Predictivism and Sample Reuse

by

Seymour Geisser*
University of Minnesota

Technical Report No.255

November, 1975

*Supported in part by U. S. Army Grant DAHC04-74-G-0216

Predictivism and Sample Reuse

Seymour Geisser
University of Minnesota

1. Introduction

The fundamental thesis of this paper is that the inferential emphasis of Statistics, theory and concomitant methodology, has been misplaced. By this is meant that the preponderance of statistical analyses deals with problems which involve inferential statements concerning parameters. The view proposed here is that this stress should be diverted to statements about observables. With regard to parameters we take the narrow view which relegates them at most to be components of a statistical model that are not capable of being observed or potentially observed. This is not necessarily to deny them their utility in many hypothetical frameworks but there has been a strong tendency to exaggerate their importance in statistical inference. Even such a compelling "parameter" as the speed of light is in some sense ostensibly capable of being measured (observed) though perhaps subject to error. In this sense it is at least a potentially observable entity. Other values which often are misdesignated as parameters are those defined as a function of a finite number of observables or potential observables which typically occur in sample survey situations. For example we may be trying to "estimate" the total response of a specific finite population by observing some random portion of that population. The unobserved responses are presumably potentially observable (or the randomization is meaningless) and it is maintained that we are basically predicting them or some function of them. This is certainly within the realm of prediction though it is generally referred to as estimating a parameter of a finite population. Hence these two previously mentioned cases, measuring some physically meaningful constant and estimating functions of observables are within the realm of predictivism. It is our contention that in other cases the introduction of a convenient parametric

statistical model seems to impel statisticians to reformulate an experimenter's often imprecisely framed question concerning the data into a parametric analysis even when the parameters are completely artificial constructs. We then proceed to foist upon the unwary client "precise" statements about these too often non-existent entities. This tendency is reinforced because we have too long been subjected to solutions to hypothetical problems which invariably begin -- "suppose we are interested in the estimation of a parametric function BLAH(θ)." This stress on parametric inference made fashionable by mathematical statisticians has been not only a comfortable posture but also a secure buttress for the preservation of the high esteem enjoyed by applied statisticians because exposure by actual observation in parametric estimation is rendered virtually impossible. Of course those who opt for predictive inference i.e. predicting observables or potential observables are at risk in that their predictions can be evaluated to a large extent by either further observation or by a sly client withholding a random portion of the data and privately assessing a statistician's prediction procedures and perhaps concurrently his reputation. Therefore much may be at stake for those who adopt the predictivistic or observabilistic or aparametric view. But its relevance is clear.

It was the burden of a previous paper Geisser (1971) to argue that most problems currently cast in terms of parametric estimation and testing could be more informatively reformulated in a predictivistic mode. A general catalogue of such problems was presented there and the Bayesian inferential approach stressed. In this paper we shall discuss the problem of prediction per se from a variety of structures ranging from high to low depending upon the amount of information infused into the model. In particular we will stress a new low structure approach termed predictive sample reuse.

2. High Structure

The high structure approach to statistical prediction involves the tight apparatus of a prior distribution for the parameters involving known hyper-parameters and a specified likelihood, i.e. a joint sampling distribution of observables, past and future, as it were. Hence we need assume that $(X_1,\ldots,X_N; X_{N+1},\ldots,X_{N+M})$ or in a more compact notation $(X^{(N)};X_{(M)})$ has joint distribution $F(x^{(N)}; x_{(M)}|\theta)$ where θ is a set of unknown parameters. Further, a prior distribution on θ, say $G(\theta|\tau)$, is also assumed where the set of hyperparameters τ is known. The posterior distribution of θ is then based on the observed $X^{(N)} = x^{(N)}$,

$$(2.1) \qquad G(\theta|x^{(N)},\tau) = \frac{F(x^{(N)}|\theta)\, G(\theta|\tau)}{F(x^{(N)}|\tau)}$$

where

$$(2.2) \qquad F(x^{(N)}|\tau) = \int F(x^{(N)}|\theta)\, d\, G(\theta|\tau).$$

This then permits the calculation of the predictive distribution of $X_{(M)}$ given $X^{(N)}$ and τ, resulting in

$$(2.3) \qquad P(x_{(M)}|x^{(N)},\tau) = \int F(x_{(M)}|x^{(N)},\theta)\, d\, G(\theta|x^{(N)},\tau)$$

where

$$(2.4) \qquad F(x_{(M)}|x^{(N)},\theta) = \frac{F(x_{(M)};x^{(N)}|\theta)}{F(x^{(N)}|\theta)} .$$

