A NOTE ON THE SELECTION OF TIME SERIES MODELS

Serena Ng * Pierre Perron †

June 2001

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.

Keywords: Information Criteria, AIC, BIC, lag length selection
JEL Classification: F30, F40, C2, C3, C5.

*(Corresponding Author) Department of Economics, Boston College, Chestnut Hill, MA 02467
Email: serena.ng@bc.edu

†Department of Economics, Boston University 270 Bay State Road, Boston, MA 02115 and C.R.D.E., Université de Montréal. Email: perron@bu.edu

1 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 a criterion such as:

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

where $\widehat{\sigma}_i^2 = T^{-1} \sum_{t=1}^{T} \widehat{e}_t^2$ is an estimate of the regression error variance for the i^{th} 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 given C_T, there is no ambiguity in how to set up the criterion function. The Akaike Information Criterion (AIC) obtains when $C_T = 2$, and the Scharwz (Bayesian) Information Criterion obtains when $C_T = \ln T$. For any $T > \exp(2)$, the penalty imposed by the BIC is larger than 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

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

is the data generating process with $e_t \sim \ iid(0, \sigma^2)$. 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:

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

The adequacy of the approximate model for the data generating process 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 (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 $\widehat{\sigma}_k^2$, C_T, and T are unambiguous, the IC can be defined in a number of ways. Specifically, let $kmax$ be the maximum number of lags deemed acceptable by a practitioner and consider

$$\min_{k=0,\ldots,kmax} IC(k) = \min_{k=0,\ldots,kmax} \ln \widehat{\sigma}_k^2 + k \frac{C_M}{M}, \tag{3}$$

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

where $\widehat{e}_{tk}$ are the least squares residuals from estimation of (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 ten variations of $IC(k)$.

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

	1	2	3	4	5	6	7	8	9	10
N	T-kmax	T-k	T-k	T-kmax	T-kmax	T-kmax	T-k	T-k	T-kmax	T-k
τ	T-kmax	T-k	T	T	T-kmax-k	T-kmax-k	T-2k	T-k	T-kmax	T-k
M	T-kmax	T-k	T	T	T-kmax-k	T-kmax	T-k	T	T-kmax-k	T-2k

Methods 1, 4, 5, 6, and 9 hold the effective number of observations fixed as k varies, namely, $N = T - kmax$. 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 k^{th} lag. On the other hand, methods 2, 3, 7, 8, and 10 make maximum use of the data since 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 k^{th} 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, $\widehat{\sigma}_k^2$ will be larger after degrees of freedom adjustment. Criteria that takes this into account should be expected to choose a smaller model, all else equal. The penalty for all ten methods converges to zero at rate T, but in finite samples, T-kmax-k< T-kmax < 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.

There appears to be no consensus in the literature on which of these variants to use. Priestley (1981) seems to suggest method 2 (p. 373). His argument requires that N does not depend on k. This, however, is invalid since he also defined N as $T-k$. In a multivariate context, Lutkepohl (1993) defines the criteria in terms of the length of the time series (p. 129), which could be T, $T-k$, or even $T - kmax$. Enders (1995) defines the criteria in terms of the number of usable observations (p. 88), but this terminology is also open to interpretation. Diebold (1997) uses the full length of

the data, T, when defining the criteria (p. 26). This is consistent with the notation of method 3. However, estimation with T observations is infeasible unless one initializes the first few lags to zero. 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]

2 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.

2.1 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 $k > p$, $\beta(k) = (\beta_1, ..., \beta_p, 0, \dots 0)'$ denote the true parameters, and $\widehat{\beta}(k) = (\widehat{\beta}_1, ..., \widehat{\beta}_k)'$ are the estimated parameters. If $p > k$, $\beta(k) = (\beta_1, \dots \beta_p)'$ and $\widehat{\beta}(k) = (\widehat{\beta}_1, \dots \widehat{\beta}_k, 0, \dots, 0)'$. Following the treatment of Gourieroux and Monfort (1995, pp. 307-309), let $f(y|\beta(k))$ be the likelihood function of the data $(y_{n+1}, ..., y_T)$ conditional on the initial observations $(y_1, ..., y_n)$. Let $N = T - n$. The Kullback distance between the true probability distribution and the estimated parametric model is $K = E_0[\ln(f(y|\beta(k))) - \ln(f(y|\widehat{\beta}(k)))]$ with sample analog:

