MONASH

WP 18/96

ISSN 1032-3813
ISBN 0 7326 0799 X

MONASH UNIVERSITY

ANCORA · IMPARO

AUSTRALIA

IMPROVED SMALL SAMPLE MODEL SELECTION PROCEDURES

Maxwell L. King
Catherine Scipione Forbes
Alan Morgan

Working Paper 18/96
September 1996

DEPARTMENT OF ECONOMETRICS

Improved Small Sample Model Selection Procedures

Maxwell L. King, Catherine Scipione Forbes,
and Alan Morgan [*]
Department of Econometrics, Monash University
Clayton, Victoria 3168 Australia

September 1996

Abstract

This paper is concerned with model selection based on penalized maximized log likelihood functions. Its main emphasis is on how these penalties might be chosen in small samples to give good statistical properties. We explore how some of the more successful principles and practices in hypothesis testing can be used to improve the properties of these model selection procedures. This leads to choosing the penalties in order to control probabilities of different models being selected. Various ways this can be achieved using simulation methods are discussed and a computer algorithm is outlined. Some illustrative Monte Carlo simulations are also reported.

Keywords: Common models; information criteria; Monte Carlo methods; Neyman-Pearson lemma; penalty functions.

1 Introduction

Often in econometrics we are forced to use data to make a choice between a number of competing alternative models. Obvious examples include deciding which regressors (and often lagged regressors) to include in a linear regression, the appropriate order of an ARMA model, and what

[*]This research was supported by an Australian Research Council grant. This paper was presented at the 1995 World Congress of the Econometrics Society, Tokyo.

lags to include in a vector autoregressive model. In general, a range of models is considered plausible and a decision is made based on how well each of the models appears to fit the observed data. Consequently model selection procedures have an important role to play in econometric modeling.

A number of different approaches to the model selection problem have been considered in the literature. One approach is the use of a series of pairwise hypotheses tests. While this approach is very common in practice, it has a number of limitations. In each step one model has to be chosen as the null hypothesis and, if the power of the test is low, this model is unfairly favored. On the other hand, if there is a lot of data and the power of the test is very high, this could disadvantage the null hypothesis model. Typically tests are constructed to make the probability of a type I error constant over the null hypothesis parameter space. It is not clear this is a desirable property when non-nested hypotheses are involved. (See King (1983, 1987) and Granger, King and White (1995).) There are also pre-testing biases that come into play when a series of tests is employed. (See King and Giles (1984) and Giles and Giles (1993).) Finally, different investigators working on the same data could easily end up with different selections purely because they performed their series of tests in different orders or used different levels of significance.

The classical hypothesis testing approach does help provide a solution to the very special case of choosing between two simple models. Here the Neyman-Pearson lemma tells us that the likelihood ratio is the most powerful statistic for discriminating between the two models. There is of course the question of what critical value should be used. While classical hypothesis testing suggests controlling the probability of a Type I error, model selection would seem to require finding a critical value that does not favor one model over the other. The familiar likelihood ratio statistic is a natural generalization of this approach to two models with unknown parameters, although again there remains the issue of choosing the critical value. Unfortunately the generalization to more than two simple models is much less obvious for model selection, although one might expect it to at least involve likelihood ratio statistics in some way. While there are well developed principles put forward in the hypothesis testing literature, such as the use of invariance, fewer guidelines are available for the model selection problem.

Interestingly, probably the most widely used method of model selection is the use of an information criterion (IC) based on minus the maximized log-likelihood function plus a penalty term. Pötscher (1991) pointed out that minimizing such an IC amounts to testing each model against all other models by means of a standard likelihood ratio test and selecting that model which is accepted against all other models; the critical values being determined by the penalty function. Not surprisingly there is little agreement about what the form of the penalty function should be. The aim of this paper is to explore how some of the more successful principles and practices in hypothesis testing might be used to solve the penalty function dilemma.

In the past, asymptotic arguments have been used to justify various choices of penalty functions. This may not always be satisfactory in small samples. For example, Grose and King (1994) have found that in some circumstances the form of the likelihood functions can have a greater influence than the data in determining which model is chosen. This suggests that a small sample approach is needed to better control the probabilities of correct selection. In looking to the hypothesis testing literature for solutions, it is obvious that there are other areas that can be improved. These include the treatment of nuisance parameters and making good use of parameter restrictions. For example, much has been written about one-sided hypothesis testing (for a recent survey, see Wu and King (1994)), but almost nothing has been written on making good use of such information in the model selection context.

The plan of the paper is as follows. Some of the issues of small sample model selection procedures are discussed in Section 2 along with a suggested method for determining a new procedure. Section 3 considers in detail how penalties might be calculated in order to control probabilities of selection. The penalties may be based on somewhat objective criteria, or more subjective considerations. An algorithm for calculating penalties based on the above ideas using Monte Carlo techniques is outlined in Section 4. A Monte Carlo comparison of the small sample performance of the suggested procedures along with that of AIC and BIC in a simple regression setting and in choosing the order of an ARMA model are presented in Section 5. The final section contains some concluding remarks.