The denominator of the above being the marginal sampling distribution of the observed random variables $X^{(N)}$. In essence, (2.3) represents the ultimate in statistical prediction and everything else is a summary of one kind or another of this distribution function. If point prediction is of interest then one might choose as a point predictor the predictive expectation of (2.3)

$$(2.5) \qquad E(X_{(M)}|X^{(N)}=x^{(N)},\tau)$$

or the median or the mode of (2.3) or whatever ensues from a particular loss function.

Often in this approach there is a necessary relaxation of the assumption that τ is known. This is generally handled in one of two ways. First it is often the case that little loss in terms of incoherence is engendered by assuming an improper prior for the hyperparameter τ. Hence a new predictive distribution is obtained by calculating

$$(2.6) \qquad P(x_{(M)}|x^{(N)}) = \int P(x_{(M)}|x^{(N)}, \tau)\, d\, G\, (\tau).$$

A second approach, usually associated with empirical Bayes procedures, is to "estimate" τ from the marginal distribution $F(x^{(N)}|\tau)$ given in (2.2) by maximum likelihood or the method of moments or any other convenient procedure. This then results in an approximate predictive distribution $P(x_{(M)}|x^{(N)}, \hat{\tau})$ and a point predictor, say, $E(x_{(M)}|x^{(N)}, \hat{\tau})$.

Historically there have also been two other high structure approaches. The first by Fisher (1956)was termed fiducial inference and the second Fraser (1968) termed structural inference. These generally require for their implementation, a much more restrictive sampling distribution and an assumption of complete ignorance concerning θ which in turn implies the absence of τ. Here one would calculate the fiducial or structural distribution $\varphi(\theta|x^{(N)})$ and then compute the predictive distribution of $X_{(M)}$,

$$(2.7) \qquad P_{\varphi}(x_{(M)}|x^{(N)}) = \int F(x_{(M)}|x^{(N)}, \theta)\ \ d\, \varphi\, (\theta|x^{(N)}).$$

This type approach is at most valid only under stringent assumptions. Many statisticians have questioned its validity entirely. Recently Barnard (1975) has developed a pivotal approach to parametric inference. His approach, as demonstrated by Hinkley (1975), can easily be adapted to a predictivistic mode by finding predictive pivots. It appears also to be capable of incorporating certain types of prior information.

3. Intermediate Structure

The classical (Neyman-Pearson) approach only assumes $(X^{(N)};X_{(M)}) \sim F(x^{(N)};x_{(M)}|\theta)$, i.e. a sampling distribution and enough structure on the distribution so that one can compute, independent of θ,

$$\Pr\left[X_{(M)} \in A(X^{(N)})\right] = p.$$

This of course in not a probability statement for $X^{(N)} = x^{(N)}$, as in the Bayes approach. Here p represents the degree of confidence that $X_{(M)} \in A\,(x^{(N)})$, p being a valid probability in the sense of the long-term frequency of repetitions from the joint set of random variables $(X^{(N)};X_{(M)})$. In other words p is the proportion of times in the long run that $X_{(M)} \in A\,(X^{(N)})$ and is interpreted as the confidence one has in $X_{(M)} \in A\,(x^{(N)})$ once $X^{(N)} = x^{(N)}$ has been observed. This is usually referred to as a tolerance interval in the statistical literature. For example, if we are dealing with the problem of predicting the $N+1$ observation X_{N+1} from the first N observations, $X_1,\ldots,X_N$ and assume that $\{X_i\}$ $i = 1,\ldots, N+1$ are iid $N(\theta,1)$ then one notes that for $\bar{X}_N = N^{-1}\sum_{i=1}^{N} X_i$

