Predictive sample reuse techniques for censored data

S. GEISSER
University of Minnesota

SUMMARY

Predictive sample reuse methods usually applied in low structure aparametric paradigms are shown to be useful in certain high structure situations when conjoined with a Bayesian approach. Particular attention is focused on the incomplete data situation for which two alternative sample reuse approaches are devised. The first involves differential weighting and the second a recursive sample reuse algorithm. There are applied to censored exponential survival data. The algorithmic approach appears to be preferable from both a computational and modelling viewpoint.

Keywords: BAYES, CONDITIONAL PREDICTIVE FUNCTION, CENSORED DATA, DISCREPANCY MEASURE, FUNGIBLE, EMPIRICAL BAYES, MAXIMUM LIKELIHOOD, METHOD OF MOMENTS, PREDICTIVE DISTRIBUTION, PREDICTIVE SAMPLE REUSE, PSEUDO-OBSERVATIONS, SAMPLE REUSE ALGORITHM.

1. PREDICTIVE SAMPLE REUSE

The predictive sample reuse (PSR) method was presented in a variety of detailed forms, Geisser (1974, 1975a), Stone (1974). Here we shall delineate it in a very simple manner appropriate to the particular applications that flow from it under discussion in later sections.

Suppose we have a set of observations $\mathbf{x}^{(N)} = (x_1,\ldots,x_N)$ and we are interested in predicting a future observation from the process generating observations of this kind. We further assume a predictive function used to forecast a potentially observable value,

$$x_{N+1} = f(\mathbf{x}^{(N)}, \alpha) \tag{1.1}$$

where α is defined as some unknown constant or set of such unknowns whose domain is Ω. Next we define a discrepancy function

$$D(\alpha) = D(d_1,\ldots,d_N, \Sigma) \tag{1.2}$$

where $d_j = d(x_j, f_j)$ represents a discrepancy between the observed value x_j and $f_j = f_j(\mathbf{x}_j^{(N-1)}, \alpha)$ which is defined as in (1.1) except that x_j has been deleted from f and $\Sigma = \Sigma(\alpha, \mathbf{x}^{(N)})$ represents some scheme of weighting the various d_j singly or jointly. For example

$$D(\alpha) = \Sigma_{j=1}^{N} a_j(\alpha) d(x_j, f_j) \tag{1.3}$$

where $a_j(\alpha)$ is the weight assigned to the j^{th} discrepancy or

$$D(\alpha) = \mathbf{d}'\Sigma\mathbf{d} \tag{1.4}$$

for $\mathbf{d}' = (d_1,\ldots,d_N)$ would be two such schemes. In most cases fungible[1] data would lead to $a_j(\alpha) = N^{-1}$ or $\Sigma = N^{-1}\mathrm{I}$. Then $D(\alpha)$ is minimized for values of α restricted to Ω which we assume yields a unique value $\hat{\alpha}$. This leads to the predictor

$$\hat{\mathbf{x}}_{N+1} = f(x^{(N)}, \hat{\alpha}) = \hat{f}. \tag{1.5}$$

For a more detailed exposition of the method involving multiple observational omissions and various schemata of omission, as well as applications, see Geisser (1974, 1975a, 1976).

In applying this method to survival or realiability data, it is quickly apparent that an inherent deficiency exists. The method as stated depends on the full knowledge of the sample values. But for this type of problem quite often our knowledge for a portion of the sample is restricted by the fact that the observations were censored at particular values. In order to remedy this lack of knowledge of fully observed values we introduce pseudo-observations. They depend on α and are determined from defined conditional predictive functions. Two procedures utilizing a pseudo-observation approach are presented. The first proposal substitutes the pseudo-observations into the discrepancy measure prior to minimization. This leads rather naturally to considering schemes whereby the censored observations are weighted differently than uncensored ones as opposed to previous applications where $a_j(\alpha) = 1$ on the basis that the data were inherently fungible. Of course, there

[1] We use the term fungible to extend the notion of exchangeable to data that are not necessarily a realization of a random set of variables. For random variables the terms are equivalent. The extension, though ill defined, conveys an attitude that one could take towards observable data for which it is inappropriate to assume that they were necessarily generated by a random process.

could arise situations where a sample of uncensored observations may require different weights because of a decision as to their treatment or a model for their generation. Here, even though we start with a scheme that treats the observations fungibly, the approach of fitting the censored observations into the predictive sample reuse framework naturally induces consideration of differential weighting schemes.

A second proposal involves the substitution of the pseudo-observations into the solutions as if all the values were fully observed and solving the requisite algorithm. Let $\mathbf{x}^{(d)} = (x_1,\ldots,x_d)$ and $\mathbf{x}^* = (x_{d+1},\ldots,x_N)$ represent respectively the completely and partially observed data sets with the understanding that the observable x_j for $j > d$ represents incomplete information of some kind on an observable entity, or when appropriate, a realization of the random variable X_j. Let $\mathbf{y} = (y_{d+1},\ldots,y_N)$ represent the set of values which would have been observed but were partially observed as $\mathbf{x}^*$; i.e. the fully realized value of X_j would have been y_j,but we were only able to record the partially observed value x_j, $j > d$. We then compute a complete solution for α, say

$$\tilde{\alpha} = \tilde{\alpha}(\mathbf{x}^{(d)},\mathbf{y}) \tag{1.6}$$

in the usual fashion, as in the fully observed case, but as a function of $\mathbf{y}$. But we need values for y_j the components of $\mathbf{y}$. We now assume a conditional predictive function for the components of $\mathbf{y}$,

$$y_j = \hat{x}_j'(\mathbf{x}^{(d)},\mathbf{x}^*, \alpha) = x_j'(\alpha); \quad j > d. \tag{1.7}$$

Now let $\mathbf{x}^*(\alpha)$ represent the set of values inserted for $\mathbf{y}$; i.e. for each component y_j we insert $x_j'(\alpha)$ in (1.6). Lastly we then have the algorithm

$$\alpha = \tilde{\alpha}(\mathbf{x}^{(d)}, \mathbf{x}^*(\alpha)) \tag{1.8}$$

which needs to be solved for α. Call the solution $\hat{\alpha}$ and one then uses this either to predict a future observation conditionally or unconditionally.

