Bernoulli **2**(4), 1996, 319–340

Prediction and asymptotics

O.E. BARNDORFF-NIELSEN[1]* and DAVID R. COX[2]

[1]*Dept. Mathematical Sciences, Aarhus University, DK-8000 Aarhus, Denmark*
[2]*Nuffield College, Oxford OX1 1NF, UK*

Prediction of an unobserved random variable is considered from a frequentist viewpoint. After a brief review of previous work, a number of examples in which an exact solution is possible are given, partly for their intrinsic interest and partly to illustrate general results. A new form of predictive density is derived accurate to the third order of asymptotic theory under ordinary repeated sampling. The formula is invariant under transformation of the observed and unobserved random variables and under reparametrization. It respects the conditionality principle and may be based on the minimal prediction sufficient statistic. Some open problems are noted.

Keywords: autoregression; invariant expansion; conditioning; prediction sufficiency; predictive density; predictive limits

1. Introduction

In the formal theory of inference, it is useful to distinguish estimation from prediction, the latter being concerned with the values of further random variables connected with the probability model assumed to underlie the data. The present paper deals with prediction from parametric models, giving first some exact results and then, from Section 5 onwards, concentrating on the role of asymptotic theory.

A fairly general formulation is as follows. A vector of observations y corresponds to a random variable Y having a density $p(y;\theta)$, where θ is unknown. It is required to predict the value of a further random variable Z having a density depending on θ. Broadly speaking, y is to be used to estimate θ, leading to an inference about Z.

In general, Y and Z are dependent so that the model specification requires the joint density $p(y,z;\theta)$. An important special case has Y and Z independent, a further special case of this arising when Y is a random sample from the same distribution as that of Z.

The latter part of the paper considers asymptotic theory in which the amount of information about θ contained in Y is large, whereas the dimension of Z is fixed, and indeed for the great majority of the discussion we take Z to be one-dimensional; this includes the case where Z is a summary statistic from a second sample. Prediction problems in which Z is multi-dimensional may often be best reformulated in terms of a number of one-dimensional targets. Where this is not reasonable, there may be some arbitrariness in the 'shape' of the prediction region to be adopted, although the shape corresponding to high predictive density with respect to some natural, even if ultimately arbitrary, dominating measure may guide the choice. See Example 4.6.

* To whom correspondence should be addressed.

1350-7265

Two other asymptotic regimes are of interest but will not be considered in detail here. In one the amounts of information about θ in Y and Z are large and comparable, as in the application to the superpopulation theory of sampling a finite population. In the second the amount of information in Z increases to infinity and the asymptotic theory of estimation is recovered as a limit.

The paper is organized as follows. In Section 2 we briefly review a number of different approaches to prediction and we indicate the route we will be taking, which is via predictive limits to a new type of predictive distribution, derived by asymptotic considerations. Section 3 briefly reviews the concept of prediction sufficiency and notes other relevant inference points. A number of examples that allow 'exact' solutions are discussed, partly for their intrinsic interest and partly as a means of assessing the performance of the asymptotic predictive density (6.8), which constitutes the main result of this paper. Sections 5 and 6 and the Appendix are devoted to the derivation of formula (6.8), and the formula is illustrated by a number of examples in Section 7. The cases of many nuisance parameters and of discrete predictands are discussed in Sections 8 and 9, respectively, and the concluding Section 10 lists a variety of open problems.

2. Formulations of prediction

There are a number of possible formulations of prediction problems even in the special case of one-dimensional Z.

Prediction may be by a point value. See, in particular, Aitchison and Dunsmore (1975) for a detailed discussion largely from a Bayesian viewpoint. Formal analysis requires the specification of a loss function, such as squared error. In a fully Bayesian formulation, i.e. one with a known prior over θ, there is, formally at least, no difficulty in principle in obtaining the posterior density of Z given $Y = y$ and this supplies predictive limits. In this approach y is fixed and an ensemble of possible values of θ is involved. Indeed, in Bayesian theory the distinction between estimation and prediction disappears, at least so far as the formalism is concerned. Geisser (1993) largely emphasizes the Bayesian approach to predictive distributions. We shall not follow this route.

Other approaches – Fisher (1956), Lauritzen (1974), Hinkley (1979) and Butler (1986; 1989), on the one hand, and Barndorff-Nielsen (1978; 1981) and Mathiasen (1979), on the other – define a 'predictive likelihood' for Z intended to have properties more or less analogous to those of likelihood for a parameter. Bjørnstad (1990) provides a valuable review of these developments.

We shall use the term 'predictive densities' for probability densities constructed from the observed data y with the direct aim of approximating the true conditional density of Z closely, in some specified sense, and this is the approach adopted here. Bjørnstad (1990) also discusses work on predictive densities up to 1990. Recently Vidoni (1995) extends an approach of Harris (1989), by use of the p^*-formula. Komaki (1996) considers construction of suitable predictive densities from the viewpoint of differential geometry and in the spirit of Amari. Harris (1989), Basu and Harris (1994) and Vidoni (1995) all measure closeness of the approximations by Kullback–Leibler distance.

The predictive densities in these papers are, however, different in nature from that to be studied in the present paper.

We may look for predictive limits, i.e. a function $c_\alpha(y)$ such that, exactly or approximately,

$$P\{Z < c_\alpha(Y); \theta\} = \alpha \tag{2.1}$$

for all θ. A set of such limits for all α supplies a fairly complete specification of knowledge about Z derivable from y. In (2.1) the probability is taken over the joint distribution of Y and Z. That is, the hypothetical frequency interpretation involves repetition of the data y, as well as of the value to be predicted z. If there is an ancillary statistic a for θ based on y, it will be reasonable to condition on its observed value.

There are broadly three approaches to satisfying (2.1), two special and 'exact' and the third general and asymptotic.

The first hinges on finding a pivotal function of Z and Y, typically of Z and the sufficient statistic for θ based on Y, the pivot having a distribution free of θ. Subject to monotonicity, a solution of (2.1) is obtained. The method yields well-known solutions to a number of standard problems of scale and location form.

The second method (Guttman 1970) involves a notional testing problem in which Z is governed by a parameter value θ^*, Y by θ, and the prediction interval consists of all those z for which the hypothesis $\theta = \theta^*$ would not be rejected at level α. This gives in particular a simple solution to what Pearson (1920) christened the 'Fundamental Problem of Applied Statistics': given the number of successes in n Bernoulli trials, how many successes will occur in a further m such trials. Fisher (1956, pp. 134–138) gave a closely related solution based directly on likelihood. Fisher's predictive likelihood has, however, the undesirable feature that it does not converge to parametric likelihood based on Y if the information in Z increases to infinity.

In the asymptotic approach, we argue initially as follows. Let $z_\alpha(y;\theta)$ be the α quantile of the conditional distribution of Z given $Y = y$. That is,

$$G(z_\alpha(y;\theta);\theta|y) = P_\theta\{Z < z_\alpha(y;\theta)|\, Y = y\} = \alpha, \tag{2.2}$$

where $G(z;\theta|y)$ is the distribution function of Z given $Y = y$; of course, if Z and Y are independent the dependence on y can be suppressed. Then if $\tilde{\theta}$ is an estimate of θ based on y, $z_\alpha(y;\tilde{\theta})$ provides an approximation to the required prediction limit. We shall generally take $\tilde{\theta}$ to be the maximum likelihood estimate $\hat{\theta}$. The objective of the asymptotic calculation is to 'improve' $z_\alpha(y;\hat{\theta})$ so as to satisfy (2.1) to a close approximation. A first approach to this (Cox 1973; see also Barndorff-Nielsen and Cox 1994, pp. 217–219; and Atwood 1984, with interesting examples) proceeds, in outline, as follows.

The starting-point is (2.2). When Z is independent of Y, we have, on replacing θ by a $\sqrt{n}$-consistent estimate $\tilde{\theta}$, that

$$P\{Z < z_\alpha(\tilde{\theta});\theta\} = \mathrm{E}\{G(z_\alpha(\tilde{\theta});\theta)\} = \mathrm{E}\{H_\alpha(\tilde{\theta},\theta)\},$$

where $H_\alpha(\tilde{\theta};\theta) = G(z_\alpha(\tilde{\theta});\theta)$ and the expectation is over the distribution of Y, i.e. of $\tilde{\theta}$. If asymptotically the bias of $\tilde{\theta}$ is $b(\theta)/n$ and the covariance matrix $c(\theta)/n$, Taylor expansion

shows that, with accuracy $O(n^{-3/2})$,

$$P\{Z < z_\alpha(\tilde{\theta}); \theta\} = \alpha + d(\theta)/n, \tag{2.3}$$

where

$$d(\theta) = [b\nabla_{\tilde{\theta}}^{\mathrm{T}} H_\alpha + \tfrac{1}{2}\,\mathrm{tr}\,\{\tilde{c}\nabla_{\tilde{\theta}}\nabla_{\tilde{\theta}}^{\mathrm{T}} H_\alpha\}]_{\tilde{\theta}\to\theta},$$

∇ denoting the gradient operator. When $\tilde{\theta}$ is the maximum likelihood estimate, $\hat{\theta}$, general formulae are available for b and c, or, if $\tilde{\theta}$ is available in explicit form, it may be simpler to calculate its bias and covariance matrix directly.