$$(3.1) \qquad \bar{X}_N - X_{N+1} \sim N(0,\ 1+N^{-1}).$$

From (3.1) we obtain

$$(3.2) \qquad \Pr\left[a \le \frac{X_{N+1}-\bar{X}_N}{\sqrt{1+N^{-1}}} \le b\right] = \Pr\left[\bar{X}_N + a\sqrt{1+N^{-1}} \le X_{N+1} \le \bar{X}_N + b\sqrt{1+N^{-1}}\right]$$

$$= \Phi(b) - \Phi(a) = p,$$

where $\Phi(y)$ is the standard normal distribution function.

While (3.2) is a probability statement, once we observe $\bar{X}_N = \bar{x}_N$ and calculate the limits, this now becomes a confidence statement and has only the restricted interpretation discussed before.

A point predictor is usually obtained by inserting in $E(X_{(M)}|X^{(N)} = x^{(N)}, \theta)$ an estimate $\hat{\theta}(x^{(N)})$ for θ - the expectation being taken over the conditional sampling distribution.

Another approach, having its roots in Fisher's work (1956), termed predictive likelihood, has recently been independently introduced by Hinkley (1975) and Lauritsen (1974). Here as in the fiducial approach, sufficiency though in an extended sense, plays the key role. It is assumed that $(X^{(N)}; X_{(M)})$ have likelihood $L(x^{(N)}; x_{(M)}|\theta)$ which admits a minimally totally sufficient reduction of the data. In the case of independent and identically distributed random variables a minimal sufficient reduction need only be available. In this latter case as pointed out by Fisher (1956), a minimal sufficient statistic is a function of the individual sufficient statistics from any portion of the entire sample. The concept of a totally sufficient statistic introduced by Lauritsen (1974) permits extension of this result to the more general case of dependence.

Let $s_N = s(X^{(N)})$ and $s_{N+M} = s(X^{(N)}, X_{(M)})$ be the set of totally sufficient statistics for θ based on the random variables to be observed and those that are to be observed and predicted, respectively. Then one can obtain, independent of θ, the conditional probability function

$$(3.3) \qquad f(s(x^{(N)})|s(x^{(N)}, x_{(M)}))$$

which is now defined as being proportional to the predictive likelihood i.e.

(3.4) $\quad f(s_N|s_{N+M}) \propto \text{prlk}\ (x^{(N)}|x_{(M)})$.

This is then treated as is the usual $L(x|\theta)$ where now $X_{(M)}$ takes on the role of θ. For the fixed value $x^{(N)}$, the predictive likelihood orders the plausibility for various values $X_{(M)} = x_{(M)}$. For a simple example, consider X_i, $i = 1,\ldots, N + M$ as Bernoulli iid random variables where $P(X_i=1) = 1-P(X_i=0) = \theta$. If r out of the first N are 1's, we can order possible predictive values for the number of 1's, say t, in the next M trials. Defining $R = \sum_{i=1}^{N} X_i$, $T = \sum_{i=1}^{M} X_{N+i}$, which are sufficient, we can compute in a simple fashion

$$(3.5) \qquad P\left[R=r \,|\, R+T = r+t\right] = \frac{\binom{N}{r}\binom{M}{t}}{\binom{N+M}{r+t}} \propto \text{prlk}\ (r|t)$$

which is used to order the plausible values for $t=0,\ldots,M$.

A point predictor can conceptually be obtained by maximizing the predictive likelihood. In the case where $M > 1$ and the random variables are iid, it is clear that $\text{prlk}\ (x_{(M)})$ will have multiple maxima due to the exchangeability of the likelihood. This must be so and should be no cause for concern. In the previous example though, there may be a unique maxima at some value of t and be adequate if t is to be predicted. It is clear, however, that if the individual $X_{N+1},\ldots, X_{N+M}$ are to be predicted and the maximum was at $t = t_o$, say, then every partition of $x_{N+1},\ldots,x_{N+M}$ into t_o 1's and $M-t_o$ 0's would also yield identical maxima of the $\text{prlk}\ (x_{(M)})$.

