A Comparison of Several Model Selection Procedures*

by

Murray K. Clayton
Seymour Geisser
Dennis E. Jennings

University of Minnesota

Technical Report No. 401
May 1982

*Research supported in part by NIH Grant GM 25271.

A Comparison Of Several Model Selection Procedures*

by

Murray K. Clayton, Seymour Geisser and Dennis E. Jennings

University of Minnesota

1. Introduction

Econometric and other statistical models are often simplifications of extremely complicated phenomena and it is a mistake to assume that any particular model is actually a true representation of the underlying process. What is assumed or hoped is that the model may be an adequate description and perhaps potentially useful for some purpose. Hence it is often puzzling why there has been so much effort, especially in the softer social sciences, devoted to "testing" parameters of a model as if they were true entities and not as in most instances, convenient artifices. A more substantial enterprise than testing should be model selection, i.e. selecting one of several alternative models such that the selected model (irrespective of its truth) would serve best some purpose of the investigator (descriptive or predictive, perhaps). Hopefully, the selected model also represents some reasonable approximation to the truth, complicated as that may be.

In theory, there is no intrinsic difficulty for the Bayesian statistician who can determine a prior probability q_k for each potential model M_k that could have generated the set of data, and who can also determine, given a particular model M_k, a proper prior density for the parameters specified by the model. A posterior probability q_k' is obtained for each model and a model selected based on all of the q_k' and a cost or penalty for choosing an incorrect model. If prediction of a future value Y is at issue than one calculates its predictive distribution as

*Research supported in part by NIH Grant GM 25271.

$$F_p(y|D) = \sum_k q_k' \, F_p(y|D, M_k) \qquad (1.1)$$

where D represents the data in hand and $F_p(y|D, M_k)$ is the predictive distribution function conditional on Y having been generated by M_k. Conceptually all predictive inferences and decisions flow from (1.1). For example a point predictor that minimizes squared error is

$$E(Y|D) = \Sigma_k q_k' \, E(Y|D, M_k) \; . \qquad (1.2)$$

Other loss functions can bring to the fore the predictive mode or median of Y.

But either for the investigator or the statistician, the situation is seldom so straightforward. As already indicated, the set of statistical models employed rarely includes an exact representation of the process, and even if it did, prior probabilities for either the various models or the parameters for any given M_k are rarely specifiable with any degree of confidence. Moreover most investigators, if not statisticians, are unsure about their preference for a loss function.

With this in mind, it appears of some importance to be able to devise reasonable and useful model selection procedures that do not depend on such tight specifications and evaluate them from more than one standpoint so as to make the rationale for their use as convincing as possible. Former reservations notwithstanding it is still of some interest (of prime interest for classical statisticians, certainly) to determine the rate at which a model selection procedure yields the true model when in fact the true model is one of the alternatives considered. A second approach--which some might consider primary--is to attempt to assess various selection procedures in terms of their predictive capacity, recognizing again the fact that, more than likely, none of the models considered is the true one and the best that can be asked of competitive model selection procedures is to find that procedure which appears to possess in some sense the best predictive capability over a broad spectrum of possibilities.

In this paper we initiate such a study of four model selection procedures with comparisons restricted to two common distribution families. We determine, by simulation, estimates of the frequency with which each chooses a "true" model and by using a prediction rule compare the selection procedures for a number of situations on the basis of squared prediction error.

In making our comparisons, we have restricted ourselves to the situation where the "true" model is among the alternatives considered in order to contrast the comparisons of correct selection rates and squared errors of prediction. Also, we decided to employ a fairly simple-minded prediction procedure or rule which was to use only the model selected in contrast to what is indicated by (1.1). It is our belief that most investigators commonly would use such a "rule" and it is of interest to make our relative comparisons on this basis.

2. Selection Criteria

Suppose we have available a set of data $\underset{\sim}{x} = (x_1, \ldots, x_N)$, presumably a realization of the random variable $\underset{\sim}{X} = (X_1, \ldots, X_N)$. Several potential models $M_1, \ldots, M_m$ are entertained as having possibly generated or offered a satisfactory explanation for the data or more directly may be useful for predicting future values from the underlying process which generated the data.

We further assume that M_k specifies a joint density $f(\underset{\sim}{x}|\underset{\sim k}{\theta}, M_k)$ where $\underset{\sim k}{\theta}$ is some set of unknown parameters.

In such situations several procedures have been put forth for choosing the most appropriate of the entertained models. Akaike (1973), using information theory, devised what has come to be known as the Akaike information criterion (AIC) which is equivalent to selecting that model which maximizes

$$A_k = e^{-p_k} f(\underset{\sim}{x}|\underset{\sim k}{\hat{\theta}}, M_k) = e^{-p_k} L_k(\underset{\sim k}{\hat{\theta}}) \qquad (2.1)$$

where $\hat{\theta}_{\sim k}$ is the maximum likelihood estimator of $\theta_{\sim k}$ and p_k represents the number of unknown parameters. For regression problems this is equivalent to the "C_p" criterion of Mallows (1973).

Schwarz (1978) developed an asymptotic expansion to an exact Bayes procedure which for nested situations assigns positive probability to lower dimensional subspaces of the parameter space. He employs the two leading terms of the expansion as a basis for selecting the model which is a posteriori most probable. His work includes as special cases an earlier effort by Jeffreys (1967) and a further development for regression situations by Zellner and Siow (1980), for large enough samples. The method is equivalent to selecting the model which maximizes

$$S_k = e^{-p_k \log\sqrt{N}} L_k(\hat{\theta}_{\sim k}) \ . \qquad (2.2)$$

This large sample Bayes (LSB) criterion assumes that there is a fixed and equal penalty (with possibly minor perturbations) for choosing the wrong model. When this latter assumption is made, Schwarz points out that Akaike's AIC cannot be asymptotically optimal. Indeed AIC is not even consistent! Although Akaike did not explicitly state what loss or penalty he considered or even if he considered one at all, it is easily shown and alluded to in Geisser and Eddy (1979) and more precisely in Geisser (1980) that AIC, for the same assumptions as Schwarz made, is also asymptotically Bayes but with penalties that depend on the sample size and the kind of selection error made. This is immediately apparent by comparing (2.1) with (2.2) and noting

$$A_k = (\sqrt{N} \ / \ e)^{p_k} S_k \ . \qquad (2.3)$$

Hence if $M_k \subset M_{k'}$ then, in the light of (2.3), the selection of M_k over $M_{k'}$ requires that

$$S_k > (\sqrt{N}/e)^d S_{k'}$$

where $d = p_{k'} - p_k$. Hence the relative penalty (the coefficient of $S_{k'}$) increases rapidly with N for $N \geq 8$. This implies that for Akaike's procedure a severe relative penalty is incurred when a "false" lower dimensional model is selected as opposed to selecting a "false" higher dimensional model. For many prediction problems involving nested models this often makes good sense.

