ISSN 1440-771X

MONASH UNIVERSITY
AUSTRALIA

DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS

Choosing Lag Lengths in Nonlinear Dynamic Models

Heather M. Anderson

Working Paper 21/2002

Choosing Lag Lengths in Nonlinear Dynamic Models

Heather M. Anderson*
Department of Econometrics
and Business Statistics
Monash University
Clayton, Victoria 3800,
AUSTRALIA

Abstract

Given that it is quite impractical to use standard model selection criteria in a nonlinear modeling context, the builders of nonlinear models often choose lag length by setting it equal to the lag length chosen for a linear autoregression of the data. This paper studies the performance of this procedure in a variety of circumstances, and then proposes some new and simple model selection procedures, based on linear approximations of the nonlinear forms. The idea here is to apply standard selection criteria to these linear approximations, rather than to autoregressions that make no provision for nonlinear behavior. A simulation study compares the properties of these proposed procedures with the properties of linear selection procedures.

Keywords: Nonlinear time series models, Neural networks, Model selection criteria, Polynomial approximations, Volterra expansions.

JEL classification: C22, C45, C51.

*Correspondence:. E-mail: Heather.Anderson@BusEco.monash.edu.au
This work was supported by the Australian Research Council (Large Grant No A00103258), and some of it was undertaken while visiting the Department of Economics at the University of California at San Diego. I would like to thank UCSD for hospitality and Farshid Vahid for many helpful comments.

1. Introduction

There has been considerable interest in nonlinear time series models in recent years, as evidenced by a growing body of studies of asymmetries in business cycles and nonlinearities in asset markets. Models that allow for state-dependent or regime-switching behaviour have been very popular, with well known examples including the Markov Switching (MS) model (Hamilton, 1989), the Current Depth of Recession (CDR) model (Beaudry and Koop, 1993), the Smooth Transition Autoregressive (STAR) model (Teräsvirta, 1994) and the Threshold Autoregressive (TAR) model (Potter, 1995). At first blush, such models have intuitive appeal and seem relatively easy to work with, although in practice they are often quite difficult to specify and estimate.

One difficulty associated with nonlinear modelling is that the researcher usually needs to determine the lag structure of the data before conducting nonlinearity tests or estimating nonlinear specifications. Researchers often choose this lag length by setting it equal to the lag length chosen for a linear autoregressive model of the data, where the latter choice is based on the partial autocorrelation function, or standard lag selection criteria such as those proposed by Akaike (1974), Hannan-Quinn (1979) or Schwartz (1978). It is recognized that these lag selection techniques will only "work" if the main features of the linear autocorrelation structure reflect the lag dependencies associated with the underlying nonlinear process, but it is quite impractical to calculate and compare model selection criteria for nonlinear specifications of different lag lengths, when each calculation requires the maximization of a potentially ill-behaved likelihood.

There is a very large literature on lag selection (see the survey article by de Goojier et al (1985)), but most of this is set in a linear context and there is comparatively little that works within the more general nonlinear framework. There is a small body of research that has used various dependence measures to construct analogues to the autocorrelation and partial autocorrelation functions that are often used in linear settings. Auestad and Tjøstheim (1990) use nonparametric estimates of conditional means and variances for this purpose, while Granger and Lin (1994) suggest the use of Pinsker's (1964) mutual information coefficients and Kendall's (1938) partial correlation coefficients. Granger et al (2001) have also worked with a dependence metric based on the Bhattacharya-

Matuisa-Hellinger measure of entropy. After simulating the distribution of this metric under the null hypothesis of independence, they suggest that it be used for identifying statistically significant lag lengths in potentially nonlinear settings.

The above procedures essentially mimic various aspects of the Box-Jenkins approach for identification, and like the Box-Jenkins methodology, their reliance on the skill and judgement of the researcher invites criticism. Another potential problem with the above procedures is that with the exception of Kendall's coefficient, they require nonparametric estimation of density functions, which is difficult for the novice and often inappropriate when dealing with small samples. The nonparametric final prediction error (FPE) criterion proposed by Tjøstheim and Auestad (1994) offers a less subjective approach to the lag selection problem, but it is nevertheless difficult to implement and impractical, given the size of typical economic data sets.

This paper looks at the problem of lag selection for nonlinear models from the viewpoint of an applied economist. It focusses on nonlinear autoregressive models because these models are popular in applied work, and the goal is to study simple and practical techniques that might be appropriate for relatively small samples of up to three hundred observations. My suggestion is to work with linear approximations to nonlinear forms and then to apply the usual lag selection criteria (for example AIC, HQ and BIC) to discriminate between such approximations. Since these approximations are linear in parameters, the calculation of selection criteria for each lag length is very straightforward. Naturally, the procedure relies on finding reasonable approximations for nonlinear functional forms. I work with second order polynomial expansions and various subsets of these expansions, and although I also experiment with neural network approximations, I find that the former seem to work better for relatively small samples.

I study both linear and nonlinear data generating processes (DGPs), with short, medium and long lag structures. These DGPs are all based on published models of macroeconomic and financial data, so that my conclusions relate to the sorts of series that econometricians actually encounter in practice. I find that when the underlying DGP is nonlinear, standard model selection criteria tend to overestimate the true lag-length. This problem is especially pronounced when AIC (which generally favours overfitting) is used, but it is also evident when the Schwartz criterion (which generally favours underfitting) is used. This suggests that when series are potentially nonlinear and the selection criteria uses only one parameter to account for each lag, higher parameter penalties are needed to account for the fact that nonlinear DGPs will typically have more than one parameter associated with each lag length. Given that "counting" parameters in a nonlinear setting can be a problematic concept, because different parameters can affect the data generating process at different points in time, I make no attempt to suggest appropriate parameter penalties here. Instead, I focus on using approximations that have a known number of parameters, and simply report on how well the application of AIC, HQ and BIC to these approximations identifies the true lag-length. My simulations suggest that lag selection based on approximations reduces the tendency to overpredict lag-length, particularly for nonlinear DGPs. Further, this reduced tendency to overpredict lag-length is accompanied by an increased tendency to underpredict lag-length.

The organization of this paper is as follows. Section 2 sets up the lag selection problem and outlines a general framework within which the researcher might tackle this problem. Here, I discuss why standard selection criteria might not work, and suggest a few approaches that might work better. In Section 3, I describe some simulation exercises designed to assess the performance of my procedures within both linear and nonlinear contexts, and then I report the results of the simulations in Section 4. Section 5 summarises and concludes.

2. Lag selection in a potentially nonlinear setting

I consider a univariate time series y_t, with history $Y_t = (\dots\ y_{-2},\, y_{-1},\, y_0,\, y_1,\, \dots\ y_{t-1})$ that has a DGP given by

$$y_t = \beta_0 + \beta' Y_{t-1}^{t-p} + \Psi(Y_{t-1}^{t-p}, \theta) + \varepsilon_t \qquad ((1))$$

where $Y_{t-1}^{t-p} = (y_{t-1},\, y_{t-2},\, \dots\ y_{t-p})$, β_0, β and θ are parameters, and $\Psi(Y_{t-1}^{t-p}, \theta)$ is a nonlinear function of its first and possibly second arguments. I assume that y_t is essentially stationary and weakly dependent as defined in Wooldridge (1994), and that although the functional form of Ψ might be initially unknown to the researcher, it satisfies conditions that will allow consistent estimation of θ once this functional form has been specified. I further assume that ε_t is a sequence of independent and identically distributed zero mean random variables with $E(\varepsilon_t^2) = \sigma^2$ being a finite constant.

The class of models defined by (1) includes the large family of exponential autoregressive (EXPAR) models discussed by Haggan and Ozaki (1981), threshold autoregressive (TAR) models (see Tong (1990)), and closely related models such as smooth transition autoregressive (STAR) models (see Teräsvirta (1994)) and current depth of recession (CDR) models (see Beaudry and Koop (1993)). In the special case given by $\sigma^2 = 0$, model (1) also includes chaotic series such as the logistic map series for which $y_t = 4y_{t-1}(1-y_{t-1})$. This series is especially interesting, because its autocorrelation and partial autocorrelation functions are equal to zero at all lags. Model (1) does not allow for autoregressive conditional heteroskedasticity (ARCH) processes, bilinear processes or hidden Markov chains, but given the practical importance of pure autoregressive processes in the applied econometrics literature, it is useful to start with these.

Our problem is to determine the lag length p, possibly prior to determining the dimension of θ and specifying the functional form of Ψ. This situation often arises when the researcher wants to test for nonlinearity, or wants to determine the type of nonlinearity present in Ψ, but the test assumes a knowledge of the lag length p. Alternatively, we might not be able to specify Ψ and estimate β and θ until we have first determined p. Even if we are willing to specify Ψ prior to determining p, we might want to avoid estimating (1) using different lags lengths, simply because nonlinear estimation is time consuming and difficult. Furthermore, the difficulty in estimating and comparing nonlinear specifications for a set of different p can be compounded if the true model is actually linear and various parameters in the specification are therefore unidentified. See Davies (1977) or Engle (1984) for further discussion of this identification problem.

