Model Selection in Threshold Models

George Kapetanios*
Churchill College, Cambridge
and
National Institute of Economic and Social Research, London

Abstract

This paper considers information criteria as model evaluation tools for nonlinear threshold models. Results concerning the consistency of information criteria in selecting the lag order of linear autoregressive models are extended to nonlinear autoregressive threshold models. Extensive Monte Carlo evidence of the small sample performance of a number of criteria is presented.

Keywords: Nonlinearity, Model Selection, Information Criteria, Threshold Models
JEL Classification: C15, C44, C52

*This paper is the result of work carried out for the author's Ph.D. thesis. I would like to thank Hashem Pesaran for his help and suggestions during the preparation of this paper which greatly improved its quality. I also thank Gary Koop and Sean Holly for helpful comments. Financial assistance from the Economic and Social Research Council is gratefully acknowledged.

1 Introduction

Model evaluation in econometrics has been carried out along two main lines. One is model selection and the other is hypothesis testing. Model selection is a decision theoretic approach. Given a set of rival models, its objective is to select the 'best' among them. This selection invariably involves the specification and estimation of a loss function which reflects the aspects of importance for the given modelling situation. The loss function is implicitly constructed from the perspective of the true data generation process which may or may not be required to belong to the set of rival models being investigated. The model which implies the smallest estimated loss is retained as the preferred specification. This approach provides unambiguous conclusions. At the end of the investigation one model is always accepted as the preferred specification. As a number of model selection procedures are derived using concepts from information theory, the set of these procedures will be collectively referred to as information criteria. This paper considers the properties of model selection using information criteria in the context of nonlinear threshold models.

Hypothesis testing starts by specifying two hypotheses usually denoted by H_0 and H_1. The objective is to consider the validity of the null hypothesis, H_0, against the evidence provided by the alternative hypothesis, H_1. Although the analysis is carried out under the assumption of the null being true, this assumption is only temporary and its validity is the focus of the investigation. Usually, H_0 may be obtained from H_1 by restricting a subset of the parameters of H_1. Then, H_0 is said to be nested within H_1. However, the case when the two hypotheses are not nested is of interest as well. The most usual instance of nonnested hypotheses testing involves the comparison of alternative parametric models. In the framework of nonlinear models, comparisons between alternative nonnested parametric models, for example between threshold autoregressive and Markov-switching models, is likely to be of interest since both classes may be used to model macroeconomic data.

It is common to consider model selection and nonnested hypothesis testing as rival procedures of model evaluation. However, it is clear that they are based on different premises. Unlike model selection, nonnested hypothesis testing makes a probabilistic statement concerning the validity of the null model against the evidence provided by the alternative model. Additionally, model selection always provides a preferred specification. On the other hand, although a single test of two nonnested models will always reject one of the two models, the asymmetric treatment of the null and alternative hypotheses suggests that both models should take the role of the null hypothesis in two different tests. This distinguishes nonnested hypotheses testing from nested hypotheses testing since, in the latter case, the nesting structure suggests the null and the alternative hypothesis. When both nonnested models take the role of the null hypothesis, it is possible that both models are rejected or accepted making the choice between them impossible. Discussions on the conceptual differences between model selection and nonnested hypotheses testing may be found in Amemiya (1980) and MacKinnon (1983).

Section 2 gives an account of the information criteria that will be considered. Their properties are presented and the basic statistical principles on which they are based are outlined. As most of the work done on model selection in econometrics has focused on linear models, it is important to consider extensions of existing theoretical results to threshold models. Sec-

tion 3 extends theoretical results concerning lag order selection, available for linear models, to threshold models. In addition to theoretical results which are, usually, of an asymptotic nature, the small sample performance of the information criteria needs to be evaluated. Therefore, Section 4 investigates the small sample performance of the criteria in selecting the lag order of threshold models. Section 5 presents Monte Carlo evidence on the small sample performance of information criteria in selecting between alternative threshold specifications. Section 6 concludes. Appendices 1 and 2 contain part of the proofs and technical discussions of the theoretical results.

2 Analysis of information criteria

A wide variety of information criteria have been proposed in the statistical and econometric literature. Most criteria are derived either from classical statistical principles starting with the pioneering work of Akaike (1973) or Bayesian statistical principles. In this paper we confine our attention to the following five criteria:

- Akaike's information criterion Akaike (1973) Akaike (1974)
- Schwarz's information criterion Schwarz (1978)
- Hannan-Quinn information criterion Hannan and Quinn (1979)
- Generalised information criterion (GIC) Takeuchi (1976), Stone (1977), Kitagawa and Konishi (1996)
- Informational complexity criterion (ICOMP) Bozdogan (1990)

The first three criteria are standard and require little discussion. The other two are less known and will be briefly discussed. All the above criteria are structurally similar since they Involve an estimate of the likelihood function of the model under consideration and a penalty term which depends directly or indirectly on the number of parameters of the model and the number of observations Other criteria available in the literature are Mallows' C_p, Mallows (1973), generalised cross-validation (CGV), Craven and Wahba (1979), Rissanen's minimum description length, Rissanen (1978) , and Shibata's prediction error criterion, Shibata (1980). In the next two subsections we will briefly present the GIC and ICOMP information criteria.

2.1 Generalised information criterion (GIC)

This information criterion was introduced by Takeuchi (1976), discussed in Stone (1977) and extended by Kitagawa and Konishi (1996). It extends the framework of AIC by dropping the assumption that the true model belongs to a parametric family of models which is the focus of investigation. Kitagawa and Konishi have extended the analysis even further by allowing for estimation methods other than maximum likelihood. It is well known that operationalising the principle of model selection based on the minimisation of the Kullback-Leibler (1951) information quantity as carried out by Akaike (1973) is equivalent to deriving an expression for the asymptotic bias of the sample log-likelihood as an estimator of the expected log-likelihood under the true model. In order to derive the expression for the penalty term of the GIC an analysis similar to that carried out by Akaike may be used but without imposing

the assumption that the true model belongs to the class of models being investigated. Let $f(\boldsymbol{\theta})$ denote the true density of each observation from the sample $(y_1, \ldots, y_T)$, $h(\boldsymbol{\gamma})$ denote the density of the observation for the generic model under investigation and let $l_T(.)$ denote the loglikelihood function. Then, the loss function of the criterion takes the form $-l_T(\hat{\boldsymbol{\theta}}) + \mathrm{Tr}(\hat{\boldsymbol{B}}\hat{\boldsymbol{A}}^{-1})$ where $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{A}}$ are estimates of $\boldsymbol{B}$ and $\boldsymbol{A}$ given by

$$\boldsymbol{A} = \mathrm{plim}_{T\to\infty}\left\{-\frac{1}{T}\frac{\partial^2 l_T(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\bigg|_{\hat{\boldsymbol{\gamma}}}\right\}, \quad \boldsymbol{B} = \lim_{T\to\infty} E\left\{\frac{1}{\sqrt{T}}\frac{\partial l_T(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}'}\bigg|_{\hat{\boldsymbol{\gamma}}}\frac{1}{\sqrt{T}}\frac{\partial l_T(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\bigg|_{\hat{\boldsymbol{\gamma}}}\right\}$$

A derivation of the penalty term may be found in Kitagawa and Konishi and Chapter 2 of Kapetanios (1998a). Now, if the competing models belong to the same parametric family with the true model then, under the true model, $\boldsymbol{A} = \boldsymbol{B}$, giving $\mathrm{Tr}(\boldsymbol{B}\boldsymbol{A}^{-1}) = k$ where k is the dimension of $\boldsymbol{\theta}$. Thus, we get again Akaike's criterion.

2.2 Informational complexity criterion (ICOMP)

ICOMP is a new information criterion which has been proposed by Bozdogan (1990). Although it is derived from principles of information theory it is a different procedure than AIC. It is based on the concept of complexity. Its aim is to provide the optimal tradeoff between the fit and the complexity of a model. Intuitively, complexity and parsimony, as represented by the number of parameters of a model, may seem related concepts. However, complexity has a specific meaning in information theory. This concept will be presented and a sketch of the derivation of the criterion will be given below.

For a random vector $\boldsymbol{y} = (y_1, \ldots, y_T)$, with joint density $f(\boldsymbol{y}) = f(y_1, \ldots, y_T)$ and marginal densities $f_1(y_1), \ldots, f_T(y_T)$, complexity is a measure of the dependency between its components. Such a measure may be constructed along the lines used in the construction of KLIC. The informational measure of dependence between $y_1, \ldots, y_T$ is given by

$$I(y_1, \ldots, y_T) = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(y_1, \ldots, y_T)\log\frac{f(y_1, \ldots, y_T)}{f_1(y_1)\ldots f_T(y_T)}dy_1\ldots dy_T$$

This is known as the expected mutual information and will be used as an initial measure of complexity. It turns out that the maximum expected mutual information of a T-dimensional vector following a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma} = [\sigma_{ij}]$, over all orthogonal transformations of $\boldsymbol{\Sigma}$, is a function of $\boldsymbol{\Sigma}$ alone[1] and is given by

$$\frac{T}{2}\log\frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{T} - \frac{1}{2}\log|\boldsymbol{\Sigma}|$$

Given the above, Bozdogan derives the maximal measure of complexity of a multivariate normal linear or nonlinear model. Such a model is assumed to have the following general form

$$\boldsymbol{y} = \boldsymbol{\Theta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is an $T\times 1$ observable random vector, $\boldsymbol{\Theta}$ is a deterministic component and $\boldsymbol{\epsilon}$ is a $T\times 1$ vector of random errors. $\boldsymbol{\Theta}$ depends on a vector of unknown parameters $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_k^0)'$

[1] See Bozdogan (1990, pp. 237-238).

whose estimate is denoted by $\hat{\boldsymbol{\theta}}$. The estimate of $\boldsymbol{\epsilon}$ is denoted by $\hat{\boldsymbol{\epsilon}}$. This is referred to as the residual of the model. Such a model is decomposed into two complexity generating subsystems. One is the set of estimated parameters, $\hat{\boldsymbol{\theta}}$ and the other is the residual. Then, the complexity of the model is the complexity of the vector $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\epsilon}})$. Assuming independent components and normally distributed, spherical residuals [2], the complexity of the model is equal to

$$\frac{k}{2}\log\frac{\mathrm{Tr}(\hat{\boldsymbol{\mathcal{F}}}^{-1})}{k} - \frac{1}{2}\log|\hat{\boldsymbol{\mathcal{F}}}^{-1}|, \quad \boldsymbol{\mathcal{F}} = \left[-E\frac{l_T(\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]^{-1}$$

where $\hat{\boldsymbol{\mathcal{F}}}$ is an estimate of $\boldsymbol{\mathcal{F}}$. As the aim of the criterion is to maximise fit and minimise complexity its final form is

$$-l_T(\hat{\boldsymbol{\theta}}) + \frac{k}{2}\log\frac{\mathrm{Tr}(\hat{\boldsymbol{\mathcal{F}}}^{-1})}{k} - \frac{1}{2}\log|\hat{\boldsymbol{\mathcal{F}}}^{-1}|$$

where $l_T(.)$ is the log-likelihood function of the model.

3 Consistency of lag order selection in threshold models

In this Section we will provide sufficient conditions for the consistency of lag order selection in threshold models using information criteria. The classes of self-exciting threshold autoregressive (SETAR) and Markov-switching models will be considered. Before proceeding with the presentation of the results we state briefly some available relevant results from the literature. In order for a criterion of the form $-l_T(\hat{\boldsymbol{\theta}}) + C_{T,k}$, with penalty term $C_{T,k}$, to be weakly consistent in the estimation of the lag order of a linear autoregressive model it is sufficient that[3]

$$C_{T,k} \to \infty \text{ as } T \to \infty \quad \text{and} \quad \lim_{T\to\infty}\frac{C_{T,k}}{T} = 0 \tag{1}$$

Sin and White (1996) provide a significant extension of the above result to general linear and nonlinear models. Abstracting from the more technical conditions needed for their results, it turns out that the above conditions are sufficient for a criterion to pick the true model with probability approaching one, assuming the true model belongs to the set of models being considered[4]. Further, a criterion will be strongly consistent for lag order selection, in linear autoregressive models[5] if, in addition to (1), its penalty term tends to infinity at a rate higher than or equal to $\log\log T$. It will be proven that the conditions needed for consistency in linear autoregressive models extend to threshold models.

[2]The latter assumption is very strong and is unlikely to hold in most practical situations. It is sufficient to note that even in a regression model where the errors are normal and spherical, the residuals are not spherical. This assumption is made by Bozdogan (1990) in the derivation of the second version of his criterion. In the first version this assumption is not made and as a result two sources of complexity must be evaluated. We will use the second definition following Bozdogan, Bearse, and Schlottman (1997).

[3]See, for example Lütkepohl (1991, pp. 131)

[4]If the true model is not considered in the model selection procedure then a criterion satisfying the above conditions will pick with probability approaching one the model with minimum KLIC. If more that one models attain this minimum then the model with the lowest dimension is chosen.

[5]This result has also been extended to more general setups by Sin and White (1996).

3.1 SETAR Models

The class of SETAR models is extensively discussed in Tong (1995). The piecewise linear structure underlying the class, indicates that extension of the results available for linear models concerning lag selection should be investigated.

Consider the following SETAR model

$$y_t = \phi_{j,0} + \phi_{j,1} y_{t-1} + \ldots + \phi_{j,p^0} y_{t-p^0} + \sigma_j \epsilon_t, \quad j = 1, \ldots, m, \; t = p^0, \ldots, T, \; \sigma_j > 0 \tag{2}$$

The model has m regimes. The process is in regime j if $r_{j-1} \leq y_{t-d} < r_j$ where d is an integer valued delay parameter. $r_0 = -\infty$ and $r_m = \infty$. $\{r_1 \ldots r_{m-1}\}$ is a strictly increasing sequence of parameters to be estimated. The number of regimes, m, and the delay parameter, d, are assumed known in our setup[6]. We note that p^0 is the true lag order for all the m regimes[7]. We will concentrate on the case $m = 2$ for simplicity. The results we will obtain can be readily extended to models with $m > 2$ regimes conditional on extending the results by Chan (1993) concerning the consistency and asymptotic normality of the parameter estimates to models with more than two regimes. Estimation is carried out by constructing a grid of possible values for $r_1 \equiv r$ and running the regressions

$$\boldsymbol{y}_j = \boldsymbol{X}_j \boldsymbol{\phi}_j + \boldsymbol{\epsilon}_j, \; j = 1, 2 \tag{3}$$

for each point in the threshold parameter grid, where $\boldsymbol{y}_j$ and $\boldsymbol{X}_j$ are a vector and matrix, respectively, containing the observations for regime j. $\boldsymbol{\phi}_j$ and $\boldsymbol{\epsilon}_j$ are the coefficient and error vectors for regime j. In matrix notation, $\boldsymbol{y}_j = (y_{j_1}, y_{j_2}, \ldots, y_{j_{T_j}})'$, $\boldsymbol{X}_j = (\boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{j_{T_j}})'$, $\boldsymbol{x}_{j_i} = (y_{j_i-1}, y_{j_i-2}, \ldots, y_{j_i-p})'$, $\boldsymbol{\phi}_j = (\phi_{j,1}, \ldots, \phi_{j,p})'$ $\boldsymbol{\epsilon}_j = (\epsilon_{j_1}, \ldots, \epsilon_{j_{T_j}})'$ and $\{j_1, j_2, \ldots, j_{T_j}\}$ are the time indices of the observations belonging to regime j, $j = 1, 2$. As we do not assume prior knowledge of p^0 we use p to denote the maintained lag order for the above regressions.

