Model Selection
IMS Lecture Notes - Monograph Series (2001) Volume 38

On Model Selection

C. R. Rao and Y. Wu

Pennsylvania State University and York University

Abstract

The task of statistical model selection is to choose a family of distributions among a possible set of families, which is the best approximation of reality manifested in the observed data. In this paper, we survey the model selection criteria discussed in statistical literature. We are mainly concerned with those used in regression analysis and time series.

C.R. Rao is Eberly Professor, Department of Statistics,Pennsylvania State University, University Park, PA 16802, U.S.A; email: crr1@psu.edu. Y.Wu is Professor, Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3; email: wuyh@mathstat.yorku.ca. Rao was partially supported by Army Research Office: DAAD 19-01-1-0563 and Wu was partially supported by the Natural Sciences and Engineering Research Council of Canada.

Contents

1 Introduction

Let $\{M_\gamma,\ \gamma \in \Gamma\}$ be candidate models for the observations, where Γ is an index set. It is possible that the true model is not included in $\{M_\gamma\}$. Based on the data, we need to select a model from $\{M_\gamma,\ \gamma \in \Gamma\}$ through a suitable model selection criterion. Many model selection procedures have been proposed in the literature. Each one is designed for a particular application.

Model selection problems are encountered almost everywhere. In linear regression analysis, it is of interest to select the right number of nonzero regression parameters. With the smallest true model, statistical inferences can be carried out more efficiently. In the analysis of time series, it is essential to know the true orders of an ARMA model. In problems of clustering, it is important to find out the number of clusters. In the signal detection, it is necessary to determine the number of true signals, and so on.

In this paper, we survey the model selection criteria discussed in statistical literature. Almost all statistical problems can be considered as model selection problems, but in this paper, we will be mainly concerned with those used in regression analysis and time series. Some interesting examples can be found in Burnham and Anderson (1998), among others.

The paper is arranged as follows: In Section 2, model selection based on hypothesis testing is examined. In Section 3, selection of a model based on the prediction errors is surveyed. In Section 4, the information theoretic criteria are discussed. In Section 5, the role of cross-validation and bootstrap methods in model selection is covered. In Section 6, Baysian approaches to model selection are described. In Section 7, studies on robust model selection are examined. In Section 8, the results on order selection in time series are presented. In Section 9, model selection criteria in categorical data analysis are explored. In Section 10, the investigation on model selection in nonparametric regression is reviewed. In Section 11, data-oriented penalties are discussed. In Section 12, the effect of prior model selection based on data on the inferential statistical analysis of the same data is examined.

In the sequel, for an index set κ, $|\kappa|$ denotes the size of κ, $\boldsymbol{c}(\kappa)$ denotes the sub-vector containing the components of the vector $\boldsymbol{c}$ that are indexed by the integers in κ and $A(\kappa)$ denotes the sub-matrix containing the columns of a matrix A that are indexed by the integers in κ, P_A is used to denote the projection operator on the linear space generated by the column vectors of A, and $\boldsymbol{Y} \sim (\boldsymbol{\mu}, \Sigma)$ is used to represent the random vector $\boldsymbol{Y}$ distributed according to a multivariate distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. For convenience, $\ell : m$ denotes the set of $\{\ell, \ell + 1, \ldots, m\}$ in this paper, where $\ell \leq m$ are positive integers.

2 Selection of a model based on hypothesis testing

A model selection procedure can be constructed based on a sequence of hypothesis tests. Assume that there is an order in the set of candidate models $\{M_i, \ i = 1, 2, \ldots\}$ such that M_i is a preferable to M_{i+1}. A sequence of hypotheses, H_{io} : M_i *holds true* versus H_{ia} : M_{i+1} *holds true*, $i = 1, 2, \ldots$, can be tested sequentially. Once H_{io} is accepted, the test procedure stops and the model M_i is selected. Another such procedure is to replace H_{ia} by H'_{ia} : *One of* $\{M_j \ : \ j > i\}$ *holds true.* It can be seen that if all the tests in two test procedures have the same thresholds, the acceptance of M_i in the second test procedure will imply that of M_i in the first test procedure.

Now assume that there is a partial order in a finite index set. Using the partial order "$\prec$", Γ can be partitioned into equivalent classes Γ_i, $i = 0, 1, \ldots.$ We further assume that some member of a subset Γ_0 has the smallest order. The problem is how to choose a model from the candidate models consisting of M_γ, $\gamma \in \Gamma$, by hypothesis testing. Suppose that the model M_{γ_1} is preferable to the model M_{γ_2} if $\gamma_1 \prec \gamma_2$. In this case, a model selection procedure can be constructed as follows: First, let $i = 0$. For each $\gamma \in \Gamma_0$, test the null hypothesis that M_γ holds true against the alternative hypothesis that one of the models $M_{\tilde{\gamma}}$, $\gamma \prec \tilde{\gamma}$ holds true. Find the largest $p-$value and if this value is less than the prechosen one, stop and select the model with that $p-$value. Otherwise, let $i = 1$, and repeat the above step with Γ_0 replaced by Γ_1. In general, if the procedure does not stop at the jth step, let $i = j + 1$ in the next step and repeat the previous step with Γ_j replaced by Γ_{j+1}.

For a better understanding of the procedures, let us consider the following linear model:

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \ldots, \tag{2.1}$$

where $\boldsymbol{x}_i$ are p-vectors, $\boldsymbol{\beta}$ is a p-vector parameter, and ε_i are random errors. Let κ denote a subset of $\{1, \ldots, p\}$. Based on the observations $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, we would like to decide whether the model (2.1) should be replaced by the following sub-model with a fixed κ:

$$Y_i = \boldsymbol{x}_i'(\kappa)\boldsymbol{\beta}(\kappa) + \varepsilon_i, \quad i = 1, 2 \ldots. \tag{2.2}$$

There are $2^p - 1$ such models.

Denote the model (2.2) by M_κ. If $\kappa_i = \{1, \ldots, i\}$, for convenience, write M_i for the model M_{κ_i}. For simplicity, consider the set of candidate models $\{M_i, \ i = 1, \ldots, p\}$. Hence, H_{io} is the hypothesis that $\beta_j = 0$ for $j > i$, H_{ia} is the hypothesis that $\beta_j = 0$ for $j > i + 1$ and $\beta_{i+1} \neq 0$, and H'_{ia} is the hypothesis that at least one of $\{\beta_j, \ j > i\}$ is not zero. Assume that the maximum likelihood estimators of unknown parameters can be worked out. Hence, the likelihood ratio statistics may be used to perform the tests. It is easy to see that the performance of the procedure is affected a great deal by the critical values of the tests. Since the exact distributions of the likelihood ratio statistics are usually unknown, their asymptotic distributions are used to compute these thresholds, whose acceptable accuracy would need a large sample. The advantage for such procedures is that the probability of overfitting is somewhat under control.

For giving a more clear view of such procedures, we further assume that ε_i, $i = 1, 2, \ldots$, are independently and identically $N(0, \sigma^2)$ distributed and $\boldsymbol{x}_i$, $i = 1, 2, \ldots$, are fixed, which can be relaxed to the condition that $\boldsymbol{\varepsilon}_n | X_n \sim N(\mathbf{0}, \sigma^2 I_n)$, where $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)'$ and $X_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$. For testing H_{io} against H_{ia}, the likelihood ratio statistic is equivalent to

$$F^{(i)} = \frac{\boldsymbol{Y}_n'[P_{X_n(1:i+1)} - P_{X_n(1:i)}]\boldsymbol{Y}_n}{\boldsymbol{Y}_n'[I_n - P_{X_n(1:i+1)}]\boldsymbol{Y}_n},$$

where $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)'$. Under H_{io}, $F^{(i)} \sim F_{1,n-(i+1)}$. The critical value is given by $F_{1,n-(i+1)}(\alpha)$, where Prob $\left(F_{1,n-(i+1)} > F_{1,n-(i+1)}(\alpha)\right) = \alpha$, and hence under M_i the probability of making a type I error, i.e. choosing M_{i+1}, is α. It can be seen that the probability of underfitting, i.e. choosing M_i when it is not a true model, depends on $\boldsymbol{\beta}(1 : i+1)$. As commented in Shao and Rao (2000), the underfitting probability converges to 0 as the sample size n increases to ∞.

In practice, *forward selection*, *backward elimination* and *stepwise regression* are popular model selection methods in linear regression. They are available in almost all of statistical software packages. In their applications, the candidate models consist of all $2^p - 1$ submodels. Controlled by one or two thresholds, the model is selected based on statistical hypothesis testing. The details can be found in Krishnaiah (1982), Miller (1990) and some textbooks on linear regression. Some authors prefer backward elimination to forward selection for the economy of effort (see Mantel 1970). But on account of the simplicity of computation and stopping rules, forward selection is recommended. For choosing the significance levels required, the most widely used level is 10% or 5%. But the overall power as well as the type I error rate are unknown unless the order of entry of the variables into the model is specified explicitly before applying any method. It is easily seen that the order of entry differs with observations. To avoid such difficulties, Aitkin (1974) and McKay (1977) proposed an application of a simultaneous testing procedure, but it requires considerable computation to obtain a set of significance levels as shown by Shibata (1986a). Note that model selection procedures based on R^2 (the square of the multiple correlation), adjusted R^2 or its equivalent MSE are also available in many statistical software packages.

Thall, Russell, and Simon (1997) proposed an algorithm, *backward elimination via repeated data splitting* (BERDS), for variable selection in regression. Initially, the data are randomly partitioned into two sets E,V, and an exhaustive backward elimination (BE) is performed in E. For each $\alpha_{\text{stay}} \in (0,\ 1)$ used in BE, the corresponding fitted model from E is validated in V by computing the sum of squared deviations of observed from predicted values. This is repeated m times, and the α^*, which minimizes the sum of the m sums of squares, is used as α_{stay} in a final BE on the entire data set. BERDS is a modification of the algorithm BECV (BE via cross-validation) proposed by Thall, Simon, and Grier (1992). Their extensive simulation study showed that, compared to BECV, BERDS has a smaller model error and higher probabilities of excluding noise variables, of selecting each of several uncorrelated true predictors, and of selecting exactly one of two or three highly correlated true predictors. Thall, Russell, and Simon (1997) also showed that BERDS is superior to standard BE with $\alpha_{\text{stay}} = .05$ or .10, and this superiority increases with the number of noise variables in the data and the degree of correlation among true predictors.

While the log likelihood ratio tests are fairly good when nested parametric hypotheses are involved, it is a different story for testing non-nested parametric hypotheses. Cox (1961, 1962) initiated research on testing separate families of hypotheses and proposed non-nested test statistics, which, according to Bera (2000), can also be viewed as Rao's score tests. Later, Williams (1970a,b) gave a different approach by directly simulating the distribution of the log likelihood ratio on a computer. According to Loh (1985), both tests are not satisfactory. Loh (1985) proposed an alternative procedure by repeated application of the parametric bootstrap method over slowly shrinking confidence regions, which, as justified by the author, is promising. There are other solutions to this kind of problems in the literature, which are not limited to parametric hypotheses.

Since in practical situations, the assumed null hypotheses are only approximations and they are almost always different from the reality, the choice of the loss function in the test theory makes its practical application logically contradictory, as commented by Akaike (1974). Bera (2000) gives a very inspiring discussion on hypothesis testing with misspecified models, which includes a historical review as well as possible future direction of research.

3 Model selection based on prediction errors

Model selection based on statistical hypothesis testing described in the last section involves many restrictions and further the choice of thresholds are open. As commented in Akaike (1969), the main difficulty in applying this kind of procedures and their relatives stems from the fact that they are essentially formulated in the form of successive tests of null hypotheses against multiple alternative hypotheses. Actually one of the alternative hypotheses is just the model one is looking for and thus it is very difficult for one to get the feeling of the possible alternative hypotheses to set reasonable significance levels. To overcome this difficulty, Akaike (1969) suggested an alternative decision theoretic approach based on prediction errors. In Akaike (1969), the final prediction error (FPE) was defined, which is the mean squared prediction error when a model fitted to the present data is applied to another independent observation, or to make a one step prediction. Based on the final prediction error, the parameters in each candidate model are estimated so that the minimum final prediction error is attained for this model, and then a model, which has the minimum final prediction error within the candidate models, is selected. We call this an FPE procedure. Note that when the purpose of model selection is for prediction, it may be wiser to choose a model based on the prediction errors than using the model selection methods discussed in Section 2.

For example, consider the model (2.2). We assume that ε_i, $i = 1, 2, \ldots$, are independently and identically distributed with mean zero and variance σ^2 and $\boldsymbol{x}_i$, $i = 1, 2, \ldots$, are fixed, which can be relaxed to the condition that $\varepsilon_n | X_n \sim (\mathbf{0}, \sigma^2 I_n)$. We also assume that X_n is of full rank. The least squares estimate of $\beta(\kappa)$ for the model (2.2) is $[X_n(\kappa)'X_n(\kappa)]^{-1}X_n(\kappa)'\boldsymbol{Y}_n$. Let $\boldsymbol{Y}_0$ consist of n new observations and $\boldsymbol{Y}_0 = X_n(\kappa)\beta(\kappa) + \varepsilon_0$ with $\varepsilon_0 \sim (\mathbf{0}, \sigma^2 I_n)$. Then the FPE, i.e. the mean squared prediction error, is given by

$$\begin{aligned} & \frac{1}{n}\mathrm{E}[(\boldsymbol{Y}_0 - \hat{\boldsymbol{Y}}_0)'(\boldsymbol{Y}_0 - \hat{\boldsymbol{Y}}_0)] \\ = \; & \frac{1}{n}\mathrm{E}[(\boldsymbol{Y}_0 - X_n(\kappa)'\beta(\kappa))]'[(\boldsymbol{Y}_0 - X_n(\kappa)'\beta(\kappa))] = \sigma^2(1 + k/n), \end{aligned} \tag{3.1}$$

where $k = |\kappa|$. Let S_κ denote the residual sum of squares under the model M_κ, and $\hat{\sigma}_\kappa^2$ denote $S_\kappa/(n-k)$. Using the unbiased estimator $\hat{\sigma}_\kappa^2$ to replace σ^2 in (3.1), we get $\hat{\sigma}_\kappa^2(1 + k/n)$, which is denoted by FPE(κ). The selected model M_{κ^*} can be obtained by minimizing FPE(κ), i.e.

$$M_{\kappa^*} = \arg\min_{M_\kappa} \mathrm{FPE}(\kappa).$$

The FPE procedure was originally derived for autoregressive time series models. A similar procedure was developed by Davisson (1965) for analyzing signal-plus-noise data. By using FPE, Akaike (1970) suggested a way to decide on the thresholds for the procedures discussed in Section 2.

Define, for the model (2.2) with $p = \infty$,

$$\mathrm{SH}(\kappa) = \hat{\sigma}_\kappa^2(n + 2k)(n - k)/n.$$

Shibata (1980) proposed a model selection criterion based on the expectation of prediction errors, which is given as follows:

$$M_{\kappa^*} = \arg\min_{M_\kappa} \mathrm{SH}(\kappa).$$

This criterion is equivalent to FPE and was shown to be asymptotically efficient.

In Akaike (1970), the modified version of the FPE procedure was proposed for improving the consistency of the FPE procedure. For the model (2.2), let

$$(\mathrm{FPE})^{\lambda}(\kappa) = S_{\kappa}(1 + k/n^{\lambda}),$$

where $0 < \lambda < 1$. A model is chosen to be the one which minimizes $(\mathrm{FPE})^{\lambda}(\kappa)$ within candidate models. We call such a procedure as the $(\mathrm{FPE})^{\lambda}$ procedure.

Even though the consistency of the $(\mathrm{FPE})^{\lambda}$ procedure is improved over the FPE procedure, it is still not consistent and the probability of overfitting is still greater than zero. Many authors have sought to modify the overfitting property by further adjusting the second term of the FPE procedure by multiplying it by δ and proposing the following FPE_{δ} procedure:

$$\min_{M_{\kappa}} \mathrm{FPE}_{\delta}(M_{\kappa}) = \hat{\sigma}_{\kappa}^{2}(1 + \delta k/n),$$

where δ may or may not depend on n.

The FPE_{δ} procedure or its equivalent was discussed by Akaike (1970, 1974), Atkinson (1980, 1981), Bhansali and Downham (1977), Shibata (1976, 1980, 1986a,b), Zhang and Krieger (1993) among others. Based on empirical evidence, some authors have suggested that the FPE_{δ} procedure with δ between 2 and 6 would do well in most situations, but such an ad hoc choice of δ seems to be lacking in theoretical justification. In view of the inconsistency of FPE and $(\mathrm{FPE})^{\lambda}$ procedures, it is unlikely that any finite δ would lead to a consistent procedure. To achieve consistency, it may be necessary to make δ dependent on sample size. Specially, it has been suggested that $\delta = \delta_n$ should satisfy $\delta_n \to \infty$ and $\delta_n/n \to 0$. See Bozdogan (1987), Nishii (1988), Rao and Wu (1987), Shao (1997), Zhao, Krishnaiah, and Bai (1986a,b) among others. By assuming normality, Venter and Steel (1992) studied the choice of the quantity δ in the FPE_{δ} procedure for selecting a member of a class of linear models having orthogonal structure. Two approaches are discussed, namely fixing the maximal estimation risk at a prescribed level and using minimax regret. In Zhang (1994) it was argued that a choice of $\delta \in [3, 4]$ would be adequate for most practical purposes, and by using decision theoretic properties of FPE_{δ}, it was shown that the incorrect models are sometimes preferable to the true model.