Standard lag selection criteria (such as AIC, HQ or BIC) applied to autoregressive time series typically assume a linear process for y_t, and then solve an optimization problem for p that simultaneously rewards the fit of the AR(p) and penalises its complexity. Higher p will improve the fit, but will also entail more complexity. The assumption that the time series is linear is not innocuous when AR(p) models are being fitted, because it implies that one of the options considered by the researcher will be the true model. The reason why these standard procedures might not work in the nonlinear context given by (1), is that by just considering linear AR(p) processes (for $0 \leq p \leq p^{\max}$), the researcher does not include the true model in his/her choice set. Tong (1990, p288) discusses this issue very briefly, and notes that when the set of models under consideration does not include the true model, the selected model may or may not be adequate, depending on how close the likelihood of the chosen model is to the likelihood of the true model.

Akaike (1985) emphasizes that the researcher's choice set of possible models should reflect his/her particular way of looking at the data, and this suggests that when models such as (1) are being considered, the family of AR(p) models might not necessarily provide the most appropriate choice set for determining lag length. This leads to the question of whether there are other families of models which might better account for the nonlinearity in (1) and thus provide a more reliable selection of p. Since Ψ is potentially difficult to estimate for different lag lengths p, and its precise functional form isn't necessarily known, it seems sensible to focus on families of models that are both simple and capture unspecified nonlinearities.

Tests for unspecified nonlinearity in the conditional mean are often based on simple linear regressions. Various examples are studied and discussed in Granger and Teräsvirta (1993) and Lee et al (1993), who claim that these tests involve "specific functions of Y_{t-1}^{t-p}, that are chosen to capture essential features of possible nonlinearities". Such functions include the duals of Volterra expansions in Y_{t-1}^{t-p} and neural networks in Y_{t-1}^{t-p}, and I discuss these further below. Two important features of each of these functions are that they can be used to approximate Ψ in model (1) and that they are linear in parameters. I use these features in the lag selection criteria that I suggest below.

Volterra expansions are discussed in Priestley (1981), and they approximate Ψ using the formula

$$\Psi(Y_{t-1}^{t-p}, \theta) \approx \Psi_0 + \sum_{k=1}^{p} \Psi_{1k} y_{t-k} + \sum_{k=1}^{p} \sum_{j=1}^{k} \Psi_{2kj} y_{t-k} y_{t-j} + \sum_{k=1}^{p} \sum_{j=1}^{k} \sum_{i=1}^{j} \Psi_{3kji} y_{t-k} y_{t-j} y_{t-i} + \cdots, \tag{2}$$

which includes squares, cross-products, cubic terms and other higher order terms to capture the nonlinearities in Ψ. When Ψ is well behaved, these polynomial expansions can be justified as Taylor series expansions around $\overline{Y}_{t-1}^{t-p}$, but (2) can often provide a good approximation of Ψ even when Ψ is ill-behaved. Such expansions can become unwieldy if p is large and the expansion includes high order terms, but expansions involving just the squares and cross-products have often proven useful in practice. In the latter case (1) is given by

$$y_t \simeq \beta_0^* + \beta'^* Y_{t-1}^{t-p} + \sum_{k=1}^{k=p} \sum_{j=1}^{j=k} \Psi_{2kj} y_{t-k} y_{t-j} + \varepsilon_t \tag{3}$$

which is linear in $(p+1) + \frac{1}{2}p(p+1)$ parameters.

Neural network approximations of Ψ are based on the intuition that any nonlinear function of Y_{t-1}^{t-p} can be approximated arbitrarily well by a linear combination of elementary nonlinear transformations of q indices of Y_{t-1}^{t-p} (i.e. $\gamma_r' Y_{t-1}^{t-p}$, $r = 1, ..., q$), for q sufficiently large (Hornik et al 1989). The approximating model of y_t is then given by:

$$y_t \simeq \beta_0^* + \beta'^* Y_{t-1}^{t-p} + \sum_{r=1}^{q} \Psi_r \phi\left(\gamma_r'(1, Y_{t-1}^{t-p})'\right) + \varepsilon_t \tag{4}$$

where ϕ is a permissible elementary function[1], and the γ_r are randomly chosen by the econometrician, independently of y_t and Y_{t-1}^{t-p}. I include a constant in ϕ because Teräsvirta et al (1993) found that this helped approximate certain types of nonlinearity. Model (4) is linear in $(1+p+q)$ parameters. Lee et al (1993) note that elements in ϕ tend to be collinear with themselves and with Y_{t-1}^{t-p}, but they resolve this difficulty by using $q^* < q$ principal components of the ϕ functions that are not collinear with Y_{t-1}^{t-p}.

The lag selection procedures that I suggest involve choosing a maximum possible lag length p^{max}, fitting the approximations of the nonlinear forms to the data for each of lags 0 to p^{max}, and then choosing the lag length p^* that minimizes AIC, HQ or BIC. Given the explanator sets used in approximations (3) and (4), the parameter penalties are $(p+1) + \frac{1}{2}p(p+1)$ in the calculations based on (3), and $(1+p+q)$ for calculations based on (4). These penalties are larger than $(p+1)$ which is used when fitting linear

[1] The elementary function, which is called the "activation function" or the "squashing function" in the neural network literature, can be any function that satisfies some continuity and denseness conditions discussed in Hornik et al (1989). The most popular one is the logistic function $\phi(z) = [1 + \exp(z)]^{-1}$.

autoregressions of order p, but we are fitting more highly parameterized models to the data. Since the approximating models can be potentially overparameterized for large p and/or q, I experiment with various subsets of explanators. For instance, when working with (3) I consider an approximation that includes just the squares (and not the cross products) of Y_{t-1}^{t-p}, so that the approximating model contains less parameters and the relevant penalty drops to $(1+2p)$. I also consider using just the first principal component of the set of $\frac{1}{2}p(p+1)$ cross products, so that the approximating models use $(2+p)$ parameters. Table 1 contains a list of the various approximating models that I consider, together with a count of how many parameters are used for each approximation.

One could work with many different versions of approximation (4). I set $q = 30$ and then use the first ten principal components of the thirty squashing functions[2]. Although Lee et al (1993) found that the first two principal components (out of ten generated) were sufficient to give neural network based nonlinearity tests power against nonlinear AR(1) and AR(2) alternatives, I suspected that more components might be needed to capture nonlinearities that had potentially longer lag structures, and this led to my choice of 10. However, to allow for more frugal approximations, I also used just the first $2p$ principal components. A comparison between these two families of models is potentially interesting, because the first keeps the number of variables in the explanator set for the nonlinearity constant while lag length increases, while the second allows the explanator set for the nonlinearity to expand with lag length.

The ability of these suggested procedures to choose a correct lag depends on whether the approximation to Ψ is "close" to Ψ, given the data. It is reasonable to expect procedures based on (3) and (4) to outperform the AR based procedures if the data has "strong" nonlinear characteristics, but that the AR based procedure might be better if our time series has very subtle nonlinearities. It is also reasonable to expect that the suggested procedures will cloud the choice of lag length when the time series is actually linear, but here one can hope that the parameter penalties that are supposed to correct for the overparameterization that occurs in this case will "do their job". These issues are studied by simulation below.

[2] I remove those components in the same basis space as the linear part of the model, prior to calculating the principal components.

3. Simulation Design

The simulation study is based on a set of DGPs that have been chosen from the applied econometrics literature. These DGPs include specifications based on Teräsvirta and Anderson's (1992) models of industrial production, Beaudry and Koop's (1993) model of US output, Rothman's (1998) models of unemployment, Anderson and Vahid's (2001) models of US output, and Martens et al's (1998) data on mispricing errors associated with S&P stocks and futures contracts. Industrial production indices, unemployment and stock returns all exhibit strong evidence of nonlinearities, whereas the nonlinearities in output are much weaker. Therefore the sample of DGPs includes processes with "strongly" nonlinear behaviour, as well as processes that are "almost linear". I also include some linear DGPs in our study, some of which are published in the above papers, and others which I obtained by estimating linear models for various economic/financial data sets and then setting my DGP parameters equal to my estimated parameters. I chose the DGPs so that they would be representative of the sorts of DGPs that econometricians encounter in practice. Some have short lag structures, others have longer lag structures, and I even include some lag structures with "holes" or "near-holes"[3] Full details of all DGPs are provided in Table 2, together with references and notes on their properties.

The error terms for our DGPs and neural network random coefficients were generated using Gauss. Error terms are drawn from the standard normal distribution and then scaled according to the standard deviation of the error term of the relevant DGP. For the neural network models I followed Lee et al (1993), rescaling all variables onto [0,1], and then drawing the hidden weights from the uniform distribution on [-2,2]. I discarded the first 1000 observations of the simulated DGPs to avoid initialization effects, and report results based on 10000 replications of the relevant DGP. I studied samples of size 100, 150, 200, 250, 300, 500, 1000 and 5000, with the last three of these being included so that I could obtain some idea of the asymptotic behaviour of the procedures. I note, however, that the results for these larger samples would also be relevant for studies of

[3] By "holes" we mean zero (or very close to zero) coefficients on intermediate lags. It is typically very hard to choose the correct lag structure, given these types of DGPs.

financial data, where samples are typically large. I report results on only a subset of the samples to conserve space, but other results are available upon request. In total, I studied five processes of order two, two of order 4, two of order five, two of order 7, and two of order nine, and the lag selection procedures considered all possible lag lengths from zero lags to ten.