For example assume that n observations are taken from each of two populations Π_1 and Π_2 where Π_i is $N(\mu_i, \sigma^2)$ and two models are entertained, namely

$$M_1: \mu_1 = \mu_2 \; ; \; \sigma^2 \text{ unspecified}$$

$$M_2: \mu_1 \neq \mu_2 \; ; \; \sigma^2 \text{ unspecified} \; .$$

Here it can easily be shown that the probabilities of correct selection, when M_1 is true, for the criteria A_k and S_k are respectively

$$C(A) = \Pr[F_{1,\, 2(n-1)} \leq 2(n-1)(e^{\frac{1}{n}}-1)]$$

and

$$C(S) = \Pr[F_{1,\, 2(n-1)} \leq 2(n-1)[(2n)^{\frac{1}{2n}}-1]$$

where $F_{a,b}$ is an F variable with a and b degrees of freedom.

It is easily shown that $C(A) \leq C(S)$ for all $n \geq 4$ and

$$\lim_{n\to\infty} C(A) = \Pr[\chi_1^2 \leq 2] \doteq .843$$

$$\lim_{n\to\infty} C(S) = 1 \; .$$

On the other hand, it is clear that the power $1-\beta$ of these criteria are such that $1-\beta_A \geq 1-\beta_S$ for all $n \geq 4$ and that both tend to 1 as n grows.

Moreover, when M_1 is true but M_2 is chosen by the selection process, the squared error of any sensible predictor of future values from Π_i will tend to σ^2 as n grows. For prediction purposes consistency of an error of the first kind may be largely irrelevant in this nested situation.

Geisser and Eddy (1979) put forth two model selection criteria based on predictive sample reuse (PSR) methods suitable for independently distributed variates such that X_j has density $f(x_j|\theta_k, M_k)$. The first, termed PSR quasi-likelihood (PSRQL) selects that model which maximizes

$$\hat{L}_k = \prod_{j=1}^{N} f(x_j|\hat{\underset{\sim}{\theta}}_{k(j)}, M_k) \tag{2.4}$$

where $\hat{\underset{\sim}{\theta}}_{k(j)}$ is the MLE of $\underset{\sim}{\theta}_k$ with x_j omitted. The second criterion, termed PSR quasi-Bayes, selects that model which maximizes

$$L_k = \prod_{j=1}^{N} f_p(x_j|\underset{\sim}{x}_{(j)}, M_k) \tag{2.5}$$

where $\underset{\sim}{x}_{(j)}$ denotes that x_j has been deleted from $\underset{\sim}{x}$ and where $f_p(\cdot)$ represents a predictive density. The latter is calculated as follows

$$f_p(x_j|\underset{\sim}{x}_{(j)}, M_k) = \int f(x_j|\underset{\sim}{\theta}_k, M_k)\, dP(\underset{\sim}{\theta}_k|\underset{\sim}{x}_{(j)}, M_k) \tag{2.6}$$

and $P(\underset{\sim}{\theta}_k|\underset{\sim}{x}_{(j)}, M_k)$ is a posterior distribution of $\underset{\sim}{\theta}_k$ based on $\underset{\sim}{x}_{(j)}$ and usually a diffuse prior on $\underset{\sim}{\theta}_k$. A complete explication of these methods appears in Geisser and Eddy (1979). In comparing models in a nested situation when a lower dimensional model M_k is assumed to be true, then for a reasonable class of priors, it can be shown that $A_{k'}$, $\hat{L}_{k'}$ and $L_{k'}$ are asymptotically equivalent for any $M_{k'}$ which contains M_k.

In order to have an idea of how these methods perform in practice we shall compare them in two cases. First we shall use the simple exponential distribution and two models. Under M_1 we assume that a dichotomously labeled random sample $\underset{\sim}{X} = (\underset{\sim}{X}_1, \underset{\sim}{X}_2)$, $X_i = (X_{i1},\ldots X_{iN_i})$ $i = 1,2$, is a set of $N_1+N_2 = N$ iid random variables each with density

$$f(x|\lambda,M_1) = \lambda e^{-\lambda x}$$

i.e. the sampling distribution does not depend on the label. Under M_2 we assume that the label is relevant and X_{ij} is a random sample from

$$f(x|\lambda_i, M_2) = \lambda_i e^{-\lambda_i x} \qquad i=1,2\,; \quad j=1,\ldots, N_i\,.$$

The AIC criterion compares

$$A_1 = (e^{N+1}\,\bar{x}^N)^{-1}$$

with

$$A_2 = (e^{N+2}\,\bar{x}_1^{N_1}\,\bar{x}_2^{N_2})^{-1}$$

where $N_i\bar{x}_i = \sum_{j=1}^{N_i} x_{ij}$, $N\bar{x} = N_1\bar{x}_1 + N_2\bar{x}_2$ for $i=1,2$.