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

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 $X_t = (y_{t-1}, ..., y_{t-k})'$ and

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

where $\widehat{\sigma}_k^2 = N^{-1} \sum_{t=n+1}^{T} \widehat{e}_{tk}^2$. Using Taylor series expansions, we have $NK = \Phi_T(k)/2 + o_p(1)$ and $N\widetilde{K} = -\Phi_T(k)/2 + o_p(1)$. Since $N(K - \widetilde{K}) = \Phi_T(k) + o_p(1)$, $\lim_{N\to\infty} E[N(K - K^*)] = 0$ for $K^* = \widetilde{K} + \Phi_T(k)$. 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

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

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

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

$$AIC(k) = \ln \widehat{\sigma}_k^2 + \frac{2k}{N}. \tag{5}$$

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

2.2 The C_p Criterion

Let $X_t = (y_{t-1}, \ldots, y_{t-p})' = (X_{1t} \;\; X_{2t})$, where $X_{1t} = (y_{t-1}, \ldots, y_{t-k})'$, $X_{2t} = (y_{t-k-1}, \ldots, y_{t-p})'$, with $Y = (y_1 \;\; y_2 \ldots y_T)'$, $X_1 = (X_{11} \;\; X_{12} \;\; \ldots X_{1T})'$ and $X_2 = (X_{21} \ldots X_{2T})'$. 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 k^{th} element. Suppose the true model is $Y = X_1\beta_1 + X_2\beta_2 + e$, 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 $\widehat{\beta}_1 - \beta_1 = (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'e$. Furthermore, $\widehat{e}_k = M_1X_2\beta_2 + M_1e$, where $M_1 = [I - X_1(X_1'X_1)^{-1}X_1']$. Then

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

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

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

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

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

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

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

[2]The developments here follow Judge, Griffiths, Hill and Lee (1980), p. 419.

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

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

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

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

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.

2.3 The FPE Criterion

The Final Prediction Error 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

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

Using the asymptotic approximation that $\sqrt{N}(\widehat{\beta}(k) - \beta(k)) \sim N(0, \sigma^2\Gamma_k^{-1})$, 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 $\widehat{\sigma}^2 = N^{-1}\sum_{t=1}^N \widehat{e}_{tk}^2$, and under normality, $N\widehat{\sigma}_k^2/\sigma^2 \sim \chi_{N-k}^2$. Since $E[N\widehat{\sigma}_k^2/\sigma^2] = (N-k)$, using $\sigma^2 \approx N\widehat{\sigma}_k^2/(N-k)$, the FPE can then be written as

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

2.4 Posterior Probability

To develop the arguments for the BIC of Schwarz (1978), we follow Chow (1983). Let $f(y|k)$ be the marginal p.d.f. for the data under a k^{th} order model, $f(k)$ be the prior density for a k^{th} 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 k^{th} 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 k^{th} order model is

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

where $f(y,\beta(k))$ is the likelihood function for the k^{th} order model with parameters $\beta(k)$. But it is also known that under regularity conditions, the posterior distribution of $\beta(k)$ is Gaussian with variance S. That is,

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

Now $|S| = N^k|\frac{1}{N}\sum_{t=n+1}^{T} X_t X_t'|$, and thus $\ln|S| = k\ln(N) + \ln(R_N)$, $R_N = \frac{1}{N}\sum_{t=n+1}^{T} X_t X_t'$. Evaluating the posterior distributions around the maximum likelihood estimator, $\widehat{\beta}(k)$, equating and rearranging terms, we have:

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

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

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

Now the first term is proportional to $(-N/2)\ln(\widehat{\sigma}_k^2)$, where $\widehat{\sigma}_k^2 = N^{-1}\sum_{t=n+1}^{T}\widehat{e}_{tk}^2$. Multiplying by $\frac{-2}{N}$, the k that maximizes the posterior of the data also minimizes:

$$BIC(k) = \ln\widehat{\sigma}_k^2 + \frac{k\ln N}{N}. \qquad (8)$$

Three assumptions are used to derive (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 $n = kmax$ as in the AIC. The third is that log likelihood function evaluated at the estimated parameters is proportional to $\widehat{\sigma}_k^2$. These are the same assumptions underlying the AIC.