4. Finite sample properties of the proposed procedures

Tables 3a, 3b, and 3c present detailed results on the performance of the AIC, Hannan-Quinn and Schwartz procedures, and Figures 1 - 4 provide visual summaries of the main patterns that seemed to emerge from these results.

Figure 1 illustrates how the standard procedures (based on the AR family of models) work for each DGP. The top row relates to linear DGPs, and illustrates with the well-known properties that (i) for small samples AIC usually dominates HQ, which usually dominates BIC; but (ii) HQ and BIC improve with sample size and eventually dominate AIC, because AIC is inconsistent, in contrast to HQ and BIC. For the AR(2) and AR(9) DGPs, the large sample properties are beginning to show for samples of only 250 - 300, while samples as small as 100 observations on the AR(5) are behaving like large samples. The latter observation is due to the relatively large coefficient on the AR(5) term, which is sufficiently large to be statistically significant in samples of 100.

Similar patterns are observed when the standard procedures are applied to nonlinear DGPs (see the remaining two rows in Figure 1). However, the latter graphs also suggest that the inconsistency of AIC becomes evident earlier (i.e. for smaller sample sizes) for nonlinear processes relative to linear processes, while the improvement in HQ and BIC with sample size seems to be more rapid.

Figure 2 illustrates the differences between standard procedures (based on the AR family of models) and approximating procedures (based on other model families), when the true DGP is linear. Since the true DGP is linear, one might expect that the former procedure will dominate, while the other procedures will have lower, but hopefully non-trivial ability to choose lag length. This seems to occur for the AR(2) and AR(9) processes, which are both “weak” in the sense that they do not generate signals that allow information criteria based on small samples to accurately select lag length. It also occurs for the AR(5), when BIC based procedures are used. However, when AIC or HQ based procedures are applied to the AR(5) process, the SQ, CR and PCC procedures almost always perform better than the AR based procedures, except when samples are very small. Given that the AR(5) process is "strong", in the sense that large sample behaviour is already evident for the AR versions of AIC and HQ when samples are relatively small, this latter finding suggests that the better performance of the nonlinear criteria is a large sample phenomenon. Possibly, this phenomenon arises because the nonlinear families of models use more parameters at each lag length than do linear models, and this make it easier to discriminate between different lag lengths. It is noteworthy that for larger samples, the nonlinear procedures also outperform linear procedures for the AR(2) and the AR(9), and the BIC nonlinear procedures also outperform the linear BIC procedures. Differences between the AR and other lines on each graph in Figure 2 measure the "cost" of applying nonlinear procedures to linear DGPs. These can be very high, especially if BIC-CR is applied to a "strong" DGP. However, this cost decreases with sample size and eventually becomes negative. The last panel in Figure 2 provides an illustration of this.

Figure 3 illustrates the differences between standard procedures (based on the AR family of models) and approximating procedures (based on other model families), when the true DGP is nonlinear. Since the true DGP is nonlinear, the hope is that the latter procedures will dominate, while the standard procedures based on the AR family will have lower ability to choose lag length. This clearly occurs for the ESTAR(2), TAR(2) and the LSTAR(5) processes illustrated in Figure 3(b), where the SQ CR and PCC procedures almost always dominate the AR based procedures. It also occurs to a lesser degree for samples of more than 200 of the CDR(2) and LSTAR(2) processes, when AIC based procedures are used. The N10 and NM2 procedures outperform the AR procedures only rarely, and while the CR procedure often dominates all others (especially for "strong" DGPs and for larger samples (>500 observations)), it is usually the worst for samples of 100. It is interesting to note that there is rarely any substantial difference between the performances of the SQ and PCC procedures.

For samples of 100, there seems to be little advantage to using a nonlinear selection criterion, even when the DGP is nonlinear. The SQ and PCC procedures sometimes work better when the true DGP is truly nonlinear, but this increase in accuracy is never more than 5% in absolute value. Given that the SQ and PCC procedures use fewer parameters to model the nonlinearity while the worst performers in small samples (CR, N10 and NM2 procedure) use many more parameters, it seems that parsimony is essential in small sample settings. Thus, AR based procedures seem best.

The picture starts to change once samples grow to about 200 observations, but this depends on the relative "strength" of the nonlinear process, and how soon the asymptotic properties of each selection procedure start to set in. Figure 3 has been roughly arranged in order of "strength", so that the "weaker" DGPs appear first in Figure 3a, and then the "stronger" DGPs appear later in Figure 3b.[4] For weak processes, lag selection based on HQ-AR or BIC-AR dominate selection based on nonlinear versions of AIC, even though the latter dominate AIC-AR. Thus, although the nonlinear versions of AIC are now working better than standard AIC, there is little point in using them, because linear versions of HQ and BIC perform better still. The same is true for the stronger processes illustrated in Figure 3b, where there can be up to a 20% improvement when AIC-CR is used rather than AIC-AR. Once again, there is little point in "capitalizing" on these relative benefits when the linear versions HQ and BIC outperform all versions of AIC, but now several of the nonlinear versions of HQ and BIC perform even better. HQ-SQ, HQ-PCC, BIC-SQ and BIC-PCC offer reliable but small improvements over HQ-AR and BIC-AR. The improvements are small, simply because the latter have accuracy rates of well over 80%. The HQ-CR and BIC-CR procedures can have even higher accuracy rates of well over 99%, but the CR criterion seems to be quite unreliable, in that it either works really well, or it doesn't work at all. This is because the CR performance curves are often shaped like a logistic curve[5] with a steep slope (γ), so that essentially

[4] Note, from Figure 1, that for the three processes in Figure 3a, (CDR(2), LSTAR(2) and ESTAR(9)), that AIC-AR is still improving with sample size and HQ-AR and BIC-AR have only just started to dominate AIC-AR. For the three processes in Figure 3b, (LSTAR(5), TAR(2) and ESTAR(2)), Figure 1 shows that HQ-AR and BIC-AR already dominate AIC-AR, which has already started to decrease with sample size

[5] A logistic function in sample size t is given by $f(t) = (1 + \exp\{-\gamma(t-c)\})^{-1}$ for $\gamma > 0$.

one has to pass a certain sample size threshold (c) before good performance is observed. For the weaker processes, this threshold has not yet been reached for samples of 300, so that CR procedures hardly work at all. For the stronger processes, the threshold occurs for samples of less than 100, so that we observe the flat part of the top of the curve, and associated good performance.

The results for samples of 5000 are not reported, but the simulations show the usual inconsistency associated with AIC in large samples. Some of this is already evident in samples of 1000. This is particularly so for the "stronger" DGPs when AIC-AR is used, but it is also observed for the CDR(2) and ES(9) processes. While the AIC-AR procedure usually shows evidence of inconsistency first (i.e. for relatively smaller samples), the performances of other nonlinear procedures based on AIC also decline after a certain point. This is true for both linear and nonlinear DGPs. I found no evidence that any of the HQ or BIC based procedures were inconsistent when applied to linear DGPs, but the performance of both linear and nonlinear procedures based on HQ or BIC became inconsistent for nonlinear DGPs. Thus, it appears that one cannot rely on standard HQ or BIC when the true DGP is nonlinear. It seems possible that one might be able to maintain consistency of HQ and BIC for nonlinear processes by using higher order approximations as the sample size grows, but this issue is not explored any further here.

Taken together, the accuracy results suggest using AIC or HQ based on AR models for small samples (of less than 150 observations). Nonlinear procedures are generally not working well for small samples, even if the true DGP is nonlinear. HQ and BIC based on the SQ and PCC nonlinear approximations can be useful for moderate samples (of 150 -300), especially if the true DGP is nonlinear. Given that typical macroeconomic data sets usually consist of forty to forty five years of quarterly observations (i.e 160 - 180 observations), this finding is potentially useful for applied macroeconomists. Unfortunately the practical question of whether these nonlinear procedures will work in any given situation depends on whether the true DGP is nonlinear, and the practitioner doesn't generally know that in advance. However, if there are good reasons to suspect nonlinearities (because, for instance, one is working with unemployment data which often shows strong evidence of nonlinearity), then it seems sensible to use the procedures based on SQ and PCC approximations. However, if working with a series that is unlikely to have strong nonlinearities (for instance GDP), then it seems best to stay with the standard AR based procedures.

While accuracy is desirable when building time series models, we need to recognize that mistakes will occur and consider whether certain types of mistakes are less costly than others. For instance, over prediction of lag lengths is a problem if we wish to forecast, while underprediction is a problem if we wish to test for and model nonlinearity. Tables 4a and 4b contain some statistics that cast light on these considerations. The same general pattern characterizes both AIC and BIC procedures. Relative to procedures based on AR models, the nonlinear criteria tend to under-predict lag length more and over-predict lag length less, although under-prediction does not seem to be a problem for many of the nonlinear models we studied. It is interesting to note that a comparison of similar series (say the AR(2) and the LS(2) which were both based on estimates from the same data) shows that underprediction is less likely when the true DGP is nonlinear. This is perhaps comforting when considering nonlinear modelling, because it is relatively easy to reduce a general nonlinear model to a more parsimonious specification, but much harder to work out from a specific to a more general specification.