The large sample Bayes procedure as given by Schwarz compares

$$S_1 = (N^{\frac{1}{2}}e^N\,\bar{x}^N)^{-1}$$

with

$$S_2 = (N\,e^N\bar{x}_1^{N_1}\bar{x}_2^{N_2})^{-1}\,.$$

The PSRQL Method compares

$$\hat{L}_1 = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \left(\frac{N-1}{N\bar{x}-x_{ij}} \right) \exp \left[- \frac{(N-1)x_{ij}}{N\bar{x}-x_{ij}} \right]$$

with

$$\hat{L}_2 = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \left(\frac{N_i-1}{N_i\bar{x}_i-x_{ij}} \right) \exp \left[- \frac{(N_i-1)x_{ij}}{N_i\bar{x}_i-x_{ij}} \right] .$$

For the PSRQB method we assume prior densities for λ_i to be of the usual non-informative type namely,

$$g(\lambda_i) \propto \lambda_i^{-1} .$$

With this formulation [Geisser and Eddy (1979)], we compare

$$L_1 = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \frac{(N-1)(N\bar{x}-x_{ij})^{N-1}}{(N\bar{x})^{N}}$$

with

$$L_2 = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \frac{(N_i-1)(N_i\bar{x}_i-x_{ij})^{N_i-1}}{(N_i\bar{x}_i)^{N_i}} .$$

For the second case we assume that the data are samples from one of two normal populations with density

$$f(x|\mu_i, \sigma_i^2) = (2\pi\sigma_i^2)^{-\frac{1}{2}} \exp \left(- \frac{1}{2\sigma_i^2} (x - \mu_i)^2 \right), \quad i=1,2 . \tag{2.7}$$

Suppose, however that there are three possible models (nested)

$$M_1: \mu_1 = \mu_2 , \ \sigma_1^2 = \sigma_2^2 \tag{2.8}$$

$$M_2: \mu_1 \neq \mu_2 , \ \sigma_1^2 = \sigma_2^2$$

$$M_3: \mu_1 \neq \mu_2 , \ \sigma_1^2 \neq \sigma_2^2 .$$

Here the AIC criterion selects the model M_k according to the largest A_k where

$$A_1 = e^{-\frac{N+4}{2}} (2\pi S^2)^{-N/2}$$

$$A_2 = e^{-\frac{N+6}{2}} (2\pi T^2)^{-N/2}$$

$$A_3 = e^{-\frac{N+8}{2}} (2\pi)^{-\frac{N}{2}} (S_1^2)^{-N_1/2} (S_2^2)^{-N_2/2}$$

where

$$NS^2 = \sum_i \sum_j (x_{ij}-\bar{x})^2$$

$$NT^2 = \sum_i \sum_j (x_{ij}-\bar{x}_i)^2$$

$$N_i S_i^2 = \sum_j (x_{ij}-\bar{x}_i)^2 \quad .$$

The LBS criterion selects M_k according to the largest of

$$S_1 = N^{-1} e^{-N/2} (2\pi S^2)^{-N/2}$$

$$S_2 = N^{-3/2} e^{-N/2} (2\pi T^2)^{-N/2}$$

$$S_3 = N^{-2} e^{-N/2} (2\pi S_1^2)^{-N_1/2} (2\pi S_2^2)^{-N_2/2} \quad .$$

For the PSRQL criterion the relevant comparison is among

$$\hat{L}_1 = \prod_{i=1}^{2}\prod_{j=1}^{N_i} (2\pi T^2_{(ij)})^{-\frac{1}{2}} \exp\left[- \frac{(x_{ij}-\bar{x}_{(ij)})^2}{2T^2_{(ij)}} \right]$$

$$\hat{L}_2 = \prod_{i=1}^{2}\prod_{j=1}^{N_i} (2\pi S^2_{(ij)})^{-\frac{1}{2}} \exp\left[- \frac{(x_{ij}-\bar{x}_{i(j)})^2}{2S^2_{(ij)}} \right]$$

$$\hat{L}_3 = \prod_{i=1}^{2}\prod_{j=1}^{N_i} (2\pi S^2_{i(j)})^{-\frac{1}{2}} \exp\left[- \frac{(x_{ij}-\bar{x}_{i(j)})^2}{2S^2_{i(j)}} \right]$$

where

$$(N-1)\bar{x}_{(ij)} = N\bar{x}-x_{ij}$$

$$(N_i-1)\bar{x}_{i(j)} = N_i\bar{x}_i-x_{ij}$$

$$(N-1)T^2_{(ij)} = \sum_{(t,k)\neq(i,j)} \sum (x_{tk}-\bar{x}_{(ij)})^2$$

$$(N-1)S^2_{(ij)} = N_i S^2_{i(j)} + N_{3-i} S^2_{3-i} \qquad i=1,2$$

$$(N_i-1)S^2_{i(j)} = \sum_{\substack{t=1 \\ t\neq j}}^{N_i} (x_{it}-\bar{x}_{i(j)})^2 \quad .$$

For the PSRQB criterion we use the usual improper prior density of the type $g(\mu_i, \sigma_i) \propto \sigma_i^{-1}$, and choose the model corresponding to the largest L_k.

Here

$$L_1 = \prod_{i=1}^{2}\prod_{j=1}^{N_i} \left[\frac{N-1}{\pi(N-2)N}\right]^{\frac{1}{2}} \frac{\Gamma[(N-1)/2]}{\Gamma[(N-2)/2]t_{(ij)}} \cdot \left[1 + \frac{(N-1)(x_{ij}-\bar{x}_{(ij)})^2}{N(N-2)t^2_{(ij)}} \right]^{-(N-2)/2}$$

where

$$(N-2)t^2_{(ij)} = (N-1)T^2_{(ij)} \quad ,$$

$$L_2 = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \left[\frac{N_i - 1}{\pi(N-3)N_i}\right]^{1/2} \frac{\Gamma[(N-2)/2]}{\Gamma[(N-3)/2]s_{(ij)}} \cdot \left[1 + \frac{(N_i-1)(x_{ij}-\bar{x}_{i(j)})^2}{N_i(N-3)s^2_{(ij)}}\right]^{-(N-2)/2}$$

where

$$(N-3)s^2_{(ij)} = (N-1)S^2_{(ij)}$$

$$(N_i-2)s^2_{i(j)} = (N_i-1)S^2_{i(j)} \quad ,$$

and

$$L_3 = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \left[\frac{N_i - 1}{\pi(N_i-2)N_i}\right]^{1/2} \frac{\Gamma[(N_i-1)/2]}{\Gamma[(N_i-2)/2]s_{i(j)}} \cdot \left[1 + \frac{(N_i-1)(x_{ij}-\bar{x}_{i(j)})^2}{N_i(N_i-2)s^2_{i(j)}}\right]^{-(N_i-1)/2} .$$

In both the normal and exponential cases discussed, the data are differentiated by a binary label and a predicted value is required for each label. A full Bayesian treatment follows similar to that in Section 1. Here we need to predict a future value for each label. Hence, the predictive density of the two values (Y_1, Y_2) is

$$f_p(y_1, y_2|D) = \sum_{k=1}^{m} q_k' \, f_p(y_1, y_2|D, M_k)$$

where q_k' is as in Section 1 and

$$f_p(y_1, y_2|D, M_k) = \int f(y_1, y_2|\underset{\sim}{\theta}_k, M_k) \, dP(\underset{\sim}{\theta}_k|D, M_k) \, .$$

All predictive inferences about Y_1, Y_2 would then flow from $f_p(y_1, y_2|D)$. In the absence of being able to execute this fully we will investigate and compare the use of the aforementioned model selection procedures with regard to correct selection rates and error of prediction when the selection method is conjoined with the simple predictive rule set forth in the previous section.

