* \in N(\theta_0)$ is in the line segment joining $\hat{\theta}_{n+s}$ and $\hat{\theta}_n$; the second equality follows because the first term of the integrand is free of θ due to sufficiency; while the third equality follows from Assumption (A1) and (A2). Part (i) of the Theorem finally follows due to Assumption (A3) and an application of Slutsky's Theorem. The proof for the approximately sufficient cases with $v \geqslant 2$ and $v = 1$ follows along the same lines.

Remark 3. Davison (1986) shows that ML estimators, for example, usually satisfy Assumption (A3). Thus that assumption in Theorem 1 is not unduly restrictive.

3. Approximate predictive likelihood

It is often the case that the sufficient or approximately sufficient estimators $\hat{\theta}_i$ do not possess tractable densities, $g(\hat{\theta}_n; \theta_0)$, thus rendering the computation of (3) analytically difficult. In such situations one can use the following result to find an approximation to $L(z \mid y_{(n)})$.

Theorem 2 (Durbin, 1980; Abril, 1986). *Let* $\hat{\theta}_n$ *be a sufficient or approximately sufficient estimator* θ, *satisfying* (1), *and not necessarily the ML estimator. Let*

$$D_n(\theta) = n\, E\{(\hat{\theta}_n - E\hat{\theta}_n)(\hat{\theta}_n - E\hat{\theta}_n)'\}$$

be positive definite and finite for each $n \in N$. *Let* $D_n(\theta) \to D(\theta_0)$ *as* $n \to \infty$ *and* $\theta \to \theta_0$. *Then under Assumptions* 1–4 *of Durbin* (1980), $g(\hat{\theta}_n; \theta_0)$ *is given by*

$$g(\hat{\theta}_n; \theta_0) = \left(\frac{1}{2\pi}\right)^{k/2} \left|\frac{D(\hat{\theta}_n)}{n}\right|^{-1/2} \frac{f(y_{(n)}; \theta_0)}{f(y_{(n)}; \hat{\theta}_n)} \times \left[1 + O_p(n^{-1})\right], \tag{5}$$

if $\hat{\theta}_n$ *is sufficient or approximately sufficient with*

$v \geq 2$, and by

$$g(\hat{\theta}_n;\ \theta_0) = \left(\frac{1}{2\pi}\right)^{k/2} \left|\frac{D(\hat{\theta}_n)}{N}\right|^{-1/2} \frac{f(y_{(n)};\ \theta_0)}{f(y_{(n)};\ \hat{\theta}_n)} \times\left[1 + O_p(n^{-1/2})\right], \tag{6}$$

if $\hat{\theta}_n$ is approximately sufficient with $v = 1$, uniformly for $\hat{\theta}_n$ in a neighborhood of θ_0.

We now state the following result that provides an approximation to the predictive likelihood, called the approximate predictive likelihood, that is easily computable.

Theorem 3. *Let $\hat{L}(z \mid y_{(n)})$ be the approximate predictive likelihood given by*

$$\hat{L}(z \mid y_{(n)}) = \frac{f(y_{(n)},\ z;\ \hat{\theta}_{n+s})}{f(y_{(n)};\ \hat{\theta}_n)} \cdot \frac{|D(\hat{\theta}_{n+s})/(n+s)|^{1/2}}{|D(\hat{\theta}_n)/n|^{1/2}}. \tag{7}$$

Then

(i) $L(z \mid y_{(n)}) = \hat{L}(z \mid y_{(n)})\left[1 + O_p(n^{-1})\right]$

if $\hat{\theta}_i$ is sufficient or approximately sufficient with $v \geq 2$, and

(ii) $L(z \mid y_{(n)}) = \hat{L}(z \mid y_{(n)})\left[1 + O_p(n^{-1/2})\right]$,

if $\hat{\theta}_i$ is approximately sufficient with $v = 1$.

Proof. The above theorem can be easily proved using (5) and (6).

Remark 4. When $\hat{\theta}_n$ is also the ML estimator, or asymptotically equivalent to it as is often the case in practice, the result of Theorem 3.1 hold with $|1/n \cdot D(\hat{\theta}_n)|^{-1/2}$ replaced by $|I_n(\theta_n)|^{1/2}$ (cf. Durbin, 1980, p. 317), where $\hat{I}_n(\hat{\theta}_n) = -\partial^2 \ln f(y(n);\ \hat{\theta}_n)/\partial\theta\partial\theta'$ is the observed information. Thus, in such cases, the ratio of determinants in (7) is replaced by $|\hat{I}_n(\hat{\theta}_n)|^{1/2}/|\hat{I}_{n+s}(\hat{\theta}_{n+s})|^{1/2}$.