The tables do not include results for the four DGPs based on financial series because all criteria (linear and nonlinear, based on AIC, HQ or BIC) had great difficulty in choosing the correct lag length. AIC based on the AR family performed best in each case, but accuracy ranged from .0359 to .1568 for the AR(4), .0137 to .1446 for the AR(7), .0355 to .1078 for the TAR(4), and .0168 to .1304 for the TAR(7), (where in each case the first figure relates to samples of 100 and the second relates to samples of 1000). Results for samples of 5000 were considerably better (ranging between 45 and 65% for AIC), but BIC results for this sample size were still small (between 2% and 8%), indicating that much larger samples would be needed before asymptotic properties become evident. In one sense, these findings are not unexpected, given the extremely weak correlation structure that is typically found in financial data. However, the results also illustrate how poorly our standard methods can work, when the true DGP has very weak properties.

5. Conclusion

This paper has studied the problem of lag selection for nonlinear models from the viewpoint of an applied economist. Two common approaches include the application of AIC, HQ or BIC to linear autoregressive models, or first specifying the nonlinearity and then applying the same criteria to a sequence of nonlinear models. I argue against the second of these because of its impracticality, but assess the first of these by means of simulation. In general I find that AIC applied to AR models works quite well for small samples even when the true model is nonlinear. In contrast, HQ and BIC perform quite poorly, unless the sample size is large.

I propose and study several lag selection criteria that might be useful in nonlinear settings. Some of these are based on polynomial approximations to the nonlinear DGP, while others are based on neural network approximations. The SQ and PPC procedures seems to improve lag selection performance, when applied to moderately small samples and used in conjunction with HQ and BIC. This offers potential when working with typical macroeconomic data sets. All procedures work well for larger samples of data which follow the sorts of nonlinear processes that are popular in macroeconomic modelling. However, since standard versions of HQ and BIC are also working well in this case, the nonlinear procedures improve lag selection only slightly. For large samples of data, both linear and nonlinear procedures are easy to implement, but will be inconsistent. Although harder to implement, non-parametric techniques (such as those suggested by Tjøstheim and Auestad (1994)) might improve accuracy.

The simulations show that the usual lag selection criteria are likely to have difficulty with typical macroeconomic and financial data sets. While the proposed procedures offer some improvement, this improvement is very limited. This leads to the conclusion that more work is needed to develop other techniques that are practical, but more helpful in small sample settings.

References

Akaike, H. (1974), "A New Look at Statistical Model Identification", IEEE Transactions on A, "A New Look at Statistical Model Identification", IEEE Transactions on Automatic Control, AC19, 716-723

Akaike, H (1985), "Prediction and Entropy", in A Celebration of Statistics: the ISI Centenary Volume (eds A.C. Atkinson and S.E. Fienberg), Springer, New York.

Anderson, H.M. and F. Vahid (2001), "Predicting the probability of a Recession with Nonlinear Autoregressive Leading Indicator Models", Macroeconomic Dynamics, 5, 482-505.

Auestad, B. and D. Tjøstheim (1990), "The Identification of Nonlinear Time Series: First Order Characterization and Order Determination", Biometrika (1990), 77, 669 - 687.

Beaudry, P. and G. Koop (1993), "Do Recessions Permanently Change Output?", Journal of Monetary Economics, 31, 149-163.

Davies, R. B. (1977). "Hypothesis Testing when a Nuisance Parameter is Present only under the Alternative Hypotheses", Biometrika,64, 247 - 254.

Engle, R.F. (1984). "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics". Chapter 24, in Z. Griliches and M. Intriligator (eds) Handbook of Econometrics, Vol 2, p775 - 826, North Holland.

Engle, R. F. (1982). "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation", Econometrica, 50, 987 - 1007.

de Gooijer, J. G., A. Bovas, A. Gould and L. Robinson (1985) "Methods for determining the order of an Autoregressive-Moving Average Process: A Survey", International Statistical Review, 53, 301 - 329.

Granger, C.W.J. and J. Lin (1994). "Using the Mutual Information Coefficient to Identify Lags in Nonlinear Models", *Journal of Time Series Analysis*, 15, 371 - 384.

Granger, C.W.J., E. Maasoumi and J. Racine, (2001), "A Dependence Metric for Possibly Nonlinear Processes", mimeo, University of California, San Diego.

Granger, C.W.J., and T. Teräsvirta (1993). *Modelling Nonlinear Economic Relationships*, Oxford University Press.

Haggan, V. and T. Ozaki (1981). "Modelling nonlinear Random Vibrations using an Amplitude-dependent autoregressive Time Series Model", Biometrika 68, 189 - 196.

Hamilton, J.D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle", *Econometrica*, 57, 357-384.

Hannan, E.J., and B. G. Quinn (1979), "The Determination of the Order of an Autoregression,", *Journal of the Royal Statistical Society*, Series B, 41, 190 - 195.

Hornik, K.M., M. Stinchcombe and H.L. White, (1989). "Universal Approximations of an Unknown Mapping and its Derivatives using Multi-Layer Feedforward Networks", *Neural Networks* 3, 551 - 560.

Kendall, M. (1938), "A New Measure of Rank Correlation ", *Biometrika*, 30, 81-89.

Lee, T-H, H. White and C.W.J. Granger (1993), "Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests", *Journal of Econometrics*, 56, 268-290.

Martens, M., P. Kofman and T.C.F. Vorst (1998), "A Threshold Error Correction Model for Intraday Futures and Index Returns", *Journal of Applied Econometrics*, 13, 245 - 264.

Potter, S.M. (1995), "A Nonlinear Approach to U.S. GNP", *Journal of Applied Econometrics*, 10, 109-126.

Preistly, M.B. (1981), *Spectral Analysis and Time Series*, Academic Press, London.

Rothman, P. (1998), "Forecasting Asymmetric Unemployment Rates", *Review of Economics and Statistics*, 80, 164 - 168.

Schwartz, G. (1978), "Estimating the Dimension of a Model", *The Annals of Statistics*, 6, 461 - 464.

Teräsvirta, T. and H.M. Anderson (1992), "Characterizing Nonlinearities in Business Cycles Using Smooth Transition Autoregressive Models", *Journal of Applied Econometrics*, 7, S119-S136.

Teräsvirta, T. (1994), "Specification, Estimation and Evaluation of Smooth Transition Autoregressive Models", *Journal of the American Statistical Association*, 89, 208-218.

Teräsvirta, T., C.F. Lin and C. W. J. Granger (1993), "Power of the Neural Network Linearity Test", *Journal of Time Series Analysis*, 14, 209 - 220.

Tiao, G.C. and R.S. Tsay (1994), "Some Advances in Nonlinear and Adaptive Modelling in Time Series", *Journal of Forecasting*, 13, 109-131.

Tjøstheim, D. and B. Auestad (1994). "Nonparametric Identification of Nonlinear Time Series: Selecting Significant Lags", *Journal of the American Statistical Association*, 89, 1410 - 1419.

Tong, H. (1990). *Non-linear Time Series Analysis: A Dynamical System Approach*, Oxford University Press.

Tsay, R. (1989) "Testing and Modeling Threshold Autoregressive Processes", *Journal of the American Statistical Association*", 84, 231-240.

Wooldridge, J. M. (1994), "Estimation and Inference for Dependent Processes", Ch. 45 in R. F. Engle and D. L. McFadden (eds.) *Handbook of Econometrics, Vol 4*, North Holland.

Table 1: Families of models used to approximate the DGP

Family	Approximating Equation	Parameters
AR	$y_t \simeq \beta_0^* + \beta'^* Y_{t-1}^{t-p}$	$(1+p)$
SQ	$y_t \simeq \beta_0^* + \beta'^* Y_{t-1}^{t-p} + \sum_{k=1}^{k=p} \Psi_{2k} y_{t-k}^2$	$(1+2p)$
CR	$y_t \simeq \beta_0^* + \beta'^* Y_{t-1}^{t-p} + \sum_{k=1}^{k=p} \sum_{j=1}^{j=k} \Psi_{2kj} y_{t-k} y_{t-j}$	$(1+p)+\frac{1}{2}p(p+1)$
PCC	$y_t \simeq \beta_0^* + \beta'^* Y_{t-1}^{t-p} + fpc\{y_{t-k} y_{t-j}$ for $1 \leq k, j \leq p\}$	$(2+p)$
N10	$y_t \simeq \beta_0^* + \beta'^* Y_{t-1}^{t-p} + f10opc$ of $\{\phi \quad \gamma_r' Y_{t-1}^{t-p}$ for $1 \leq r \leq 30\}$	$(11+p)$
NM2	$y_t \simeq \beta_0^* + \beta'^* Y_{t-1}^{t-p} + f(2p)opc$ of $\{\phi \quad \gamma_r' Y_{t-1}^{t-p}$ for $1 \leq r \leq 30\}$	$(1+3p)$

Note 1: fpc is the first principal component of the bracketed explanator set.

Note 2: f10opc takes the first 10 principal components of the bracketed explanator set, othogonal to the linear explanator set.

Note 3: f(2p)opc takes the first 2p principal components of the bracketed explanator set, othogonal to the linear explanator set.

