OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 67, 1 (2005) 0305-9049

PRACTITIONERS' CORNER

A Note on the Selection of Time Series Models

SERENA NG* and PIERRE PERRON†

*Department of Economics, University of Michigan, Ann Arbor, MI, USA
(e-mail: serena.ng@umich.edu)
†Department of Economics, Boston University, Boston, MA, USA
(e-mail: perron@bu.edu)

Abstract

We consider issues related to the order of an autoregression selected using information criteria. We study the sensitivity of the estimated order to (i) whether the effective number of observations is held fixed when estimating models of different order, (ii) whether the estimate of the variance is adjusted for degrees of freedom, and (iii) how the penalty for overfitting is defined in relation to the total sample size. Simulations show that the lag length selected by both the Akaike and the Schwarz information criteria are sensitive to these parameters in finite samples. The methods that give the most precise estimates are those that hold the effective sample size fixed across models to be compared. Theoretical considerations reveal that this is indeed necessary for valid model comparisons. Guides to robust model selection are provided.

I. Motivation

Consider the regression model $y_t = x_t'\beta + e_t$ where x_t is a vector of p strictly exogenous regressors for $t = 1, \ldots, T$. If we were to determine the optimal number of regressors, we could set it to be the global minimizer of an information criterion (IC) such as:

$$\mathrm{IC}(i) = \ln \hat{\sigma}_i^2 + k_i \frac{C_T}{T}$$

JEL Classification number: C22.

where

$$\hat{\sigma}_i^2 = T^{-1} \sum_{t=1}^{T} \hat{e}_t^2$$

is an estimate of the regression error variance for the ith model, k_i is the number of regressors in that model, C_T/T is the penalty attached to an additional regressor, and T is the number of observations available. If p regressors were available, we have a total of 2^p models to consider. The problem is computationally burdensome, but for a given C_T, there is no ambiguity in how to set up the criterion function. The Akaike information criterion (AIC) is obtained when $C_T = 2$, and the Scharwz (Bayesian) information criterion (BIC), when $C_T = \ln T$. For any $T > \exp(2)$, the penalty imposed by the BIC is larger than that for the AIC. The IC is very general, and can be justified in a number of ways as we discuss below.

Time series data are correlated over time, and it is widely popular to capture the serial dependence in the data by autoregressive models. Suppose

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + e_t \tag{1}$$

is the data-generating process (DGP) with $e_t \sim$ i.i.d.$(0, \sigma^2)$. If p is finite, y_t is a finite-order AR(p) process. If y_t has moving-average components, p is infinite. We do not know p, and we cannot estimate an infinite number of parameters from a finite sample of T observations. Instead, we consider an autoregressive model of order k:

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + e_{tk}. \tag{2}$$

The adequacy of the approximate model for the DGP depends on the choice of k. Because the regressors in the autoregression are ordered by time, many of the 2^k permutations can be dismissed, and in this regard, the model selection problem in autoregressions is much simpler than the strictly exogenous regressors case. However, because lagged observations are required, the data available for the estimation of equation (2) are less than T. A regression that uses observations $n + 1$ to T would have an effective sample size of $N = T - n$. Therefore, unlike in the case of strictly exogenous regressors when the definitions of $\hat{\sigma}_k^2$, C_T, and T are unambiguous, the IC can be defined in a number of ways. Specifically, let $k_{\max}$ be the maximum number of lags deemed acceptable by a practitioner and consider

$$\min_{k=0,\ldots,k_{\max}} \mathrm{IC}(k) = \min_{k=0,\ldots,k_{\max}} \ln \hat{\sigma}_k^2(\tau) + k \frac{C_M}{M} \tag{3}$$

$$\hat{\sigma}_k^2(\tau) = \frac{1}{\tau} \sum_{t=n+1}^{T} \hat{e}_{tk}^2$$

where $\hat{e}_{tk}$ are the least squares residuals from estimation of equation (2). Although it would be tempting to exploit the largest sample possible and to

use an unbiased estimator of σ^2 in estimations, these choices may not be desirable from the point of view of model comparison.

This paper considers the sensitivity of the lag length selected by the AIC and the BIC to different choices for n, τ, and M. The latter affects the severity of the penalty. The former two determine how the goodness-of-fit is measured. We consider 10 variations of IC(k) based upon regressions estimated from $t = n + 1, \ldots, T$, with $N = T - n$:

$$\textit{Methods considered: } IC(k) = \ln[\tfrac{1}{\tau}\textstyle\sum_{t=n+1}^{T} \hat{e}_{tk}^2] + k\frac{C_M}{M}$$

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
N	$T-k_{\max}$	$T-k$	$T-k$	$T-k_{\max}$	$T-k_{\max}$	$T-k_{\max}$	$T-k$	$T-k$	$T-k_{\max}$	$T-k$
τ	$T-k_{\max}$	$T-k$	T	T	$T-k_{\max}-k$	$T-k_{\max}-k$	$T-2k$	$T-k$	$T-k_{\max}$	$T-k$
M	$T-k_{\max}$	$T-k$	T	T	$T-k_{\max}-k$	$T-k_{\max}$	$T-k$	T	$T-k_{\max}-k$	$T-2k$

Methods 1, 4, 5, 6, and 9 hold the effective number of observations fixed as k varies, namely, $N = T - k_{\max}$. Hence the difference in the sum of squared residuals between a model with k lags and one with $k - 1$ lags is purely the effect of adding the kth lag. Methods 2, 3, 7, 8, and 10 make maximum use of the data as a model with shorter lags will need fewer initial values and the regression uses observations $t = k + 1, \ldots, T$ with $N = T - k$. However, the sum of squared residuals between a model with k lags and one with $k - 1$ lags will differ not only because of the effect of adding the kth lag, but also because the smaller model is estimated with a larger effective sample size. Hayashi (2000) refers to these as cases of 'elastic' samples.

Apart from the degrees of freedom adjustment in the estimation of σ^2, methods 6, 7, and 8 are identical to methods 1, 2, and 3, respectively, in all other respects. Clearly, $\hat{\sigma}_k^2$ will be larger after degrees of freedom adjustment. Criteria that take this into account should be expected to choose a smaller model, all else equal. The penalty for all 10 methods converges to 0 at rate T, but in finite samples, $T - k_{\max} - k < T - k_{\max} < T - k < T$. Thus, of all the methods, method 5 puts the heaviest penalty on an extra lag and is expected to choose the most parsimonious model for a given C_M.

A quick review of textbooks produce no definitive guide. Priestley (1981; p. 373) seems to suggest method 2. His argument requires that N does not depend on k. This, however, is invalid as he also defined N as $T - k$. In a multivariate context, Lutkepohl (1993; p.129) defines the criteria in terms of the length of the time series, which could be T, $T - k$, or even $T - k_{\max}$. Diebold (1997; p. 26) uses the full length of the data, T, when defining the criteria. This is consistent with the notation of method 3. However, estimation with T observations is infeasible unless one initializes the first few lags to 0. The definition is therefore not useful

in practice. Hayashi (2000) noted several possibilities when implementing information criteria, but no particular recommendation was made. The software Eviews (1997), which is used to provide examples in many textbooks, presents an AIC and BIC individually for each k, which is consistent with method 2.[1] Enders (1995; p. 88) defines the criteria in terms of the number of usable observations, but pointed out that the sample size should be held fixed. This suggests method 1. In view of all these different recommended specifications, there is a need to clarify this issue. Hence, while no new theoretical results will be offered in this note, our simulations and theoretical overview will provide useful guides to practitioners.