2. AN APPLICATION--UNCENSORED CASE

The application of these ideas for forecasting in a particular survival or reliability data situation will be presented where the predictive sample reuse technique is used in partial conjunction with a Bayesian approach. Initially we shall assume the entire fine structure of an exponential survival distribution *cum* gamma prior distribution on the exponential parameter. Subsequently

the predictive distribution of a future observation from the process is obtained. In the gamma prior we essentially assume one of the hyperparameters known (or guessed) and the other unknown. An estimate for the latter is produced by the predictive sample reuse method essentially as a by-product of deriving a point predictor. The question of censored data, where ambiguity exists in the execution of the predictive sample reuse method is treated in the next section and tentatively resolved by the ploy of pseudo-observations that are supplied from a partial Bayesian or other structure.

The utilization of the approximate predictive distribution;i.e. with one hyperparameter estimated, as a forecasting tool is valid to the extent of the appropriateness of the fine structure assumptions with uncertainty commensurate with the roughness of the approximation. On the other hand the predictor itself may be useful considerably beyond the bounds of the initial structure assumed in that it may be robust as a point predictor for a variety of possible structures. Further it may be most useful in a low structure situation, where any specific distributional assumptions are fraught with peril.

Suppose we have a random sample $X_1,\ldots,X_N$ on an exponential random variable X whose density is

$$f(x|\mu) = \mu e^{-\mu x}, \quad \mu > 0, \quad x > 0. \tag{2.1}$$

If our prior objective or subjective information is subsumed in a prior density for μ,

$$p(\mu) \propto \mu^{\delta-1} e^{-\gamma\mu}, \quad \gamma > 0, \delta > 0 \tag{2.2}$$

and we are interested in predicting a value x_{N+1} for the random future observation X_{N+1} given the previous N observations $\mathbf{x}^{(N)}$, say, then the predictive density for X_{N+1} is easily calculated to be, for $x_{N+1} > 0$,

$$\begin{aligned} f(x_{N+1}|\mathbf{x}^{(N)}) &= \int p(\mu|\mathbf{x}^{(N)}) f(x_{n+1}|\mu)\, d\mu \\ &= (N+\delta)(N\bar{x}+\gamma)^{N+\delta}/(N\bar{x}+\gamma+x_{N+1})^{N+\delta+1} \end{aligned} \tag{2.3}$$

where $\bar{x}$ is the sample mean and $p(\mu|x^{(N)})$ is the posterior density of μ given the previous N observations $x^{(N)}$. Hence our forecast about X_{N+1} involves the hyperparameters γ and δ which enter the problem via the distribution of the parameter μ. Before any observations are taken one can also find the predictive (marginal) density of the generic variable X, namely

$$f(x) = \int f(x|\mu)\, p(\mu)\, d\mu = \delta\gamma^{\delta}/(\gamma+x)^{\delta+1}, \quad x > 0. \tag{2.4}$$

Hence it is convenient and perhaps more appropriate to think about these hyperparameters in terms of predicting X before any observations are taken rather than in how they modulate the assumed prior distribution of μ. Therefore, prior to the sample, we have

$$\begin{aligned} E(X) &= \gamma/(\delta - 1) = g \\ \text{Var}\,(X) &= \delta\gamma^2/(\delta-2)(\delta-1)^2 = g^2(1+\alpha)/(1-\alpha) \end{aligned} \tag{2.5}$$

where $\alpha = (\delta-1)^{-1}$.

Clearly Var(X) exists for $0<\alpha<1$, and E(X) exists for $\alpha>0$ while the distribution exists for all $\alpha \notin [-1,0]$. Hence if one could frame his prior opinions about the potentially observable values of X in terms of its expectation and variance then one can easily execute the whole predictive process by solving for the appropriate values δ and γ from (2.5) and substituting them in (2.3).

It is to be noted that (2.3) and (2.4) were obtained from (2.1) and (2.2). However, for the predictivist who would prefer to start from (2.1) and (2.4) in terms of convenience of framing his predictions this is somewhat awkward. Interestingly enough in this case starting with $f(x|\mu)$ and $f(x)$ is sufficient to obtain $p(\mu)$ and $f(x_{N+1}|\bar{x})$ which is a more logical and appealing approach for the predictivist. This is true here because $f(x)$ is the unique Laplace transform of $\mu^{-1}p(\mu)$.

Now as we mentioned previously making all of these assumptions yields the requisite information for making probability statements about a future value provided that one has specified values for g and α. However while one may often be willing to hazard a guess at g, one may be far less willing to specify a value for α.

We now shall apply the predictive sample reuse method in order that the data itself should yield a value for α once g has been assumed.

If we had already observed $\mathbf{X}^{(N)} = \mathbf{x}^{(N)}$ and wished to predict a future value for X_{N+1}, we could use the posterior expectation of X_{N+1} obtained from the predictive density given by (2.3). This is easily calculated to be

$$E(X_{N+1}) = (N\bar{x} + \gamma)/(N + \delta - 1) = (\alpha N\bar{x} + g)/(\alpha N + 1) = f. \tag{2.6}$$

Note that when $\delta \to 1$ and $\gamma \to 0$, we obtain the usual predictor $\bar{x}$.

In terms of the predictive sample reuse method, Geisser (1975), equation (2.6) may be utilized as a predictive function. In order to supply a value for α we apply the method using one-at-a-time omissions and a squared discrepancy as follows: The average squared discrepancy is

$$D(\alpha) = N^{-1}\sum_i (f_i - x_i)^2 = N^{-1}\sum_i \left(\frac{\alpha(N\text{-}1)\bar{x}_i + g}{\alpha(N\text{-}1) + 1} - x_i\right)^2 \tag{2.7}$$

where f_i and $\bar{x}_i$ are defined respectively as the predictive function and the sample average with x_i omitted. In order to find a suitable α, we minimize $D(\alpha)$ with respect to α for $\alpha \geq 0$. (Note again that for the density given by (2.4), Var (X) exists only for $0 < \alpha < 1$, although the distribution for X exists for $\delta > 0$ and hence for all $\alpha \notin [-1,0]$. Nevertheless we shall not restrict ourselves to $\alpha > 0$ although this is essentially the range on α for which the prior mean exists), but also include $\alpha = 0$, a value, which is possible when γ is a function of α and $\alpha \to g$ as $\alpha \to 0$.)