Remark 5. We may also mention that $\hat{L}(z \mid y_{(n)})$ is related to Davison's (1986) approximate predictive likelihood which is defined as

$$f^D(z \mid y_{(n)}) = \frac{f(y_{(n)},\ z;\ \hat{\theta}^*_{n+s}) \cdot \lambda(\hat{\theta}^*_{n+s}) \cdot |\hat{I}_n(\hat{\theta}^*_n)|^{1/2}}{f(y_{(n)};\ \hat{\theta}^*_n) \cdot \lambda(\hat{\theta}^*_n) \cdot |\hat{I}^*_{n+s}(\hat{\theta}^*_{n+s})|^{1/2}}, \tag{8}$$

where $\lambda(\cdot)$ is as in (4), $\hat{\theta}^*_i$ is the *ML type* estimate based on the product $f(y_{(i)};\ \theta) \cdot \lambda(\theta)$, and $\hat{I}_i(\hat{\theta}^*_i)$ is the corresponding observed information matrix based on $f(y_{(i)};\ \theta) \cdot \lambda(\theta)$, $i = n,\ n+s$. This definition is obtained by using the Bayes set-up and expanding (4). Only when $\hat{\theta}$, the estimator used in (7) is also the ML estimator, and $\lambda(\theta)$ is constant in a neighborhood of θ_0, will (7) and (8) be identical as functions of z.

4. An example

Consider the pth order non-circular autoregressive process (cf. Anderson, 1971, p. 164)

$$y_n = \alpha_1 y_{n-1} + \cdots + \alpha_p y_{n-p} + \varepsilon_n, \quad n = 1, 2, \ldots, \tag{9}$$

where ε_n are iid $N(0, \sigma^2)$, the roots of $\alpha(L) = 1 - \alpha_1 L - \cdots - \alpha_p L^p$ all lie outside the unit circle, and the initial conditions $y_{(0)} = (y_0, y_{-1}, \ldots, y_{-p+1})'$ are known. Let $\theta = (\alpha', \sigma^2)'$, where $\alpha = (\alpha_1, \ldots, \alpha_p)'$. Clearly, for all $n \geq 1$, the conditional pdf of Y_n given $Y_{(n-1)}$ is

$$h(y_n;\ \theta \mid y(n-1)) = (2\pi)^{-1/2}\ \sigma^{-1} \times \exp\left\{\frac{-1}{2\sigma^2}(y_n - \alpha_1 y_{n-1} - \cdots - \alpha_p y_{n-p})^2\right\}.$$

Let the 'data' matrices be $X_{(n)} = (x_1, \ldots, x_n)'$, $X_{(s)} = (\omega_{n+1}, \ldots, \omega_{n+s})'$, $X_{(n+s)} = (X'_{(n)},\ X'_{(s)})'$, where $x'_i = (y_{i-1}, \ldots, y_{i-p})$, $i = 1, \ldots, n$, and $\omega'_j = (z_{j-1}, \ldots, z_{j-p})$, $j = n+1, \ldots, n+s$. Notice that some elements of $X_{(s)}$ are the unobserved future observations z_{n+j}, $j = 1, \ldots, s-1$. Then, due to (9), the density function of $Y_{(i)}$, $i = n,\ n+s$ conditional on $y_{(0)}$, is

$$f(y_{(i)};\ \theta) = (2\pi)^{-i/2}\ \sigma^{-i} \times \exp\left\{\frac{-1}{2\sigma^2}\| y_{(i)} - X_{(i)}\alpha \|^2\right\}, \tag{10}$$

where $\|\cdot\|$ denotes the usual Euclidean norm. If we let

$$\hat{\theta}_i = (\hat{\alpha}_i, \hat{\sigma}_i^2),$$

where

$$\hat{\alpha}_i = (X_{(i)}'X_{(i)})^{-1}X_{(i)}'y_{(i)} \quad \text{and}$$

$$i\hat{\sigma}_i^2 = \| y_{(i)} - X_{(i)}\hat{\alpha}_i \|^2$$

then the exponent of (10) is

$$i\hat{\sigma}_i^2 + (\hat{\alpha}_i - \alpha)'(X_{(i)}'X_{(i)})(\hat{\alpha}_i - \alpha),$$

thus implying that $\hat{\theta}_i$ is sufficient, $i = n,\ n+s$. They are also the ML estimators. Because these estimators do not possess tractable sampling distribution, we derive the approximate predictive likelihood. Given that

$$|\hat{I}_i(\hat{\theta}_i)|^{1/2} = \left(\frac{i}{2}\right)^{1/2} \hat{\sigma}_i^{-(p+2)} |X_{(i)}'X_{(i)}|^{1/2},$$