II. Some theoretical considerations

This section considers the guidance provided by theory. The criteria considered are all based on large sample approximations, but in ways that imply specific choices of M, n and τ.

The Akaike information criterion

We first consider the derivation of the AIC for data generated by a finite-order AR(p) with normal errors. The regression model has k lags. If $k > p$, $\beta(k) = (\beta_1, \ldots, \beta_p, 0, \ldots, 0)'$ denote the true parameters, and $\hat{\beta}(k) = (\hat{\beta}_1, \ldots, \hat{\beta}_k)'$ are the estimated parameters. If $p > k$, $\beta(k) = (\beta_1, \ldots, \beta_p)'$ and $\hat{\beta}(k) = (\hat{\beta}_1, \ldots, \hat{\beta}_k, 0, \ldots, 0)'$. Following the treatment of Gourieroux and Monfort (1995, pp. 307–309), let $f(y|\beta(k))$ be the likelihood function of the data $(y_{n+1}, \ldots, y_T)$ conditional on the initial observations $(y_1, \ldots, y_n)$. Let $N = T - n$. The Kullback distance between the true probability distribution and the estimated parametric model is

$$K = E_0[\ln(f(y|\beta(k))) - \ln(f(y|\hat{\beta}(k)))]$$

with sample analog:

$$\tilde{K} = N^{-1} \sum_{t=n+1}^{T} \ln(f(y_t|\beta(k))) - N^{-1} \sum_{t=n+1}^{T} \ln(f(y_t|\hat{\beta}(k))).$$

Akaike's suggestion was to find a K^* such that

$$\lim_{T\to\infty} E[N(K - K^*)] = 0$$

so that K^* is unbiased for K to order N^{-1}. Let $X_t = (y_{t-1}, \ldots, y_{t-k})'$ and

[1]Correspondence with the Eviews support group confirms this to be the case.

$$\Phi_T(k) = (1/\hat{\sigma}_k^2)(\hat{\beta}(k) - \beta(k))' \sum_{t=n+1}^{T} X_t X_t' (\hat{\beta}(k) - \beta(k))$$

where

$$\hat{\sigma}_k^2 = N^{-1} \sum_{t=n+1}^{T} \hat{e}_{tk}^2.$$

Using Taylor series expansions, we have

$$NK = \Phi_T(k)/2 + o_p(1) \quad \text{and} \quad N\tilde{K} = -\Phi_T(k)/2 + o_p(1).$$

As

$$N(K - \tilde{K}) = \Phi_T(k) + o_p(1), \quad \lim_{N\to\infty} E[N(K - K^*)] = 0$$

for $K^* = \tilde{K} + \Phi_T(k)$. Furthermore, $\Phi_T(k)$ converges to a χ^2 random variable with k degrees of freedom. Hence a K^* that will satisfy

$$\lim_{T\to\infty} E[N(K - K^*)] = 0$$

is

$$K^* = N^{-1} \sum_{t=n+1}^{T} \ln(f(y_t|\beta(k))) - N^{-1} \sum_{t=n+1}^{T} \ln(f(y_t|\hat{\beta}(k))) + k. \qquad (4)$$

Under normality, the second term is proportional to $-(N/2)\ln(\hat{\sigma}_k^2)$. Thus, if the first term is common to all models, minimizing K^* with respect to k is equivalent to finding the minimizer of:

$$\text{AIC}(k) = \ln \hat{\sigma}_k^2 + \frac{2k}{N}. \qquad (5)$$

Note the two assumptions leading to equation (5). The first is the commonality of the first term in equation (4) to all models, which can be true only if n is held fixed across models to be considered. The second is the use of the maximum likelihood estimator of σ^2 in place of the second term of equation (4), implying $\tau = N$.

The C_p criterion

Let

$$Y = (y_1, y_2, \ldots, y_T)', \quad X_1 = (X_{11}\, X_{12}\, \ldots X_{1T})', \quad X_2 = (X_{21} \ldots X_{2T})',$$

with

$$X_t = (y_{t-1}, \ldots, y_{t-p})' = (X_{1t}\ X_{2t}),$$

where

$$X_{1t} = (y_{t-1}, \ldots, y_{t-k})', \quad X_{2t} = (y_{t-k-1}, \ldots, y_{t-p})'.$$

In what follows, it is understood that $X_{2t} = 0$ if $k \geq p$. Let $\beta = (\beta_1\ \beta_2)$, where the partition is also at the kth element. Suppose the true model is

$$Y = X_1\beta_1 + X_2\beta_2 + e, \quad \text{with } E(e_t^2) = \sigma^2$$

and we estimate the model $Y = X_1\beta_1 + e_k$. If X_1 and X_2 have the same number of observations in the time dimension, then

$$\hat{\beta}_1 - \beta_1 = (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'e.$$

Furthermore,

$$\hat{e}_k = M_1X_2\beta_2 + M_1e, \quad \text{where } M_1 = [I - X_1(X_1'X_1)^{-1}X_1'].$$

Then

$$E[\hat{e}_k'\hat{e}_k] = E[\tau\hat{\sigma}_k^2] = \beta_2'X_2'M_1X_2\beta_2 + \sigma^2\text{tr}(M_1) = \beta_2'X_2'M_1X_2\beta_2 + (N-k)\sigma^2.$$

The mean-squared prediction error of a model with k regressors is[2]

$$\begin{aligned} E[\text{mse}(X_1\hat{\beta}_1, X\beta)] &= E[(X_1\hat{\beta}_1 - X\beta)'(X_1\hat{\beta}_1 - X\beta)] \\ &= \sigma^2k + \beta_2'X_2'M_1X_2\beta_2 \\ &= \sigma^2k + E[\tau\hat{\sigma}_k^2] - (N-k)\sigma^2 \\ \therefore \frac{E[\text{mse}(X_1\hat{\beta}_1, X\beta)]}{\sigma^2} &= k + \frac{E[\tau\hat{\sigma}_k^2]}{\sigma^2} - (N-k) \\ &= \frac{E[\tau\hat{\sigma}_k^2]}{\sigma^2} + 2k - N. \end{aligned}$$

The C_p criterion of Mallows (1973) replaces σ^2 by a consistent estimate (say, $\hat{\sigma}^2$) that is the same across models to be compared giving

$$C_p = \frac{\tau\hat{\sigma}_k^2}{\hat{\sigma}^2} + (2k - N). \tag{6}$$

Lemma 1. If N is the same across models, then the C_p yields the same minimizer as

$$C_p^* = \frac{\tau\hat{\sigma}_k^2}{\hat{\sigma}^2} + 2k.$$

Furthermore, $\frac{1}{\tau}C_p^*$ yields the same minimizer as

$$SC_p^* = \ln\hat{\sigma}_k^2 + \frac{2k}{\tau}.$$

The first result is obvious. The second result follows by noting that for any $\hat{\sigma}^2$ that does not depend on k, the SC_p^* (scaled C_p^*) yields the same minimizer as

[2]The developments here follow Judge *et al.* (1980; p. 419).

$$\ln \hat{\sigma}_k^2 - \ln \hat{\sigma}^2 + \frac{2k}{\tau} = \ln(1 + \hat{\sigma}_k^2/\hat{\sigma}^2 - 1) + \frac{2k}{\tau} \approx \frac{\hat{\sigma}_k^2}{\hat{\sigma}^2} - 1 + \frac{2k}{\tau}.$$

But this is simply $\frac{1}{\tau} C_p^* - 1$, and hence has the same minimizer as $\frac{1}{\tau} C_p^*$. Note, however, that these derivations are valid only if X_1 and X_2 have the same number of observations. In our notation, this again suggests $N = T - k_{\max}$ with $\tau = M$, but does not prescribe a particular value for τ.