We can easily evaluate

$$D(\alpha) = [(N\text{-}1)s^2(\alpha N + 1)^2 + N(g-\bar{x})^2] \,/\, N[\alpha(N\text{-}1) + 1]^2, \tag{2.8}$$

where $s^2 = (N\text{-}1)^{-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$. Taking the derivative with respect to α and setting this equal to zero yields the solution

$$\begin{aligned} \hat{\alpha} &= (t^2\text{-}1)/N \qquad && \text{for } t^2 > 1 \\ \hat{\alpha} &= 0 && \text{if } t^2 \leq 1 \end{aligned} \tag{2.9}$$

where $t^2 = N(g-\bar{x})^2/s^2$. Hence this yields the predictor

$$\begin{aligned} f(\hat{\alpha}) = \hat{f} &= [(t^2\text{-}1)\bar{x} + g]/t^2 \qquad && \text{if } t^2 > 1 \\ f(\hat{\alpha}) &= g && \text{if } t^2 \leq 1 \end{aligned} \tag{2.10}$$

Of course for the strict Bayesian the use of $\hat{\alpha}$ and its derived value $\hat{\delta}$ contradicts the fundamental canon of Bayesianism that the prior hyperparameters should not depend on the data. However it should serve as an approximate solution to the problem in the sense that the unknown hyperparameter δ is replaced by $\hat{\delta}$ if $\hat{\alpha} > 0$ in (2.3), given the high structure assumptions. This problem and method for solution was first proposed by Geisser (1975b) with further commentary, Geisser (1976, 1980).

It may also be mentioned that the predictor $\hat{f}$ can also be conceived as totally independent of the Bayesian process and the likelihood when obtained from this approach in the sense that we have merely chosen f as a point predictor for X_{N+1} and have ascertained $\hat{f}$ by a squared discrepancy measure. We also note that the predictive function f is basically a linear combination of the mean $\bar{x}$ and the prior guess g with weights αN and 1. There are

undoubtedly other models that can lead to forecasting the next observation as linear combinations of a prior mean and the sample mean when the predictive expectation of a future observation is utilized. In this regard then one could define a predictive function that is a linear combination of the mean and a guessed value g

$$f^* = \alpha^* x + (1-\alpha^*)g, \quad 0 \leq \alpha^* \leq 1 \tag{2.11}$$

This yields, for squared discrepancy and one-at-a-time omissions, Geisser (1975a),

$$\begin{aligned} \alpha^* &= (t^2-1)/[t^2+(N-1)^{-1}] \qquad \text{for } t^2 > 1, \\ &= 0, \qquad\qquad\qquad\qquad \text{for } t^2 \leq 1 \end{aligned} \tag{2.12}$$

Hence

$$\begin{aligned} \hat{f}^* &= [(t^2-1)x + N(N-1)^{-1}g]/[t^2+(N-1)^{-1}], \qquad \text{for } t^2 \geq 1 \\ &= g \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \text{if } t^2 < 1 \end{aligned} \tag{2.13}$$

Clearly $\alpha^* = \alpha N/(\alpha N+1)$ for $\alpha \geq 0$ in terms of the transformed predictive function. On the other hand $\hat{\alpha}^* < \hat{\alpha}N/(\hat{\alpha}N+1)$, for $t^2 > 1$, the estimation procedure not being invariant under such a transformation. However they will be quite close as they are asymptotically equivalent for large N. Comparison of $\hat{f}$ with $\hat{f}^*$ reveals they also converge for large N, but slightly more weight is attached to $\bar{x}$ in $\hat{f}$ than in $\hat{f}^*$.

In summary then, in the assumed presence of the high initial structure f should be preferable, but for robustness to other structures leading approximately to the aforementioned linear combination, f^* may be preferable. In any event the difference is negligible for large N. In the absence of any distributional assumptions both predictors are viable methods for having something to say about the prediction of future observations.

3. CENSORED DATA

In many cases especially in survival or reliabilitity studies the experiment is usually terminated before all of the subjects or units have expired or failed. Suppose the experiment is such that for d of the observations, failure times are recorded as $x_1, \ldots, x_d$, while the remaining N-d observations have survived but were censored at values $x_{d+1}, \ldots, x_N$. Hence

$$L(\mu) = \Pi_{i=1}^{d} f(x_i|\mu)\, \Pi_{i=d+1}^{N} [1\text{-}F(x_i|\mu)]$$

where $F(x_i|\mu)$ is the distribution function of X_i. For the exponential case, clearly

$$L(\mu) \propto \mu^d e^{-\mu[dx_d + (N-d)x_{N-d}]} \tag{3.1}$$

where $x_d = d^{-1}\Sigma_1^d x_i$ and $x_{N-d} = (N-d)^{-1}\Sigma_{i=1}^{N-d} x_{d+i}$. From (3.1) and (2.2) we can obtain first the posterior density of μ and then, as previously, the predictive density for a future observation X_{N+1},

$$\begin{aligned} &f(x_{N+1}|\mathbf{x}^{(d)},\mathbf{x}^{(N-d)}) \\ &\quad = (d+\delta)(dx_d + (N-d)x_{N-d} + \gamma)^{d+\delta}/(dx_d + (N-d)x_{N-d} + \gamma + x_{N+1})^{d+\delta+1} \end{aligned} \tag{3.2}$$

where $\mathbf{x}^{(d)}$ represents the observations whose failure times are recorded and $\mathbf{x}^{(N-d)}$ the censored observations. Further the predictive expectation, to be used as the predictive function, is

$$\begin{aligned} E(X_{N+1}) &= [dx_d + (N-d)\bar{x}_{N-d} + \gamma]/(d + \delta - 1) \\ &= [(dx_d + (N-d)x_{N-d})\alpha + g]/(\alpha d + 1) = f. \end{aligned} \tag{3.3}$$

Note that for $\delta \to 1$ and $\gamma \to 0$ we obtain the usual predictor

$$x_d + d^{-1}(N-d)x_{N-d}.$$

Due to censoring there is difficulty in appropriately executing the predictive sample reuse method. One tentative solution is to generate $N-d$ pseudo-observations having values x'_{d+i}, $i = 1,\ldots, N-d$, say. These are the presumed failure times for the censored observations $x_{d+1},\ldots,x_N$. We shall take as the pseudo value x'_{d+i}, the expectation of the predictive distribution of X_{d+i} given $X_{d+i} > x_{d+i}$, the censored value. More precisely the likelihood in (3.1) is used but with x_{d+i} omitted; i.e., based on all the observations but x_{d+i}. This is then combined with the prior density of μ whence the posterior density of μ is obtained and subsequently the predictive density of X_{d+i} computed. From this we then compute the conditional density of X_{d+i} given $X_{d+i} > \mathbf{x}_{d+i}$

$$f(x|X_{d+i} > \mathbf{x}_{d+i}) = \frac{(d+\delta)(d\bar{x}_d + (N-d)\bar{x}_{N-d} + \gamma)^{d+\delta}}{(d\bar{x}_d + (N-d)\bar{x}_{N-d} + \gamma + x - x_{d+i})^{d+\delta+1}} \tag{3.4}$$