For a variety of interesting applications of predictive likelihood to standard statistical situations, the reader is referred to Hinkley (1975).

4. Low Structure and Assessment

Before actually discussing techniques available in low structure situations it will be useful to review a very old and informal method of considerable value in comparing point predictors. Suppose several predictors are suggested for a set of data, then a fruitful comparison of them may be accomplished by a validation technique. The sample $x^{(N)}$ is randomly divided into two parts $x^{(N-n)}=(x_1,\ldots,x_{N-n})$ and $x^{(n)} = (x_{N-n+1},\ldots,x_N)$ called the construction sample and the validation sample respectively. Assume also that associated with each sample point x_i is a known value z_i. The data analyst then computes the competing predictors from the construction sample obtaining, say, $\hat{x}_{ji}(x^{(N-n)}, z^{(N-n)}; z_j) = \hat{x}_{ji}$ as the i^{th} predictor for the value x_j at known value z_j, $j = N-n+1,\ldots, N$; $i = 1,\ldots, K$ where K represents the number of predictors to be compared, and $z^{(N-n)} = (z_1,\ldots,z_{N-n})$. First the residuals $\hat{x}_{ji} - x_j = r_{ji}$ are computed and then the empirical distribution functions of residuals are plotted for each predictor. A comparison of these empirical distribution functions will shed much light in determining which predictor is most appropriate. Sometimes when the validation sample is not very large a relevant summary measure of the predictive discrepancy is adequate for comparison. For example we might compute the predictive mean squared error $s_i^2 = (N-n)^{-1} \sum_{j=N-n+1}^{N} r_{ij}^2 \quad i=1,\ldots,K.$

This procedure is generally useful only when a reasonably large number of observations is at hand. This is often not the case. Also the procedure seems inefficient in that it does not extract all of the information in the data. To overcome this a technique which is referred to as simple cross-validation may be

substituted.

Let $x_j^{(N-1)} = (x_1,\ldots,x_{j-1}, x_{j+1},\ldots,x_N)$ with corresponding $Z_j^{(N-1)} = (z_1,\ldots,z_{j-1}, z_{j+1},\ldots,z_j)$ be the data set with the j^{th} observation omitted. Now for each predictive function we compute the predictor $\hat{x}_{ji} = \hat{x}_{ji}\,(x_j^{(N-1)}, Z_j^{(N-1)}, z_j)$ for the omitted observation x_j and repeat this for $j=1,\ldots,N$ for each predictor obtaining $r_{ij} = \hat{x}_{ji} - x_j$. Similarly as in the validation set up, we are in a position to compare for each predictor its empirical distribution function or a relevant summary measure of predictive discrepancy. However in the case of simple cross validation we have N residuals for each predictor instead of n as in the validation case. One caution is in order -- in the validation case the residuals are dependent only by virture of the same predictive function while in the simple cross-validation some further algebraic dependence creeps in as a result of using the data repetitively. On the other hand the simple cross-validation assessment uses all of the data while the validation assessment only uses a sample of the data. Notwithstanding, the cross-validatory assessment procedure is certainly very useful for the comparison of predictors generated from various structural assumptions as the basic dependence is the same for all of them.

However there are situations where specification of a particular sampling distribution and the resultant predictor based on such assumptions may be fraught with peril. When a particular sampling paradigm becomes difficult or impossible to identify, and yet prediction is necessary, data analytic techniques based on minimal assumptions need come to the fore. One such technique, termed predictive sample reuse (PSR), Geisser (1974a,1975a) or cross-validatory choice, Stone (1974a), is currently a leading candidate for a satisfactory resolution

of this low structure case. It may also be of service in what are basically higher structure situations as we will detail later. First of all the PSR method, when flexibly used, is very likely to be robust for a variety of sampling paradigms. A second feature is that it simulates the predictive process upon itself in some optimal fashion often using some structural hints. It is even capable in one of its manifestations of comparing a variety of approaches. Essentially the goal is to predict a future observation or set of such, or some function of them. For the purposes of this exposition we shall restrict ourselves to a single future observation with a form arbitrarily chosen for predicting it as

