Statistical Science
1990, Vol. 5, No. 1, pp. 242–265

Predictive Likelihood: A Review

Jan F. Bjørnstad

Abstract. The concept of predictive likelihood is reviewed and studied. The emphasis is on comparing and clarifying some of the more important predictive likelihoods suggested in the literature. A unified modification and simplification of the sufficiency-based predictive likelihoods is suggested. Other predictive likelihoods discussed include the profile predictive likelihood and various modifications of it.

Key words and phrases: Prediction, likelihood, approximations, consistency, invariance, predictive intervals.

1. INTRODUCTION

Prediction of the value of an unobserved or future random variable is a fundamental problem in statistics. From a Bayesian point of view, it is solved in a straightforward manner by finding the posterior predictive density of the unobserved random variable given the data. If one does not want to pay the Bayesian price of having to determine a prior, no unifying basis for prediction has existed until recently. In the last few years, however, attempts have been made to develop a non-Bayesian likelihood approach to prediction via the concept of predictive likelihood. In this paper, we compare and study some of the more important predictive likelihoods that have been proposed in the literature in order to try to describe the current state of affairs. The main perspective is to discuss the concept of predictive likelihood as a foundation for prediction analysis rather than considering various prediction methods.

Let $Y = y$ be the data. The problem is to predict the unobserved value z of Z. The inference is usually in terms of a confidence interval and/or a predictor for z. It is assumed that (Y, Z) has a probability density, with respect to Lebesgue measure or mass function (pdf) $f_\theta(y, z)$, where θ is the unknown parameter vector. In general we shall let $f_\theta(\cdot)$ or $f(\cdot)$ denote the pdf of the enclosed variables, and $f_\theta(\cdot \mid \cdot)$ or $f(\cdot \mid \cdot)$ denotes the conditional pdf of the enclosed variables. $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ based on the data y, and $\hat{\theta}_z$ is the MLE based on (y, z). The MLE based on z alone is denoted by $\tilde{\theta}$. It is assumed that $Y = (X_1, \cdots, X_n)$, the unobserved sample is $Y' = (X_1', \cdots, X_m')$ and Z is some function of Y', like $\sum X_i'$ or Y' itself. $R = r(Y, Z)$ is a minimal sufficient statistic for (Y, Z).

The fundamental point in the prediction problem is that we are dealing with two unknown quantities, z and θ, and the primary aim is to gain information about z with θ playing the role of a nuisance parameter. Berger and Wolpert (1984) formulate a likelihood principle for prediction, stating that all evidence about (z, θ) is contained in the joint likelihood function

$$l_y(z, \theta) = f_\theta(y, z).$$

With l_y as a basis, the objective is to develop a likelihood for z, $L(z \mid y)$, by eliminating θ from l_y. Any such likelihood will be called a predictive likelihood. We see that likelihood prediction must deal with the problem of nuisance parameters, and different ways of eliminating θ from l_y give rise to different predictive likelihoods.

It follows that the concept of predictive likelihood is rather vague, reflected by this review which presents 14 versions of $L(z \mid y)$. However, many of these are quite similar, and all the different versions are based essentially on one of the following three operations on l_y: integration, maximization or conditioning. In comparison, the Bayes approach is equivalent to integrating l_y with respect to a prior on θ.

An example of a predictive likelihood is the so-called profile predictive likelihood, $L_p(z \mid y) = \sup_\theta f_\theta(y, z) = l_y(z, \hat{\theta}_z)$, first studied by Mathiasen (1979). The Bayes posterior predictive density with flat prior, $f_0(z \mid y)$, can be thought of as an integrated (marginal) likelihood, since $f_0(z \mid y) \propto \int l_y(z, \theta)\, d\theta$.

EXAMPLE 1. Consider X_i, X_j''s independent $N(\theta, \sigma_0^2)$, where σ_0^2 is known, and let $Z = \sum X_j'/m$. Then L_p and f_0 result in the same likelihood,

$$L_0 \sim N\left(\bar{x}, \left(\frac{1}{m} + \frac{1}{n}\right)\sigma_0^2\right), \quad \bar{x} = \frac{\sum x_i}{n}.$$

Jan F. Bjørnstad is Professor of Statistics, The University of Tromsø, Institute of Mathematical and Physical Sciences, P.O. Box 953, N-9001 Tromsø, Norway.

EXAMPLE 2. Assume that all X_i's and X_j''s are independent Bernoulli variables with success probability θ and consider $Z = \sum X_j'$. Let $S = \sum X_i$. Then the Bayes predictive density with flat prior is given by

$$f_0(z \mid y) = \frac{\binom{m}{z}\binom{n}{s}}{\binom{m+n}{s+z}} \frac{n+1}{m+n+1}, \quad \text{for } 0 \le z \le m;$$

i.e., the posterior distribution of Z is negative hypergeometric. The profile predictive likelihood is proportional to

$$\binom{m}{z}(s+z)^{s+z}(n+m-s-z)^{n+m-s-z}.$$

Stirling's approximation of L_p results in

$$L_p(z \mid y) \propto \frac{f_0(z \mid y)}{[\hat{\theta}_z(1-\hat{\theta}_z)]^{1/2}}, \quad \hat{\theta}_z = \frac{s+z}{n+m}.$$

This means that, relative to f_0, L_p assigns higher likelihood for z as $\hat{\theta}_z$ approaches 0 or 1.

There is one major difference between predictive likelihood and ordinary parametric likelihood. The parametric likelihood of θ or θ' has no meaning, while the predictive likelihood of a value $z \in B$ can be defined. For instance, in the discrete case, $L_p(z \in B \mid y) = \sup_\theta P_\theta(Y = y \cap Z \in B) = \sup_\theta \{\sum_{z \in B} l_y(z, \theta)\}$. We note that f_0 is additive, but L_p is not. In general, any Bayesian predictive density is additive, while the predictive likelihoods suggested in the literature typically are not. In this review, only the estimative predictive likelihood, L_e, and the bootstrap predictive likelihood, L^*, (both defined in Section 2) are additive. To illustrate these properties, consider again Example 2. Let $m = n = 2$ and assume $s = 1$. Consider $B = \{0, 1\}$. Then $f_0(z \in B \mid y) = f_0(z = 0 \mid y) + f_0(z = 1 \mid y) = .3 + .4 = .7$. Normalized to be a probability distribution, $L_p(z \mid y) = a \cdot \sup_\theta f_\theta(y, z)$, with $a = 128/43$, which gives $L_p(z = 0 \mid y) = .314$ and $L_p(z = 1 \mid y) = .372$. In contrast, $L_p(z \in B \mid y) = a \cdot \sup_\theta \{f_\theta(y, 0) + f_\theta(y, 1)\} = .600$.

From the fact that two predictive likelihoods are equivalent if they are proportional (in z) to each other, we can normalize L to be a probability distribution. For instance, if $\theta = \theta_0$ is known, the unique normalized predictive likelihood is $l_y(z, \theta_0)/f_{\theta_0}(y) = f_{\theta_0}(z \mid y)$. In general, this enables us to describe the predictive likelihood in common distributional terms, as has also been customary in the literature, and makes it easier to compare different predictive likelihoods.

Also, one can consider a predictive likelihood as an estimate of $f_\theta(z \mid y)$ (as done by Lejeune and Faulkenberry, 1982; Levy and Perng, 1986; and Harris, 1989). Then normalizing it to be a probability distribution is indeed the prudent thing to do.

A normalized L will better fulfill the desired purpose of serving as a basis for the prediction analysis. L should play a similar role to the one played by the posterior predictive density for the Bayesian approach, for example as a tool in constructing predictors and confidence regions for z. With a normalized L we can do exactly that. A confidence region P_y for z, called a $(1 - \alpha)$ predictive region and containing the z-values with highest likelihood, can be constructed in the following way,

$$P_y = \{z: L(z \mid y) \geqq k_\alpha\}, \tag{1.1}$$

where k_α is determined such that

$$\int_{P_y} L(z \mid y)\, dz \left(\sum_{z \in P_y} L(z \mid y) \text{ in discrete case} \right) = 1 - \alpha.$$

As an analogue to parametric inference, one possible predictor that does not depend on L being normalized is the maximum likelihood predictor (MLP), $\hat{z}_{\text{ml}}$, the value of z that maximizes $L(z \mid y)$. However, as the uniform and exponential models in Examples 4 and 7 show, the MLP can be undefined (in the sense of not being unique) or obviously unreasonable even in simple models. This is quite contrary to the behavior of MLE in parametric inference. With L normalized an alternative predictor is the mean of L, the analogue of the usual Bayes predictor. The mean of L will be called the predictive expectation of Z and denoted by $E_p(Z)$.

The above considerations show that, not only is it justified, but also advantageous to normalize the predictive likelihood. It also follows that one way to evaluate a predictive likelihood L is to study how well L performs the task of generating predictors and prediction regions. Examples of such evaluations are given in Lejeune and Faulkenberry (1982) and Butler (1989).

A predictive likelihood L should satisfy two fundamental properties. First, as pointed out by Butler (1986, Rejoinder), L should be invariant to a 1–1 reparametrization of the model. That is, the form in which a model is presented should not affect the derivation of L. L_p has this property which must be regarded as a rather basic property of a predictive likelihood. All but one of the predictive likelihoods considered in this review are parameter invariant.