The aim of using the information criterion is to obtain an estimate of the true lag order p^0. It is, thus, assumed that the maximum lag order checked through an information criterion is P where $p^0 \leq P$. The following assumptions are made.

Assumption 1 *The process* $\{(y_t, y_{t-1}, \ldots, y_{t-p^0+1})\}$ *satisfying (2), viewed as a Markov chain, admits a unique invariant measure* $\pi(.)$ *such that* $\ni K, \; \rho < 1, \forall \boldsymbol{z} \in \mathbb{R}^{p^0}$ *and* $t \geq 1$, $||P^t(\boldsymbol{z}, \boldsymbol{A}) - \pi(\boldsymbol{A})|| \leq K(1 + |\boldsymbol{z}|)\rho^t$, *where* $P^t(.,.)$ *is the* t*-step transition probability,* $||.||$ *denotes the total variation norm and* $|.|$ *denotes the Euclidean norm.*

Assumption 2 ϵ_t *is absolutely continuous with a uniformly continuous, positive probability density function and finite fourth moment.*

Assumption 3 y_t *is stationary with finite fourth moment.*

Assumption 4 *The autoregressive function is discontinuous.*

[6] The number of regimes is usually dictated by theory or preliminary examination of the data. The assumption that d is known may be dropped without affecting the asymptotic results (see Chan (1993).

[7] The superscript 0 indicates true lag order values.

Remark 1 *The above assumptions are taken from Chan (1993) and are sufficient for strong consistency of all the parameters and asymptotic normality of the autoregressive parameters and* σ_j, $j = 1, 2$. *As mentioned in Chan (1993), Assumption 1 is stronger than geometric ergodicity. But if Assumption 2 holds and* $\max_{j=1,2} \sum_{i=1}^{p^0} \phi_{j,i} < 1$, *Assumption 1 is obtained following Chan, Petruccelli, Tong, and Woolford (1985) and Chan (1989)*[8]. *Further, Assumption 3 is obtained by Assumption 1, if the marginal distribution of* y_1, $\pi_1(.)$, *is the marginal distribution of the first element of a vector random variable with distribution* $\pi(.)$.

Remark 2 *It is obvious that the distribution* $\pi(.)$ *places positive probability mass on both partitions of the state space* $\mathbb{R}$ *of* y_t, *defined by the threshold parameter,* r *(See Remark B(ii) of Chan (1993)). As a result the number of observations in regime* j, $j = 1, 2$ *rises at rate* T *and it follows that* $\lim_{T\to\infty} \frac{T_1}{T} = b, \quad 0 < b < 1, \quad a.s.$

In the above setup we want to provide necessary and sufficient conditions for weak and strong consistency of lag order selection through information criteria. All of the criteria considered in Section 2 are likelihood based and introduce penalty terms to promote model parsimony. The penalty term may depend on the number of observations and depends either directly or indirectly on the dimension of the parameter vector of the model. To provide a general treatment of lag selection through information criteria we will denote the penalty term by $C_{T,k}$ where k is the dimension of the parameter vector. Note that $k = 2p+4$ for two regime SETAR models[9]. We also define k_j, $j = 1, 2$ to be the dimension of the parameter vector for regime j. We will additionally make the following two assumptions

Assumption 5 *For* $p < p^0$, *the estimate of* r, $\hat{r}$, *converges almost surely to some constant* r^*.

Assumption 6 *For given* r *and* p, $|C_{T,k} - (C_{T_1,k_1} + C_{T_2,k_2})| < C$ *almost surely, for all* $p = 1, \ldots, P$, *where* C *is a positive constant and* C_{T_j,k_j}, $j = 1, 2$ *is the penalty term that applies to the observations in regime* j *in (3). Further, for given* T *and* k *but different* r *it is assumed that the difference in the penalty terms, for a given criterion, is again almost surely bounded.*

Remark 3 *It is easy to see that for all standard criteria (Akaike, Schwarz and Hannan-Quinn) Assumption 6 holds. A proof of this statement for the first part of Assumption 6 may be found in Remark 8 of Appendix 1.*

Then the following results hold.

Theorem 1 *If* $\{y_t\}$ *is generated according to the SETAR model defined by (2), then, under Assumptions 1-6, the estimate of the lag order* p^0, $\hat{p}$, *obtained through an information criterion with penalty term* $C_{T,k}$, *is weakly consistent (i.e. converges to its true value in probability) if, and only if, the following conditions hold*

$$1.\ C_{T,k} \xrightarrow{p} \infty \qquad 2.\ \frac{C_{T,k}}{T} \xrightarrow{p} 0 \qquad 3.\ \textit{If}\ \ k^1 > k^2\ \ \textit{then}\ \ C_{T,k^1} - C_{T,k^2} \xrightarrow{p} \infty$$

[8]See also Chan, Petruccelli, Tong, and Woolford (1985) for sufficient conditions for ergodicity of a first order SETAR model.

[9]Disregarding the number of threshold parameters which remains constant throughout the search for the lag order.

Theorem 2 *If $\{y_t\}$ is generated according to the SETAR model defined by (2) then, under Assumptions 1-6, the estimate of the lag order p^0, $\hat{p}$, obtained through an information criterion with penalty term $C_{T,k}$, is strongly consistent (i.e. converges to its true value almost surely) if, and only if, the following conditions hold*

1. $\frac{C_{T,k}}{\log\log T} \stackrel{a.s.}{\to} C$ *where* C *is a constant greater than* k *or* $\frac{C_{T,k}}{\log\log T} \stackrel{a.s.}{\to} \infty$.
2. $\frac{C_{T,k}}{T} \stackrel{a.s.}{\to} 0$
3. *If* $k^1 > k^2$ *then* $\frac{C_{T,k^1} - C_{T,k^2}}{\log\log T} \stackrel{a.s.}{\to} C$, *where* C *is constant greater than* $k^1 - k^2$, *or* $\frac{C_{T,k^1} - C_{T,k^2}}{\log\log T} \stackrel{a.s.}{\to} \infty$

Remark 4 *The general theory developed by Sin and White (1996) does not apply in the case of SETAR models as the likelihood function is not continuous with respect to the threshold parameters.*

The proof for both Theorems is given in Appendix 1. For ease of exposition the proof distinguishes between the case where r is known and the case where r is estimated. The aim is to prove that an information criterion whose penalty term satisfies the conditions of the Theorems, minimised over $p = 1, \ldots, P$, obtains its minimum at p^0 in probability for Theorem 1 and almost surely for Theorem 2. The cases $p < p^0$ and $p > p^0$ are distinguished. For $p < p^0$, it is sufficient to show that the change in the likelihood, arising out of an increase in p, dominates the change in the penalty term in probability and almost surely. For $p > p^0$, the opposite must be shown to hold.

Remark 5 *The setup we are considering restricts all regimes to have the same lag order. If we wish to relax this assumption the following consistent procedure may be used. Assume a common lag order and use a consistent (weakly or strongly) information criterion to obtain its estimate, $\hat{p}$. This will, asymptotically, be equal to the maximum true lag order over all the regimes. Then, using the estimate of the threshold parameter obtained above, search within each regime, using the information criterion, over $p = 1, \ldots, \hat{p}$. The estimate obtained will be consistent for the true lag order of the regime. See Remark 11 in Appendix 1 for a justification of this procedure.*

3.2 Markov-switching models

In this subsection we will examine lag selection for Markov-Switching models. A brief review of this class will be given first. The class was introduced and analysed by Hamilton (1988, 1989, 1990, 1994, 1996) . In these models the switch between regimes is regulated by an unobserved Markov chain[10]. The presence of the unobserved Markov chain makes estimation of the model more difficult. Hamilton (1989) provides a nonlinear filter which draws inferences about the Markov chain and produces the conditional likelihood of the model which is used

[10]The most usual case involves a two-state Markov chain where the transition matrix of the chain is made up of constant parameters. However, extensions to more states have also been investigated, see Hamilton (1990). Further, the transition probabilities have been allowed to depend on the duration of the period during which the system has been in a given regime (see Durland and Mccurdy (1994)) or on a vector of exogenous variables (see Filardo (1994)).

for ML estimation of the parameters[11].

Consider the following Markov-Switching model

$$y_t = \phi_{j,0} + \phi_{j,1} y_{t-1} + \ldots + \phi_{j,p^0} y_{t-p^0} + \sigma_j \epsilon_t, \quad j = 1, \ldots, m, \; t = p^0, \ldots, T, \; \sigma_j > 0 \qquad (4)$$

The model has m regimes. The process is in regime j if $S_t = j$ where $\{S_t\}$ is an m-state first order Markov chain with transition matrix P. For example, for $m = 2$, $P = \begin{bmatrix} q_1 & 1-q_1 \\ 1-q_2 & q_2 \end{bmatrix}$. where $0 \leq q_1 \leq 1$ and $0 \leq q_2 \leq 1$. The number of regimes, m, is assumed known in our setup. For simplicity we will concentrate on the case of $m = 2$. As before, p^0 is the true lag order for both regimes. The maximum lag order checked through the information criterion is P where $p^0 \leq P$. The following assumptions are made.

Assumption 7 *q_1 and q_2 are bounded away from 0 and 1.*

Assumption 8 *For $j = 1, 2$, the roots of $1 - \phi_{j,1} z - \ldots - \phi_{j,p^0} z = 0$ lie outside the unit circle.*

Assumption 9 *$\{\epsilon_t\}$ is an i.i.d. sequence of random variables with finite $2 + \delta$ moment where $\delta > 0$.*

Remark 6 *Assumption 7 ensures that the Markov chain $\{S_t\}$ is ergodic. Therefore, by example 2 in Chapter 20 of Billingsley (1968), $\{S_t\}$ is a uniformly mixing sequence of arbitrary large size. For simplicity we will also assume that the initial distribution of the Markov chain is also the invariant distribution. Trivially, by Assumption 9, $\{\epsilon_t\}$ is a uniformly mixing sequence of arbitrarily large size.*

Unlike SETAR models, Markov-Switching models may be treated under the framework of Sin and White (1996). Unfortunately, in order to use the results of this paper a number of complicated regularity conditions are needed. These are presented in Appendix F of Kapetanios (1998a). The conditions are needed to establish pointwise and uniform laws of large numbers (LLN), central limits theorems (CLT) and laws of iterated logarithm (LIL) for y_t and its derivatives which are required to establish weak and strong consistency for the information criteria. The key to deriving the limit laws, needed to apply the results of Shin and White, lies in proving that $\{y_t\}$ is a near epoque dependent process[12] on $\{\epsilon_t, S_t\}$. Under Assumption 8, we can easily prove this. The proof may be found in Appendix 2. Since $\{\epsilon_t, S_t\}$ is a mixing process of arbitrarily large size, it then follows that $\{y_t\}$ is a mixingale[13] and the results on limit laws for mixingales may then be applied.

Following the above discussion we state the following results

Theorem 3 *If $\{y_t\}$ is generated according to the Markov-switching model defined by (4), then, under Assumptions 7-9 and the regularity conditions in Appendix F of Kapetanios (1998b),*

[11] As the likelihood of the model is often ill-behaved, suffering from singularities and multiple local maxima, Hamilton (1990) has suggested an analytical EM algorithm for parameter estimation. More details about the estimation of Markov-Switching models may be found in Kapetanios (1998b).

[12] The definition of a near epoque dependent processes is given in Appendix 2. For examples of near epoque dependent processes see Gallant and White (1988, pp. 27-31).

[13] For the definition of a mixingale see Appendix 2.

the estimate of the lag order p^0, $\hat{p}$, obtained through an information criterion with penalty term $C_{T,k}$, is weakly consistent, if the following conditions hold

$$1.\ C_{T,k} \xrightarrow{p} \infty \qquad 2.\ \frac{C_{T,k}}{T} \xrightarrow{p} 0 \qquad 3.\ \textit{If}\ \ k^1 > k^2\ \ \textit{then}\ \ C_{T,k^1} - C_{T,k^2} \xrightarrow{p} \infty$$

Theorem 4 *If $\{y_t\}$ is generated according to the Markov-switching model defined by (4), then, under Assumptions 7-9, and the regularity conditions in Appendix F of Kapetanios (1998b), the estimate of the lag order p^0, $\hat{p}$, obtained through an information criterion with penalty term $C_{T,k}$, is strongly consistent if the following conditions hold*

1. *$\frac{C_{T,k}}{\log\log T} \xrightarrow{a.s.} C$ where C is a constant greater than k or $\frac{C_{T,k}}{\log\log T} \xrightarrow{a.s.} \infty$.*
2. *$\frac{C_{T,k}}{T} \xrightarrow{a.s.} 0$*
3. *If $k^1 > k^2$ then $\frac{C_{T,k^1} - C_{T,k^2}}{\log\log T} \xrightarrow{a.s.} C$, where C is a constant greater than $k^1 - k^2$, or $\frac{C_{T,k^1} - C_{T,k^2}}{\log\log T} \xrightarrow{a.s.} \infty$*

Theorems 3 and 4 may be proven by using Propositions 4.2(a), 4.2(c), 5.2(a) and Corollary 5.4(b), of Sin and White (1996). The proofs of the Theorems involve supplying the relevant regularity conditions for the limit laws needed for the theorems in Sin and White (1996) to hold. These are provided in Appendix 3.

4 Small sample properties of lag order selection

The theoretical results obtained in Section 3 hold asymptotically. Therefore, it is necessary to investigate the properties of lag order selection in small samples.

4.1 SETAR models

The structure of the Monte Carlo experiments is as follows: The SETAR models have two regimes. Four true data generating processes (DGP) are used. These are described in Table 1. The parameters in this table, refer to the SETAR model given by equation (2). For all the DGPs the true value of r is 0. The first and the third DGPs have coefficients with diminishing absolute values for higher order lag coefficients whereas the second and the fourth DGPs have coefficients which do not fall in absolute value with the order of the lag. The signs of the coefficients and the intercepts are chosen so as to approximate an upward trending series whose differences follow a SETAR model, as the estimates of the proportion of observations belonging to each of the two regimes, given in Table 1, show. A number of macroeconomic series have been modelled similarly in the literature (See for example Potter (1995)). The absolute values of the coefficients are on purpose small to investigate the performance of the criteria for weak threshold autoregressions and to minimise small sample estimation bias. For all DGPs, $p = 1, \ldots, 6$. r is estimated by grid search. The grid contains 21 points centered around the true value. Of course, in practise the grid cannot be centered around the true

value of r, since the threshold parameter is unknown. However, as the same value of the maximised log-likelihood, obtained through the grid search, is used by all criteria, their relative performance should not be greatly affected. Indeed, limited experimentation with an alternative grid structure where quantiles of the Monte Carlo samples are used to construct the grid, indicates that the results are not affected. The delay parameter, d, takes the values 1 and 2. The true value is 1. T takes the values 150, 200, 400 and 600. The rest of the design of the experiments is common for all experiments in this paper. The error terms are constructed to be zero mean normal variates. For each replication a sample of size $T + 200$ is initially generated. The first 200 observations of each sample are discarded to minimise the effect of initial conditions[14]. For each of the DGPs and for each T, 400 replications are carried out.