2.5 Overview

To relate the 10 methods to the theoretical discussions, the AIC and BIC both require $M = N$, both require $\ln\widehat{\sigma}_k^2$ 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 $n = kmax$. 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 a SC_p^* with $C_M = 2$. This includes methods 1, 4, and 5. The lnFPE obtains by letting τ=N and M=N-k. Thus, methods 9 and 10 are consistent with the theoretical underpinnings of the lnFPE. Of the 10 methods considered, methods 2, 3, 6, 7, 8 bear no immediate relation to well-known criteria in the literature.

3 Simulations

To assess the empirical properties of the 10 methods considered, we simulate data from 25 time series processes detailed in Appendix A. 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 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 a ARMA(1,1) model, $(1 - \phi L)y_t = (1 + \theta L)e_t$, and consider a truncated version of its infinite autoregressive representation is $\sum_{i=0}^{\infty}(\phi + \theta)(-\theta)^i y_{t-i} = e_t$. 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 $|\beta_{p+1}| < .1$. The parameterizations allow us to assess situations when the autoregressive coefficients decline at a geometric rate. We also consider ten cases with ARCH errors. In cases 21-25, $p = 0$, and we assess if ARCH errors affect the lag length selected for the autoregressions. In cases 26-35, we estimate autoregressions in y_t^2 so the IC is used to select the order of ARCH processes. Tables 1-4 report results for DGPs 1-25. Tables 5-8 reports the corresponding results for DGPs 26-35.

Simulations were performed using Gauss for T=100 and 250, with $kmax$ set to int$[10(T/100)^{1/4}]$. We only report results for $T = 100$. Results for T=250 are available on request. Table 1 reports the average k selected by the AIC and BIC over 5000 simulations, Table 2 reports the probability of selecting the true model, while Table 3 reports the standard errors. Differences between the AIC and the BIC in DGPs 1 to 20 are along the lines documented in the 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 model 19 with $p = 8$, the AIC truncates at six lags on average with $\beta_6 = .26$. The BIC, on the other hand, truncates at three lags with $\beta_3 = .51$.[3] Results relating to ARCH errors are lesser known and are noteworthy in their own right. As seen from Tables 1 and 2, the AIC tends to mistreat ARCH effects for serial correlation and often selects lags larger than zero. For a given p, the probability that the AIC and BIC correctly select the order of an ARCH(p) is lower than the probability that the criteria can correctly select an AR(p) process (compare Table

[3]In results for T=250 (not reported), the k chosen by the BIC is still small.

2 with 6).

Our main interest is in the sensitivity of the methods with respect to N, τ, and M. Of the three parameters, the estimates are more 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-kmax is more precise even though it uses fewer observations in estimating models with $k <$kmax 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 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 why, recall that the basis of the IC is to trade-off good fit against parsimony. Let $RSS_k = \sum_{t=n+1}^{T} \widehat{e}_{tk}^2$, so that $\widehat{\sigma}_k^2 = \frac{RSS}{\tau}$. Then

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

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. Second, 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 3.

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 C_M constant 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 be invariant to k is desirable. This, however, is of secondary importance relative to fixing N, since 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.

We also rank the methods by the average mean-squared error and by the probability of selecting the true model. The results are reported in Table 4. 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 $\lim_{T\to\infty} P(\widehat{k} = p) = 1$ when p is finite[4], 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.[5] 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-kmax. Holding the sample size fixed is thus crucial in model comparisons.

4 Conclusion

Lag length selection is frequently required in time series analysis. This paper shows that how the AIC and BIC are formulated 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 methods 2, 3 or 7. Neither is desirable from an empirical standpoint. Theory dictates that the penalty factor must increase in k. In practice, there is some leeway in how M and τ are defined to make this condition hold. Our simulations show that the methods that give the most precise estimates are those that hold N fixed across models to be compared. Theoretical considerations reveal that this is indeed necessary for valid model comparisons.