The second basic requirement deals with asymptotic consistency properties and was first discussed by Hinkley (1979) and Mathiasen (1979). When Y and Z are independent these properties can be formulated as follows for a normalized L:

$$L(z \mid Y) \xrightarrow{P} f_\theta(z) \quad \text{as } n \to \infty. \tag{1.2}$$

Assume Z is sufficient for Y', with respect to θ. Then there exists $\{a_m\}$ such that

$$a_m L(Z \mid y) \xrightarrow{P} f_\theta(y) \quad \text{as } m \to \infty, \tag{1.3}$$

where a_m may depend on y, but is independent of z. (1.2) follows since θ is known in the limit as $n \to \infty$ and L should converge to the proper normalized predictive likelihood in this case, $f_\theta(z)$. When $m \to \infty$, predicting Z becomes equivalent to estimating θ, and (an equivalent version of) L should therefore converge to the likelihood function of θ, $f_\theta(y)$.

EXAMPLE 1 (continued). It is readily seen that L_p and f_0, given by L_0, satisfy (1.2) and (1.3) with

$$a_m = a = (\sqrt{2\pi}\sigma_0)^{-(n-1)} n^{-1/2} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}.$$

Further we note that

$$\hat{z}_{ml} = E_p(Z) = \bar{x}.$$

EXAMPLE 2 (continued). By applying Stirling's asymptotic result, $x!e^x x^{-x}(2\pi x)^{-1/2} \to 1$ as $x \to \infty$, it is straightforward to show that f_0 satisfies (1.2) and (1.3) with $a_m = \binom{n}{s}^{-1}(m + n + 1)/(n + 1)$. Also L_p satisfies (1.2). However, as $m \to \infty$ there exists $\{b_m\}$, independent of z, such that $b_m L_p(Z \mid y) \xrightarrow{P} f_\theta(y)/[\theta(1 - \theta)]^{1/2}$. Hence, there exists no sequence $\{a_m\}$, independent of z, such that (1.3) holds, and L_p is not quite consistent in m.

Based on f_0, $E_p(Z) = m(s + 1)/(n + 2)$. For L_p, $E_p(Z)$ must be computed numerically for each case. For instance with $m = 2$ and $n = 3$, L_p results in $E_p(Z) = .19, .75, 1.25, 1.81$ while f_0 gives .4, .8, 1.2, 1.6 for $s = 0, 1, 2, 3$, respectively. This illustrates the fact that L_p gives higher likelihood to $z = 0, 2$ when $s = 0, 3$ than does f_0.

The history of predictive likelihood is short. Although the concept seems to have initially been suggested by Fisher (1956) in the binomial case, the first paper on the subject was Lauritzen (1974) considering discrete random variables. Hinkley (1979) introduces the term "predictive likelihood," thereby expressing the need and desire for a likelihood-based approach to prediction. They propose very similar predictive likelihoods. Mathiasen (1979) studies several prediction functions including the profile predictive likelihood. This likelihood has also been considered by Lejeune and Faulkenberry (1982) and Levy and Perng (1984). Butler (1986) defines a conditional predictive likelihood, from a geometric point of view, which is closely related to the proposals by Lauritzen and Hinkley. They are all based on conditioning on the minimal sufficient R, and only applicable when R provides a genuine reduction of the data. Leonard (1982), Davison (1986) and Tierney and Kadane (1986) all suggest Laplace's approximation to the Bayes posterior predictive distribution.

Davison proposes as a predictive likelihood this approximate Bayes predictive distribution with uniform prior. This predictive likelihood can be regarded as a modification of L_p and is applicable also when R does not provide a genuine reduction. It is, however, not parameter invariant. Leonard (1982) suggests tranforming θ to the parameter ρ which gives the best approximate normality to $f(\rho \mid y, z)$, and then using the approximate Bayes predictive distribution with uniform prior on ρ. This predictive likelihood will then be parameter invariant. Butler (1986, Rejoinder) proposes a parameter invariant approximate conditional predictive likelihood which can also be regarded as an adjusted profile predictive likelihood, and a third modification of L_p is considered in Butler (1989). These two suggestions by Butler (defined later by (3.4) and (3.5)) seem to be the only parameter invariant predictive likelihoods which will work reasonably well in practice in situations where there are (1) no genuine reduction of the data by sufficiency and (2) a large number of unknown parameters. Harris (1989) considers what he calls a bootstrap predictive distribution, which amounts to integrating $l_y(z, \hat{\theta})$ with respect to the distribution of $\hat{\theta}$, at $\theta = \hat{\theta}$.

We shall in this paper concentrate on the profile predictive likelihood and its modifications, and various conditional predictive likelihoods, based on conditioning on R. Butler (1986) also develops a marginal predictive likelihood based on ancillary statistics. This idea is closely related to the traditional approach to prediction and will not be considered here. Other predictive likelihoods based on ancillary statistics have been suggested by Hinkley (1979) and Davison (1986). Neither will these be considered here.

In Section 2, a historical view of the development of sufficiency-based predictive likelihood is taken. We compare the proposed predictive likelihoods and an attempt is made to clarify some of the current misconceptions about these prediction functions. A unified simplification, L_c, is suggested. Section 3 deals with the profile predictive likelihood (which seems to play a central role in prediction), its properties and some modifications of L_p.

In addition to being parameter invariant and asymptotically consistent, a predictive likelihood should also be invariant under scale changes of z. L_p, L_c and the four predictive likelihoods suggested by Davison (1986), Leonard (1982), Butler (1989) and Harris (1989) all possess this invariance. However, the conditional and approximate conditional predictive likelihoods proposed by Butler (1986) are not

invariant under scale changes of the predictand, as shown by Examples 5 and 6 in Sections 2 and 3.

In Section 4 we compare, in the normal case, some of the predictive likelihoods considered earlier by investigating properties of the predictive intervals they generate. These intervals all satisfy (1.1). Also considered is the prediction function L_{pc} presented by Barndorff-Nielsen (1980), the profile of the joint credibility function $l_y(z, \theta)/\sup_z f_\theta(z \mid y)$. L_{pc} is parameter invariant and invariant under scale changes of z. Some final comments are made, including a discussion about model robustness.

2. SUFFICIENCY-BASED PREDICTIVE LIKELIHOODS

Before we start discussing the various suggestions, let us briefly mention the most primitive direct approach. The estimative approach to prediction is to substitute θ by $\hat{\theta}$ in $l_y(z, \theta)$. Normalized this becomes

$$L_e(z \mid y) = f_{\hat{\theta}}(z \mid y).$$

(see, for example, Rao, 1977). This is based on the consideration that if θ is known then $f_\theta(z \mid y)$ is the unique normalized predictive likelihood, and when θ is unknown it seems reasonable to substitute θ with an estimate. However, as emphasized in Aitchison and Dunsmore (1975) and by Butler (1986), L_e will be misleadingly precise since it in effect assumes $\theta = \hat{\theta}$ and does not account for the uncertainty in the knowledge of θ. Section 4 will clearly illustrate that L_e is inadequate for prediction.

Lauritzen (1974) considers discrete (Y, Z) and suggests as a predictor for z the value $\hat{z}$ that maximizes

$$L_1(z \mid y) = f(y \mid r(y, z)). \tag{2.1}$$

Hence $\hat{z} = \hat{z}_{ml}$, the MLP based on L_1, and $L_1(z \mid y)$ is inherently used to assess the "likelihood" of the value z in light of the data y.

Hinkley (1979) deals with continuous as well as discrete random variables, essentially extending and slightly modifying L_1 to cover both cases, although requiring some conditions. This important paper is the first major effort at constructing a likelihood-type basis for prediction, and has since served as an inspiration for much of the research in the area. Let us first assume that Y and Z are independent, and let S, T be the minimal sufficient reductions of Y, Z respectively. Then R is a function $r(S, T)$ of (S, T), where we here use $r(\cdot, \cdot)$ as a generic symbol for the minimal sufficient reduction of the enclosed variables. Hence, $r(y, z) = r(S(y), T(z))$. It is assumed that

$$T \text{ is determined by } (S, R);$$
$$T = \varphi(S, R). \tag{2.2}$$

Hinkley then defines the predictive likelihood for $T = t$ given that $S = s$ by (2.1), i.e.,

$$L_2(t \mid s) = f(s \mid r(s, t)). \tag{2.3}$$

The predictive likelihood of z is then

$$L_2(z \mid s) = f(z \mid t)L_2(t \mid s); \quad t = t(z).$$

Finally, the predictive likelihood of z given y is

$$L_2(z \mid y) = f(y \mid s)L_2(z \mid s); \quad s = s(y). \tag{2.4}$$

In the discrete case, when (2.2) holds, we have that

$$L_2(z \mid y) = f(y, z \mid r(y, z)). \tag{2.5}$$

Also Butler (1986) suggests (2.5) for the discrete case. (2.5) also holds for $L_1(z \mid y)$ when $T = Z$. However, as will be seen later, when $T \neq Z$, L_1 and L_2 are not equivalent. In the general case with Y, Z not necessarily independent, let S again be sufficient for Y, but let T now denote a function of (Y, Z) such that R is a function $r(S, T)$ of (S, T). Assuming that (2.2) holds for this T and that

$$\text{the minimal sufficient reduction of } Z \text{ is determined by } (S, T), \tag{2.6}$$

Hinkley lets $L_2(t \mid s)$ in (2.3) be the predictive likelihood of $T = t$ given s and lets

$$L_2(z \mid s) = f(z \mid s, t) \cdot L_2(t \mid s). \tag{2.7}$$

Here $f(z \mid s, t) = f(s, z)/f(s, t)$ and $t = t(s, z)$.