$$(4.1) \qquad x = x(X,Z,z;\alpha) \quad \alpha \in \Omega$$

where α is some set of unknown values, $X = (x_1,\ldots,x_N)$ represents a sample of size N and with each x_i is associated a known z_i, and $Z = (z_1,\ldots,z_N)$. It must be stressed that in this approach α is not a platonic ideal nor in any sense a true value of paramount importance. It is to be regarded as merely a convenient way of forming a predictive function. Let $P_i^{(N-n)}$ represent the i^{th} partition of the sample $N-n$ retained and n omitted observations $0 < n \leq M$, where M is the largest integer such that the predictive function (4.1) can be formed with $N-M$ observations. More precisely, the observational set X and the set Z with which it is associated are partitioned such that

$$(4.2) \qquad P_i^{(N-n)} = (X_{ir}^{(N-n)},\ Z_{ir}^{(N-n)};\quad X_{io}^{(n)},\ Z_{io}^{(n)})$$

is the i^{th} partition belonging to a set Γ of partitions relevant to a particular schema of observational omissions where $(X_{ir}^{(N-n)},\ Z_{ir}^{(N-n)})$ and $(X_{io}^{(n)},\ Z_{io}^{(n)})$ represent the $N-n$ retained and n omitted data sets, respectively. Let the total number of such partitions be $P(N, n, \Gamma)$, or simply P. The specified

predictive function is then applied to the retained observations for prediction of the omitted observations for each partition with the unknown set of values α estimated by means of optimizing an average discrepancy measure, say,

$$(4.3) \qquad D_{N,n}(\alpha) = P^{-1}n^{-1} \sum_{i\in\Gamma} d(X_{io}^{(n)}, \hat{X}_{io}^{(n)}(X_{ir}^{(N-n)}, Z_{ir}^{(N-n)}, Z_{io}^{(n)}; \alpha))$$

where each element in the set $\hat{X}_{io}^{(n)}$ is the form of the predictive function and d is a measure of the discrepancy of the set of values $X_{io}^{(n)}$ from the set of predicted values $\hat{X}_{io}^{(n)}$ for given α. $D_{N,n}(\alpha)$ is then optimized with respect to α in some sense. On the basis that this leads to a solution say, $\hat{\alpha}$, we obtain the predictor $\hat{x} = x(X,Z,z;\hat{\alpha}) = \hat{f}$.

When predictive functions are to be compared irrespective of their generation one can use a cross-validatory assessment. For a given discrepancy measure we could consider for the i^{th} partition the set of retained observations and associated values $(X_{ir}^{(N-n)}, Z_{ir}^{(N-n)})$ and partition this into two sets $(X_{irr}^{(N-2n)}, Z_{irr}^{(N-2n)}; X_{iro}^{(n)}, Z_{iro}^{(n)})$. From this reduced set of $N-n$ observations and associated values we would, as previously, obtain an $\hat{\alpha}_i$ and compute the discrepancy (not necessarily based on the same d as was used to obtain the predictor) between the values predicted for the n omitted observations and the actual observations themselves. Repeating this for each i we would then compute an overall discrepancy measure

$$(4.4) \qquad D_{N-n}^{*} = P^{-1}n^{-1} \sum_{i\in\Gamma} d(X_{io}^{(n)}, \hat{X}_{io}^{(n)}(X_{ir}^{(N-n)}, Z_{ir}^{(N-n)}, Z_{io}^{(n)}; \hat{\alpha}_i))$$

for each predictive function. This measure then would be relevant to assessing either different predictive functions or various estimators of α in terms of predictive discrepancy for the same predictive functions. We also note that comparisons other than the average D_{N-n}^{*} can be utilized, e.g., empirical

distributions of the discrepancy can be compared for several predictors. A variety of applications of PSR can be found in the following papers, Geisser (1974a, 1974b, 1975a, 1975b), Stone (1974a, 1974b). Here we shall only present one such very simple application involving a data based predictor which is to be combined with limited prior information. Let the predictive function be

$$(4.5) \qquad f = \alpha\, h\,(X) + (1-\alpha)\, g \qquad 0 \le \alpha \le 1$$

where g represents a prior guess at the value to be predicted and $h\,(X)$ the data based predictor. We shall use the squared discrepancy measure, with a one-at-a-time omission schema so that