Mallows (1973) took a different approach to model selection criterion in a linear regression problem. Consider the model (2.2) and assume that the conditions made previously hold true. The fitted regression subset at the point $\boldsymbol{x}_i$ is given by

$$\hat{Y}_{i,\kappa} = \boldsymbol{x}_i(\kappa)\hat{\boldsymbol{\beta}}_{\kappa},$$

where $\hat{\boldsymbol{\beta}}_{\kappa}$ is the least squares estimate of $\boldsymbol{\beta}(\kappa)$ under the model (2.2). If $\mathrm{E}(\hat{Y}_{i,\kappa}) = \mu_{i,\kappa}$, then $\mu_{i,\kappa}$ generally differs from $\boldsymbol{x}_i(\kappa)'\hat{\boldsymbol{\beta}}_{\kappa}$ because of possible bias in the M_{κ}. Let $\mathrm{E}(Y_i) =$

θ_i. Then

$$\begin{aligned}\mathrm{E}[\hat{Y}_{i,\kappa}-\theta_i)^2] &= \mathrm{Var}(\hat{Y}_{i,\kappa})+(\mu_{i,\kappa}-\theta_i)^2\\ &= \sigma^2\boldsymbol{x}_i(\kappa)'(X_n(\kappa)'X_n(\kappa))^{-1}\boldsymbol{x}_i(\kappa)+(\mu_{i,\kappa}-\theta_i)^2,\end{aligned}$$

which implies that

$$\begin{aligned}\Delta_\kappa &= \frac{1}{\sigma^2}\mathrm{E}\left[\sum_{i=1}^n(\hat{Y}_{i,\kappa}-\theta_i)^2\right]\\ &= \frac{1}{\sigma^2}\left[\sum_{i=1}^n\sigma^2\boldsymbol{x}_i(\kappa)'(X_n(\kappa)'X_n(\kappa))^{-1}\boldsymbol{x}_i(\kappa)+\sum_{i=1}^n(\mu_{i,\kappa}-\theta_i)^2\right]\\ &= k+\frac{1}{\sigma^2}\sum_{i=1}^n(\mu_{i,\kappa}-\theta_i)^2.\end{aligned}$$

Since $\mathrm{E}(S_\kappa)=\sigma^2(n-k)+\sum_{i=1}^n(\mu_{i,\kappa}-\theta_i)^2$, Δ_κ can be estimated by

$$C_\kappa=\frac{S_\kappa}{\hat{\sigma}^2_{1:p}}+2k-n.$$

However, the quantity C_κ is not an unbiased estimator of Δ_κ. Mallows (1995) suggested that any candidate model where $C_\kappa < k$ should be carefully examined as a potential best model. This procedure is called C_p criterion.

Shibata (1980) showed that FPE is an asymptotically efficient procedure. Since FPE and C_p procedures are asymptotically equivalent (see, e.g., Nishii 1984), C_p is also asymptotically efficient. Note that both procedures are not consistent.

Considering a sequence of models with kth model given by (2.2) for $\kappa = 1:k$, Breiman and Freedman (1983) proposed the following criterion also based on the expectation of prediction errors:

$$M_{1:\hat{k}}=\arg\min_{M_{1:k},\,k\le n/2}\hat{\sigma}_{1:k}(1+k/(n-1-k)).$$

This criterion was shown to be asymptotically efficient.

In a series of papers, Rissanen introduced minimum description length (MDL) principle as a process of searching for models and model classes with the shortest code length (see Rissanen 1989). The application of the MDL principle with predictive code length is called the predictive MDL (PMDL). Consider the linear model (2.2). Based on his PMDL principle, Rissanen (1986a,b,c) proposed a new criterion that selects the model which minimizes

$$PLS(\kappa)=\sum_{i=m+1}^n(y_i-\boldsymbol{x}_i'(\kappa)\hat{\beta}_\kappa^{(i-1)})^2,$$

where $\hat{\beta}_\kappa^{(j)}$ is the least squares estimate based on $\{y_i,\boldsymbol{x}_i(\kappa);\ i\le j\}$ and m is the first integer j so that $\hat{\beta}_\kappa^{(j)}$ is uniquely defined. Since $(y_i-\boldsymbol{x}_i'(\kappa)\hat{\beta}^{(i-1)}(\kappa))^2$ is the square of the prediction error at stage i, this criterion is called the predictive least squares (PLS) principle, which is strongly consistent. A drawback with PLS is that the data must be ordered, and the result may depend on the particular order selected. Hence, the symmetric PLS was proposed. See Rissanen (1989) for more details.

If one first uses a criterion based on data to select a set of regressors and then estimates the regression coefficients, such a popular strategy is called an s/e procedure. Foster and George (1994) proposed a measure for the evaluation of variable selection procedures in multiple regression. This measure, which is called risk inflation, is the maximum possible increase in risk of the consequent s/e procedure due to selecting rather than knowing the "correct" predictors. The risk inflation is obtained as the ratio of risk of a s/e estimator to the risk of the ideal (but unavailable) selection/estimation estimator which uses only the "correct" predictors. Consider the models (2.2). In the case of orthogonal predictors, the authors argued that compared to overall inclusion, AIC, C_p and BIC offer smaller risk inflation and hence they proposed a model selection procedure (RIC) as follows:

$$M_{\kappa^*} = \arg\min_{M_\kappa}[S_\kappa + k\hat{\sigma}^2(2\log p)],$$

which, as they stated in their paper, substantially improves on AIC, C_p and BIC and is close to optimal. For the general case, it is unfortunate that the model selection procedure based on the risk inflation depends on the correlation structure of the predictors. See Foster and George (1994) for details.

In a linear regression model, for attenuating possible excessive modelling biases, a large number of predictors are usually introduced at the initial stage of modelling. To enhance predictability and to select significant variables, one usually applies stepwise deletion, subset selection and ridge regression. While these three methods are useful in practice, they ignore stochastic errors inherited in the previous stages of variable selections (see, e.g., Fan and Li 2001). Tibshirani (1996) proposed a new approach, called least absolute shrinkage and selection operator (LASSO), which simultaneously selects variables and estimates parameters. By using the LASSO, some regression coefficients are shrinked and others are set to be zero. According to Tibshirani (1996), LASSO retains good features of both subset selection and ridge regression, and can be applied to generalized linear models, besides, the LASSO estimate is also a Bayes estimate. As a matter of fact, the LASSO is closely related to penalized likelihood with the L_1 penalty. Fan and Li (2001) generalized the LASSO method by proposing the penalized likelihood with a smoothly clipped absolute deviation (SCAD) penalty function along with a unified algorithm backed up by statistical theory, which resulted in an estimator with good statistical properties. Their approach includes the LASSO as its special case. Their simulation results showed that their method compared favorably with other approaches as an automatic variable selection technique. As shown in their paper, by the advantage of simultaneous selection of the variables and estimation of parameters, they were able to give a simple estimated standard error formula, which was tested to be accurate enough for practical applications.

Recently Breiman (1996) studied how to stabilize an unstable model selection procedure in linear regression. Such problems are very important and need further investigation.

4 Information theoretic criteria

Let $z_1, \ldots, z_n$ be n independent observations on a random vector $\boldsymbol{Z}$ with probability density function $g(\boldsymbol{z})$. Consider parametric family of density functions is $\{f_{\boldsymbol{\theta}}(\boldsymbol{z}),\ \boldsymbol{\theta} \in \Theta\}$ with a vector parameter $\boldsymbol{\theta}$ and parameter space $\Theta \subset R^m$ for which the average log-likelihood is given by

$$\frac{1}{n}\sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(z_i), \tag{4.1}$$

where log denotes the natural logarithms. As n increases to infinity, this average tends to

$$S(g; f_{\boldsymbol{\theta}}) = \int g(\boldsymbol{z}) \log f_{\boldsymbol{\theta}}(\boldsymbol{z})\, d\boldsymbol{z},$$

with probability 1, where the existence of the integral is assumed. The difference

$$K(g; f_{\boldsymbol{\theta}}) = S(g; g) - S(g; f_{\boldsymbol{\theta}})$$

is known as the Kullback-Leibler distance (information) between $g(\boldsymbol{z})$ and $f_{\boldsymbol{\theta}}(\boldsymbol{z})$ and takes positive values, unless $f_{\boldsymbol{\theta}}(\boldsymbol{z}) = g(\boldsymbol{z})$ holds almost everywhere. Hence $S(g; f_{\boldsymbol{\theta}})$ is reasonable for defining a best fitting model by its maximization or, from the analogy to the concept of entropy, by minimizing $-S(g; f_{\boldsymbol{\theta}})$. Maximizing (4.1) with respect to $\boldsymbol{\theta}$ leads to the MLE $\hat{\boldsymbol{\theta}}$.

Consider the case that $g(\boldsymbol{z}) = f_{\boldsymbol{\theta}_0}(\boldsymbol{z})$, where $\boldsymbol{\theta}_0 \in \Theta$. When $\boldsymbol{\theta}$ is sufficiently close to $\boldsymbol{\theta}_0$,

$$K(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}}) \approx (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' J (\boldsymbol{\theta} - \boldsymbol{\theta}_0)/2,$$

where J is the Fisher information matrix. When the MLE $\hat{\boldsymbol{\theta}}$ lies very close to $\boldsymbol{\theta}_0$, $K(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}})$ can be approximately measured by $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' J (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/2$. Under certain regularity conditions $n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' J (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically distributed as chi-square with k degrees of freedom, and $\mathrm{E}[2nK(f_{\boldsymbol{\theta}_0}, f_{\hat{\boldsymbol{\theta}}})] \approx n(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' J (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + k$, where k is the number of independent parameters. By using

$$2\left(\sum_{i=1}^{n} \log f_{\boldsymbol{\theta}_0}(z_i) - \sum_{i=1}^{n} \log f_{\hat{\boldsymbol{\theta}}}(z_i)\right)$$

to approximate $n(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'J(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, a correction is needed for the downward bias due to replacing $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$. Akaike (1973) added k as the correction and introduced the famous AIC criterion: Let

$$\text{AIC}(\boldsymbol{\theta}) = -2\log(\text{maximum likelihood}) + 2k, \tag{4.2}$$

where k is as defined above. The selected model is

$$M_{\boldsymbol{\theta}^*} = \arg\min_{M_{\boldsymbol{\theta}}} \text{AIC}(\boldsymbol{\theta}).$$

The justification of the correction k can be found in Akaike (1973), Linhart and Zucchini (1986) and Sakamoto, Ishiguro, and Kitagawa (1986) among others. Note that AIC is, in final analysis, based on the concept of minimizing the expected Kullback-Leibler distance (see, e.g., Sawa 1978, Sugiura 1978). It is worth mentioning that information theory (see, e.g., Guiasu 1977) has been a discipline only since the mid-1940s and covers a variety of theories and methods that are fundamental to many of the sciences.

For the model (2.2), assuming that the errors are $N(0, \sigma^2)$ distributed, AIC can be expressed as

$$\text{AIC}(\kappa) = n\log(S_\kappa/n) + 2k,$$

where S_κ is defined as before and $k = |\kappa|$.

Assuming that the errors have a multivariate normal distribution, Fujikoshi and Satoh (1997) proposed modified AIC and C_p for selecting multivariate linear regression models by reducing the bias of estimation of Akaike- and Mallows-type risks when the collection of candidate models includes both underspecified and overspecified models. Their simulation study showed that both modified AIC and C_p provided better approximations to their risk functions, and better model selection, than AIC and C_p.

For model selection in settings where the observed data are incomplete, Shimodaira (1994) proposed a natural extension of AIC, called predictive divergence for incomplete observation model criterion (PDIO). Cavanaugh and Shumway (1998) derived a variant of AIC based on the motivation provided by Shmodaira (1994), which can be evaluated using only complete-data tools, readily available through the EM algorithm and the supplemented EM algorithm. The authors compared their criterion with AIC and PDIO by simulation. The results showed that Cavanaugh and Shumway's criterion was less prone to overfitting than AIC and less prone to underfitting than PDIO.

Shibata (1980) has shown that AIC is asymptotically efficient. However, AIC is not consistent. Note that AIC, FPE and C_p are asymptotically equivalent (see, e.g., Nishii 1984). For small samples, many researchers have shown that AIC leads to overfitting (see, e.g., Hurvich and Tsai 1989). For improving on AIC, Sugiura (1978) and Hurvich and

Tsai (1989) derived AICc by estimating the expected Kullback-Leibler distance directly in regression models, where a second order bias adjustment was made, and the criterion is given as follows:

$$\mathrm{AICc}(\boldsymbol{\theta}) = -2\log(\text{maximum likelihood}) + 2k\left(\frac{n}{n-k-1}\right) = \mathrm{AIC}(\boldsymbol{\theta}) + \frac{2k(k+1)}{n-k-1}, \quad (4.3)$$

where k denotes the number of free parameters in the candidate model. The model for which AICc is smallest is chosen. From (4.3), it can be seen that AICc has an additional bias correction term. If n is large with respect to k, then the second order correction is negligible and AIC should perform well, which implies that AICc and AIC are asymptotically equivalent and hence AICc is asymptotically efficient but not consistent. Findley (1985) noted that the study of the bias correction is of interest in itself; the exact small sample bias correction term varies by type of models involved.

Denote a model selection criterion by MSC. Model A (with k variables) will be considered better than Model B (with $k + \ell$ variables) if MSC(B) > MSC(A). Define the signal as E[MSC(B) – MSC(A)] and the noise as the standard deviation of the difference denoted by sd(MSC(B) – MSC(A)). Then the signal-to-noise ratio is defined as E[MSC(B) – MSC(A)]/sd(MSC(B) – MSC(A)). See McQuarrie and Tsai (1998) for more details. AIC has a weak signal-to-noise ratio (see, e.g., McQuarrie and Tsai 1998) and hence it tends to overfit. In contrast, AICc's has better signal-to-noise ratio so that AICc should perform well regarding the overfitting. The performance of model selection criteria with weak signal-to-noise ratios could be improved if their signal-to-noise ratios could be strengthened. Unfortunately, there is no single appropriate correction for all criteria.

For the model (2.2), AICu was proposed by McQuarrie, Shumway, and Tsai (1997), where S_κ/n of AIC term in (4.3) was replaced by $S_\kappa/(n-k)$ and the other term remains the same, which provides better model choices than AICc for moderate to large sample sizes except when the true model is of infinite order.

For improving on the inconsistency of AIC criterion, Akaike (1978) and Schwarz (1978) introduced equivalent consistent model selection criteria conceived from a Bayesian perspective. Schwarz derived SIC for selecting models in the Koopman-Darmois family, while Akaike derived his model selection criterion BIC for the problem of selecting a model in linear regression. The two procedures introduced about the same time are equivalent. See McQuarrie and Tsai (1998) for more details. In Schwarz (1978), it was assumed that the observations come from a Koopman-Darmois family with density of the form

$$f(x, \boldsymbol{\theta}) = \exp(\boldsymbol{y}(x)'\boldsymbol{\theta} - b(\boldsymbol{\theta})),$$

where $\boldsymbol{\theta} \in \Theta$, a convex subset of R^p, and $\boldsymbol{y}$ is a p-dimensional sufficient statistic for $\boldsymbol{\theta}$.

Since the exact distribution of the prior need not be known for the sake of asymptotic nature of SIC, it suffices to assume that the prior is of the form $\sum \gamma_j \mu_j$, where γ_j is the prior probability for model M_j, and μ_j is the conditional prior of $\boldsymbol{\theta}$ given M_j. Further, Schwarz assumed a fixed loss for selecting the wrong model. As stated in Schwarz (1978), the Bayes solution consists of selecting the model with a high posterior probability. In large samples, this posterior probability can be approximated by a Taylor expansion. Schwarz found its first term to be the log of the MLE for the model M_j and its second term was of the form $k \log(n)$ where k is the dimension of the model and n is the sample size. The remaining terms in the Taylor expansion were shown to be bounded and hence could be ignored in large samples. The SIC is given as follows: Let

$$\mathrm{SIC}(\boldsymbol{\theta}) = -2 \log(\text{maximum likelihood}) + k \log(n), \tag{4.4}$$

and choose the model for which SIC is smallest. It can be seen that the $2k$ term in AIC is replaced by $k \log(n)$ in SIC, which places a much stronger penalty for overfitting. When the parameters in SIC are estimated based on MDL principle, the resulting criterion is called MDL, which was derived in Rissanen (1978, 1983) under the assumption that there is no prior knowledge about $\boldsymbol{\theta}$.

BIC or SIC is strongly consistent but not asymptotically efficient. For small sample sizes, the chance of underfitting should not be overlooked. For improving the underfitting, it is natural to ask if $\log n$ can be replaced by a function of n which approaches infinity not as fast as $\log n$ when n tends to infinity. This function can not be constant. In Hannan and Quinn (1979), they argued, by applying the law of the iterated logarithm, that $\log n$ can be replaced by $c \log \log n$ with $c > 2$ in SIC without losing strong consistency. We call this new criterion HQ. When applying it, the underfitting is improved but does not vanish. It is unfortunate that a consistent model selection criterion usually tends to underfit when sample size is not large enough. All one can do is to find a consistent model selection criterion such that the underfitting is at its lowermost level. HQ meets such requirement.

For improving the small-sample performance of SIC, McQuarrie (1999) used the relationship between AIC and AICc to derive its small-sample correction denoted by SICc. He showed that SICc overfits less frequently than SIC, performs better in small samples and is asymptotically equivalent to SIC.