The FPE criterion

The final prediction error (FPE) criterion developed by Akaike (1969) is based on minimizing the one-step-ahead prediction error. For a model with k lags, define $\beta(k) = (\beta_1, \beta_2, \ldots, \beta_k)'$, and $X_t = (y_{t-1}, \ldots, y_{t-k})'$. Given a sample of T observations, the one-step-ahead mean-squared prediction error is

$$E(y_{T+1} - \hat{\beta}(k)' X_T)^2 = \sigma^2 + \sigma^2 E[(\hat{\beta}(k) - \beta(k))' X_T X_T' (\hat{\beta}(k) - \beta(k))].$$

Using the asymptotic approximation that

$$\sqrt{N}(\hat{\beta}(k) - \beta(k)) \sim N(0, \sigma^2 \Gamma_k^{-1}), \quad \text{where } \Gamma_k = E[X_T X_T'],$$

N times the second term reduces to the expectation a χ^2 random variable with k degrees of freedom, giving FPE $= \sigma^2(1 + k/N)$. The maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = N^{-1} \sum_{t=1}^{N} \hat{e}_{tk}^2,$$

and under normality, $N\hat{\sigma}_k^2/\sigma^2 \sim \chi_{N-k}^2$. As $E[N\hat{\sigma}_k^2/\sigma^2] = (N - k)$, using $\sigma^2 \approx N\hat{\sigma}_k^2/(N - k)$, the FPE can then be written as

$$\begin{aligned} \text{FPE} &= \hat{\sigma}_k^2 \frac{N + k}{N - k} \\ \ln \text{FPE} &\approx \ln \hat{\sigma}_k^2 + \ln\left(1 + \frac{N + k - (N - k)}{N - k}\right) \\ &\approx \ln \hat{\sigma}_k^2 + \frac{2k}{N - k}. \end{aligned}$$

In our notation, this criterion also prescribes $N = T - k_{\max}$, but specifies $\tau = N$ and $M = N - k$.

Posterior probability

To develop the arguments for the BIC as defined in Schwarz (1978), we follow Chow (1983). Let $f(y|k)$ be the marginal probability density function for the data under a kth order model, $f(k)$ be the prior density for a kth order

model, and $f(y)$ be the marginal density of the data. Given observations $y = (y_{n+1}, \ldots, y_T)$, the posterior probability of a kth order model is $f(k|y) = f(k)f(y|k)/f(y)$. If $f(y)$ and $f(k)$ are the same for all k, then maximizing $f(k|y)$ is equivalent to maximizing $f(y|k)$. To evaluate $f(y|k)$, we use the fact that the log posterior density of β in a kth order model is

$$\ln f(\beta(k)|y,k) = \ln f(y,\beta(k)) + \ln f(\beta(k)|k) - \ln f(y|k)$$

where $f(y,\ \beta(k))$ is the likelihood function for the kth order model with parameters $\beta(k)$. But it is also known that under regularity conditions, the posterior distribution of $\beta(k)$ is Gaussian with inverse variance S; i.e.

$$f(\beta(k)|y,k) = (2\pi)^{-k/2}|S|^{1/2}\exp[-\tfrac{1}{2}(\beta(k) - \hat{\beta}(k))')S(\beta(k) - \hat{\beta}(k))] \times (1 + O(N^{-1/2})).$$

Now evaluate the posterior distributions around the maximum likelihood estimator, $\hat{\beta}(k)$, and approximate $\ln(S)$ by $k\ln(N) + \ln(R_N)$, where

$$R_N = \frac{1}{N}\sum_{t=n+1}^{T} X_t X_t'.$$

After rearranging terms, we have:

$$\ln f(y|k) \approx \ln f(y,\hat{\beta}(k)) - \frac{k}{2}\ln(N) - \frac{1}{2}\ln R_N + \ln f(\hat{\beta}(k)|k) + \frac{k\ln(2\pi)}{2} + O_p(N^{-1/2}). \tag{7}$$

If we use the first two terms of equation (7), the usual approximation for exponential families, we have

$$\ln f(y|k) \approx \ln f(y,\hat{\beta}(k)) - \frac{k}{2}\ln N.$$

Now the first term is proportional to $(-N/2)\ln(\hat{\sigma}_k^2)$, where

$$\hat{\sigma}_k^2 = N^{-1}\sum_{t=n+1}^{T}\hat{e}_{tk}^2.$$

Multiplying by $-(2/N)$, the k that maximizes the posterior of the data also minimizes:

$$\mathrm{BIC}(k) = \ln\hat{\sigma}_k^2 + \frac{k\ln N}{N}. \tag{8}$$

Three assumptions are used to derive equation (7). The first is that the prior is the same for all models, but this does not depend on n or τ. The second is that $f(y)$ and R_N are the same across models, which in turn requires that $n = k_{\max}$

(or $N = T - k_{\max}$), the same as for the AIC. The third is that log-likelihood function evaluated at the estimated parameters is proportional to $\hat{\sigma}_k^2$. These are the same assumptions underlying the AIC, i.e. $\tau = M = N$.

Overview

To relate the 10 methods to the theoretical discussions, the AIC and BIC both require $M = N$, and ln $\hat{\sigma}_k^2$ to be the maximum likelihood estimator with $\tau = N$, and both hold n (and thus N) fixed across models. Allowing for lagged observations, the largest sample in which n can be held fixed is to set $n = k_{\max}$. Taking all conditions into account, only method 1 satisfies all these conditions. Note that adjusting τ for degrees of freedom would be incompatible with the AIC or the BIC.

When N does not depend on k and $M = \tau$, the IC can be seen as an SC_p^* with $C_M = 2$. This includes methods 1, 4, and 5. The ln FPE is obtained by letting $\tau = N$ and $M = N - k$. Thus, methods 9 and 10 are consistent with the theoretical underpinnings of the ln FPE. Of the 10 methods considered, methods 2, 3, 6, 7, 8 bear no immediate relation to well-known criteria in the literature.

III. Simulations

To assess the empirical properties of the 10 methods considered, we simulate data from 25 time series processes detailed in Table 1. The first 12 are simple finite-order AR models. But information criteria are often used in cases when the true model is of higher order. For example, a stationary and invertible autoregressive moving average (ARMA) process has an infinite autoregressive representation. We do not consider such models in the simulations because the true value of p is not admissible by design. Instead, we start with an ARMA(1,1) model, $(1 - \phi L)y_t = (1 + \theta L)e_t$, whose infinite autoregressive representation is

$$\sum_{i=0}^{\infty}(\phi + \theta)(-\theta)^i y_{t-i} = e_t.$$

We then consider a truncated version of it. Specifically, case 13 to case 20 are finite-order autoregressive processes with p coefficients identical to the first p terms in the infinite autoregressive representations of ARMA(1,1) processes, where the truncation point p is chosen such that $|\beta_{p+1}| < 0.1$. The parameterizations allow us to assess situations when the autoregressive coefficients decline at a geometric rate. We also consider five cases (21–25) with ARCH errors. In these cases, we estimate autoregressions in y_t^2 so the IC is used to select the order of ARCH processes.