$i = n,\ n+s,$

the approximate predictive likelihood of Z given $Y_{(n)} = y_{(n)}$ is (using (7), and Remark 4)

$$\hat{L}(z \mid y_{(n)}) = \frac{(n+s)^{(s+n-p-3)/2}}{n^{(n-p-3)/2}} \cdot \frac{e^{-s/2}}{(2\pi)^{s/2}} \cdot \frac{|X_{(n)}'X_{(n)}|^{1/2}}{|X_{(n+s)}'X_{(n+s)}|^{1/2}} \cdot \frac{[(n+s)\,\hat{\sigma}_{n+s}^2]^{-(s+v-2)/2}}{[n\,\hat{\sigma}_n^2]^{-(v-2)/2}} \tag{11}$$

which (aside from terms not depending on the future observations)

$$= |X_{(n+s)}'X_{(n+s)}|^{-1/2} \times \left[\hat{n}\sigma_n^2 + (z - X_{(s)}\hat{\alpha}_n)' \times M^{-1}(z - X_{(s)}\hat{\alpha}_n)\right]^{-(s+v-2)/2}$$

where $M = [I_s + X_{(s)}(X_{(n)}'X_{(n)})^{-1}X_{(s)}']$, I_s is the identity matrix of order s, and $v = n - p$.

Note that although the approximate predictive likelihood resembles the pdf of a multivariate t distribution with $v - 2$ degrees of freedom, it is not one because the matrix M contains z_{n+j}, $j = 1, \ldots, s-1$.

For purposes of comparison we also compute the Bayes prediction density and Davisons' predictive likelihood. Suppose that prior information on α is diffuse while that on $\tau^2 = 1/\sigma^2$ is Gamma $(v_a/2, v_b/2)$, and independent of α. Thus $\lambda(\theta) \propto \alpha(\tau^2)^{v_a/2-1} e^{-v_b\tau^2/2}$. Then Davisons' predictive likelihood is

$$f^D(z \mid y(n)) = \frac{(n+s+v_a-2)^{(n+s+v_a-p-1)/2}}{(n+v_a-2)^{(n+v_a-p-1)/2}} \cdot \frac{e^{-s/2}}{(2\pi)^{s/2}} \cdot \frac{|X_{(n)}'X_{(n)}|^{1/2}}{|X_{(n+s)}'X_{(n+s)}|^{1/2}} \cdot \frac{[(n+s)\hat{\sigma}_{n+s}^2 + v_b]^{-(s+v+v_a)/2}}{[n\hat{\sigma}_n^2 + v_b]^{-(v+v_a)/2}} \tag{12}$$

while the Bayes prediction density is

$$f^B(z \mid y_{(n)}) = \pi^{-s/2} \cdot \frac{\Gamma((s+v+v_a)/2)}{\Gamma((v-v_a)/2)} \cdot \frac{|X_{(n)}'X_{(n)}|^{1/2}}{|X_{(n+s)}'X_{(n+s)}|^{1/2}} \cdot \frac{[(n+s)\hat{\sigma}_{n+s}^2 + v_b]^{-(s+v+v_a)/2}}{[n\hat{\sigma}_n^2 + v_b]^{-(v+v_a)/2}}$$
$$= \hat{L}(z \mid y_{(n)})[1 + O_p(n^{-1})]$$
$$= f^D(z \mid y_{(n)})[1 + O_p(n^{-1})], \quad p \neq 2, \tag{13}$$

as $n \to \infty$, after considerable simplification.

5. Conclusion

This paper proposes a new definition of the predictive likelihood, and derives an $O_p(n^{-1})$ approximation to the predictive likelihood that is easily computable. The new definition is based on estimators that may be only approximately sufficient and hence can be applied to a large class of models including those with dependent observations.

References

Abril, J.C. (1985) Approximations for densities of approximately sufficient statistics, Manuscript.

Anderson, T.W. (1971), *The Statistical Analysis of Time Series* (John Wiley, New York).

Butler, R.W. (1986), Predictive likelihood inference with applications, *J. Roy. Statist. Soc. Ser. B.* **48**, 1–38.

Chib, S. (1985), Some contributions to likelihood based prediction methods, Ph.D. Dissertation, UC Santa Barbara.

Cooley, T.F. and W.R. Parke (1985), Asymptotic predictive likelihood, Manuscript.

Davison, A.C. (1986), Approximate Predictive Likelihood, *Biometrika*, **13**, 323–332.

Durbin, J. (1980), Approximations for densities of sufficient estimators, *Biometrika* **67**, 311–333.

Hinkley, D.V. (1979), Predictive likelihood, *Annals of Statistics* **4**, 718–728.

Lauritzen, S.L. (1974), Sufficiency, prediction and extreme models, *Scand. J. Statistics* **1**, 128–134.