3. Simple Exponential Models

A simulation was performed to compare model selection rates and predictive squared error of the four procedures when used in the exponential setting described in Section 2. The relative size of the parameters for the two populations was varied by setting $\lambda_1 = 1$ and $\lambda_2 = 1(.5)2.5$. For each value of λ_2 a sample of size n was generated for each population[1], the model selection procedures were applied and the selected model noted. This was repeated 10,000 times and the proportion of correct selections for each procedure was recorded. The results for selected values of n ranging from 4 to 20 appear in Table 1. Our choices of λ_2, n, and the number of replications were aided by a preliminary simulation study by Seber (1979).

As noted in Table 1, these estimates of the probability of correct model selection have standard errors ranging from .004 to .005. When comparing different procedures for the same value of λ_2 and n, the standard error of a difference of estimated probabilities has been reduced by computing the estimates from the same data sets. In the discussion that follows in this paper, a difference will be called "significant" if it exceeds twice the standard error of that difference. We intend only to use this as a benchmark for making comparisons, and do not imply that a formal test is being made.

The behavior of the LSB procedure is distinct from the other three procedures. It is the only one of the procedures which is consistent, that is, its

---

[1]Generated by the inverse cummulative method with uniform pseudorandom numbers supplied by the University of Minnesota Computing Center's routine RAN2F.

correct selection rate goes to 1 as n increases under M_1. When M_1 is true, the probability of correct selection for the AIC, PSRQB and PSRQL procedures has an asymptotic value of .843 as noted in section 2. The first section of Table 1 reflects this asymptotic behavior. The LSB procedure makes more correct selections under M_1 and fewer correct selections under M_2 than the other procedures.

The other three methods act more similarly with the AIC perhaps preferred; the AIC is never significantly worse than PSRQB and is significantly better than PSRQL in all but a few parameter configurations. Under M_1, AIC is significantly better than PSRQL and PSRQB for small sample sizes. The PSRQL criterion is clearly the poorest under M_1 for sample sizes of 4 and 8, but is best at $\lambda_2 = 1.5$ for the same sample sizes. These are parameter assignments which are close to M_1 and thus the superiority of the PSRQL criterion there is not surprising in light of its poor performance under M_1. To see this, observe that under M_1, it selects M_2 more frequently than the other procedures. Not surprisingly, as we move slightly away from M_1, it still picks M_2 more often, thus doing better than the others. However, as we continue to move farther from M_1, or as we increase the sample size, AIC and the PSRQB method become clearly better.

Differences between AIC and PSRQB are small under M_2, although for two combinations ($\lambda_2 = 1.5$ with $n = 20$ and $\lambda_2 = 2.5$ with $n = 8$) AIC is significantly better than the PSRQB method. In no situation was the PSRQB method significantly better than the AIC.

Various prediction problems could arise in this setting; the one which we will consider here was discussed in section 2. The problem supposes a need to predict a future observation from each population (under M_1 they are identical populations) with loss measured as the sum of squared prediction errors. Also, we use the outcome of the model selection process to determine our predictions.

Specifically, the problem examined is to predict future observations labeled as Y_1 and Y_2 with Y_i distributed as an observation from the $i^{\underline{th}}$ population. If our predictors are $\hat{y}_1$ and $\hat{y}_2$, then the expected squared prediction error is $E[(Y_1-\hat{y}_1)^2 + (Y_2-\hat{y}_2)^2]$. Since $E(Y_i-\hat{y}_i)^2 = \text{Var } Y_i + E(\hat{y}_i-EY_i)^2$, we need only estimate the last term in the simulation. Doing so increases the accuracy of the estimate and eliminates the need to generate new observations, Y_i. Following Geisser and Eddy (1979), if M_1 is selected then we set $\hat{y}_1 = \hat{y}_2 = \bar{x}$, the grand mean, as the predictors; if M_2 is selected then $\hat{y}_i = \bar{x}_i$ for $i = 1,2$ are used. Table 2 gives the estimates of the expected squared error of prediction (SEP) for the cases examined.

Examination of the table shows the LSB method doing better under M_1 and usually poorer under M_2. The other three procedures seem indistinguishable with regard to prediction error. Surprisingly this is true even in those situations when one procedure clearly makes more correct model selections. Thus, only LSB yields significantly different prediction errors and the magnitude of these differences is not consequential. The largest relative difference in squared error is only 1.5% which occurs when $\lambda_2 = 2.0$ and $n = 20$. One might argue that in comparing estimates of SEP that they should be expressed in the same units as the original observations. We can do so by comparing the square roots of the estimated SEPs. In this case, the relative difference of transformed estimates is .8%. This corresponds to a 17 percentage point difference in probability of correct selection for this parameter assignment.

Three other columns are included in Table 2 to give some guidance in judging how well we are doing or can hope to do. The expected squared error of prediction when λ_i is known (SEPK) represents the unavoidable part of the loss which occurs due to the random variability of the observations to be predicted. Consequently, it can be viewed as the asymptotic limit of

the prediction error for given values of λ_1 and λ_2 as n increases. The prediction error which would occur if we knew the correct model (SEPC) but did not know the actual values of λ_1 and λ_2 represents how well a "foolproof" model selection procedure would do. The error incurred if we always use the wrong model (SEPI) is also given. In some cases (λ_2=1.5 with n = 4,8 or 12 and λ_2=2.0 with n = 4), it is better to be always wrong than always right! Of course, it is no easier to be always wrong than always right.