TABLE 1

DGP for models 1–20; $y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + e_t$, $e_t \sim$ *i.i.d. N(0, 1),* $y_0 = y_{-1} = \cdots = y_{-p} = 0$

DGP	p	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	θ	ϕ
1	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	1	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	1	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	1	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	1	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	1	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	2	0.40	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	2	1.10	−0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	2	1.30	−0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	3	0.30	0.20	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	3	0.10	0.20	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	4	0.20	−0.50	0.40	0.50	0.00	0.00	0.00	0.00	0.00	0.00
13	8	1.20	−0.96	0.77	−0.61	0.49	−0.39	0.31	−0.25	0.80	0.40
14	2	1.00	−0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.80
15	4	1.30	−0.65	0.33	−0.16	0.00	0.00	0.00	0.00	0.50	0.80
16	2	0.60	0.18	0.00	0.00	0.00	0.00	0.00	0.00	−0.30	0.90
17	2	0.55	0.22	0.00	0.00	0.00	0.00	0.00	0.00	−0.40	0.95
18	3	0.50	−0.25	0.13	0.00	0.00	0.00	0.00	0.00	0.50	0.00
19	8	0.80	−0.64	0.51	−0.41	0.33	−0.26	0.21	−0.17	0.80	0.00
20	2	−0.40	−0.16	0.00	0.00	0.00	0.00	0.00	0.00	−0.40	0.00

DGP	p	β_1	β_2	β_3	β_4
DGP for models 21–25; $y_t = \sqrt{h_t} e_t$, $e_t \sim$ *i.i.d. N(0, 1),* $h_t^2 = 1 + \beta_1 h_{t-1}^2 + \cdots + \beta_m h_{t-m}^2$					
21	0	0.50	0.00	0.00	0.00
22	0	0.80	0.00	0.00	0.00
23	0	0.50	0.40	0.00	0.00
24	0	0.20	0.30	0.40	0.00
25	0	0.20	0.30	0.40	0.05

Notes: For models 13–20, $\beta_i = (\phi + \theta)(-\theta)^i$ subject to the constraint that $|\beta_i| > 0.1$.

Simulations were performed using Gauss for $T = 100$, 250, 500, and 1,000 with $k_{\max}$ set to int$[10(T/100)^{1/4}]$. We report only the tabulated results for $T = 100$. Results for other sample sizes will be summarized below. Table 2 reports the average k selected by the AIC and BIC over 5,000 simulations, Table 3 reports the probability of selecting the true model, while Table 4 reports the standard errors. Each row gives results for the 10 methods considered. For example, the entries in row 5 of Table 3 are the probabilities that methods 1–10 select the correct model when the DGP is model 5 as described in Table 1.

The AIC variant of method 3 selects the correct model with probability 0.19, while method 9 achieves a probability of 0.73. Differences between the AIC and the BIC in DGPs 1 to 20 are along the lines documented in the

TABLE 2

Average k selected

DGP	*p/model*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Variant of AIC											
1	0	0.87	1.50	5.36	1.20	0.20	0.23	0.40	1.85	0.69	1.22
2	1	1.58	2.17	5.74	1.92	0.84	0.88	1.05	2.51	1.39	1.90
3	1	1.83	2.39	5.81	2.14	1.18	1.21	1.38	2.72	1.61	2.11
4	1	1.85	2.42	5.81	2.16	1.19	1.21	1.39	2.74	1.65	2.12
5	1	1.86	2.42	5.86	2.15	1.19	1.22	1.40	2.73	1.65	2.14
6	1	1.87	2.43	5.88	2.18	1.19	1.22	1.40	2.77	1.64	2.15
7	2	2.35	2.87	6.09	2.64	1.57	1.62	1.81	3.22	2.12	2.53
8	2	2.78	3.25	6.25	3.04	2.13	2.16	2.37	3.57	2.56	2.97
9	2	2.81	3.29	6.27	3.04	2.13	2.18	2.37	3.61	2.57	3.00
10	3	2.59	3.19	6.28	2.92	1.75	1.81	2.03	3.56	2.37	2.87
11	3	3.38	3.90	6.63	3.69	2.39	2.46	2.73	4.23	3.12	3.59
12	4	4.73	5.16	7.25	4.95	4.16	4.20	4.41	5.48	4.49	4.87
13	8	7.04	7.12	8.65	7.36	5.08	5.50	5.66	7.51	6.53	6.67
14	2	2.51	3.02	6.17	2.79	1.77	1.82	2.00	3.35	2.30	2.71
15	4	3.76	4.24	6.85	4.07	2.82	2.90	3.12	4.59	3.50	3.91
16	2	2.28	2.80	6.07	2.57	1.51	1.55	1.75	3.16	2.06	2.46
17	2	2.42	2.94	6.13	2.73	1.67	1.71	1.91	3.31	2.21	2.63
18	3	2.79	3.35	6.31	3.11	1.95	2.00	2.20	3.72	2.54	3.01
19	8	5.83	6.03	8.20	6.23	3.81	4.12	4.36	6.49	5.27	5.51
20	2	2.35	2.87	6.03	2.66	1.60	1.63	1.83	3.22	2.13	2.56
21	0	1.16	1.75	4.74	1.47	0.43	0.45	0.78	2.02	0.98	1.52
22	0	1.48	2.07	4.45	1.77	0.71	0.74	1.12	2.30	1.31	1.85
23	0	1.98	2.05	4.57	2.35	0.91	0.97	1.08	2.30	1.75	1.85
24	0	2.14	1.74	4.52	2.53	0.86	0.92	0.75	1.96	1.84	1.49
25	0	2.27	1.55	4.20	2.72	0.92	1.00	0.68	1.76	1.96	1.35
Variant of BIC											
1	0	0.06	0.12	0.31	0.07	0.03	0.03	4.47	0.13	0.05	0.11
2	1	0.58	0.67	0.95	0.63	0.46	0.46	4.80	0.69	0.57	0.65
3	1	1.05	1.11	1.32	1.07	1.01	1.01	4.90	1.12	1.04	1.09
4	1	1.05	1.10	1.31	1.07	1.02	1.02	4.93	1.12	1.04	1.09
5	1	1.05	1.11	1.31	1.07	1.02	1.02	4.94	1.13	1.05	1.10
6	1	1.06	1.11	1.32	1.08	1.02	1.02	5.00	1.13	1.05	1.10
7	2	1.31	1.42	1.73	1.37	1.19	1.20	5.23	1.45	1.30	1.40
8	2	1.97	2.03	2.31	2.00	1.88	1.89	5.40	2.06	1.95	2.02
9	2	1.96	2.04	2.32	1.99	1.88	1.89	5.42	2.08	1.94	2.02
10	3	1.38	1.51	1.93	1.46	1.16	1.18	5.42	1.55	1.34	1.46
11	3	1.75	1.95	2.60	1.91	1.29	1.34	5.92	2.01	1.69	1.89
12	4	4.04	4.11	4.40	4.07	3.98	3.99	6.60	4.17	4.02	4.07
13	8	3.95	4.09	5.69	4.34	3.01	3.19	8.29	4.48	3.60	3.74
14	2	1.51	1.59	1.92	1.56	1.37	1.39	5.29	1.62	1.49	1.57
15	4	2.43	2.54	3.06	2.52	2.20	2.23	6.15	2.62	2.38	2.46
16	2	1.28	1.37	1.66	1.32	1.19	1.20	5.22	1.40	1.26	1.35
17	2	1.41	1.51	1.82	1.46	1.29	1.30	5.31	1.54	1.39	1.48