There seems to be a widely held belief in the literature that L_1 and L_2, given by (2.4), are identically the same. The following example illustrates that such is not the case when the minimal sufficient $T \neq Z$.

EXAMPLE 3. As in Example 2, assume that all X_i's and X_j''s are independent Bernoulli variables with success probability θ, but now let $Z = Y'$. Then the minimal sufficient reductions are $S = \sum X_i$, $T = \sum X_j'$ and $R = S + T$. Here $L_2(z \mid y) = \binom{m+n}{s+t}^{-1}$ while on the other hand

$$L_1(z \mid y) = \frac{P(Y = y \cap T = t)}{P(R = s + t)} = \binom{m}{t} \Big/ \binom{m + n}{s + t},$$

and these are not equivalent. That they can, in fact, give quite different qualitative results is seen by considering $m = n = 2$ and $s = 1$. If these predictive likelihoods are normalized to be probability distributions we get the likelihood values for the four possible values of z shown in Table 1.

It is clear that $t = 1$ is the most likely outcome, for both L_1 and L_2, but otherwise L_1 and L_2 give quite different results. We note that $L_2(z \mid y)$ is consistent with the fact that $L_2(t \mid s = 1) = (0.3, 0.4, 0.3)$ for $t = (0, 1, 2)$ while $L_1(z \mid y)$ is not consistent with $L_1(t \mid s = 1) = L_2(t \mid s = 1)$. The difference between L_1

TABLE 1
L_1 and L_2 in the binomial case, $m = n = 2$ and $s = 1$

z	(0, 0)	(0, 1)	(1, 0)	(1, 1)
$L_1(z \mid y)$	3/14	4/14	4/14	3/14
$L_2(z \mid y)$	3/10	2/10	2/10	3/10

and L_2 is, of course, the factor $f(z \mid t)$. We also observe that the Bayes posterior predictive density $f(z \mid y)$ with flat prior on θ equals $L_2(z \mid y)$, while L_1 is not a Bayes predictive distribution for any conjugate prior. The last statement follows since for any such prior we have that

$$f((0, 1) \mid y) = f((1, 0) \mid y) < \max\{f((0, 0) \mid y), f((1, 1) \mid y)\}.$$

The Bayesian uninformative approach may seem like a possible non-Bayesian likelihood solution. However, a word of caution is in order here. There is typically no unique ignorance prior for θ. In this binomial case it can be very difficult to make a reasonable choice. Besides the flat prior, another possible noninformative prior is the Jeffreys prior, $\pi^{-1}\theta^{-1/2}(1-\theta)^{-1/2}$. This means that the uninformative approach necessarily involves choosing a prior and hence is a Bayesian approach, while the rationale for the likelihood approach is to provide a prediction function without reference to a prior.

The Bayesian uninformative approach suffers from the arbitrariness in the choice of ignorance prior, and the likelihood approach will give no unique answer either. This shows that there is a certain amount of arbitrariness in how we can predict, and that the problem of predicting the number of successes, phrased as "the fundamental problem of practical statistics" by Pearson (1920), is still subject to debate.

Consider now the independence case and definition (2.3) of L_2. It seems unnecessary to restrict L_2 only to cases where (2.2) holds. For given t, s let $r = r(s, t)$. Then $L_2(t \mid s) = f(s \mid r)$ is always well-defined and constant on the surface $\{t' : r(s, t') = r\}$. If, however, (2.2) does hold and (S, R) has a joint density then, with $R = (R_1, \cdots, R_p)$ and $T = (T_1, \cdots, T_p)$, $f(s, r(s, t)) = f(s, t)/||\partial r/\partial t||$. Here $|\partial r/\partial t|$ is the determinant of the $p \times p$-matrix of partial derivatives $\partial r_i/\partial t_j$. This implies that $L_2(t \mid s) = f(s, t \mid r(s, t))/||\partial r/\partial t||$ and

$$L_2(z \mid y) = \frac{f(y, z \mid r(y, z))}{\left|\left|\frac{\partial r}{\partial t}\right|\right|} = \frac{f(y)f(z)}{f(r(y, z))\left|\left|\frac{\partial r}{\partial t}\right|\right|}. \tag{2.8}$$

Typically, R, S, T can be chosen such that R is linear in T. Then $L_2(z \mid y)$ is equivalent to $f(y, z \mid r(y, z))$.

A general problem with L_2 is that it depends on the choice of R, which is not unique, since any one–one transformation of R is also minimal sufficient. So L_2 is really a "sufficiency-based" predictive likelihood, and a more accurate notation would be $L_2^{(R)}$ to indicate this dependency on R. Another serious problem with L_2 arises when for some values of (y, z), $L_2(t \mid s)$ is a probability and for other values of (z, y) it is a density, as illustrated by the next example.

EXAMPLE 4. Let $X_1, \cdots, X_n, Z$ be independent $U(0, \theta)$. Here $S = \max_{1 \leqq i \leqq n} X_i$, $R = \max(S, Z)$, and $P(S = r \mid r) = n/(n+1)$ and $P(S \leqq s \mid r) = s^n/(n+1)r^n$ if $s < r$. Hence, the conditional pdf is a probability for $s = r$ and a density when $s < r$. We find

$$L_2(z \mid s) = \begin{cases} n/(n+1) & \text{if } 0 \leqq z \leqq s \\ ns^{n-1}/(n+1)z^n & \text{if } z > s. \end{cases} \tag{2.9}$$

Unless $s = 1$, L_2 has a jump at $z = s$. This indicates that L_2 only works in the continuous case when $f(s \mid r)$ is a regular density.

For the definition (2.7) of L_2 in the general case the conditions seem superfluous as in the independent case, since $L_2(t \mid s)$ is clearly well-defined even when (2.2) and/or (2.6) do not hold. In fact, Hinkley (1979) applies this definition to two examples and in neither one are (2.2) and (2.6) both satisfied. One problem with this definition is the seemingly arbitrary choice of T indicating that $L_2(z \mid s)$ (for a fixed R) may not be uniquely defined. If we in Example 4 use $T = 0$ or Z according to $Z \leqq S$ or $Z > S$ then, as shown by Hinkley (1979),

$$L_2(z \mid s) = \begin{cases} n/(n+1)s & \text{if } 0 \leqq z \leqq s, \\ ns^{n-1}/(n+1)z^n & \text{if } z > s. \end{cases} \tag{2.10}$$

This seems a more appropriate predictive likelihood than (2.9), but implies at the same time that, at least unless (2.2) holds, $L_2(z \mid s)$ is no longer uniquely defined.

In view of (2.5), (2.8) and the remark after (2.8), it seems at this point natural to suggest the following modification of L_1 and L_2, whether or not Y, Z are independent:

$$L_c(z \mid y) = f(y, z \mid r(y, z)) = f_\theta(y, z)/f_\theta(r(y, z)).$$

If S and T are sufficient for Y and Z, respectively, with $R = r(S, T)$ then $L_c(t \mid s) = f(s, t \mid r(s, t))$ and $L_c(z \mid y) = f(y, z \mid s, t)L_c(t \mid s)$ with $s = s(y)$, $t = t(z)$, independent of the choice of (S, T). However, there

is really no need to introduce (S, T) other than for simplifying the derivation of $L_c(z \mid y)$.

In Example 3, the binomial case, L_c equals L_2 and is therefore the Bayes predictive distribution with flat prior. For the binomial model, the Bayes predictive distribution under any conjugate prior equals a ratio of gamma functions. We see now that this exact functional form is also justified by the sufficiency-based likelihood approach.

We note that L_c is parameter invariant. As with L_2, however, L_c is not invariant with respect to choice of R in the continuous case. Hence, a more appropriate notation would be $L_c^{(R)}$, but for simplicity we shall use L_c. However, when deriving L_c one should always state which R is being used. As we shall see later (in (2.15)), Butler (1986) shows how to derive a "canonical" sufficiency-based predictive likelihood that does not suffer from this lack of invariance.

In Example 4 straightforward calculation shows that $L_c(z \mid s)$ is given by (2.10). In this uniform model, the normalized L_c becomes

$$L_c(z \mid y) = \begin{cases} (n-1)/ns & \text{if } z \leqq s \\ (n-1)s^{n-1}/nz^n & \text{if } z > s. \end{cases}$$

In order to derive a predictor for z from L_c, we first note that the MLP is not unique. A natural choice of predictor based on L_c is the mean, $E_p(Z) = (n-1)s/2(n-2)$. $E_p(Z)$ is approximately equal to $E_{\hat\theta}(Z) = s/2$.

Consider now the case where a prior $f(\theta)$ is available. Let $f_m(\cdot) = \int f_\theta(\cdot) f(\theta)\, d\theta$ and let $l(\theta \mid s) = f_\theta(s)$ be the parametric likelihood. Then the posterior density is $f(\theta \mid s) = l(\theta \mid s) f(\theta)/f_m(s)$.