$$(4.6) \qquad D_{N,1}(\alpha) = N^{-1} \sum_{j=1}^{N} (\alpha h_j + (1-\alpha)\, g - x_j)^2$$

where h_j is of the form h, but based on $N-1$ observations, i.e. x_j has been omitted. Maximization of $D_{N,1}(\alpha)$ with respect to α yields

$$(4.7) \qquad \begin{cases} \hat{f} = h & \text{if } \hat{\alpha} \ge 1 \\ \quad = g & \text{if } \hat{\alpha} \le 0 \\ \quad = \hat{\alpha}\, h + (1-\hat{\alpha})g & \text{otherwise} \end{cases}$$

where

$$(4.8) \qquad \hat{\alpha} = \frac{\sum_{j=1}^{N}(h_j-g)(x_j-g)}{\sum_{j=1}^{N}(h_j-g)^2}.$$

In particular if $h = \bar{x}$ then for $s^2 = (N-1)^{-1}\sum_{j=1}^{N}(x_j-\bar{x})^2$ and $t^2 = N(\bar{x}-g)^2/s^2$

$$(4.9)\quad \begin{cases} \hat{\alpha} = \dfrac{t^2 - 1}{t^2 + (N-1)^{-1}} & \text{if } t^2 > 1 \\ = 0 & \text{otherwise.} \end{cases}$$

This procedure has the property that if the sample mean is within one sample standard deviation of the mean from the prior guess g one uses g otherwise one uses the linear combination. Further as the distance between the sample mean and g increases relative to the sample standard deviation, greater weight is attached to the sample mean. Moreover as N increases the predictor tends asymptotically to the sample mean.

In many applications it would appear that observational omissions one-at-a-time are appropriate. However there are some applications where this may not be the case. This point and others involving various schemata of omissions and choice of relevant partitions are discussed in Geisser (1975a).

There have also been various attempts to extend PSR point prediction to sets, intervals and regions. It is not yet clear as to how satisfactory any of these methods are. Pertinent references are Geisser (1974b), Hinkley (1975), Butler and Rothman (1975).

5. An Application

We now illustrate how some of the previous methodology might be applied in practice to what may be termed a simple survival situation. Suppose we have a random sample $X_1,\ldots,X_N$ on an exponential random variable X whose density is

$$(5.1) \qquad f(x|\mu) = \mu e^{-\mu x} \quad \mu > 0,\ x > 0.$$

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

$$(5.2) \qquad p(\mu) \propto \mu^{\delta-1} e^{-\gamma\mu}, \quad \gamma > 0,\ \delta > 0.$$

Here μ takes the place of θ in the high structure Bayesian approach and $\tau = (\delta,\gamma)$. Our interest is in predicting a value x_{N+1} for the random future observation X_{N+1} given the previous N observations $x^{(N)}$, say. Then the predictive density for X_{N+1} is easily calculated to be

$$(5.3) \qquad f(x_{N+1}|x^{(n)}) = \int p(\mu|x^{(N)})\ f(x_{N+1}|\mu)\,d\mu$$

$$= (N+\delta)(N\bar{x}+\gamma)^{N+\delta}/(N\bar{x}+\gamma+x_{N+1})^{N+\delta+1} \qquad z > 0,$$

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

$$(5.4) \qquad f(x) = \int f(x|\mu)p(\mu)\,d\mu$$

$$= \delta\gamma^{\delta}/(\gamma+x)^{\delta+1}, \quad x > 0.$$

Hence it is convenient and more appropriate from the predictive view 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

$$(5.5) \quad \begin{cases} E(X) = \gamma/(\delta - 1) = g \\ Var(X) = \delta\gamma^2/(\delta - 2)(\delta - 1)^2 = g^2(1 + \alpha)/(1 - \alpha) \end{cases}$$

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 (5.5) and substituting them in (5.3).

It is to be noted that (5.3) and (5.4) were obtained from (5.1) and (5.2). However, for the predictivist who would prefer to start from (5.1) and (5.4) 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 possible here because $f(x)$ is the unique Laplace transform of $\mu^{-1}p(\mu)$.