Sometimes the correction term in (2.3) may be so small that further consideration is unnecessary, but in general we aim to modify $z_\alpha(\tilde{\theta})$ so as to get closer to the target value α. This can be done in various ways. The most direct is to replace $z_\alpha(\tilde{\theta})$ by $z_{\alpha_1}(\tilde{\theta})$, where

$$\alpha_1 = \alpha + d(\tilde{\theta})/n. \tag{2.4}$$

The second approach is to modify $z_\alpha(\tilde{\theta})$ to $z_\alpha(\tilde{\theta}) + z_\alpha^*(\tilde{\theta})/n$ and to choose $z_\alpha^*(\tilde{\theta})/n$ to absorb the additional term in (2.3).

The third is to 'guess' the form of the predictive density, typically a rather long-tailed modification of the distribution of Z, to expand this distribution with the distribution of Z as leading term and then to calculate the prediction limit to match (2.1).

If Y and Z are not independent, but there is a low-dimensional statistic W, so that the quantile is $z_\alpha(\theta; w)$, the argument proceeds by first conditioning on W. Typically, because W is of low dimension, the asymptotic covariance of $\hat{\theta}$ is unaffected by the conditioning but the $O(n^{-1})$ bias of $\hat{\theta}$ will be changed; in some of our examples the availability of simple explicit forms for $\hat{\theta}$ and W enable the conditional bias to be evaluated directly. We thereby obtain prediction limits having the desired properties conditionally on W and therefore also unconditionally.

While this method is direct, flexible, and elementary, the form of the answer is unenlightening and there is typically some lack of invariance with respect to choice of the parameter θ, although not, at least in the form (2.3), with respect to transformation of Z. Also we have not here taken the conditionality principle into account.

In the majority of cases, use of the maximum likelihood estimate $\hat{\theta}$ is natural and Sections 5–7 of the paper are concerned with a solution based explicitly on the maximum likelihood estimate, on invariant calculations with the likelihood function, and on the conditionality principle.

It may be convenient to present the results of (2.1) via a formal predictive density for Z, $\vec{g}(z|y)$ say, integration of which will generate the predictive limits of (2.1). That is,

$$\int_{-\infty}^{c_\alpha(y)} \vec{g}(u; y)\,\mathrm{d}u = \alpha$$

for all α. In simple cases the form of $\vec{g}$ may be directly recognizable; in general, it can be obtained via an expression of α as a function of c.

An important result of the present paper is an asymptotic expression for $\vec{g}$, given as formula (6.8).

3. Prediction sufficiency and inference

From one point of view an important distinction is between problems in which (Y, Z) are or are not independent. From a more applied perspective, the dependent problems can be separated into forecasting problems, in which Z is a future observation in a stochastic system that generated Y, and those where Z is some latent feature of the model generating Y, these being broadly of empirical Bayes type. In both dependent cases, a crucial aspect is the form of $p(z; \theta | y)$, the conditional density of Z given $Y = y$. In some, although not all, cases this involves y only via a low-dimensional component w, say, i.e.

$$p(z; \theta | y) = p(z; \theta | w).$$

Whether W is low-dimensional or not we call W *transitive* for Z, a terminology consistent with Bahadur (1954). Note that w has no necessary connection with the estimation of θ. In general, the solution of a prediction problem will depend not only on the transitive statistic w but also on some further part v of the data y. We shall say that (v, w) is *prediction sufficient* with respect to the solution in question. The mathematical definition of prediction sufficiency that we shall use, and which goes back to Kolmogorov (1942) and Bahadur (1954), says in essence that a statistic T (a function of Y alone) is prediction sufficient for Z provided it is transitive and sufficient for Y with respect to inference on θ.

Under mild regularity conditions, T is (minimal) prediction sufficient for Z if and only if T is (minimal) sufficient for the class of all conditional distributions of Y given Z, this class considered as a parametric family with both θ and z as parameters (Barndorff-Nielsen and Skibinsky 1963; Skibinsky 1967, Theorem 2).

We use a prediction sufficient reduction wherever possible.

When the conditional distribution of z given y does, in fact, depend on θ through ψ only, it will usually be advisable in the calculations of predictive limits and predictive densities to use not the likelihood function for θ based on y but rather an adjusted profile likelihood, and to condition on any ancillary statistic, a (Barndorff-Nielsen and Cox 1994; Barndorff-Nielsen 1995).

Example 3.1. The relevance of conditioning on ancillaries is well illustrated by the case where $Y = (Y_1, \ldots, Y_n)$ with $Y_1, \ldots, Y_n$ and Z independent and identically following the uniform distribution on the interval $(\theta, \theta + 1)$. Here, denoting the order statistic of Y by $(Y_{(1)}, \ldots, Y_{(n)})$, we have that $W = (Y_{(1)}, Y_{(n)})$ is prediction sufficient and $a = Y_{(n)} - Y_{(1)}$ is ancillary, and the precision with which we are able to predict Z depends crucially on the size of a, with high precision if a is close to 1 and low precision if a is close to 0.

4. Some touchstone examples

We now give examples which illustrate a number of types of prediction problem for which formally 'exact' solutions are available. In particular, it is important that the general asymptotic procedure to be developed later reproduces exactly or very nearly these 'exact' results.

Example 4.1 *Problems with pivots.* Let $Y_1, \ldots, Y_n$ be a random sample from a normal distribution of unknown mean μ and unit variance and let Z be an independent future observation from that distribution. Then $Z - \bar{Y}$ is a pivot with a normal distribution of zero mean and variance $1 + n^{-1}$, leading (Fisher 1935) to

$$c_\alpha(y) = \bar{y} + k_\alpha(1 + n^{-1})^{1/2},$$

where $\Phi(k_\alpha) = \alpha$. Note that $z_\alpha(\mu) = \mu + k_\alpha$. The predictive density is normal with mean $\bar{y}$ and variance $1 + n^{-1}$; in the corresponding problem with unknown variance it follows the Student t distribution with $n - 1$ degrees of freedom. Note that the predictive density is not in the same family as the density of Z, but it is in a related family with inflated dispersion.

A similar argument applies if Z is exponentially distributed with mean μ. Suppose that on the basis of data Y, an unbiased estimate $\hat{\mu}$ of μ is constructed having a gamma distribution, say, with degrees of freedom d_n. Then, assuming that Y and Z are independent, $Z/\hat{\mu}$ has the standard F distribution with $(2, d_n)$ degrees of freedom from which predictive limits follow. Again the prediction distribution is not a member of the family for Z.

If Y consists of a random sample of size n from the exponential distribution of mean μ, then $\hat{\mu}$ is the sample mean and $d_n = 2n$. Much more generally, however, $\hat{\mu}$ might be constructed via some kind of regression analysis with, typically, the effective degrees of freedom d_n obtained in some approximate way.

Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed in the location model $f(y - \theta)$, where f is known and θ is unknown, and that Z is independent of $Y = (Y_1, \ldots, Y_n)$ and has density $h(z - \theta)$, a possibly different location model with the same θ. Conditionally on the ancillary sample configuration $a = (y_1 - \hat{\theta}, \ldots, y_n - \hat{\theta})$, the maximum likelihood estimator $\hat{\theta}$ has a known distribution, obtained by normalizing the likelihood function to integrate to 1. Thus the pivot $Z - \hat{\theta}$ has a known density found by convolving $h(\cdot)$ with the likelihood function.

Example 4.2 *Maximum of m normal observations.* Suppose that $f(\cdot)$ is standard normal and that Z is the largest of m future observations from the same distribution. Thus the density of Z is

$$m\varphi(z - \theta)\{\Phi(z - \theta)\}^{m-1},$$

and the density of the pivot, in this case $Z - \bar{Y}$, is

$$m\sqrt{n}\int_{-\infty}^{\infty} \varphi(\sqrt{n}(x - u))\varphi(u)\{\Phi(u)\}^{m-1}\,\mathrm{d}u.$$

This, or the corresponding distribution function

$$m\int_{-\infty}^{\infty} \Phi(\sqrt{n}(x - u))\varphi(u)\{\Phi(u)\}^{m-1}\,\mathrm{d}u,$$

can be evaluated numerically. Asymptotic expansions can be made for large n or even for large m, although it is known that the latter converges very slowly.