2 Small sample issues

In order to explore the relevant model selection issues, we consider first the elementary case of choosing between two simple models. Let G_1 and G_2 be probability distributions possessing densities with respect to a common measure μ denoted g_1 and g_2 respectively, and let $\mathbf{x} = \{x_1, \ldots, x_n\}$ denote an observed sample of size n. Here G_1 and G_2 correspond to two distinct model choices. The Neyman-Pearson fundamental lemma states that the most powerful test of G_1 against G_2 rejects G_1 in favor of G_2 only if

$$\ln g_1(\mathrm{x}) - \ln g_2(\mathrm{x}) < p, \tag{1}$$

where the critical value p is selected so that

$$P_{G_1}(\text{reject } G_1) = \alpha, \tag{2}$$

for some prespecified (and typically small) level α. Here $P_G(E)$ denotes the probability of event E under the distribution G. For hypothesis testing, controlling the probability of incorrectly rejecting the null hypothesis, in this case G_1, takes priority over controlling the probability of incorrectly concluding G_1. In fact, the probability of choosing G_2 may be very low, as is typically the case with very small samples. However, this test is still used because it is the best that can be done among all tests with fixed level α.

From a hypothesis testing point of view this favoring of the null hypothesis may have some justification. However, from a model selection point of view this does not seem to be desirable, particularly in this simple problem. Rather, the main objective of a good model selection procedure should be to make the correct decision as often as possible, without unnecessarily favoring one model over another. One way to achieve this objective is to choose p so that

$$P_{G_2}(\ln g_1(\mathrm{x}) - \ln g_2(\mathrm{x}) < p) = P_{G_1}(\ln g_1(\mathrm{x}) - \ln g_2(\mathrm{x}) \geq p), \tag{3}$$

or equivalently,

$$P_{G_2}(\text{choose } G_2) = P_{G_1}(\text{choose } G_1). \tag{4}$$

It is likely that for small sample sizes these probabilities are small, and that their value will increase with n. In addition, it is likely that the value of p itself will depend on the sample size.

Indeed it seems valuable to be able to calculate how 'powerful' a given model selection procedure is for a particular sample size.

We have identified some of the important issues in the specific case of comparing two simple models. In most model selection problems, however, not only are there more than two distributions under consideration, but also each distribution is often indexed by an unknown parameter. For the case where there are more than two models under consideration and none depend on an unknown parameter vector, we can easily generalize the above. Consider probability distributions $G_1, \ldots, G_m$, where $m > 2$, with density functions with respect to some dominating measure μ, denoted $g_1, \ldots, g_m$, respectively. An obvious generalization of the simple versus simple model selection procedure is to choose model G_i only if

$$\ln g_i(\mathbf{x}) - p_i > \ln g_j(\mathbf{x}) - p_j \text{ for all } j \neq i. \tag{5}$$

As only $m-1$ critical values, or penalty functions, p_i are required, we can set $p_1 = 0$ and determine the $m-1$ remaining p_i values from the added condition

$$P_{G_1}(\text{choose } G_1) = P_{G_2}(\text{choose } G_2) = \ldots = P_{G_m}(\text{choose } G_m) \tag{6}$$

where

$$P_{G_i}(\text{choose } G_i) = P_{G_i}(\ln g_i(\mathbf{x}) - p_i > \ln g_j(\mathbf{x}) - p_j \text{ for all } j \neq i). \tag{7}$$

The difficulty lies in generalizing this notion further to the case where the models depend on unknown parameters. A natural generalization of the difference of logs of density functions as in (5) is the log of the likelihood ratio statistic. We begin with some notation and definitions. The general model selection problem consists of choosing between models $M_1, \ldots, M_m$, corresponding to probability distributions $G_1[\theta_1], \ldots, G_m[\theta_m]$. Here $\theta_i \in \Theta_i$ represents an unknown parameter vector, and typically $\Theta_i \subseteq \mathbb{R}^{q_i}$. Let $L_i(\theta_i) : \Theta_i \rightarrow \mathbb{R}$ denote the log-likelihood function of model M_i based on a sample of size n and let L_i denote the value of the maximized log-likelihood function for model M_i. That is,