Now as we mentioned previously positing 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. So in further analysis of this problem we may be in a position such that some of the parameters of τ are assumed known and others unknown. Assume then that g is known but not α.

One approach for estimating α or δ is from the marginal density

$$(5.6) \qquad f(x_1,\ldots,x_N|\delta,\gamma) = \int f(x_1,\ldots,x_N|\mu)p(\mu|\delta,\gamma)\,d\mu$$

$$= \frac{\Gamma(N+\delta)\,\gamma^{\delta}}{\Gamma(\delta)\,[N\bar{x}+\gamma]^{N+\delta}}$$

Since we assume $g = \frac{\gamma}{\delta-1}$ is known we let $Y_i = g^{-1}X_i$ and obtain for $N\bar{y} = \sum_{i=1}^{N} y_i$

$$(5.7) \qquad f(Y_1,\ldots,Y_N|\delta) = \frac{\Gamma(N+\delta)\,(\delta-1)^{\delta}}{\Gamma(\delta)\,[N\bar{y}+\delta-1]^{N+\delta}}$$

Clearly $\sum_{i=1}^{N} Y_i = S$ is sufficient for δ in the above likelihood. The density of S is then easily obtained to be

$$(5.8) \qquad f(s|\delta) = \frac{(\delta-1)^{\delta}\,\Gamma(N+\delta)\,s^{N-1}}{\Gamma(N)\,\Gamma(\delta)\,(s+\delta-1)^{N+\delta}}$$

which implies that $\alpha S \sim \beta_2(\alpha S; N, \delta)$ a Beta distribution of the second kind. The method of moments essentially fails here to yield a sensible estimate e.g. $E(S) = N$, which is uninformative relative to δ or α. Use of higher moments tends to restrict the range of $\hat{\delta}$ and renders it unreasonable as an estimator. The reason that moment estimators are basically inappropriate here is that they assume the existence of the moments used and hence tend to presume a restriction on the range of δ, whose restriction on the outset is $\delta > 1$. One can use however maximum likelihood estimation. Hence we calculate

$$(5.9) \qquad \frac{\partial \log f}{\partial \delta} = \log\frac{\delta-1}{s+\delta-1} + \frac{\delta}{\delta-1} + \frac{1}{\delta} + \frac{1}{\delta+1} + \cdots + \frac{1}{N-1+\delta} - \frac{N+\delta}{s+\delta-1}$$

and one would have to find by one means or another $\hat{\delta}$ satisfying $\frac{\partial \log f}{\partial \delta} = 0$. An explicit solution for $\hat{\delta}$ seems impossible to achieve. One can approximate (5.9) by using the Euler-Maclauren sum formula so that we obtain for large N

$$(5.10)\qquad \frac{\partial \log f}{\partial \delta} \doteq \frac{\delta}{\delta-1} - \log\frac{\delta}{\delta-1} + \log\frac{N+\delta}{s+\delta-1} - \frac{N+\delta}{s+\delta-1} + \frac{1}{2\delta} - \frac{1}{2(\delta+N)}$$

This is still quite formidable and when set equal to zero still does not yield an explicit solution for δ.

We now show how PSR may be of service even in this high structure situation. Suppose we were to predict a single value x_{N+1} from (5.3) using the predictive mean

$$(5.11)\qquad E(X_{N+1}|\bar{X} = \bar{x}) = (\alpha N\bar{x} + g)/(\alpha N+1).$$