If we look more carefully at one of the cases where SEPI is less than SEPC, such as $\lambda_2 = 1.5$ and $n = 4$, the problem is a little clearer. In this case all of the model selection procedures have estimated squared errors of prediction which are less than SEPC and greater than SEPI. To see why this might be the case, consider using a predictor for Y_1 of the form $a\bar{x}+(1-a)\bar{x}_1$. One can show that for this parameter assignment the optimal predictor of this form has $a = \frac{16}{17}$. This predictor is very close to $\bar{x}$, which is used under SEPI. These cases emphasize that prediction is not coincident with model selection and that optimal squared error predictors need not be based on the acceptance of a single model, but on a combination of them--this, of course, is well known.

4. Normal Models

The correct selection rates and predictive squared error of the four selection criteria were also compared using simulated normal data (2.7). We consider a variety of combinations of the parameter values which conform to the possible model restrictions (2.8). For simplicity, we fix $\mu_1 = 0$, $\sigma_1 = 1$, and vary μ_2 and σ_2. Tables 3 and 4 contain the parameter configurations used.

Similar to Section 3, for each given parameter assignment a sample of size n $(=N_1=N_2)$ was generated from each population[2], the model selection criteria were applied, and the selected model was noted. We repeated this process 2000 times[3] and noted the proportion of correct model selections by each procedure. These results for $n = 10$ and $n = 20$ appear in Table 3. For reasons explained below, we also noted the proportion of selections of M_1; these data also appear in Table 3.

The estimates in Table 3 have standard errors which lie between .008 and .012. As in Section 3, the estimates within a given row of Table 3 are based on the same data sets. This yields estimates which are positively correlated, thus reducing the standard error of a difference in estimated probabilities. Typical standard errors for a difference range from .002 to .008.

Generally speaking, the AIC is best in terms of frequency of correct model selection, with some interesting exceptions. Under M_1, $(\mu_2,\sigma_2) = (0,1)$, the LSB criterion is constructed to be consistent, and so it is not surprising that it should dominate the other selection criteria when the sample size is large. It is perhaps a little surprising that it is clearly better for n as small as 20. When n is 10 no other method is significantly better. The other instances where the LSB method dominates the remaining methods are when $\sigma_2 = 1$ $(=\sigma_1)$ and μ_2 is large ($n=10$, $\mu_2 = 2,4$ and $n = 20$, $\mu_2 = 1,2,4$). To see why this might be so, note that due to the penalty structure imposed in the LSB method it tends to choose lower dimensional models. Thus, when $\sigma_2 = 1$ and μ_2 is large, while both the AIC and LSB procedures are fairly successful in determining that M_1 is incorrect, the AIC tends to pick M_3 too frequently.

---

[2]Using the University of Minnesota Computing Center's pseudorandom normal variate generator NORMAL.

[3]Except as noted in Table 4, where 10,000 repetitions were used to reduce the standard error of the estimates contained therein.

This is borne out by computing the proportion of times M_3 is chosen by each method. (These calculations can be made from the data in Table 3.) On the other hand, when M_3 is the true model, the AIC performs better than the LSB criterion. This results because the latter method tends to prefer lower dimensional models and thus frequently avoids choosing the correct model, M_3.

The situation becomes more difficult when one attempts to rank all four selection criteria in terms of their ability to select the correct model. Generally we see that the PSRQL and PSRQB methods have correct selection rates which lie between those of the AIC and LSB method. Some exceptions to this arise when $\sigma_2 = 1$ and μ_2 is intermediate in size. This is not surprising, for in this region dominance passes from the AIC to the LSB criterion. Other exceptions arise when $\sigma_2 = 2$, $n = 10$, and $\mu_1 = 1,2$ or 4; cases where the PSRQL procedure performs quite poorly. We also observe that when M_2 is correct then the PSRQL procedure dominates the PSRQB method; this is generally reversed when M_2 is false (exceptions are $\sigma_2 = 1.5$, $n = 20$, $\mu_2 = .25,.50$). It is also evident from Table 3 that, while the PSRQB, PSRQL, and AIC are asymptotically equivalent, for small samples there can be large differences between their correct selection rates.

A somewhat different situation arises when we compare the four methods with respect to prediction error. As in Section 3, our interest here is in the expected squared error of prediction (SEP) where the predictors are based on model selection. If, for a given data set, a model selection procedure picks M_1, then we set the predictors $\hat{y}_1=\hat{y}_2=\bar{x}$, the grand mean. On the other hand, if the selection procedure picks M_2 or M_3, then we set $\hat{y}_1=\bar{x}_1$ and $\hat{y}_2=\bar{x}_2$, the individual sample means. Also, as in Section 3, we estimate the SEP by estimating the quantity $E(\hat{y}_1-\mu_1)^2+\sigma_1^2 + E(\hat{y}_2-\mu_2)^2+\sigma_2^2$. Table 4 contains the estimates of the SEP for each parameter combination and model selection criterion. The table also contains the standard error of each estimated SEP. Note that if, for a given parameter configuration, the selection criteria always agree to

reject M_1, then $(\hat{y}_1,\hat{y}_2) = (\bar{x}_1,\bar{x}_2)$ always, and so the estimate of the SEP will be the same. This occurs, for example, when $\mu_2=4$, $\sigma_2=2$, and $n=20$ (see Tables 3 and 4).

In any case, the four procedures behave quite similarly in terms of prediction error. In particular, even though the standard errors of their estimates of SEP are small, it is difficult to determine a unique best procedure for a given parameter specification. With this difficulty in mind, we note that for the most part the AIC procedure either yields the smallest point estimate of the SEP, or that estimate is not significantly different from the smallest point estimate of the SEP for a given assignment of parameter values. In the case where this fails ($\mu_2=0$, $\sigma_2=1$; $\mu_2=.25$, $\sigma_2=1$, $n=10$; and $\mu_2=.25$, $\sigma_2=1.5$) the AIC is in fact the worst performing procedure, and one might prefer one of the PSRQB, PSRQL, or LSB procedures. Indeed, in the cases noted, these three perform similarly, and except when $\mu_2=0$, $\sigma_2=1$, and $n=20$, they do not differ significantly from each other. (When $\mu_2=0$, $\sigma_2=1$, $n=20$, the PSRQL method performs significantly worse than the PSRQB and LSB methods.) On the other hand, when the AIC is the best performing procedure, then the other three methods tend to differ for small values of μ_2 (where the PSRQL method performs better than the LSB or PSRQB methods) and for moderate values of μ_2 (where the LSB procedure dominates the PSRQL and PSRQB methods). When μ_2 is large, all four methods yield essentially the same results.