Consider the model (2.1). A framework is called prediction with repeated refitting if it allows model selection at each time, i.e., a model is chosen on the basis of the data available at time t, and the model selected is used to predict Y_{t+1}, while a framework is called prediction without refitting if a model is chosen on the basis of the training sample, and then the model selected is used to predict. Under the frame of finite-dimensional normal regression models, Speed and Yu (1993) compared model selection criteria according to prediction errors based upon prediction with refitting, and prediction without refitting and showed that Rissanen's accumulated prediction error and stochastic complexity criteria, AIC, SIC, and the FPE criteria achieve both low bounds for prediction with refitting and without refitting.

AIC was derived under the assumptions that (i) the estimation is by maximum likelihood and (ii) the parametric family of distributions includes the true model. Could these assumptions be somehow relaxed? Let

$$b(G) = \mathrm{E}_G\left[\frac{1}{n}\sum_{i=1}^{n}\log f_{\hat{\boldsymbol{\theta}}}(z_i) - \int \log f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{u})\, dG(\boldsymbol{u})\right],$$

where the expectation is taken over the true distribution G and $\hat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$. Without assuming that the true distribution belongs to the specified parametric family of probability distributions, $b(G)$ is asymptotically given by

$$b(G) = \frac{1}{n}\mathrm{tr}\{J(G)^{-1}I(G)\} + O(n^{-2}),$$

where $J(G)$ and $I(G)$ are defined by

$$J(G) = -\mathrm{E}_G\left[\frac{\partial^2 \log f_{\boldsymbol{\theta}}(z)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \quad \text{and} \quad I(G) = -\mathrm{E}_G\left[\frac{\partial \log f_{\boldsymbol{\theta}}(z)}{\partial\boldsymbol{\theta}}\frac{\partial \log f_{\boldsymbol{\theta}}(z)}{\partial\boldsymbol{\theta}'}\right].$$

Denote the bias corrected log likelihood by

$$\mathrm{TIC}(\boldsymbol{\theta}) = -2\log(\text{maximum likelihood}) + 2\mathrm{tr}\{\hat{J}(G)^{-1}\hat{I}(G)\},$$

where $\hat{J}(G)$ and $\hat{I}(G)$ are respectively consistent estimates of $J(G)$ and $I(G)$, and choose the model for which TIC is smallest. This criterion is called TIC and was originally introduced by Takeuchi (1976) and also Stone (1977a), and later discussed extensively by Shibata (1989) and Konishi (1999). When the true model is included in the set of candidate models, $b(G)$ can be reduced to

$$b(G) = \frac{k}{n} + O(n^{-2}),$$

where k is the number of free parameters in the model, and TIC becomes AIC. If none of the candidate models is close to the true model, TIC is an alternative if sample size is large.

A generalized information criterion (GIC) was introduced in Konishi and Kitagawa (1996) by estimating the same Kullback-Leibler distance as in AIC while relaxing both the assumptions (i) and (ii). If the bias $b(G)$ can be estimated by appropriate procedures, then the bias corrected log likelihood is given by

$$\mathrm{GIC}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\log f_{\hat{\boldsymbol{\theta}}}(z_i) - \hat{b}(G),$$

where $\hat{\boldsymbol{\theta}}$ may be obtained by maximum likelihood, penalized likelihood or robust procedures. The estimated bias $\hat{b}(G)$ is generally given as an asymptotic bias and an approximation to $b(G)$. A model is selected such that the GIC is smallest. Konishi and Kitagawa (1996) employed a functional estimator $\hat{\boldsymbol{\theta}} = \boldsymbol{t}(\hat{G})$ with Fisher consistency and approximated $b(G)$ by a function of the empirical influence function of the estimator and the score function of the parametric model. They obtained the GIC in the form

$$\text{GIC}(\boldsymbol{\theta}) = -2\sum_{i=1}^{n} \log f_{\hat{\boldsymbol{\theta}}}(z_i) + \frac{2}{n}\sum_{i=1}^{n} \text{tr}\left\{ \boldsymbol{t}^{(1)}(z_i; \hat{G}) \frac{\partial \log f_{\boldsymbol{\theta}}(z_i)}{\partial \boldsymbol{\theta}'}\Big|_{\hat{\boldsymbol{\theta}}} \right\}.$$

Here $\boldsymbol{t}^{(1)}(z_i;\hat{G}) = (t_1^{(1)}(z_i;\hat{G}), \ldots, t_k^{(1)}(z_i;\hat{G}))'$ and $t_i^{(1)}(z_i;\hat{G})$ is the empirical influence function defined by

$$t_i^{(1)}(z_i;\hat{G}) = \lim_{\varepsilon \to 0} \frac{t_i((1-\varepsilon)\hat{G} + \varepsilon\delta_i) - t_i(\hat{G})}{\varepsilon}$$

with δ_i being a point mass at z_i. Note that AIC and TIC are special cases of GIC.

In Bozdogan (1987), CAICF (C denoting "consistent" and F denoting the use of Fisher information matrix) was proposed. Let

$$CAICF(\boldsymbol{\theta}) = -2\log(\text{maximum likelihood}) + k[\log(n) + 2] + \log|\hat{J}|.$$

CAICF criterion chooses a model for which CAICF is smallest. In Bozdogan (1988), an information theoretic measure of complexity called ICOMP for model selection for general multivariate linear and nonlinear structural models was proposed. The author claimed that ICOMP takes the spirit of AIC, but it is a different procedure than AIC in the sense that ICOMP is based on the entropic characterization of the measure of complexity of a model and that such a formulation provides a criterion of goodness of fit of a model. For a multivariate normal linear and nonlinear structural models, ICOMP is defined by

$$\begin{aligned} \text{ICOMP}(\boldsymbol{\theta}) &= -2\log(\text{maximum likelihood}) \\ &+ \left\{ k\log\left(\frac{\text{tr}\hat{\Sigma}_k}{k}\right) - \log|\hat{R}_k| \right\} + \left\{ n\log\left(\frac{\text{tr}\hat{R}}{n}\right) - \log|\hat{\Sigma}_k| \right\} \end{aligned} \tag{4.5}$$

where $\hat{\Sigma}_k$ is the estimated covariance matrix and $\hat{R}$ is the model residuals. A model with minimum ICOMP is chosen to be the best model among all candidate models. The author argued that minimization of ICOMP provides a trade-off between the accuracy of the estimated parameters, as measured by the interactions among the parameters, and the independent normal errors. The author asserted that ICOMP leads to a parsimonious description of the fitted model. As commented in Burnham and Anderson (1998), neither CAICF or ICOMP are invariant to 1-to-1 transformations of the parameters, and this feature would seem to limit their application. From (4.5), it can be seen:

1. The second term generally has the order of $\log n$. When the eigenvalues of the estimated covariance matrix are asymptotically proportionally identical with certain rate, it may tend to zero, which may cause serious overfitting. This happens in the balanced ANOVA.

2. The third term vanishes in case errors are homogeneous or may have order of n (e.g., $\mathrm{Var}(\varepsilon_{2i} = 2\mathrm{Var}(\varepsilon_{2i-1}))$, which may cause serious underfitting.

In Wei (1992) a model selection criterion FIC was proposed for linear regression based on Fisher information. Consider the linear model (2.2). Assume that $\varepsilon_i \sim N(0, \sigma^2)$ and that $\boldsymbol{x}_i$ is $\sigma(\varepsilon_1, \ldots, \varepsilon_{i-1})$ measurable. Then the conditional Fisher information matrix for $\boldsymbol{\beta}(\kappa)$ is $\sigma^{-2} \sum_{i=1}^n \boldsymbol{x}_i(\kappa)\boldsymbol{x}_i'(\kappa)$. The quantity $|\sum_{i=1}^n \boldsymbol{x}_i(\kappa)\boldsymbol{x}_i'(\kappa)|$ can be interpreted as the amount of information about $\boldsymbol{\beta}(\kappa)$. Denote

$$\mathrm{FIC}(\kappa) = n\hat{\sigma}_\kappa^2 + \hat{\sigma}^2 \log |\sum_{i=1}^n \boldsymbol{x}_i(\kappa)\boldsymbol{x}_i'(\kappa)|.$$

A model is selected for which FIC is the smallest. In FIC, the redundant information by introducing a spurious variable is used to represent its penalty. Compared with PLS (predictive least squares), the author argued that FIC is permutation invariant, easy to compute, no initialization problem is involved and is strongly consistent and, further, FIC seems to have better small sample performance.

A widely used procedure for inference about parameters of interest in the presence of nuisance parameters is based on the profile log-likelihood function. However, this procedure may give inconsistent or inefficient estimates. Since the profile log-likelihood itself is not a log-likelihood, Shi and Tsai (1998a) argued that one must consider conditional log-likelihood, modified profile log-likelihood, or marginal log-likelihood as alternative approaches. For simplicity, they proposed a model selection criterion based on marginal log-likelihood for linear regression. They first obtained an unbiased estimator of the expected Kullback-Leibler information of the marginal log-likelihood function of the fitted model and then derived the modified information criterion (MIC) based on it. Under some conditions, MIC is shown to be strongly consistent. Based on their simulation results, they indicated that MIC not only outperforms the efficient criteria AIC, AICc, FPE and C_p, but is superior (or comparable) to the consistent criteria BIC and FIC in both small and large sample sizes.

Consider the model (2.2). First let the set of all candidate models consist of $\{M_j\}$ where $M_j = M_{1:j}$. Denote $S_j = S_{1:j}$. Define

(1) $G_n^{(1)}(k) = S_k + kC_nS_p/(n-p), \quad k = 1, \ldots, p;$

(2) $G_n^{(2)}(k) = S_k + kC_n, \quad k = 1, \ldots, p;$

(3) $G_n^{(3)}(k) = n \log S_k + kC_n, \quad k = 1, \dots, p.$

Rao and Wu (1989) and Bai, Rao, and Wu (1999) proposed the following selection rules based on $G_n^{(\ell)}$'s; the selected model is defined by $M_{\hat{k}_n}$ for which

$$G_n^{(\ell)}(\hat{k}_n) = \min_{1 \le k \le p} G_n^{(\ell)}(k). \tag{4.6}$$

The selection procedures defined above are called the general information criteria for linear regression (GIC-LR). Note that the equivalent criteria can be found in literature (e.g., Nishii 1984, Pötscher 1989, Shao 1997). Assume that C_n is a function of n satisfying the conditions

$$\frac{C_n}{n} \to 0, \quad \frac{C_n}{\log \log n} \to \infty. \tag{4.7}$$

It was shown in Rao and Wu (1989) and Bai, Rao, and Wu (1999) among others that under mild conditions, these criteria are strongly consistent.

Now consider the general situation where the set of all candidate models consist of all possible 2^p submodels. For each $1 \le i \le p$, denote

$$\boldsymbol{\beta}_{-i} = (\beta_1, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_p)'$$

and

$$X_{n,-i} = X_n(\{1 : i-1\} \bigcup \{i+1 : p\}).$$

Consider the model

$$y_n = X_{n,-i}\boldsymbol{\beta}_{-i} + \varepsilon_n. \tag{4.8}$$

Write the corresponding usual residual sum of squares by S_{-i}. In order to determine if the ith element of $\boldsymbol{\beta}$ is zero, we need only compare two models, one is the full model (2.1) and the other is the reduced model (4.8). Define, for $1 \le i \le p$,

(1) $G_n^{(1)}(-i) = S_{-i} - S_p - C_n S_p/(n-p);$

(2) $G_n^{(2)}(-i) = S_{-i} - S_p - C_n;$

(3) $G_n^{(3)}(-i) = n(\log S_k - \log S_p) - C_n.$

Then, choose the model as

$$\beta_i = 0 \quad \text{if} \quad G_n^{(\ell)}(-i) \le 0 \quad \text{and} \quad \beta_i \ne 0 \quad \text{if} \quad G_n^{(\ell)}(-i) > 0$$
$$i = 1, \dots, p.$$

Assume that C_n is chosen in accordance with the condition (4.7). It was shown in Rao and Wu (1989) and Bai, Rao, and Wu (1999) that under mild conditions, these criteria are strongly consistent. The advantage of such selection procedures is that it needs only the computation of $p + 1$ residual sums of squares instead of 2^p residual sums of squares.

When p is large, this method can also be applied to the criteria discussed previously. Its disadvantage is that the underfitting may not be at nominal level. As an alternative method (Zheng and Loh 1995), the explanatory variables may be ordered using the t statistic. Using this method may need extra work in computing the t statistics and sorting. If all important predictors are significantly non-negligible, every method can give good detection of the smallest true model. The problem arises when some predictors are very critical. There are cases where a predictor is shown to be important when just this predictor is examined, and it becomes unimportant when the effect of some other predictors is eliminated (of course it is not for the limiting case). Therefore, a variable which is in fact more important than another may be excluded by wrong ordering. Both methods proposed by Zheng and Loh (1995) and Bai, Rao, and Wu (1999) respectively may face this problem. It needs deeper investigation to determine which performs better.

It is worth while mentioning that the assumption of normality and the assumption that the errors are identically distributed are not necessary for the criteria to be strongly consistent in this example. See Bai, Rao, and Wu (1999) for details.

In the problem of selecting a linear model to approximate the true unknown regression model, some necessary and/or sufficient conditions were established by Shao (1997) for the asymptotic validity of various model selection procedures such as AIC, C_p, FPE$_\alpha$, BIC, SIC, GIC-LR, etc.. It was found that these selection procedures can be classified into three classes according to their asymptotic behavior. Under some fairly weak conditions, Shao (1997) showed that the selection procedures in one class are asymptotically valid if there exist fixed-dimension correct models; the selection procedures in another class are asymptotically valid if no fixed-dimension correct model exists. The procedures in the third class are compromises of the procedures in the first two classes.

Since the general information criterion for linear regression is consistent, it is of interest to know its convergence rate. In Shao (1998), some convergence rates for the error probabilities of the criterion, in terms of C_n and the order of the design matrix, were established. The author argued that the rates obtained there are sharper than the existing ones in the literature (e.g., Zhang 1993b) when the distribution of the response variable is non-normal.

5 Cross-validation, bootstrap and related model selection methods

Cross-validation is a method for model selection in terms of the predictive ability of the models. Suppose that n data points are available. A model is to be selected from a class of models. First, hold one data point and use the rest of $n-1$ data points to fit a model. Then check the predictive ability of the model in terms of the withheld data point. Perform this procedure for all data points. Select the model with the best average predictive ability. This procedure is described as the LOO (leave one out) method. For details, see Stone (1974, 1977a,b), Geisser (1975), Efron (1983, 1986), Picard and Cook (1984), Herzberg and Tsukanov (1986) and Li (1987) among others. Note that Allen's PRESS is equivalent to cross-validation (Allen 1974).

The cross-validation can be generalized as follows. Instead of choosing one data point for assessing the predictive ability, k data points are reserved for that purpose. The rest of $n_k = n - k$ data points are used to fit the model. There are ${}_nC_k$ different ways to partition the data set. The generalized cross-validation selects the model with the best average predictive ability among different ways of data splitting. It is easy to see that the computational complexity of this method increases as k increases.

Herzberg and Tsukanov (1986) did some simulation comparisons between the cross validation procedures with $k \equiv 1$ and $k \equiv 2$. They found that the leave-two-out cross-validation is sometimes better than the LOO cross-validation, although the two procedures are asymptotically equivalent in theory. When the number of predictors in any regression model under consideration is fixed, this type of cross-validation is not consistent and it can be shown that it is equivalent to Akaike information criterion. See also Geisser (1975), Burman (1989), and Zhang (1993a). It will not be the case if k is chosen to depend on n. For emphasizing this dependence, write k as $k(n)$. Shao (1993) showed that $k(n)/n \to 1$ as $n \to \infty$ is needed to guarantee that the cross-validation is asymptotically correct. When $k(n)$ is large, the amount of computation required to use the cross-validation seems impractical. Shao (1993) suggested several approaches (e.g., the balanced incomplete CV($k(n)$) and Monte Carlo CV($k(n)$)) to remedy it, and examined their performances in a simulation study. Wu, Tam, Li, and Zen (1999) have used Shao's method to estimate the number of super imposed exponential signals.

When the number of predictors in the regression model under consideration increases as n increases, the story is different. Li (1987) showed that under some conditions, the LOO cross-validation is consistent and is asymptotically optimal in some sense.

The bootstrap is a data resampling method for estimating or approximating the sampling distribution of a statistic and its characteristics. The general application of bootstrap method to model selection can be found in Linhart and Zucchini (1986). Recent developments in this area include bootstraping the mean squared prediction error (Shao 1996 and McQuarrie and Tsai 1998) and constructing bootstrapped estimate for the Kullback-Leibler discrepancy (Shibata 1997).

Breiman (1992) introduced a data-driven model selection based on the little bootstrap. Consider the model (2.2). Define the prediction error as

$$\mathrm{E}\|\boldsymbol{Y}_0 - X_n(\kappa)'\beta(\kappa)\|^2 = N\sigma^2 + \|\mathrm{E}(\boldsymbol{Y}) - X_n(\kappa)'\beta(\kappa)\|^2,$$

where $\boldsymbol{Y}_0$ denotes the vector of n new observations. Write

$$\mathrm{ME} = \|\mathrm{E}(\boldsymbol{Y}) - X_n(\kappa)'\beta(\kappa)\|^2,$$

which is the error in fitting the true model. The following procedure describes how to get the little bootstrap ME estimate:

1. 1. Fit the full model (2.1), getting S_p and $\hat\sigma^2$. Do the variable selection, getting the sequence of subsets of indices $\kappa_0, \kappa_1, \ldots, \kappa_p$, and the values of S_{κ_j}, where $\kappa_0 = \emptyset$.