More generally, provided f is standard normal, the density of the pivot $Z - \bar{Y}$ is

$$\sqrt{n}\int_{-\infty}^{+\infty} \varphi(\sqrt{n}(x-u))h(u)\,\mathrm{d}u, \tag{4.1}$$

where, as above, h denotes the density of $Z - \theta$. Laplace's method for the asymptotic evaluation of integrals shows that, to order $O(n^{-2})$, the density (4.1) equals

$$h(x)\{1 + \tfrac{1}{2}n^{-1}h''(x)/h(x)\}$$

and, letting $k(x) = \log h(x)$, this in turn is asymptotically equivalent to

$$\{1 + \tfrac{1}{2}n^{-1}k''(x)\}h(x + \tfrac{1}{2}n^{-1}k'(x)), \tag{4.2}$$

a form which will tie in with later derivations.

Example 4.3 *Autoregression with unknown mean.* Let $Y_1, \ldots, Y_n$ form a first-order Gaussian autoregression of known innovation variance σ^2 and correlation parameter ρ, and with unknown mean μ, Y_1 having the stationary distribution of the process, i.e. being normal with mean μ and variance $\sigma^2/(1-\rho^2)$. Let Z be the next observation, Y_{n+1}. The argument can be trivially adapted for predicting Y_{n+s} $(s > 0)$. Now, given $Y = y$,

$$\mathrm{E}(Z \mid Y = y) = \mu + \rho(y_n - \mu), \qquad \mathrm{var}\,(Z \mid Y = y) = \sigma^2.$$

Thus

$$z_\alpha(y; \mu) = \mu + \rho(y_n - \mu) + k_\alpha\sigma.$$

We have next to 'remove' the parameter μ. Now, the maximum likelihood estimate of μ is

$$\hat{\mu} = \{y_1 + (1-\rho)(y_2 + \cdots + y_{n-1}) + y_n\}/(n - n\rho + 2\rho).$$

Then we consider the pivot $Z - \rho Y_n - (1-\rho)\hat{\mu}$ which has mean zero. Its variance is most easily calculated by expressing it as a function of Z, Y_n and $Y_1 + (1-\rho)(Y_2 + \cdots + Y_{n-1})$ and noting that the first and third random variable are conditionally independent given the second. The result is that the variance of the pivot is

$$\sigma^2\left[1 + \left(\frac{1-\rho}{n - n\rho + 2\rho}\right)^2\{n - 1 + (1+\rho)^2\}\right] = \sigma^2 r_n^2(\rho),$$

say, so that the prediction limit is

$$c_\alpha(y) = \hat{\mu} + \rho(y_n - \hat{\mu}) + k_\alpha\sigma r_n(\rho).$$

Note that if $\rho \uparrow 1$ then $c_\alpha(y) \to y_n + k_\alpha\sigma$, as is clear from first principles, and that for $\rho \in (-1, 1)$ and n large, $r_n^2(\rho) \sim 1 + n^{-1}$.

In the present example, y_n is transitive and $(\hat{\mu}, y_n)$ is prediction sufficient.

Example 4.4 *Empirical Bayes with random effects model.* Here we consider one of the simplest problems of empirical Bayes form. Let $Y_1, \ldots, Y_n$ be independently normally distributed with variance σ_{w}^2 and means $\mu_1, \ldots, \mu_n$; and let $\mu_1, \ldots, \mu_n$ be independent and normally distributed with mean μ and variance σ_{b}^2. Thus the unconditional distribution of

Y_i is normal with mean μ and variance $\sigma_w^2 + \sigma_b^2$. Suppose we are interested in μ_1, so that $Z = \mu_1$. We regard σ_w^2 and σ_b^2 as known and μ as unknown. Now given $Y = y$ and letting $\tau^2 = (\sigma_w^{-2} + \sigma_b^{-2})^{-1}$, Z given y is normally distributed with $\mathrm{E}(Z \mid Y = y) = \tau^2(y_1/\sigma_w^2 + \mu/\sigma_b^2)$, $\mathrm{var}(Z \mid Y = y) = \tau^2$, from which $z_\alpha(y; \mu)$ follows directly. Note that the transitive statistic w in the general formulation is y_1 which is nearly unconnected with the efficient estimation of μ. The maximum likelihood estimate of μ is $\hat{\mu} = (Y_1 + \cdots + Y_n)/n$ and the covariance matrix of $(Z, Y_1, \hat{\mu})$ is easily obtained, for example

$$\mathrm{cov}(\hat{\mu}, Y_1) = (\sigma_w^2 + \sigma_b^2)/n, \qquad \mathrm{cov}(Z, Y_1) = \sigma_b^2.$$

Thus we find that

$$Z - \tau^2(Y_1/\sigma_w^2 + \hat{\mu}/\sigma_b^2)$$

has mean zero and variance $\tau^2\{1 + (\sigma_w^2/\sigma_b^2)n^{-1}\}$, from which prediction limits are obtained. Note that this gives simple answers directly obtained from first principles in the three special cases $n = 1, \sigma_w^2 \to 0, \sigma_b^2 \to 0$. Provided that $\sigma_w^2/\sigma_b^2 \ll n$, the effect of having to estimate μ is small.

The pair $(\hat{\mu}, y_1)$ constitutes a prediction sufficient statistic.

In applications, for each j there would typically be a number of independent observations with the same μ_j; we then take Y_j to be their mean. The variance σ_w^2 will, if the number of replicate observations is large, be estimated with a large number of degrees of freedom and to regard it as known may be quite reasonable. Allowance for errors of estimation in σ_b^2 is typically needed.

Example 4.5 *Random walk*. Suppose that a random walk of unknown drift μ is observed with error and that it is required to estimate the 'true' position of the process at the end of the period of observation. That is, we observe

$$Y_j = j\mu + (X_1 + \cdots + X_j) + U_j$$

$(j = 1, \ldots, n)$, where $\{X_1, \ldots, X_n; U_1, \ldots, U_n\}$ are independently normally distributed with zero mean and variances respectively σ_x^2, σ_u^2, which initially we regard as known; let $Z = n\mu + (X_1 + \cdots + X_n)$ be the value to be predicted. Note that this example is different from the above in that Z depends not only on unobserved random variables but also on the unknown parameter.

Let $e_n = (1, 2, \ldots, n)$ and $\lambda = \sigma_u^2/\sigma_x^2$, write $\sigma_x^2 C_n$ for the covariance matrix of the random walk $S_j = X_1 + \cdots + X_j$, so that the covariance matrix of Y is $\sigma_x^2(C_n + \lambda I_n) = \sigma_x^2 V_n$, say, and note that the covariance of Z with the vector having components S_j is $\sigma_x^2 e_n$.

Then conditionally on $Y = y$, Z is normally distributed with mean

$$\hat{\mu}_z = n\mu + e_n V_n^{-1}(y - \mu e_n)^{\mathrm{T}}.$$

To estimate μ we take the maximum likelihood estimate

$$\hat{\mu} = (e_n V_n^{-1} e_n^{\mathrm{T}})^{-1} e_n V_n^{-1} y^{\mathrm{T}},$$

leading to the consideration of

$$Z - n\hat{\mu} - e_n V_n^{-1}(y - \hat{\mu} e_n)^{\mathrm{T}} = Z - n\hat{\mu}.$$

Now

$$\text{var}\,(Z) = n\sigma_x^2, \qquad \text{cov}\,(Z, \hat{\mu}) = \sigma_x^2, \qquad \text{var}\,(\hat{\mu}) = (e_n V_n^{-1} e_n^{\mathrm{T}})^{-1}\sigma_x^2,$$

so that $Z - n\hat{\mu}$ has expectation zero and variance

$$n\sigma_x^2\{n(e_n V_n^{-1} e_n^{\mathrm{T}})^{-1} - 1\},$$

from which the required limits follow.

Thus $\hat{\mu}$ is prediction sufficient for Z. If λ were unknown there would be no exact reduction by transitivity, because of the dependence of V_n on λ.

Example 4.6 *Bivariate normal predictands*. Suppose that it is required to predict a pair of future values $Z = (Z_1, Z_2)$ from a normal distribution of mean μ. In some contexts it would be fruitful to specify the prediction in terms of $\frac{1}{2}(Z_1 + Z_2)$ and $Z_1 - Z_2$ or in terms of $\max(Z_1, Z_2)$ and $\min(Z_1, Z_2)$, but in the absence of such special conditions, note that for known μ the regions of high density are discs centred on (μ, μ). Such discs have the property that points outside have lower probability density than those inside. The radius r_α needed to achieve a specified coverage probability α is given via the exponential distribution of $(Z_1 - \mu)^2 + (Z_2 - \mu)^2$ as $r_\alpha = -2\log\alpha$.

If, when μ is unknown, we take a disc with centre $(\hat{\mu}, \hat{\mu})$ and radius r_a', the conditional coverage probability is $C_2(r_a'; \|\mu - \hat{\mu}\|)$, derived from the non-central chi-square distribution function C_2 with non-centrality parameter $\|\mu - \hat{\mu}\|$ and two degrees of freedom. The unconditional coverage probability is the expectation over the known semi-normal distribution of $\|\mu - \hat{\mu}\|$ and in principle an exact value of r_a' can be computed.

Note that if the underlying distribution is exponential the corresponding prediction region for a pair of values is a triangle.