As emphasized by Hinkley (1979), it is important that a predictive likelihood plays a similar role to the predictive posterior density $f(t \mid s)$ as $l(\theta \mid s)$ does to $f(\theta \mid s)$. Assuming $f_\theta(s, t) > 0 \Rightarrow f_\theta(r(s, t)) > 0$ we have

$$f(t \mid s) = L_c(t \mid s) f_m(r(s, t))/f_m(s),$$

showing a similar factorization. This does not in general hold for L_2.

Let us now consider the exponential family where the X_i and X_j''s are independent with common pdf

$$f_\theta(x) = \exp\left\{ \sum_{i=1}^{k} c_i(\theta) U_i(x) + d(\theta) + b(x) \right\}. \tag{2.11}$$

Here $\theta = (\theta_1, \cdots, \theta_k)$ and $c = (c_1, \cdots, c_k)$ is assumed to be a continuous function of θ.

Letting $U = (U_1, \cdots, U_k)$, the sufficient statistics for Y, Z and (Y, Z) are S, T and $R = S + T$ where $S = \sum_{i=1}^n U(X_i)$ and $T = \sum_{j=1}^m U(X_j')$. In the continuous case, it is assumed that the densities for S and T exist (with respect to the Lebesgue measure on $\mathbb{R}^k$) which typically means that $m, n \geqq k$.

The two desirable consistency properties (1.2) and (1.3) (with $a_m = 1$) of a predictive likelihood L are (i) $L(t \mid S) \xrightarrow{P} f_\theta(t)$ as $n \to \infty$, and (ii) $L(T \mid s) \xrightarrow{P} f_\theta(s)$ as $m \to \infty$.

Recall that, in this context, $\tilde\theta$ is the MLE of θ based on T alone. The asymptotic properties of L_c can be summarized using Theorem 5.4 from Mathiasen (1979) and the fact that $\hat\theta - \theta = O_p(n^{-1/2})$ and $\tilde\theta - \theta = O_p(m^{-1/2})$. Under some regularity conditions on the family (2.11) (see Mathiasen, 1979) we find that:

(a) As $n \to \infty$,

$$\begin{aligned} L_c(t \mid S) &= f_{\hat\theta}(t) + O_p(n^{-1}) \\ &= f_\theta(t) + O_p(n^{-1/2}). \end{aligned}$$

(b) As $m \to \infty$,

$$\begin{aligned} L_c(T \mid s) &= f_{\tilde\theta}(s) + O_p(m^{-1}) \\ &= f_\theta(s) + O_p(m^{-1/2}). \end{aligned} \tag{2.12}$$

In particular, L_c satisfies the two consistency requirements. (2.12) generalizes the result for $k = 1$ from Hinkley (1979).

EXAMPLE 5 (The linear model). Y, Z are independent and

$$\underset{n\times 1}{Y} = \underset{n\times p}{C} \cdot \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon},$$

$\varepsilon_1, \cdots, \varepsilon_n$ are independent $N(0, \sigma^2)$,

$$\underset{m\times 1}{Z} = \underset{m\times p}{C_0} \cdot \beta + \underset{m\times 1}{\varepsilon_0},$$

$\varepsilon_{01}, \cdots, \varepsilon_{0m}$ are independent $N(0, \sigma^2)$.

Full rank of C is assumed.

We shall consider $L_c(z \mid y)$, normalized as a probability distribution. Let $t_\nu^{(k)}(A)$ denote the k-dimensional multivariate t-distribution with ν degrees of freedom and with variance-covariance matrix A, and let $(\hat\beta, \hat\sigma^2)$ be the maximum likelihood estimate of (β, σ^2) based on Y. Define $V = C_0(C'C)^{-1}C_0' + I$. Let $\hat\beta_z$ be the MLE of β and let RSS_z be the residual sum of squares, based on (y, z). With $R = (\hat\beta_z, RSS_z)$ we find, after some algebra, that

$$L_c(z \mid y) \propto \left(1 + \frac{(z - C_0\hat\beta)' V^{-1} (z - C_0\hat\beta)}{n\hat\sigma^2}\right)^{-(n-p-2+m)/2},$$

which implies that $L_c(z \mid y)$ is such that

$$(n-p-2)^{1/2}(Z - C_0\hat\beta)/\sqrt{n}\hat\sigma \sim t_{n-p-2}^{(m)}(V). \tag{2.13}$$

We note that the usual frequentist predictive pivot is $\sqrt{n-p}(Z - C_0\hat\beta)/\sqrt{n}\hat\sigma$, distributed as $t_{n-p}^{(m)}(V)$.

We observe that the MLP and the vector of predictive expectations both equal $C_0\hat{\beta}$.

Using that $\Gamma(x) = e^{-x} \cdot x^{x-1/2}(2\pi)^{1/2}(1 + O(x^{-1}))$ and $(1 + (x/n))^n = e^x + O((1/n))$ it can be shown that

$$L_c(z \mid Y) = f_{\hat{\theta}}(z) + O_p(n^{-1}) = f_\theta(z) + O_p(n^{-1/2})$$

provided $C'C/n \to D$, a positive definite matrix.

For $m = 1$, (2.13) means that

$$\sqrt{n-p-2}(Z - C_0\hat{\beta})/\sqrt{nV}\hat{\sigma} \sim t_{n-p-2}. \tag{2.14}$$

By utilizing the connection between the problem of marginalizing l_y and the problem of removing nuisance parameters in parametric likelihood, Butler (1986) introduces an interesting and novel geometric viewpoint, inspired by the work of Kalbfleisch and Sprott (1970, 1973) on parametric likelihood. Assuming the existence of orthonormal coordinates $u(y, z)$ that are locally orthogonal to $r(y, z)$ such that $(y, z) \to (u, r)$ is one–one, the predictive likelihood is then defined as $L_I(z \mid y) = f(u(y, z) \mid r(y, z))$. Assume z is q-dimensional. Let $r = (r_1, \cdots r_p)$ and let J be the $p \times (n + q)$ matrix of partial derivatives of r with respect to y, z. Then, for the continuous case

$$L_I(z \mid y) = L_c(z \mid y) \mid JJ' \mid^{-1/2}. \tag{2.15}$$

For the discrete case the J-factor disappears and Butler (1986) suggests L_c as the predictive likelihood. We note that such an orthogonal and normed u may not exist globally. For example, let Y, Z be independent $N(0, \sigma^2)$. Then $r(y, z) = y^2 + z^2$ and no normed u orthogonal to r exists (see Appendix). However, u always exists locally, which is all that is needed to define L_I. If the transformation $(y, z) \to (u, r)$ is not differentiable everywhere, then (2.15) is not directly applicable and $f(u \mid r)$ may differ from (2.15). Consider again Example 4 with s as the data. Here, $u = \min(s, z)$ and $r = \max(s, z)$, and the transformation is differentiable if $z \neq s$. When $z \neq s$, $\mid JJ \mid = 1$ and (2.15) equals L_c. However, $f(u \mid r) = n(r^{n-1} + u^{n-1})/(n + 1)r^n$, which is not equivalent to L_c as a function of z. We also note that $(s, z) \to (u, r)$ is one–one only locally, if $z \neq s$, and not globally. Compared to L_c, $f(u \mid r)$ has some disadvantages since it cannot be normalized. Therefore, $E_p(Z)$ is not applicable as a predictor for z and the MLP equals s which is not a sensible predictor. L_I is invariant with respect to choice of R. As mentioned earlier, L_c does not have this property, so L_I can be regarded as the invariant version of L_c. L_I is also parameter invariant, since L_c is. Usually the J-factor changes L_c only slightly, e.g., in normal models it typically has the effect of adding one degree of freedom to L_c. In Example 5, with $m = 1$, L_I is such that

$$\sqrt{n-p-1}(Z - C_0\hat{\beta})/\sqrt{nV}\hat{\sigma} \sim t_{n-p-1}, \tag{2.16}$$

while from (2.14) L_c yields a t_{n-p-2}-distribution.

It should be mentioned that it is possible, with a nontraditional choice of R, to get a substantial contribution from the J-factor. For details we refer to Example 6.

Levy and Perng (1986) consider for this linear model the class Ψ of prediction functions that depend on z only through $w = (z - C_0\hat{\beta})/\sqrt{n}\hat{\sigma}$. Let $L_0(z \mid y)$ be such that $\sqrt{n-p}\,W/\sqrt{V} \sim t_{n-p}$. It is shown that L_0 is optimal for estimating $f_\theta(z)$ in the sense that it minimizes, uniformly in θ, $E_\theta\{\log(f_\theta(Z)/L(Z \mid Y))\}$, for $L \in \Psi$. The optimal L_0 gives the same answer as the traditional frequentist approach, and L_I is approximately optimal in Ψ in this information-theoretic sense. It should in this connection be pointed out that in deriving L_I we did not restrict the prediction problem to consider only w.

Let us finish this section by considering a way to adjust the estimative approach, L_e, to account for the uncertainty in $\hat{\theta}$. Harris (1989) considers independent Y, Z and integrates L_e with respect to the distribution of $\hat{\theta}$, computed at $\theta = \hat{\theta}$:

$$L^*(z \mid y) = E_{\hat{\theta}}\{L_e(z \mid Y)\} = \int f_t(z) f_{\hat{\theta}}(\hat{\theta} = t)\, dt \cdot \left(\sum_t f_t(z) f_{\hat{\theta}}(\hat{\theta} = t) \text{ in discrete case} \right).$$

Harris calls this the bootstrap predictive distribution and shows it is typically an improvement over L_e when considered as an estimate of $f_\theta(z)$. L^* will usually not be on a closed form and can be rather complicated to compute numerically in simple models. Besides comparing it to L_e, properties of L^* have not been studied. It is readily seen that L^* is parameter invariant and invariant under scale changes of z. L^* does not work well, however, in the $U(0, \theta)$-case of Example 4 where it will give zero likelihood to all $z > s$. Still, this is clearly an interesting concept and deserves further attention.