[4]Geweke and Meese (1981) showed that for consistent model selection, we need $C_M/M \to 0$, and $TC_M/M \to \infty$ for estimators of β that are $\sqrt{T}$ consistent. The second condition is not met by the AIC.

[5]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.

References

Akaike, H. (1969), Fitting Autoregressions for Predictions, *Annals of the Institute of Statistical Mathematics* **21**, 243–247.

Chow, G. (1983), *Econometrics*, Mcgraw Hill.

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

Enders, W. (1995), *Applied Econometric Time Series*, Wiley, New York.

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

Geweke, J. and Meese, R. (1981), Estimating Regression Models of Finite but Unknown Order, *International Economic Review* **23:1**, 55–70.

Hayashi, F. (2000), *Econometrics*, Princeton Unversity Press, Princeton, N.J.

Judge, G., Griffiths, W., Hill, R. and Lee, T. (1980), *The Theory and Practice of Econometrics*, John Wiley and Sons, New York.

Lutkepohl, H. (1993), *Introduction to Multiple Time Series*, Springer Verlag, Berlin.

Mallows, C. L. (1973), Come Comments on C_p, *Technometrics* **15**, 661–675.

Priestley, M. B. (1981), *Spectral Analysis and Time Series*, Vol. 1, Academic Press, New York.

Schwarz, G. (1978), Estimating the Dimension of a Model, *The Annals of Statistics* **6**, 461–464.

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

Table A: DGP for Models 1-25: $y_t = \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + e_t$, $e_t \sim iidN(0,1)$, $y_0 = y_{-1} = \ldots y_{-p} = 0$.

Model	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 for Models 21-35: $y_t = \sqrt{h_t} e_t$, $e_t \sim N(0,1)$, $h_t^2 = 1 + \beta_1 h_{t-1}^2 + \ldots + \beta_p h_{t-p}^2$

Model	p	β_1	β_2	β_3	β_4
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
26	0	0.00	0.00	0.00	0.00
27	1	0.25	0.00	0.00	0.00
28	1	0.50	0.00	0.00	0.00
29	1	0.80	0.00	0.00	0.00
30	1	0.90	0.00	0.00	0.00
31	1	0.95	0.00	0.00	0.00
32	2	0.40	0.40	0.00	0.00
33	2	0.60	0.30	0.00	0.00
34	3	0.30	0.20	0.40	0.00
35	4	0.30	0.10	0.10	0.40

- For Models 13-20, $\beta_i = (\phi + \theta)(-\theta)^i$ subject to the constraint that $|\beta_i| > .1$.
- For models 1-25, autoregressions are constructed for y_t.
- For models 26-35, autoregressions are constructed for y_t^2.

Table 1: Average k selected

AIC											
Model	p	1	2	3	4	5	6	7	8	9	10
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
BIC											
Model	p	1	2	3	4	5	6	7	8	9	10
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
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

Table 2: $P(\widehat{k} = p)$

AIC											
Model	p	1	2	3	4	5	6	7	8	9	10
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
BIC											
Model	p	1	2	3	4	5	6	7	8	9	10
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
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

Table 3: Standard Errors of $\widehat{k}$

AIC											
Model	p	1	2	3	4	5	6	7	8	9	10
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
BIC											
Model	p	1	2	3	4	5	6	7	8	9	10
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
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

Table 4a: Rankings of the 10 methods for AIC:

	Models									
MSE	All		1-12		13-20		1-20		21-25	
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
$P(\widehat{k}=p)$	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

Table 4b: Rankings of the 10 methods for BIC

	Models									
MSE	All		1-12		13-20		1-20		21-25	
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
$P(\widehat{k}=p)$	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
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

Given a class of models, the first column is the MSE, and the second column is $P(\widehat{k}=p)$. Let $\widehat{k}$ be the k chosen on average by a given criterion (i.e. results in Table 1). Then the MSE for that criterion is $\frac{1}{J}\sum_{i=1}^{J}(\widehat{k}_i - p)^2$, where $J = 5000$ is the number of replications .