5. Approximate predictive limits

So far we have distinguished between random variables and their observed values by upper- and lower-case letters, but from here on it is convenient to drop this distinction and use lower case for both.

Consistent with the discussion relating to formulae (2.1)–(2.4), for any given $\alpha \in (0, 1)$ we now seek an approximate solution $q_\alpha = q_\alpha(\theta) = q_\alpha(\theta, a)$ to the equation

$$\int G(q_\alpha$$

(\^\theta ,a);\theta {\kern 1}|{\kern 1}\^\theta , a)p(\^\theta ; \theta {\kern 1}|{\kern 1}a)\, {\rm d}\^\theta = \alpha .\eqno (5.1)

In (5.1), $G = G(z; \theta|\hat{\theta}, a)$ is the conditional distribution function of z given $(\hat{\theta}, a)$ and $p(\hat{\theta}; \theta|a)$ is the conditional density of $\hat{\theta}$ given a. Furthermore, $(\hat{\theta}, a)$ is supposed to be a prediction sufficient reduction of y with a ancillary, either exactly or to the appropriate

degrees of approximation. In applications $(\hat{\theta}, a)$ may typically be obtained by first determining a minimal prediction sufficient statistic, by means of the characterization in terms of minimal sufficiency mentioned in Section 3, followed by a smooth one-to-one transformation. For an illustration, see Example 7.4 below.

The asymptotic solution to be discussed is invariant with respect to transformations of the random variable Z to be predicted as well as to transformations of the parametrization given by θ, and it is generally correct to order $O(n^{-3/2})$ when y constitutes a random sample of order n.

To specify the solution we need the concept of mixed log-likelihood derivatives. Since $(\hat{\theta}, a)$ is assumed to be prediction sufficient it follows that it is sufficient, and we may therefore consider the log-likelihood function l for θ based on y as depending on y through $(\hat{\theta}, a)$ only, which we express by writing $l = l(\theta; \hat{\theta}, a)$. Our calculations will be conditional on the ancillary a which may therefore be considered as a constant. Mixed log-likelihood derivatives are now defined as the partial derivatives of any order with respect to components of either θ or $\hat{\theta}$. Denoting generic coordinates of θ and $\hat{\theta}$ by $\theta^r, \theta^s, \ldots$ and $\hat{\theta}^r, \hat{\theta}^s, \ldots$, with $r, s, \ldots$ running from 1 to d, the dimension of θ, we use the notation

$$l_{rs;t} = \partial^3 l / \partial\theta^r \partial\theta^s \partial\hat{\theta}^t, \qquad l_{r;st} = \partial^3 l / \partial\theta^r \partial\hat{\theta}^s \partial\hat{\theta}^t.$$

For a detailed discussion of properties and applications of mixed log-likelihood (or model) derivatives, see Barndorff-Nielsen and Cox (1994, Section 6.2).

If $q = q(\hat{\theta}, a, z, \theta)$ is any function of the data $(\hat{\theta}, a)$, the predictand z and the parameter θ, we write $\mathbf{q}$ for the quantity obtained from q by substituting θ for $\hat{\theta}$, i.e. $\mathbf{q} = q(\theta, a, z, \theta)$.

For use below, we note the expression

$$\mathbf{l}_{u;rs} + \mathbf{l}_{rsu} + \mathbf{l}_{ru;s} + \mathbf{l}_{su;r} = 0, \tag{5.2}$$

a special case of the balance relations for the observed likelihood yoke (cf. formula (5.13) of Barndorff-Nielsen and Cox 1994, p. 148).

Since the value of the ancillary a is kept fixed in our calculations, we shall often suppress a in our notation, for instance writing $q_\alpha(\hat{\theta})$ rather than $q_\alpha(\hat{\theta}, a)$ and $G(q_\alpha(\theta); \theta \mid \theta)$ rather than $G(q_\alpha(\theta, a); \theta \mid \theta, a)$.

The derivation of the approximate solution to (5.1) is given in the Appendix, as a special case of a somewhat more general problem. The resulting expression is defined indirectly as the solution q_α to the equation

$$G(q_\alpha(\theta); \theta \mid \theta) = \alpha - R(\theta). \tag{5.3}$$

The solution q_α approximates the solution of (5.1) to the appropriate order of $O(n^{-3/2})$. The correction term $R(\theta)$, which is of order $O(n^{-1})$ under repeated sampling, is given by

$$R(\theta) = Q(q_{o\alpha}(\theta); \theta), \tag{5.4}$$

where $q_{o\alpha} = q_{o\alpha}(\theta)$ is the solution to the first-order equation

$$G(q_{o\alpha}(\theta); \theta \mid \theta) = \alpha \tag{5.5}$$

and where, writing j^{rs} for a generic element of the inverse j^{-1} of the observed information

$j = [j_{rs}] = -[l_{rs}]$,

$$Q(z;\theta) = \tfrac{1}{2}\{\mathbf{H}_{rs} - \mathbf{H}_t \mathbf{l}_{urs}\mathbf{j}^{tu}\}\mathbf{j}^{rs}$$
$$= \tfrac{1}{2}\{\mathbf{H}_{rs} - \mathbf{H}_t(\mathbf{l}_{rsu} + \mathbf{l}_{ru;s} + \mathbf{l}_{su;r})\mathbf{j}^{tu}\}\mathbf{j}^{rs}. \tag{5.6}$$

Furthermore, the quantities $\mathbf{H}_r$ and $\mathbf{H}_{rs}$ are defined as follows. Let H_r and H_{rs} be given by

$$H_r = -G_{;r;} \tag{5.7}$$

$$H_{rs} = g^{-1}(G_{;r;} + G_{;;r})g_{;s;}[2] - G_{;rs;} - G_{;r;s}[2], \tag{5.8}$$

where g denotes the conditional density of z given $(\hat{\theta}, a)$ and where

$$\begin{aligned} g_{;r;} &= \partial g(z;\theta|\hat{\theta})/\partial\theta^r, \\ G_{;r;} &= \partial G(z;\theta|\hat{\theta})/\partial\theta^r, \qquad & G_{;;r} &= \partial G(z;\theta|\hat{\theta})/\partial\hat{\theta}^r \\ G_{;r;s} &= \partial^2 G(z;\theta|\hat{\theta})/\partial\theta^r\partial\hat{\theta}^s, \qquad & G_{;rs;} &= \partial^2 G(z;\theta|\hat{\theta})/\partial\hat{\theta}^r\partial\hat{\theta}^s. \end{aligned} \tag{5.9}$$

Moreover, [2] indicates the sum of two terms, obtained by permutation of the indices (r and s) involved. The quantities $\mathbf{H}_r$ and $\mathbf{H}_{rs}$ entering (5.6) are then obtained from (5.7) and (5.8) by substituting θ for $\hat{\theta}$.

If y and z are independent, so that G does not depend on $\hat{\theta}$ and a, then H_r and H_{rs} simplify, in obvious notation, to

$$H_r = -G_{;r} \qquad H_{rs} = g^{-1}G_{;r}g_{;s}[2] - G_{;rs}.$$

Note, incidentally, that if both the future observation z and the parameter θ are one-dimensional then $-G_{;r}$ is determined by substituting z by $q_{o\alpha}(\theta)$ and y by $(\hat{\theta}, a)$ in

$$-\partial G(z;\theta|y)/\partial\theta,$$

the latter being the 'fiducial density' for θ given z and conditional on y (cf., for instance, Pedersen 1978).

Without changing the order of approximation in (5.3), the quantities $\mathbf{j}^{rs}$ and $\mathbf{l}_{t;rs}$ can be replaced by their expected counterparts i^{rs} and $i_{t,rs} + i_{r,s,t}$, where $i_{t,rs} = \mathrm{E}\{l_t l_{rs}\}$ and $i_{r,s,t} = \mathrm{E}\{l_r l_s l_t\}$ (cf. Barndorff-Nielsen and Cox 1994, p. 160). In that case the correction term $R(\theta)$ changes to $\bar{R}(\theta) = \bar{Q}(q_{o\alpha}(\theta);\theta)$, where

$$\bar{Q}(z;\theta) = \tfrac{1}{2}\{\mathbf{H}_{rs} - \mathbf{H}_t(i_{u,rs} + i_{u,r,s})i^{tu}\}i^{rs}.$$

Like $R(\theta)$, the correction $\bar{R}(\theta)$ is invariant under reparametrizations. Note that if y and z are independent then it is not necessary to specify the ancillary a in order to calculate $\bar{Q}$.