We present the percentage frequencies of lag orders selected for all DGPs and for $T = 200$ in Tables 2-5. The standard errors of the estimated percentage frequencies are given in parentheses[15]. To save space, the actual frequencies of lag orders selected for all experiments are presented graphically in Figures 1 to 6 at the end of the paper. Each histogram in these Figures has twelve bars. The first six correspond to $d = 1$, $p = 1, \ldots, 6$, whereas the last six correspond to $d = 2$, $p = 1, \ldots, 6$. To facilitate the legibility of the Figures, the axes are kept constant over the five criteria for given experiments but vary between experiments.

GIC has a very similar performance to AIC. This, of course, is to be expected given the fact that GIC is a generalisation of AIC under less stringent assumptions and the fact that the basic assumption underlying AIC is satisfied[16]. It is obvious that for DGPs 2 and 4, as the number of observations increases, SC and HQ pick the right order almost perfectly, whereas AIC and GIC overestimate it for large T at expected. However, at smaller samples, AIC performs better compared to SC which underestimates the order considerably. HQ performs better overall for DGPs 2 and 4. As far as DGPs 1 and 3 are concerned it is obvious that all criteria, apart from ICOMP underestimate the order. This is to be expected given the very small and diminishing absolute values of the coefficients. However, AIC is doing slightly better with HQ improving for larger samples. It is obvious that the tendency of AIC to overestimate the order helps. ICOMP performs poorly in smaller samples especially for DGPs whose true lag order is 2. Its performance improves for $T = 400$ and 600. Overall, HQ seems to be the best criterion both at small and large samples.

We also comment on the estimation of the delay parameter, d, which, given the estimation framework, may take the values 1 and 2. Note that the true value is 1. The correct value of the delay parameter is chosen more often for all the criteria with the exception of ICOMP in DGPs 1 and 3 and smaller sample sizes and in one case of SC. Furthermore, the frequency profiles of the lag orders selected are similar for $d = 1$ and $d = 2$.

Kapetanios (1998a) investigates briefly the accuracy of the estimates of the parameters of the models, obtained during the Monte Carlo simulations. The estimates do not appear to suffer from large biases apart from the threshold parameter which is upwards biased[17].

[14]The starting values are set to zero.

[15]The standard errors are given by $100\sqrt{N^{-1}\hat{\pi}(1-\hat{\pi})}$ where $\hat{\pi}$ is the estimated percentage frequency divided by 100 and N is the number of replications for the Monte Carlo experiment.

[16]AIC is valid if the true model belongs to the parametric family of models being considered.

[17]See Kapetanios(1998a) for a more extensive Monte Carlo investigation of the threshold parameter estima-

4.2 Markov-switching models

The Monte Carlo simulations conducted to investigate lag order selection for Markov models have a similar structure with those presented for SETAR models. Again, four DGPs are considered. The main features of these DGPs are presented in Table 1. The parameters of this table refer to the Markov-switching model given by equation (4).

The specification of the autoregressive functions for each regime is similar to that of SETAR models. DGPs 1 and 3 have coefficients with small absolute values which are decreasing in the lag order while DGPs 2 and 4 have larger coefficients in absolute value which remain large for higher lag orders. The transition probabilities are both set equal to 0.5, making both regimes equally likely to occur. The experiments are carried out for 200 and 400 observations. As for SETAR models, $p = 1, \ldots, 6$. Estimation is carried out using the maximum likelihood routines of GAUSS 3.2.35[18]. The percentage frequencies of lag orders selected for all DGPs and $T = 200$ are presented in Table 6. All results are presented graphically in Figures 7 to 9. In these Figures, all histograms have six bars for $p = 1, \ldots, 6$.

As before HQ performs best in selecting the lag order followed by SC. They both do well for DGPs 2 and 4 at 200 and 400 observations. For DGPs 1 and 3 HQ and SC underestimate the true lag. AIC overestimates the true lag order for most DGPs. GIC performs similarly to AIC as expected. ICOMP significantly overestimates the true lag order for all DGPs. Once again we conclude that HQ and to a lesser degree SC are the best choices for lag selection in Markov-switching models. We also note that the performance of the criteria may be adversely affected by the small sample behaviour of the maximum likelihood estimator. A recent paper by Psaradakis and Sola (1998) provides Monte Carlo evidence which suggests that conventional asymptotic approximations for the distribution of the ML estimates are poor for small samples in Markov-switching models.

5 Selection between alternative threshold models

As new classes of nonlinear threshold models are being developed, it is important to consider formal methods of selection between alternative nonlinear models as opposed to the prevalent practise of picking, for a variety of ad hoc reasons, a class of threshold models and working within the framework of that class only.

In theory, AIC, SC and HQ[19] are not, strictly speaking, applicable in this context since the assumption concerning model selection within a parametric family is not satisfied. On the other hand, ICOMP and GIC should be useful tools as they are not bound by such strict assumptions. Three Monte Carlo experiments are carried out to investigate these issues. The experiments consider a SETAR, a Markov-switching and an EDTAR model. The EDTAR model is given by equations

$$I_{f,t} = \mathbf{1}(y_t < y_t^{\tau} - r_f), \quad I_{\text{cor},t} = \mathbf{1}(I_{f,t} + I_{c,t} = 0), \quad I_{c,t} = \mathbf{1}(y_t > y_t^{\tau} + r_c), \quad r_c, r_f > 0 \qquad (5)$$

tor in SETAR models.

[18]The user defined routines for GAUSS are modified versions of the GAUSS programs by van Norden and Vigfusson (1996) (see also Gable, van Norden, and Vigfusson (1995)).

[19]Note that HQ was originally suggested as a tool for lag selection for linear models.

$$F_t = \sum_{i=0}^{p_r} \left[(y^{\tau}_{t-i} - r_f - y_{t-i}) \prod_{j=0}^{i} I_{f,t-j} \right], \quad C_t = \sum_{i=0}^{p_e} \left[(y_{t-i} - y^{\tau}_{t-i} - r_c) \prod_{j=0}^{i} I_{c,t-j} \right] \tag{6}$$

$$x_t = \phi_0 + \Phi(L)x_t + \theta_f F_{t-1} + \theta_c C_{t-1} + h_t \epsilon_t \tag{7}$$

where $x_t = \Delta y_t$, $h_t = \sigma_{\text{cor}} I_{\text{cor},t-1} + \sigma_f I_{f,t-1} + \sigma_c I_{c,t-1}$, $\{\epsilon_t\}$ is an i.i.d. sequence of disturbances and $\mathbf{1}(.)$ denotes the indicator function. p_r, p_e are the lag orders for the effects of past deviations from the trend on the current x_t and $\Phi(L)$ is a lag polynomial of order p. y^{τ}_t is the unobserved trend process of y_t estimated by a recursive Hodrick-Prescott filter. More details about the EDTAR model may be found in Kapetanios (1999). The parameter values for the generation of the Monte Carlo samples are given in Table 7.

The autoregressive structure of the regimes of the SETAR and Markov-switching models is the same so as to minimise the distance between the two models. The experiments are carried out for samples of 200 observations only, as the treatment of Markov and EDTAR models is, computationally, very intensive. For each of the three experiments a different specification from Table 7 is used as the true DGP. The percentage frequencies of models selected for each criterion are given in Table 8.

Under a Markov DGP all criteria perform well choosing the true model most often. ICOMP performs best choosing the true model 96 % of the time. GIC has the least convincing performance and chooses the Markov model 87 % of the time. When a SETAR DGP is considered the performance of the criteria deteriorates significantly. All the criteria, apart from ICOMP, still pick the SETAR model most often but the highest selection frequency is obtained by AIC which picks the SETAR model 61.25 % of the time. ICOMP picks the Markov model more often than the SETAR model casting doubts on its impressive performance for the Markov DGP. It is likely that the inclusion of q_1 and q_2 in the covariance matrix, accentuates its block diagonality, compared to the SETAR covariance matrix, and reduces complexity. Under the EDTAR DGP all criteria perform impressively. The highest selection frequency is by ICOMP which picks the EDTAR model 99.5 % of the time. The lowest is by AIC which picks the true model 92 % of the time. The performance of the criteria in the case of the EDTAR model may be due to the fact that the EDTAR model involves three regimes unlike the other two models. Additionally, the EDTAR model has lower dimension that either the SETAR or the Markov-switching model. In general it should be expected that the more distant, in terms of KLIC, two models are the easier it will be for the criteria to pick the true one. In general AIC, HQ and SC perform slightly better than GIC and significantly better than ICOMP. Given the results of Sin and White (1996), SC and HQ should to be preferred since they are strongly consistent. Unfortunately, these results are not valid in this case, since the assumptions concerning continuity and differentiability of the likelihood functions are not satisfied in the case of SETAR and EDTAR models.

6 Conclusion

In this paper the role of information criteria in the analysis of threshold models was investigated. Theoretical results concerning consistency of lag selection which are known for linear models have been extended to threshold autoregressive and Markov-switching models. Monte

Carlo evidence on the small sample performance of a number of criteria in lag order selection and selection across different classes of threshold models was presented. As always, the conclusions are conditional on the design of the specifications. Nevertheless, we can conclude that standard information criteria have an important role to play in model selection for nonlinear threshold models. Other information criteria such as GIC and especially ICOMP have proven less reliable. The overall relative performance of ICOMP leads to the conclusion that it is of limited potential in model selection for nonlinear threshold models.

References

Akaike, H. (1973): "Information Theory and an Extension of the Maximum Likelihood Principle," in *2nd International Symposium on Information Theory*, pp. 267–281.

——— (1974): "A New Look at the Statistical Model Identification," *I.E.E.E. Trans.Auto. Control*, AC-19, 716–723.

Amemiya, T. (1980): "Selection of Regressors," *International Economic Review*, 21(2), 331–354.

Billingsley, P. (1968): *Convergence of Probability Measures*. Wiley.

Bozdogan, H. (1990): "On the Information-Based Measure of Covariance Complexity and its Application to the Evaluation of Multivariate Linear Models," *Communications in Statistics, Theory and Methods*, 19(1), 221–278.

Bozdogan, H., P. M. Bearse, and A. M. Schlottman (1997): "Empirical Econometric Modelling of Food Consumption Using a New Informational Complexity Approach (with discussion)," *Journal of Applied Econometrics*, 12(5), 563–592.

Chan, K. S. (1989): "A Note on the Geometric Ergodicity of a Markov chain," *Advances in Applied Probability*, 21, 702–704.

——— (1993): "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Model," *Annals of Statistics*, 21(1), 520–533.

Chan, K. S., J. D. Petruccelli, H. Tong, and S. W. Woolford (1985): "A Multiple Threshold AR(1) Model," *Journal of Applied Probability*, 22, 267–279.

Craven, P., and G. Wahba (1979): "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalised Cross-Validation," *Numerical Mathematics*, 31, 377–403.

Davidson, J. (1994): *Stochastic Limit Theory*. Oxford University Press.

Durland, J. M., and T. H. Mccurdy (1994): "Duration Dependent Transition in a Markov Model of United States GNP Growth," *Journal of Business and Economic Statistics*, 12(3), 279–288.

Gable, J., S. van Norden, and R. Vigfusson (1995): "Analytical Derivatives for Markov Switching Models," Working Paper 95-7, Bank of Canada.

Gallant, A. R., and H. White (1988): *A Unified Theory for Estimation and Inference for Nonlinear Dynamic Models*. Basil Blackwell.

HAMILTON, J. D. (1988): "Rational Expectations Econometric Analysis of Changes in Regime," *Journal of Economic Dynamics and Control*, 12, 385–423.

——— (1989): "A new Approach to the Economic Analysis of Time Series," *Econometrica*, 57(2), 357–384.

——— (1990): "Analysis of Time Series Subject to Changes in Regime," *Journal of Econometrics*, 45, 39–70.

——— (1994): *Time Series Analysis*. Princeton University Press.

——— (1996): "Specification Testing in Markov-switching Times Series Models," *Journal of Econometrics*, 70, 127–157.

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.

KAPETANIOS, G. (1998a): "Essays on the Econometric Analysis of Threshold Models," Ph.D. Thesis, University of Cambridge.

——— (1998b): "A Review of Nonlinear Dynamic Models in Econometrics," Unpublished Manuscript, University of Cambridge.

——— (1999): "Threshold Models for Trended Time Series," Working Paper, Department of Applied Economics, University of Cambridge.

KITAGAWA, G., AND S. KONISHI (1996): "Generalised Information Criteria in Model Selection," *Biometrika*, 83(4), 875–890.

MACKINNON, J. G. (1983): "Model Specification Tests against Nonnested Alternatives (With Discussion)," *Econometric Reviews*, 2(1), 85–158.

MALLOWS, C. L. (1973): "Some Comments on C_p," *Technometrics*, 15, 661–675.

POTTER, S. (1995): "A nonlinear approach to US GNP," *Journal of Applied Econometrics*, 10(2), 109–125.

PSARADAKIS, Z., AND M. SOLA (1998): "Finite-Sample Properties of the Maximum Likelihood Estimator in Autoregressive Models with Markov switching," *Journal of econometrics*, 86, 369–386.

RISSANEN, J. (1978): "Modelling by Shortest Data Description," *Automatica*, 14, 465–471.

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, pp. 461–464.

SHIBATA, R. (1976): "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion," *Biometrika*, 63(1), 117–126.

——— (1980): "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process," *Annals of Statistics*, 8, 147–164.

SIN, C. Y., AND H. WHITE (1996): "Information Criteria for Selecting Possibly Misspecified Parametric Models," *Journal of Econometrics*, 71(1–2), 207–225.

STONE, M. (1977): "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *Journal of the Royal Statistical Society, Series B*, 39, 44–47.

TAKEUCHI, K. (1976): "Distribution of Information Statistics and criteria for Adequacy of Models," *Mathematical Sciences*, 153, 12–28.

TONG, H. (1995): *Nonlinear time series: A dynamical system approach.* Oxford University Press.

VAN NORDEN, S., AND R. VIGFUSSON (1996): "Regime-Switching Models: A Guide to the Bank of Canada Gauss Procedures," Working Paper 96-3, Bank of Canada.