Apply the PSR method for the estimation of α using (5.11) as a predictive function and squared discrepancy with one-at-a-time omission schema so that

$$(5.12)\qquad D_{N,1}(\alpha) = N^{-1}\sum_{j=1}^{N}\left(\frac{\alpha(N-1)\bar{x}_j+g}{\alpha(N-1)+1} - x_j\right)^2$$

where $\bar{x}_j$ is the mean of the observation with x_j omitted. Minimization of $D_{N,1}(\alpha)$ with respect to α yields

$$(5.13)\qquad \begin{cases} \hat{\alpha} = \dfrac{t^2-1}{N} & \text{for } t^2 > 1 \\ \hat{\alpha} = 0 & \text{for } t^2 \le 1 \end{cases}$$

where $t^2 = N(g-\bar{x})^2/s^2$ and $s^2 = N^{-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$. Hence PSR may be used to generate estimates even in the high structure case. On the other hand using (5.11) and (5.12) as a predictive function and discrepancy measure respectively yields a PSR predictor

$$(5.14)\qquad \hat{x}_{N+1} = (\hat{\alpha} N\bar{x} + g)/(\hat{\alpha} N+1)$$

that does not strictly depend on high structure assumptions. In fact it may be robust for a variety of high structure assumptions which result in a predictive expectation approximately equal to (5.11). Actually if one did not use any high

structure hint for a predictive function for this problem but merely used a convex combination of sample mean and prior guess

$$(5.15) \qquad x_{N+1} = \alpha^* \bar{x} + (1-\alpha^*) \qquad 0 \le \alpha^* \le 1,$$

then the result for $\hat{\alpha}^*$ was already obtained in section 4 as

$$(5.16) \quad \begin{cases} \hat{x}_{N+1} = \dfrac{(t^2-1)\,\bar{x} + \dfrac{N}{N-1}\,g}{t^2 + (N-1)^{-1}} & \text{if } t^2 > 1 \\ = g & \text{if } t^2 \le 1 \end{cases}$$

This may be contrasted with (5.14) when the value for $\hat{\alpha}$ is inserted which turns out to be

$$(5.17) \quad \begin{cases} \hat{x}_{N+1} = \dfrac{(t^2-1)\,\bar{x} + g}{t^2} & t^2 > 1 \\ = g & t^2 \le 1. \end{cases}$$

The predictor in (5.17) is weighted slightly more towards $\bar{x}$ than (5.16), but in fact they are asymptotically equivalent to order N^{-1}. In any practical example there would probably not be much to choose between them.

It is also to be noted that the intermediate structures are difficult or impossible to apply in situations such as this one where there may be some prior information that should be taken into account.

6. Remarks

A somewhat abbreviated exposition of the predictivistic view has been presented. This view is not a mode of inference as such but can be implemented from a variety of inferential modes. It stems from the attitude that inferences should be restricted to potentially observable entities unless compelling reasons to contrary exist. In conformance with this view we have presented various ways, arising from different standpoints, of implementing the predictive approach. In particular a recently developed low structure approach PSR has also been delineated in somewhat greater detail, which should be of great value in many situations and need be added, we believe, to the toolkit of every statistician.

References

Barnard, G. A. (1975). New foundations of statistical inference. Unpublished lecture notes, University of Minnesota.

Butler, R. and Rothman, E. D. (1975). Intervals based on reuse of the sample. Technical Report 56, University of Michigan.

Fisher, R. A. (1956). Statistical Methods and Scientific Inference. (1st Ed.) New York: Hafner.

Fraser, D. A. S. (1968). The Structure of Inference, New York: John Wiley and Sons.

Geisser, S. (1971). The inferential use of predictive distributions. In Foundations of Statistical Inference (V. P. Godambe and D. A. Sprott, eds.) 456-69. Toronto, Montreal: Holt, Rinehart and Winston.

Geisser, S. (1974a). A predictive approach to the random effect model. Biometrika, 61, 101-107.

Geisser, S. (1974b). A new approach to the fundamental problem of applied statistics. Technical Report 235, University of Minnesota.

Geisser, S. (1975a). The predictive sample reuse method with applications. J. Amer. Statist. Assoc. 70, 350, 320-328.

Geisser, S. (1975b). Bayesianism, predictive sample reuse, pseudo observations, and survival. Bulletin of the International Statistical Institute (to appear).

Hinkley, D. V. (1975). Predictive Inference. Technical Report 254, University of Minnesota.

Lauritzen, S. L. (1974). Sufficiency, prediction and extreme models. Scand. J. Statist. 1, 128-34.

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

Stone, M. (1974b). Cross-validation and multinomial prediction. Biometrika, 61, 3, 509-515.