$$L_i = \sup_{\theta_i \in \Theta_i} L_i(\theta_i). \tag{8}$$

Model selection procedures based on a comparison of log likelihood functions are typically referred to as IC procedures in the literature. Most IC model selection procedures define a penalty term for each model, denoted p_i, and select the model with the largest value of $L_i - p_i$ (or equivalently the smallest $-L_i + p_i$ value). Two well known examples of IC procedures are AIC, where $p_i = q_i$ and BIC, where $p_i = \frac{1}{2} q_i \ln n$. Both AIC and BIC heavily penalize models with a large number of parameters. This is often advocated by asymptotic arguments and/or appealing to the principle of parsimony. However, in the case where the models under consideration have the same number of parameters, both AIC and BIC reduce to a comparison of log-likelihood values alone as the penalty functions are all equal. As was detailed in Grose and King (1994), the use of an IC in the simple case of comparing AR(1) versus MA(1) regression disturbances, the probabilities of correctly choosing each model were far from equal, particularly for smaller sample sizes. Our goal should be to choose $p_i, i = 2, \ldots, m$, so as to control the probabilities of correct selection but in such a way that no one model is favored unknowingly. The principle issue then is exactly how this can be achieved.

Some econometricians feel uncomfortable using IC based model selection procedures because they feel it can lead to data mining. This is the problem of having a very large number of models to choose from and only a limited sample of data to make the choice. Consequently the probability of the chosen model being the true model can be very small indeed. Deficiencies in the chosen model typically come to light when the model is used for out of sample forecasting. Another important issue is how to guard against, or reduce the problems of, data mining. The hypothesis testing literature suggests that nonsample information, such as knowledge of the signs of unknown parameters, should be used wherever possible with the aim of increasing the probability of correct choice.

There is a rich literature on one-sided and multivariate one-sided hypothesis testing but almost nothing is written on one-sided model selection procedures. Hughes and King (1994) have derived a multivariate one-sided (or partially one-sided) AIC procedure which involves smaller penalties for inequality restricted parameters. Their simulation results suggest that when one-sided information is known, using appropriate inequality constrained estimates in place of the usual unrestricted estimates typically helps improve the small sample properties of the IC procedures. However, there is a tendency to penalize too heavily for extra parameters, hence the need to modify the penalty functions.

It would also be helpful if the number of parameters could be reduced. In most model selection problems there are at least some nuisance parameters in the sense that they appear in each model and are therefore not in dispute. There is now a rich literature (see for example Kalbfleish and Sprott (1970), Basu (1977), Cox and Reid (1987) and Tunnicliffe Wilson (1989)) on how to deal with these nuisance parameters. The main message is that nuisance parameters can cause classical maximum likelihood estimators to be biased. Elimination of these nuisance parameters through invariance or the use of marginal or other modified likelihood functions does appear to result in improved small sample properties, particularly for asymptotic likelihood based test procedures such as the likelihood ratio test. It seems obvious that such techniques should be used to improve IC model selection procedures. There is one study that has taken up this point. Grose and King (1994) have shown that replacing the classical likelihood function with the marginal likelihood function, or equivalently the likelihood function of the maximal invariant statistic, in their case of choosing between AR(1) and MA(1) regression disturbances, results in better small sample properties.

Another way to guard against data mining would be to have a calculated measure that reflects the level of confidence in the final selection. This may involve calculating probabilities of correct selection, or conditional probabilities of correct selection given the choice made. It would therefore be a bonus if such probabilities could easily be calculated as a part of the model selection procedure.

On the positive side, the amount of computer power available to econometricians is continually increasing, and there is probably no reason to doubt that it will continue to increase in the future. We can therefore ask questions such as 'What procedure would we like to use?' rather than 'What procedure can we use given our current computing constraints?'. Simulation methods provide a powerful tool for evaluating complex probabilities should they be required for any procedure. In addition, simulation methods can be used to evaluate any model selection procedure in any set of circumstances. Unfortunately, when any set of procedures is compared empirically in this manner, usually there is no clear cut answer to the question of which is best. This leads to the critical question of how we should evaluate and compare the small sample performances of different model selection procedures.

3 Controlling probabilities of correct selection

In this section we consider small sample based methods for choosing the penalty functions in an IC model selection procedure so as to control the probabilities of selection. Based on the notation introduced for the m model selection problem in Section 2, our aim is to provide methods for determining $p_i, i = 2, \ldots, m$. One can view these penalties as similar to $m-1$ critical values which in hypothesis testing would be determined by probability equations. Hence, if probabilities of selection are to be controlled, we need $m-1$ equations to solve for our $m-1$ unknowns. There are many ways in which these equations could be defined. Some are outlined below.