Appendix 1:Proof of Theorems 1 and 2

The proofs for both Theorems entail common elements. A single account will be given and any analysis needed for a specific Theorem will be given at the appropriate point. The proof will concentrate on the case $m = 2$. The results may be extended to cases of more than two regimes conditionally on extending the results of Chan (1993) concerning consistency to more than two regimes. For the first part of this proof we will assume r to be known. Modifications needed, when this assumption is relaxed, are given at the second part of the proof. If r is known the analysis by Sin and White (1996) is relevant. However, we choose to provide our own setup which will be used in the case where r is estimated. The analysis used by Shibata (1976) to discuss weak consistency is not valid since the resulting model is not an autoregression. As mentioned in the main part of the paper, estimation of the coefficients of the model is carried out through two OLS regressions for the two regimes. These regressions are given by (3). The notation introduced for these regressions will be used throughout. In addition, we define the standard regression idempotent matrix $\boldsymbol{M}_j = \boldsymbol{I}_{T_j} - \boldsymbol{X}_j'(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j$ which will be used later. To motivate the use of the likelihood based information criteria we note the equivalence of maximum likelihood estimation under error normality and conditional least squares and use the normal likelihood in the specification of the criteria. The concentrated conditional log-likelihood of the SETAR model is then given by[20]

$$l_T(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = const - \frac{T_1}{2}\log\{\frac{1}{T_1}(\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\phi}_1)'(\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\phi}_1)\} - \frac{T_2}{2}\log\{\frac{1}{T_2}(\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\phi}_2)'(\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\phi}_2)\} \tag{8}$$

In this context we define the following loss function associated with the information criterion with penalty term $C_{T,k}$ which when minimised over all possible lag orders gives the estimated lag order according to the information criterion.

$$IC_T(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, C_{T,k}) = \frac{T_1}{2}\log\{\frac{(\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\phi}_1)'(\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\phi}_1)}{T_1}\} + \frac{T_2}{2}\log\{\frac{(\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\phi}_2)'(\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\phi}_2)}{T_2}\} + C_{T,k} \tag{9}$$

We now claim that minimising (9) over $p = 1, \ldots, P$ is asymptotically equivalent to minimising the two quantities below over $p = 1, \ldots, P$ using observations from each of the two regimes for each quantity. We also claim that this equivalence holds both in probability and almost surely. This claim will be proven at a later stage.

$$IC_{T_1,1}(\boldsymbol{\phi}_1, C_{T_1,k_1}) = \frac{T_1}{2}\log\{\frac{1}{T_1}(\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\phi}_1)'(\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\phi}_1)\} + C_{T_1,k_1} \tag{10}$$

$$IC_{T_2,2}(\boldsymbol{\phi}_2, C_{T_2,k_2}) = \frac{T_2}{2}\log\{\frac{1}{T_2}(\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\phi}_2)'(\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\phi}_2)\} + C_{T_2,k_2} \tag{11}$$

[20] log denotes natural logarithms.

Remark 7 *This equivalence holds only asymptotically. For small samples the conclusions reached by minimising (9) as opposed to (10) or (11) will not be the same when the Schwarz and Hannan-Quinn criteria are used.*

Remark 8 *As noted in Remark 3 we now prove that the first part of Assumption 6 holds for the standard information criteria (Akaike, Schwarz and Hannan-Quinn). For Akaike we have* $C_{T,k} = k$, $C_{T_1,k_1} = k_1$ *and* $C_{T_2,k_2} = k_2$. *Since* $k_1 + k_2 = k$, *the first part of the assumption holds for Akaike's criterion. For Schwarz's criterion,* $C_{T,k} = \frac{k}{2}\log T$, $C_{T_1,k_1} = \frac{k_1}{2}\log T_1$ *and* $C_{T_2,k_2} = \frac{k_2}{2}\log T_2$. *But* $C_{T_1,k_1} + C_{T_2,k_2} = \frac{k_1}{2}\log T + \frac{k_2}{2}\log T + \frac{k_1}{2}\log\frac{T_1}{T} + \frac{k_2}{2}\log\frac{T_2}{T}$ *which is asymptotically equal to* $\frac{k}{2}\log T + \frac{k_1}{2}\log b + \frac{k_2}{2}\log(1-b)$. *But* $\frac{k_1}{2}\log b + \frac{k_2}{2}\log(1-b)$ *is bounded by Remark 2. For the Hannan-Quinn criterion,* $C_{T,k} = k\log\log T$, $C_{T_1,k_1} = k_1\log\log T_1$ *and* $C_{T_2,k_2} = k_2\log\log T_2$. *But* $k_1\log\log T_1 + k_2\log\log T_2 = k\log\log T + k_1\log\frac{\log T_1}{\log T} + k_2\log\frac{\log T_2}{\log T}$. *Using L'Hopital's rule for limits of fractions we get that the limits of* $k_1\log\frac{\log T_1}{\log T}$ *and* $k_2\log\frac{\log T_2}{\log T}$ *are* $k_1\log\frac{1}{b}$ *and* $k_2\log\frac{1}{1-b}$ *respectively and therefore bounded.*

The decomposition of the likelihood in terms of regimes permits the search for the lag order for each regime independently since a different set of observations is used for the search in each regime. Thus, $IC_{T_1,1}$ or $IC_{T_2,2}$ may be minimised over p.

Following the above argument, we can concentrate on regime 1. The same argument can be applied to the second regime. To reduce notational burden we drop the subscript indicating the regime from the coefficient and error vectors and data matrices. From now on when the matrices $\boldsymbol{X}$ and $\boldsymbol{M}$ have superscript 0, they are constructed using the true lag order p^0. If they have no superscript then they are constructed using lag order $p \neq p^0$. When the coefficient vector $\boldsymbol{\phi}$ has superscript 0 then it refers to a model using the true lag order. Hats indicate estimated parameters. At first we consider the case where $p < p^0$. Then, weak consistency requires that

$$\lim_{T_1\to\infty} P\{IC_{T_1,1}(\hat{\boldsymbol{\phi}}, C_{T_1,k_1}) - IC_{T_1,1}(\hat{\boldsymbol{\phi}}^0, C_{T_1,k_1^0}) < 0\} = 0 \tag{12}$$

By substitution, using standard regression results and after some algebra this becomes

$$\lim_{T_1\to\infty} P\left\{\frac{\frac{1}{T_1}(\boldsymbol{\phi}^{0\prime}\boldsymbol{X}^{0\prime}\boldsymbol{M}\boldsymbol{X}^0\boldsymbol{\phi}^0 + \boldsymbol{\epsilon}'\boldsymbol{M}\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}'\boldsymbol{M}\boldsymbol{X}^0\boldsymbol{\phi}^0)}{\frac{1}{T_1}\boldsymbol{\epsilon}'\boldsymbol{M}^0\boldsymbol{\epsilon}} < e^{\frac{C_{T_1,k_1^0} - C_{T_1,k_1}}{T_1/2}}\right\} = 0 \tag{13}$$

But, $\boldsymbol{X} \subset \boldsymbol{X}^0$. Also idempotency implies positive-definiteness for $\boldsymbol{M}$. As a result

$$\boldsymbol{\phi}^{0\prime}\boldsymbol{X}^{0\prime}\boldsymbol{M}\boldsymbol{X}^0\boldsymbol{\phi}^0 > 0 \text{ in probability} \tag{14}$$

for all sample sizes. By the assumed stationarity and ergodicity of the model, we get

$$T_1^{-1}\boldsymbol{X}^{0\prime}\boldsymbol{X}^0 \xrightarrow{p} \boldsymbol{\Sigma}^0, \quad T_1^{-1}\boldsymbol{X}'\boldsymbol{X} \xrightarrow{p} \boldsymbol{\Sigma}, \quad T_1^{-1}\boldsymbol{X}'\boldsymbol{X}^0 \xrightarrow{p} \boldsymbol{\Xi} \tag{15}$$

where $\boldsymbol{\Sigma}^0$ and $\boldsymbol{\Sigma}$ are positive definite matrices. Further, by the i.i.d. assumption on ϵ_t we have

$$T_1^{-1}\boldsymbol{X}^{0\prime}\boldsymbol{\epsilon} \to_p 0, \quad T_1^{-1}\boldsymbol{X}'\boldsymbol{\epsilon} \to_p 0 \text{ and } T_1^{-1}\boldsymbol{\epsilon}'\boldsymbol{M}^0\boldsymbol{\epsilon} \to_p \sigma_1^2, \quad T_1^{-1}\boldsymbol{\epsilon}'\boldsymbol{M}\boldsymbol{\epsilon} \to_p \sigma_1^2 \tag{16}$$

For all $\delta > 0$ there exists a constant K_0 such that for all $T_1 > K_0$, the RHS of the inequality in (13) is less than $1+\delta$. This is because of the second condition of the Theorem. Now, given (14) and (16), for a fixed $\varepsilon > 0$ and for all sufficiently small $\delta > 0$, there exists another constant, K_1, such that for all $T_1 > K_1$ the following holds

$$P\left\{\frac{\frac{1}{T_1}(\boldsymbol{\phi}^{0\prime}\boldsymbol{X}^{0\prime}\boldsymbol{M}\boldsymbol{X}^0\boldsymbol{\phi}^0 + \boldsymbol{\epsilon}'\boldsymbol{M}\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}'\boldsymbol{M}\boldsymbol{X}^0\boldsymbol{\phi}^0)}{\frac{1}{T_1}\boldsymbol{\epsilon}'\boldsymbol{M}^0\boldsymbol{\epsilon}} < 1+\delta\right\} < \varepsilon \tag{17}$$

Noting that ε was arbitrary, we have proven weak consistency for $p < p^0$. Now we consider strong consistency. (10) may also be written as

$$IC_{T_1,1}(C_{T_1,k_1}) = \frac{T_1}{2}\log\hat{\sigma}_p^2 + C_{T_1,k_1} \tag{18}$$

where $\hat{\sigma}_p^2$ is the residual sum of squares of the regression of y_t on $y_{t-1},\dots y_{t-p}$, involving observations in regime 1, divided by the number of observation in that regime. A well known regression result (see for example Cramer (1946) p. 307) states that

$$\hat{\sigma}_p^2 = \hat{\sigma}_0^2(1-\hat{\rho}_{1|0}^2)(1-\hat{\rho}_{2|1}^2)\dots(1-\hat{\rho}_{p|p-1}^2) \tag{19}$$

where $\hat{\sigma}_0^2$ is the sample variance, and $\hat{\rho}_{i|i-1}$, $i = 1,\dots,p$, is the estimated partial correlation coefficient between y_t and y_{t-i} (i.e. the correlation coefficient between the residuals of the regression of y_t on $y_{t-1},\dots,y_{t-i+1}$ and the residuals of the regression of y_{t-i} on $y_{t-1},\dots,y_{t-i+1}$). Using the above result the objective function of the information criterion may be written as

$$IC_{T_1,1}(C_{T_1,k_1}) = \frac{T_1}{2}\log\hat{\sigma}_0^2 + \frac{T_1}{2}\sum_{j=1}^{p}\log(1-\hat{\rho}_{j|j-1}^2) + C_{T_1,k_1}$$

For $p < p^0$ strong convergence of the parameter estimates in the regression of of y_t on $y_{t-1},\dots,y_{t-i+1}$ and the regression of y_{t-i} on $y_{t-1},\dots,y_{t-i+1}$, guarantees that $\hat{\rho}_{i|i-1}$ estimates consistently the true partial correlation coefficient $\rho_{i|i-1}$. But $\rho_{i|i-1}$ is nonzero for $p < p_0$ and less than 1. Thus, $\log(1-\hat{\rho}_{j|j-1}^2)$ is strictly negative, and, by the second condition of Theorem 2, $IC_{T_1,1}(C_{T_1,k_1})$ cannot be minimised for $p < p^0$ proving strong consistency for $p < p^0$.

Remark 9 *The same conclusion would also follow if we observe that (14), and (15)-(16) hold almost surely as well as in probability implying that the event in the probability expression (17) occurs almost surely for sufficiently large T_1.*

Now we want to prove weak and strong consistency for $p > p^0$. Using (18) and (19) we need to analyse the behaviour of $\log(1-\hat{\rho}_{p|p-1}^2)$ for $p > p^0$. We concentrate on $\hat{\rho}_{p|p-1}$

$$\hat{\rho}_{p|p-1} = \frac{\frac{1}{T_1}\sum_{t=p}^{T_1}\left[y_t - \sum_{i=1}^{p-1}\hat{\phi}_i y_{t-i}\right]\left[y_{t-p} - \sum_{i=1}^{p-1}\hat{\phi}_i^* y_{t-i}\right]}{\hat{\sigma}_{p-1}\hat{\sigma}_{p-1}^*}$$

$\hat{\sigma}_{p-1}^{*2}$ is equal to the residual sum of squares of the regression of y_{t-p} on $y_{t-1},\dots,y_{t-p+1}$ divided by the number of observations and $\hat{\phi}_i^*$ are the estimated coefficients of that regression. The denominator of the above expression converges strongly to some positive constant. Thus, we only need to consider the numerator. Ergodicity and stationarity of $\{y_t\}$ imply that $\hat{\phi}_i^*$, $i = 1,\dots,p-1$, converge strongly to some constants, say ϕ_i^*, at the rate of $T_1^{\frac{1}{2}}$. Thus, putting $\delta_i = \hat{\phi}_i - \phi_i$ and $\delta_i^* = \hat{\phi}_i^* - \phi_i^*$, $i = 1,\dots,p-1$, the numerator of $\hat{\rho}_{p|p-1}$ is given by

$$\frac{1}{T_1}\sum_{t=p}^{T_1}\left[y_t - \sum_{i=1}^{p-1}\phi_i y_{t-i} + \sum_{i=1}^{p-1}\delta_i y_{t-i}\right]\left[y_{t-p} - \sum_{i=1}^{p-1}\phi_i^* y_{t-i} + \sum_{i=1}^{p-1}\delta_i^* y_{t-1}\right]$$

Since the errors are serially uncorrelated and orthogonal to past values of y_t, it follows that $y_t - \sum_{i=1}^{p-1}\phi_i y_{t-i}$ is uncorrelated with y_{t-i} $i = 1,\dots,p-1$, $p > p^0$, and also that $y_{t-p} - \sum_{i=1}^{p-1}\phi_i^* y_{t-i}$ is uncorrelated with y_{t-i}. This implies that, for $j = 1,\dots,p-1$,

$$\frac{1}{T_1}\sum_{t=p}^{T_1}\left[y_{t-p} - \sum_{i=1}^{p-1}\phi_i^* y_{t-i}\right]y_{t-j} = O_p(T_1^{-\frac{1}{2}}) \quad \text{and} \quad \frac{1}{T_1}\sum_{t=p}^{T_1}\left[y_t - \sum_{i=1}^{p-1}\phi_i y_{t-i}\right]y_{t-j} = O_p(T_1^{-\frac{1}{2}})$$

Thus, the numerator becomes[21]

$$\frac{1}{T_1}\sum_{t=p}^{T_1}\left[y_t-\sum_{i=1}^{p-1}\phi_i y_{t-i}\right]\left[y_{t-p}-\sum_{i=1}^{p-1}\phi_i^* y_{t-i}\right]+O_p(T_1^{-1})$$

We only need to consider

$$\left[y_t-\sum_{i=1}^{p-1}\phi_i y_{t-i}\right]\left[y_{t-p}-\sum_{i=1}^{p-1}\phi_i^* y_{t-i}\right] \tag{20}$$

But, as the two quantities involved in (20) are uncorrelated for $p > p^0$, this forms a stationary, ergodic square integrable martingale difference sequence. This implies that a central limit theorem holds[22] for $\sqrt{T_1}\hat{\rho}_{p|p-1}$ implying that $\hat{\rho}_{p|p-1} = O_p(T_1^{-1/2})$ or equivalently that $\log(1-\hat{\rho}^2_{p|p-1}) = \log(1-O_p(T_1^{-1})) = O_p(T_1^{-1})$. Then, $\frac{T_1}{2}\sum_{j=p^0+1}^{p}\log(1-\hat{\rho}^2_{j|j-1}) = O_p(1)$ implying the sufficiency of conditions 1 and 3 in Theorem 1 for weak consistency of lag selection. For strong consistency we note that by Heyde and Scott (1973) a law of iterated logarithm holds for the martingale difference in (20) implying that

$$\hat{\rho}_{p|p-1} = \zeta_p(T_1)T_1^{-\frac{1}{2}}(2\log\log T_1)^{\frac{1}{2}},\ a.s.\ \text{ for } p > p_0$$

where $\limsup \zeta_p(T_1) = 1$ and $\liminf \zeta_p(T_1) = -1$. It then follows that $\log(1-\hat{\rho}^2_{p|p-1})+2T_1^{-1}\log\log T_1 > 0$ a.s. for $p > p_0$ implying the sufficiency of conditions 1 and 3 of Theorem 2 for strong consistency of lag selection.