2. Generate $\{\varepsilon_{1i}\}$, $i = 1, \ldots, n$, as i.i.d. $N(0, t^2\hat\sigma^2)$ and form the new y data
$$\tilde{\boldsymbol{y}} = \boldsymbol{y} + \boldsymbol{\varepsilon}_1,$$
where $\boldsymbol{\varepsilon}_1 = (\varepsilon_{11}, \ldots, \varepsilon_{1n})'$ and $t > 0$.

3. Using the data $(\tilde y_i, \boldsymbol{x}_i)$, find the subset sequence $\{\tilde\kappa_j\}$ following the same procedure as in Step 1, and compute the predictors $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{y}}_{\tilde\kappa_j}$ based on the full model and the model $M_{\tilde\kappa_j}$.

4. Calculate
$$\frac{1}{t^2}\boldsymbol{\varepsilon}_1'(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{\tilde\kappa_j}).$$

5. Repeat Steps 2, 3, and 4 a number of times and average the quantities computed in Step 4. Denote the average by $B_t(j)$.

6. The little bootstrap estimate is
$$\widehat{\mathrm{ME}}(\kappa_j) = S_{\kappa_j} - S_p + p\hat\sigma^2 - 2B_t(j).$$

Brieman proposed to select $M_{\tilde\kappa_{j*}}$ if

$$M_{\tilde\kappa_{j*}} = \arg\min_{M_{\tilde\kappa_{j*}}} \widehat{\mathrm{ME}}(\kappa_j).$$

In his paper, he commented on the choice of t and compared his method with C_p and replicate data method by simulation. His simulation results indicate that his method is better than C_p and he argued that all selection methods not based on data reuse give highly biased results and poor subset selection.

Recently the covariance inflation criterion (CIC) was proposed in Tibshirani and Knight (1999a), which adjusts the training error by applying the model selection procedure to permuted versions of the data set, to measure the covariance between the predicted values and the responses. In doing so, this criterion captures the inherent variability associated with an adaptive procedure, such as best subset regression. The CIC can be applied in the prediction problems such as regression and classification, and to nonlinear, adaptive prediction rules.

Consider the models (2.2) with squared error and fixed $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. Denote $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i)$, $i = 1, \dots, n$, and $\boldsymbol{z}' = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$. Let $\mu_i = \mathrm{E}(Y_i|\boldsymbol{x}_i)$, $\sigma^2 = \mathrm{Var}(Y_i|\boldsymbol{x}_i)$, and the conditional distribution of $Y_i|\boldsymbol{x}_i$ be F_{μ_i}. On the basis of $\boldsymbol{z}$ a model M_κ is chosen and the corresponding prediction rule $\eta_{\boldsymbol{z}}(\boldsymbol{x}, M_\kappa)$, indexed by a tuning parameter κ, is formulated.

The true error of the rule $\eta_{\boldsymbol{z}}(\boldsymbol{x}, M_\kappa)$ is

$$\mathrm{Err}(\kappa) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{\mu_i}\{Y_i^* - \eta_{\boldsymbol{z}}(\boldsymbol{x}_i, M_\kappa)\}^2,$$

where $Y_i^* \sim F_{\mu_i}$ with the training set $\boldsymbol{z}$ fixed. This is sampling error and the training error (or apparent error) is

$$\overline{\mathrm{err}}(\kappa) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{\mu_i}\{y_i - \eta_{\boldsymbol{z}}(\boldsymbol{x}_i, M_\kappa)\}^2.$$

Note that $\overline{\mathrm{err}}$ tends to be biased downwards as an estimate of Err because the training set $\boldsymbol{z}$ is used twice, both to construct the rule and to test it.

let $\hat{\sigma}^2$ be an estimate of the noise variance σ^2 and let

$$\hat{\sigma}_y^2 = \sum_{i=1}^{n}(y_i - \bar{y})/(n-1).$$

The covariance inflation criterion (CIC) is defined by

$$\mathrm{CIC}(\kappa) = \overline{\mathrm{err}}(\kappa) + \frac{2}{n}\frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}\sum_{i=1}^{n}\mathrm{Cov}^0(Y_i^*, \eta_{\boldsymbol{z}^*}(\boldsymbol{x}_i, M_\kappa^*)) + \frac{2}{n}\hat{\sigma}^2.$$

A model is chosen to be the one which minimizes CIC(κ). The notation Cov^0 indicates covariance under the permutation distribution of $\boldsymbol{x}$ and y: $\boldsymbol{x}_i^* = (y_i^*, \boldsymbol{x}_i)$ with $y_1^*, \dots, y_n^*$ a sample drawn without replacement from $y_1, \dots, y_n$ and the $\boldsymbol{x}_i$ fixed. Here M_κ^* is the model, given a tuning parameter κ, chosen from the permuted data. Tibshirani and Knight (1999a) argued that the idea behind this definition is that, the harder the data are fitted, the more $\eta_{\boldsymbol{z}}(\boldsymbol{x}_i, M_\kappa)$ affects its own prediction, and hence the greater the optimism in $\overline{\mathrm{err}}(\kappa)$. Since in practice it is hard to compute all the permutations, the authors suggested taking a sample of them. Note that instead of sampling the responses without replacement, they can be sampled with replacement, and given an independent bootstrap distribution for the predictors and responses. See Tibshirani and Knight (1999a) for details. The authors commented that even if the little bootstrap procedure looks similar to the CIC in the context of linear regression, the uncertainty in the choice of t makes it difficult to generalize the little bootstrap to other settings. In contrast, the CIC can be defined for general prediction models for regression and classification, which was presented in Tibshirani and Knight (1999a). The CIC was also compared with AIC, BIC and RIC (the risk inflation criterion) in the paper.

Brieman (1996) showed how one can use the bootstrap for the more primary purpose of producing a better estimator. Breiman's bagging procedure applies a given estimator $\hat{\boldsymbol{\theta}}$ to each of B bootstrap samples, and then averages the B values to produce a new estimator $\tilde{\boldsymbol{\theta}}$. In a number of experiments involving trees, subset selection and ridge regression, Breiman showed that the bagged estimate $\tilde{\boldsymbol{\theta}}$ often has smaller mean squared error than the original $\hat{\boldsymbol{\theta}}$. The largest gains occurred for unstable estimators $\hat{\boldsymbol{\theta}}$, like subset selection and trees, for which small changes in the data can produce large changes in the estimate. The improvement in mean squared error is mostly due to a reduction in variance. As commented in Tibshirani and Knight (1999b), unfortunately the averaging process that produces the bagged estimate $\tilde{\boldsymbol{\theta}}$ also destroys any simple structure that is present in the original estimate $\hat{\boldsymbol{\theta}}$. A different use of the bootstrap was proposed in Tibshirani and Knight (1999b). They used bootstrap samples to provide candidate models for the model search. They argued that this has the advantage that it preserves the structure of the estimator while still inducing stability. Let $\boldsymbol{z} = (z_1, \ldots, z_n)$ be a training sample i.i.d. from a distribution F. Suppose that there is a model for the data that depends on a set of parameters $\boldsymbol{\theta}$. From the training sample, it is assumed that $\boldsymbol{\theta}$ is to be estimated by minimizing a target criterion

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} R(\boldsymbol{z}, \boldsymbol{\theta}).$$

Suppose also that there is a (possibly different) working criterion R_0 for which minimization is convenient. Tibshirani and Knight (1999b) proposed to estimate $\boldsymbol{\theta}$ by drawing bootstrap samples $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^B$, estimating $\hat{\boldsymbol{\theta}}$ via R_0 from each sample

$$\hat{\boldsymbol{\theta}}^b = \arg\min_{\boldsymbol{\theta}} R_0(\boldsymbol{z}^b, \boldsymbol{\theta}),$$

and then choosing $\hat{\boldsymbol{\theta}}$ as the value among the $\hat{\boldsymbol{\theta}}^b$ producing the smallest value of $R(\boldsymbol{z}, \theta)$:

$$\hat{\boldsymbol{\theta}}^B = \hat{\boldsymbol{\theta}}^{*b} \quad \text{where } \hat{\boldsymbol{\theta}}^{*b} = \arg\min_{b} R(\boldsymbol{z}, \hat{\boldsymbol{\theta}}^b).$$

As a convention, the original sample $\boldsymbol{z}$ is included among the B bootstrap samples. This procedure is called "Bumping" for Bootstrap Umbrella of Model Parameters. The value

$\hat{\boldsymbol{\theta}}^B$ is the bumping estimate of $\boldsymbol{\theta}$. The authors argued that the bumping provides a convenient method for finding better local minima, for resistant fitting, and for optimization under constraints.

6 Baysian approach to model selection

For the model (2.1), assume that $\boldsymbol{Y}|\boldsymbol{\beta},\sigma \sim N(X\boldsymbol{\beta},\sigma^2 I)$. Let $\boldsymbol{\tau} = (\tau_1,\ldots,\tau_p)'$ where $\tau_i = 0$ or 1 if β_i is small or large, respectively. The size of the $\boldsymbol{\tau}$th subset is denoted as $q_{\boldsymbol{\tau}} \equiv \boldsymbol{\tau}'\mathbf{1}$. Since the appropriate value of $\boldsymbol{\tau}$ is unknown, the uncertainty underlying variable selection is modelled by a hierarchical mixture prior $\pi(\boldsymbol{\beta},\sigma,\boldsymbol{\tau}) = \pi(\boldsymbol{\beta}|\sigma,\boldsymbol{\tau})\pi(\sigma|\boldsymbol{\tau})\pi(\boldsymbol{\tau})$. For this hierarchical setup, the marginal posterior distribution $\pi(\boldsymbol{\tau}|Y)$ contains the relevant information for variable selection. Based on the data Y, the posterior $\pi(\boldsymbol{\tau}|Y)$ updates the prior probabilities on each of the 2^p possible values of $\boldsymbol{\tau}$. Identifying each $\boldsymbol{\tau}$ with a submodel via $(\tau_i = 1) \Leftrightarrow (X(i)$ is included), those $\boldsymbol{\tau}$ with higher posterior probability $\pi(\boldsymbol{\tau}|Y)$ identify the more "promising" submodels, that is those supported most by the data and the prior distribution.

For identifying "promising" subsets of predictors for the model (2.1), a Bayesian procedure, called stochastic search variable selection (SSVS), was proposed in George and McCulloch (1993), which specifies a hierarchical Bayes mixture prior which uses the data to assign larger posterior probability to the more promising models. To avoid the overwhelming burden of calculating the posterior probabilities of all 2^p models, SSVS uses the Gibbs sampler to simulate a sample from the posterior distribution. The Gibbs sampler is effectively used to search for promising models rather than compute the entire posterior. The key to the potential of SSVS is the fast and efficient simulation of the Gibbs sampler.

George and McCulloch (1997) described a variety of approaches to Bayesian variable selection which includes SSVS as a special case. These approaches all use hierarchical mixture priors to describe the uncertainty present in variable selection problems. In the paper, hyperparameter settings which base selection on practical significance, and the implications of using mixtures with point priors were discussed. The authors showed that conjugate versions of these priors yield expressions for the posterior which can sometimes be sequentially computed using efficient updating schemes. According to the paper, when p is moderate (less than about 25), performing such sequential updating in a Gray Code order yields a feasible approach for exhaustive evaluation of all 2^p posterior probabilities, and for larger values of p, Markov chain Monte Carlo (MCMC) methods, such as the Gibbs sampler or the Metropolis-Hastings algorithms, can exploit such updating schemes to rapidly search for high probability models. The authors observed that estimation of normalization constants would provide improved posterior estimates of individual model probabilities and the total visited probability. In their paper, nonconjugate and conjugate MCMC implementations are compared on three simulated sample problems. They also illustrated the application of Bayesian variable selection to a real problem involving $p = 200$ potential regressors.

As discussed in George and McCulloch (1997), a variety of approaches for Bayesian model selection can be put into the following categories: 1. Prior specification and 2. Posterior computation. The prior specification corresponding to the removal of a predictor can be obtained by either a continuous distribution on β_i with high concentration at 0, or assigning an atom of probability to the event $\beta_i = 0$. Another distinguishing characteristic of prior specification is the difference between nonconjugate and conjugate forms for the coefficient priors. Nonconjugate forms offer the advantage of precise specification of a nonzero threshold of practical significance, and appear to allow for more efficient MCMC exploration with approximately uncorrelated predictors. Conjugate forms offer the advantage of analytical simplification which allows for exhaustive posterior evaluation in moderately sized problems (p less than about 25). In problems with large p where evaluation of posterior probabilities is not feasible, conjugate forms allow for exact calculation of relative posterior probabilities and estimates of total visited probability by MCMC posterior exploration. Furthermore, conjugate forms appear to allow for more efficient MCMC exploration with more correlated designs. For the purpose of posterior exploration, a large variety of MCMC algorithms can be constructed based on the Gibbs sampler and Metropolis-Hastings algorithms.

With prediction as the goal, Geisser (1993) considered it more appropriate to average predictions over the posterior distribution rather than using predictions from any single model. The potential of prediction averaging in the context of variable selection uncertainty has been nicely illustrated by Clyde, DeSimone, and Parmigiani (1996) and Raftery, Madigan, and Hoeting (1997) among others. In practice, there exists a situation where a single model is needed for prediction (e.g., Wakefield and Bennett 1996).

In Laud and Ibrahim (1995), a predictive Baysian viewpoint was advocated to avoid the specification of prior probabilities for the candidate models and the detailed interpretation of the parameters in each model. Consider probability models for the observable $\boldsymbol{y}$ conditioned on each model M_γ with the associated parameter vector $\boldsymbol{\theta}^{(M_\gamma)}$:

$$p(\boldsymbol{y}|M_\gamma, \boldsymbol{\theta}^{(M_\gamma)}), \quad \boldsymbol{\theta}^{(M_\gamma)} \in \Theta^{(M_\gamma)}, \; \gamma \in \Gamma,$$

where $\Theta^{(M_\gamma)}$ is the parameter space for the model M_γ and Γ is the index set. Suppose that a prior $\pi(\boldsymbol{\theta}^{(M_\gamma)}|M_\gamma)$ has been specified for each $\boldsymbol{\theta}^{(M_\gamma)}$, $\gamma \in \Gamma$. The posterior for $\boldsymbol{\theta}^{(M_\gamma)}$ under each model M_γ, given data $\boldsymbol{Y} = \boldsymbol{y}$, is given by $\pi(\boldsymbol{\theta}^{(M_\gamma)}|\boldsymbol{y},\, M_\gamma)$. Now envision replicating the entire experiment and denote by $\boldsymbol{Z}$ the vector of responses that might result. The predictive density for $\boldsymbol{Z}$ under model M_γ is

$$p(z|M_\gamma, \boldsymbol{y}) = \int p(z|M_\gamma, \boldsymbol{\theta}^{(M_\gamma)})\pi(\boldsymbol{\theta}^{(M_\gamma)}|\boldsymbol{y},\ M_\gamma)\ d\boldsymbol{\theta}^{(M_\gamma)}.$$

This density was called the predictive density of a replicate experiment (PDRE) in Laud and Ibrahim (1995). The replicative experiment is an imaginary device that puts the predictive density to inferential use, adapting the philosophy advocated in Geisser (1971). The imagined replication makes $\boldsymbol{y}$ and $\boldsymbol{Z}$ comparable, in fact exchangeable a priori. Moreover, the parameters in the model play a minimal role under replication. It seems clear that good models, among those under consideration, should make predictions close to what has been observed for an identical experiment. With this motivation, Laud and Ibrahim (1995) proposed three criteria. Consider

$$L_{M_\gamma} = \sqrt{\mathrm{E}\{(\boldsymbol{Z}-\boldsymbol{y})'(\boldsymbol{Z}-\boldsymbol{y})\}},$$

where the expectation is taken with respect to the PDRE. Good models will have small values of L_{M_γ}, which results in the first criterion. Based on PDRE itself, the second criterion is formulated considering that small values of $(\mathrm{PDRE})^{-1/n}$ indicate good models. The third criterion is based on the Kullback-Leibler distance between two predictive densities. Using these criteria, they implemented their methodology for three common problems arising in normal linear models: variable subset selection, selection of a transformation of predictor variables and estimation of a parametric variance function.