6. Predictive distributions

Based on the prediction confidence limits $q_\alpha(\hat{\theta})$ defined in Section 5, we now introduce functions $\vec{G}(z|y)$ and $\vec{g}(z|y)$ which may be considered as a predictive distribution function

and a predictive density, respectively, for the unobserved random variable z with distribution function $G = G(z;\theta|y)$ and density $g(z;\theta|y)$. The idea is to think, for prediction purposes, of the unobserved random variable z as if it had a distribution function $\vec{G}(z|y)$ satisfying the equation

$$\vec{G}(q_\alpha(\hat{\theta})|y) = \alpha, \tag{6.1}$$

which is to hold for all $\alpha \in (0,1)$. In other words, $\vec{G}(c|y)$ is, for any real c, the prediction confidence with which we can state that $z \leq c$ in the sense of (2.1). The predictive density $\vec{g}$ is then, in principle, defined as the derivative of $\vec{G}(z|y)$ with respect to z.

To express $\vec{g}$ explicitly, to the relevant asymptotic accuracy, we differentiate (6.1) with respect to α, thus obtaining

$$\vec{g}(q_\alpha(\hat{\theta})|y) = \{q_\alpha(\hat{\theta})_{/\alpha}\}^{-1}.$$

We proceed to determine $q_\alpha(\theta)_{/\alpha}$ via differentiation through $q_{o\alpha}(\theta)$, the first-order approximation to $q_\alpha(\theta)$. Thus, rather than conceiving of $q_\alpha(\theta)$ as a function of α and θ, we think of it as a function of $q_{o\alpha}(\theta)$ and θ.

We have

$$q_\alpha(\theta)_{/\alpha} = q_\alpha(\theta)_{/q_{o\alpha}(\theta)} \cdot q_{o\alpha}(\theta)_{/\alpha}.$$

By definition, $G(q_{o\alpha}(\theta);\theta|\theta) = \alpha$, and hence

$$q_{o\alpha}(\theta)_{/\alpha} = g(q_{o\alpha}(\theta);\theta|\theta)^{-1}.$$

Furthermore, by differentiation with respect to $q_{o\alpha}(\theta)$ of the defining relation for $q_\alpha(\theta)$, i.e.

$$G(q_\alpha(\theta);\theta|\theta) = G(q_{o\alpha}(\theta);\theta|\theta) - Q(q_{o\alpha}(\theta);\theta), \tag{6.2}$$

we find

$$q_\alpha(\theta)_{/q_{o\alpha}(\theta)} = g(q_\alpha(\theta);\theta|\theta)^{-1}\{g(q_{o\alpha}(\theta);\theta|\theta) - \dot{Q}(q_{o\alpha}(\theta);\theta)\}. \tag{6.3}$$

Here and elsewhere, a dot above a function symbol means the derivative with respect to the first argument. We proceed to rewrite this expression by Taylor-expanding both $g(q_\alpha(\theta);\theta|\theta)$ and $G(q_\alpha(\theta);\theta|\theta)$ about $q_{o\alpha}(\theta)$, obtaining (to first order)

$$g(q_\alpha(\theta);\theta|\theta) = g(q_{o\alpha}(\theta);\theta|\theta) + (q_\alpha - q_{o\alpha})\dot{g}(q_{o\alpha}(\theta);\theta|\theta) \tag{6.4}$$

and

$$G(q_\alpha(\theta);\theta|\theta) = G(q_{o\alpha}(\theta);\theta|\theta) + (q_\alpha - q_{o\alpha})g(q_{o\alpha}(\theta);\theta|\theta). \tag{6.5}$$

In view of (6.2), the latter formula may be re-expressed as

$$q_\alpha - q_{o\alpha} = -g^{-1}Q, \tag{6.6}$$

where g and Q are short for $g(q_{o\alpha}(\theta);\theta|\theta)$ and $Q(q_{o\alpha}(\theta);\theta)$. Using (6.6) to eliminate $q_\alpha - q_{o\alpha}$ from (6.4) and inserting the resulting expression for $g(q_\alpha(\theta);\theta|\theta)$ in (6.3), we have

$$\{q_\alpha(\theta)_{/q_{o\alpha}(\theta)}\}^{-1} = \frac{1 - g^{-1}Q\dot{g}/g}{1 - g^{-1}\dot{Q}}.$$

Hence, letting

$$r(z;\theta|y) = Q(z;\theta)/g(z;\theta|y), \tag{6.7}$$

using formula (6.6) again and dropping a term of order $O(n^{-3/2})$, the predictive density may be expressed as

$$\vec{g}(z|y) = (1 + \hat{r}_{/z})g(z + \hat{r};\hat{\theta}|y). \tag{6.8}$$

Note, in particular, that the integral of $\vec{g}(z|y)$ with respect to z is, as one would hope, equal to 1.

It should also be noted that the above derivation of expression (6.8) for the predictive density $\vec{g}(z|y)$ does not rely on the explicit formula for Q given in Section 5. Essentially, all we are using is that $q_\alpha(\hat{\theta})$ satisfies equation (6.1) approximately and that an equation of the form (6.2) holds with $Q(q_{o\alpha}(\theta);\theta)$ as a (relatively small) remainder term. Thus, conceivably, the formula for $\vec{g}$ may be useful with other approximations of the prediction quantile $q_\alpha(\theta)$ than that considered in this paper.

7. Examples

In the following we denote the density of a normal distribution with mean μ and variance σ^2 by $\varphi(\cdot;\mu,\sigma^2)$.

Example 7.1 *Location models.* Suppose that y and z are independent and that both follow a one-dimensional location model with location parameter μ. Further, let the density for y be symmetric around μ, in which case $\mathbf{l}_{t;rs} = 0$. Then r, as given by (6.7), is of the simple form

$$r = \tfrac{1}{2} l_\mu(\mu;z)\hat{j}^{-1},$$

i.e. minus the score for μ based on z, and

$$\vec{g}(z|y) = \{1 + \tfrac{1}{2} l_{\mu\mu}(\hat{\mu};z)\hat{j}^{-1}\}g(z - \hat{\mu} - \tfrac{1}{2} l_\mu(\hat{\mu};z)\hat{j}^{-1}). \tag{7.1}$$

In particular, if $y \sim N(\mu, n^{-1}\sigma^2)$ and $z \sim N(\mu,\sigma^2)$, with σ^2 known, then

$$\vec{g}(z|y) = \varphi(z;y,\sigma^2(1 - \tfrac{1}{2}n^{-1})^{-2})$$

which, to order $O(n^{-1})$, agrees with the classical solution given by $z - y \sim N(0,\sigma^2(1 + n^{-1}))$; cf. Example 4.1.

Note further that if $y \sim N(\mu, n^{-1}\sigma^2)$ then (7.1) agrees to order $O(n^{-2})$ with the exact density of $z - \hat{\mu}$; cf. formula (4.2).

Example 7.2 *Autoregression with unknown mean.* Under the autoregression model considered in Example 4.3, $a = (y_1 - \hat{\mu}, \ldots, y_n - \hat{\mu})$ is ancillary and we have, writing

$$\xi = (1 - \rho)\mu + \rho(\hat{\mu} + a_n),$$

that

$$G(z;\mu\,|\,\hat{\mu},a)=\Phi(\sigma^{-1}(z-\xi)),\qquad g=-\sigma^{-1}\varphi(\sigma^{-1}(z-\xi))$$

and

$$G_{;\mu;}=-(1-\rho)k,\qquad G_{;;\hat{\mu}}=-\rho k,\qquad g_{;\mu;}=\sigma^{-2}(1-\rho)(z-\xi)k$$
$$G_{;\mu\mu;}=-\sigma^{-2}(1-\rho)^2(z-\xi)k,\qquad G_{;\mu;\hat{\mu}}=\sigma^{-2}\rho(1-\rho)(z-\xi)k.$$

Furthermore,

$$j=\sigma^{-1}(1-\rho)\{1+\rho+(n-1)(1-\rho)\}.$$

It follows that

$$r(z;\mu)=-\frac{1}{2}\left(n-1+\frac{1+\rho}{1-\rho}\right)^{-1}(z-\xi)$$

and (6.8) takes the form

$$\vec{g}(z\,|\,y)=\varphi\left(z;\hat{\xi},\sigma^2\left\{1-\frac{1}{2}\left(n-1+\frac{1+\rho}{1-\rho}\right)^{-1}\right\}^{-2}\right).$$

In other words, the predictive density for z is as if z given y followed the normal distribution with mean $\rho y_n+(1-\rho)\hat{\mu}$ and variance

$$\sigma^2\left\{1-\frac{1}{2}\left(n-1+\frac{1+\rho}{1-\rho}\right)^{-1}\right\}^{-2}=\sigma^2 s_n^2(\rho),$$

say. The mean is the same as that of the 'exact' solution, considered in Example 4.3, while the ratio of the variances, $s_n^2(\rho)/r_n^2(\rho)$, tends to 1 for $\rho\uparrow 1$ and behaves like $1+O(n^{-2})$ for $n\to\infty$.