TABLE 2
(continued)

DGP	*p/model*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
18	3	1.59	1.70	2.12	1.66	1.37	1.39	5.54	1.75	1.56	1.66
19	8	2.91	3.10	4.36	3.17	2.34	2.42	7.66	3.34	2.73	2.90
20	2	1.30	1.39	1.73	1.36	1.14	1.15	5.23	1.42	1.28	1.36
21	0	0.22	0.37	0.68	0.24	0.14	0.14	4.05	0.40	0.21	0.34
22	0	0.42	0.64	1.00	0.47	0.29	0.30	3.87	0.69	0.40	0.60
23	0	0.49	0.55	0.98	0.57	0.34	0.35	4.02	0.62	0.47	0.51
24	0	0.39	0.33	0.65	0.48	0.23	0.25	3.82	0.36	0.36	0.30
25	0	0.42	0.31	0.58	0.51	0.25	0.26	3.53	0.33	0.38	0.29

literature. On average, the AIC overparameterizes low-order AR models, while the BIC abandons information at lags shorter than p more often. For example, for DGP 19 with model 8, the AIC truncates at six lags on average with $\beta_6 = 0.26$. The BIC, on the other hand, truncates at three lags with $\beta_3 = 0.51$.[3] For some models, the AIC and the BIC have low probabilities of selecting the correct model. As discussed in Hendry and Krolzig (2002), although a model selection criterion may have good asymptotic properties under specific assumptions, one cannot assume that it always has good finite sample properties.

Our main interest is in the sensitivity of the methods with respect to N, τ, and M. Of the three parameters, the estimates are most robust to variations in M. Changing M from $T - k$ (method 2) to T (method 8) or to $T - 2k$ (method 10) apparently makes only small differences. The AIC is especially sensitive to whether or not N is held fixed. Method 3, for example, with $N = T - k$ provides estimates that are both mean and median biased. But for the same τ, method 4 with $N = T - k_{\max}$ is more precise although it uses fewer observations in estimating models with $k < k_{\max}$ lags. Furthermore, changing τ from T (method 3) to $T - k$ (method 8) can yield sharp changes in the estimates if we do not hold N fixed. Although the BIC is generally more robust to different choices of N, differences between methods remain apparent. Method 7 overestimates p in much the same way method 3 does under the AIC, and the BIC estimates are in this case also mean and median biased. Interestingly, method 7 works well under the AIC but not the BIC, implying that how N, τ, and M affects the IC also depends on the choice of C_M.

The simulation results thus show that the properties of the criteria can differ quite substantially across methods especially with respect to whether N depends on k. To further understand this, recall that the basis of the IC is to trade-off good fit against parsimony. Let

[3]For $T = 250$ and higher (not reported), the k chosen by the BIC is still small.

TABLE 3

Probability that $\hat{k} = p$

DGP	*p/model*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Variant of AIC											
1	0	0.70	0.57	0.19	0.64	0.88	0.88	0.81	0.55	0.73	0.60
2	1	0.55	0.46	0.17	0.51	0.59	0.59	0.56	0.43	0.57	0.48
3	1	0.70	0.58	0.19	0.63	0.88	0.87	0.81	0.54	0.73	0.62
4	1	0.70	0.58	0.19	0.63	0.89	0.88	0.81	0.54	0.73	0.61
5	1	0.70	0.58	0.19	0.64	0.88	0.87	0.81	0.54	0.73	0.61
6	1	0.70	0.58	0.19	0.64	0.88	0.87	0.81	0.54	0.74	0.61
7	2	0.41	0.36	0.15	0.40	0.39	0.39	0.38	0.34	0.42	0.39
8	2	0.69	0.58	0.20	0.64	0.84	0.83	0.77	0.54	0.73	0.62
9	2	0.68	0.58	0.20	0.64	0.84	0.83	0.77	0.53	0.72	0.62
10	3	0.16	0.17	0.10	0.17	0.12	0.12	0.14	0.16	0.17	0.17
11	3	0.57	0.47	0.19	0.54	0.59	0.59	0.55	0.44	0.60	0.51
12	4	0.70	0.58	0.21	0.65	0.88	0.87	0.79	0.52	0.76	0.64
13	8	0.47	0.40	0.34	0.49	0.24	0.30	0.27	0.40	0.43	0.38
14	2	0.51	0.44	0.17	0.48	0.53	0.53	0.50	0.41	0.53	0.47
15	4	0.35	0.31	0.16	0.35	0.27	0.28	0.27	0.28	0.36	0.32
16	2	0.36	0.32	0.14	0.35	0.32	0.33	0.33	0.30	0.37	0.34
17	2	0.46	0.41	0.16	0.44	0.46	0.46	0.44	0.38	0.47	0.43
18	3	0.23	0.21	0.11	0.23	0.17	0.17	0.18	0.20	0.23	0.22
19	8	0.26	0.22	0.26	0.28	0.07	0.10	0.10	0.23	0.21	0.19
20	2	0.43	0.37	0.16	0.41	0.41	0.41	0.40	0.34	0.45	0.39
21	0	0.53	0.54	0.23	0.48	0.72	0.71	0.72	0.51	0.55	0.56
22	0	0.41	0.44	0.21	0.37	0.58	0.58	0.60	0.42	0.43	0.45
23	0	0.38	0.47	0.23	0.33	0.58	0.57	0.65	0.45	0.40	0.49
24	0	0.42	0.57	0.27	0.37	0.65	0.64	0.74	0.55	0.45	0.59
25	0	0.41	0.58	0.29	0.35	0.64	0.63	0.74	0.56	0.43	0.59
Variant of BIC											
1	0	0.96	0.92	0.84	0.95	0.98	0.98	0.23	0.92	0.96	0.92
2	1	0.50	0.52	0.55	0.53	0.43	0.43	0.22	0.52	0.50	0.52
3	1	0.95	0.92	0.83	0.94	0.96	0.96	0.25	0.91	0.96	0.92
4	1	0.96	0.92	0.84	0.95	0.98	0.98	0.25	0.92	0.96	0.93
5	1	0.96	0.92	0.84	0.95	0.98	0.98	0.25	0.91	0.96	0.93
6	1	0.95	0.92	0.84	0.94	0.98	0.98	0.24	0.91	0.96	0.93
7	2	0.29	0.32	0.37	0.31	0.21	0.22	0.19	0.32	0.28	0.31
8	2	0.86	0.85	0.79	0.86	0.84	0.84	0.27	0.84	0.86	0.85
9	2	0.86	0.84	0.78	0.86	0.83	0.83	0.27	0.83	0.86	0.84
10	3	0.06	0.08	0.13	0.07	0.02	0.03	0.13	0.08	0.05	0.07
11	3	0.45	0.48	0.54	0.49	0.33	0.34	0.25	0.48	0.44	0.47
12	4	0.93	0.90	0.79	0.92	0.94	0.94	0.30	0.88	0.94	0.91
13	8	0.11	0.11	0.27	0.15	0.02	0.03	0.40	0.15	0.06	0.07
14	2	0.44	0.45	0.49	0.46	0.35	0.36	0.23	0.46	0.43	0.44
15	4	0.16	0.17	0.26	0.18	0.08	0.09	0.21	0.19	0.14	0.16
16	2	0.23	0.27	0.32	0.26	0.17	0.18	0.18	0.27	0.23	0.26
17	2	0.35	0.38	0.43	0.39	0.27	0.28	0.21	0.38	0.34	0.37