Suppose that we are considering two models, M_1 and M_2. Let $p(\boldsymbol{y}|M_i)$ and $p(M_i)$ be respectively the distribution of the data $\boldsymbol{Y}$ and the prior probability of the model M_i, $i = 1, 2$. The posterior probabilities of M_i, $i = 1, 2$, are given by

$$p(M_i|\boldsymbol{y}) = p(\boldsymbol{y}|M_i)p(M_i)/p(\boldsymbol{y}). \tag{6.1}$$

The posterior odds in favor of model M_1 over alternative M_2 are

$$\frac{p(M_1|\boldsymbol{y})}{p(M_2|\boldsymbol{y})} = \left(\frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)}\right)\left(\frac{p(M_1)}{p(M_2)}\right). \tag{6.2}$$

Let $\pi_i(\boldsymbol{\theta}_i)$ be the prior distributions of the d_i-dimensional parameter vector $\boldsymbol{\theta}_i$ under the models M_i, $i = 1, 2$. Expressing $p(\boldsymbol{y}|M_i)$ as the average of the usual likelihood $p(\boldsymbol{y}|\boldsymbol{\theta}_i)$ over the parameter space, we have

$$p(\boldsymbol{y}|M_i) = \int p(\boldsymbol{y}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\ d\boldsymbol{\theta}_i,$$

which, together with (6.1), implies that

$$p(M_i|\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\ d\boldsymbol{\theta}_i p(M_i)/p(\boldsymbol{y}).$$

Hence, (6.2) can be expressed as

$$\frac{p(M_1|\boldsymbol{y})}{p(M_2|\boldsymbol{y})} = \left(\frac{\int p(\boldsymbol{y}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1}{\int p(\boldsymbol{y}|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2)\, d\boldsymbol{\theta}_2} \right) \left(\frac{p(M_1)}{p(M_2)} \right).$$

Since it is often that the prior odds $p(M_1)/p(M_2)$ is 1, the ratio

$$B_{12} = \frac{\int p(\boldsymbol{y}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1}{\int p(\boldsymbol{y}|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2)\, d\boldsymbol{\theta}_2}$$

is defined as the Bayes factor. B_{12} can be viewed as the "weighted" likelihood ratio of M_1 to M_2 and hence can be interpreted solely in terms of comparative support of the data for the two models (see Kass and Raftery 1995). Computing B_{12} requires specification of $\pi_i(\boldsymbol{\theta}_i)$, $i = 1, 2$. Often in Bayesian analysis, one can use noninformative (or default) priors. Commonly used priors are the "uniform" prior, the Jeffreys prior, and the reference prior (see, e.g., Berger and Bernardo 1992). Since the Bayes factor compares model M_1 to alternative M_2, it has been used for model selection.

Akaike (1983) mentioned that model comparisons based on the AIC are asymptotically equivalent to those based on Bayes factors. As commented in Kass and Raftery (1995), this is true only if the precision of the prior is comparable to that of the likelihood, but not in the more usual situation where prior information is small relative to the information provided by the data. In the latter more usual situation, the SIC indicates that the model with the highest posterior probabilities is the one that minimizes SIC($\boldsymbol{\theta}$) given in (4.4).

It is unfortunate that the Bayes factors typically depend rather strongly on the prior information and several problems arise using the Bayes factor when prior information is weak (see, e.g., Berger and Pericchi 1996a,b, De Santis and Spezzaferri 1997, Kass and Raftery 1995). As commented in De Santis and Spezzaferri (1997), assigning a diffuse proper prior to the parameters $\boldsymbol{\theta}_i$ is critical because the flatter the prior is, the more penalized the corresponding model M_i is. Furthermore, when the distribution $\pi_i(\boldsymbol{\theta}_i)$ is improper and defined only up to arbitrary constants, the Bayes factor itself is a multiple of these arbitrary constants. In this situation of weak prior information, several authors have proposed the use of partial Bayes factors (see, among others, Berger and Pericchi 1996a,b, O'Hagan 1995). The idea is to use part of the data as a training sample to update the prior distributions of the models and the remainder of the data to compute the Bayes factor.

Assume $\boldsymbol{y} = (\boldsymbol{y}(m)', \boldsymbol{y}(n-m)')'$, where $\boldsymbol{y}(m)$ is a proper training sample of size m, that is, a subsample such that $0 < \int p(\boldsymbol{y}(m)|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i < \infty$, $i = 1, 2$. The training sample is minimal if it does not contain subsets that are proper training samples. The partial Bayes factor for model M_1 against model M_2 is then defined as

$$B_{12}(m) = \frac{\int p(\boldsymbol{y}(n-m)|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1|\boldsymbol{y}(m))\,d\boldsymbol{\theta}_1}{\int p(\boldsymbol{y}(n-m)|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2|\boldsymbol{y}(m))\,d\boldsymbol{\theta}_2}, \tag{6.3}$$

where $\pi_i(\boldsymbol{\theta}_i|\boldsymbol{y}(m))$ is the posterior distribution of the parameter $\boldsymbol{\theta}_i$, $i = 1, 2$. O'Hagan (1995) showed that the partial Bayes factor is less sensitive to the priors than the Bayes factor. By (6.3), it can be seen that the partial Bayes factor does not depend on absolute values of prior distributions but on their relative values and on the other hand, the partial Bayes factor depends on the specific $\boldsymbol{y}(m)$ chosen. As described in De Santis and Spezzaferri (1997), in finite sample, when the size m of the training sample increases, the sensitivity of the partial Bayes factor to prior distributions decreases, but at the same time its discriminatory power decreases. To eliminate the dependence of the partial Bayes factor on $\boldsymbol{y}(m)$ and to increase its stability, Berger and Pericchi (1996a,b) suggested averaging the partial Bayes factor corresponding to all the possible training samples and obtained intrinsic Bayes factor. A simple alternative that avoids averaging is the fractional Bayes factor proposed in O'Hagan (1995), which is given as

$$B_{12}^{\mathrm{F}}(m) = B_{12}\left(\frac{\int [p(\boldsymbol{y}|\boldsymbol{\theta}_2)]^{m/n}\pi_2(\boldsymbol{\theta}_2)\,d\boldsymbol{\theta}_2}{\int [p(\boldsymbol{y}|\boldsymbol{\theta}_1)]^{m/n}\pi_1(\boldsymbol{\theta}_1)\,d\boldsymbol{\theta}_1}\right).$$

The fractional Bayes factor has an asymptotic motivation: if m and n are both large, the likelihood based on $\boldsymbol{y}(m)$ is approximated by the one based on $\boldsymbol{y}$, raised to the power m/n. The comparison of the intrinsic Bayes factor and the fractional Bayes factor can be found in Berger and Pericchi (1996a,b) and De Santis and Spezzaferri (1997) among others.

7 Robust model selection

For the methods discussed in previous sections, it can be seen that there is an involvement of the distribution information of the models, direct, indirect, or as vehicle. How can we cope with the case when there are departures from the distributional assumptions or there exist outliers in the data at hand? Robust model selection criteria have been proposed for this purpose.

According to Qian and Künsch (1999), the following three issues should be taken into consideration when proposing a robust criterion. First, it should take into account the possibility that observations of both response and predictors may contain gross errors.

Therefore, the criterion should be somewhat robust to small number of outliers or small change in all of the data. Second, the criterion should be consistent if a finite-dimensional true model exists. Third, the effect of the signal-to-noise ratio on the empirical performance of the criterion needs to be taken care of.

Back to the literature, Ronchetti (1985) proposed a robust version of AIC, called AICR. In AICR, the first term of (4.2) is replaced by the sum of the discrepancy functions computed at the M-estimate and the second term is replaced by a product of α and the number of free parameters, where the choice of α follows from the asymptotic equivalence of the AIC given in Stone (1977a). It is easy to see that robustness of the AICR depends on the robustness of the M-estimation. Hampel (1983) suggested a modified version of it. Härdle (1987) investigated the properties of a selection criterion for regression which is asymptotically equivalent to the AICR and showed that it is asymptotically efficient. A similar idea was used by Martin (1980) and Behrens (1991) for autoregressive models. For general parametric models AICR was discussed in Ronchetti (1997).

Antoch (1986, 1987) introduced an algorithm to perform variable selection, where α-trimmed least squares estimators for parameters are computed for all possible submodels and then compared to the same estimator obtained in the full model. The submodels which lead to estimates whcih are "indistinguishable" from that of the full model are considered acceptable.

In Hurvich and Tsai (1990) a small sample criterion for the selection of least absolute deviations regression models was developed. Their criterion provides an exactly unbiased estimator for the expected Kullback-Leibler information when the error distribution is double exponential and the model is not underfitted. The selection procedure performs better than both AIC and AICR with the L_1-norm discrepancy function.

Recently, Ronchetti and Staudte (1994) presented a robust version of Mallow's C_p, denoted RC_p, which can be used with a large variety of robust estimators for the parameters, including M-estimators, bounded influence estimators, and one-step M-estimators with a high breakdown starting point. RC_p chooses the models that fit the majority of the data by taking into account the presence of outliers and possible departures from the normality assumption on the error distribution. Later, Sommer and Staudte (1995) implemented RC_p for Mallows-type estimators so that leverage points as well as outliers can be downweighted. Some examples can be found in both the papers to support the applications of RC_p. Another robust version of C_p can be derived from the Wald test as proposed by Sommer and Huggins (1996). Consider a set of candidate models $\{M_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta\}$, where the candidate models are indexed by their parameter vector $\boldsymbol{\theta}$ and Θ is the parameter vector space in $\mathcal{R}^p$. Let $\hat{\boldsymbol{\theta}}$ be an M-estimator of $\boldsymbol{\theta}$. Define $W_\kappa = W_{\boldsymbol{\theta}(\kappa)} - p + 2k$, where $W_{\boldsymbol{\theta}(\kappa)} = n\hat{\boldsymbol{\theta}}(\kappa)'\widehat{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}(\kappa)\hat{\boldsymbol{\theta}}(\kappa)$, the Wald test statistic for testing the null hypothesis that $\boldsymbol{\theta}(\kappa) = \mathbf{0}$, and κ denotes a subset of $\{1, \ldots, p\}$. Then a model is selected for which W_κ is the smallest. An advantage of such a model selection criterion is that it is easily adapted to other nonadditive error model structures.

Using Shao's cross-validation methods (Shao 1993) for choice of variables as a starting point, a robust algorithm for model selection was proposed in Ronchetti, Field, and Blanchard (1997). Since Shao's techniques are based on least squares, they are sensitive to outliers. The authors developed their robust procedure using the same ideas of cross-validation as Shao but use estimators that are of optimal bounded influence for prediction. They demonstrated the effectiveness of their robust procedure in providing protection against outliers both in a simulation study and in a real example and contrasted the results with those obtained by Shao's method, demonstrating a substantial improvement in choosing the correct model in the presence of outliers with little loss of efficiency at the normal model.

A robust version of the Schwartz criterion was proposed in Machado (1993). It was shown in his paper that under some assumptions, the smallest true model would be selected with probability approaching one as $n \to \infty$. Consider the model (2.2). In Burman and Nolad (1995), an M-estimation-based Akaike-type criterion was presented, where the penalty term is the product of the number of free parameters and the estimate of C_ρ. The C_ρ is given by

$$C_\rho = \frac{\sum\limits_{i=1}^{n} \mathrm{var}[\rho_1(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0)]}{\sum\limits_{i=1}^{n} R_2[\mathrm{E}(Y_i) - \boldsymbol{x}_i'\boldsymbol{\beta}_0]},$$

where ρ is a convex discrepancy function with a unique minimum at 0, and twice differentiable in expectation, ρ_1 is the derivative of ρ, R_2 is the second derivative of $\mathrm{E}[\rho(\varepsilon + t)]$, and $\boldsymbol{\beta}_0$ is the minimizer of $\sum_{i=1}^{n} \mathrm{E}\rho(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$. Many examples were given in the paper for the applications of the criterion. Based on the newly developed theory of stochastic complexity (Qian and Künsch 1998) in linear regression, Qian and Künsch (1999) proposed a model selection procedure. This criterion is itself a model selection procedure based on M-estimation, where the discrepancy function for the M-estimation was the Huber's function defined as

$$\rho_c(t) = \begin{cases} \frac{1}{2}t^2, & |t| < c, \\ c|t| - \frac{1}{2}c^2, & |t| \geq c. \end{cases}$$

Under some conditions, the criterion was shown to be strongly consistent in the paper. An extensive simulation study, which compares their method with several robust model selection procedures, can also be found there. By approximating the expected discrepancy functions, a criterion was suggested in general in Shi and Tsai (1998b). For linear regression, the authors proposed three criteria based on M-estimation, called AICR*, AICcR* and AICcR, respectively, where AICcR* was obtained by following the similar approaches as in the derivations of AICc. Consider the model (2.2). Define

$$R_n(\kappa) = \sum_{i=1}^{n} \rho(y_i - x_i'(\kappa)\hat{\beta}_\kappa) + q(k)C_n, \tag{7.1}$$

where $\hat{\beta}_\kappa$ is the M-estimator for the model M_κ corresponding to the discrepancy ρ, i.e.

$$\sum_{i=1}^{n} \rho(y_i - x_i'(\kappa)\hat{\beta}_\kappa) = \min_{\beta(\kappa)} \sum_{i=1}^{n} \rho(y_i - x_i'(\kappa)\beta(\kappa)), \tag{7.2}$$

and $q(k)$ is a strictly increasing function of k and C_n is a function of n. It can be seen that in (7.1), the first term is a generalization of a minimum negative log likelihood function and the second term is the penalty on the use of models involving more parameters. Wu and Zen (1999) introduced the selection criterion called Criterion R based on $R_n(\kappa)$ under which M_{κ^*} is selected such that

$$R_n(\kappa^*) = \min_{\kappa} R_n(\kappa), \tag{7.3}$$

where C_n is such that $n^{-1}C_n \to 0$ and $(\log\log n)^{-1}C_n \to \infty$ as $n \to \infty$. This criterion includes many classical model selection criteria as special cases, and it is shown to be strongly consistent in their paper.

A general form of M-estimation was proposed in Bai and Wu (1997), where the discrepancy functions may be convex functions or differences of convex functions. The model covered all linear and nonlinear regression models, AR time series, EIVR models, etc. as its special cases. Hence it is worth while to construct a model selection criterion based on it.

It is also of interest to examine methods to assess influence in model selection problems. Léger and Altman (1993) examined the use of "leave-one-out" measure of changes in predicted values to assess influence of individual observations in model building. They suggested this measure considering multicollinearity among the independent variables. It seems to us that other measures can also be proposed and studied.

8 Order selection in time series

Consider the following autoregressive model of order p:

$$x_t - \Phi_1 x_{t-1} - \cdots - \Phi_p x_{t-p} = z_t, \tag{8.1}$$

where $\Phi_1, \dots, \Phi_p$ are real $m \times m$ matrices, and $\{z_t\}$ is the error process.

The autoregressive models are the popular models for time series data. In previous sections, we have discussed many model selection methods and some of them, e.g., FPE, $(\mathrm{FPE})^\lambda$, FPE_α and HQ, were originally derived for autoregressive models. In the analysis of autoregressive models, it is of interest to know the order of the optimal autoregressive model. Hence, a criterion is needed to fulfill this task. Generally speaking, most model selection methods, which work for linear regression, apply to the selection of the order of an autoregressive model. In the framework of stationary autoregressive models, Hannan and Quinn (1979) proved the strong consistency of the order estimators obtained by HQ for the univariate case, and Quinn (1980) obtained a similar result for the multivariate case. For the nonstationary autoregressive models with independently and identically distributed errors, weak consistency of the order estimators was established independently by Paulsen (1984) and Tsay (1984). Paulsen (1984) also discussed the multivariate case. The nonstationarity considered in both papers arises from the fact that the characteristic polynomial is allowed to have roots not only outside but also on the unit circle. Paulsen and Tjøstheim (1985) also discussed the case of nonstationarity where autoregressive scheme is stable but the error process is allowed to have a nonconstant variance. Pötscher (1989) gave strong consistency results for order estimation in nonstationary autoregressive models under the assumptions weaker than those employed in Paulsen (1984), Tsay (1984) and Paulsen and Tjøstheim (1985). He assumed the error process to be a martingale difference with respect to $\{\mathcal{F}_t\}$, where $\mathcal{F}_t$ is the σ-algebra generated by $\{x_s,\ s \le t\}$ in the model (8.1) with $m = 1$.

Using asymptotic efficiency as the criterion, Hurvich and Tsai (1989, 1993) studied the order estimation in the autoregression models without assuming the bound of the possible orders. They proposed a bias-corrected version of AIC (AICc), which works well in this case. The correction is of particular use when the sample size is small, or when the number of fitted parameters constitutes a large fraction of the sample size. The corrected method is asymptotically efficient if the true model is infinite dimensional. Furthermore, when the true model is of finite dimension, the method is found to provide better model order choices than any other asymptotically efficient method. Applications to nonstationary autoregressive and mixed autoregressive moving average models are also discussed there. Hurvich, Shumway, and Tsai (1990) suggested another order estimator which provides somewhat better model selection than the previous one if none of the candidate model dimensions exceeds one-half the sample size and provides a much better model selection than the previous one if some of the candidate models have large dimension and the sample size is small, when the autoregression models are estimated by maximum likelihood.

An order selection procedure based on the subsampling method can be found in Fukuchi (1999). He proposed to select a time series model empirically from a set of possibly nonnested and misspecified models by using estimated risk of prediction as a selection criterion. The author argued that compared with information theoretic criteria, his procedure was free of the problem of penalty selection. However the choice of subsample size will affect the performance of the procedure. According to Fukuchi (1999), the method of subset selection of stochastic regressors based on cross validation by Yao and Tong (1994) seems to be extendable to the model selection problem considered in the paper.

Suppose the goal is to make long range prediction, e.g., h-step forecasts, where we need to predict x_{n+h} from the time series $x_1, \ldots, x_n$. A simple method can be given based on the "plug-in" method (see Box and Jenkins 1970), in which an initial k-th order autoregression is chosen with k itself selected by an order selection criterion, and the multistep forecasts are obtained from this initial model fitted to the time series by repeatedly iterating the model and replacing the unknown future values by their own forecasts. Whittle (1963) observed that the plug-in method is optimal in a least-squares sense if the fitted model coincides with that generating the time series, or in a somewhat restricted sense, for prediction only one step ahead. Since all fitted models may be incorrect in practice, this observation suggests that for multistep prediction the plug-in method may not work well. A different approach may be desirable.