3. THE PROFILE PREDICTIVE LIKELIHOOD AND MODIFICATIONS

Mathiasen (1979) considers the case where Y and Z are independent and looks at several prediction functions: L_1, one based on the plausibility function (see also Barndorff-Nielsen, 1980), and the following likelihood-based function

$$L_p(z \mid y) = l_y(z, \hat{\theta}_z) = \sup_\theta f_\theta(y, z).$$

L_p is also considered by Lejeune and Faulkenberry (1982) and Levy and Perng (1984) and is, of course, well-defined also when Y, Z are not independent. L_p is motivated by an intuitive appealing idea. With z as the "parameter of interest" and θ as the nuisance parameter the most likely value of θ, given (y, z), is determined and $L_p(z \mid y)$ is the resulting likelihood. This corresponds to the profile likelihood in parametric inference (see, e.g., Kalbfleisch and Sprott, 1970) and L_p will be called the profile predictive likelihood. If θ has high dimension L_p can be misleadingly precise as mentioned by Butler (1986) and Aitkin (1986). This was also noted by Kalbfleisch and Sprott (1970) for the parametric profile likelihood. Otherwise, this predictive function can be applied in most situations. Also, as mentioned in Section 1, L_p is clearly parameter invariant.

EXAMPLE 6. Let the X_i's and X_j''s be independent $N(\mu, \sigma^2)$, and $Z = \sum X_j'$. Then, from Lejeune and Faulkenberry (1982), $L_p(z \mid y)$ is such that

$$\frac{Z - m\bar{x}}{(m + (m^2/n))^{1/2}\hat{\sigma}} \sim t_n.$$

Here $\hat{\sigma}^2$ is maximum likelihood estimate of σ^2 based on y and $\bar{x} = \sum x_i/n$. Let $\hat{\mu}_z = (n\bar{X} + Z)/(n + m)$ be the MLE of μ based on (Y, Z), and let $RSS_z = \sum (X_i - \hat{\mu}_z)^2 + (1/m)(Z - m\hat{\mu}_z)^2$. With $R = (\hat{\mu}_z, RSS_z)$, $L_c(z \mid y)$ is found to give

$$\left(\frac{n-3}{n}\right)^{1/2} \frac{(Z - m\bar{x})}{(m + (m^2/n))^{1/2}\hat{\sigma}} \sim t_{n-3}.$$

The R-invariant modification of L_c results in

$$L_I(z \mid y) \propto \left[1 + \frac{(z - m\bar{x})^2}{m(m+n)\hat{\sigma}^2}\right]^{-(n-2)/2} \cdot \left[1 + \beta \frac{(z - m\bar{x})^2}{m(m+n)\hat{\sigma}^2}\right]^{-1/2} \tag{3.1}$$

where $\beta = (m + n)/(m + mn)$. When $m = 1$, $\beta = 1$ and L_I is such that

$$\left(\frac{n-2}{n}\right)^{1/2} \frac{(Z - \bar{x})}{(1 + (1/n))^{1/2}\hat{\sigma}} \sim t_{n-2}.$$

When $m \geq 2$, L_I does not lead to a t-distribution. We note that $0 < \beta \leq 1$. Since the zero-value corresponds to a t_{n-3}-distribution for $\sqrt{n-3}(Z - m\bar{x})/\sqrt{m(m+n)}\hat{\sigma}$, we see that L_I differs only slightly from L_c, one might say the difference is "less than" one degree of freedom.

At this point we should note that with another choice of R, L_c and L_I may differ drastically. Suppose we choose $R = (\hat{\mu}_z, RSS_z^{1000})$. Then the $|JJ'|^{-1/2}$-factor is enormous, L_c is nonsensical, while L_I of course does not change.

The result (3.1) reveals a problem with L_I that does not exist with L_p or L_c A predictive likelihood L is invariant under scale changes of z, $z \to cz = z_t$, where c is a constant, if the normalized $L(z \mid y)$ is the same whether L is based on (y, z) or (y, z_t). L_p and L_c are invariant in this way. It seems that this form of prediction invariance must be regarded as a rather fundamental type of invariance for a predictive likelihood. This normal example shows, however, that L_I is not in general invariant under scale changes of z. This is seen by transforming z to $z_t = z/\sqrt{m}$. Then (Y, Z_t) satisfies the linear model in Example 5. Hence, from (2.16) we get that $L_I(z_t \mid y)$ is such that $\sqrt{n-2}(Z_t - \sqrt{m}\bar{x})/\sqrt{m+n}\hat{\sigma} \sim t_{n-2}$, i.e.,

$$\left(\frac{n-2}{n}\right)^{1/2} \frac{(Z - m\bar{x})}{(m + (m^2/n))^{1/2}\hat{\sigma}} \sim t_{n-2}. \tag{3.2}$$

We shall call this the transformed L_I, denoted by L_I^t, for predicting z. Since L_I^t differs from L_I, it means that L_I is not invariant under $z \to z/\sqrt{m}$.

For all predictive likelihoods considered in this example we see that $\hat{z}_{\text{ml}} = E_p(Z) = m\bar{x}$.

An interesting feature of L_p is its close connection to the Bayesian posterior predictive density with flat prior, first observed by Leonard (1982). Consider the case where $Z = Y'$ and all the X_i's and X_j''s are independent, with $\theta = (\theta_1, \cdots, \theta_k)$. Let $I(\theta)$ and $I^z(\theta)$ be the observed information-matrices based on y and (y, z), respectively. That is, $I^z(\theta) = \{I_{ij}^z(\theta)\}^5$ with $I_{ij}^z(\theta) = -\partial^2 \log f_\theta(y, z)/\partial\theta_i\partial\theta_j$, and similar for $I(\theta)$. Davison (1986) shows that, provided $f_\theta(y)$ and $f_\theta(y, z)$ have well-defined modes as functions of θ, Laplace's approximation method for integrals gives that the Bayesian posterior $f(z \mid y)$ with flat prior equals

$$L_{a1}(z \mid y)\left\{1 + O_p\left(\frac{1}{m+n}\right)\right\} \Big/ \{1 + O_p(n^{-1})\},$$

where

$$L_{a1}(z \mid y) = \frac{L_p(z \mid y)\,|\,I(\hat{\theta})\,|^{1/2}}{f_{\hat{\theta}}(y)\,|\,I^z(\hat{\theta}_z)\,|^{1/2}} \propto L_p(z \mid y)\,|\,I^z(\hat{\theta}_z)\,|^{-1/2}, \tag{3.3}$$

is suggested by Davison (1986) as a predictive likelihood. Leonard (1982) suggests (3.3) for the reparametrization $\theta \to \rho$ that obtains the best approximate normality to $f(\rho \mid y, z)$. Let us denote this parameter invariant predictive likelihood by L_{a1}^t. These two authors, as well as Tierney and Kadane (1986), suggest also for a general prior the same Laplace approximation to the Bayesian predictive distribution. We see that L_{a1} and L_{a1}^t can also be regarded as modifications of L_p. L_{a1} and L_{a1}^t are both invariant under scale changes of z.

One rather fundamental problem with L_{a1} is that it is not invariant to a one–one reparametrization of the model. Butler (1986, Rejoinder) points this out and suggests instead an approximate conditional predictive likelihood, L_{a2}, that is parameter invariant, and can also be regarded as a modification of L_p. $L_{a2}(z \mid y)$ is an approximation of the conditional density of the locally orthonormal coordinate u to $\hat{\theta}_z$ in (y, z)-space, given $\hat{\theta}_z$, and can be expressed as

$$L_{a2}(z \mid y) = \frac{L_p(z \mid y) \mid I^z(\hat{\theta}_z) \mid^{1/2}}{\mid H_z H_z' \mid^{1/2}}. \tag{3.4}$$

Here $H_z = H_z(\hat{\theta}_z)$, and $H_z(\theta)$ is the $k \times (n + m)$ matrix of second-order partial derivatives of $\log f_\theta(y, z)$ with respect to θ and (y, z). Barnard (1986) proposes the use of pivotal distributions for prediction. Butler (1989) shows that, under certain conditions, L_{a2} can be regarded as an approximate predictive pivotal distribution. Also, from Butler (1989), L_{a2} is a saddlepoint approximation to L_I in the case of a regular exponential family. In the normal cases of Examples 5 and 6, L_{a2} equals L_I, which implies that L_{a2}, like L_I, is not in general invariant under scale changes of z.