The above covers the sufficiency part of the proof. The necessity of condition 2 for both Theorems is obvious from what has been said above. The necessity of conditions 1 and 3 is obtained as follows. By similar arguments to those used above we can show that the change in the likelihood arising out of including one extra lag is asymptotically distributed as a χ^2-variate when $p > p^0$. This implies that any criterion whose penalty term does not tend to infinity with the number of observations cannot be weakly consistent, since it will overestimate with positive probability, asymptotically, the lag order. For strong consistency, we have that for any criterion whose penalty term $C_{T_1,p}$ does not satisfy conditions 1 and 3 of Theorem 2, $\log(1-\hat{\rho}^2_{p|p-1}) + 2T_1^{-1}\log\log T_1 > 0$ does not hold almost surely for $p > p_0$.

Remark 10 *We need to provide a justification for the validity of using the decomposition given in (10)-(11). For $p < p^0$, and under the conditions of Theorem 2 it has been shown that $IC_{T_j,j}(\hat{\boldsymbol{\phi}}_j^0, C_{T_j,k_j^0}) - IC_{T_j,j}(\hat{\boldsymbol{\phi}}_j, C_{T_j,k_j})$, $j = 1,2$, are almost surely negative and $O_{a.s.}(T)$. For $p > p^0$ the same quantities are again negative almost surely and $O_{a.s.}(\log\log T)$. But, by Assumption 6, $IC_T(\hat{\boldsymbol{\phi}}_1, \hat{\boldsymbol{\phi}}_2, C_{T,k}) - IC_{T_1,1}(\hat{\boldsymbol{\phi}}_1, C_{T_1,k_1}) - IC_{T_2,2}(\hat{\boldsymbol{\phi}}_2, C_{T_2,k_2})$ is almost surely bounded for all p. The same holds in probability. As a result the decomposition is justified.*

When r is not known but estimated the above arguments need to be extended. In this case the decomposition of IC in terms of regimes cannot be used. The cases $p < p^0$ and $p > p^0$ need to be considered The second case implies that r is estimated strongly consistently and thus what has been said above holds. Note that the rate of convergence of the estimate of r, $\hat{r}$ to its true value is T (See Chan (1993)). A lower rate would have invalidated the argument developed above. In the first case,

[21] Note that all the results hold both a.s and in probability.

[22] See Davidson (1994) pp. 383-385

consistency of the estimate of r is not guaranteed. We will assume that consistency does not hold since if it did what has been said above would hold. We reintroduce the regime subscript. Since the regime based decomposition does not hold, (12) is replaced by

$$\lim_{T\to\infty} P\{IC_T(\hat{\boldsymbol{\phi}}_1, \hat{\boldsymbol{\phi}}_2, \hat{r},\ C_{T,k}) - IC_T(\hat{\boldsymbol{\phi}}_1^0, \hat{\boldsymbol{\phi}}_2^0, \hat{r}^0, C_{T,k^0}) < 0\} = 0 \tag{21}$$

where $\hat{r}^0$ denotes the estimate of r for p^0 and $\hat{r}$ denotes the estimate of r for $p \neq p^0$. Note that implicit dependence on the threshold parameter is introduced. We define $l_j(\boldsymbol{\phi}_j, r) = \frac{T_j}{2}\log\{\frac{1}{T_j}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\phi}_j)'(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\phi}_j)\}$, $j = 1, 2$, to be the contribution to the log-likelihood[23] involved in $IC_T(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, r)$ from regime j. Then, we can write (21) as

$$\lim_{T\to\infty} P\{l_1(\hat{\boldsymbol{\phi}}_1, \hat{r}) + l_2(\hat{\boldsymbol{\phi}}_2, \hat{r}) + C_{T,k} - l_1(\hat{\boldsymbol{\phi}}_1^0, \hat{r}^0) - l_2(\hat{\boldsymbol{\phi}}_2^0, \hat{r}^0) - C_{T,k^0} < 0\} = 0 \tag{22}$$

This is equivalent to

$$\lim_{T\to\infty} P\{l_1(\hat{\boldsymbol{\phi}}_1, \hat{r}) + l_2(\hat{\boldsymbol{\phi}}_2, \hat{r}) - l_1(\hat{\boldsymbol{\phi}}_1^0, \hat{r}) - l_2(\hat{\boldsymbol{\phi}}_2^0, \hat{r}) + C_{T,k} - \tag{23}$$

$$l_1(\hat{\boldsymbol{\phi}}_1^0, \hat{r}^0) - l_2(\hat{\boldsymbol{\phi}}_2^0, \hat{r}^0) + l_1(\hat{\boldsymbol{\phi}}_1^0, \hat{r}) + l_2(\hat{\boldsymbol{\phi}}_2^0, \hat{r}) - C_{T,k^0} < 0\} = 0$$

It is sufficient to show the following

$$\lim_{T\to\infty} P\{l_1(\hat{\boldsymbol{\phi}}_1^0, \hat{r}) + l_2(\hat{\boldsymbol{\phi}}_2^0, \hat{r}) - l_1(\hat{\boldsymbol{\phi}}_1^0, \hat{r}^0) - l_2(\hat{\boldsymbol{\phi}}_2^0, \hat{r}^0) < C_{T,k^0} - C_{T,k}\} = 0 \tag{24}$$

$$\lim_{T\to\infty} P\{l_1(\hat{\boldsymbol{\phi}}_1, \hat{r}) + l_2(\hat{\boldsymbol{\phi}}_2, \hat{r}) - l_1(\hat{\boldsymbol{\phi}}_1^0, \hat{r}) - l_2(\hat{\boldsymbol{\phi}}_2^0, \hat{r}) < 0\} = 0 \tag{25}$$

Consider (24) first. The RHS of the inequality in the probability expression is $o_p(T)$, by the conditions of Theorem 1. The LHS is the log-likelihood of the model under the consistent estimate of r minus the log-likelihood of the model under an inconsistent estimate of r. Both log-likelihoods are obtained under the true lag order. By Chan (1993), it follows that the LHS is $O_p(T)$. Thus, (24) is proven. Now we turn to (25). We note that for all the terms in the expression in the probability the same value of the threshold parameter is involved. We denote the number of observations belonging to regime j under $\hat{r}$ by T_j^*, $j = 1, 2$. By substitution and rearranging terms, the inequality inside the probability in (25) becomes

$$T_1^* \log\left\{\frac{\frac{1}{T_1^*}\boldsymbol{y}_1^{*\prime}\boldsymbol{M}_1^*\boldsymbol{y}_1^*}{\frac{1}{T_1^*}\boldsymbol{y}_1^{*\prime}\boldsymbol{M}_1^{0*}\boldsymbol{y}_1^*}\right\} + T_2^* \log\left\{\frac{\frac{1}{T_2^*}\boldsymbol{y}_2^{*\prime}\boldsymbol{M}_2^*\boldsymbol{y}_2^*}{\frac{1}{T_2^*}\boldsymbol{y}_2^{*\prime}\boldsymbol{M}_2^{0*}\boldsymbol{y}_2^*}\right\} < 0 \tag{26}$$

where stars indicate that the vectors or matrices are constructed using $\hat{r}$. But the argument of both logarithms in (26) is a ratio of residual sums of squares where in both the numerator and the denominator the dependent variable is the same. As the set of regressors in the denominator include the set of regressors in the numerator, standard regression analysis states that their ratio is greater than one. Thus the LHS of (26) is positive in probability proving that (26) holds. The above concerned weak consistency. Strong consistency is obtained by noting that the event in the probability expression in (24) occurs almost surely for sufficiently large T and that the LHS of (26) is almost surely positive.

[23] We choose to denote the contribution to the likelihood by $l_j(\boldsymbol{\phi}_j, r)$ instead of $l_{T_j,j}(\boldsymbol{\phi}_j, r)$ to reduce the notational burden, although, of course, this contribution depends on the sample size.

Remark 11 *The above treatment assumed that both regimes have a common lag order. In some situations this may be considered too restrictive. In Remark 5 we proposed a procedure for obtaining lag orders under the assumption that the lag order differs across regimes. In the case where r is estimated we advocate assuming a common lag order for all regimes, at first. Then from what has been said in this Appendix we realise that the maximum lag order over all regimes will be chosen. To see that let the maintained lag order be p and let* $p < p_j^0$ *for some j where* p_j^0 *denotes the true lag order of regime j. Then, under the conditions of either Theorem 1 or 2, the rise in likelihood resulting from considering a higher p will dominate the rise in the penalty term. This will keep happening for as long as* $p < p_j^0$ *for some j. Once the maximum lag order has been obtained we can start searching for the lag orders of individual regimes using the estimates of the threshold parameters that have been obtained in the first stage of the search. These estimates will clearly be consistent. Then the analysis presented in the first part of this proof where r was assumed known is relevant and the conditions of Theorems 1 and 2 are sufficient for weak and strong consistency of lag order selection for individual regimes.*

Appendix 2:The Markov-Switching Model as a NED Process

In this Appendix we prove that a process following the Markov-switching model is NED (see also Gallant and White (1988, pp. 98)). Below we define NED processes.

Definition 1 *For a, possibly vector valued, stochastic process* $\{\boldsymbol{z}_t\}_{-\infty}^{\infty}$ *on a probability space* $(\Omega, \mathcal{F}, P)$*, let* $\mathcal{F}_{t-m}^{t+m} = \sigma(\boldsymbol{z}_{t-m}, \ldots, \boldsymbol{z}_{t+m})$*, such that* $\{\mathcal{F}_{t-m}^{t+m}\}_{m=0}^{\infty}$ *is an increasing sequence of* σ*-fields*[24]*. If, for* $\upsilon > 0$*, a sequence of integrable random variables* $\{y_t\}_{-\infty}^{\infty}$ *satisfies*

$$sup_t ||y_t - E(y_t|\mathcal{F}_{t-m}^{t+m})||_\upsilon \equiv v_m$$

and $v_m = O(m^{-\alpha})$*, then* y_t *will be said near epoque dependent in* L_υ*-norm (*L_υ*-NED) of size* $-\alpha$ *on* $\{\boldsymbol{z}_t\}_{-\infty}^{\infty}$*, where* $||.||_\upsilon$ *denotes* L_υ*-norm.*

This definition is taken from Davidson (1994) and generalises previous definitions by considering L_υ-norms, $\upsilon \geq 1$, instead of the L_2 norm. The class of NED processes is useful because it includes a number of processes widely encountered in econometrics such as linear and many nonlinear autoregressive processes. The NED property focuses on the relationship between the process $\{y_t\}$ and the underlying process $\{\boldsymbol{z}_t\}$. On its own it is of little use. However, when the underlying process, $\{\boldsymbol{z}_t\}$, is mixing, the NED property may be used to extend results on limit laws which hold for mixing processes to the process $\{y_t\}$ which may not be mixing. The fact which permits this extension is that NED processes on mixing processes are, under regularity conditions, mixingales[25]. Therefore, we can apply results on limit laws available for mixingales to NED processes.

To see that a process following the Markov-switching model is NED we investigate the two regime simple model given below

$$y_t = \begin{cases} \phi_1 y_{t-1} + \epsilon_t & \text{if } S_t = 1 \\ \phi_2 y_{t-1} + \epsilon_t & \text{if } S_t = 2 \end{cases} \tag{27}$$

[24]For a random variable x, we denote by $\sigma(x)$ the intersection of all σ-fields of the sample space Ω, with respect to which x is measurable.

[25]Given a probability space $(\Omega, \mathcal{F}, P)$, the sequence of $\{y_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ where $\{\mathcal{F}_t\}$ is an increasing sequence of σ-subfields of $\mathcal{F}$ and $\{y_t\}$ is a sequence of integrable random variables, is called an L_υ-mixingale if, for $\upsilon \geq 1$, there exist sequences of nonnegative constants $\{c_t\}_{-\infty}^{\infty}$ and $\{\zeta_m\}_{-\infty}^{\infty}$ such that $\zeta_m \to 0$ as $m \to \infty$, $||E(y_t|\mathcal{F}_{t-m})||_\upsilon \leq c_t\zeta_m$ and $||y_t - E(y_t|\mathcal{F}_{t-m})||_\upsilon \leq c_t\zeta_{m+1}$. As for the definition of NED process, this definition is taken from Davidson (1994) where again L_υ-norms, $\upsilon \geq 1$, instead of the L_2 norm is used.

where S_t is a Markov chain specified as in (4). Extension to higher order lag structures makes no difference for what follows. In this case the underlying process, $\{\boldsymbol{z}_t\}$, is given by $\{\epsilon_t, S_t\}$. Then the following Theorem may be proven

Theorem 5 *Under Assumptions 8 and 9 the process defined in (27) is L_2-NED of arbitrarily large size on $\{\epsilon_t, S_t\}$.*

Proof of Theorem 5

(27) may be written as $y_t = \sum_{\tau=0}^{\infty} \left(\prod_{u=0, S_{t-u}=1}^{u=\tau} \phi_1 \prod_{v=0, S_{t-v}=2}^{v=\tau} \phi_2 \right) \epsilon_{t-\tau}$. But, defining $\phi_{\max}$ to be the coefficient with the maximum absolute value between ϕ_1 and ϕ_2, noting that $E|\epsilon_t| < \infty$, that $|\phi_{\max}| < 1$ from Assumption 8 and that the conditional expectation is the minimum squared error predictor of y_t we have

$$||y_t - E(y_t|\mathcal{F}_{t-m}^{t+m})||_2 \leq \left|\left| y_t - \sum_{\tau=0}^{m} \left(\prod_{\substack{u=0 \\ S_{t-u}=1}}^{u=\tau} \phi_1 \prod_{\substack{v=0 \\ S_{t-v}=1}}^{v=\tau} \phi_2 \right) \epsilon_{t-\tau} \right|\right|_2 = \left|\left| \sum_{\tau=m+1}^{\infty} \left(\prod_{\substack{u=0 \\ S_{t-u}=1}}^{u=\tau} \phi_1 \prod_{\substack{v=0 \\ S_{t-v}=1}}^{v=\tau} \phi_2 \right) \epsilon_{t-\tau} \right|\right|_2 \leq$$

$$\left|\left| \sum_{\tau=m+1}^{\infty} \phi_{\max}^{\tau} \epsilon_{t-\tau} \right|\right|_2 \leq |\phi_{\max}|^m \sum_{\tau=1}^{\infty} |\phi_{\max}|^{\tau} ||\epsilon_{t-\tau-m}||_2 = \frac{|\phi_{\max}|^{m+1} ||\epsilon_t||_2}{1 - |\phi_{\max}|}$$

Consequently $v_m \to 0$ as $m \to \infty$, and more specifically $v_m = O(m^{-\gamma})$ where γ is a arbitrarily large.