There has been much interest recently in the question of using lead-time (h) dependent estimates or model for multistep prediction of a time series. It is easy to see that such study involves solving an order selection problem, which is essential for forecasting. Earlier references advocating lead-time dependent model selection and/or parameter estimation for multistep forecasting include Findley (1983), Tiao and Xu (1993) and Lin and Granger (1994). In Bhansali (1996), a direct procedure involving a linear least-squares regression of x_{t+h} on $x_t, \ldots, x_{t-k+1}$ was used for estimating the prediction constants, with $k = \tilde{k}_h$, say, treated as a random variable whose value is selected anew for each h by an order selection criterion. He showed that the order selection by suitable h-step generalizations of the AIC and FPE criteria or their equivalents are asymptotically efficient for h-step prediction as the bound is attained in the limit if $\tilde{k}_h$ is selected by any of these criteria. The comparison between the plug-in method and the direct procedure can be found in Bhansali (1997).

In Hurvich and Tsai (1997), a version of the corrected AIC (AICc) was developed for the selection of an h-step-ahead linear predictor for a weakly stationary time series in discrete time. A motivation for this criterion was provided in terms of a generalized Kullback-Leibler distance which is minimized at the optimal h-step predictor, and which is equivalent to the ordinary Kullback-Leibler distance when $h = 1$. In their simulation study, it was found that if the sample size is small and the predictor coefficients are estimated by Burg's method (Burg 1978), AICc typically outperforms both the ordinary AIC and FPE for h-step prediction. Note that Chen, Davis, Brockwell, and Bai (1993) presented simulation results that for a finite order autoregressive process, the Burg estimator can perform very poorly in small samples if the model order used in the estimator greatly exceeds the true order: they found that the Yule-Walker estimator performs much better at these high orders. Hurvich and Tsai (1996) argued that the reason for them to use Burg's estimator is that if the AICc is used it is rare that a large model order is selected. They presented evidence to indicate that Burg estimation can produce much better selected predictors than Yule-Walker estimation.

Liu (1996) investigated the simultaneous multistep forecasts. First, a univariate autoregressive model is translated into a constrained multivariate regression model. Based on this transformation, it was shown that the model selection procedures derived from one-step ahead forecasts also keeps some optimality in the sense of multistep forecasts. The author obtained the multistep versions of BIC, FIC, and C_p.

Now consider the following autoregressive moving average (ARMA) model:

$$x_t - \Phi_1 x_{t-1} - \cdots - \Phi_p x_{t-p} = z_t + \Psi_1 z_{t-1} + \cdots + \Psi_q z_{t-q}, \qquad (8.2)$$

where $\Phi_1, \ldots, \Phi_p$ and $\Psi_1, \ldots, \Psi_q$ are real $m \times m$ matrices, and $\{z_t\}$ is the error process. The orders of this model are p and q.

The monograph by Choi (1992) gives a comprehensive survey and an extensive bibliography on the ARMA model identification. As pointed out by Choi, the FPE, AIC, BIC, SIC, HQ, MDL, and PLS and similar procedures can be used to select the orders of the ARMA models. Lai and Lee (1997) expanded the list and they extended Rissanen's accumulated prediction error criterion and Wei's FIC from linear to general stochastic regression models, which includes ARMA models as its special case. They showed that these criteria are strongly consistent under certain conditions.

Zhang and Wang (1994) proposed the order determination quantity (ODQ) as a new way to solve order estimation problems for the model (8.2) with $m = 1$. The ODQ is defined as

$$\mathrm{ODQ}_n(p,q) = n\hat{\sigma}_n^2(p,q) - n\hat{\sigma}_n^2(p^*,q^*) - a_n$$

where $\hat{\sigma}^2(\cdot,\cdot)$ denotes an estimate of the common variance of the noise sequence; $0 < p \le p^*$, $0 < q \le q^*$; (p^*, q^*) is an upper bound of the unknown true order (p_0, q_0), which can be arbitrarily large but fixed and is supposed to be known *a priori*; n is the sample size; and $a_n > 0$ is a data-dependent constant such that $a_n/n \to 0$ and $a_n/(\log n)^\gamma \to \infty$ almost surely, where $\gamma = 1$ for pure autoregressive models and $\gamma \ge 1$ is a nonrandom constant to be specified in the future for general ARMA models. $(\hat{p}, \hat{q})$ is determined to satisfy

$$\mathrm{ODQ}_n(\hat{p}-1, \hat{q}) > 0, \;\; \mathrm{ODQ}_n(\hat{p}, \hat{q}-1) > 0 \text{ and } \mathrm{ODQ}_n(\hat{p}, \hat{q}) < 0,$$

instead of by minimizing $\mathrm{ODQ}(p, q)$ over (p, q). The authors argued that theoretical analysis and simulation showed that the ODQ has higher identifiability for unknown true orders, provides clear separation points and requires less computational effort than the order estimation criteria such as AIC, BIC, HQ, PLS, etc. Under certain conditions, it was shown that ODQ is strongly consistent for unstable autoregressive processes.

Note that if an ARMA model is invertible, it can be approximated by an autoregressive model of order m for large m. If the set of candidate models consists of autoregressive models while the true model is an ARMA model, an efficient order selection criterion is recommended against a consistent order selection criterion. A simulation study for the case that the true model is a moving average model while the set of candidate models consists of autoregressive models can be found in McQuarrie and Tsai (1998).

The following nonlinear time series model was studied in Chen, McCulloch, and Tsay (1997):

$$\begin{aligned} x_t &= f(x_{t-1}, \cdots, x_{t-p}; a_{t-1}, \ldots, a_{t-q}; \boldsymbol{\beta}_f) + a_t, \\ a_t &= g_t z_t, \\ g_t &= g(x_{t-1}, \ldots, x_{t-u}; a_{t-1}, \ldots, a_{t-v}; g_{t-1}, \ldots, g_{t-w}; \boldsymbol{\beta}_g), \end{aligned} \tag{8.3}$$

where x_t is a univariate time series, $f(\cdot)$ and $g(\cdot)$ are two known functions with finite dimensional parameter vectors $\boldsymbol{\beta}_f$ and $\boldsymbol{\beta}_g$, respectively, p, q, u, v, and w are non-negative integers, and $\{z_i\}$ is a sequence of independent and identically distributed random variables with mean zero and variance one. The function $g(\cdot)$ is assumed to be positive; it governs the evolution of the volatility of the innovational series a_t. Examples of the model so defined were presented there. In that paper, they claimed that there was little discussion of model selection across different classes of nonlinear models and that much work on model selection in the literature focuses on nested models for which the traditional maximum likelihood ratio tests or Rao's score tests or information criterion functions apply. It is easy to see that for non-nested models, model discrimination becomes much more involved, especially when the competing models are nonlinear. Li (1993) adopted the idea of separate families of hypotheses of Cox (1962) and proposed a test statistic for discriminating between bilinear and threshold models. Chen, McCulloch, Tsay (1997) argued that Li's test was closely related to the method of selecting a model with smaller residual variance and was not applicable to other nonlinear models. They proposed an approach that is, as they asserted, widely applicable in univariate time series analysis for linear or nonlinear models, and can discriminate between non-nested nonlinear models.

Their approach is based on Gibbs sampling and in particular they treated starting values of the time series and the innovational series as parameters and considered the conditional likelihood function of a parameter given the others. The approach also requires some prior specification, which is the probability that an individual observation is generated by a specified model given that both observations adjacent in time are generated by the same model. The drawback is that it may take substantial computing time in some applications.

The order determination criteria can also be used to test the hypotheses of white noise model against autoregressive models (see Pukkila and Krishinaiah 1988a, b) for details). Based on this idea, a procedure for identifying ARMA models was proposed in Pukkila, Koreisha, and Kallinen (1990).

The selection of a model when the candidate models are some general stochastic models needs some investigation. It is certain that the model selection methods, e.g., AIC, BIC and HQ, can be adopted to this case, but their performances may not be satisfactory. Better selection procedures need to be explored.

9 Model selection in categorical data analysis

Data collected in the social sciences for measuring attitudes and opinions on various issues and demographic characteristics such as gender, race, and social class and in biomedical sciences to measure such factors as severity of an injury, degree of recovery from surgery, and stage of a disease are categorical. Categorical data also arises in other sciences.

For categorical data, the problems such as checking independence of attributes, selection of optimal explanatory variables, and selection of an optimal categorization are of special interest. Appropriate procedures to solve these problems by AIC can be found in Sakamoto (1991) among others. Sakamoto (1991) also proposed ABIC, an extension of AIC, for evaluating Baysian binary regression models.

Generalized linear models were introduced by Nelder and Wedderburn (1972). This family contains important models for categorical data such as logit and probit models for quantal responses, loglinear models and multinomial response models for counts, as well as linear regression and analysis of variance models for continuous response variables. A generalized linear model is specified by three components: a random component, a systematic component, and a link. For such a model, the deviance or Pearson's chi-square goodness-of-fit statistic may be used to measure the fit.

When a generalized linear model selection problem is about selecting optimal explanatory variables, it is not hard to see that the model selection methods discussed in Sections 2-4 can adjust themselves to serve generalized linear model selection problems (e.g., Agresti 1990, Bai, Krishnaiah, Sambamoorthi, and Zhao 1992, Christensen 1997, Hosmer, Jovanovic, and Lemeshow 1989, McCullagh and Nelder 1989, Pregibon 1979). For example, consider a set of loglinear models. In this case, stepwise procedures can be performed by starting with an initial model and then using rules for adding or deleting terms to arrive at a final model. Note that a model selection procedure may be improved or modified to adapt to the situation. As an example, the backward elimination is modified for controlling the experimentwise error rate (Aitkin 1978, 1979).

The generalized linear models usually do not include a dispersion parameter. McCullagh and Nelder (1989) suggested that it is often wise to assume that a dispersion parameter is present in the model unless the data or prior information indicate otherwise. Hurvich and Tsai (1995) generalized the AICc to an extended quasi-likelihood model, which includes the generalized linear model with a dispersion parameter as a special case.

Qian, Gabor, and Gupta (1996) considered the problem of selecting a model with the best predictive ability in a class of generalized linear models. A predictive least quasi-deviance criterion was proposed to measure the predictive ability of a model. This criterion is obtained by applying the idea of the predictive minimum description length principle and the theory of quasi-likelihood functions. The resulting predictive quasi-deviance function is an extension of the predictive stochastic complexity of the model. Under rather weak conditions the authors showed that the predictive least quasi-deviance method is consistent. Also, the authors showed that the selected model converges to the optimal model in expectation. The method was then modified for finite sample applications. Examples and simulation results were presented in the paper. There is still much work to be done in this direction.

Random effect models are useful for explanations of overdispersion, correlation and subject-specific inference. Hence generalized linear models with random effects are very desirable in practice. The choice of a model in such cases needs some study.

10 Model selection in nonparametric regression

In nonparametric regression, local polynomial, kernel and smoothing spline methods among others have been used to construct nonparametric estimates of smooth regression functions (see, e.g., Fan and Gijbels 1996 and Simonoff 1996). These estimators use a smoothing parameter to control the amount of smoothing performed on a given data set, where the parameter is chosen using a selection criterion. Many methods have been proposed for selecting the parameter.

As commented in Loader (1999), the methods for bandwidth selection for procedures such as kernel density estimation and local regression can be divided into two broad classes. One of the classes includes classical methods such as cross-validation, Mallows' C_p, AIC, etc. while the other class contains plug-in methods. In a plug-in method, the bias of an estimate $\hat{f}$ is written as a function of the unknown f, and usually approximated through Taylor series expansions, and a pilot estimate of f is then "plugged in" to derive an estimate of the bias and hence an estimate of the mean integrated squared error, and then the optimal bandwidth minimizes this estimated measure of fit. According to Loader (1999), substantial "evidence" has been collected to establish superior performance of modern plug-in methods in comparison to methods such as cross validation; this has ranged from detailed analysis of rates of convergence, to simulations, to superior performance on real datasets. Loader (1999) took a detailed look at some of this evidence, looking into the sources of differences. He argued that his findings challenge the claimed superiority of plug-in methods on several fronts. First, plug-in methods are heavily dependent on arbitrary specification of pilot bandwidths and fail when this specification is wrong. Second, the often-quoted variability and undersmoothing of cross validation simply reflects the uncertainty of bandwidth selection; plug-in methods reflect this uncertainty by oversmoothing and missing important features in complicated situations. Third, in terms of the asymptotic theory, plug-in methods use available curvature information in an inefficient manner, resulting in inefficient estimates. Asymptotically, the plug-in based estimates are beaten by their own pilot estimates.

Recently an interesting approach for selecting the smoothing parameters of nonparametric regression estimators was proposed in Hart and Yi (1998). Their method was based on one-sided cross-validation instead of ordinary cross-validation. The authors argued that by using one-sided cross-validation their method retains the nature of ordinary cross-validation and has much better statistical properties. It was shown that statistical properties of their method are comparable to those of a plug-in methods.

However, due to great variability and a tendency to under smooth, the "classical" criteria, such as generalized cross-validation and AIC are not ideal for selecting the smoothing parameter. Hurvich, Simonoff, and Tsai (1998) addressed these problems by proposing a nonparametric version of their AICc criterion. The authors argued that AICc, unlike plug-in methods, can be used to choose smoothing parameters for any linear smoother, including local quadratic and smoothing spline estimators, and AICc is competitive with plug-in methods for choosing smoothing parameters, and also performs well when a plug-in approach fails or is unavailable. Since in some applications neither parametric nor nonparametric estimation may give a reasonable fit to the data, Shi and Tsai (1999) and Simonoff and Tasi (1999) obtained AICc for semiparametric regression models.

Consider the selection of a hard wavelet threshold for recovery of a signal embedded in additive Gaussian white noise, a problem closely related to that of selecting a subset model in orthogonal normal linear regression. The existing approaches, such as AIC, Donoho and Johnstone's universal method (Donoho and Johnstone 1994), Nason's cross-validatory method (Nason 1996), etc. were presented in McQuarrie and Tsai (1998). A computationally efficient algorithm for implementing Nason's method can also be found there. Hurvich and Tsai (1998) proposed a data-dependent method of hard threshold selection based on a cross-validatory version of AICc, which, like universal thresholding and Nason's method, can be implemented in $O(n \log n)$ operations (where n is the sample size). The simulation results presented in McQuarrie and Tsai (1998) showed that both of the cross-validatory methods outperform universal thresholding.

As another approach for using wavelet decompositions to select a regression model, Antoniadis, Gijbels, and Grégoire (1997) suggested the determination of the number of nonzero coefficients in the vector of wavelet coefficients based on the idea of MDL. They pointed out that the class of functions tested by their criterion allowed them to approximate quite efficiently alternatives composed by complicated functions with inhomogeneous smoothness.

In the theory of linear models, the concept of degrees of freedom plays an important role. This concept is often used for measurement of model complexity, for obtaining an unbiased estimate of the error variance, and for comparison of different models. A concept of generalized degrees of freedom (GDF) that is applicable to complex modelling procedures was developed in Ye (1998). The definition is based on the sum of the sensitivities of each fitted value to perturbation in the corresponding observed value. The concept is nonasymptotic in nature and does not require analytic knowledge of the modelling procedures. The concept of GDF offers a unified framework under which complex and highly irregular modelling procedures can be analyzed in the same way as classical linear models. Besides, there is an interesting connection between the GDF and the half-normal plot.

Consider a response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$,

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2 I),$$

where σ^2 is assumed to be known and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ is an $n \times 1$ mean vector. Define a modelling procedure $\mathcal{M}$ as a mapping from R^n to R^n that produces a set of fitted values $\hat{\boldsymbol{\mu}} \equiv \hat{\boldsymbol{\mu}}(\boldsymbol{Y})$ from $\boldsymbol{Y}$. Note that $\boldsymbol{\mu}$ often depends on some observed covariates, and so does the modelling procedure $\mathcal{M}$. The GDF for modelling $\mathcal{M}$ are given by $D(\mathcal{M}) = \sum_{i=1}^{n} h_i^{\mathcal{M}}(\boldsymbol{\mu})$, where

$$h_i^{\mathcal{M}}(\boldsymbol{\mu}) = \frac{\partial \mathrm{E}_{\boldsymbol{\mu}}[\hat{\mu}_i(\boldsymbol{Y})]}{\partial \mu_i}.$$

In classical linear models, the GDF reduces to the standard degrees of freedom. If $\mathcal{M}$ is a linear smoother, then $D(\mathcal{M})$ reduces to the trace of the smoothing matrix. Efron (1986) obtained the concept of "expected optimism" by using the average of the covariance form of Y_i and $\hat{\mu}_i$. Ye (1998) argued that the covariance form is less intuitive and more difficult to analyze and estimate. The author pointed out that $\mathrm{E}_{\boldsymbol{\mu}}[\hat{\mu}_i(\boldsymbol{Y})]$ is an infinitely differentiable function of $\hat{\boldsymbol{\mu}}$, which has three implications: 1. GDF is defined even when $\hat{\mu}_i$ is highly irregular or even discontinuous. 2. Because $h_i^{\mathcal{M}}(\boldsymbol{\mu})$ is also infinitely differentiable, it can be estimated with its value $h_i^{\mathcal{M}}(\boldsymbol{Y})$. 3. Because $\mathrm{E}_{\boldsymbol{\mu}}[\hat{\mu}_i(\boldsymbol{Y})]$ can be viewed as a smoothing of the fitted value $\hat{\mu}_i$, what is important is the global behavior of $\hat{\mu}_i$, not the local behavior. For estimating $D(\mathcal{M})$, an algorithm, which is based on Monte Carlo method, was provided in Ye (1998). The algorithm is given as follows:

1. Repeat $t = 1, \dots, T$;
2. Generate $\boldsymbol{\Delta}_t = (\delta_{t1}, \dots, \delta_{tn})$ for the density $\prod(1/\tau^n)\phi(\delta_{ti}/\tau)$;
3. Evaluate $\hat{\mu}_i(\boldsymbol{Y} + \boldsymbol{\Delta}_t)$ based on the modelling procedure $\mathcal{M}$;
4. Calculate $\hat{h}_i$ as the regression slope from

$$\hat{\mu}_i(\boldsymbol{Y} + \boldsymbol{\Delta}_t) = \alpha + \hat{h}_i \delta_{ti}, \quad t = 1, \dots, T,$$

$D(\mathcal{M})$ is estimated by $\hat{D}(\mathcal{M}) = \sum_i \hat{h}_i$. The parameter T determines the number of perturbations. It was suggested that $T \geq n$ in Ye (1998). $\hat{D}(\mathcal{M})$ depends on a tuning parameter τ, which the author referred to as the perturbation size. According to Ye (1998), it can be shown that $\hat{D}(\mathcal{M}) \to D(\mathcal{M})$ if $\tau \to 0$. This approach is similar to the little bootstrap of Brieman (1992).