Of course, L_p, L_{a1}^t, L_{a1} and L_{a2} are all applicable even when sufficiency does not provide a genuine reduction of the data, and also when Y and Z are not independent. A third parameter-invariant adjustment of L_p, L_{a3}, that is also invariant under scale changes of z, is proposed by Butler (1989). L_{a3} is a predictive analogue of the modified profile likelihood in parametric inference, suggested by Barndorff-Nielsen (1983), and obtained as an approximation of $f_\theta(z \mid \hat{\theta}_z)$. Assuming the transformation $(z, \hat{\theta}) \to (z, \hat{\theta}_z)$ is one–one, L_{a3} is given by

$$L_{a3}(z \mid y) = L_p(z \mid y) \mid I^z(\hat{\theta}_z) \mid^{-1/2} \left\| \frac{\partial \hat{\theta}}{\partial \hat{\theta}_z} \right\|. \tag{3.5}$$

Here, $\partial\hat{\theta}/\partial\hat{\theta}_z$ is the matrix of partial derivatives of $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_k)$.

The expressions (3.3)–(3.5) and the accompanying discussion indicate that L_p plays a rather fundamental role in prediction. L_{a3} is clearly the most natural way of making L_{a1} parameter invariant. Also, Butler (1989) shows that, if $X_1, X_2, \cdots, X_n, Z$ are iid with density belonging to a regular exponential family, then L_{a3} is Laplace's approximation to the Bayesian posterior $f(z \mid y)$ with flat prior on $E_\theta(X_i)$.

L_{a2} is not z-invariant in general, and L_{a3} is not applicable when $\hat{\theta}$ is not a function of $(\hat{\theta}_z, z)$ (e.g., the uniform case of Example 4). Still, in situations where there are (1) no low dimensional sufficient statistics and (2) a large number of parameters, it seems that the only parameter invariant predictive likelihoods which work in practical terms are L_{a2} and L_{a3}.

The asymptotic properties of L_p, for the exponential model (2.11), are similar to those of L_c. More precisely, again from Mathiasen (1979, Theorem 5.4), we get

(3.6) (a) As $n \to \infty$,

$$L_p(t \mid S) = f_{\hat{\theta}}(S) f_{\hat{\theta}}(t)\{1 + O_p(n^{-1})\}.$$

(b) As $m \to \infty$,

$$L_p(T \mid s) = f_{\hat{\theta}}(s) f_{\hat{\theta}}(T)\{1 + O_p(m^{-1})\}.$$

(3.6) implies that $L_p(t_1 \mid S)/L_p(t_2 \mid S) \xrightarrow{P} f_\theta(t_1)/f_\theta(t_2)$. Lejeune and Faulkenberry (1982) show that this can be strengthened to almost sure convergence. Also, from part (a) of (3.6), it follows that a proper normalizing constant for L_p as $n \to \infty$ is $f_{\hat{\theta}}(S)$ and

$$\frac{L_p(t \mid S)}{f_{\hat{\theta}}(S)} = f_{\hat{\theta}}(t) + O_p(n^{-1}) = f_\theta(t) + O_p(n^{-1/2}),$$

implying, from (2.12), that L_c and L_p are asymptotically equivalent as $n \to \infty$.

We can also use L_p and L_{a3} to approximate L_c. For the model (2.11), Mathiasen (1979) shows that, provided some regularity conditions hold and R has a density,

$$f_{\hat{\theta}_z}(r) = \frac{1}{[2\pi(m + n)]^{k/2}} \frac{1}{\mid V_{\hat{\theta}_z} \mid^{1/2}} \left(1 + O_p\left(\frac{1}{m + n}\right)\right).$$

Here, V_θ is the variance-covariance matrix for U. Let us now reparametrize (2.11) to canonical form $\theta \to \eta = (c_1(\theta), \cdots, c_k(\theta))$. Then $V_{\hat{\theta}_z} = V_{\hat{\eta}_z}$ and $V_\eta = I^z(\eta)/(n + m)$, where now

$$I_{ij}^z(\eta) = -\partial^2 \log f_\eta(y, z)/\partial\eta_i \partial\eta_j.$$

Moreover, from Butler (1989), in this case $L_{a3}(z \mid y) = L_p(z \mid y) \mid I^z(\hat{\eta}_z) \mid^{1/2}$. It follows that $L_c(z \mid y) = L_{a3}(z \mid y)\{1 + O_p(1/(m + n))\}$. Another approximation to L_c is L_{a1}, given by (3.3), which Davison (1986) claims is accurate to $O_p(1/(m + n))$ in many cases, although no general result to this effect exists for L_{a1}. Also, L_{a3} is parameter invariant in contrast to L_{a1}.

EXAMPLE 6 (continued). Let $m = 1$. It is readily shown that the normalized L_{a3} is identical to L_c. With $\theta = (\mu, \sigma)$, then L_{a1} is such that $(n - 2)^{1/2}(Z - \tilde{x})/\{(n + 1)^{1/2}\hat{\sigma}\} \sim t_{n-2}$. If we use $\theta = (\mu, \sigma^2)$ as parameters L_{a1} leads to $(n - 3)^{1/2}(Z - \tilde{x})/\{(n + 1)^{1/2}\hat{\sigma}\} \sim t_{n-3}$, illustrating the lack of parameter invariance for L_{a1}. As mentioned earlier, L_{a2} is identical to L_I.

EXAMPLE 7. Let $X_1, \cdots, X_n, Z$ be independent with common pdf $f_\theta(x) = (1/\theta)e^{-x/\theta}$, and let $S = \sum X_i$. All predictive likelihoods we consider are in normalized form as probability distributions. With $R = S + Z$, $L_c(z \mid y) = (n - 1)s^{n-1}/(s + z)^n$, while $L_p(z \mid y) = ns^n/(s + z)^{n+1}$. Hence, L_c is such that $(n - 1)Z/s \sim F_{2,2(n-1)}$, and L_p is such that $nZ/s \sim F_{2,2n}$. For comparison, we note that the frequentist

pivot nZ/s is distributed as $F_{2,2n} \cdot L_I$ and L_{ai}, for $i = 1, 2, 3$, all equal L_c. For L_{a1}^t, Leonard (1982) suggests the transformation $\rho = \log(1/\theta)$. Then L_{a1}^t equals L_p. For these predictive likelihoods we note that the MLP $\hat{z}_{\text{ml}} = 0$, illustrating how poor $\hat{z}_{\text{ml}}$ can be as a predictor even in a simple parametric model. An alternative predictor is $E_p(Z)$ which equals $s/(n-1)$ for L_p and $s/(n-2)$ for L_c.

If the model is reparametrized as $\theta e^{-\theta x}$ we find that $L_{a1}(z \mid y) = (n+1)s^{n+1}/(s+z)^{n+2}$, again illustrating that L_{a1} depends on how one chooses to index the family of distributions.

As these examples illustrate the L_{ai}'s are typically very accurate approximations to L_c and L_I, again underlining the central role of L_p.

A situation of general interest is $Z = \sum X_j'$ when $X_1', \cdots, X_m'$ are independent with common distribution and Y, Z are independent. As we have seen the various predictive likelihoods are all rather slight modifications of L_p, so let us now concentrate on L_p. Finding $f_\theta(z)$ can be rather complicated and in some cases next to impossible. The idea is now to approximate $L_p(z \mid y)$ by first approximating $f_\theta(z)$ and then taking sup. The normal distribution is, of course, one approximation, but a much better one is the saddle-point approximation (see Barndorff-Nielsen and Cox (1979) for regularity conditions). Let $M_\theta(\lambda) = E_\theta(e^{\lambda X_1'})$ and $K_\theta(\lambda) = \log M_\theta(\lambda)$. $\hat{\lambda} = \hat{\lambda}_\theta$ is the solution of the saddle-point equation $mK_\theta'(\hat{\lambda}) = z$, where $K_\theta'(\hat{\lambda}) = \partial K/\partial\lambda$. Then $f_\theta(z) = f_\theta^*(z)\{1 + O_p(m^{-1})\}$ where the saddle-point approximation f^* is given by

$$f_\theta^*(z) = \frac{\exp\{mK_\theta(\hat{\lambda}) - z\hat{\lambda}\}}{\{2\pi m K_\theta''(\hat{\lambda})\}^{1/2}}.$$

$f_\theta^*(z)$ is amazingly accurate and usually much more so than the $O_p(m^{-1})$-term seems to indicate. If $X_j' \sim N(\mu, \sigma^2)$ $f_\theta^*(z) \equiv f_\theta(z)$, and if X_j' is gamma distributed, the normalized f_θ^* equals f_θ (noted first by Daniels, 1954). Essentially the same happens in the binomial and Poisson cases where $f_\theta^*(z)$ amounts to using Stirling's approximation for $\binom{m}{z}$ and $z!$ respectively.

The approximate profile predictive likelihood is now

$$L_p^*(z \mid y) = \sup_\theta \left(f_\theta(y) f_\theta^*(z)\right).$$

We note that L_p^* is parameter invariant and invariant under scale changes of z. L_p^* (normalized) equals L_p in the normal and gamma cases. In Example 2, the binomial case with $Z = \sum X_j'$, L_p^* differ from L_p only in the evaluation of $\binom{m}{z}$, using Stirling's formula. The Laplace approximation L_{a1} to the Bayesian predictive distribution $f_0(z \mid y)$ with flat prior and the two parameter invariant adjustments of L_{a1}, L_{a2} and L_{a3} are all identically the same (normalized) and proportional to

$$\binom{m}{z}(s+z)^{s+z+1/2}(n+m-s-z)^{n+m-s-z+1/2}, \quad s = \sum x_i. \tag{3.7}$$

Moreover, Stirling's approximation of (3.7) gives $f_0(z \mid y)$ which also equals L_c in this case. So the approximation to $f_0(z \mid y)$ provided by Laplace's method is virtually as accurate as the saddle-point approximation to L_p.