Our first suggestion involves evaluating probabilities of selecting each of the models at the same parameter point. This has the advantage that the probabilities must sum to one. Because of this constraint, $m-1$ equations can be found setting $m-1$ such probabilities equal to predetermined values. There is of course the question of which parameter point to use and what the predetermined probabilities should be. One solution to the former question is to use the so called 'minimal' or 'common' model. This occurs when the parameters under dispute in each model are set to a constant (typically zero) and the same model always results. This latter model is the minimal or common model at which we propose the probabilities of selection should be controlled. In order to illustrate, consider the simple setting of choosing between m regression models

$$\begin{aligned} M_1 : y_t &= \beta_1 x_{1t} + u_{1t}, & u_{1t} &\sim IN(0, \sigma_1^2) \\ M_2 : y_t &= \beta_2 x_{2t} + u_{2t}, & u_{2t} &\sim IN(0, \sigma_2^2) \\ &\vdots \\ M_m : y_t &= \beta_m x_{mt} + u_{mt}, & u_{mt} &\sim IN(0, \sigma_m^2). \end{aligned} \tag{9}$$

$G_i[\beta_i, \sigma_i]$ is a multivariate Normal distribution with a mean whose t^{th} component is $\beta_i x_{it}$ and covariance matrix $\sigma_i^2 \mathcal{I}_n$, where $\mathcal{I}_n$ is the $n \times n$ identity matrix. Notice that when all $\beta_i^* = 0.0$ and $\sigma_1^2 = \ldots = \sigma_m^2 = \sigma^2$, say, the different regression models M_i describe the same data generating process (DGP). That is, under these conditions, all of models can be considered to be 'true' simultaneously. Setting $p_1 = 0$, we define the penalties $p_2, \ldots, p_m$ according to

$$P_{G_i[\beta_i^*, \sigma_i^2]}(\text{ choose } M_i) = \frac{1}{m} \tag{10}$$

for $i = 1, \ldots, m$. It may seem that the nuisance parameter σ^2 needs to be determined before these probabilities can be calculated. Fortunately, because the problem of choosing between $M_1, \ldots, M_m$ is invariant to changes in the scale of y_t, the probabilities in (10) are invariant to the value of σ^2 and hence can be calculated for any convenient value such as $\sigma^2 = 1$. In the particular case of $m = 2$, p_2 could be found using Imhof's (1961) algorithm, but this does not generalize easily to wider classes of models. An obvious alternative is to use Monte Carlo simulation, which has the advantage of being easily extended for $m > 2$ in nested problems as well as many nonnested situations outside of this simple regression setting. What is required is that all m models contain a common model. That is, there is some point in the parameter space for each model, $\theta_i = \theta_i^c$, where the different models can be considered to represent the same DGP.

The other important consideration is what predetermined values should the probabilities of selection under the common model be set to. In the example (9) above, the choice of $1/m$ does seem sensible. This is because at the common model we are essentially indifferent between the m models. They are either equally correct or equally incorrect, depending on one's viewpoint. As one of the β_i values moves away from zero, we could expect the probability of choosing the i^{th} model to increase (from $1/m$) while the probabilities of choosing the j^{th} model ($j \neq i$) would be expected to decline. Hence this choice of $1/m$ (and assuming our expectations are true) always gives the correct model the highest probability of correct selection which is a highly desirable property.

The difficulty with this approach comes when some alternative models are nested within other models. For example, suppose $m = 3$, M_1 and M_2 are as before, but now

$$M_3 : y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + u_{3t}, \qquad u_{3t} \sim \mathrm{IN}(0, \sigma_3^2). \tag{11}$$

Then $y_t = u_{it}$ is still the common model but we may no longer be indifferent between M_1, M_2 and M_3. One might view M_3 as "less correct" than M_1 and M_2 when the common model is the true model. This problem can be overcome by setting the selection probabilities under the common model to $r/2$ for models M_1 and M_2 and to $1 - r$ for model M_3. It should be up to the individual user to set r. We suggest a large probability for r in this setting such as $r = 0.9$. This comes from considering the case in which the common model is one of the competing models.

Then one might prefer a larger probability (such as 0.9) of choosing it when true.

Unfortunately the common model approach has a degree of arbitrariness when nested alternative models are involved which makes it less than attractive. Also there are settings in which it is not possible to find a common model.

Another approach is to control probabilities of correct selection at representative fixed points for each of the competing models. This has an element of the point optimal testing methodology about it and results in m probabilities while only $m-1$ equations are needed to determine the penalty values $p_2, \ldots, p_m$. An obvious solution is to find those penalty values such that the probabilities of correct selection at the chosen parameter points are equal; the latter being justified by the desire not to favor a particular model. Alternatively, certain models can be favored explicitly if we choose $p_2, \ldots, p_m$ by equating the appropriate predetermined proportions of the probabilities of correct selection for each model at their chosen representative values.

The question then is how should the representative fixed points be chosen. One approach is to leave this entirely to the user. One can view the choice of penalty values $p_2, \ldots, p_m$ as the choice of a particular set of selection probabilities over the entire parameter spaces of each of the models. Rather than choosing the largely unknown selection probabilities that come with 'brand name' penalties such as AIC or BIC, the user chooses the set of selection probabilities by setting them at predetermined points. Here an important by-product is the actual probabilities themselves which gives the user an important measure of the accuracy of the procedure.