When comparing the criteria in terms of SEP it is important to note that even in those cases where the selection procedures do yield significantly different SEP estimates, the relative difference is small. For example, if $\mu_2=1$, $\sigma_2=1$, $n=10$, then the smallest point estimate of the SEP is 2.242 provided by the AIC and the largest is 2.303 provided by the PSRQB method; a relative difference of 2.7%. Examination of Table 4 shows that this is an extreme case; the largest relative difference for a given parameter

configuration in all other cases is less than 2.0%. If we compare the square roots of estimates of the SEP thus transforming them to the same units as the original observations we find that the largest relative difference for a given parameter configuration is less than 1.4%.

Since the predictive process depends only on whether M_1 is or is not chosen, it is appropriate to compare the corresponding relative difference in selection rates of M_1 for the selection criteria. For example, when $\mu_2 = 1$, $\sigma_2 = 1$, $n = 10$, the PSRQB method picks M_1 most frequently (30.5% of the time) and the AIC picks M_1 least frequently (16.9% of the time). As in the exponential sampling situation we see here that while the selection criteria might differ considerably in their rates of selection of M_1, from a practical point of view they yield essentially the same estimates of the SEP.

The similar behavior of the selection criteria in terms of prediction seems to be a consequence of the following phenomenon. Suppose one were faced with predicting Y_1 and Y_2, and suppose one thought M_1 were correct. If σ_2 is known, then the expected prediction error is $E(Y_1-\bar{x})^2 + E(Y_2-\bar{x})^2 = (\sigma_2^2+1)(1+1/(2n))+\mu_2^2/2 = PE(M_1)$, say. If one thought M_1 were incorrect, then the pair $(\hat{y}_1,\hat{y}_2) = (\bar{x}_1,\bar{x}_2)$ would be used, and the expected prediction error would be $E(Y_1-\bar{x}_1)^2 + E(Y_2-\bar{x}_2)^2 = (\sigma_2^2+1)(1+1/n) = PE(\bar{M}_1)$, say. It is apparent that $PE(M_1)$ can exceed $PE(\bar{M}_1)$ by a large amount, and so if μ_2 is large, the prediction error can be large if the predicting pair $(\hat{y}_1,\hat{y}_2) = (\bar{x},\bar{x})$ is inappropriately used. However, we see from Table 3 that in the cases where μ_2 is large, and hence where a potentially large prediction error could be made, the selection procedures rarely, if ever, choose M_1. That is, in the case where a potentially large error could arise, the selection procedures are able to discriminate well enough between M_1 and the other models to avoid such an error. On the other hand, when μ_2 is close to but not equal to zero, the selection procedures generally err in picking M_1 frequently. This does not appreciably increase the prediction error, for if μ_2 is small then $PE(M_1)$ and

$PE(\bar{M}_1)$ are close in value. In fact, such incorrect choices of M_1 can be helpful, since if μ_2 is close enough to zero, ($\mu_2^2 < (\sigma_2^2+1)/n$), then $PE(\bar{M}_1)$ is larger than $PE(M_1)$.

That is, there is a smaller predictive error incurred if $(\hat{y}_1,\hat{y}_2) = (\bar{x},\bar{x})$ is used, even though M_1 is known not to be the correct model. Such a situation arises in the cases studied when $\mu_1 = .25$, with $(\sigma_2,n) = (1,10)$, $(1.5,10)$ or $(1.5,20)$, and so in these cases one prefers a procedure which picks M_1 frequently. (A similar situation was noted in Section 3.) This explains why the LSB, PSRQL, and PSRQB procedures dominate the AIC for those parameter assignments. Outside of these cases, however, when $\mu_2 \neq 0$ then the AIC dominates the others in terms of picking M_1 less frequently, and accordingly weakly dominates them in terms of estimated SEP.

5. Remarks

One of the principal reasons for building models of statistical data is to enable one to predict new observations thought to arise from the same source. Frequently such predictions are made by first choosing a model according to some criterion, and then basing the predictions upon the chosen model. We have seen, for the models and model selection criteria examined in this study, that a particular measure of predictive capacity, namely squared error, appears to be fairly stable when predictions are made in such a manner. Under the same circumstances, the correct selection rates for those model selection procedures can vary appreciably. As a consequence, it would seem that, for predictive purposes, efforts to find "optimal" model selection procedures based on error rates may be misguided. Instead, it is more pertinent and sensible to devote one's time to constructing models for data and assessing the predictive adequacy of such models.

This precept makes even more sense when it is realized that the models typically considered are recognized to be artifacts, and the "true" model is not likely to be included among the model alternatives considered. In

that case it seems pointless to evaluate model selection criteria on the basis of fictional error rates. Nonetheless, it is still meaningful to compare the predictive capability of various models, even if none is true.

It would therefore be of interest to see if the model selection criteria examined herein exhibit the same stability with regard to prediction error when the data are generated by a process which is not among the model alternatives considered. This would indicate how "robust" the prediction procedure used in this study is to variation in both the source of the data and the model selection procedure used. If one is going to base predictions on model selection, then it would be desirable to have a model selection procedure whose predictive ability did not depend critically on the source of the data.

One such possibility is a sample reuse procedure for low structure situations which was proposed by Geisser and Eddy (1979) and was devised to have good predictive squared error properties for the prediction rule used here.

Suppose the data is $\underset{\sim}{x} = (\underset{\sim}{x}_1, \underset{\sim}{x}_2)$ with $\underset{\sim}{x}_i = (\underset{\sim}{x}_{i1}, \ldots, x_{iN_i})$ and M_1 asserts that the label $i=1,2$ is irrelevant for prediction and M_2 asserts that it is. The low structure rule is to calculate

$$D_1 = N^{-1} \sum_i \sum_j (x_{ij} - \bar{x}_{(ij)})^2$$
$$D_2 = N^{-1} \sum_i \sum_j (x_{ij} - \bar{x}_{i(j)})^2 \qquad (5.1)$$

and predict (Y_1, Y_2) as $(\bar{x}, \bar{x})$ if $D_1 < D_2$ and $(\bar{x}_1, \bar{x}_2)$ if $D_2 < D_1$.