We finish this section by returning to Example 5, the linear model. Levy and Perng (1984) have shown that

$$L_p(z \mid y) \propto \left(1 + \frac{(z - C_0\hat{\beta})' V^{-1}(z - C_0\hat{\beta})}{n\hat{\sigma}^2}\right)^{-(m+n)/2}$$

and

$$L_p(z \mid y) = f_{\hat{\theta}}(z) + O_p(n^{-1}) = f_\theta(z) + O_p(n^{-1/2}),$$

provided $C'C/n \to D$, a positive definite matrix. This implies that $\sqrt{n}(Z - C_0\hat{\beta}/\sqrt{n}\hat{\sigma} \sim t_n^{(m)}(V)$ and that L_p and L_c have the same asymptotic (in n with p fixed) property. However, L_p ignores the number of parameters in the degrees of freedom and can be excessively accurate if p is large compared to n. With $m = 1$, L_p is such that $(Z - C_0\hat{\beta})/\sqrt{V}\hat{\sigma} \sim t_n$. In this case, the two parameter-invariant modifications of L_p suggested by Butler (1986, Rejoinder; 1989) are found to be $L_{a2}(z \mid y) = L_I(z \mid y)$ and $L_{a3}(z \mid y) = L_c(z \mid y)$. Hence, from (2.14) and (2.16), L_{a2} leads to a t_{n-p-1}-distribution while L_{a3} leads to $n - p - 2$ degrees of freedom, and they both adjust L_p by taking into account the number of parameters in the model.

4. COMPARISONS AND FINAL COMMENTS

One way to compare different predictive likelihoods is to see what kind of predictive intervals they generate. Assuming a given predictive likelihood L is normalized to be a probability distribution in z, a $(1-\alpha)$ predictive interval is given by $[\hat{z}(\alpha_1), \hat{z}(\alpha_2)]$, $\alpha_2 - \alpha_1 = 1 - \alpha$, where $\hat{z}(\alpha_i)$ is the α_i-quantile of L. The confidence level is $\text{Cl}(\theta) = P_\theta\{\hat{z}(\alpha_1, Y) \le Z \le \hat{z}(\alpha_2, Y)\}$. Although we have no guarantee that Cl is close to $1 - \alpha$ we do expect $[\hat{z}(\alpha_1), \hat{z}(\alpha_2)]$ to be an informative interval for Z. Lejeune and Faulkenberry (1982) consider L_p for binomial and Poisson sampling and show that $[\hat{z}(\alpha/2), \hat{z}(1-\alpha/2)]$ has $\text{Cl}(\theta)$ very close to $1 - \alpha$. Let us now consider the normal model in Example 6, with X_i, X_j' independent $N(\mu, \sigma^2)$ and $Z = \sum X_j'$. Let $u(\varepsilon)$, $t_\nu(\varepsilon)$ be upper ε-quantiles in the $N(0,1)$ and the t_ν-distribution, respectively. Since $L_e(z \mid y)$ is

TABLE 2
Confidence levels for P_c, P_I^t and P_p

$1-\alpha$	n											
	Cl_c				Cl_p				Cl_I^t			
	5	10	20	50	5	10	20	50	5	10	20	50
0.90	0.986	0.940	0.919	0.907	0.854	0.880	0.890	0.896	0.947	0.920	0.909	0.904
0.95	0.996	0.975	0.962	0.955	0.917	0.936	0.944	0.948	0.979	0.963	0.956	0.952

$N(m\bar{x}, m\hat{\sigma}^2)$, it follows from Example 6 that the symmetric $(1-\alpha)$ predictive intervals for L_c, L_p and L_e, denoted by P_c, P_p, P_e, are all of the form

$$m\bar{x} \pm d\sqrt{m}\hat{\sigma}, \tag{4.1}$$

where the constant d is equal to

$$d_c = t_{n-3}(\alpha/2)\{(n/(n-3))(1+m/n)\}^{1/2},$$

$d_p = t_n(\alpha/2)(1+m/n)^{1/2}$ and $d_e = u(\alpha/2)$ for P_c, P_p and P_e, respectively. The frequentist interval P_f has, of course, d equal to

$$d_f = t_{n-1}(\alpha/2)\{(n/(n-1))(1+m/n)\}^{1/2}.$$

As we have seen from (3.1), L_I does not lead to a t-distribution, but differs only slightly from L_c. However, the $(1-\alpha)$ symmetric P_I^t interval based on L_I^t, given by (3.2), is also of the form (4.1) with d equal to

$$d_I^t = t_{n-2}\left(\frac{\alpha}{2}\right)\left\{\left(\frac{n}{n-2}\right)\left(1+\frac{m}{n}\right)\right\}^{1/2}$$

The approximate conditional predictive likelihood L_{a2} equals L_I, given by (3.1). (3.1) will give intervals close to the intervals based on the t_{n-2}- and t_{n-3}-distributions. To simplify and unify the comparison, we shall therefore consider the interval P_I^t instead of the one derived from (3.1). We also note that L_{a3} equals L_c.

There are in the literature, it seems, only three predictive likelihoods that are parameter invariant, invariant under scale changes of z and applicable in most parametric models. Two of these are L_p and L_{a3}. The third one is a prediction function suggested by Barndorff-Nielsen (1980) which can also be considered as a slight modification of L_p. The following joint credibility function for z and θ is used:

$$C_y(z,\theta) = \frac{f_\theta(y,z)}{\sup_{z'} f_\theta(z=z'\,|\,y)}.$$

The suggested predictive likelihood is then the profile of C_y,

$$L_{pc}(z\,|\,y) = \sup_\theta C_y(z,\theta).$$

(Strictly speaking, L_{pc} is not a predictive *likelihood* since it is based on C_y instead of l_y.) In this normal case, L_{pc} is such that

$$\sqrt{n-1}(Z-m\bar{x})/\{\sqrt{m(m+n)}\hat{\sigma}\} \sim t_{n-1}.$$

Hence the predictive interval based on L_{pc} equals P_f.

All intervals considered satisfy (1.1) by containing the z-values with highest likelihood.

Now, $P_e \subset P_p \subset P_f \subset P_I^t \subset P_c$ and $\mathrm{Cl}(\theta)$ is independent of θ for all intervals. Since $\mathrm{Cl}_f = 1-\alpha$, $\mathrm{Cl}_c > 1-\alpha$ and $\mathrm{Cl}_I^t > 1-\alpha$ while $\mathrm{Cl}_p < 1-\alpha$ and $\mathrm{Cl}_e < 1-\alpha$. To illustrate how much Cl_c, Cl_I^t, Cl_p differ from $1-\alpha$, we consider the cases $1-\alpha = 0.90$, 0.95, and observe that these confidence levels only depend on n (Table 2).

Cl_e depends on both m and n. That the accuracy in L_e is very misleading is clearly illustrated by Table 3.

An important property is the conditional level given the data, $C_\theta(y) = P_\theta\{\hat{z}(\alpha_1, y) \le Z \le \hat{z}(\alpha_2, y)\,|\,y\}$. Following the terminology in Aitchison and Dunsmore (1975), $C_\theta(y)$ is called the cover of the interval. A measure of the quality of an interval's coverage is the guarantee of coverage $1-\alpha$, defined as $g(1-\alpha) = \inf_\theta P_\theta(C_\theta(Y) \ge 1-\alpha)$. The distribution of the cover is independent of θ for all the intervals. Let $u = n^{1/2}(\bar{x}-\mu)/\sigma$ and $v = n\hat{\sigma}^2/\sigma^2$, and define $C(u,v) = \Phi\{(m/n)^{1/2}u + d(v/n)^{1/2}\} - \Phi\{(m/n)^{1/2}u - d(v/n)^{1/2}\}$. Here $\Phi(x)$ is the cdf of $N(0,1)$. Then $C(u,v)$ with appropriate d is the cover for the various intervals.

Consider first the asymptotic case: $n \to \infty$ and $m/n \to \lambda > 0$. Then $g_e(1-\alpha) \to 0$ while for P_c, P_I^t, P_p and P_f

$$g(1-\alpha) \to g_\lambda(1-\alpha) = P(C_0(U) \ge 1-\alpha),$$

where

$$C_0(u) = \Phi\{\lambda^{1/2}u + u(\alpha/2)(1+\lambda)^{1/2}\} - \Phi\{\lambda^{1/2}u - u(\alpha/2)(1+\lambda)^{1/2}\}.$$

TABLE 3
Confidence level for P_e, $n = 10$

$1-\alpha$	m					
	1	2	5	10	20	100
0.90	0.829	0.812	0.765	0.701	0.609	0.351
0.95	0.890	0.876	0.837	0.779	0.689	0.411

TABLE 4
The asymptotic guarantee of 95% coverage for the 95% predictive intervals from L_c, L_I^t, L_p and for P_f

λ	0.1	0.2	0.5	1	2	3	10	50	100
$g_\lambda(0.95)$	0.687	0.692	0.709	0.740	0.784	0.811	0.875	0.919	0.929

$g_\lambda(1-\alpha)$ can easily be calculated. Table 4 is for $g_\lambda(0.95)$.

We also see that $\lim_{\lambda\to\infty} g_\lambda(1-\alpha) = 1-\alpha$. This is relevant and interesting for cases where $m \gg n$, e.g., survey sampling.