Table 2: DGPs used in the simulation studies

AR(2) $y_t = 0.49 + 0.25y_{t-1} + 0.13y_{t-2} + \varepsilon_t$ with $\varepsilon_t \sim N(0, 0.89^2)$

AR(5) $y_t = 0.005+0.935y_{t-1}+0.055y_{t-2}-0.049y_{t-3}-0.609y_{t-4}+0.417y_{t-5}+\varepsilon_t$ with $\varepsilon_t \sim N(0, 0.027^2)$

AR(9) $y_t = 0.008 + 1.423y_{t-1} - 0.7347y_{t-2} + 0.3375y_{t-3} - 0.6423y_{t-4} + 0.5348y_{t-5} - 0.1115y_{t-6} + 0.0409y_{t-7} - 0.2685y_{t-8} + 0.1837y_{t-9} + \varepsilon_t$ with $\varepsilon_t \sim N(0, 0.022015^2)$.

CDR(2) $y_t = 0.35+0.24y_{t-1}+0.22y_{t-2}+0.20CRD_{t-1}+\varepsilon_t$ with $CDR_t = \max\{CDR_{t-1}, y_t\}-y_t$ and $\varepsilon_t \sim N(0, 0.89^2)$.

ES(2) $y_t = 0.325y_{t-1} - 1.777y_{t-2} + f_t \times (1.219y_{t-1} + 1.124y_{t-2}) + \varepsilon_t$ with $f_t = (1 - \exp\{-10.230 \times 200(y_{t-1})^2\})$ and ε_t ~ $N(0, 0.0576^2)$.

TAR(2) $y_t = 0.0529 + 1.349y_{t-1} - 1.665y_{t-2} + f_t \times (1.646y_{t-1} - 0.733y_{t-2}) + \varepsilon_t$ with $f_t = (1)(y_{t-1} < 0.062)$ and ε_t ~ $N(0, 0.063^2)$

LS(2) $y_t = -1.51 - 1.41y_{t-2} + f_t \times (2.04 + 0.26y_{t-1} + 1.50y_{t-2}) + \varepsilon_t$ with $f_t = (1 + \exp\{-11(y_{t-2} + 0.55)\})^{-1}$ and ε_t ~ $N(0, 0.89^2)$.

LS(5) $y_t = -0.030+0.64y_{t-1}-0.29y_{t-2}-0.64y_{t-4}+f_t\times(0.044+0.49y_{t-2}+0.45y_{t-5})+\varepsilon_t$ with $f_t = (1 + \exp\{-7.3 \times 21.6(y_{t-1} + 0.015)\})^{-1}$ and $\varepsilon_t \sim N(0, 0.0231^2)$.

ES(9) $y_t = 0.0075 + 3.03y_{t-1} - 1.31y_{t-2} - \Delta 0.49y_{t-4} + f_t \times (-1.68y_{t-1} + 0.87y_{t-2} - \Delta 0.30y_{t-8}) + \varepsilon_t$ with $f_t = (1 - \exp\{-1.54 \times 196(y_{t-1} + 0.082)^2\})$ and $\varepsilon_t \sim N(0, 0.0185^2)$.

AR(4) $y_t = 0.0033 + 0.8679y_{t-1} + 0.0429y_{t-2} + 0.0228y_{t-3} + 0.0348y_{t-4} + \varepsilon_t$ with $\varepsilon_t \sim N(0, 0.02856^2)$.

AR(7) $y_t = 0.00085+0.8976y_{t-1}-0.0142y_{t-2}-0.0073y_{t-3}-0.0002y_{t-4}+0.0121y_{t-5}+0.0011y_{t-6} + 0.0372y_{t-7} + \varepsilon_t$ with $\varepsilon_t \sim N(0, 0.0296240^2)$.

TAR(4) $y_t = I(y_{t-1} < -0.090)(0.0031 + 0.6098y_{t-1} + 0.3577y_{t-2} - 0.1996y_{t-3} + 0.1682y_{t-4}) + I(-0.090 \leq y_{t-1} < 0.062)(0.0025 + 0.8916y_{t-1} + 0.0124y_{t-2} - 0.0061y_{t-3} + 0.0220y_{t-4}) + I(0.062 \leq y_{t-1})(0.008 + 0.8547y_{t-1} + 0.0142y_{t-2} - 0.0048y_{t-3} + 0.0251y_{t-4}) + \varepsilon_t$ with $\varepsilon_t \sim N(0, 0.0248^2)$.

TAR(7) $y_t = I(y_{t-1} < -0.073)(-0.0161 + 0.6748y_{t-1} - 0.0578y_{t-2} + 0.0362y_{t-3} + 0.10321y_{t-4} - 0.0244y_{t-5} + 0.0182y_{t-6} + 0.1147y_{t-7}) + I(-0.073 \leq y_{t-1} < 0.072)(0.0002 + 0.9311y_{t-1} - 0.0048y_{t-2} - 0.0154y_{t-3} + 0.02119y_{t-4} + 0.0003y_{t-5} + 0.0016y_{t-6} + 0.0164y_{t-7}) + I(0.072 \leq y_{t-1})(0.0159 + 0.8185y_{t-1} - 0.0292y_{t-2} - 0.004275y_{t-3} - 0.0695y_{t-4} + 0.0803y_{t-5} - 0.0222y_{t-6} + 0.060y_{t-7}) + \varepsilon_t$ with $\varepsilon_t \sim N(0, 0.0294^2)$.

Notes and Sources:

The AR(2), CDR(2) and LS(2) are based on $\Delta lnGDP$ for USA (see Anderson and Vahid (2001)). These DGPs are "weak" in that coefficients and/or evidence of nonlinearity don't become statistically significant until the sample is large.

The TAR(2) and ES(2) are based on log linear detrended unemployment for the USA (Rothman (1999)). These DGPs are "strong", in that coefficients and evidence of nonlinearity are statistically significant, regardless of sample size.

The AR(5) and LS(5) are based on fourth differences of the logarithms of industrial production for Belgium (see Teräsvirta and Anderson (1992)). Both are "strong" DGPs, although LS(5) has a "hole" at lag 3. The AR(9) and ES(9) are based on similarly transformed data for the USA and Japan. Both are moderately "strong" DGPs, but the ES(9) process for Japan has "holes" (no structure for lags 6 and 7, and restrictions for lags 5 and 9).

The AR(4), AR(7), TAR(4) and TAR(7) are based on data for mispricing errors associated with the S&P 500 index and matching futures contracts. See Martens et al (1998). As is typical for financial data, the lag structure is "weak". Evidence of nonlinearity is strong, but the corresponding threshold models contain many "holes".

Table 3a: Performance of AIC based criteria

(proportion of times the correct lag is picked)

	Data Generating Process (samples of 100)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.2302**	**.7013**	**.2464**	**.3287**	.7050	.6982	**.4010**	.6526	**.5030**
SQ	.1740	.6930	.1457	.2509	.7426	.7430	.3516	**.6772**	.2912
CR	.1049	.2725	.0531	.1472	.5389	.5215	.2205	.2721	.0715
PCC	.1740	.6964	.1489	.2512	**.7567**	**.7441**	.3453	.6719	.3043
N10	.1389	.4436	.1706	.1867	.6720	.6225	.3001	.4199	.2690
NM2	.1357	.4393	.1476	.1770	.6074	.6741	.3204	.3297	.2524

	Data Generating Process (samples of 150)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.3413**	.7216	**.4224**	**.4719**	.7093	.6941	**.5311**	.6426	**.7058**
SQ	.2746	.7666	.2671	.4142	.7711	.7531	.5094	.7249	.5094
CR	.2421	**.7720**	.0547	.3745	**.8359**	**.8285**	.4878	**.7476**	.1465
PCC	.2763	.7641	.2692	.4138	.7727	.7576	.5051	.7041	.5445
N10	.2167	.5128	.2593	.3196	.6926	.6104	.4471	.4864	.3827
NM2	.2180	.5672	.1916	.3166	.6712	.7592	.4876	.3536	.4000

	Data Generating Process (samples of 200)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.4194**	.7257	**.5559**	**.5605**	.7111	.6886	.6164	.6280	**.7754**
SQ	.3571	.7824	.3882	.5222	.7766	.7553	.6124	.7267	.6522
CR	.3200	**.8674**	.0601	.5107	**.8552**	**.8477**	**.6199**	**.8252**	.2324
PCC	.3538	.7807	.3871	.5251	.7807	.7556	.6061	.7088	.6863
N10	.2748	.5263	.3551	.4193	.6995	.5956	.5511	.5385	.4371
NM2	.2818	.6050	.2515	.4129	.6764	.7958	.5918	.3216	.5398

(continued)

Table 3a: Performance of AIC based criteria (continued)

(proportion of times the correct lag is picked)