Appendix 3: Regularity Conditions and Proofs for Theorems 3 and 4

In this Appendix we provide the technical regularity conditions needed for Theorems 3 and 4 and the proofs of the Theorems. The specification of the regularity conditions requires the following definitions. The sequence $\{y_t\}$ is defined on a generic probability space $(\Omega, \mathcal{F}, P)$. Let $\boldsymbol{y}_{t,p} = (y_t, \ldots, y_{t-p})'$ Let $\boldsymbol{\psi}_p$, $\boldsymbol{\Psi}_p$, $v_{t,p}(\boldsymbol{y}_{t,p}, \boldsymbol{\psi}_p)$ and $V_{T,p}(\boldsymbol{\psi}_p) = \sum_{t=1}^{T} v_{t,p}(\boldsymbol{y}_{t,p}, \boldsymbol{\psi}_p)$ denote a generic vector containing the parameters of the model, the parameter space, the log-likelihood[26] of observation t and the log-likelihood of the whole model under a maintained lag order p. Also, let $u_t(\boldsymbol{y}_{t,p})$ and μ denote the density of $\boldsymbol{y}_{t,p}$ and a measure dominating the marginal distribution of $\boldsymbol{y}_{t,p}$, $t = 1, \ldots, T$, respectively. To simplify notation, the symbols ∇ and ∇^2 are used to denote the gradient and the hessian of a function, respectively. In what follows expectations are taken with respect to the true distribution of $\boldsymbol{y}_{t,p}$.

Assumption 10 *$V_{T,p}(\boldsymbol{\psi}_p)$ is measurable-$\mathcal{F}$ and twice continuously differentiable on $\boldsymbol{\Psi}_p$ almost everywhere, for all $\boldsymbol{\psi}_p \in \boldsymbol{\Psi}_p$ for $p = 1, \ldots, P$. For all $\boldsymbol{\psi}_p \in \boldsymbol{\Psi}_p$, $E(V_{T,p}(\boldsymbol{\psi}_p))$ exists and defines an almost surely twice continuously differentiable function on $\boldsymbol{\Psi}_p$. Finally, the integral and differentiation operator in the above expectation are interchangeable.*

Assumption 11 *The parameters of the model are uniquely identified for $p = 1, \ldots, P$.*

Assumption 12 *The parameter vector which attains the supremum of the expectation of the log-likelihood of the model for $p = 1, \ldots, P$, denoted $\boldsymbol{\psi}_p^*$, lies in the interior of $\boldsymbol{\Psi}_p$.*

[26]For details on how to obtain the log-likelihood for observation t see Hamilton (1989).

Assumption 13 *$\boldsymbol{\Psi}_p$, $p = 1, \ldots, P$, is compact.*

Assumption 14 *For all points in $\boldsymbol{\Psi}_p$ lying in an open sphere of radius $\varepsilon > 0$ centered at $\boldsymbol{\psi}_p^*$, $T^{-1}\nabla^2 V_{T,p}(\boldsymbol{\psi}_p)$ is asymptotically bounded away from zero almost surely and $E[T^{-1}\nabla^2 V_{T,p}(\boldsymbol{\psi}_p)]$ is asymptotically bounded almost surely, for $p = 1, \ldots, P$.*

Assumption 15 *For the sequence $\{v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*)\}$ and the sequences of elements $\{\nabla_i v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*)\}$ and $\{\nabla^2_{i,j} v_{t,p}(\boldsymbol{\psi}_p^*)\}$ of the processes $\{\nabla v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*)\}$ and $\{\nabla^2 v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*)\}$, the following holds almost everywhere for all t and $\boldsymbol{y}^1$, $\boldsymbol{y}^2$, $p = 1, \ldots, P$, $i, j = 1, \ldots, dim(\boldsymbol{\psi}_p)$, where B_t, $B_{t,i}$ and $B_{t,i,j}$ are finite constants*

$$|v_{t,p}(\boldsymbol{y}^1, \boldsymbol{\psi}_p^*) - v_{t,p}(\boldsymbol{y}^2, \boldsymbol{\psi}_p^*)| \leq B_t \sum_{l=1}^{l=p+1} |y_l^1 - y_l^2|$$

$$|\nabla_i v_{t,p}(\boldsymbol{y}^1, \boldsymbol{\psi}_p^*) - \nabla_i v_{t,p}(\boldsymbol{y}^2, \boldsymbol{\psi}_p^*)| \leq B_{t,i} \sum_{l=1}^{l=p+1} |y_l^1 - y_l^2|$$

and

$$|\nabla^2_{i,j} v_{t,p}(\boldsymbol{y}^1, \boldsymbol{\psi}_p^*) - \nabla^2_{i,j} v_{t,p}(\boldsymbol{y}^2, \boldsymbol{\psi}_p^*)| \leq B_{t,i,j} \sum_{l=1}^{l=p+1} |y_l^1 - y_l^2|$$

Assumption 16 *There exist a sequence of positive constants $\{c_{t,1}\}$, $c_{t,1} \to \infty$, such that $\{[v_{t,p}(\boldsymbol{y}_{t,p}, \boldsymbol{\psi}_p^*) - E(v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*))]/c_{t,1}\}$ and $\{[\nabla^2_{i,j} v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*) - E(\nabla^2_{i,j} v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*))]/c_{t,1}\}$ are uniformly L_2-bounded, and*

$$\sum_{t=1}^{\infty} \left|\left| \frac{v_{t,p}(\boldsymbol{y}_{t,p}, \boldsymbol{\psi}_p^*) - E(v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*))}{c_{t,1}} \right|\right|_2^2 < \infty$$

$$\sum_{t=1}^{\infty} \left|\left| \frac{\nabla^2_{i,j} v_{t,p}(\boldsymbol{\psi}_p^*) - E(\nabla^2_{i,j} v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*))}{c_{t,1}} \right|\right|_2^2 < \infty$$

for $i, j = 1, \ldots, dim(\boldsymbol{\psi}_p)$, $p = 1, \ldots, P$.

Assumption 17 *$E(\nabla_i v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*)) = 0$, $\{\nabla_i v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*)\}$ is L_2-bounded and $0 < \lim_{T\to\infty} T^{-1}\sigma^2_{T,p,i} < \infty$, where $\sigma^2_{T,p,i} = Var\left[\sum_{t=1}^{T} \nabla_i v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*)\right]$, for all t, $i = 1, \ldots, dim(\boldsymbol{\psi}_p)$, $p = 1, \ldots, P$.*

Assumption 18 *$\sigma^2_{T,p,i} < \infty$ where $\sigma^2_{T,p,i}$ is defined in Assumption 17. There exists a sequence of positive constants $c_{t,2}$ such that*

$$sup_t \left|\left| \frac{\nabla_i v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*) - E(\nabla_i v_{t,p}(\boldsymbol{y}_{t,p}\boldsymbol{\psi}_p^*))}{c_{t,2}} \right|\right|_{2+\delta}, \quad \delta > 0$$

and $sup_T T[\max_{1\leq t\leq T}\{c_{t,2}\}]^2 < \infty$, $i = 1, \ldots, dim(\boldsymbol{\psi}_p)$, $p = 1, \ldots, P$.

Assumption 19 *$u_t(\boldsymbol{y}_{t,p})$ is continuous for all t.*

Assumption 20

$$\int sup_{t\geq 1,\boldsymbol{\psi}_p\in\boldsymbol{\Psi}_p}|v_{t,p}(\boldsymbol{y},\boldsymbol{\psi}_p)|u_t(\boldsymbol{y})\mu(d\boldsymbol{y})<\infty,\quad p=1,\ldots,P$$

.

Assumption 21 *For each element of* $\nabla^2 v_{t,p}(\boldsymbol{y}_{t,p},\boldsymbol{\psi}_p)$ $\nabla^2_{i,j}v_{t,p}(\boldsymbol{y}_{t,p},\boldsymbol{\psi}_p)$, $i,j=1,\ldots,dim(\boldsymbol{\psi}_p)$

$$\int sup_{t\geq 1,\boldsymbol{\psi}_p\in\boldsymbol{\Psi}_p}|\nabla^2_{i,j}v_{t,p}(\boldsymbol{y},\boldsymbol{\psi}_p)|u_t(\boldsymbol{y})\mu(d\boldsymbol{y})<\infty,\quad p=1,\ldots,P$$

.

Assumption 22 *1. For* $p<p^0$

$$E\left\{\frac{1}{T}\sum_{t=1}^{T}\left[v_{t,p}(y_t,\boldsymbol{\phi}^*_{p^0})-v_{t,p}(y_t,\boldsymbol{\phi}^*_p)\right]\right\}>0$$

for all sufficiently large T.

2. For $p>p^0$

$$\frac{1}{T}\sum_{t=1}^{T}\left[v_{t,p}(y_t,\boldsymbol{\phi}^*_{p^0})-v_{t,p}(y_t,\boldsymbol{\phi}^*_p)\right]=O_p(T^{-1})$$

3. For $p>p^0$

$$E\left\{\frac{1}{T}\sum_{t=1}^{T}\left[v_{t,p}(y_t,\boldsymbol{\phi}^*_{p^0})-v_{t,p}(y_t,\boldsymbol{\phi}^*_p)\right]\right\}=0$$

for all sufficiently large T.

Assumption 23 *The information matrix equality holds for* $p\geq p^0$.

Assumptions 10-13 are standard regularity and identifiability conditions. Assumption 15 provides a uniform Lipschitz condition for the gradient and Hessian of $v_{t,p}(\boldsymbol{y}_{t,p},\boldsymbol{\psi}_p)$. Assumptions 16-18 are needed for establishing limit laws for y_t and the gradient and hessian of $v_{t,p}(\boldsymbol{y}_{t,p},\boldsymbol{\psi}_p)$. Finally, Assumptions 19-21 are needed for obtaining uniform laws of large numbers (ULLN) from their pointwise counterparts.

In what follows we take Assumptions 7-9, and therefore the conclusions of Theorem 5, as given. Assumptions 10-12 provide Assumption A of Sin and White (1996). By Theorem 17.12 of Davidson (1994) and given that y_t is a L_2-NED process of arbitrarily large size, Assumption 15 ensures that $v_t(\boldsymbol{y}_{t,p},\boldsymbol{\psi}_p)$ and every element of its hessian are L_2-NED processes of arbitrarily large size as well. Therefore, by Theorem 20.20 of Davidson (1994) and Assumption 16, $v_t(\boldsymbol{y}_{t,p},\boldsymbol{\psi}_p)$ and every element of its hessian obey a pointwise strong law of large numbers (SLLN)[27]. Using Corollary 3 of Andrews

[27]For weak consistency of an information criterion only a weak law of large numbers (WLLN) needs to be obtained. This may obtained under less stringent conditions than a SLLN. However, the difference in the conditions lies mainly in the required sizes for the NED process and the underlying mixing processes. Therefore, given that all processes we consider are of arbitrarily large sizes we will not pursue the distinction further.

Table 1: DGPs for Monte Carlo experiments on lag selection

	SETAR Model			
	DGP 1	DGP 2	DGP 3	DGP 4
d	1	1	1	1
r	0	0	0	0
p^0	2	2	3	3
$\phi_{1,0}$	0.2	0.2	0.2	0.2
$\phi_{1,1}$	0.2	0.2	0.2	0.2
$\phi_{1,2}$	0.1	0.2	-0.1	-0.2
$\phi_{1,3}$			0.1	0.2
$\phi_{2,0}$	0.4	0.4	0.4	0.4
$\phi_{2,1}$	0.3	0.3	0.3	0.3
$\phi_{2,2}$	0.05	-0.3	0.1	0.3
$\phi_{2,3}$			0.05	-0.3
σ_1^2	1.5	1.5	1.5	1.5
σ_2^2	1	1	1	1
$\hat{b}^a$	0.30	0.37	0.27	0.31

	Markov Switching Model			
	DGP 1	DGP 2	DGP 3	DGP 4
p^0	2	2	3	3
q_1	0.5	0.5	0.5	0.5
q_2	0.5	0.5	0.5	0.5
$\phi_{1,0}$	0.5	0.5	0.5	0.5
$\phi_{1,1}$	0.3	0.4	0.2	0.5
$\phi_{1,2}$	0.1	0.4	0.2	0.5
$\phi_{1,3}$			0.1	-0.5
$\phi_{2,0}$	1	1	1	1
$\phi_{2,1}$	0.2	0.4	0.1	0.6
$\phi_{2,2}$	-0.1	-0.5	-0.1	-0.4
$\phi_{2,3}$			0.05	0.6
σ_1^2	1	1	1	1
σ_2^2	1	1	1	1

[a] $\hat{b}$ is a Monte Carlo estimate of the proportion of observations in regime 1 under the given DGP.

(1987) and Assumptions 13, 19, 20 and 21, uniform SLLNs are obtained for these processes. Further, Assumption 17 with Corollary AIII.3 of Sin and White (1992) and 18 with Corollary 24.7 of Davidson (1994) provide a LIL and a CLT, respectively, for each element of $\nabla v_t(\boldsymbol{y}_{t,p}, \boldsymbol{\psi}_p)$.

Combining Assumption A of Sin and White with the uniform LLNs for $v_t(\boldsymbol{y}_{t,p}, \boldsymbol{\psi}_p)$ and its Hessian, the CLT for each element of $\nabla v_t(\boldsymbol{y}_{t,p}, \boldsymbol{\psi}_p)$, Assumptions 14, 22 (i),(ii) and the conditions of Theorem 3 we obtain the conclusions of Proposition 4.2(a),(c) of Sin and White. This proves Theorem 3.

For Theorem 4 we need to obtain Proposition 5.2(a) and Corollary 5.4(b) of Sin and White. These are obtained through Assumption A of Sin and White, the uniform SLLNs for $v_t(\boldsymbol{y}_{t,p}, \boldsymbol{\psi}_p)$ and its Hessian, the LIL for each element of $\nabla v_t(\boldsymbol{y}_{t,p}, \boldsymbol{\psi}_p)$, Assumptions 14, 22 (i),(iii), 23 and the conditions of Theorem 4. This concludes the proof of Theorems 3 and 4.