TABLE 3

(continued)

DGP	*p/model*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
18	3	0.09	0.12	0.17	0.11	0.05	0.05	0.15	0.12	0.09	0.11
19	8	0.02	0.03	0.11	0.03	0.00	0.00	0.28	0.04	0.01	0.02
20	2	0.31	0.33	0.38	0.33	0.23	0.23	0.20	0.33	0.30	0.33
21	0	0.83	0.83	0.75	0.81	0.88	0.88	0.27	0.83	0.83	0.84
22	0	0.71	0.72	0.64	0.69	0.78	0.78	0.24	0.72	0.72	0.73
23	0	0.72	0.76	0.68	0.70	0.79	0.79	0.26	0.75	0.73	0.77
24	0	0.80	0.85	0.77	0.77	0.87	0.87	0.31	0.85	0.81	0.86
25	0	0.80	0.85	0.77	0.77	0.86	0.86	0.33	0.84	0.80	0.85

$$RSS_k = \sum_{t=n+1}^{T} \hat{e}_{tk}^2,$$

so that $\hat{\sigma}_k^2 = RSS_k/\tau$. Then

$$\mathrm{IC}(k) = \ln(RSS_k) - \ln(\tau) + kC_M/M. \tag{9}$$

Two observations can be made. First, the well-known result in least squares regression that RSS_k is non-increasing in k pre-supposes that the sample size is held fixed as k increases. This is not necessarily the case when the sample size is elastic. Secondly, if τ depends on k, then $kC_M/M - \ln(\tau)$ can be seen as the effective penalty for k regressors. The penalty becomes non-linear in k in ways that depend on both M and τ. The two considerations together imply that there could exist choices of τ, M, and N such that the IC bears unpredictable relations with k, and in consequence, produce unstable choices of p. Method 3 under the AIC and method 7 under the BIC appear to be such cases, as seen from the standard errors reported in Table 4.

Equation (9) makes clear that the effective penalty for model comparison is the term $kC_M/M - \ln(\tau)$, which depends on C_M. A method that works for the AIC with constant C_M may not work for the BIC that allows C_M to vary. Indeed, such is the case with method 7. To the extent that the penalty reflects our preference for parsimony, there is no unique choice for M and τ. One can nonetheless ensure that the penalty moves with k in the most predictable way possible, and in this regard, letting M and τ to be invariant to k is desirable. This, however, is of secondary importance relative to fixing N, as by ensuring that RSS_k is indeed non-increasing in k, we also ensure that the goodness-of-fit of two models are properly compared. Holding N fixed in model comparisons is theoretically desirable and is recommended in applications.

To better highlight the fact that holding N fixed is desirable in model selection and that a method that works well for the AIC need not work well for the BIC, we now consider a response surface analysis based on the 20 DGPs

TABLE 4

Standard error of $\hat{k}$

DGP	*p/model*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Variant of AIC											
1	0	1.84	2.48	3.82	2.23	0.70	0.79	1.14	2.85	1.51	2.11
2	1	1.76	2.34	3.50	2.11	0.80	0.86	1.16	2.69	1.46	2.02
3	1	1.73	2.27	3.43	2.06	0.62	0.71	1.05	2.61	1.36	1.93
4	1	1.73	2.30	3.44	2.08	0.64	0.72	1.09	2.61	1.41	1.95
5	1	1.78	2.30	3.43	2.07	0.66	0.72	1.10	2.61	1.42	1.96
6	1	1.80	2.32	3.43	2.12	0.66	0.74	1.09	2.65	1.42	1.98
7	2	1.80	2.25	3.19	2.05	0.80	0.90	1.20	2.56	1.48	1.83
8	2	1.66	2.09	3.06	1.92	0.64	0.73	1.12	2.37	1.33	1.76
9	2	1.69	2.12	3.06	1.93	0.66	0.78	1.12	2.40	1.35	1.80
10	3	1.84	2.30	3.10	2.11	1.00	1.07	1.36	2.59	1.56	1.98
11	3	1.85	2.19	2.80	2.03	1.39	1.42	1.62	2.42	1.57	1.91
12	4	1.43	1.78	2.31	1.63	0.60	0.69	1.04	1.99	1.12	1.50
13	8	2.00	2.09	1.40	1.88	2.13	2.22	2.31	2.01	2.07	2.15
14	2	1.76	2.22	3.13	2.00	0.82	0.89	1.22	2.50	1.44	1.87
15	4	1.78	2.15	2.60	1.98	1.08	1.16	1.45	2.36	1.54	1.91
16	2	1.80	2.27	3.22	2.07	0.80	0.87	1.18	2.59	1.49	1.87
17	2	1.75	2.22	3.16	2.04	0.82	0.89	1.19	2.54	1.48	1.84
18	3	1.79	2.21	3.01	2.04	0.96	1.03	1.31	2.51	1.47	1.91
19	8	2.31	2.40	1.87	2.28	1.81	2.00	2.18	2.44	2.22	2.31
20	2	1.76	2.20	3.21	2.05	0.83	0.89	1.22	2.52	1.43	1.86
21	0	1.88	2.71	3.85	2.20	0.86	0.90	1.80	2.99	1.57	2.47
22	0	1.95	2.85	3.71	2.24	1.10	1.15	2.10	3.05	1.71	2.62
23	0	2.36	2.89	3.80	2.58	1.42	1.50	2.17	3.10	2.12	2.71
24	0	2.64	2.80	3.91	2.85	1.53	1.63	1.78	3.02	2.33	2.51
25	0	2.76	2.58	3.88	2.98	1.62	1.76	1.63	2.82	2.46	2.33
Variant of BIC											
1	0	0.29	0.49	0.98	0.34	0.19	0.20	3.68	0.54	0.28	0.44
2	1	0.60	0.69	1.08	0.62	0.54	0.54	3.37	0.74	0.59	0.66
3	1	0.30	0.45	0.95	0.36	0.21	0.23	3.31	0.51	0.28	0.39
4	1	0.25	0.42	0.96	0.36	0.16	0.17	3.31	0.52	0.24	0.38
5	1	0.26	0.44	0.96	0.35	0.16	0.17	3.32	0.50	0.24	0.39
6	1	0.29	0.43	0.99	0.37	0.16	0.17	3.32	0.51	0.26	0.38
7	2	0.57	0.69	1.12	0.63	0.50	0.51	3.10	0.75	0.56	0.66
8	2	0.44	0.53	1.06	0.47	0.40	0.41	2.96	0.63	0.42	0.50
9	2	0.44	0.56	1.07	0.46	0.41	0.42	2.97	0.63	0.41	0.52
10	3	0.79	0.91	1.31	0.84	0.71	0.72	3.01	0.95	0.76	0.84
11	3	1.42	1.45	1.63	1.42	1.36	1.38	2.73	1.48	1.40	1.44
12	4	0.37	0.54	1.05	0.42	0.29	0.31	2.27	0.65	0.32	0.43
13	8	1.85	1.93	2.36	2.02	1.18	1.38	1.55	2.14	1.58	1.68
14	2	0.60	0.70	1.16	0.64	0.51	0.52	3.04	0.74	0.58	0.67
15	4	0.85	0.98	1.43	0.91	0.64	0.67	2.58	1.06	0.79	0.87
16	2	0.52	0.65	1.11	0.57	0.41	0.42	3.12	0.70	0.49	0.61
17	2	0.57	0.69	1.14	0.61	0.48	0.49	3.07	0.75	0.55	0.65