	Data Generating Process (samples of 250)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.4896**	.7240	**.6640**	.6015	.7125	.6796	.6611	.6163	**.7856**
SQ	.4295	.7820	.5064	.5928	.7849	.7648	.6720	.7291	.7142
CR	.3917	**.9005**	.0757	**.6150**	**.8642**	**.8543**	**.6944**	**.8512**	.3533
PCC	.4323	.7797	.5100	.5930	.7847	.7542	.6709	.6994	.7452
N10	.3410	.5352	.4080	.5037	.7051	.5775	.6179	.5782	.4752
NP2	.3399	.6220	.3228	.4937	.6845	.8213	.6638	.2804	.6483
	Data Generating Process (samples of 300)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.5410**	.7328	.7254	.6291	.7107	.6728	.6829	.5984	**.7885**
SQ	.4931	.7826	**.7951**	.6447	.7838	.7650	.7162	.7259	.7348
CR	.4677	**.9127**	.1010	**.6942**	**.8696**	**.8459**	**.7499**	**.8600**	.4729
PCC	.4936	.7871	.6073	.6453	.7845	.7580	.7091	.6921	.7575
N10	.4033	.5345	.4862	.5639	.6981	.5691	.6687	.6062	.4750
NM2	.4019	.6349	.3932	.5544	.6764	.8333	.7205	.2428	.7172
	Data Generating Process (samples of 1000)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	.7141	.7300	.8492	.6083	.7076	.5286	.7104	.3896	.7053
SQ	.7828	.7963	.8638	.6181	.7921	.6917	.7835	.6657	.5096
CR	**.8577**	**.9365**	.8097	**.7572**	**.8807**	.7818	.8565	**.8109**	**.7725**
PCC	.7870	.7922	**.8672**	.3858	.7939	.6658	.7721	.5076	.6964
N10	.7336	.5264	.6970	.5801	.6171	.5533	.7411	.7215	.3270
NM2	.7124	.6668	.6923	.5634	.5499	**.9264**	**.8704**	.0524	.9158

See Table 1 for a description of model families and Table 2 for the DGPs

Table 3b: Performance of HQ based criteria

(proportion of times the correct lag length is picked)

	Data Generating Process (samples of 100)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.1857**	**.8312**	**.1122**	**.2894**	.8761	.8680	**.3676**	**.8008**	**.3285**
SQ	.0868	.7733	.0244	.1461	.9240	.9231	.2394	.7896	.0906
CR	.0530	.3199	.0018	.0903	**.9630**	**.9638**	.1780	.4424	.0085
PCC	.0866	.7742	.0258	.1472	.9296	.9242	.2320	.7895	.1068
N10	.0627	.6081	.0471	.1004	.8819	.8479	.1913	.5400	.1130
NM2	.0529	.5579	.0148	.0780	.9047	.9259	.1980	.4447	.0589

	Data Generating Process (samples of 150)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.2910**	.8860	**.2356**	**.4697**	.8916	.8771	**.5419**	.8397	**.6154**
SQ	.1488	.9172	.0558	.2812	.9499	.9402	.4097	**.9120**	.2463
CR	.0839	.6865	.0000	.1854	**.9801**	**.9807**	.3126	.8025	.0039
PCC	.1464	**.9180**	.0560	.2809	.9468	.9382	.3984	.9022	.2801
N10	.0956	.7763	.0949	.1852	.9102	.8591	.3342	.6669	.2258
NM2	.0862	.8188	.0178	.1562	.9281	.9577	.3517	.5588	.1206

	Data Generating Process (samples of 200)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.3852**	.9050	**.3689**	**.6010**	.9045	.8826	**.6744**	.8468	**.7893**
SQ	.2115	.9494	.1107	.4167	.9557	.9465	.5428	.9317	.4529
CR	.1252	.9001	.0000	.2986	**.9855**	**.9824**	.4367	**.9499**	.0051
PCC	.2122	**.9498**	.1098	.4246	.9555	.9443	.5350	.9193	.4948
N10	.1288	.8206	.1558	.2861	.9195	.8485	.4567	.7132	.3315
NM2	.1290	.9018	.0347	.2555	.9407	.9698	.5001	.5671	.2456

(continued)

Table 3b: Performance of HQ based criteria (continued)

(proportion of times the correct lag is picked)

	Data Generating Process (samples of 250)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.4685**	.9067	**.4902**	**.6998**	.9080	.8809	**.7623**	.8446	**.8603**
SQ	.2770	.9586	.1849	.5440	.9603	.9507	.6600	.9395	.6325
CR	.1677	**.9739**	.0000	.4071	**.9892**	**.9853**	.5470	**.9851**	.0099
PCC	.2761	.9570	.1842	.5498	.9587	.9474	.6507	.9189	.6800
N10	.1707	.8365	.2250	.3874	.9292	.8407	.5715	.7319	.4040
NP2	.1774	.9220	.0550	.3654	.9451	.9733	.6185	.5320	.4087

	Data Generating Process (samples of 300)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.5459**	.9114	.6037	**.7733**	.9126	.8767	**.8145**	.8439	**.8843**
SQ	.3528	.9600	**.6124**	.6546	.9626	.9514	.7418	.9441	.7557
CR	.2213	**.9942**	.0000	.5236	**.9911**	**.9859**	.6474	**.9936**	.0227
PCC	.3513	.9612	.2704	.6564	.9656	.9491	.7374	.9176	.8008
N10	.2182	.8458	.3225	.5015	.9330	.8234	.6668	.7479	.4650
NM2	.2234	.9304	.0967	.4789	.9472	.9767	.7138	.4886	.5521

	Data Generating Process (samples of 1000)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.9171**	.9381	.9528	.8686	.9329	.8021	.9358	.7561	.8626
SQ	.8982	.9751	.9578	.9118	.9775	.9351	.9731	.9427	.7878
CR	.8107	**.9996**	.0223	**.9772**	**.9956**	.9815	**.9904**	**.9954**	**.9269**
PCC	.9001	.9754	**.9598**	.8976	.9741	.9221	.9699	.8219	.9061
N10	.7998	.8571	.8300	.9613	.9275	.7735	.9684	.7988	.3890
NM2	.7931	.9650	.7961	.9297	.9291	**.9937**	.9788	.1267	.9495

See Table 1 for a description of model families and Table 2 for the DGPs

Table 3c: Performance of BIC based criteria

(proportion of times the correct lag is picked)

	Data Generating Process (samples of 100)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.1021**	**.8473**	**.0215**	**.1790**	.9576	.9527	**.2468**	**.8297**	**.1107**
SQ	.0223	.5578	.0008	.0430	.9863	.9859	.0886	.6326	.0064
CR	.0072	.0246	.0000	.0165	**.9969**	**.9973**	.0438	.0751	.0000
PCC	.0221	.5558	.0005	.0426	.9866	.9847	.0876	.6388	.0008
N10	.0131	.4459	.0022	.0227	.9673	.9584	.0597	.4746	.0179
NM2	.0057	.2407	.0001	.0109	.9846	.9893	.0468	.2712	.0010

	Data Generating Process (samples of 150)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.1700**	**.9521**	**.0591**	**.3218**	.9660	.9618	**.4222**	**.9378**	**.3364**
SQ	.0410	.8672	.0012	.0939	.9921	.9885	.1873	.9025	.0317
CR	.0129	.1225	.0000	.0372	**.9989**	**.9990**	.0953	.2678	.0000
PCC	.0404	.8660	.0010	.0963	.9927	.9881	.1791	.9058	.0385
N10	.0180	.7737	.0056	.0427	.9767	.9651	.1195	.7196	.0515
NM2	.0111	.5885	.0000	.0253	.9931	.9953	.1126	.5237	.0027

	Data Generating Process (samples of 200)								
Model Family	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.2358**	.9722	**.1211**	**.4665**	.9719	.9634	**.5622**	.9522	**.5920**
SQ	.0646	.9722	.0036	.1666	.9941	.9919	.2953	**.9770**	.1026
CR	.0193	.3411	.0000	.0667	**.9993**	**.9996**	.1577	.5586	.0000
PCC	.0645	**.9728**	.0028	.1683	.9949	.9913	.2851	.9735	.1209
N10	.0243	.9187	.0131	.0720	.9801	.9639	.1801	.8235	.1170
NM2	.0181	.8437	.0000	.0513	.9942	.9970	.1861	.6708	.0118

(continued)

Table 3c: Performance of BIC based criteria (continued)

(proportion of times the correct lag is picked)

Model Family	Data Generating Process (samples of 250)								
	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.3073**	.9763	**.2028**	**.5858**	.9758	.9664	**.6775**	.9553	**.7718**
SQ	.0912	**.9925**	.0068	.2599	.9959	.9932	.4019	**.9892**	.2258
CR	.0281	.6098	.0000	.1172	**.9997**	**.9996**	.2311	.8068	.0000
PCC	.0906	.9912	.0072	.2627	.9963	.9923	.3946	.9842	.2541
N10	.0340	.9565	.0240	.1148	.9864	.9633	.2574	.8590	.2190
NP2	.0268	.9561	.0000	.0888	.9968	.9985	.2878	.7371	.0391

Model Family	Data Generating Process (samples of 300)								
	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.3809**	.9808	**.2915**	**.6948**	.9798	.9648	**.7640**	.9552	**.8746**
SQ	.1281	.9954	.1370	.3588	.9973	.9930	.5027	**.9919**	.3781
CR	.0390	.8275	.0000	.1750	**.9999**	**.9995**	.3083	.9379	.0000
PCC	.1270	**.9955**	.0173	.3643	.9972	.9927	.4938	.9874	.4166
N10	.0426	.9638	.0502	.1696	.9877	.9601	.3389	.8642	.3000
NM2	.0376	.9887	.0003	.1431	.9970	.9986	.3898	.7529	.1026

Model Family	Data Generating Process (samples of 1000)								
	AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
AR	**.9126**	.9915	**.9694**	.9690	.9880	.9389	**.9885**	.9517	**.9550**
SQ	.6652	.9987	.6874	**.9802**	.9991	.9930	.9871	.9953	.9528
CR	.3839	**1.000**	.0000	.9350	**1.000**	**.9997**	.9578	**1.000**	.0510
PCC	.6662	.9988	.6889	.9779	.9986	.9915	.9876	.9644	.9857
N10	.3706	.9749	.7670	.9174	.9923	.9452	.9623	.8824	.4790
NM2	.3853	.9998	.2467	.9233	.9975	.9996	.9829	.3626	.9710