Table 5: Average k selected for ARCH Models

AIC											
Model	p	1	2	3	4	5	6	7	8	9	10
26	0	0.66	1.63	3.85	0.91	0.16	0.18	0.83	1.84	0.54	1.46
27	1	1.23	2.04	3.91	1.47	0.66	0.68	1.28	2.23	1.11	1.89
28	1	1.67	2.39	3.80	1.88	1.11	1.13	1.77	2.52	1.54	2.24
29	1	2.02	2.62	3.72	2.23	1.45	1.50	2.14	2.73	1.91	2.53
30	1	2.17	2.79	3.81	2.36	1.58	1.63	2.33	2.90	2.05	2.69
31	1	2.25	2.95	3.96	2.45	1.64	1.68	2.51	3.07	2.11	2.85
32	2	2.77	3.35	4.77	3.02	1.93	1.99	2.67	3.50	2.57	3.20
33	2	2.80	3.49	4.77	3.04	1.98	2.04	2.90	3.63	2.61	3.34
34	3	3.77	4.44	5.92	4.07	2.64	2.75	3.63	4.64	3.50	4.24
35	4	4.13	4.78	6.45	4.47	2.79	2.94	3.82	5.01	3.81	4.54
BIC											
Model	p	1	2	3	4	5	6	7	8	9	10
26	0	0.05	0.40	0.74	0.06	0.02	0.03	3.32	0.44	0.05	0.37
27	1	0.43	0.84	1.20	0.46	0.34	0.35	3.45	0.88	0.42	0.80
28	1	0.85	1.36	1.69	0.91	0.74	0.75	3.46	1.40	0.84	1.30
29	1	1.18	1.79	2.09	1.24	1.04	1.06	3.47	1.84	1.15	1.74
30	1	1.27	1.97	2.27	1.33	1.12	1.14	3.56	2.01	1.25	1.91
31	1	1.33	2.12	2.44	1.40	1.18	1.20	3.73	2.18	1.30	2.08
32	2	1.50	2.15	2.59	1.59	1.24	1.27	4.43	2.21	1.46	2.09
33	2	1.57	2.38	2.82	1.65	1.34	1.36	4.46	2.45	1.52	2.31
34	3	1.98	2.92	3.52	2.13	1.59	1.64	5.56	3.03	1.91	2.81
35	4	1.98	2.93	3.66	2.15	1.48	1.55	6.03	3.05	1.86	2.82

Table 6: $P(\widehat{k} = p)$ for ARCH Models

AIC											
Model	p	1	2	3	4	5	6	7	8	9	10
26	0	0.76	0.61	0.35	0.71	0.91	0.91	0.77	0.59	0.78	0.63
27	1	0.38	0.34	0.23	0.36	0.39	0.39	0.36	0.33	0.39	0.35
28	1	0.52	0.47	0.32	0.49	0.59	0.59	0.54	0.45	0.53	0.48
29	1	0.51	0.46	0.33	0.48	0.61	0.60	0.54	0.45	0.52	0.47
30	1	0.49	0.43	0.31	0.46	0.59	0.59	0.51	0.42	0.51	0.44
31	1	0.48	0.40	0.28	0.45	0.59	0.58	0.48	0.39	0.50	0.41
32	2	0.29	0.27	0.19	0.28	0.33	0.32	0.29	0.26	0.31	0.28
33	2	0.23	0.21	0.16	0.23	0.26	0.26	0.23	0.21	0.25	0.22
34	3	0.28	0.26	0.17	0.27	0.32	0.32	0.29	0.25	0.30	0.28
35	4	0.25	0.23	0.16	0.24	0.26	0.25	0.21	0.22	0.27	0.24
BIC											
Model	p	1	2	3	4	5	6	7	8	9	10
26	0	0.97	0.86	0.79	0.96	0.98	0.98	0.38	0.85	0.97	0.86
27	1	0.33	0.32	0.35	0.35	0.29	0.29	0.26	0.32	0.33	0.33
28	1	0.59	0.55	0.54	0.60	0.56	0.56	0.34	0.54	0.60	0.55
29	1	0.66	0.58	0.54	0.65	0.66	0.66	0.35	0.58	0.66	0.59
30	1	0.65	0.56	0.52	0.64	0.67	0.66	0.33	0.56	0.65	0.57
31	1	0.64	0.53	0.49	0.63	0.66	0.66	0.30	0.53	0.65	0.54
32	2	0.30	0.28	0.29	0.31	0.28	0.28	0.21	0.28	0.30	0.29
33	2	0.23	0.23	0.23	0.24	0.22	0.22	0.17	0.22	0.23	0.23
34	3	0.29	0.26	0.28	0.30	0.24	0.24	0.19	0.26	0.29	0.26
35	4	0.19	0.18	0.20	0.20	0.14	0.15	0.18	0.18	0.18	0.18