Another approach to choosing the representative point θ_i^* is to fix its value at some 'distance' away from θ_i^c corresponding to the common model. For the above regression problem $\theta_i^c = (\beta_i^c = 0.0, \sigma_i^{2c} = 1.0)$. It is well known that under model M_i, with parameter $\theta_i = \theta_i^c$, that the least squares estimator

$$\hat{\beta}_i = \sum_{t=1}^{n} x_{it} y_t / \sum_{t=1}^{n} x_{it}^2 \tag{12}$$

has a Normal distribution with mean 0.0 and variance

$$\mathrm{var}(\hat{\beta}_i) = (\sum_{t=1}^{n} x_{it}^2)^{-1}. \tag{13}$$

One suggestion would be to set

$$\beta_i^* \quad = \quad \beta_i^c + 3\sqrt{\mathrm{var}(\hat{\beta}_i)} \tag{14}$$

$$= \quad \beta_i^c + 3\sqrt{(\sum_{t=1}^{n} x_{it}^2)^{-1}} \tag{15}$$

and $\sigma_i^{2*} = 1.0$, for $i = 1, \ldots, m$. The penalty functions are again determined by solving

$$P_{G_i[\theta_i^*]}(\text{ choose } M_i) = P_{G_j[\theta_j^*]}(\text{ choose } M_j) \tag{16}$$

for $p_2, \ldots, p_m$, with $p_1 = 0$. In this case θ_i^* was chosen so that the 'distance' between θ_i^* and θ_i^c is 3 standard deviations of $\hat{\theta}_i$ under the distribution $G_i[\theta_i^c]$. Three standard deviations was chosen simply because at that value of θ_i^* it is unlikely that another model would generate data that would mistakingly appear to have come from $G_i[\theta_i^*]$.

For models with high dimensional parameters and in cases that cannot be reduced by invariance or other arguments, there may be many such θ_i^* points at a specified distance away from θ_i^c. One suggestion is to calculate the penalties $p_2, \ldots, p_m$, based on the average probability of correct selection. In regular problems, the maximum likelihood estimator of θ_i under model M_i has an asymptotic q_i-variate Normal distribution with mean vector θ_i^c and covariance matrix equal to the inverse of the Fisher information matrix evaluated at the common model, $\mathcal{J}_i^c$. In this case, any points on the ellipsoid

$$\mathcal{E}_i = \{\theta_i \in \Theta_i : \theta_i' \mathcal{J}_i^c \theta_i = k^2\} \tag{17}$$

are in some (asymptotic) sense equidistant from θ_i^c. We can then calculate penalties based on the averaged probabilities of correct selection

$$\int_{\mathcal{E}_i} P_{G_i[\theta]}(\text{ choose } M_i) d\theta_i. \tag{18}$$

The value of k can be taken to be any number, however we think something of the order of $k = 2$ or $k = 3$ will be sufficient for discriminating models. Following from (18), we could define other functions of the probabilities in terms of defining the penalties.

4 Main algorithm

In this section we discuss the algorithm for calculating penalties based on the ideas in Section 3. This algorithm applies to the case when the m models $M_1, \ldots, M_m$ depend on unknown parameters $\theta_1, \ldots, \theta_m$, respectively, as well as the case when no unknown parameters are involved. In particular, we will discuss algorithms for calculating the penalty function based on fixing probabilities of selection according to the common model approach, when $\theta_i^* = \theta_i^c$, and also according to fixing $\theta_i^* \in \Theta_i$ at points not corresponding to a common model.

Our main approach is to use simulation methods for calculating the probabilities of correct selection. We simulate $k = 1, \ldots, R$ data sets Y_{lk} with sample size n from each of the probability distributions $G_l[\theta_l^*]$, for $l = 1, \ldots, m$, and calculate the maximized log-likelihoods, L_i^{lk}, for each model $i = 1, \ldots, m$ under each data set for $k = 1, \ldots, R$. Let

$$1(L_i^{lk} - p_i > L_j^{lk} - p_j \text{ for all } j \neq i \text{ and } i = l) \tag{19}$$

denote the indicator function for the event that model M_i is correctly chosen for data set Y_{ik}. The penalties are found such that the empirical probability of correctly choosing model M_i

$$\hat{P}_{G_i[\theta_i^*]}(\text{choose } M_i) = \frac{1}{R}\sum_{k=1}^{R} 1(L_i^{lk} - p_i > L_j^{lk} - p_j \text{ for all } j \neq i \text{ and } i = l) \tag{20}$$