Example 7.3 *Empirical Bayes with random effects model.* In the situation of Example 4.4, the maximum likelihood estimate is $\hat{\mu}=\bar{y}$ and the vector $a=(y_1-\hat{\mu},\ldots,y_n-\hat{\mu})$ is ancillary and independent of $\hat{\mu}$. Writing

$$\tau^2=(\sigma_{\mathrm{w}}^{-2}+\sigma_{\mathrm{b}}^{-2})^{-1}$$
$$\xi=\tau^2\{(\hat{\mu}+a_1)/\sigma_{\mathrm{w}}^2+\mu/\sigma_{\mathrm{b}}^2\}$$

for the conditional variance and mean of μ_1 given y, we have

$$G(\mu_1;\mu\,|\,\hat{\mu},a)=\Phi(\tau^{-1}(\mu_1-\xi))$$

and hence $g=\tau^{-1}\varphi(\tau^{-1}(\mu_1-\xi))$ and

$$G_{;\mu;}=-\tau^2\sigma_{\mathrm{b}}^{-2}g,\qquad G_{;;\hat{\mu}}=-\tau^2\sigma_{\mathrm{w}}^{-2}g,\qquad g_{;\mu;}=(\mu_1-\xi)\sigma_{\mathrm{b}}^{-2}g$$
$$G_{;\mu\mu;}=-(\mu_1-\xi)\tau^2\sigma_{\mathrm{b}}^{-4}g,\qquad G_{;\mu;\hat{\mu}}=-(\mu_1-\xi)\tau^2\sigma_{\mathrm{b}}^{-2}\sigma_{\mathrm{w}}^{-2}g$$

(recall the convention for the use of bold letters introduced in Section 5).

Noting also that the observed information on μ is $n(\sigma_w^2 + \sigma_b^2)^{-1}$, we obtain

$$r(\mu_1; \mu) = -\tfrac{1}{2} n^{-1} (\mu_1 - \xi)(\sigma_w^2/\sigma_b^2).$$

It follows that the predictive density (6.8) is

$$\vec{g}(\mu_1 | y) = \varphi(\mu_1; \hat{\xi}, \tau^2 \{1 - \tfrac{1}{2}(\sigma_w^2/\sigma_b^2) n^{-1}\}^{-2}). \tag{7.2}$$

This agrees, to order $O(n^{-2})$, with the solution derived in Example 4.4.

It is noteworthy that, while the asymptotic arguments behind the derivation of the general expression (6.8) for $\vec{g}$ relate to $n \to \infty$, the way that σ_w^2 and σ_b^2 enter formula (7.2) is nicely in line with the exact solution in Example 4.4. In this connection, note that, on a priori grounds it could be seen that the coefficient of n^{-1} must be dimensionless and therefore a function of σ_w^2/σ_b^2 and, moreover, that it must reduce to zero if $\sigma_w^2 = 0$ when the leading term is exact.

As the final example in this section we consider a case that does not admit an 'exact' solution.

Example 7.4 *Autoregression with unknown regression coefficient.* Let $y_0, y_1, \ldots, y_n, \ldots$ be an autoregressive process with $y_0 = 0$ and $y_i | y_{i-1} \sim N(\rho y_{i-1}, \sigma^2)$, the variance σ^2 being taken as known, for simplicity. Suppose that $y = (y_1, \ldots, y_n)$ has been observed and that it is desired to predict y_{n+1}.

To determine the minimal prediction sufficient statistic we consider the conditional model for y given y_{n+1} whose probability density function is of the form

$$A(\rho, y_{n+1}) B(y) \exp\{\varphi(\rho, y_{n+1}) \cdot t(y)\}, \tag{7.3}$$

with

$$\varphi(\rho, y_{n+1}) = \sigma^{-2}(\rho, y_{n+1}\rho, -\tfrac{1}{2}\rho^2) \tag{7.4}$$

$$t(y) = \left(\sum_{i=1}^{n} y_{i-1} y_i, y_n, \sum_{i=1}^{n+1} y_{i-1}^2 \right). \tag{7.5}$$

Considering both ρ and y_{n+1} as parameters, we have that (7.3) constitutes a $(3, 2)$ exponential model with (7.5) as minimal sufficient statistic. Hence $t(y)$ or, equivalently

$$u = \left(\sum_{i=1}^{n} y_{i-1} y_i, \sum_{i=1}^{n} y_{i-1}^2, y_n \right) \tag{7.6}$$

is minimal prediction sufficient for y_{n+1}. Note that (7.6) consists of the minimal sufficient statistic v based on y, i.e. $v = \left(\sum_{i=1}^{n} y_{i-1} y_i, \sum_{i=1}^{n} y_{i-1}^2 \right)$, and the minimal transitive statistic y_n.

Next we transform u to $(\hat{\rho}, a_0, a_1)$ where a_0 is the score ancillary (or Efron–Hinkley ancillary; cf. Barndorff-Nielsen and Cox 1994, Section 8.2) for the $(2, 1)$ exponential model determined by y and where

$$a_1 = \Phi(y_n; 0, \sigma^2(1 - \hat{\rho}^{2n})/(1 - \hat{\rho}^2)). \tag{7.7}$$

In defining a_1 we have used the fact that the marginal distribution of y_n is $N(0, \sigma^2(1-\rho^{2n})/(1-\rho^2))$, ensuring that a_1 is indeed asymptotically distribution constant. The formula for a_0 is

$$a_0 = (\hat{j}/\hat{i} - 1)/\hat{\gamma} \tag{7.8}$$

where $j = \sigma^{-2}\sum_{i=1}^{n} y_{i-1}^2$ is the observed information based on y, i the expected information and γ^2 the variance of j/i.

Writing

$$g = (2\pi)^{-1/2}\sigma^{-1}\exp\{-\tfrac{1}{2}(y_{n+1} - \rho y_n)^2\},$$

the quantities (5.9) needed to calculate Q (cf. formulae (6.7), (6.8) and (5.6)) are

$$g_{;\rho;} = \sigma^{-2}y_n(y_{n+1} - \rho y_n)g, \qquad G_{;\rho;} = -y_n g, \qquad G_{;;\hat{\rho}} = -\rho g y_{n/\hat{\rho}},$$
$$G_{;\rho\rho;} = -\sigma^{-2}y_n^2(y_{n+1} - \rho y_n)g, \qquad G_{;\rho;\hat{\rho}} = -\{1 + \sigma^{-2}\rho y_n(y_{n+1} - \rho y_n)\}g y_{n/\hat{\rho}}.$$

Furthermore, we have to determine $l_{\rho;\hat{\rho}\hat{\rho}}$. For this we note that the score is given by

$$l_\rho = \sigma^{-2}\sum_{i=1}^{n} y_{i-1}(y_i - \rho y_{i-1}) = (\rho - \hat{\rho})\hat{j}.$$

It follows that

$$l_{\rho;\hat{\rho}\hat{\rho}} = -2\hat{j}_{/\hat{\rho}} + (\rho - \hat{\rho})\hat{j}_{/\hat{\rho}\hat{\rho}}$$

and hence

$$\hat{\mathbf{I}}_{\rho;\hat{\rho}\hat{\rho}} = -2\hat{j}_{/\hat{\rho}}.$$

Collecting terms, we find

$$H_\rho = y_n g, \qquad H_{\rho\rho} = \{2y_{n/\hat{\rho}} - \sigma^{-2}y_n^2(y_{n+1} - \rho y_n)\}g$$

and

$$r = \hat{j}^{-1}\{y_{n/\hat{\rho}} - \tfrac{1}{2}\sigma^{-2}y_n^2(y_{n+1} - \hat{\rho}y_n) - \hat{j}_{/\hat{\rho}}y_n\}.$$

The derivatives $y_{n/\hat{\rho}}$ and $\hat{j}_{/\hat{\rho}}$ are for a_0 and a_1 held fixed and are determined by differentiation of (7.7) and (7.8). Writing $\tau^2 = \sigma^2(1-\rho^{2n})/(1-\rho^2)$, we find

$$y_{n/\hat{\rho}} = \tfrac{1}{2}y_n\hat{\tau}^{-2}(\hat{\tau}^2)_{/\hat{\rho}}$$

and

$$\hat{j}_{/\hat{\rho}} = \hat{j}\hat{i}^{-1}\hat{i}_{/\hat{\rho}} + \tfrac{1}{2}(\hat{j} - \hat{i})\hat{\gamma}^{-2}(\hat{\gamma}^2)_{/\hat{\rho}}.$$

8. Many nuisance parameters

Models with a large number of nuisance parameters raise difficulties for all forms of asymptotic statistical inference, although these difficulties are often best expressed via an

empirical Bayes formulation. We consider briefly problems where this is not pertinent. Let the observed random variable Y, of dimension n, have a distribution depending on incidental parameters $\chi_1, \ldots, \chi_{m_n}$ and a structural parameter ψ. In the cases of interest here, as the dimension n of Y increases so does m_n and consistent estimation of ψ is possible, whereas the same is not true in general for the components of $\chi = (\chi_1, \ldots, \chi_{m_n})$.

There are two broad classes of prediction problems according as the random variable Z to be predicted has a distribution depending only on ψ or one depending on some of the χs as well as possibly on ψ. In the latter case there are difficulties in a general treatment, especially when Z depends only on a small number of the components of χ which may be poorly estimated. Use of the best available estimates of the bias and variance of the estimated χs will be needed if the route of Section 4 is followed. In one familiar instance there is an exact solution.