Further computation yields

$$E(X_{d+i} \mid X_{d+i} > x_{d+i}) = [(d+\delta\text{-}1)x_{d+i} + dx_d + (N\text{-}d)x_{N-d} + \gamma]/(d+\delta\text{-}1) \qquad (3.5)$$
$$= x_{d+i} + \frac{(dx_d + (N\text{-}d)x_{N-d})\alpha + g}{\alpha d + 1} = x'_{d+i},$$

and

$$\mathrm{Var}(X_{d+i} \mid X_{d+i} > x_{d+i}) = \frac{(dx_d + (N\text{-}d)x_{N-d} + \gamma)^2(d+\delta)}{(d+\delta\text{-}1)^2(d+\delta\text{-}2)} = \frac{d+\delta}{d+\delta\text{-}2} f^2 \qquad (3.6)$$

the latter being independent of i.

Now in executing the sample reuse method with predictive function given by (3.3) using the actual observations $x_1,\ldots,x_d$ and the pseudo observations $x'_{d+1},\ldots,x'_N$ given by (3.5) it seems sensible to give the pseudo-observations a weight that differs from that assigned to the uncensored observations in contradisctinction to an unweighted and consequently inadequate solution, Geisser (1975b). We note that

$$\mathrm{Var}(X_i \mid \mu) = \mu^{-2} \qquad \text{for } i = 1,\ldots,d. \qquad (3.7)$$

Since μ is unknown we shall compute

$$E_\mu[\mathrm{Var}(X_i \mid \mu)] = E_\mu[\mu^{-2}] \qquad (3.8)$$

over the posterior distribution of μ. This results in

$$E_\mu(\mu^{-2}) = \frac{(dx_d + (N\text{-}d)x_{N-d} + \gamma)^2}{(d+\delta\text{-}1)(d+\delta\text{-}2)} = \frac{d+\delta\text{-}1}{d+\delta\text{-}2} f^2 \qquad (3.9)$$

where f is as defined in (3.3).

We can define a weighted discrepancy for $d > 1$, $N\text{-}d > 1$ as follows:

$$D(\alpha) = E_\mu^{-1}(\mu^{-2}) \sum_{j=1}^{d} \left(\frac{[(d\text{-}1)x_{d,j} + (N\text{-}d)x_{N-d}]\alpha + g}{\alpha d + 1} - x_j \right)^2 \qquad (3.10)$$
$$+ [\mathrm{Var}(X \mid X > x_{d+i})]^{-1} \sum_{k=d+1}^{N} \left(\frac{[dx_d + (N\text{-}1\text{-}d)x_{N-d,k}]\alpha + g}{\alpha d + 1} - x'_k \right)^2$$

where $x_{d,j}$ and $x_{N-d,k}$ are respectively the sample means of d-1 uncensored observations omitting x_j and the mean of N-1-d censored observations omitting x_k.

After some algebraic manipulation we obtain

$$D(\alpha) = \frac{(d\text{-}1)s_d^2(\alpha d+1)^3 + d(g\text{-}x_d+\alpha(N\text{-}d)x_{N\text{-}d})^2(\alpha d+1)}{[\alpha(d\text{-}1)+1][(d\bar{x}_d+(N\text{-}d)\bar{x}_{N\text{-}d})\alpha+g]^2}$$

$$+ \frac{[\alpha(d+1)+1][\alpha(d\text{-}1)+1]}{[(d\bar{x}_d+(N\text{-}d)\bar{x}_{\text{-}d})\alpha+g]^2}\,\Sigma_{j=d+1}^{N}x_j^2. \tag{3.11}$$

The solution then for α is obtained by differentiating (3.11) with respect to α and setting it equal to zero. This will result in a polynomial in α, whose roots are stationary points. After discarding negative and complex roots, the positive roots α, say, need be compared with D(0) and D(∞) to ascertain the global minumum for $\alpha \geq 0$.

For $d=1$ and $N>2$ only the second term in (3.11) obtains and formal minimization in this case yields $\alpha=\infty$, so that $f=Nx$, the usual predictor in this case.

For $d>1$ and $N=d+1$ only the first term in (3.11) obtains. Minimization then follows in the same manner as in the discussion for $d>1$ and $N\text{-}d>1$.

It is to be noted that in the weighting we merely used terms that reflected variation. Perhaps a more appropriate weighting scheme would also include covariation among those values that are correlated. As a step in this direction we can take cognizance of the covariance among the pseudo-observations.

A simple calculation reveals that the joint predictive density of X_{d+i} and X_{d+j} $i\neq j = 1, \ldots, N\text{-}d$ conditional on $X_{d+i} > x_{d+i}$ and $X_{d+j} > x_{d+j}$ is

$$f(z,w\,|\,X_{d+i}>x_{d+i},X_{d+j}>x_{d+j}) = \frac{(d+\delta)(d+\delta+1)(d\bar{x}_d+(N\text{-}d)\bar{x}_{N\text{-}d}+\gamma)^{d+\delta}}{(d\bar{x}_d+(N\text{-}d)\bar{x}_{N\text{-}d}+z\text{-}x_{d+i}+w\text{-}x_{d+j})^{d+\delta+2}} \tag{3.12}$$

whence we calculate

$$\mathrm{Cov}(X_{d+i}X_{d+j}\,|\,X_{d+i}>X_{d+j}>x_{d+j}) = (d+\delta)^{-1}\,\mathrm{Var}(X_{d+i}\,|\,X_{d+i}>x_{d+i}), \quad \text{for } i\neq j,\ i,j=1,\ldots,N\text{-}d \tag{3.13}$$

Use of this alters the second term in (3.11) to

$$\frac{[\alpha(d+1)+1][\alpha(d\text{-}1)+1]}{(\alpha N+1)(\alpha d+1)[(d\bar{x}_d+(N\text{-}d)\bar{x}_{N\text{-}d})\alpha+g]^2} \times [\alpha(N\text{-}1)+1)\,\Sigma_{j=1}^{N-D}x_{d+j}^2 - 2\alpha\Sigma_{i>j}^{N-D}x_ix_j] \tag{3.14}$$