TABLE 4

(continued)

DGP	*p/model*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
18	3	0.75	0.88	1.29	0.79	0.66	0.68	2.93	0.93	0.74	0.84
19	8	1.35	1.51	2.24	1.54	0.94	1.04	2.05	1.74	1.17	1.32
20	2	0.64	0.71	1.15	0.66	0.60	0.61	3.09	0.77	0.63	0.67
21	0	0.54	1.19	1.70	0.58	0.41	0.42	3.67	1.27	0.53	1.09
22	0	0.79	1.56	2.02	0.86	0.63	0.64	3.53	1.65	0.76	1.47
23	0	0.99	1.45	2.09	1.10	0.78	0.80	3.62	1.60	0.94	1.36
24	0	0.96	1.10	1.68	1.11	0.73	0.76	3.69	1.19	0.91	1.00
25	0	1.03	0.97	1.51	1.16	0.76	0.78	3.62	1.03	0.94	0.91

with no ARCH effects. Simulations were carried out for $T = 100$, 250, 500, and 1,000 thus yielding 800 observations (10 methods, 20 DGPs and four sample sizes). The dependent variables are (i) the log odds of the true order (i.e. $\ln(f_i/(1 - f_i))$ where f_i is the simulated probability of selecting the correct order for the ith specification), and (ii) the log mean-squared error (MSE). The regressors are based on the following dummy variables: N_1, a dummy variable that takes the value 1 if $N = T - k_{\max}$ (0 otherwise); N_2, a dummy variable that takes the value 1 if $N = T - k$ (0 otherwise); τ_1, a dummy variable that takes the value 1 if $\tau = N$ (0 otherwise), M_1, a dummy variable that takes the value 1 if $M = N$ (0 otherwise). After an extensive search, we settled on the following regressors in all four cases (AIC and BIC for both dependent variables): a constant, N_1, $N_1 \times \tau_1$, $N_2 \times \tau_1$, $N_1 \times M_1$, $N_2 \times M_1$, $M_1 \times \tau_1$, p, p^2, T^{-1}, T^{-2} and p/T. It is important to note that our aim is to show how using different specifications for N, M and τ affects the probability of selecting the correct model and the MSE of the estimate of the autoregressive order. Also, some of our designs do not ensure a consistent estimate even for BIC (e.g. method 7 for which the probability of selecting the correct order is at most 20% even when $T = 1{,}000$). Hence, we cannot impose restrictions such that the regression function implies that the probability of selecting the correct model goes to 1 or that the MSE goes to 0 as the sample size increases. Nevertheless, the specification used allows us to highlight useful conclusions.

The results are reported in Table 5. Evidently, for both the AIC and the BIC, $P(\hat{k} = p)$ rises and the MSE falls when $N = T - k_{\max}$, but the effect is much larger for the AIC. Whereas having $\tau = N$ or $M = N$ when $N = T - k_{\max}$ is desirable for the AIC, this is not the case for the BIC (see the coefficients on $N_1 \times \tau_1$ and $N_1 \times M_1$, respectively). The effect of setting $M = \tau = N$ when $N \neq T - k_{\max}$ is an increase in the log of MSE of the AIC by 0.849. But if $N = T - k_{\max}$, the net effect is a reduction in log MSE, as $-1.000 + 0.849 < 0$. For the BIC, having $\tau = M = T - k_{\max}$ yields a

TABLE 5

Response surface analysis

	AIC		*BIC*	
Regressor	$log\frac{P(\hat{k}=p)}{1-P(\hat{k}=p)}$	*log(MSE)*	$log\frac{P(\hat{k}=p)}{1-P(\hat{k}=p)}$	*log(MSE)*
N_1	1.944 (25.53)	−1.000 (23.44)	0.809 (4.31)	−1.142 (18.31)
$N_1 \times \tau_1$	0.111 (1.53)	0.169 (4.62)	−0.467 (2.09)	0.484 (6.89)
$N_2 \times \tau_1$	1.120 (17.34)	−0.317 (10.52)	0.544 (3.13)	−0.787 (13.89)
$N_1 \times M_1$	0.794 (8.69)	−0.572 (13.59)	−0.434 (1.82)	0.341 (4.42)
$N_2 \times M_1$	1.753 (21.79)	−0.981 (30.33)	−1.521 (7.66)	1.342 (14.76)
$M_1 \times \tau_1$	−1.363 (15.59)	0.849 (24.10)	1.070 (4.27)	−0.904 (10.06)
p	−0.201 (4.70)	−0.043 (2.06)	−1.247 (10.90	0.243 (6.03)
$1/T$	−21.441 (0.68)	−121.14 (6.84)	806.18 (8.44)	429.59 (12.24)
p^2	0.028 (6.27)	−0.001 (0.36)	0.099 (7.62)	−0.010 (2.30)
$1/T^2$	−1,579.2 (0.59)	8,299.1 (5.87)	−76,810.0 (9.49)	−27,811.0 (9.67)
p/T	−25.295 (6.19)	7.444 (3.86)	−16.42 (1.78)	2.483 (0.74)
C	−0.471 (4.00)	1.535 (25.90)	1.894 (7.11)	−1.856 (18.02)
n	800	800	800	800
R^2	0.66	0.71	0.43	0.72

Notes: $N_1 = 1$ if $N = T - k_{\max}$; $N_2 = 1$ if $N = T - k$; $\tau_1 = 1$ if $\tau = N$; $M_1 = 1$ if $M = N$. Robust *t*-statistics in parenthesis. The response surface is performed using results from DGPs 1 to 20 with T = 100, 250, 500, and 1,000, giving 800 observations in the response surface analysis.

reduction in MSE of −2.042, with the largest gain coming from setting $M = \tau$. These results are consistent with our casual observation that the model selection criteria are better behaved when $N = T - k_{\max}$.

We also rank the methods by the average MSE and by the probability of selecting the true model. The results are reported in Table 6. Rankings are reported for all models (column 1), models 1–12 (column 2), models 13–20 (column 3), models 1–20 (column 4), and models 21–25 (column 5). These groupings are chosen to highlight the fact that the AIC and BIC are better suited for different data types. For low-order AR models, methods 5 and 6 are best for the AIC, while 1, 4, and 9 are best for the BIC. Although in theory, the AIC does not have the property that $\lim_{T\to\infty} P(\hat{k} = p) = 1$ when p is finite, for the models being considered, the AIC apparently performs quite well overall. Differences between the AIC and the BIC are more marked in models 13–20. In such cases, the AIC performs noticeably better especially when methods 1, 4 and 9 are used.[4] Whether one uses the AIC or the BIC in selecting the order of ARCH processes, methods 5 and 6 are clearly the best. A feature common to methods 1, 4, 5, 6 and 9 is $N = T - k_{\max}$. Holding the sample size fixed is thus crucial in model comparisons.