Table 7: $SE(\widehat{k})$, ARCH Models

AIC											
Model	p	1	2	3	4	5	6	7	8	9	10
26	0	1.61	2.75	3.82	1.96	0.62	0.69	1.99	2.96	1.36	2.55
27	1	1.69	2.65	3.51	1.95	0.89	0.94	2.03	2.83	1.47	2.49
28	1	1.70	2.59	3.29	1.95	1.02	1.08	2.16	2.70	1.49	2.43
29	1	1.87	2.57	3.13	2.08	1.30	1.36	2.30	2.66	1.71	2.49
30	1	2.00	2.63	3.11	2.18	1.44	1.52	2.40	2.72	1.84	2.54
31	1	2.09	2.68	3.11	2.24	1.52	1.58	2.48	2.76	1.92	2.59
32	2	2.25	2.77	3.17	2.40	1.66	1.72	2.52	2.86	2.06	2.67
33	2	2.33	2.87	3.19	2.46	1.72	1.80	2.66	2.95	2.15	2.76
34	3	2.59	2.95	3.06	2.68	2.09	2.20	2.86	3.03	2.41	2.86
35	4	2.83	3.10	3.01	2.88	2.37	2.50	3.08	3.16	2.66	3.03
BIC											
26	0	0.30	1.29	1.91	0.35	0.19	0.21	3.58	1.40	0.29	1.23
27	1	0.64	1.58	1.99	0.67	0.54	0.55	3.30	1.66	0.61	1.49
28	1	0.77	1.82	2.13	0.83	0.67	0.69	3.10	1.88	0.75	1.73
29	1	1.04	2.07	2.29	1.10	0.91	0.95	2.96	2.12	1.00	2.01
30	1	1.18	2.18	2.40	1.22	1.00	1.04	2.96	2.23	1.12	2.11
31	1	1.29	2.26	2.47	1.35	1.12	1.15	2.98	2.33	1.21	2.22
32	2	1.39	2.26	2.52	1.45	1.18	1.22	3.04	2.33	1.32	2.20
33	2	1.48	2.42	2.66	1.53	1.28	1.31	3.08	2.48	1.40	2.35
34	3	1.88	2.71	2.89	1.94	1.65	1.72	3.00	2.77	1.79	2.61
35	4	2.13	2.91	3.11	2.20	1.83	1.90	3.01	2.98	2.01	2.83

Table 8: Rankings of the 10 methods for ARCH processes

	MSE		$P(\widehat{k}=p)$		MSE		$P(\widehat{k}=p)$	
	AIC				BIC			
1	2.67	5	0.49	5	11.03	5	0.46	5
2	2.89	6	0.48	6	11.32	6	0.46	6
3	4.27	9	0.43	9	15.80	7	0.42	7
4	5.22	1	0.42	7	17.89	9	0.35	9
5	6.36	4	0.42	1	18.62	1	0.34	1
6	6.95	7	0.40	4	19.34	4	0.33	4
7	8.82	10	0.38	10	22.09	10	0.33	10
8	9.82	2	0.37	2	22.72	2	0.33	2
9	10.88	8	0.36	8	23.31	8	0.32	8
10	19.05	3	0.25	3	38.96	3	0.19	3