When, as is often the case, all of the N-d observations are censored at the same value, say x_o, then (3.14) simplifies to

$$\frac{[\alpha(d+1)+1]^2[\alpha(d\text{-}1)+1](\text{N-}d)x_o^2}{(\alpha\text{N}+1)\,[(dx_d+(\text{N-}d)x_o)\alpha+g]^2} \tag{3.15}$$

This term is then $[\alpha(d+1)+1]/[\alpha\text{N}+1]$ times the second term in (3.11), indicating roughly the disminished effect of the contribution of the portion of $D(\alpha)$ involving the pseudo-observations by taking into account their covariance structure. Of course this further complicates arriving at a solution for α and it is not clear just how significant the resulting improvement would be.

The most complex weighting scheme would also attempt to take into account covariation between uncensored observations and pseudo-observations. Now for $i=1, \ldots, d$, $j=d+1,\ldots, \text{N}$; $X_j' = X_j + (\text{N}\,\alpha\,\text{X} + g)/(\alpha\, d + 1)$

$$\text{Cov}(X_i,X_j'|\mu) = \frac{\alpha}{\alpha\, d + 1}\,\text{V}(X_i|\mu) = \frac{\alpha\mu^{-2}}{\alpha\, d + 1}. \tag{3.16}$$

Again using (3.9) we find that

$$E_\mu[\text{Cov}(X_i,X_j'|\mu] = f^2/(d+\delta\text{-}2). \tag{3.17}$$

Hence we may use as a weighting matrix the inverse of the NxN partitioned matrix

$$V = f^2/(\text{d}+\delta\text{-}2)\begin{pmatrix} \text{d} & \text{N-d} \\ (\text{d}+\delta\text{-}2)\text{I} & \text{J}_{12} \\ \text{J}_{21} & (d+\delta\text{-}1)\text{I}+\text{J}_{22} \end{pmatrix}\begin{matrix} d \\ \\ \text{N-}d \end{matrix} \tag{3.18}$$

where J_{ij} is a matrix all of whose entries are unity. The inverse of V can readily be displayed by letting $U = f^2(\alpha d+1)[\alpha(d\text{-}1)+1]^{-1}V^{-1}$ with partitions similar to V so that

$$U_{11} = \text{I} + \frac{(\text{N-}d)\text{J}_{11}}{(d+\delta\text{-}1)(\text{N}+\delta\text{-}1)\text{-}d(\text{N-}d)},$$

$$U_{ij} = \frac{-(d+\delta\text{-}1)\text{J}_{ij}}{(d+\delta\text{-}1)(\text{N}+\delta\text{-}1)\text{-}d(\text{N-}d)}, \text{ for } i \neq j \tag{3.19}$$

$$U_{22} = I\,\frac{(\delta\text{-}1)J_{22}}{(d+\delta\text{-}1)^2+(\delta\text{-}1)(N\text{-}d)}.$$

Now for $d>1$ and $N\text{-}d>1$, let

$$\begin{aligned}\triangle_j &= f_j\text{-}x_j \qquad \text{for } j = 1,\ldots,d \\ &= f_j\text{-}x_j' \qquad \text{for } j = d+1,\ldots,N\end{aligned} \tag{3.20}$$

where again f_j is the predictive expectation f omitting the j^{th} observation. Further, letting $\triangle' = (\triangle_1,\ldots,\triangle_N)$ we can now define

$$D(\alpha) = \triangle' V^{-1} \triangle$$

and minimize it for $\alpha > 0$. Again evaluation of $D(\alpha)$ leads to rather complicated algebra which we shall omit.

Once a solution $\hat{\alpha}$ is rendered we can convert it to obtain the approximate predictive distribution of a future observation or just use $\hat{f}$ as a point predictor.

For the second kind of predictive function

$$f^* = \alpha^*(\bar{x}_d + d^{-1}(N\text{-}d)\bar{x}_{N-d}) + (1\text{-}\alpha^*)\, g = \alpha^* h + (1\text{-}\alpha^*)\, g \tag{3.21}$$

which does not lean as much on the previous high structure assumptions, we use as pseudo-observations

$$x'_{d+i} = x_{d+i} + \bar{x}_d + d^{-1}(N\text{-}d)\bar{x}_{N-d} = x_{d+i} + h. \tag{3.22}$$

This is akin to frequentist prediction since using x'_{d+i}, $i = 1,\ldots,N\text{-}d$ as actual observations in conjunction with $x_1,\ldots,x_d$ preserves the frequentist predictor, $\bar{x}_d + d^{-1}(N\text{-}d)\bar{x}_{N-d}$, as this is the average of both uncensored values and pseudo-observations. Now (3.22) can also be obtained by letting $\delta \to 1$ and $\gamma \to 0$ in (3.5).

Here the simplest weighted squared discrepancy measure neglecting covariation but not variances is

$$D(\alpha^*) \propto \Sigma_{j=1}^{d}\,(f_j^*\text{-}x_j)^2 + \frac{d}{d+1}\Sigma_{j=d+1}^{N}(f_j^*\text{-}x_j)^2 \tag{3.23}$$

where f_j^* is f^* as in (3.21) but with x_j omitted. The weighting here is again closer to a frequentist approach although it also can be obtained from (3.6) and (3.9) by letting $\delta \to 1$. Let $f_j^* = \alpha^* h_j + (1-\alpha^*)g$ so that

$$h_j = (d\text{-}1)^{-1}(d\bar{x}_d + (N\text{-}d)\bar{x}_{N-D} - x_j) \quad \text{for } j = 1,\ldots,d \tag{3.24}$$
$$= \bar{x}_d + d^{-1}[(N\text{-}d)\bar{x}_{N-d} - x_j] \quad \text{for } j = d+1,\ldots,N$$

then the minimization of $D(\alpha^*)$ with respect to α^* yields

$$\hat{\alpha}^* = \frac{\sum_{j=1}^{d}(h_j\text{-}g)(x_j\text{-}g) + d(d+1)^{-1}\sum_{j=d+1}^{N}(h_j\text{-}g)(x_j\text{-}g)}{\sum_{j=1}^{d}(h_j\text{-}g)^2 + d(d+1)^{-1}\sum_{j=d+1}^{N}(h_j\text{-}g)^2} \quad \text{for } 0 \le \hat{\alpha}^* \le 1$$
$$= 1 \qquad \text{for } \hat{\alpha}^* > 1$$
$$= 0 \qquad \text{for } \hat{\alpha}^* < 0. \tag{3.25}$$