Table 2: Percentage frequencies of the lag order selected for SETAR models (DGP 1, T=200, N=400) (standard errors of the frequency estimates are given in parentheses)

		Information Criteria				
d[a]	p[b]	AIC	SC	HQ	GIC	ICOMP
1	1	$47.75_{(2.50)}$	$75.00_{(2.17)}$	$69.50_{(2.30)}$	$46.00_{(2.49)}$	$14.50_{(1.76)}$
	2	$12.25_{(1.64)}$	$0.50_{(0.35)}$	$5.50_{(1.14)}$	$14.00_{(1.73)}$	$6.50_{(1.23)}$
	3	$7.25_{(1.30)}$	$0.00_{(N/A)}$	$0.75_{(0.43)}$	$7.50_{(1.32)}$	$4.50_{(1.04)}$
	4	$5.00_{(1.09)}$	$0.00_{(N/A)}$	$1.25_{(0.56)}$	$3.75_{(0.95)}$	$6.00_{(1.19)}$
	5	$2.00_{(0.70)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$2.50_{(0.78)}$	$7.75_{(1.34)}$
	6	$3.00_{(0.85)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$4.00_{(0.98)}$	$13.25_{(1.70)}$
2	1	$10.50_{(1.53)}$	$24.00_{(2.14)}$	$21.00_{(2.04)}$	$9.50_{(1.47)}$	$21.25_{(2.05)}$
	2	$5.25_{(1.12)}$	$0.50_{(0.35)}$	$1.75_{(0.66)}$	$4.75_{(1.06)}$	$2.25_{(0.74)}$
	3	$2.75_{(0.82)}$	$0.00_{(N/A)}$	$0.25_{(0.25)}$	$3.25_{(0.89)}$	$1.50_{(0.61)}$
	4	$0.50_{(0.35)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$0.75_{(0.43)}$	$5.25_{(1.12)}$
	5	$1.50_{(0.61)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$1.50_{(0.61)}$	$7.25_{(1.30)}$
	6	$2.25_{(0.74)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$2.50_{(0.78)}$	$10.00_{(1.50)}$

[a]Delay Parameter
[b]Lag Order

Table 3: Percentage frequencies of the lag order selected for SETAR models (DGP 2, T=200, N=400)

		Information Criteria				
d	p	AIC	SC	HQ	GIC	ICOMP
1	1	$1.50_{(0.61)}$	$9.75_{(1.48)}$	$3.00_{(0.85)}$	$1.50_{(0.61)}$	$0.00_{(N/A)}$
	2	$55.25_{(2.49)}$	$50.75_{(2.50)}$	$65.25_{(2.38)}$	$54.00_{(2.49)}$	$20.50_{(2.02)}$
	3	$12.00_{(1.62)}$	$1.00_{(0.50)}$	$4.75_{(1.06)}$	$12.75_{(1.67)}$	$11.00_{(1.56)}$
	4	$5.50_{(1.14)}$	$0.00_{(N/A)}$	$1.75_{(0.66)}$	$6.25_{(1.21)}$	$8.00_{(1.36)}$
	5	$4.25_{(1.01)}$	$0.00_{(N/A)}$	$0.50_{(0.35)}$	$4.25_{(1.01)}$	$12.75_{(1.67)}$
	6	$5.00_{(1.09)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$5.25_{(1.12)}$	$24.50_{(2.15)}$
2	1	$5.75_{(1.16)}$	$36.00_{(2.40)}$	$16.75_{(1.87)}$	$4.75_{(1.06)}$	$5.25_{(1.12)}$
	2	$5.25_{(1.12)}$	$2.25_{(0.74)}$	$6.75_{(1.25)}$	$5.25_{(1.12)}$	$0.50_{(0.35)}$
	3	$1.50_{(0.61)}$	$0.25_{(0.25)}$	$0.75_{(0.43)}$	$1.75_{(0.66)}$	$1.25_{(0.56)}$
	4	$1.50_{(0.61)}$	$0.00_{(N/A)}$	$0.50_{(0.35)}$	$1.25_{(0.56)}$	$1.50_{(0.61)}$
	5	$1.00_{(0.50)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$1.75_{(0.66)}$	$5.00_{(1.09)}$
	6	$1.50_{(0.61)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$1.25_{(0.56)}$	$9.75_{(1.48)}$

Table 4: Percentage frequencies of the lag order selected for SETAR models (DGP 3, T=200, N=400)

		Information Criteria				
d	p	AIC	SC	HQ	GIC	ICOMP
1	1	$25.75_{(2.19)}$	$55.00_{(2.49)}$	$44.00_{(2.48)}$	$24.25_{(2.14)}$	$5.25_{(1.12)}$
	2	$20.00_{(2.00)}$	$3.75_{(0.95)}$	$14.00_{(1.73)}$	$21.25_{(2.05)}$	$7.00_{(1.28)}$
	3	$11.50_{(1.60)}$	$0.50_{(0.35)}$	$5.00_{(1.09)}$	$12.25_{(1.64)}$	$9.25_{(1.45)}$
	4	$4.00_{(0.98)}$	$0.00_{(N/A)}$	$0.75_{(0.43)}$	$4.75_{(1.06)}$	$9.00_{(1.43)}$
	5	$3.25_{(0.89)}$	$0.00_{(N/A)}$	$0.75_{(0.43)}$	$4.25_{(1.01)}$	$9.00_{(1.43)}$
	6	$3.25_{(0.89)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$3.25_{(0.89)}$	$13.50_{(1.71)}$
2	1	$15.75_{(1.82)}$	$40.50_{(2.45)}$	$29.75_{(2.29)}$	$14.25_{(1.75)}$	$19.00_{(1.96)}$
	2	$5.25_{(1.12)}$	$0.25_{(0.25)}$	$3.00_{(0.85)}$	$4.50_{(1.04)}$	$3.25_{(0.89)}$
	3	$4.00_{(0.98)}$	$0.00_{(N/A)}$	$1.75_{(0.66)}$	$3.50_{(0.92)}$	$2.75_{(0.82)}$
	4	$2.75_{(0.82)}$	$0.00_{(N/A)}$	$0.25_{(0.25)}$	$3.25_{(0.89)}$	$2.75_{(0.82)}$
	5	$2.00_{(0.70)}$	$0.00_{(N/A)}$	$0.50_{(0.35)}$	$2.25_{(0.74)}$	$7.25_{(1.30)}$
	6	$2.50_{(0.78)}$	$0.00_{(N/A)}$	$0.25_{(0.25)}$	$2.25_{(0.74)}$	$12.00_{(1.62)}$

Table 5: Percentage frequencies of the lag order selected for SETAR models (DGP 4, T=200, N=400)

		Information Criteria				
d	p	AIC	SC	HQ	GIC	ICOMP
1	1	$0.75_{(0.43)}$	$13.25_{(1.70)}$	$2.75_{(0.82)}$	$0.75_{(0.43)}$	$0.00_{(N/A)}$
	2	$2.00_{(0.70)}$	$5.75_{(1.16)}$	$4.75_{(1.06)}$	$1.50_{(0.61)}$	$0.00_{(N/A)}$
	3	$61.25_{(2.44)}$	$57.50_{(2.47)}$	$73.75_{(2.20)}$	$60.00_{(2.45)}$	$31.25_{(2.32)}$
	4	$13.50_{(1.71)}$	$1.00_{(0.50)}$	$6.50_{(1.23)}$	$13.25_{(1.70)}$	$12.50_{(1.65)}$
	5	$9.50_{(1.47)}$	$0.00_{(N/A)}$	$0.75_{(0.43)}$	$9.50_{(1.47)}$	$19.75_{(1.99)}$
	6	$7.00_{(1.28)}$	$0.00_{(N/A)}$	$0.75_{(0.43)}$	$8.50_{(1.39)}$	$24.00_{(2.14)}$
2	1	$1.00_{(0.50)}$	$20.25_{(2.01)}$	$6.50_{(1.23)}$	$1.00_{(0.50)}$	$0.75_{(0.43)}$
	2	$0.25_{(0.25)}$	$1.00_{(0.50)}$	$1.25_{(0.56)}$	$0.75_{(0.43)}$	$0.00_{(N/A)}$
	3	$2.75_{(0.82)}$	$1.25_{(0.56)}$	$3.00_{(0.85)}$	$2.75_{(0.82)}$	$1.00_{(0.50)}$
	4	$0.25_{(0.25)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$0.25_{(0.25)}$	$1.00_{(0.50)}$
	5	$1.25_{(0.56)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$1.25_{(0.56)}$	$4.00_{(0.98)}$
	6	$0.50_{(0.35)}$	$0.00_{(N/A)}$	$0.00_{(N/A)}$	$0.50_{(0.35)}$	$5.75_{(1.16)}$

Table 6: Percentage frequencies of the lag order selected for Markov-switching models (T=200, N=400)

		Information Criteria				
DGP	p	AIC	SC	HQ	GIC	ICOMP
1	1	$17.75_{(1.91)}$	$75.75_{(2.14)}$	$45.00_{(2.49)}$	$21.75_{(2.06)}$	$11.00_{(1.56)}$
	2	$14.25_{(1.75)}$	$9.25_{(1.45)}$	$13.25_{(1.70)}$	$16.50_{(1.86)}$	$9.75_{(1.48)}$
	3	$12.00_{(1.62)}$	$5.50_{(1.14)}$	$8.25_{(1.38)}$	$12.00_{(1.62)}$	$8.75_{(1.41)}$
	4	$16.00_{(1.83)}$	$3.50_{(0.92)}$	$9.50_{(1.47)}$	$14.00_{(1.73)}$	$13.75_{(1.72)}$
	5	$17.25_{(1.89)}$	$2.50_{(0.78)}$	$10.00_{(1.50)}$	$15.75_{(1.82)}$	$20.00_{(2.00)}$
	6	$22.75_{(2.10)}$	$3.50_{(0.92)}$	$14.00_{(1.73)}$	$20.00_{(2.00)}$	$36.75_{(2.41)}$
2	1	$5.75_{(1.16)}$	$33.75_{(2.36)}$	$14.75_{(1.77)}$	$8.75_{(1.41)}$	$4.25_{(1.01)}$
	2	$41.50_{(2.46)}$	$57.50_{(2.47)}$	$59.75_{(2.45)}$	$35.00_{(2.38)}$	$12.25_{(1.64)}$
	3	$11.75_{(1.61)}$	$3.00_{(0.85)}$	$8.50_{(1.39)}$	$16.75_{(1.87)}$	$9.25_{(1.45)}$
	4	$11.00_{(1.56)}$	$2.50_{(0.78)}$	$4.50_{(1.04)}$	$12.25_{(1.64)}$	$14.00_{(1.73)}$
	5	$11.00_{(1.56)}$	$1.50_{(0.61)}$	$5.75_{(1.16)}$	$10.75_{(1.55)}$	$22.25_{(2.08)}$
	6	$19.00_{(1.96)}$	$1.75_{(0.66)}$	$6.75_{(1.25)}$	$16.50_{(1.86)}$	$38.00_{(2.43)}$
3	1	$15.50_{(1.81)}$	$73.00_{(2.22)}$	$38.75_{(2.44)}$	$18.50_{(1.94)}$	$8.75_{(1.41)}$
	2	$12.25_{(1.64)}$	$9.25_{(1.45)}$	$14.50_{(1.76)}$	$14.25_{(1.75)}$	$9.00_{(1.43)}$
	3	$13.75_{(1.72)}$	$7.75_{(1.34)}$	$11.25_{(1.58)}$	$15.50_{(1.81)}$	$10.75_{(1.55)}$
	4	$15.25_{(1.80)}$	$3.50_{(0.92)}$	$11.25_{(1.58)}$	$14.50_{(1.76)}$	$13.75_{(1.72)}$
	5	$15.50_{(1.81)}$	$2.75_{(0.82)}$	$8.50_{(1.39)}$	$15.50_{(1.81)}$	$23.75_{(2.13)}$
	6	$27.75_{(2.24)}$	$3.75_{(0.95)}$	$15.75_{(1.82)}$	$21.75_{(2.06)}$	$34.00_{(2.37)}$
4	1	$0.25_{(0.25)}$	$12.25_{(1.64)}$	$2.00_{(0.70)}$	$1.00_{(0.50)}$	$0.25_{(0.25)}$
	2	$0.75_{(0.43)}$	$1.75_{(0.66)}$	$1.50_{(0.61)}$	$1.50_{(0.61)}$	$1.00_{(0.50)}$
	3	$63.50_{(2.41)}$	$83.25_{(1.87)}$	$86.00_{(1.73)}$	$59.50_{(2.45)}$	$21.00_{(2.04)}$
	4	$17.50_{(1.90)}$	$1.75_{(0.66)}$	$6.50_{(1.23)}$	$18.00_{(1.92)}$	$19.75_{(1.99)}$
	5	$8.50_{(1.39)}$	$0.75_{(0.43)}$	$2.50_{(0.78)}$	$10.50_{(1.53)}$	$21.00_{(2.04)}$
	6	$9.50_{(1.47)}$	$0.25_{(0.25)}$	$1.50_{(0.61)}$	$9.50_{(1.47)}$	$37.00_{(2.41)}$

Table 7: Monte Carlo DGPs for model selection between alternative threshold models

	DGP 1	DGP 2
	M-S[a] Model	SETAR Model
m	2	2
r		0
p^0	3	3
q_1	0.5	
q_2	0.5	
$\phi_{1,0}$	0.5	0.5
$\phi_{1,1}$	0.5	0.5
$\phi_{1,2}$	0.5	0.5
$\phi_{1,3}$	-0.5	-0.5
$\phi_{2,0}$	1	1
$\phi_{2,1}$	0.6	0.6
$\phi_{2,2}$	-0.4	-0.4
$\phi_{2,3}$	0.6	0.6
σ_1^2	2.25	2.25
σ_2^2	2.25	2.25

[a]Markov-switching

	DGP 3
	EDTAR Model
p_r	3
p_e	3
p	3
ϕ_0	0.5
ϕ_1	0.4
ϕ_2	0.4
ϕ_3	-0.4
θ_f	0.3
θ_c	-0.3
σ_f^2	1
σ_c^2	1
σ_{cor}^2	1

Table 8: Percentage frequencies of the model selected (T=200, N=400)

		Information Criteria				
DGP[a]	Model Selected	AIC	SC	HQ	GIC	ICOMP
1	1	$94.00_{(1.19)}$	$92.00_{(1.36)}$	$93.25_{(1.25)}$	$87.00_{(1.68)}$	$96.00_{(0.98)}$
	2	$5.00_{(1.09)}$	$4.25_{(1.01)}$	$4.50_{(1.04)}$	$4.75_{(1.06)}$	$0.75_{(0.43)}$
	3	$1.00_{(0.50)}$	$3.75_{(0.95)}$	$2.25_{(0.74)}$	$8.25_{(1.37)}$	$3.25_{(0.89)}$
2	1	$35.50_{(2.39)}$	$33.75_{(2.36)}$	$35.25_{(2.39)}$	$38.75_{(2.43)}$	$47.00_{(2.49)}$
	2	$61.25_{(2.43)}$	$59.25_{(2.46)}$	$60.50_{(2.44)}$	$47.00_{(2.49)}$	$39.75_{(2.45)}$
	3	$3.25_{(0.89)}$	$7.00_{(1.27)}$	$4.25_{(1.01)}$	$14.25_{(1.75)}$	$13.25_{(1.69)}$
3	1	$7.00_{(1.27)}$	$5.25_{(1.11)}$	$6.25_{(1.21)}$	$5.25_{(1.11)}$	$0.25_{(0.25)}$
	2	$1.00_{(0.50)}$	$0.25_{(0.25)}$	$1.00_{(0.50)}$	$0.25_{(0.25)}$	$0.25_{(0.25)}$
	3	$92.00_{(1.36)}$	$94.50_{(1.14)}$	$92.75_{(1.30)}$	$94.50_{(1.14)}$	$99.50_{(0.35)}$