Based on Ye (1998), the concept of GDF allows complex modelling procedures to be analyzed in a way similar to the analysis of classical linear models, and is independent of the sample size constraint and the complexity of the modelling procedure. Let $\{\mathcal{M}_\gamma,\ \gamma \in \Gamma\}$ be a set of modelling procedures, where Γ is an index set. He proposed two model selection procedures:

1. Let

$$A_e(\mathcal{M}_\gamma) = \|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_\gamma\|^2 - n\sigma^2 + 2D(\mathcal{M})\sigma^2,$$

and minimize $A_e(\mathcal{M}_\gamma)$ with respect to $\mathcal{M}_\gamma$. This is called the extended AIC (EAIC).

2. Let

$$\mathrm{YGCV}(\mathcal{M}_\gamma) = \|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_\gamma\|^2/(n - D(\mathcal{M}))^2,$$

and minimize $\mathrm{YGCV}(\mathcal{M}_\gamma)$ with respect to $\mathcal{M}_\gamma$. This is the analog of GCV in Graven and Wahba (1979). It is referred to as YGCV. The advantage of this criterion is that it does not assume a known σ^2.

When σ^2 is unknown, an estimate of it may be obtained by

$$s^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\|^2/(n - D(\mathcal{M})).$$

Based on Ye (1998), EAIC and YGCV can be used to compare the performance of a nonparametric regression to a linear model as a way of diagnosing the adequacy of the linear model, or of selecting the most suitable nonparametric regression procedures among various alternatives, such as classification and regression trees, projection pursuit regression, and artificial neural networks. The criteria can also be used for selection of variables in nonparametric regression settings (Bickel and Zhang 1992 and Zhang 1991) by treating variable selection as a special case of selecting modelling procedures.

11 Data-oriented penalty

Let a model selection criterion be the sum of two components, where the first component measures the model fit and the other component is used for penalizing the model complexity. Among them, many have a data-independent penalty fixed or as an unbounded function of the sample size. In the first case, a procedure tends to overfit and yields less prediction error and in the later case, a procedure is quite stable with the penalty falling in some interval varied with the sample size. The choice of the penalty will affect the performance of such a model selection criterion. A theoretical study on the effect of a penalty of a classical linear model selection criterion can be found in Shao (1997). Hence, it is needed to find a good data-oriented penalty so that a procedure with its use will perform well. Such efforts can be found in Rao and Wu (1989), Wei (1992) and Shao (1998) among others.

Recently, a promising model selection procedure has been proposed by J. S. Rao-Tibshirani by searching for an optimal penalty in terms of minimizing an objective function via methods such as the cross-validation (see J. S. Rao 1994). Bai, Rao, and Wu (1999) proposed a method of constructing a data-oriented penalty. For selecting a model from candidate models given in (2.2), consider the model selection criterion $G_n^{(1)}$ defined in (4.6). A data-oriented procedure to select the penalty C_n proposed in Bai, Rao, and Wu (1999) is as follows:

1. Compute any consistent estimate $\tilde{\beta}_n = (\tilde{\beta}_{1n}, \dots, \tilde{\beta}_{pn})'$ of β and the mean squared error $\tilde{\sigma}_p^2$. For instance, $\tilde{\beta}_n$ can be chosen to be the least square estimate of β.

2. Compute $\hat{\varepsilon}_n = \boldsymbol{y}_n - X_n\tilde{\beta}_n$.

3. Let $\overline{\beta}_n = (\overline{\beta}_{1n}, \dots, \overline{\beta}_{pn})'$ be defined as follows:

$$\overline{\beta}_{in} = \begin{cases} \tilde{\beta}_{in}, & \text{if } |\tilde{\beta}_{in}| \geq \eta, \\ \\ \eta \text{sign}(\tilde{\beta}_{in}), & \text{if } |\tilde{\beta}_{in}| < \eta, \end{cases} \qquad \text{for } i = 1, \dots, p,$$

   where the constant η is a suitable chosen threshold value.

4. Let

$$\boldsymbol{u}_n(h) = X_n(h)\overline{\beta}_n(h) + \hat{\varepsilon}_n, \quad h = 1, \dots, p.$$

   Compute

$$D_n(q,h) = \overline{S}_q(h) - \overline{S}_h(h), \quad q = 0, 1, \dots, p$$

   where $\overline{S}_q(h) = (\boldsymbol{u}_n(h))'(I - P_q)\boldsymbol{u}_n(h)$. It can be shown that $\overline{S}_p(h) = S_p$ if $\overline{\beta}_n = \tilde{\beta}_n$ Define

$$\Delta_h = \min_{q<h} \left\{ \frac{D_n(q,h)}{(h-q)\tilde{\sigma}_p^2} \right\}.$$

5. Define

$$C_n^{(R)} = \frac{\text{average of } \{\Delta_h,\ h = 1, \dots, p\}}{1 + \sqrt{\lfloor 0.01n \rfloor}},$$

   where $\lfloor b \rfloor$ denotes the integer part of b.

Then choose $C_n^{(R)}$ as the penalty C_n in $G_n^{(1)}$. It was shown in Bai et al. (1992) that $C_n^{(R)}$ asymptotically satisfies the conditions given in (4.7), while it works well for small to moderate sample sizes. Similar results are true for the other two criteria $G_n^{(\ell)}$, $\ell = 2, 3$, defined in (4.6). Based on the same idea, Wu (2001) proposed a data-oriented penalty for Criterion R (7.3).

Shao and J. S. Rao (2000) combined hypothesis testing and model selection together and proposed a particular choice of the penalty parameter C_n for the model selection criterion $G_n^{(1)}$ defined in (4.6). The authors demonstrated that the new model selection procedure inherits good properties from both approaches, i.e., its overfitting and underfitting probabilities converge to 0 as $n \to \infty$, and when n is fixed, its overfitting probability is controlled to be approximately under a pre-assigned level of significance.

Consider the candidate models of form (2.2). Let $\mathcal{M}_c$ denote the set of true models in $\{M_\kappa\} = \mathcal{M}$ and M_{κ_0} denote the optimal model, which is the model in $\mathcal{M}_c$ with the smallest dimension. If M_{κ^*} be the model selected by the model selection criterion $G_n^{(1)}$, $p_1 = P(M_{\kappa^*} \in \mathcal{M}_c)$ is the error probability of overfitting and $p_2 = P(M_{\kappa^*} \in \mathcal{M} - \mathcal{M}_c)$ is the error probability of underfitting, which depend on C_n. Let δ be a pre-assigned level of significance. Shao and J. S. Rao (2000) proposed the choice of C_n so that $p_1 \leq \min\{\delta,\ 1/\sqrt{n}\}$ (approximately) holds and otherwise as small as possible to minimize p_2. Simulation studies were presented in the paper. A real data analysis can also be found there.

There are still many interesting open problems. For example, how can one find a good data-oriented penalty if the number of available models is infinite? Another problem is how to find a good data-oriented penalty when the true model is not included in a given class of models. Certainly, we can still proceed as if the data were from one of the available models. Then what is the consequence of such a procedure? Since the data are also used to choose the penalty, what is its impact on the inferences performed on such selected model?

12 Statistical analysis after model selection

Once a model is selected, it is usual that statistical analysis is carried out based on the selected model and often the analysis will be done as if the selected model is the true model. It is easy to see that any statistical analysis based on the selected model is affected by the nature of the true model, candidate models and the model selection procedure.

It is already well known that the model selection procedure can severely affect the validity of standard regression procedures. Rencher and Pun (1980) demonstrated that a model selected by the best subset regression method tends to have an inflated value of R^2. Miller (1990) showed that, if one starts with a model selected from the data, then regression estimators may be biased and standard hypothesis tests may not be valid. Breiman (1992) has shown that models selected by classical data-driven methods can produce strongly biased estimates of mean squared prediction error, while both little bootstrap and cross-validation can produce relatively unbiased estimates of mean squared prediction error for data-selected submodels. Reviews of some of the difficulties induced by variable selection can be found in Bankroft and Han (1977), Burnham and Anderson (1998), Chatfield (1995), Cohen and Sackrowitz (1987) and Miller (1990) among others.

Cohen and Sackrowitz (1987) studied the problem of inference following a model selection and proposed a decision-theory approach for it, where the loss function includes components for model selection as well as for inference and allows for flexibility in emphasis on one or the other. Assuming that the data are normally distributed, Hurvich and Tsai (1990) presented Monte Carlo results on the coverage rates of confidence regions for the regression parameters, conditional on the selected model order. Their findings showed that conditional coverage rates are much smaller than the nominal coverage rates, obtained by assuming that the model is known in advance. In a very general context, Pötscher (1991) established the asymptotic properties of estimators of parameters based on a statistical model which has been selected via a model selection procedure. The asymptotic distributions of the estimators are obtained and the effects of the model selection process are illustrated numerically using the example of a distributed lag model. An important potential application of such results is to the generation of confidence regions for the parameters of interest. Kabaila (1995) demonstrated that a great deal of care must be exercised in any attempt at such an application. The author examined the effect of model selection procedures on confidence and prediction regions and emphasized that consistent estimation of the order of the model need not necessarily lead to confidence and prediction regions with optimal properties. As Pötscher (1991) noted, the asymptotic properties are of little value unless they hold for realized sample sizes. Pötscher and Novák (1998), by using simulation, studied the small sample distribution of estimators of parameters based on a statistical model which has been selected via a model selection procedure, and, in particular, evaluated the accuracy of the approximation provided by the asymptotic distribution in small samples. Zhang (1992) gave related results for the linear regression model, mainly concentrating on first and second moment properties of the estimators. Note that in Pötscher (1991), model selection procedures under investigation were testing procedures for variable selection while in Zhang (1992), FPE_α was used instead. In Pötscher and Novák (1998), both procedures were considered.

Based on Ye (1998), EAIC and YGCV can be used to evaluate the effect induced by various selection procedures, such as variable selection in linear models, and bandwidth selection in nonparametric regression. The author applied the proposed framework to measure the effect of variable selection in linear models, leading to corrections of selection bias in various goodness-of-fit statistics.

It is often that in statistical practice, the same data are used for formulating, fitting and checking a model, which may lead to inaccurate summaries and overconfident decisions. A Bayesian view of the problem can be found in Draper (1995). By virtue of recent computational advances, he discussed a Baysian approach to solving this problem and examined its implementation in some examples. But the approach introduced another problem of how to choose a prior, which is associated with Bayesian approaches in general.

The problem is actually very profound. Whether or not an answer is satisfactory depends on one's belief. Some people believe in the existence of a true model and others regard it as a fiction. The related discussion can be found in Chatfield (1995, 1996), and Draper (1995) among others.

Based on the above investigation, it can be seen that extensive research on the impact of model selection on statistical analysis is in great demand. Much work needs to be done on this direction.

13 Conclusions

In this paper, numerous model selection procedures are discussed. They are developed based on hypothesis testing, prediction errors, information measurement, MDL, resampling methods, Bayesian approach, etc.. Besides, the impact of carrying out statistical analysis after model selection is surveyed.

As demonstrated in this paper, the research on model selection is of great importance from both theoretical and practical points of view. It is not hard to see that the area of model selection is rich in problems, which are waiting to be solved. Some of them are already mentioned in previous sections.

In the end, we wish to emphasize that the model we use to analyze a data set depends on the specific questions to be answered. There are instances where different models may have to be used on the same data to answer different questions. Also, it is a good statistical practice to analyze the data under different possible models to answer a specific question to see how different or robust the answers are.

Finally, the model we select by using any of the methods described in the paper will depend on the sample size. Take for instance a simple regression problem for predicting a response measurement y using a set of p predictor variables x. We may have a sample , $(y_1, x_1), \ldots, (y_n, x_n)$, of n observations and the use of a selection method may indicate that a subset of x may be sufficient to predict y. If we have a larger sample, the same selection method may indicate that a larger subset of x will provide better prediction (see Rao 1987).

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis.* Wiley, New York.

Aitkin, M. A. (1974). Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics* **16** 221-227.

Aitkin, M. A. (1978). The analysis of unbalanced cross-classifications. With discussion. *J. Roy. Statist. Soc. Ser. A* **141**, 195-223.

Aitkin, M. (1979). A simultaneous test procedure for contingency tables. *Appl. Statist.* **28**, 233-242.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16**, 125-127.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21**, 243-247.

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203-217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd Internat. Symp. on Information Theory* (Petrov, B.N. and Czàki, F., eds.) 267–281, Akademiai Kiadò, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control* AC-19, 716-723.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* **30** 9-14.

Akaike, H. (1983). Information measures and model selection. *Bull. Int. Statist. Inst.* **50**, 277-290.

Antoch, J. (1986). Algorithmic development in variable selection procedure. *Proc. of COMPSTAT* 83-90, Physica-Verlag, Heidelberg.

Antoch, J. (1987). Variable selection in linear models based on trimmed least squares estimator. In *Statistical Data Analysis Based on the L_1-norm and Related Methods* (Y. Dodge, ed.) 231-245, North Holland.

Antoniadis, A, Gijbels, I. and Grégoire, G. (1997). Model selection using wavelet decomposition and applications. *Biometrika* **84**, 751-763.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika 67*, 413-418.

Atkinson, A. C. (1981). Likelihood ratios, posterior odds and information criteria. *J. Econometrics* **16**, 15-20.

Bai, Z. D., Krishnaiah, P. R., Sambamoorthi, N. and Zhao, L. C. (1992). Model selection for log-linear models. *Sankhyā Ser. B* **54**, 200-219.

Berger, J. O. and Bernardo, J. M. (1992). On the development of the reference prior method. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 25-44, Oxford Univ. Press.

Berger, J. O. and Pericchi, L. R. (1996a). The intrinsic Bayes factor for model selection and prediction. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 35-60, Oxford Univ. Press.

Berger, J. O. and Pericchi, L. R. (1996b). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91**, 109-122.

Bhansali, R. J. (1996). Asymptotically efficient autoregressive model selection for multistep prediction. *Ann. Inst. Statist. Math.* **48**, 577-602.

Bhansali, R. J. (1997). Direct autoregressive predictors for multistep prediction: Order selection and performance relative to the plug in predictors. *Statist. Sinica* **7**, 425-450.

Bhansali, R. J. and Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika* **64**, 547-551.

Bickel, P. and Zhang, P. (1992). Variable selection in nonparametric regression with categorical covariates. *J. Amer. Statist. Assoc.* **87**, 90-97.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345-370.

Bozdogan, H. (1988). ICOMP: A new model selection criterion. In *Classification and Related Methods of Data Analysis* (Hans H. Bock, ed.) 599-608, Amsterdam, North Holland.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Amer. Statist. Assoc.* **87**, 738-754.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350-2383.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78**, 131-136.

Burg, J. P. (1978). A new analysis technique for time series data. In *Modern Spectrum Analysis* (D. G. Childers, ed.) 42-48, IEEE Press, New York.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-hold cross-validation, and repeated learning-testing methods. *Biometrika* **76**, 503-514.

Burman, P., Nolan, D. (1995). A general Akaike-type criterion for model selection in robust regression. *Biometrika* **82**, 877-886.

Burnham, K. P. and Anderson D. R. (1998). *Model Selection and Inference A practical Information-Theoretic Approach.* Springer-Verlag, New York.

Cavanaugh, J. E. and Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *J. Statist. Plann. Inference* **67**, 45-65.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). *J. Roy. Statist. Soc. Ser. A* **158**, 419-466.

Chatfield, C. (1996). Model uncertainty and forecast accuracy. *J. Forecasting* **15**, 495-508.

Chen, C. H., Davis, R. A., Brockwell, P. J. and Bai, Z. D. (1993). Order determination for autoregressive processes using resampling methods. *Statist. Sinica*, **3**, 481-500.

Chen, C. W. S., McCulloch, R. E. and Tsay, R. S. (1997). A unified approach to estimating and modelling linear and nonlinear time series. *Statist. Sinica*, **7**, 451-472.

Choi, B. S. (1992). *ARMA Model Identification.* Springer-Verlag, New York.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*, 2nd ed. Springer-Verlag, New York.

Clyde, M. A., DeSimone, H. and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *J. Amer. Statist. Assoc.* **91**, 1197-1208.

Cohen, A. and Sackrowitz, H. B. (1987). An approach to inference following model selection with applications to transformation-based and adaptive inference. *J. Amer. Statist. Assoc.* **82**, 1123-1130.

Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1**, 105-123.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. Roy. Statist. Soc. Ser. B* **24**, 406-424.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerical Mathematics* **31**, 377-403.

Davisson, L. D. (1965). The prediction error of stationary Gaussian time series of unknown covariance. *IEEE Trans. Inform. Theory* **11**, 527-532.

De Santis, F. and Spezzaferri, F. (1997). Alternative Bayes factors for model selection.

Canad. J. Statist. **25**, 503-515.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc. B* **57**, 45-97.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316-331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 461-470.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Fan, J. and Li, R. (2001). Variable selection via penalized likelihood. *J. Amer. Statist. Assoc.* To appear.

Findley, D. F. (1983). On the use of multiple models for multi-period forecasting. In *Proc. Bus. Econ. Statist. Sect.* pp. 528-531. Washington, D. C., Amer. Statist. Assoc.

Findley, D. F. (1985). On the unbiasedness property of AIC for exact or approximating linear stochastic time series models. *J. Time Ser. Anal.* **6** 229-252.

Foster, D. and George, E. (1994). the risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947-1975.

Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika* **84**, 707-716.

Fukuchi, J.-I. (1999). Subsampling and model selection in time series analysis. *Biometrika* **86**, 591-604.

Geisser, S. (1971). The inferential use of predictive distributions. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 456-469, Holt, Rinehart and Winston, Toronto.

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

Geisser, S. (1993). *Predictive Inference: an Introduction.* Chapman and Hall, New York.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881-889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7**, 339-373.

Guiasu, S. (1977). *Information Theory with Applications.* McGraw-Hill, New York.

Hampel, F. R. (1983). Some aspects of model choice in robust statistics. *Proc. Int. Statist. Inst., 44th Session*, Book 2, 767-771, Madrid.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41**, 190-195.

Härdle, W. (1987). An effective selection of regression variables when the error distribution is incorrectly specified. *Ann. Inst. Statist. Math.* **39**, 533-548.

Hart, J. D. and Yi, S. (1998). One-sided cross-validation. *J. Amer. Statist. Assoc.* **93**, 620-631.

Herzberg, A. M. and Tsukanov, A. V. (1986) The design of experiments for model selection. *Proc. 1st World Congress Bernoulli Soc.* **2**, 175-178.

Hosmer, D. W., Jovanovic, B. and Lemeshow, S. (1989). Best subsets logistic regression. *Biometrics* **45**, 1265-1270.

Hurvich, C. M., Shumway, R. and Tsai, C. L. (1990). Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. *Biometrika* **77**, 709-719.

Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Stat. Soc. Ser. B* **60**, 271-293.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.

Hurvich, C. M. and Tsai, C. L. (1990). Model selection for least absolute deviations regression in small samples. *Statist. Probab. Lett.* **9**, 259-265.

Hurvich, C. M. and Tsai, C. L. (1993). A corrected Akaike information criterion for vector autoregressive model selection. *J. Time Ser. Anal.* **14**, 271-279.

Hurvich, C. M. and Tsai, C. L. (1995). Model selection for extended quasi-likelihood in small samples. *Biometrics* **51**, 1077-1084.

Hurvich, C. M. and Tsai, C. L. (1996). The impact of unsuspected serial correlations on model selection in linear regression. *Statist. Probab. Lett.* **27**, 115-126.

Hurvich, C. M. and Tsai, C. L. (1997). Selection of a multistep linear predictor for short time series. *Statist. Sinica*, **7**, 395-406.

Hurvich, C. M. and Tsai, C. L. (1998). A crossvalidatory AIC for hard wavelet thresholding in spatially adaptive function estimation. *Biometrika* **85**, 701-710.

Kabaila, P. V. (1995). The effect of model selection on confidence regions and prediction regions. *Econometric Theory* **11**, 537-549.

Kass, R. E. and Raftery, A. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

Konishi, S. (1999). Statistical model evaluation and information criteria multivariate analysis. In *Design of Experiments, and Survey Sampling* (S. Ghosh, ed.), 369-399, Marcel Dekker, Inc.

Konishi, S. and Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika* **83**, 875-890.

Krishnaiah, P. R. (1982). Selection of variables under univariate regression models. In *Handbook of Statistics II* (P. R. Krishnaiah, ed.), 805-820, North-Holland, Amsterdam.

Lai, T. L. and Lee, C. P. (1997). Information and prediction criteria for model selection in stochastic regression and ARMA models. *Statist. Sinica*, **7**, 285-310.

Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *J. Roy. Statist. Soc. Ser. B* **57**, 247-262.

Léger, C. and Altman, N. (1993). Assessing influence in variable selection problems. *J. Amer. Statist. Assoc.* **88**, 547-556.

Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15**, 958-975.

Li, W. K. (1993). A simple one degree of freedom test for non-linear time series model discrimination. *Statist. Sinica* **3**, 245-254.

Lin, J.-L. and Granger, C. W. J. (1994). Forecasting from non-linear models in practice. *J. Forecasting* **13**, 1-9.

Linhart, H. and Zucchini, W. (1986). *Model Selection.* Wiley.

Liu, S. I. (1996). Model selection for multiperiod forecasts. *Biometrika* **83**, 861-873.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Ann. Statist.* **27**, 415-438.

Loh, W. (1985). A new method for testing separate families of hypotheses. *J. Amer. Statist. Assoc.* **80**, 362-368.

Machado, J. A. F. (1993). Robust model selection and M-estimation. *Econometric Theory* **9**, 478-493.

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

Mallows, C. L. (1995). More comments on C_p. *Technometrics* **37**, 362-372.

Mantel, N. (1970). Why stepdown procedures in variable selection? *Technometrics* **12** 621-625.

Martin, R. D. (1980). Robust estimation of autoregressive models. In *Direction in Time Series* (D. R. Brillinger and G. C. Tiao, eds.), 228-262, IMS, Hayward, CA.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

McKay, R. J. (1977). Variable selection in multivariate regression: an application of simultaneous test procedures. *J. R. Statist. Soc. B* **39** 371-380.

McQuarrie, A. (1999). A small-sample correction for the Schwarz SIC model selection criterion. *Statist. Probab. Lett.* **44**, 79-86.

McQuarrie, A., Shumway, R. and Tsai, C. L. (1997). The model selection criterion AICu. *Statist. Probab. Lett.* **34**, 285-292.

McQuarrie, A. and Tsai, C. L. (1998). *Regression And Time Series Model Selection.* World Scientific, Singapore.

Miller, A. J. (1990). *Subset Selection in Regression.* Chapman and Hall, London.

Nason, G. P. (1996). Wavelet regression by cross-validation. *J. Roy. Statist. Soc. Ser. B* **58**, 463-479.

Nelder, J. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135**, 370-384.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, 758-765.

Nishii, R. (1988). Maximum likelihood principle and model selection when the true model unspecified. *J. Multivariate Anal.* **27**, 392-403.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. Roy. Statist. Soc. Ser. B* **57**, 99-138.

Paulsen, J. (1984). Order determination of multivariate autoregressive time series with unit roots. *J. Time Ser. Anal.* **5**, 115-127.

Paulsen, J. and Tjøstheim, D. (1985). Least squares estimates and order determination procedures for autoregressive processes with a time dependent variance. *J. Time Ser. Anal.* **6**, 117-133.

Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models. *J. Amer. Statist. Assoc.* **79**, 575-583.

Pötscher, B. M. (1989). Model selection under nonstationarity: autoregressive models and stochastic linear regression models. *Ann. Statist.* **17**, 1257-1274.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory* **7**, 163-185.

Pötscher, B. M. and Novák, A. J. (1998). The distribution of estimators after model selection: large and small sample results. *J. Statist. Comput. Simulation* **60**, 19-56.

Pregibon, D. (1979). Data analytic methods for generalized linear models. Ph. D. dissertation, Univ.. of Toronto, Canada.

Pukkila, T., Koreisha, S. and Kallinen, A. (1990). The identification of ARMA models. *Biometrika* **77**, 537–548.

Pukkila, T. and Krishnaiah, P. R. (1988a). On the use of autoregressive order determination criteria in univariate white noise tests. *IEEE Trans. Acoust., Speech and Signal Processing* **36**, 764-774.

Pukkila, T. and Krishnaiah, P. R.(1988b). On the use of autoregressive order determination criteria in multivariate white noise tests. *IEEE Trans. Acoust., Speech and Signal Processing* **36**, 1396-1403.

Qian, G., Gabor, G. and Gupta, R. P. (1996). Generalized linear model selection by the predictive least quasi-deviance criterion. *Biometrika* **83**, 41-54.

Qian, G. and Künsch, H. R. (1998). On model selection via stochastic complexity in robust linear regression. *J. Statist. Plann. Inference* **75**, 91-116.

Quinn, B. G. (1980). Order determination for multivariate autoregression. *J. Roy. Statist. Soc. Ser. B* **42**, 182-185.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* **92**, 179-191.

Rao, C. R. (1987). Prediction of future observations in growth curve models. *Statist. Sci.* **2**, 434-471.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76**, 369-374.

Rao, J. S. (1994). Adaptive subset selection via cost-optimization using resampling methods in linear regression models. Ph.D. dissertation, Univ. of Toronto, Canada.

Rencher, A. C. and Pun, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics* **22**, 49-53.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica* **14**, 465-471.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11**, 416-431.

Rissanen, J. (1986a). Stochastic complexity and modelling. *Ann. Statist.* **14**, 1080-1100.

Rissanen, J. (1986b). A predictive least squares principle. *IMA J. Math. Control. Inform.* **3**, 211-222.

Rissanen, J. (1986c). Order estimation by accumulated prediction errors. In *Essays in Time Series and Allied Processes* (J. Gani and M. B. Priestly, eds.). *J. Appl. Probab.* **23A** 55-61.

Rissanen, J. (1989). Stochastic Complexity in Statistical Inquiry. World Scientific, Singapore.

Ronchetti, E. (1985). Robust model selection in regression. *Statist. Prob. Lett.* **3**, 21-23.

Ronchetti, E. (1997). Robustness aspects of model choice. *Statist. Sinica* **7**, 327-338.

Ronchetti, E., Field, C. and Blanchard, W. (1997). Robust linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **92**, 1017-1023.

Ronchetti, E. and Staudte, R. G. (1994). A robust version of Mallows's C_p. *J. Amer. Statist. Assoc.* **89**, 550-559.

Sakamoto, Y. , Ishiguro, M. and Kitagawa, G. (1986). *Akaike Information Criterion Statistics.* KTK Scientific Publishers, Tokyo.

Sakamoto, Y. (1991). *Categorical Data Analysis by AIC.* KTK Scientific Publishers, Tokyo.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica* **46**, 1273-1291.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.* **91** (1996), 655-665.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica*, **7**, 221-264.

Shao, J. (1998). Convergence rates of the generalized information criterion. *J. Nonparametric Statist.* **9**, 217-225.

Shao, J. and Rao, J. S. (2000). The GIC for model selection: A hypothesis testing approach. *J. Statist. Plann. Inference* **88**, 215-231.

Shi, P. and Tsai, C. L. (1998a). On the use of marginal likelihood in model selection. Technical Report, Graduate School of Management, Univ. California, Davis.

Shi, P. and Tsai, C. L. (1998b). A note on the unification of the Akaike information criterion. *J. R. Stat. Soc. Ser. B* **60**, 551-558.

Shi, P. and Tsai, C. L. (1999). Semiparametric regression model selections. *J. Statist. Plann. Inference* **77**, 119-139.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117-126.

Shibata, R. (1980). Asymptotic efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8**, 147-164.

Shibata, R. (1986a). Selection of regression variables. In *Encyclopedia of Statistical Sciences VII* (S. Kotz and N. L. Johnson, eds.), 709-714, John Wiley & Sons.

Shibata, R. (1986b). Selection of the number of regression variables; a minimax choice of generalized FPE. *Ann. Inst. Statist. Math.* **38**, 459-474.

Shibata, R. (1989). Statistical aspects of model selection. In *From Data to Model* (J. C. Willems, ed.), 215-240, Springer-Verlag.

Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. *Statist. Sinica* **7**, 375-394.

Shimodaira, H. (1994). A new criterion for selecting models from partially observed data. In *Selecting Models from Data: Artificial Intelligence and Statistics IV, Lecture Notes in Statistics* (P. Cheeseman and R. W. Oldford, eds.), 21-29, Springer, New York.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer-Verlag.

Simonoff, J. S. and Tsai, C. L. (1999). Semiparametric and additive model selection Using an Improved AIC Criterion. *J. Comput. Graph. Statist.* **8**, 22-40.

Sommer, S. and Huggins, R. M. (1996). Variable selection using the Wald test and a robust C_p. *Appl. Statist.* **45**, 15-29.

Sommer, S. and Staudte, R. G. (1995). Robust variable selection in regression in the presence of outliers and leverage points. *Austral. J. Statist.* **37**, 323-336.

Speed, T. P. and Yu, B. (1993). Model selection and prediction: normal regression. *Ann. Inst. Statist. Math.* **45**, 35-54.

Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction. *J. Roy. Statist. Soc. Ser. B.* **36**, 111-133.

Stone, M. (1977a). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B.* **39**, 44-47.

Stone, M. (1977b). Asymptotics for and against cross-validation. *Biometrika* **64**, 29-38.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist.- Theory Meth.* **7** 13-26.

Simonoff, J. S. and Tsai, C. L. (1999). Semiparametric and additive model selection Using an Improved AIC Criterion. *J. Comput. Graph. Statist.* **8**, 22-40.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku* (Mathematic Sciences) **153**, 12-18. (In Japanese).

Thall, P. F., Simon, R., and Grier, D. A. (1992). Test-based variable selection via cross-validation. *J. Comput. Graph. Statist.* **1**, 41-61.

Thall, P. F., Russell, K. E. and Simon, R. M. (1997). Variable selection in regression via repeated data splitting. *J. Comput. Graph. Statist.* **6**, 1-34.

Tiao, G. C. and Xu. D. (1993). Robustness of MLE for multi-step predictions: the exponential smoothing case. *Biometrika* **80**, 623-641.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Tibshirani, R. and Knight, K. (1999a). The covariance inflation criterion for adaptive model selection. *J. Roy. Statist. Soc. Ser. B.* **61**, 529-546.

Tibshirani, R. and Knight, K. (1999b). Model search and inference via bootstrap bumping. *J. Comput. Graph. Statist.* **8**, 671-686.

Tsay, R. S. (1984) Order selection in nonstationary autoregressive models. *Ann. Statist.* **12**, 1425-1433.

Venter, J. H. and Steel, S. J. (1992) Some contributions to selection and estimation in the normal linear model. *Ann. Inst. Statist. Math.* **44**, 281-297.

Wakefield, J. C. and Bennett, J. E. (1996). The Bayesian modelling of covariates for population pharmacokinetic models. *J. Amer. Statist. Assoc.* **91**, 917-927.

Wei, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20**, 1-42.

Whittle, P. (1963). *Prediction and Regulation by Linear Least-Square Methods.* English Univ. Press, London.

Williams, D. A. (1970a). Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics* **28**, 23-32.

Williams, D. A. (1970b). Discussion of "A method for discriminating between models," by A. C. Atkinson. *J. Roy. Statist. Soc. Ser. B* **32**, 350.

Wu, Y. (2001). An M-estimation-based model selection criterion with a data-oriented penalty. *J. Statist. Comput. Simulation.* To appear.

Wu, Y., Tam, K. W., Li, F. and Zen, M. M. (1999). A note on estimating the number of super imposed exponential signals by the cross-validation approach. Technical Report, Center for Multivariate Analysis, Penn. State Univ.

Wu, Y. and Zen, M. (1999). A strongly consistent linear model selection procedure based on M-estimation. *Probability Theory and Related Fields* **113**, 599-625.

Yao, Q. and Tong, H. (1994). On subset selection in non-parametric stochastic regression. *Statist. Sinica* **4**, 51-70.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93**, 120-131.

Zhang, H. M. and Wang, P. (1994). A new way to estimate orders in time series. *J. Time Ser. Anal.* **15**, 545-559.

Zhang, P. (1991). Variable selection in nonparametric regression with continuous covariates. *Ann. Statist.* **19**, 1869-1882.

Zhang, P. (1992). Inference after variable selection in linear regression models. *Biometrika* **79**, 741-746.

Zhang, P. (1993a). Model selection via multifold cross validation. *Ann. Statist.* **21**, 299-313.

Zhang, P. (1993b). On the convergence rate of model selection criteria. *Commun. Statist.-Theory Meth.*, **22**, 2765-2775.

Zhang, P. (1994). On the choice of penalty term in generalized FPE criterion. In *Selecting Models from Data* (P. Cheeseman and R. W. Oldford, eds.), 41-49, Springer-Verlag, New York.

Zhang, P. and Krieger, A. M. (1993) Appropriate penalties in the final prediction error criterion: a decision theoretic approach. *Statist. Probab. Lett.* **18**, 169–177.

Zhao, L. C., Krishnaiah, P. R. and Bai, Z. D. (1986a). On detection of the number of signals in presence of white noise. *J. Multivariate Anal.* **20**, 1-25.

Zhao, L. C., Krishnaiah, P. R. and Bai, Z. D. (1986b). On determination of the number of signals when the noise covariance matrix is arbitrary. *J. Multivariate Anal.* **20**, 26-49.

Zheng, X. and Loh, W. (1995). Consistent variable selection in linear models. *J. Am. Statist. Assoc.* **90**, 151-156.