If one uses a scheme with no weighting at all then

$$\hat{\alpha}^* = \frac{(d\text{-}1)N(h\text{-}g)^2 + (h\text{-}g)d^{-1}(N\text{-}d)\bar{x}_{N-d} - d^{-1}(N\text{-}d)^2\bar{x}_{N-d}^2 - (d\text{-}1)s_d^2 - (d\text{-}1)d^{-1}\sum_{d+1}^{N}x_j^2}{(d^2\text{-}1)(h\text{-}g)^2 + 2(h\text{-}g)(N\text{-}d)\bar{x}_{N-d} + (d\text{-}1)^{-1}d^{-1}\bar{x}_{N-d}^2 + s_d^2 + (d\text{-}1)d^{-2}\sum_{d+1}^{N}x_j^2}$$

$$= 0 \qquad \text{if } \hat{\alpha}^* \le 0$$
$$= 1 \qquad \text{if } \hat{\alpha} \le 1. \tag{3.26}$$

A slightly different solution can be obtained by altering the function h. Previously h was defined as the sum of all the observations censored and uncensored, divided by the number of uncensored observations. We also noted that h was the mean of the uncensored values and the pseudo-observations.

Hence we could change the definition of h to this mean value which keeps invariant the value of the predictive function for given α. However h_j would now be altered to

$$h_j' = (N\text{-}1)^{-1}[N\bar{x}_d + (N\text{-}d)N\,d^{-1}\,\bar{x}_{N-d} - x_j] \quad \text{for } j = 1,\ldots,d \tag{3.27}$$

$$= \bar{x}_d + (N\text{-}d)d^{-1}\bar{x}_{N-d} - (N\text{-}1)^{-1}x_j \qquad \text{for } j = d+1,\ldots,N.$$

The solution for α^* is now obtained by substituting h_j' for h_j in (3.25).

An unweighted solution in this case is, Geisser (1975b),

$$\hat{\alpha}^* = \frac{N(g-h)^2 - A}{N(g-h)^2 + (N-1)^{-1}A} \text{ for } \hat{\alpha} > 0 \qquad (3.28)$$
$$= 0 \quad \text{for } \hat{\alpha} \leq 0$$

where

$$(N-1)A = (d-1)s_d^2 + d^{-1}(N-d)^2\,\bar{x}_{N-d}^2 + \Sigma_{j=d+1}^{N} x_j^2. \qquad (3.29)$$

However, though very simple, this does not appear to be a very satisfactory solution to the problem.

In both (3.24) and (3.27) it is required that $d > 1$ and $N-d > 1$. If $d = 1$ and $N > 2$ then the solution for α^* is the ratio of the second terms in (3.25) utilizing either h_j or h_j' respectively. For $d > 1$, $N = d+1$, the solution is the ratio of the first terms.

4. THE ALTERNATIVE APPROACH-SAMPLE REUSE ALGORITHMS

The second general approach described in Section 1 is both conceptually easier to apply and more readily facilitates arithmetic solutions. We now apply it to the censored situation of the previous section. Using (2.9)

$$\tilde{\alpha} = \frac{t^2(\alpha) - 1}{N} \qquad (4.1)$$

where from (3.5)

$$x_j'(\alpha) = x_j + \frac{(d\bar{x}_d + (N-d)\bar{x}_{N-d})\alpha + g}{\alpha d + 1} \qquad j > d. \qquad (4.2)$$

Let

$$\bar{x}(\alpha) = \tfrac{1}{N}\left[\Sigma_{j=1}^{d} x_j + \Sigma_{j=d+1}^{N} x_j'(\alpha)\right] = \bar{x} + \frac{(N-d)}{N}\left(\frac{N\bar{x}\alpha + g}{\alpha d + 1}\right) \qquad (4.3)$$

where $N\bar{x} = \Sigma_{\alpha=1}^{N} x_j$. Let

$$\beta = \frac{N-d}{N}\left(\frac{N\bar{x}\alpha + g}{\alpha d + 1}\right) \tag{4.4}$$

$$\begin{aligned}(N\text{-}1)s^2(\alpha) &= \Sigma_{j=1}^{d}(x_j\text{-}\bar{x}\text{-}\beta)^2 + \Sigma_{j=d+1}^{N}(x_j + \beta\text{-}\bar{x}\text{-}\beta)^2 \\ &= (N\text{-}1)s^2 + d\beta^2 - 2\beta d(\bar{x}_d\text{-}\bar{x})\end{aligned} \tag{4.5}$$

where $(N\text{-}1)s^2 = \Sigma_{j=1}^{N}(x_j\text{-}\bar{x})^2$. Now by definition

$$t^2(\alpha) = \frac{N(\bar{x}(\alpha)\text{-}g)^2}{s^2(\alpha)} \; . \tag{4.6}$$

Hence substituting (4.6) in (4.1) and solving for α in terms of β; i.e.,

$$N\alpha + 1 = \frac{(N\text{-}d)(g\text{-}\bar{x}\text{-}\beta)}{d\beta\text{-}(N\text{-}d)\bar{x}} \tag{4.7}$$

we obtain a quadratic equation in β

$$a\beta^2 + b\beta + c = 0 \tag{4.8}$$

where

$$\begin{aligned}a &= d(N^2\text{-}d)/(N\text{-}1) \\ b &= 2(N\text{-}d)d(\bar{x}\text{-}x_d)(N\text{-}1)^{-1} + dN(\bar{x}\text{-}g)\text{-}N(N\text{-}d)\bar{x} \\ c &= (N\text{-}d)s^2 + N(N\text{-}d)\bar{x}(g\text{-}\bar{x}) \; .\end{aligned} \tag{4.9}$$

After obtaining the solution $\hat{\beta}$ we solve for $\hat{\alpha}$ from (4.7) and substituting this in (4.2) we obtain the conditional predictor $x(\hat{\alpha})$ and setting $x_j = 0$ the unconditional predictor.