It is of interest to note for the normal models Π_1 and Π_2 discussed in section 2, that this low structure procedure yields probability of correct selection under M_1 as

$$\Pr[F_{1,2(n-1)} \leq 2 + \frac{1}{2(n-1)}]$$

where $n=N_1=N_2$. This correct selection probability is always larger than that for the AIC procedure although it tends to the same limit. In addition to possibly being more robust, this procedure directly faces the predictive aspect of model selection. For the most part, this predictive aspect has been neglected in the development and evaluation of model selection procedures.

In summary, we would feel secure in conjecturing that if the "true or approximately true model" is included amongst the alternatives considered, all reasonable model selection procedures will possess rather similar predictive capabilities. Although it is considerably more useful to devise more appropriate models for data than to devise trivially more efficient selection procedures, we also believe that the area of robust model selection approaches is still in need of further development and investigation.

References

Akaike, Hirotugu (1973), Information Theory and an Extension of the Maximum Likelihood Principle, in *Proceedings of the 2nd International Symposium on Information Theory*, eds. B.N. Petrov and F. Czaki, Budapest: Akademiai Kiado, 267-281.

Geisser, S. (1980). Discussion on Hypothesis Testing, *Bayesian Statistics*, ed. Bernardo, J.M. et al., University Press, Valencia, 636.

Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.*, 74, 153-160.

Jeffreys, H. (1967). *Theory of Probability*, Oxford. Clarendon Press.

Mallows, C.L. (1973). Some comments on C_p, *Technometrics*, 15, 661-675.

Schwarz, Gideon (1978). Estimating the Dimension of a Model, *Annals of Statistics*, 6, 461-464.

Seber, C. (1979). Power comparisons of the quasi-Bayes and quasi-likelihood criteria for selection of exponential models. Preliminary Report, unpublished.

Zellner, A. and Siow, A. (1980). Posterior odds ratio for selected regression hypotheses. *Bayesian Statistics*, ed. Bernardo, J.M. et al., University Press, Valencia, 585-618.

Table 1. Estimated Probability of Correct Selection Between Populations Specified by $f_1 = e^{-x}$ and $f_2 = \lambda_2 e^{-\lambda_2 x}$ for Four Selection Criteria.[4]

		Selection Criterion			
λ_2	Size of each Sample	PSRQL	PSRQB	AIC	LSB
1.0	4	.799(.794)*	.827(.821)*	.832	.841
	8	.816(.830)*	.827(.837)*	.830	.894
	12	.834	.827	.835	.920
	20	.837	.837	.840	.942
1.5	4	.262	.237	.232	.223
	8	.291	.284	.282	.201
	12	.343	.346	.345	.214
	20	.446	.451	.456	.260
2.0	4	.342	.334	.339	.327
	8	.479	.489	.489	.385
	12	.590	.600	.599	.455
	20	.763	.770	.771	.602
2.5	4	.440	.445	.444	.430
	8	.633	.654	.661	.560
	12	.768	.784	.787	.664
	20	.918	.926	.928	.833

[4]Based on 10,000 samples of each sample size and standard error of tabular entries estimated to be between .004 and .005 .

*Values in parentheses are estimates from Geisser and Eddy (1979) using same procedure.

PSRQL; Predictive sample reuse quasi-likelihood.

PSRQB; Predictive sample reuse quasi-Bayes.

AIC; Akaike information criterion.

LSB; Large sample Bayes.

Table 2. Estimates of Expected Squared Error of Prediction (SEP) and Their Standard Error Based on Use of Four Selection Criteria.[5]

λ_2	Size of each Sample	Selection Criterion PSRQL	PSRQB	AIC	LSB	SEPC[6]	SEPI[7]	SEPK[8]
1	4	2.376(.005)	2.383(.005)	2.380(.006)	2.375(.005)	2.250	2.500	2.00
	8	2.195(.004)	2.196(.004)	2.196(.004)	2.177(.004)	2.125	2.250	
	12	2.131(.002)	2.133(.002)	2.133(.002)	2.115(.002)	2.083	2.167	
	20	2.078(.001)	2.079(.001)	2.079(.001)	2.065(.001)	2.050	2.100	
1.5	4	1.763(.005)	1.768(.005)	1.770(.005)	1.768(.005)	1.806	1 681	1.44
	8	1.620(.002)	1.622(.002)	1.623(.002)	1.620(.002)	1.625	1.590	
	12	1.573(.001)	1.573(.001)	1.574(.001)	1.575(.002)	1.565	1.560	
	20	1.529(.001)	1.529(.001)	1.529(.001)	1.537(.001)	1.517	1.536	
2.0	4	1.576(.004)	1.583(.005)	1.585(.005)	1.585(.005)	1.563	1.531	1.25
	8	1.438(.002)	1.438(.002)	1.438(.002)	1.449(.002)	1.406	1.453	
	12	1.384(.002)	1.383(.002)	1.382(.002)	1.399(.002)	1.354	1.427	
	20	1.329(.001)	1.329(.001)	1.328(.001)	1.348(.001)	1.313	1.406	
2.5	4	1.492(.005)	1.496(.005)	1.498(.005)	1.500(.005)	1.450	1.485	1.16
	8	1.346(.002)	1.342(.002)	1.341(.002)	1.358(.002)	1.305	1.413	
	12	1.283(.002)	1.281(.002)	1.280(.002)	1.300(.002)	1.257	1.388	
	20	1.226(.001)	1.225(.001)	1.225(.001)	1.239(.001)	1.218	1.369	

[5]Based on samples used in Table 1.

[6]Expected squared error of prediction (SEPC) when all selections are correct.

[7]Expected squared error of prediction (SEPI) when all selections are incorrect.

[8]Expected squared error of prediction when λ_i is known (SEPK).