Example 8.1 *Normal-theory linear model.* Let Y follow the normal-theory linear model $\mathrm{E}(Y) = x\beta$, where β, which corresponds to χ in the general formulation, is of high dimension, and where the error variance σ^2 corresponds to ψ. If the predictand Z is a future observation with $\mathrm{E}(Z) = x_0\beta$, where x_0 is a given covariate vector, then for known (β, σ^2), the prediction limits for Z have the form $x_0\beta + k_a\sigma$; whereas if σ is known and β unknown the limits are

$$x_0\beta + k_\alpha\sigma\{1 + x_0^{\mathrm{T}}(x^{\mathrm{T}}x)^{-1}x_0\}^{1/2};$$

and with σ also unknown and replaced by the residual root mean square, k_α is replaced in 'exact' theory by the Student t_α. The two key issues, in terms of general theory, are that it is entirely possible that the second term in the square root is as great as or greater than 1 and that σ is not estimated by ordinary maximum likelihood.

There are many generalizations for linear and nonlinear models inside and outside the class of exponential families. We consider just one instance.

Example 8.2 *Prediction of Weibull distributed failure times.* Suppose that Y_{ij} $(i = 1, \ldots, m; j = 1, \ldots, r)$ have independent Weibull distributions with common index ψ and separate rate parameters $\chi_1, \ldots, \chi_m$, i.e. Y_{ij} has the survivor function $\exp\{-(\chi_i y)^\psi\}$. Suppose that Z, the value to be predicted, has the distribution with rate parameter χ_1. The most challenging situation has r small and m large. Given the parameters, the upper α prediction limit for Z is

$$z_\alpha(\chi, \psi) = \chi_1^{-1}(-\log\alpha)^{1/\psi}.$$

For known ψ we have a simple transformation of Example 4.2, with

$$\hat{\chi}_{1\psi}^{-1} = \Sigma Y_{ij}^{\psi}/r,$$

so that an initial approximation to the prediction limit is

$$\{(\Sigma Y_{1j}^{\tilde{\psi}}/r)(-\log\alpha)\}^{1/\tilde{\psi}},$$

$\tilde{\psi}$ being a suitable estimate of ψ.

There are now a number of difficulties which we shall not address in detail. First, especially if r is small compared with m, $\tilde{\psi}$ should be based on an adjusted log-likelihood function and the quantities associated with that will be required. Even if we were to use ordinary maximum likelihood to estimate ψ, some extensive calculations would be required. A useful practical solution if r is small and m is large is to note that errors in estimating ψ will be negligible compared with those in estimating χ_1. Hence we may apply Example 4.2 as if ψ were known, i.e. use the estimate $(\Sigma Y_{1j}^{\hat{\psi}}/r)^{1/\hat{\psi}}$ multiplied by the upper α point of the variance ratio distribution with $(2, 2r)$ degrees of freedom.

9. Discrete predictands

There are some special considerations when the variable to be predicted is discrete, shown in most extreme form by the prediction of a binary outcome Z specified by a model in which

$$P(Z = 1) = p(\theta), \qquad P(Z = 0) = 1 - p(\theta),$$

and based on observations of a random variable Y with log-likelihood $l(\theta)$ and maximum likelihood estimate $\hat{\theta}$. We suppose that Y and Z are independent.

We can take two broad approaches to the prediction of Z. The first and more cautious is to provide upper and lower confidence limits for the parametric function $p(\theta)$. The second, which gives a simpler answer and which may be more satisfactory in a genuinely repetitive situation, is to find a function $\tilde{p}(y)$ such that, exactly or approximately,

$$\mathrm{E}\{Z - \tilde{p}(y); \theta\} = 0,$$

where the expectation is over Y and Z.

Thus we need an unbiased estimate of $p(\theta)$, and $p(\hat{\theta})$ is a natural first approximation. Sometimes an 'exact' form can be obtained for $\tilde{p}(y)$, but for a general asymptotic discussion we modify $p(\hat{\theta})$. This may be best done by writing

$$\tilde{p}(y) = p(\hat{\theta} + \hat{\epsilon}),$$

where ϵ is to be determined. Expansion similar to that leading to (2.3) gives

$$\epsilon(\theta)\nabla^{\mathrm{T}} p(\theta) = -b(\theta)\nabla^{\mathrm{T}} p(\theta) - \tfrac{1}{2}\,\mathrm{tr}\,\{v(\theta)\nabla\nabla^{\mathrm{T}} p(\theta)\},$$

where b and v denote the bias and variance of $\hat{\theta}$. This is a scalar equation for what in general is a vector $\epsilon(\theta)$. There are thus many solutions; often a convenient one will be obtained by setting all the components of $\epsilon(\theta)$ except one equal to 0.

Example 9.1 *Poisson process.* Let Y have a Poisson distribution of large mean $\mu = nt_0\theta$, corresponding to the number of points in a Poisson process with a total 'exposure' time nt_0. Suppose we are interested in the event that there are no points in a further independent period t_0, so that $p(\theta) = \exp(-t_0\theta)$. The crude estimate is $p(\hat{\theta}) = \mathrm{e}^{-y/n}$ and the 'exact' solution, obtained by Rao–Blackwellization, is

$$\tilde{p}(\hat{\theta}) = (1 - n^{-1})^y.$$

The asymptotic solution is

$$\exp\left\{-\frac{1}{n}\left(1+\frac{1}{2n}\right)y\right\}$$

which agrees with $\tilde{p}(\hat{\theta})$ to order n^{-2}.

10. Some open problems

While the procedure for prediction given in this paper is in some sense very general, nevertheless the illustrative examples discussed are all relatively simple. Quite a number of further issues thus remain, some concerned with general principles and others with the difficulties of handling specific problems.

Prominent among the former are prediction of quantities that depend not only on unobserved random variables but also on all or part of the unknown parameters (a simple case is considered in Example 4.5), and the need to handle situations in which an adjusted profile likelihood is a basis for estimating unknown parameters, preferable to ordinary likelihood. Expansions analogous to those of Section 5 and the Appendix are needed. Further general issues relate to the approximate ancillary statistics required in the general development and which, for some at least of the specific examples, need explicit formulation. How critical is the precise formulation? (Not very, we surmise.)

Another point concerns the relation with a Bayesian approach. What is the Bayes prior, if any, that would generate $\vec{g}(z\,|\,y)$ as posterior?

Among the more specific issues needing development are tractable methods for handling problems with several unknown parameters, for example the first-order autoregression with all parameters unknown; the relation with calibration (cf., for instance, Brown 1993); the relation with state-space modelling, of which Example 4.4 is a very special case; prediction of variance changes (volatility) in the kind of nonlinear models (ARCH, GARCH, SV, ...) useful in mathematical finance (see the review paper by Shephard 1995); and the special problems of prediction in spatial and spatial-temporal models, for example of rainfall fields.

A major extension would be required to deal with semi-parametric settings.

Appendix. Derivation of formulae (5.3)–(5.9)

By the technique of invariant Taylor series in terms of the observed likelihood yoke (cf. Barndorff-Nielsen and Cox 1994, Section 5.6), a function f of θ can be expanded as

$$\begin{aligned} f(\hat{\theta}) = {} & f(\theta) + f_r(\theta)\mathbf{j}^{rs} l_s(\theta;\hat{\theta}) \\ & + \tfrac{1}{2}\{f_{rs}(\theta) - f_t(\theta)\mathbf{j}^{tu}\mathbf{I}_{u;rs}\}\mathbf{j}^{rv}\mathbf{j}^{sw} l_v(\theta;\hat{\theta}) l_w(\theta;\hat{\theta}) + \cdots . \end{aligned} \tag{A.1}$$

Here, as previously, we have suppressed the dependence on the ancillary in the notation, and the quantities in bold are as defined in Section 5.

A variety of problems in prediction and estimation can be formulated as follows. An exact or approximate solution q is sought to the equation

$$\int K(q(\hat{\theta}, a); \theta; \hat{\theta} \mid a) p(\hat{\theta}; \theta \mid a) \, \mathrm{d}\hat{\theta} = \eta(\theta). \tag{A.2}$$

Here $K = K(\gamma; \theta; \hat{\theta})$ and η are known functions, q is a real-valued function of θ, and $p(\hat{\theta}; \theta \mid a)$ is the conditional density of $\hat{\theta}$ given a. That is, we require a conditionally (on a) unbiased estimate of $\eta(\theta)$, the estimate to be a function of the maximum likelihood estimate $\hat{\theta}$. We shall derive an approximate solution to this problem. The solution is invariant with respect to transformations of the parametrization given by θ and is generally correct to order $O(n^{-3/2})$ when y constitutes a random sample of order n. The solution presupposes that K is a monotone function of its first argument γ (for which the function q has been inserted above).