This approach can also be applied to the case given by equations (2.11) and (2.12), namely $f^* = \alpha^*\bar{x} + (1\text{-}\alpha^*)g$ for $0 \leq \alpha^* \leq 1$

$$\alpha^* = (t^2(\alpha^*) - 1)/[t^2(\alpha^*) + (N\text{-}1)^{-1}] \tag{4.10}$$

$$t^2(\alpha^*) = \frac{N(\bar{x}(\alpha^*) - g)^2}{s^2(\alpha^*)} \tag{4.11}$$

where the assumed conditional predictor is

$$x_j'(\alpha^*) = x_j + Nd^{-1}\bar{x}\alpha^* + (1\text{-}\alpha^*)g \tag{4.12}$$

so that

$$x(\alpha^*) = x + (N\text{-}d)N^{-1}[Nd^{-1}x\alpha^* + (1\text{-}\alpha^*)g] \tag{4.13}$$

and

$$(N\text{-}1)s^2(\alpha^*) = (N\text{-}1)s^2 + d\beta^{*2}\text{-}2\beta^* d(x_d\text{-}x)$$

where

$$N\beta^* = (N\text{-}d)\,(Nd^{-1}\bar{x}\alpha^* + (1\text{-}\alpha^*)g) = (N\text{-}d)\,(zd^{-1}\alpha^* + g) \tag{4.14}$$

or

$$zd^{-1}(N\text{-}d)\alpha^* = N\beta^*\text{-}(N\text{-}d)g$$

for $z = Nx - dg$.

Hence solutions for α^*, say $\hat{\alpha}^*$, are obtained from the cubic equation

$$(N\text{-}1)(1\text{-}\alpha^*)(d + (N\text{-}d^2)\alpha^*)z^2 = Nd^2(N\text{-}1 + \alpha^*)s^2(\alpha^*). \tag{4.15}$$

Only one value of the cubic will be appropriate for a fixed x, s^2 and g. Substitution of the appropriate $\hat{\alpha}^*$ in (4.12) yields the conditional predictor $x(\hat{\alpha}^*)$ and setting $x_j = 0$ yields the unconditional predictor.

We now illustrate this approach with some data obtained from Gnedenko, Belyayev and Solovyev (1969, p. 176). A sample of 100 items are tested and time to failure recorded for each up until 500 time units have elapsed (the actual time unit is not given). It is found that during this period 89 items have survived and the recorded failure times for the other 11 are; 31, 49, 90, 135, 161, 249, 323, 353, 383, 436, 477. The total time on test in undetermined units, is 47,187 (inaccurately given as 47,147 by the authors).

Figure I represents a plot of the predicted value of a future time to failure comparing (4.12), substituting $\hat{\alpha}^*$ for α^*, as a function of g, an apriori guessed value, with (4.2), substituting $\hat{\alpha}$ for α, which derives from the more highly structured predictive approach. The two curves exhibit similar shapes except that the interval for disregarding the data is more than twice as wide for the high structured case and the approach to completely disregarding the guess is far slower.

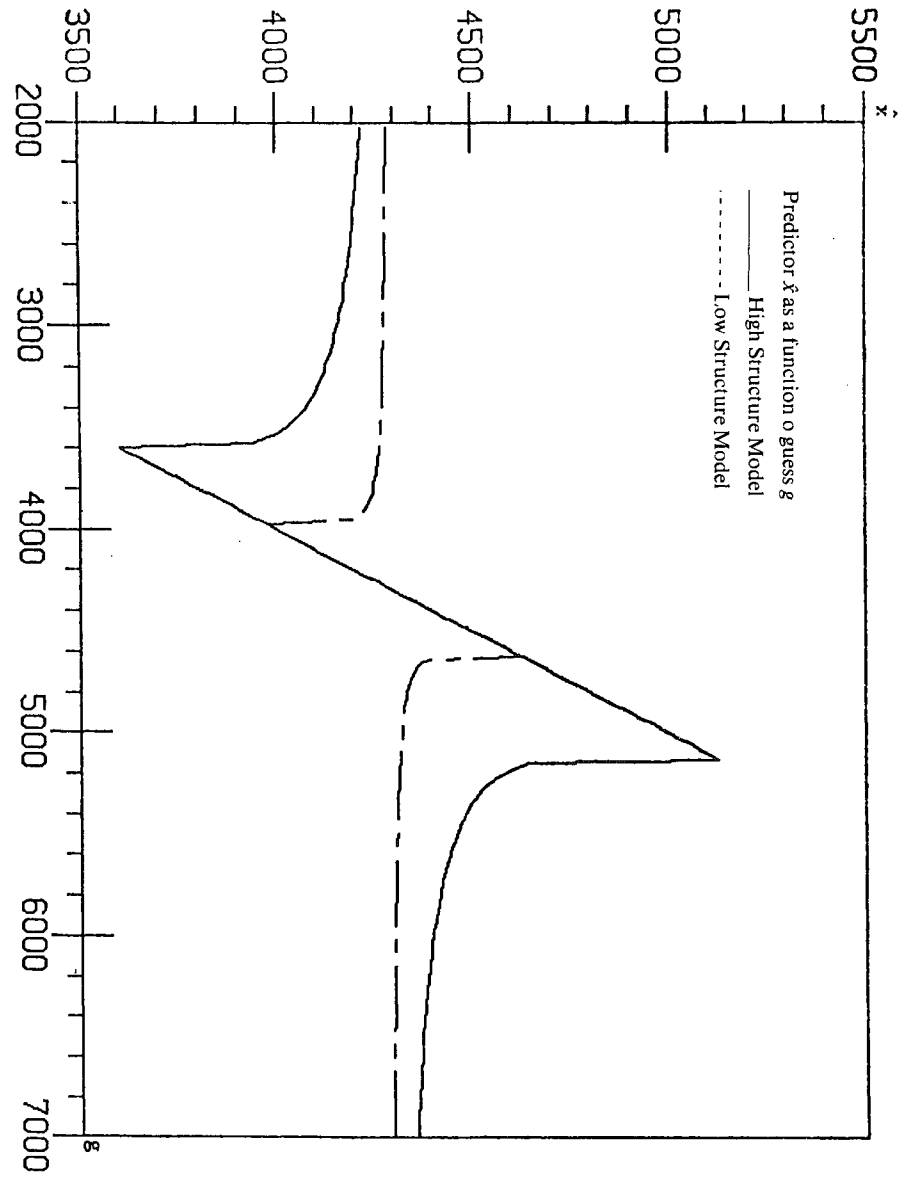
3500
4000
4500
5000
5500
x
2000
3000
4000
5000
6000
7000
g
Predictor x as a function o guess g
High Structure Model
Low Structure Model

FIGURE 1

Figure II demonstrates how the estimated predictive density of a future observation varies as a function of g using the high structure model. Note that values of g from 3,700 throgh 5,000 result in $\hat{\alpha}=0$ and consequently the density is exponential while for other values of g the density is of the beta form given by (3.2). This accounts for some of the minor perturbations.