is controlled to the desired value

$$c_i = P_{G_i[\theta_i^*]}(\text{choose } M_i). \tag{21}$$

In the case when a common model is used, $\sum_{i=1}^{m} c_i = 1$. If the penalties are to be constructed so that the probabilities of correct selection are all equal, then $c_i = 1/m$ for all $i = 1, \ldots, m$. In the case where the user specifies the values of the c_i directly, the values are the same throughout the entire algorithm. In the case when a common model is not used and we are equating probabilities of correct selection, the common value of $c_i = c$, for $i = 1, \ldots, m$, is calculated within the algorithm as

$$c = \frac{1}{m}\sum_{i=1}^{m} \hat{P}_{G_i[\theta_i^*]}. \tag{22}$$

Using either approach, the algorithm stops when for all of the differences d_i

$$\max_{i=1,\ldots,m} \mid d_i \mid < \tau, \tag{23}$$

where

$$d_i = \hat{P}_{G_i[\theta_i^*]} - c_i \tag{24}$$

for some preselected tolerance level τ.

The general algorithm proceeds as follows:

1. initialise penalty $p_i = 0$ for $i = 1, \ldots, m$ and stepsize $s_i = 1.0$ for $i = 1, \ldots, m$;

2. generate data Y_{lk} for models $l = 1, \ldots, m$ and replications $k = 1, \ldots, R$;

3. calculate the $i = 1, \ldots, m$ maximised log-likelihood values L_i^{lk}, for each data set Y_{lk} for $l = 1, \ldots, m$, and $k = 1, \ldots, R$;

4. calculate $\hat{P}_{G_i[\theta_i^*]}$, according to (20); c_i according to (22) (if relevant); and d_i according to (24), for $i = 1, \ldots, m$;

5. if $\max_{i=1,\ldots,m} \mid d_i \mid < \tau$ then return $p_2, \ldots, p_m$, and stop;

6. define J such that $\max_{i=1,\ldots,m} \mid d_i \mid = \mid d_J \mid$;

7. if $J = 1$ then adjust the other $i = 2, \ldots, m$ penalties according to

   $p_i^* = p_i - sign(d_1) * s_1$

   leaving $p_1^* = 0$;

8. if $J \neq 1$ then adjust p_J only according to

   $p_J^* = p_J + sign(d_J) * s_J$;

   and leaving $p_i^* = p_i$ for $i \neq J$;

9. calculate $\hat{P}^*_{G_i[\theta_i^*]}$, c_i^* and d_i^* based on p_i^* for $i = 1, \ldots, m$;

10. if adjustment to penalty function is too large, change the stepsize by $s_J = s_J/10$;

11. reset $p_i = p_i^*$ and $d_i = d_i^*$ for $i = 1, \ldots, m$, and go to 5.

A few comments regarding the above algorithm are in order. First, we set M_1 to be the model with the fewest number of unknown parameters, if possible. In the nested situation this will result in all nonnegative penalties. Second, the number of Monte Carlo repetitions R affects the precision of $\hat{P}_{G_i[\theta_i^*]}$, and hence is related to whether or not the convergence criterion can be achieved. We recommend convergence be declared if the maximum distance between $\hat{P}_{G_i[\theta_i^*]}$ and the $P_{G_i[\theta_i^*]}$ desired to be less than $\tau = 2/R$. In some cases a tighter bound may be possible. Also note that the data sets Y_{lk} generated under each model need only be simulated once in Step 2, and if $\theta_i^* = \theta_i^c$, then only Y_{1k} need be generated and set $Y_{lk} = Y_{1k}$ for $l = 2, \ldots, m$. The strategy for modifying the penalties beginning with Step 6 is an adaptive procedure that changes the penalty associated with the largest absolute difference $\mid d_i \mid$. However, since p_1 is constrained to be zero, if $\mid d_1 \mid$ is the largest of the $\mid d_i \mid$ values, then all of the other penalties are reduced (or increased) to increase (or reduce) the relative penalty for model M_1. The step sizes are reduced by a factor of 10 when overcorrecting of the penalties occurs.

5 Monte Carlo simulation results

5.1 Monte Carlo simulations

To test our small sample methods of generating penalties, we used Monte Carlo simulation to evaluate the probabilities of correct selection for our new IC model selection procedures along with those of the AIC and BIC procedures. In each of the following examples, a sample size of $n = 30$ was used. We first considered a four model ($m = 4$) nested linear regression problem and calculated penalties based on setting probabilities of correct selection under a common model. Each model is of the form

$$M_i : Y = X_i\beta_i + U \tag{25}$$

where

$$\begin{aligned}
X_1 &= [x_1] \\
X_2 &= [x_1 : x_2] \\
X_3 &= [x_1 : x_2 : x_3] \\
X_4 &= [x_1 : x_2 : x_3 : x_4]
\end{aligned}$$

and

x_1 is an $n \times 1$ vector of ones

x_2 is an $n \times 1$ vector containing a quarterly seasonal dummy variable

x_3 is an $n \times 1$ vector containing a stationary AR(1) series with $\rho = .5$

x_4 is an $n \times 1$ vector containing a stationary AR(1) series plus a linear trend
with zero intercept and slope of .25