Table 3. Estimated Probability of Correct Selection (First Value) and that M_1 will be Selected (Second Value) for Populations Specified by $f_1 = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}x^2]$ and $f_2 = \frac{1}{\sigma_2\sqrt{2\pi}} \exp[-\frac{1}{2\sigma_2^2}(x-\mu_2)^2]$ assuming that if $\mu_2 = 0$, then $\sigma_2^2 = 1$, for Four Selection Criteria.[9]

(μ_2,σ_2)	Size of each Sample	Selection Criterion PSRQL	PSRQB	AIC	LSB
(0,1)	10	.824;.824	.872;.872	.737;.737	.871;.871
	20	.812;.812	.873;.873	.767;.767	.917;.917
(.25,1)	10	.154;.773	.102;.833	.197;.684	.133;.818
	20	.215;.701	.152;.775	.245;.649	.136;.841
(.50,1)	10	.265;.650	.211;.703	.317;.547	.252;.692
	20	.436;.448	.355;.538	.463;.402	.359;.602
(.75,1)	10	.417;.464	.364;.510	.480;.343	.429;.486
	20	.684;.193	.609;.262	.697;.146	.638;.315
(1.00,1)	10	.598;.284	.550;.305	.654;.169	.636;.272
	20	.795;.050	.754;.077	.782;.034	.833;.103
(2.00,1)	10	.843;.000	.820;.000	.803;.000	.885;.000
	20	.851;.000	.838;.000	.831;.000	.943;.000
(4.00,1)	10	.854;.000	.818;.000	.802;.000	.877;.000
	20	.854;.000	.840;.000	.834;.000	.942;.000
(.25,1.5)	10	.217;.685	.220;.708	.321;.550	.183;.724
	20	.457;.440	.431;.510	.521;.378	.258;.677
(.50,1.5)	10	.231;.603	.238;.630	.344;.461	.190;.641
	20	.494;.348	.483;.398	.568;.268	.307;.547
(.75,1.5)	10	.253;.484	.282;.503	.363;.341	.228;.499
	20	.530;.204	.539;.249	.602;.149	.366;.361
(1.00,1.5)	10	.282;.371	.320;.382	.399;.234	.256;.365
	20	.549;.091	.580;.129	.629;.057	.393;.202
(2.00,1.5)	10	.354;.044	.412;.038	.452;.009	.312;.027
	20	.571;.000	.614;.000	.633;.000	.424;.000
(4.00,1.5)	10	.352;.001	.413;.000	.452;.000	.319;.000
	20	.578;.000	.624;.000	.639;.000	.439;.000
(1.00,2)	10	.535;.294	.602;.282	.699;.165	.552;.289
	20	.891;.047	.902;.053	.931;.026	.811;.108
(2.00,2)	10	.598;.074	.686;.059	.736;.022	.618;.060
	20	.907;.002	.937;.002	.942;.001	.870;.003
(4.00,2)	10	.635;.006	.728;.000	.757;.000	.660;.000
	20	.923;.000	.944;.000	.949;.000	.874;.000

[9] Based on 2,000 samples of each sample size and standard error of tabular entries estimated to be between .008 and .012.

Table 4. Estimates of Expected Squared Error of Prediction (SEP) and Their Standard Error Based on Use of Four Selection Criteria.[10]

(μ_2,σ_2)	Size of each Sample	Selection Criterion PSRQL	PSRQB	AIC	LSB	SEPC
(0,1)	10	2.140(.004)	2.132(.004)	2.156(.004)	2.135(.004)	2.100
	20	2.076(.002)	2.070(.002)	2.080(.002)	2.066(.002)	2.050
(.25,1)	10	2.176(.004)	2.172(.004)	2.185(.004)	2.176(.004)	2.200
	20	2.101(.002)	2.100(.002)	2.102(.002)	2.098(.002)	2.100
(.50,1)	10	2.231(.004)	2.236(.004)	2.223(.004)	2.236(.004)	2.200
	20	2.130(.002)	2.141(.002)	2.125(.002)	2.148(.002)	2.100
(.75,1)	10	2.290(.005)	2.303(.005)	2.260(.005)	2.296(.005)	2.200
	20	2.132(.003)	2.150(.003)	2.121(.003)	2.163(.003)	2.100
(1,1)	10	2.295(.006)	2.303(.006)	2.242(.005)	2.287(.006)	2.200
	20	2.116(.003)	2.126(.003)	2.110(.003)	2.136(.004)	2.100
(2,1)	10	2.210(.006)	2.200(.005)	2.198(.005)	2.201(.005)	2.200
	20	2.101(.002)	2.101(.002)	2.101(.002)	2.101(.002)	2.100
(4,1)	10	2.202(.005)	2.202(.005)	2.202(.005)	2.202(.005)	2.200
	20	2.100(.002)	2.100(.002)	2.100(.002)	2.100(.002)	2.100
(.25,1.5)	10	3.528(.007)	3.526(.007)	3.550(.008)	3.528(.007)	3.575
	20	3.408(.004)	3.404(.004)	3.410(.004)	3.398(.004)	3.413
(.50,1.5)	10	3.581(.007)	3.584(.007)	3.582(.007)	3.584(.007)	3.575
	20	3.432(.004)	3.436(.004)	3.426(.004)	3.447(.004)	3.413
(.75,1.5)	10	3.650(.007)	3.657(.007)	3.622(.008)	3.655(.007)	3.575
	20	3.445(.004)	3.455(.004)	3.433(.004)	3.480(.004)	3.413
(1,1.5)	10	3.693(.008)	3.700(.008)	3.636(.008)	3.690(.008)	3.575
	20	3.437(.005)	3.452(.005)	3.424(.004)	3.480(.005)	3.413
(2,1.5)	10*	3.645(.005)	3.617(.005)	3.581(.004)	3.604(.004)	3.575
	20	3.412(.004)	3.412(.004)	3.412(.004)	3.412(.004)	3.413
(4,1.5)	10*	3.583(.004)	3.573(.003)	3.573(.003)	3.573(.003)	3.575
	20	3.415(.004)	3.415(.004)	3.415(.004)	3.415(.004)	3.413
(1,2)	10*	5.575(.005)	5.571(.005)	5.530(.006)	5.571(.005)	5.500
	20	5.272(.007)	5.274(.007)	5.265(.007)	5.291(.007)	5.250
(2,2)	10*	5.615(.007)	5.570(.007)	5.516(.006)	5.559(.007)	5.500
	20	5.249(.007)	5.249(.007)	5.248(.007)	5.251(.007)	5.250
(4,2)	10*	5.548(.008)	5.506(.006)	5.506(.006)	5.506(.006)	5.500
	20	5.260(.007)	5.260(.007)	5.260(.007)	5.260(.007)	5.250

[10]Based on samples in Table 3, except rows marked with an asterisk, which are based on 10,000 replications.