In the prediction context $\eta(\theta)$ will typically be constant, equal to a confidence coefficient α, and $K(\cdot\,; \theta; \hat{\theta} \mid a)$ will be of the form

$$K(\cdot\,; \theta; \hat{\theta} \mid a) = G(\cdot\,; \theta \mid \hat{\theta}, a), \tag{A.3}$$

where $G(\cdot\,; \theta \mid \hat{\theta}, a)$ is the conditional distribution function of the random variable z to be predicted, conditional on $(\hat{\theta}, a)$ or, equivalently, on y. When y and z are independent (A.3) takes the form $K(\cdot\,; \theta) = G(\cdot\,; \theta)$.

Applying (A.1) to $K(q(\hat{\theta}); \theta; \hat{\theta})$ considered as a function of $\hat{\theta}$ (and where, as before, we have suppressed the dependence on the ancillary) and then taking the mean under $p(\hat{\theta}; \theta \mid a)$, we obtain the approximation

$$\eta(\theta) = K(q(\theta); \theta; \theta) + \tfrac{1}{2} \{ \mathbf{H}_{rs} - \mathbf{H}_t \mathbf{l}_{u;rs} \mathbf{j}^{tu} \} \mathbf{j}^{rs}, \tag{A.4}$$

where

$$H_r = k q_r + K_{;;r} \tag{A.5}$$

$$H_{rs} = \dot{k} q_r q_s + k q_{rs} + k_{;;s} q_r[2] + K_{;;rs}. \tag{A.6}$$

In arriving at (A.4) we have used the fact that to first order $\mathbf{j} = i$, the latter being the expected information on θ determined by y. The notation employed is, moreover, as follows:

$$k = \partial K(\gamma; \theta; \hat{\theta}) / \partial \gamma, \qquad \dot{k} = \partial k / \partial \gamma,$$

$$q_r = \partial q / \partial \theta^r, \qquad K_{;;r} = \partial K(\gamma; \theta; \hat{\theta}) / \partial \hat{\theta}^r,$$

etc., with the further convention that in (A.4) the quantity $q(\theta)$ has been inserted for γ (while the bold symbols indicate that θ has been substituted for $\hat{\theta}$; cf. Section 5).

In order to eliminate q_r and q_{rs} from (A.5) and (A.6), we note that to first order (and we only need accuracy to that order),

$$\eta(\theta) = K(q(\theta); \theta; \theta), \tag{A.7}$$

so that

$$\eta_r = \mathbf{k} q_r + \mathbf{K}_{;r;} + \mathbf{K}_{;;r} \tag{A.8}$$

and

$$\eta_{rs} = \dot{\mathbf{k}} q_r q_s + \mathbf{k}_{;s;} q_r[2] + \mathbf{k}_{;;s} q_r[2] + \mathbf{k} q_{rs} + \boldsymbol{K}_{;rs;} + \boldsymbol{K}_{;r;s}[2] + \boldsymbol{K}_{;;rs}, \tag{A.9}$$

where

$$\begin{aligned} \eta_r &= \partial\eta/\partial\theta^r, \qquad \eta_{rs} = \partial^2\eta/\partial\theta^r\theta^s, \\ k_{;r;} &= \partial k(\gamma;\theta;\hat{\theta})/\partial\theta^r, \qquad k_{;;r} = \partial k(\gamma;\theta;\hat{\theta})/\partial\hat{\theta}^r, \\ K_{;r;} &= \partial K(\gamma;\theta;\hat{\theta})/\partial\theta^r, \qquad K_{;;rs} = \partial^2 K(\gamma;\theta;\hat{\theta})/\partial\hat{\theta}^r\partial\hat{\theta}^s, \text{ etc.} \end{aligned} \tag{A.10}$$

Solving equations (A.8) and (A.9) for q_r and q_{rs} and inserting in (A.5) and (A.6), we finally obtain

$$\mathbf{H}_r = \eta_r - \mathbf{K}_{;r;} \tag{A.11}$$

$$\mathbf{H}_{rs} = \eta_{rs} - \mathbf{k}^{-1}\eta_r \mathbf{k}_{;s;}[2] + \mathbf{k}^{-1}(\mathbf{K}_{;r;} + \mathbf{K}_{;;r})\mathbf{k}_{;s;}[2] - \mathbf{K}_{;rs;} - \mathbf{K}_{;r;s}[2]. \tag{A.12}$$

The announced approximate solution to equation (A.2) is obtained by solving (A.4) for $q = q(\theta) = q(\theta, a)$, with $\mathbf{H}_r$ and $\mathbf{H}_{rs}$ given by (A.11) and (A.12).

In applications to prediction the function η is usually constant, equal to an α in the interval $(0, 1)$, and K is of the form (A.3); then (A.11) and (A.12) become

$$\mathbf{H}_r = -\mathbf{G}_{;r;} \tag{A.13}$$

$$\mathbf{H}_{rs} = \mathbf{g}^{-1}(\mathbf{G}_{;r;} + \mathbf{G}_{;;r})\mathbf{g}_{;s;}[2] - \mathbf{G}_{;rs;} - \mathbf{G}_{;r;s}[2]. \tag{A.14}$$

We then denote the solution to (A.4) by $q_\alpha(\theta)$ where, again, we have suppressed the possible dependence on a. The formulae (A.4) and (A.13)–(A.14) are reproduced as (5.3) and (5.8)–(5.9) in Section 5.

Acknowledgements

A major part of this paper was established during a stay at CIMAT (Guanajuato, Mexico) in early 1995, and it is a pleasure to thank CIMAT and our host, Victor Perez-Abreu, for excellent working conditions and warm hospitality. We are also grateful to Paolo Vidoni and to the referees for helpful comments on an earlier version of the paper. The work behind the paper was supported by NATO grant 5-2-05/RG 930085.

References

Aitchison, J. and Dunsmore, I.R. (1975) *Statistical Prediction Analysis*. Cambridge: Cambridge University Press.

Atwood, C.L. (1984) Approximate tolerance intervals, based on maximum likelihood estimates. *J. Amer. Statist. Assoc.*, **79**, 459–465.

Bahadur, R.R. (1954) Sufficiency and statistical decision functions. *Ann. Math. Statist.*, **25**, 423–462.

Barndorff-Nielsen, O.E. (1978) *Information and Exponential Families*. Chichester: Wiley.

Barndorff-Nielsen, O.E. (1981) Likelihood prediction. *Sympos. Math.*, **XXV**, 11–24.
Barndorff-Nielsen, O.E. (1995) Stable and invariant adjusted profile likelihood and directed likelihood for curved exponential models. *Biometrika*, **82**, 489–499.
Barndorff-Nielsen, O.E. and Cox, D.R. (1994) *Inference and Asymptotics*. London: Chapman & Hall.
Barndorff-Nielsen, O.E. and Skibinsky, M. (1963) Adequate subfields and almost sufficiency. Applied Mathematics Publication 329, Brookhaven National Laboratory, New York.
Basu, A. and Harris, I.R. (1994) Robust predictive distributions for exponential families. *Biometrika*, **81**, 790–794.
Bjørnstad, J.F. (1990): Predictive likelihood: a review. (With discussion.) *Statist. Sci.*, **5**, 242–265.
Brown, P.J. (1993) *Measurement, Regression and Calibration*. Oxford: Oxford University Press.
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. (1973) Prediction intervals and empirical Bayes confidence intervals. In J. Gani (ed.), *Perspectives in Probability and Statistics*, pp. 47–55. London: Academic Press.
Fisher, R.A. (1935) The fiducial argument in statistical inference. *Ann. Eugenics*, **6**, 391–398.
Fisher, R.A. (1956) *Statistical Methods and Scientific Inference*. London: Oliver and Boyd.
Geisser, S. (1993) *Predictive Inference: An Introduction*. New York: Chapman & Hall.
Guttman, I. (1970) *Statistical Tolerance Regions: Classical and Bayesian*. London: Griffin.
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.
Kolmogorov, A. (1942) Sur l'estimation statistique des paramètres de la loi de Gauss. *Bull. Acad. Sci. URSS, Ser. Math.*, **6**, 3–32.
Komaki, F. (1996) On asymptotic properties of predictive distributions. *Biometrika*, **83**, 299–314.
Lauritzen, S.L. (1974) Sufficiency, prediction and extreme models. *Scand. J. Statist.*, **1**, 128–134.
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.
Pedersen, J.G. (1978) Fiducial inference. *Internat. Statist. Rev.*, **46**, 147–170.
Shephard, N. (1995) Statistical aspects of ARCH and stochastic volatility. In D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen (eds), *Likelihood, Time Series with Econometric and Other Applications*, pp. 1–67. London: Chapman & Hall.
Skibinsky, M. (1967) Adequate subfields and sufficiency. *Ann. Math. Statist.*, **38**, 155–161.
Vidoni, P. (1995) A simple predictive density based on the p^*-formula. *Biometrika*, **82**, 855–864.

Received June 1995 and revised April 1996.