See Table 1 for a description of model families and Table 2 for the DGPs

Table 4a: Under and over prediction when AIC procedures are used

(proportion of times under (U), correctly (C) or over (O) predicted)

Model Family		Data Generating Process (samples of 100)								
		AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
	U	.5617	.0162	.6738	.4137	.0000	.0000	.3700	.0154	.3509
AR	C	.2302	.7013	.2464	.3287	.7050	.6982	.4010	.6526	.5030
	O	.2081	.2825	.0798	.2576	.2950	.3018	.2290	.3320	.1461
	U	.6878	.0433	.7888	.5724	.0000	.0000	.4778	.0341	.4117
SQ	C	.1740	.6930	.1457	.2509	.7426	.7430	.3518	.6772	.2912
	O	.1382	.2637	.0655	.1767	.2574	.2570	.1704	.2887	.1205
	U	.5530	.0995	.4399	.4831	.0000	.0000	.3806	.0621	.3201
CR	C	.1049	.2725	.0531	.1472	.5389	.5215	.2205	.2721	.0651
	O	.3421	.6280	.5070	.3697	.4611	.4785	.3989	.6658	.6148

Model Family		Data Generating Process (samples of 200)								
		AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
	U	.3402	.0000	.3148	.1339	.0000	.0000	.1175	.0000	.0280
AR	C	.4194	.7257	.5559	.5605	.7111	.6886	.6164	.6280	.7754
	O	.2404	.2743	.1293	.3056	.2889	.3114	.2661	.3289	.1966
	U	.4822	.0001	.5092	.2517	.0000	.0000	.1897	.0002	.1395
SQ	C	.3571	.7824	.3882	.5222	.7766	.7553	.6124	.7267	.6522
	O	.1607	.2175	.1026	.2261	.2234	.2447	.1979	.2731	.2083
	U	.5917	.0044	.9114	.3504	.0000	.0000	.2510	.0016	.6149
CR	C	.3200	.8674	.0601	.5107	.8552	.8477	.6119	.8252	.2369
	O	.0883	.1292	.0285	.1389	.1448	.1523	.1371	.1732	.1482

(continued)

Table 4a: Prediction when AIC criteria are used (continued)

(proportion of times under (U), correctly (C) or over (O) predicted)

Model Family		Data Generating Process (samples of 300)								
		AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
	U	.2023	.0000	.1365	.0416	.0000	.0000	.0383	.0000	.0021
AR	C	.5410	.7328	.7254	.6291	.7107	.6728	.6829	.5984	.7885
	O	.2567	.2672	.1381	.3293	.2893	.3272	.2788	.4016	.2094
	U	.3316	.0000	.2779	.0934	.0000	.0000	.0701	.0000	.0176
SQ	C	.4931	.7826	.5998	.6447	.7838	.7650	.7169	.7259	.7348
	O	.1753	.2174	.1223	.2619	.2162	.2350	.2130	.2741	.2476
	U	.4380	.0000	.8840	.1551	.0000	.0000	.1093	.0000	.3752
CR	C	.4677	.9127	.1010	.6942	.8698	.8459	.7499	.8600	.4785
	O	.0943	.0873	.0180	.1750	.1302	.1541	.1408	.1400	.1463

Table 4b: Under and over prediction when BIC procedures are used

(proportion of times under (U), correctly (C) or over (O) predicted))

Model Family		Data Generating Process (samples of 100)								
		AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
	U	.8844	.1100	.9760	.8006	.0000	.0000	.7335	.1125	.8910
AR	C	.1021	.8473	.0215	.1790	.9576	.9527	.2468	.8297	.1107
	O	.0135	.0427	.0025	.0204	.0424	.0473	.0197	.0578	.0083
	U	.9759	.4289	.9992	.9554	.0000	.0000	.9082	.3508	.9923
SQ	C	.0223	.5578	.0008	.0430	.9863	.9859	.0886	.6326	.0064
	O	.0018	.0133	.0000	.0016	.0137	.0141	.0032	.0166	.0013
	U	.9928	.9751	1.000	.9835	.0000	.0000	.9560	.9245	.9999
CR	C	.0072	.0246	.0000	.0165	.9969	.9973	.0438	.0751	.0000
	O	.0000	.0003	.0000	.0000	.0031	.0027	.0002	.0005	.0001

(continued)

Table 4b: Prediction when BIC criteria are used (continued)

(proportion of times under (U), correctly (C) or over (O) predicted)

Model Family		Data Generating Process (samples of 200)								
		AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
	U	.7513	.0008	.8736	.5079	.0000	.0000	.4190	.0014	.3806
AR	C	.2358	.9722	.1211	.4665	.9719	.9634	.5622	.9522	.5920
	O	.0129	.0270	.0053	.0256	.0281	.0366	.0188	.0464	.0274
	U	.9343	.0205	.9961	.8306	.0000	.0000	.7023	.0129	.8913
SQ	C	.0646	.9722	.0036	.1666	.9941	.9919	.2953	.9770	.1026
	O	.0011	.0073	.0003	.0028	.0059	.0081	.0024	.0101	.0061
	U	.9807	.6589	1.000	.9323	.0000	.0000	.8422	.4414	1.000
CR	C	.0193	.3411	.0000	.0667	.9993	.9996	.1577	.5586	.0000
	O	.0000	.0000	.0000	.0000	.0007	.0004	.0001	.0000	.0000

Model Family		Data Generating Process (samples of 300)								
		AR(2)	AR(5)	AR(9)	CDR(2)	ES(2)	TAR(2)	LS(2)	LS(5)	ES(9)
	U	.6079	.0000	.7016	.2777	.0000	.0000	.2169	.0000	.0859
AR	C	.3809	.9808	.2915	.6948	.9798	.9648	.7640	.9552	.8746
	O	.0112	.0192	.0069	.0275	.0202	.0452	.0191	.0448	.0394
	U	.8711	.0003	.9832	.6372	.0000	.0000	.4942	.0002	.6088
SQ	C	.1281	.9954	.0167	.3588	.9973	.9930	.5027	.9919	.3781
	O	.0008	.0043	.0001	.0002	.0027	.0070	.0031	.0079	.0131
	U	.9610	.1725	1.000	.8249	.0000	.0000	.6913	.0620	1.000
CR	C	.0390	.8275	.0000	.1750	.9999	.9995	.3083	.9379	.0000
	O	.0000	.0000	.0000	.0000	.0001	.0005	.0004	.0001	.0000

Figure 1: Performance of Standard Model Selection Criteria
(Proportion of Correct Picks vs Sample Size)