For fixed m, n $g(1-\alpha)$ is more difficult to calculate. If $m/n \ll 1$ we can use approximations developed by Howe (1969). Two cases are given in Table 5.

This example suggests that L_e is practically useless for prediction while the other predictive likelihoods mentioned in this section have good properties in terms of prediction intervals in this normal case, especially when m/n and n are large. This is a situation that occurs often in survey sampling.

Lauritzen (1986) makes the point that exact methods can be very sensitive to the exact formulation of the model. The conditional likelihoods L_c and L_I, being tied to sufficiency, are clearly sensitive to model formulation, since two similar models can have vastly different sufficiency structure. The following illustration shows that L_p and the two modifications L_{a2}, L_{a3} tend to be of a more robust nature, with respect to the model.

Let Log(μ, τ) denote the logistic distribution with mean μ and variance τ. Consider the following two models:

I: Y, Z are independent $N(\mu, 1)$.

II: Y, Z are independent Log(μ, 1).

These two distributions are very similar (see, for example, Johnson and Kotz, 1970), but the two models have very different sufficiency structure. For Model 1, L_p, L_c,L_I, L_{a2} and L_{a3} all give the same predictive likelihood, $N(y, 2)$. In Model 2, sufficiency provides no reduction and $L_c(z \mid y) = L_I(z \mid y) \equiv 1$, i.e., L_c and L_I do not work here. Also in Model 2, $\hat{\mu}_z = \frac{1}{2}(y+z)$, which leads to

$$L_p(z \mid y) = \frac{a^2 e^{-a(z-y)}}{[1+e^{-(a/2)(z-y)}]^4}, \quad a = \pi/\sqrt{3}.$$

TABLE 5
Approximations to g(0.95) for P_c, P_I^t, P_f, P_p and P_e

(n, m)	$g_c(0.95)$	$g_I^t(0.95)$	$g_f(0.95)$	$g_p(0.95)$	$g_e(0.95)$
(10, 1)	0.85	0.76	0.66	0.56	0.29
(10, 2)	0.84	0.75	0.66	0.56	0.26

Normalized to be a density $L_p(z \mid y)$ equals

$$\frac{3ae^{-a(z-y)}}{[1+e^{-(a/2)(z-y)}]^4}.$$

This somewhat resembles the Log(y, 2)-distribution.

L_{a2} and L_{a3} give the same predictive likelihood which in normalized form equals

$$\frac{(4/\sqrt{3})e^{-(3/4)a(z-y)}}{[1+e^{-(a/2)(z-y)}]^3}.$$

This is slightly wider than L_p and is, in fact, extremely close to the $N(y, 2)$-distribution.

Let us at this point mention the approach suggested by Fisher (1956) for the binomial case (Example 3), $L_3(t \mid s) = L_p(t \mid s)/\{f_\theta(s) f_\theta(t)\} \propto L_p(t \mid s)/f_\theta(t)$. This suggests using $L_3(z \mid y) = L_p(z \mid y)/\{f_\theta(y) f_\theta(z)\}$ in general when Y, Z are independent. It follows that $L_3(z \mid y) = L_3(t \mid s)$, missing the factor $f(z \mid t)$. More importantly, L_3 can break down in simple situations like Example 3 (where $f_\theta(z) = \infty$). Mathiasen (1979) and Barndorff-Nielsen (1980) consider other aspects of L_3 that illustrate its inadequacy.

Finally in this section we consider the usual frequentist approach to prediction. It consists of finding a pivotal statistic $U = U(Y, Z)$, i.e., U is ancillary, and then constructing a prediction region for Z based on the (pivotal) distribution of U. As mentioned earlier, the approximate conditional predictive likelihood L_{a2}, given by (3.4), is in certain situations an approximation to the pdf of such an ancillary statistic.

In normal models the pivotal distributions and the usual predictive likelihoods are typically quite similar. However, as indicated by Barndorff-Nielsen (1980), this is really more the exception than the rule. When pivotal solutions cannot be obtained one can use, if n is large and m is small, the approximate method suggested by Cox (1975). This approach, however, tends to get complicated in realistic examples.

There are situations, e.g., in time series, where pivotal solutions are not available, but where predictive likelihood can be used. Some examples are given by Barndorff-Nielsen (1980), one being the first-order autoregressive process, using L_{pc}. In general, time series and forecasting constitutes a major area of application for predictive likelihood. This should be a

fruitful area of future research, although the technical and numerical problems could be quite complex.

APPENDIX: ON THE NONEXISTENCE OF ORTHOGONAL AND NORMED u

Let $(y, z) \in \mathbb{R}^2$ and $r(y, z) = y^2 + z^2$. Define $K = (\partial u/\partial y, \partial u/\partial z)$ and $J = (\partial r/\partial y, \partial r/\partial z) = 2(y, z)$. According to Butler (1986), $u = u(y, z)$ must satisfy two requirements:

(i) $$KK' = 1 \quad \text{and}$$

(ii) $$KJ' = 0.$$

Letting $u_1 = \partial u/\partial y$ and $u_2 = \partial u/\partial z$ (i) and (ii) can be expressed as

(i) $$u_1^2 + u_2^2 = 1 \quad \text{and}$$

(ii) $$yu_1 + zu_2 = 0.$$

The solutions are $u_1^0 = \pm z/(y^2 + z^2)^{1/2}$ and $u_2^0 = \mp y/(y^2 + z^2)^{1/2}$. Hence, from u_1^0 we have that

$$u(y, z) = \pm z \log[C_1(z)\{y + (y^2 + z^2)^{1/2}\} + C_2(z)].$$

It is clear that $\partial u/\partial z \neq u_2^0$ and hence no u satisfying (i) and (ii) exists.

It should be mentioned that an orthogonal u is $u = \tan^{-1}(z/y)$. However, (u, r) is then not one–one with (y, z).

ACKNOWLEDGMENTS

The author would like to express his appreciation for the inspiring and thought-provoking comments given by the former Executive Editor, Morris DeGroot. This work was partly done while the author was visiting the Center for Statistical Sciences, The University of Texas at Austin. Helpful and inspiring discussions with Ron Butler, Anthony Davison and David Hinkley are gratefully acknowledged. The author would also like to thank three referees and the Executive Editor, Carl N. Morris, for many penetrating comments that led to substantial improvements. Special thanks go to one referee who read several versions of the paper. This research was supported in part by the Norwegian Research Council for Science and the Humanities.

REFERENCES

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

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

BARNARD, G. A. (1986). Comment on "Predictive likelihood inference with applications" by R. W. Butler. *J. Roy. Statist. Soc. Ser. B* **48** 1–38.

BARNDORFF-NIELSEN, O. (1980). Likelihood prediction. Istituto Nazionale di alta Matematica. *Symposia Mathematica* **25** 11–24.

BARNDORFF-NIELSEN, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365.

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.

BERGER, J. O. and WOLPERT, R. L. (1984). *The Likelihood Principle*. IMS, Hayward, Calif.

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

BUTLER, R. W. (1989). Approximate predictive pivots and densities. *Biometrika* **76** 489–501.

COX, D. R. (1975). Prediction intervals and empirical Bayes confidence intervals. In *Perspectives in Probability and Statistics* (J. Gani, ed.). 47–55. Academic, London

DANIELS, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25** 631–650.

DAVISON, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73** 323–332.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, London.

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

HINKLEY, D. V. (1979). Predictive likelihood. *Ann. Statist.* **7** 718–728. Corrigendum **8** 694.

HOWE, W. G. (1969). Two-sided tolerance limits for normal populations: Some improvements. *J. Amer. Statist. Assoc.* **64** 610–620.

JOHNSON, N. I. and KOTZ, S. (1970). *Continuous Univariate Distributions* **2**. Houghton Mifflin, Boston.

KALBFLEISCH, J. D. and SPROTT, D. A. (1970). Applications of likelihood methods to models involving large numbers of parameters (with discussion). *J. Roy. Statist. Soc. Ser. B* **32** 175–208.

KALBFLEISCH, J. D. and SPROTT, D. A. (1973). Marginal and conditional likelihoods. *Sankhyā Ser. A* **35** 311–328.

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

LAURITZEN, S. L. (1986). Comment on "Predictive likelihood inference with applications" by R. W. Butler. *J. Roy. Statist. Soc. Ser. B* **48** 1–38.

LEJEUNE, M. and FAULKENBERRY, G. D. (1982). A simple predictive density function. *J. Amer. Statist. Assoc.* **77** 654–657.

LEONARD, T. (1982). Comment on "A simple predictive density function" by M. Lejeune and G. D. Faulkenberry. *J. Amer. Statist. Assoc.* **77** 657–658.

LEVY, M. S. and PERNG, S. K. (1984). A maximum likelihood prediction function for the linear model with consistency results. *Comm. Statist. A—Theory Methods* **13** 1257–1273.

LEVY, M. S. and PERNG, S. K. (1986). An optimal prediction function for the normal linear model. *J. Amer. Statist. Assoc.* **81** 196–198.

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

PEARSON, K. (1920). The fundamental problem of practical statistics. *Biometrika* **13** 1–16.

RAO, C. R. (1977). Prediction of future observations with special reference to linear models. In *Multivariate Analysis IV* (P. R. Krishnaiah, ed.) 193–208. North-Holland, Amsterdam.

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86.