[4]When p is infinite and assuming Gaussian errors, Shibata (1980) showed that the AIC achieves an asymptotic lower bound of the mean squared prediction errors.

TABLE 6

Rankings of the 10 variants of the AIC and the BIC

Variant	*DGP*									
MSE	*All*		*1–12*		*13–20*		*1–20*		*21–25*	
AIC										
1	2.53	6	0.83	5	4.06	9	2.48	6	2.41	5
2	2.54	5	0.94	6	4.32	1	2.58	5	2.74	6
3	3.25	7	1.59	7	4.80	6	2.95	7	4.45	7
4	3.79	9	2.32	9	4.86	4	3.02	9	6.87	9
5	4.88	1	3.56	1	4.99	7	3.87	1	8.93	1
6	5.59	10	4.50	10	5.06	10	4.72	10	9.03	10
7	6.33	4	5.10	4	5.21	5	5.01	4	11.06	2
8	7.22	2	6.47	2	5.96	2	6.26	2	11.63	4
9	9.18	8	8.80	8	7.18	8	8.15	8	13.30	8
10	27.24	3	29.83	3	18.57	3	25.33	3	34.90	3
$P(\hat{k} = p)$	*All*		*1–12*		*13–20*		*1–20*		*21–25*	
1	0.57	5	0.72	5	0.38	1	0.56	6	0.69	7
2	0.57	6	0.72	6	0.38	9	0.56	5	0.63	5
3	0.56	7	0.67	7	0.38	4	0.53	9	0.63	6
4	0.52	9	0.64	9	0.34	10	0.53	7	0.54	10
5	0.50	1	0.60	1	0.33	2	0.52	1	0.52	2
6	0.48	10	0.56	4	0.32	6	0.49	4	0.50	8
7	0.47	4	0.54	10	0.32	8	0.46	10	0.45	9
8	0.45	2	0.51	2	0.31	7	0.44	2	0.43	1
9	0.43	8	0.47	8	0.31	5	0.41	8	0.38	4
10	0.20	3	0.18	3	0.19	3	0.18	3	0.25	3
Variant	*DGP*									
MSE	*All*		*1–12*		*13–20*		*1–20*		*21–25*	
BIC										
1	2.65	4	0.74	4	4.99	3	2.85	3	0.52	5
2	2.80	1	0.76	1	6.30	8	3.02	4	0.56	6
3	2.86	8	0.76	9	6.43	4	3.04	8	0.83	9
4	2.91	2	0.77	10	6.74	2	3.18	2	0.93	1
5	2.93	9	0.80	2	7.02	1	3.26	1	1.19	4
6	2.98	10	0.86	8	7.17	10	3.33	10	1.59	10
7	3.07	3	0.88	6	7.48	9	3.45	9	1.83	2
8	3.22	6	0.90	5	8.40	6	3.89	6	2.13	8
9	3.31	5	1.42	3	8.67	5	4.00	5	3.92	3
10	20.92	7	22.43	7	14.20	7	19.14	7	28.07	7
$P(\hat{k} = p)$	*All*		*1–12*		*13–20*		*1–20*		*21–25*	
1	0.58	8	0.73	4	0.30	3	0.53	4	0.84	5
2	0.58	2	0.73	1	0.24	8	0.53	3	0.83	6
3	0.58	4	0.73	9	0.24	4	0.52	8	0.81	10
4	0.58	10	0.72	10	0.23	7	0.52	1	0.81	2

TABLE 6

(continued)

$P(\hat{k} = p)$	*All*		*1–12*		*13–20*		*1–20*		*21–25*	
5	0.57	1	0.71	2	0.23	2	0.52	2	0.80	8
6	0.57	9	0.71	8	0.22	10	0.52	10	0.78	9
7	0.57	3	0.71	6	0.21	1	0.52	9	0.77	1
8	0.56	6	0.71	5	0.20	9	0.49	6	0.75	4
9	0.55	5	0.68	3	0.16	6	0.48	5	0.72	3
10	0.24	7	0.24	7	0.15	5	0.23	7	0.28	7

Notes: Given a class of models, the first column is the MSE, and the second column is $P(\hat{k} = p)$. Let $\hat{k}$ be the k chosen on average by a given criterion (i.e. results in Table 1). Then the MSE for that criterion is $\frac{1}{J}\sum_{i=1}^{J}(\hat{k}_i - p)^2$, where $J = 5{,}000$ is the number of replications.

IV. Conclusion

Lag length selection is frequently required in time series analysis. This paper shows that the formulation of AIC and BIC can affect the precision and variability of the selected lag order. Textbooks that define the penalty functions as C_T/T can quite easily be misinterpreted as method 2 (which uses the maximum number of observations $N = T - k$, and scale the penalty and the sum of squared residuals accordingly), method 3 (which again estimates each autoregression using the maximum number of observations, $N = T - k$, but scales the penalty and the sum of squared residuals by T), or method 7 (which instead scales the penalty by $T - k$ and the sum of squared residuals by $T - 2k$). Neither is desirable from a practical standpoint. Theory dictates that the penalty factor must increase in k. In practice, there is some leeway in how the scaling on the penalty, M, and the degrees of freedom adjustment of the estimate of the variance, τ, are defined to make this condition hold. Our simulations show that the methods that give the most precise estimates are those that hold the number of effective observations, N, fixed across models to be compared. Theoretical considerations reveal that this is indeed necessary for valid model comparisons.

Final Manuscript Received: April 2004

References

Akaike, H. (1969). 'Fitting autoregressions for predictions', *Annals of the Institute of Statistical Mathematics*, Vol. 21, pp. 243–247.

Chow, G. (1983). *Econometrics*, McGraw Hill, New York, NY.

Diebold, F. X. (1997). *Elements of Forecasting*, South Western Publishing, Cincinatti, OH.

Enders, W. (1995). *Applied Econometric Time Series*, Wiley, NewYork, NY.

Eviews (1997). *User's Guide,* QMS Software, Irvine, CA.

Hayashi, F. (2000). *Econometrics*, Princeton Unversity Press, Princeton, NJ.
Hendry, D. F. and Krolzig, H. M. (2002). 'The properties of automatic Gets modelling', Unpublished Manuscript, Oxford University, Oxford.
Judge, G., Griffiths, W., Hill, R. and Lee, T. (1980). *The Theory and Practice of Econometrics*, John Wiley and Sons, New York, NY.
Lutkepohl, H. (1993). *Introduction to Multiple Time Series*, Springer Verlag, Berlin.
Mallows, C. L. (1973). 'Come comments on C_p', *Technometrics*, Vol. 15, pp. 661–675.
Priestley, M. B. (1981). *Spectral Analysis and Time Series*, Vol. 1, Academic Press, NewYork, NY.
Schwarz, G. (1978). 'Estimating the dimension of a model', *The Annals of Statistics*, Vol. 6, pp. 461–464.
Shibata, R. (1980). 'Asymptotic efficient selection of the order of the model for estimating parameters of a linear process', *Annals of Statistics*, Vol. 8, pp. 147–164.