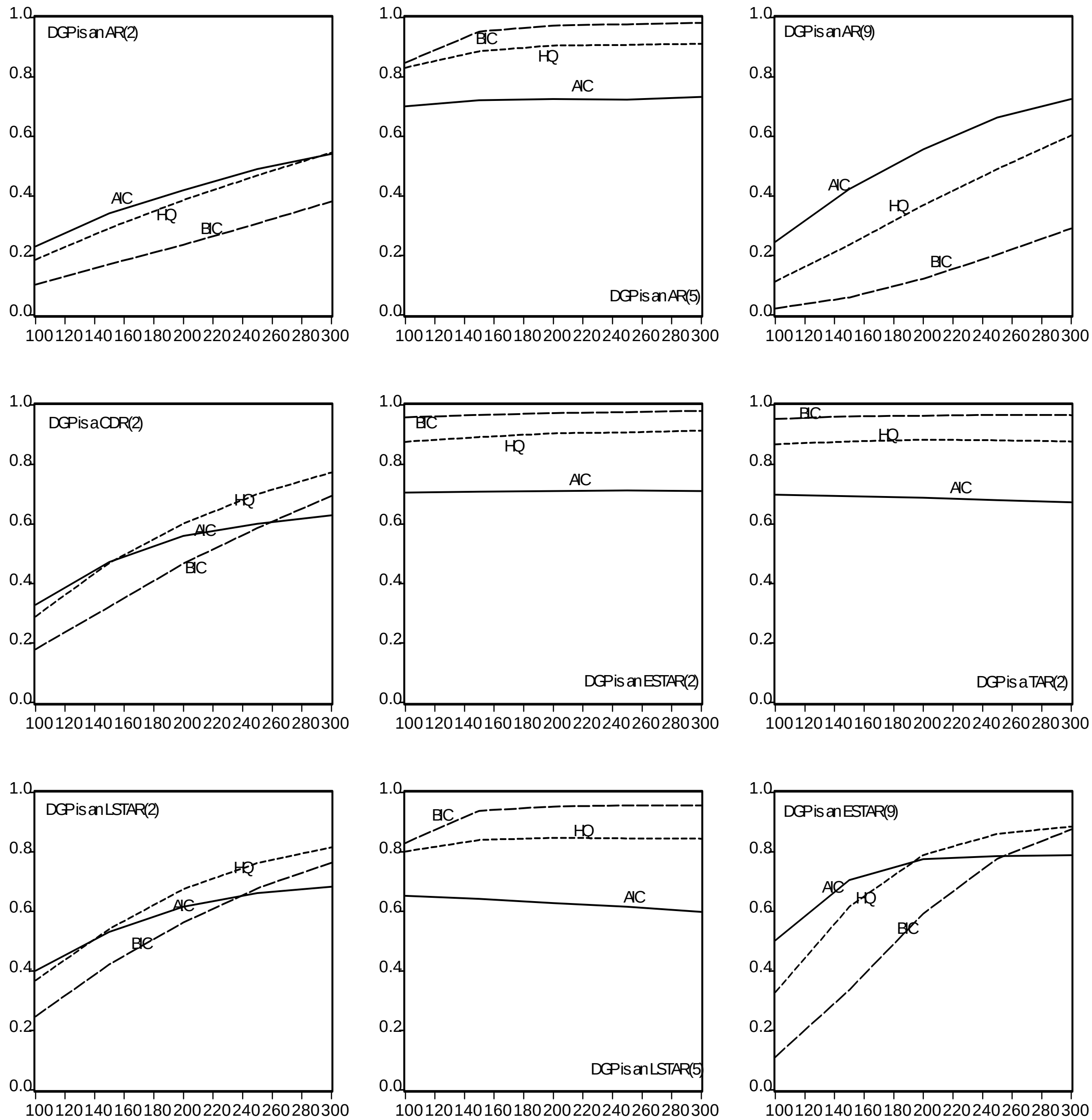

Figure 2: Performance of Selection Criteria when True DGP is Linear
(Proportion of Correct Picks vs Sample Size)

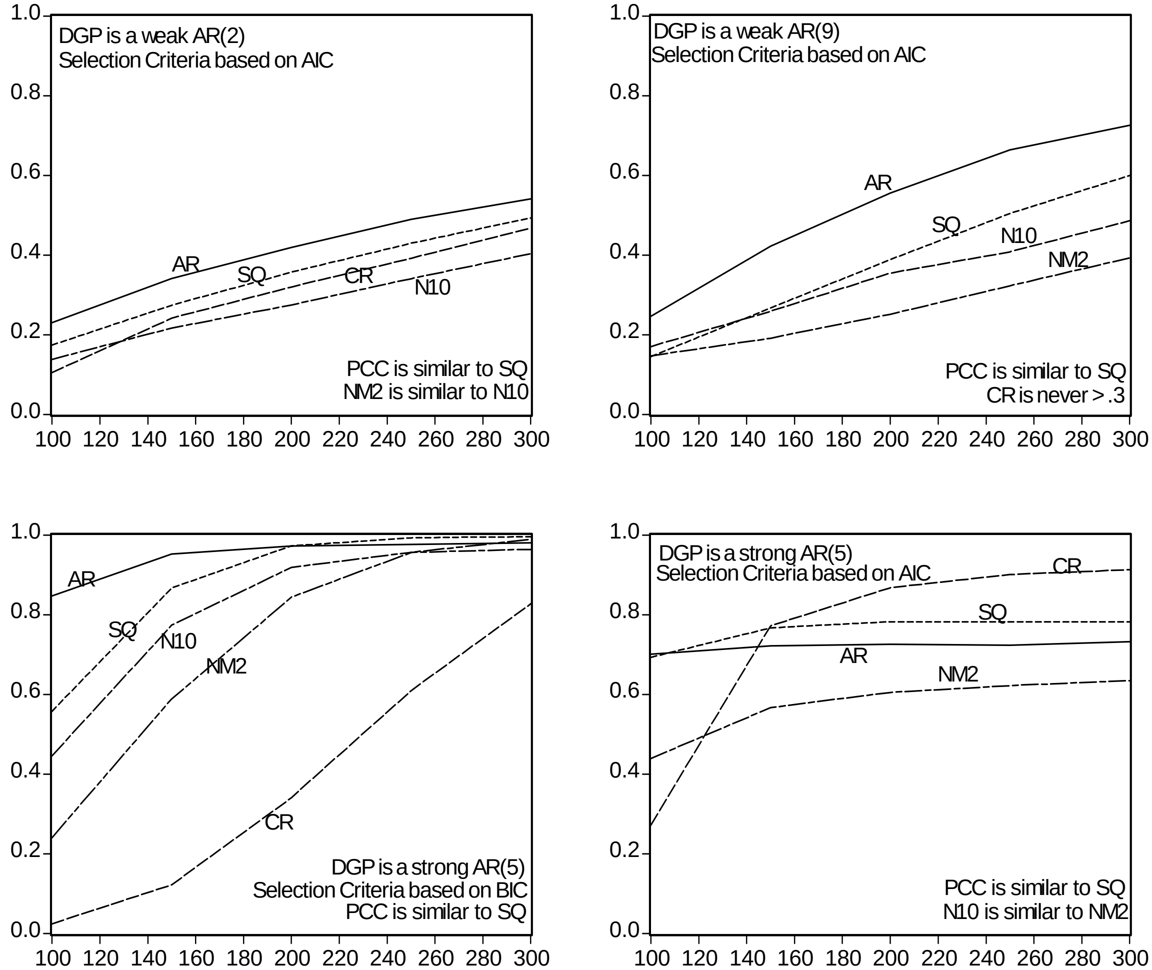

Figure 3a: Performance of Model Selection Criteria when True DGP is Nonlinear
(Proportion of Correct Picks vs Sample Size)

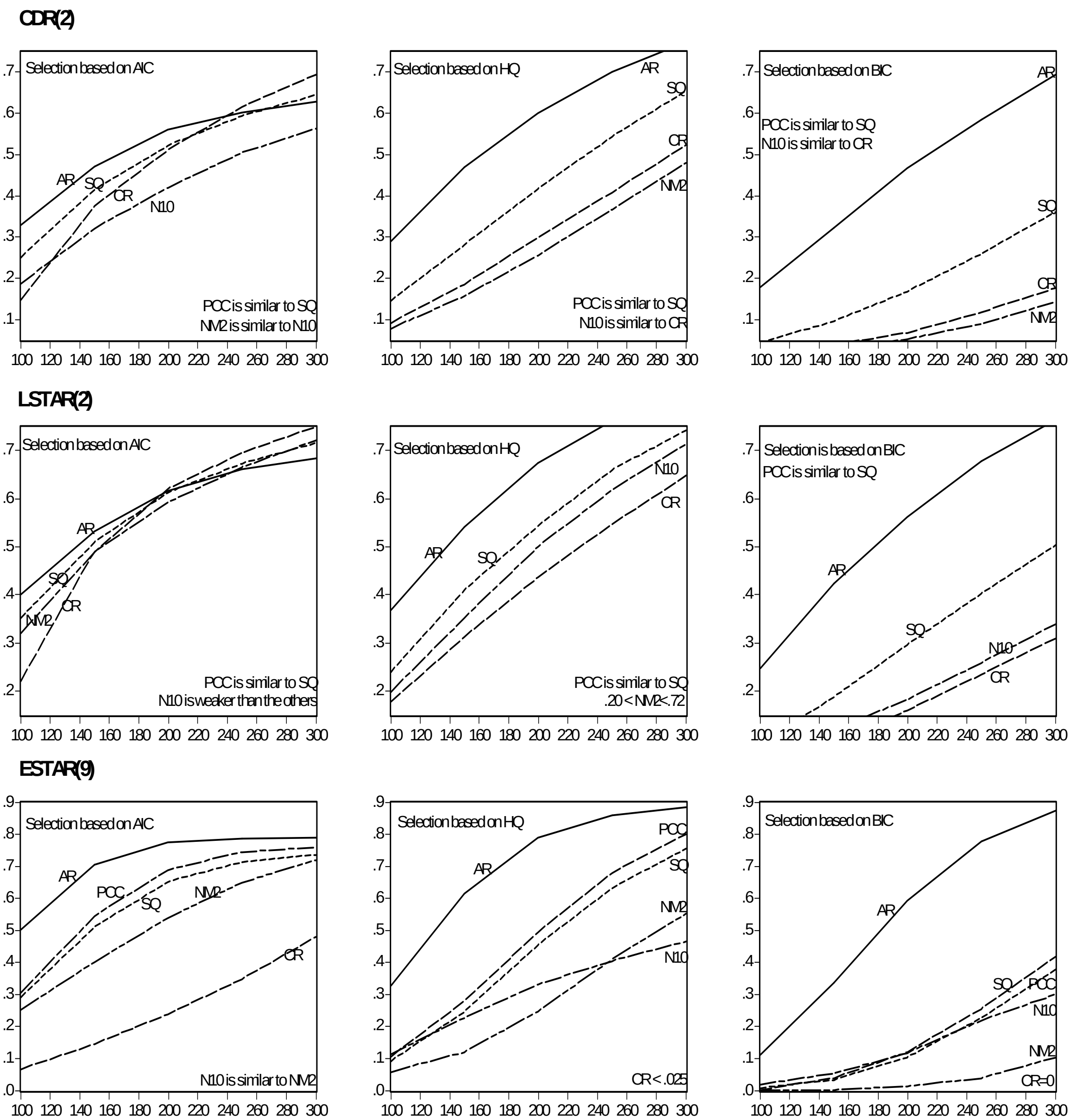

Figure 3b: Performance of Model Selection Criteria when True DGP is Nonlinear
(Proportion of Correct Picks vs Sample Size)