The disturbances in U are made up of a vector of independent standard normal random variables. We used the white noise model $Y = U$ as the common model since when each regression coefficient is zero, each model reduces to white noise. The penalties were calculated using our algorithm with $R = 2000$ replications. Table 1 shows two sets of c_i values used to control the probabilities of correct selection under the common model, while Table 2 displays the corresponding penalty functions, along with the AIC and BIC penalties. Note that the AIC and BIC penalties are adjusted so that the penalties associated with M_1 are all zero. Using the calculated penalties from Table 2 a Monte Carlo simulation using 1000 repetitions was used to assess the power of the two model selection procedures along with the AIC and BIC procedures. The results are presented in Table 3.

A similar study was completed for the case of choosing between white noise, MA(1) and AR(1) models. White noise again was the common model used to define the penalty functions based on $R = 2000$ in this nonnested situation for the two different sets of c_i probabilities given in Table 4. Table 5 presents the corresponding calculated penalties. To study the resulting probabilities of correct selection using these penalty functions, a Monte Carlo simulation was completed for a range of MA(1) parameter values, $\theta = \{.1, .3, .5, .7, .9\}$, as well as a range of AR(1) parameter values, $\rho = \{.1, .3, .5, .7, .9\}$. The results using 1000 repetitions are given in Table 6.

Next we illustrate the calculating of penalties under the alternative model set up. In this case we again compared white noise, MA(1), and AR(1) models by equating the probabilities of success. Two different sets of penalties were generated, once with both θ and ρ set equal to $2\sqrt{1/n} = 0.3651$ and once with both parameters set equal to $3\sqrt{1/n} = 0.5477$. These values correspond to 2 and 3 times the asymptotic standard error of the maximum likelihood parameter estimates under the assumption that the parameter values are zero. The resulting penalties are given in Table 7 and the Monte Carlo probabilities of correct selection for a range of θ and ρ values are given in Table 8.

Table 1: Probabilities of correct selection used to determine penalties in nested regression under common model of white noise

	M_1	M_2	M_3	M_4
	$\beta_1 = 0$	$\beta_2 = (0,0)'$	$\beta_3 = (0,0,0)'$	$\beta_4 = (0,0,0,0)'$
set 1	0.25	0.25	0.25	0.25
set 2	0.7025	0.1760	0.0780	0.0435

Table 2: Penalty functions for nested regression under common model of white noise

	p_1	p_2	p_3	p_4
set 1	0.0000	0.1992	0.6174	1.1824
set 2	0.0000	0.8159	1.9999	3.2999
AIC	0.0000	1.0000	2.0000	3.0000
BIC	0.0000	1.7006	3.4012	5.1018

Table 3: Monte Carlo probabilities of correct selection for nested regression using penalties calculated under white noise common model

	$Y = X_1\beta_1 + U$	$Y = X_2\beta_2 + U$	$Y = X_3\beta_3 + U$	$Y = X_4\beta_4 + U$
	$\beta_1 = 0.6$	$\beta_2 = (.6, .25)'$	$\beta_3 = (.6, .25, .225)'$	$\beta_4 = (.6, .25, .225, .05)$
set 1	0.230	0.248	0.472	0.426
set 2	0.696	0.242	0.312	0.168
AIC	0.710	0.176	0.312	0.204
BIC	0.896	0.106	0.186	0.070

Table 4: Probabilities used to determine penalties under white noise vs. MA(1) vs. AR(1) model selection problem under common model of white noise

	white noise	MA(1)	AR(1)
set 1	0.3333	0.3333	0.3333
set 2	0.90	0.05	0.05

Table 5: Penalty functions for white noise vs. MA(1) vs. AR(1) under common model of white noise

	white noise	MA(1)	AR(1)
	p_1	p_2	p_3
set 1	0.0000	0.1226	0.1020
set 2	0.0000	1.7000	1.5100
AIC	0.0000	1.0000	1.0000
BIC	0.0000	1.7006	1.7006

Table 6: Monte Carlo probabilities of correct selection when selecting between white noise vs. MA(1) vs. AR(1) using penalties calculated under common model of white noise

	white	MA(1)					AR(1)				
	noise	.1	.3	.5	.7	.9	.1	.3	.5	.7	.9
set 1	0.343	0.360	0.543	0.772	0.905	0.976	0.368	0.507	0.697	0.895	0.992
set 2	0.880	0.060	0.239	0.593	0.855	0.969	0.078	0.334	0.646	0.899	0.991
AIC	0.787	0.114	0.389	0.711	0.898	0.974	0.116	0.342	0.647	0.888	0.991
BIC	0.893	0.070	0.273	0.633	0.876	0.972	0.055	0.274	0.594	0.880	0.990