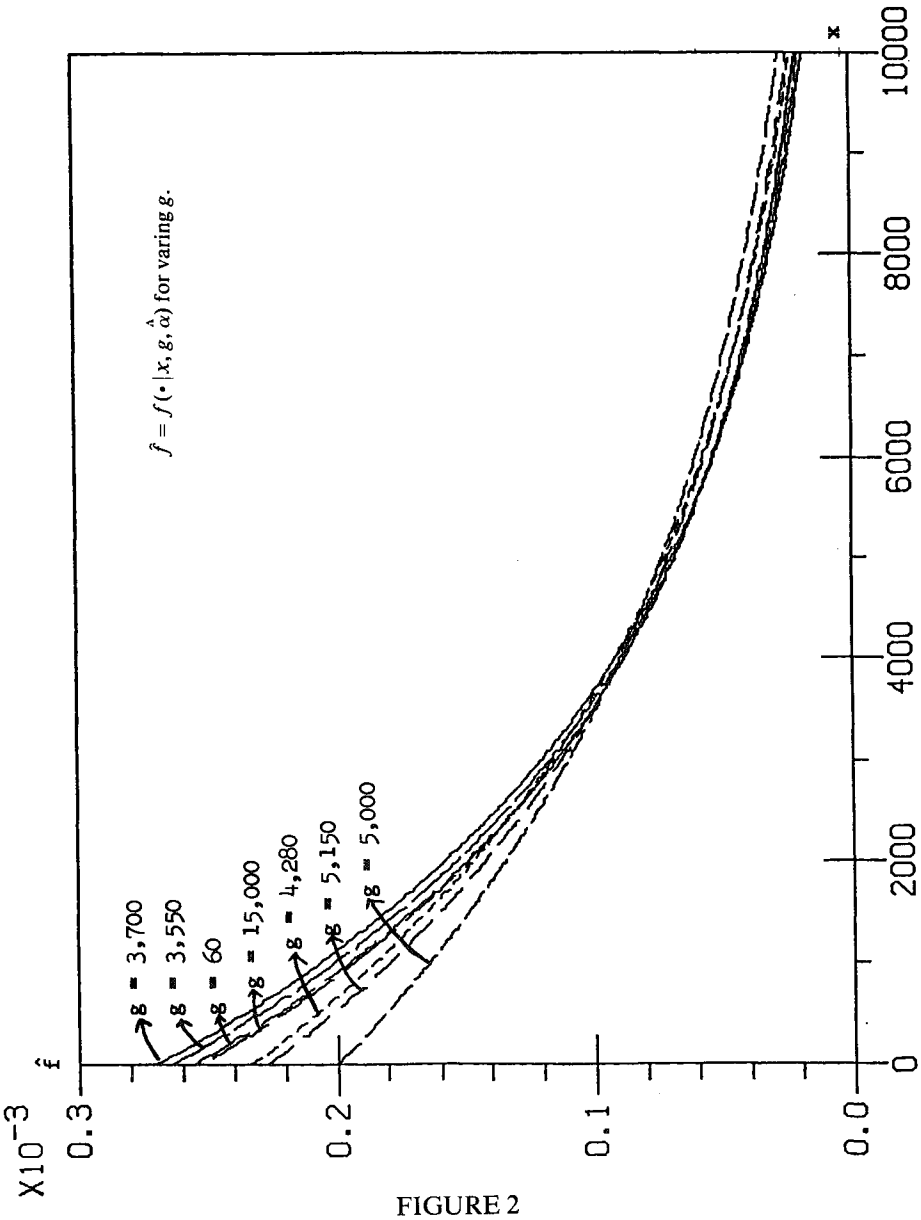

FIGURE 2

Table I gives the shortest .9 probability interval (90th percentile points) for a future value of x for varying g from the estimated predictive distribution.

TABLE I

90th Percentile Point of $F(.|x, g, \alpha)$ to Nearest Integer

g	60	3,550	3,700	4,280	4,290	5,000	5,150	15,000
$\hat{\alpha}$	9.8179	.1259	0	0	0	0	.9409	62.3680
pc point	9,890	9,226	8,520	9,855	9,878	11,513	10,151	10,018

Guesses that are widely discrepant with the data such as 60 and 15,000 are largely ignored and yield percentiles close to that of $g = 4290$, a guess equivalent to the data predictor. Reversals in percentile points for such values as 3,550 and 3,700 are accounted for by the same phenomenon occuring in Figure I and to a lesser extent to the change in the form of the distribution function.

ACKNOWLEDGEMENT

I would like to thank Douglas Kess for the computations and graphs appearing in the previous section. This work was supported by NIH grant GM25271.

REFERENCES

GEISSER, S. (1971) The inferential use of predictive distributions. *Foundations of Statistical Inference.* (B.P. Godambe and D.A. Sprott, eds.) 456-69. Toronto, Montreal: Holt, Rinehard and Winston.

— (1974) A predictive approach to the random effect model. *Biometrika* **61**, 101-107.

— (1975a) The predictive sample reuse method with application. *J. Amer. Statist. Assoc.* **70**, 320-328.

— (1975b) Bayesianism, predictive sample reuse, pseudo-observations, and survival. *Bull. Internat. Statist. Inst.* **40**, 285-289.

— (1975c) A new approach to the fundamental problem of applied statistics. *Sankhya* **37, B** 385-397.

— (1976). Predictivism and sample reuse. *Proc. 21st Design of Experiments Conference* ARO Report 76-2, 385-397.

— (1980) A predictivistic primer. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys.* 363-381 (*A. Zellner, ed*) Amsterdam: North Holland.

GNEDENKO, B.V., BELYAYEV, YU. K. and SOLOVYEV, A.D. (1969) *Mathematical Methods of Reliability Theory*. New York: Academic Press.

STONE, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion), *J. Roy. Statist. Soc. B*, **36**, 111-147