Table 7: Penalty functions for white noise vs. MA(1) vs. AR(1) with probabilities of correct selection set under alternative models.

	white noise	MA(1)	AR(1)
	p_1	p_2	p_3
2s.e.	0.0000	0.3908	0.3495
3s.e	0.0000	0.8150	0.9440
AIC	0.0000	1.0000	1.0000
BIC	0.0000	1.7006	1.7006

Table 8: Monte Carlo probabilities of correct selection when selecting between white noise vs. MA(1) vs. AR(1) using penalties calculated under alternative models

	white	MA(1)					AR(1)				
	noise	.1	.3	.5	.7	.9	.1	.3	.5	.7	.9
2 s.e.	0.563	0.235	0.486	0.729	0.909	0.979	0.256	0.481	0.685	0.883	0.984
3 s.e.	0.760	0.196	0.456	0.748	0.919	0.980	0.091	0.320	0.710	0.849	0.979
AIC	0.801	0.140	0.389	0.703	0.902	0.978	0.109	0.358	0.646	0.873	0.984
BIC	0.908	0.075	0.273	0.621	0.887	0.974	0.059	0.271	0.616	0.864	0.984

The calculated results for the regression problem (Table 3) only show probabilities of correct selection at one point per model. This can be a little misleading. However they do demonstrate that probabilities of correct selection can be controlled. Increasing one probability typically results in other probabilities declining. It is clear to us the first strategy of setting all selection probabilities under the common model equal in this very nested selection problem is less than satisfactory. While higher probabilities of correct selection result for more complex models, this is achieved at the expense of lower probabilities of correctly choosing the simplest model.

The time series results in Tables 6 and 8 illustrate the tradeoffs that are typically involved in setting penalties in IC selection procedures. High probabilities of correct selection of the white noise model come at the expense of lower probabilities of correct selection of the AR(1) or MA(1) models when the associated parameters are small.

In both cases there is no one set of probabilities of correct selection that is clearly better than the other probabilities. Which set of penalties should be chosen is really up to the user.

6 Concluding remarks

We have discussed some of the issues involved in model selection, particularly when viewed in light of our experience in hypothesis testing. The main point of this paper has been to demonstrate how the user can set the penalties in IC model selection procedures and hence take control of probabilities of selection. An important by-product of one of the approaches is the estimation of some probabilities of correct selection which provide a useful measure of how reliable the overall procedure is. Taking control of probabilities of correct selection is one matter, but how that control should be exercised is an issue we are continuing to research. The present paper has given some tentative suggestions. In the future we hope to be able to present more convincing strategies, particularly concerning the control of average probabilities.

References

[1] Basu, D. (1977) 'On the elimination of nuisance parameters,' *Journal of the American Statistical Association,* 72, 355-366.

[2] Cox, D. R. and N. Reid (1987) 'Parameter orthogonality and approximate conditional inference,' *Journal of the Royal Statistical Society B,* 41, 113-147.

[3] Giles, J. A. and D. E. A. Giles (1993) 'Pre-test estimation and testing in econometrics: Recent developments,' *Journal of Economic Surveys,* 7, 145-197.

[4] Granger, C. J. W., M. L. King and H. White (1995) 'Testing economic theories and the use of model selection criteria,' *Journal of Econometrics,* 67, 173-187.

[5] Grose, S. D. and M. L. King (1994) 'The use of information criteria for model selection between models with equal numbers of parameters,' paper presented at the 1994 Australasian Meeting of the Econometric Society.

[6] Hughes, A. and M. L. King (1994) 'One-sided model selection procedures,' paper presented at the 1994 Australasian Meeting of the Econometric Society.

[7] Imhof, P. J. (1961) 'Computing the distribution of quadratic forms in normal variables', *Biometrika,* 48, 419-426.

[8] Kalbfleish, J. and D. Sprott (1970) 'Application of likelihood methods to models involving large numbers of parameters' (with discussion), *Journal of the Royal Statistical Society B,* 32, 175-208.

[9] King, M. L. (1983) 'Testing for autoregressive against moving average errors in the linear regression model,' *Journal of Econometrics,* 21, 35-51.

[10] King, M. L. (1987) 'Towards a theory of point optimal testing,' *Econometric Reviews,* 6, 169-218.

[11] King, M. L. and D. E. A. Giles (1984) 'Autocorrelation pre-testing in the linear model: Estimation, testing and prediction,' *Journal of Econometrics,* 25, 35-48.

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

[13] Tunnicliffe Wilson, G. (1989) 'On the use of marginal likelihood in the time series model estimation' *Journal of the Royal Statistical Society B,* 51, 15-27.

[14] Wu, P. X. and M. L. King (1994) 'One-sided hypothesis testing in econometrics: A survey,' *Pakistan Journal of Statistics,* 10, 261-300.