[a]DGP 1: Markov, DGP 2: SETAR, DGP 3: EDTAR

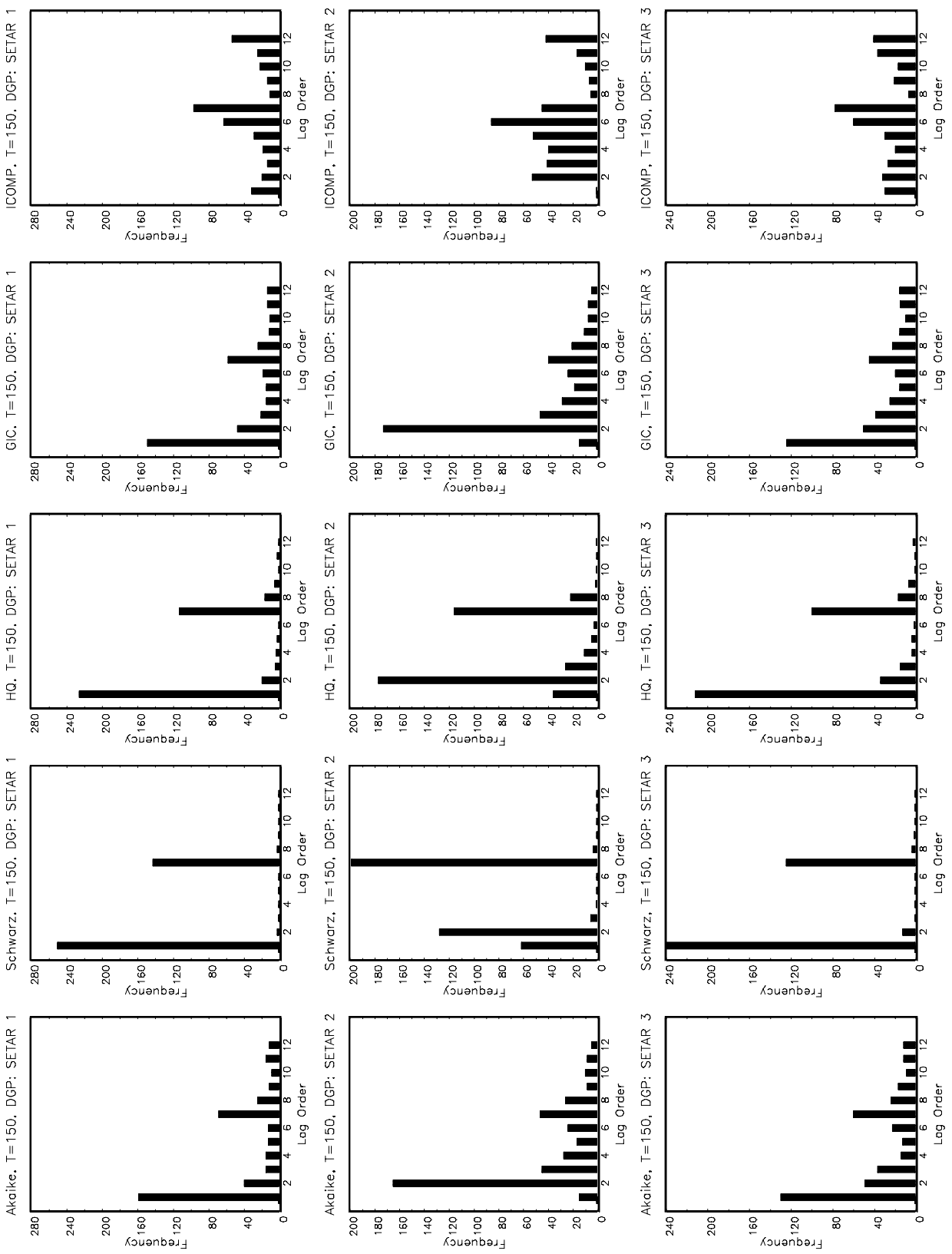

Figure 1: Lag selection in SETAR models. DGP 1,2,3: T=150

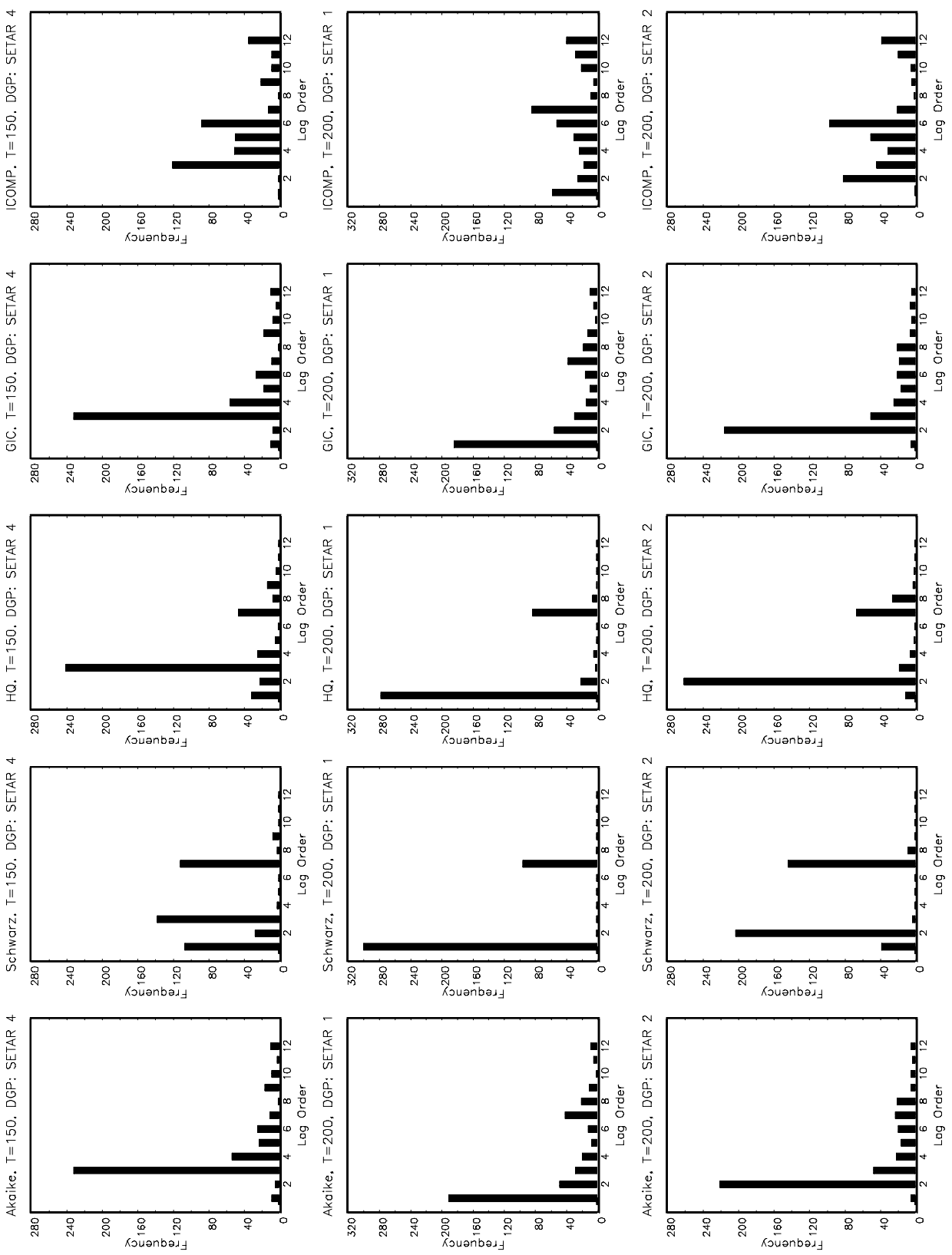

Figure 2: Lag selection in SETAR models. DGP 4: T=150; DGP 1,2: T=200

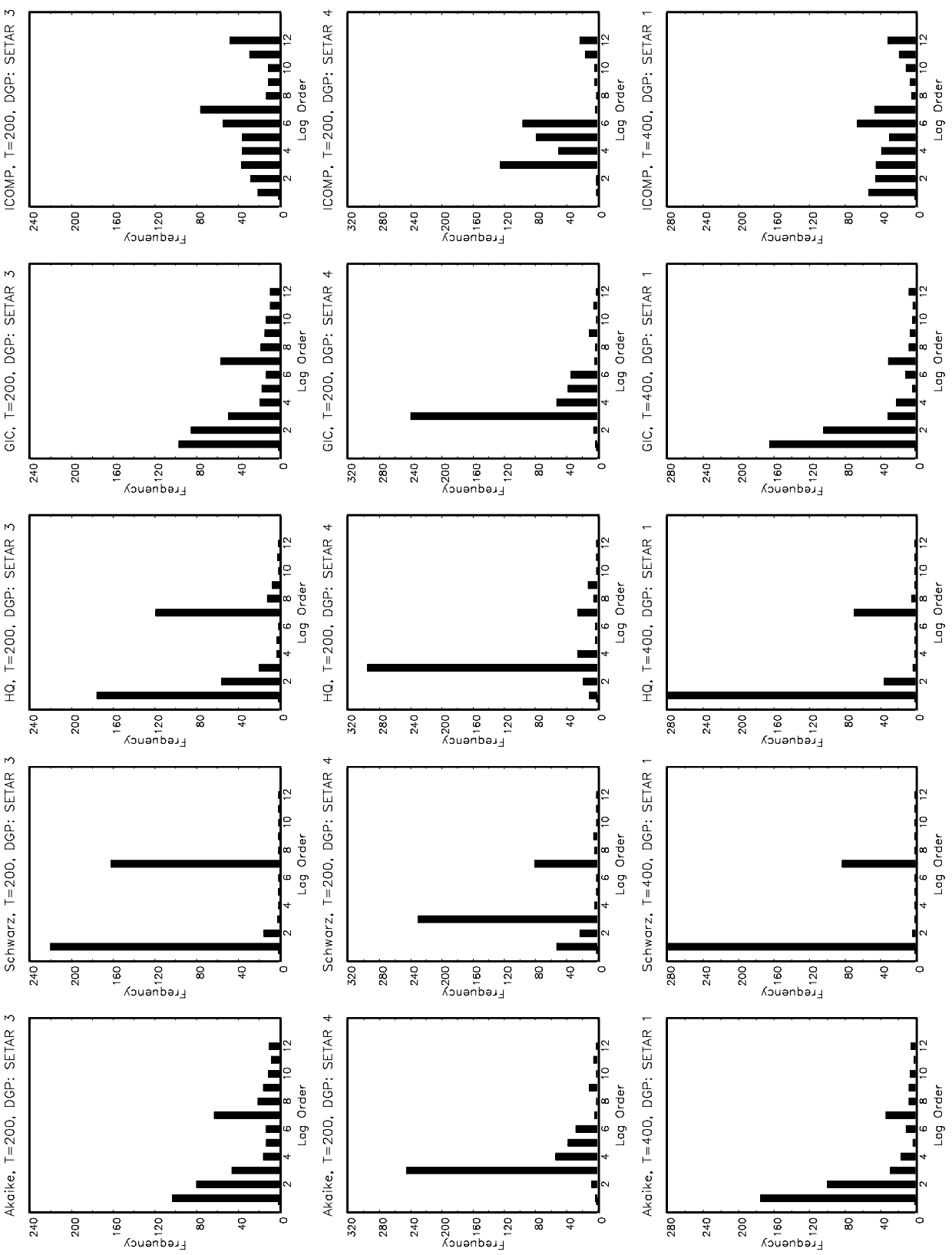

Figure 3: Lag selection in SETAR models. DGP 3,4: T=200; DGP 1: T=400

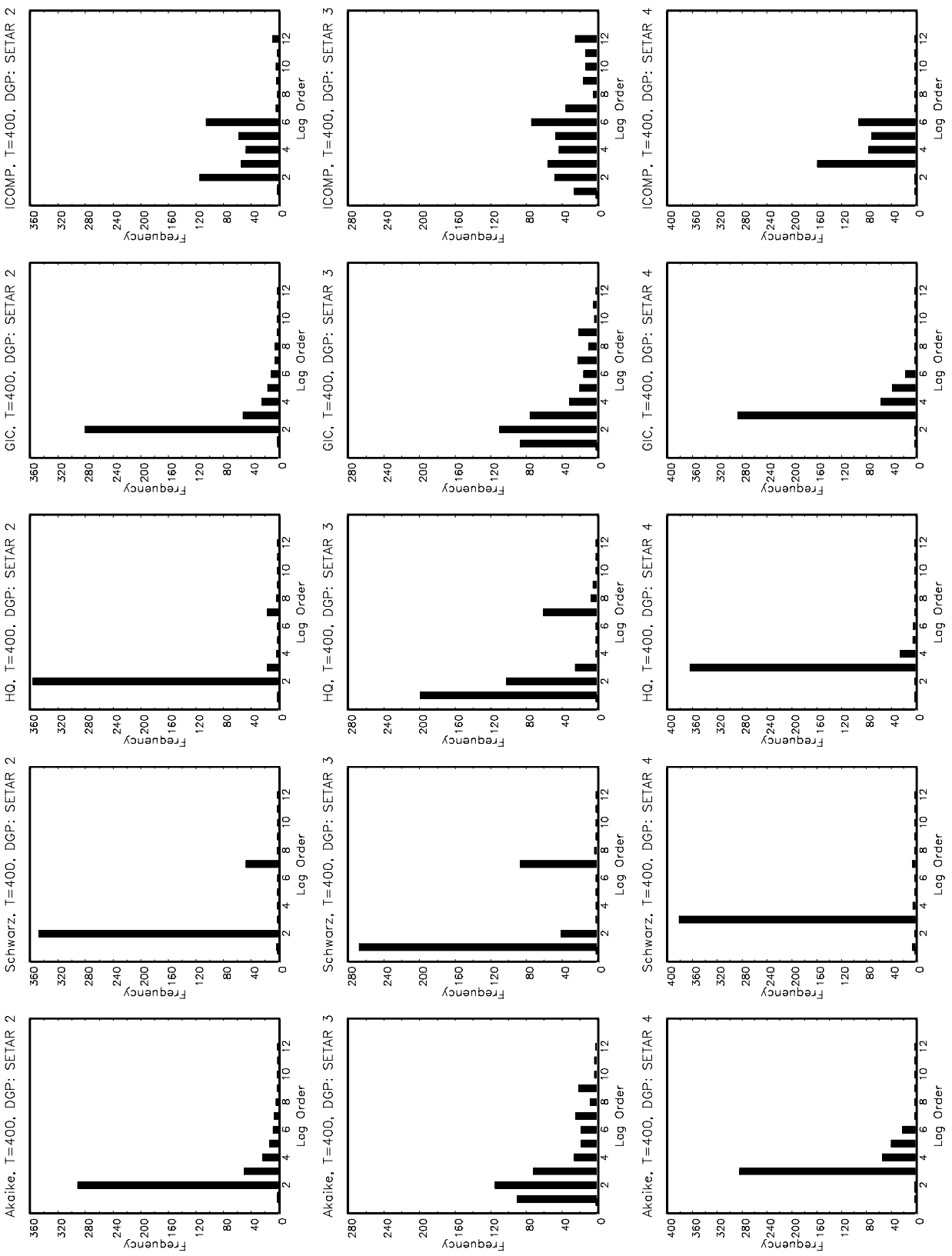

Figure 4: Lag selection in SETAR models. DGP 2,3,4: T=400

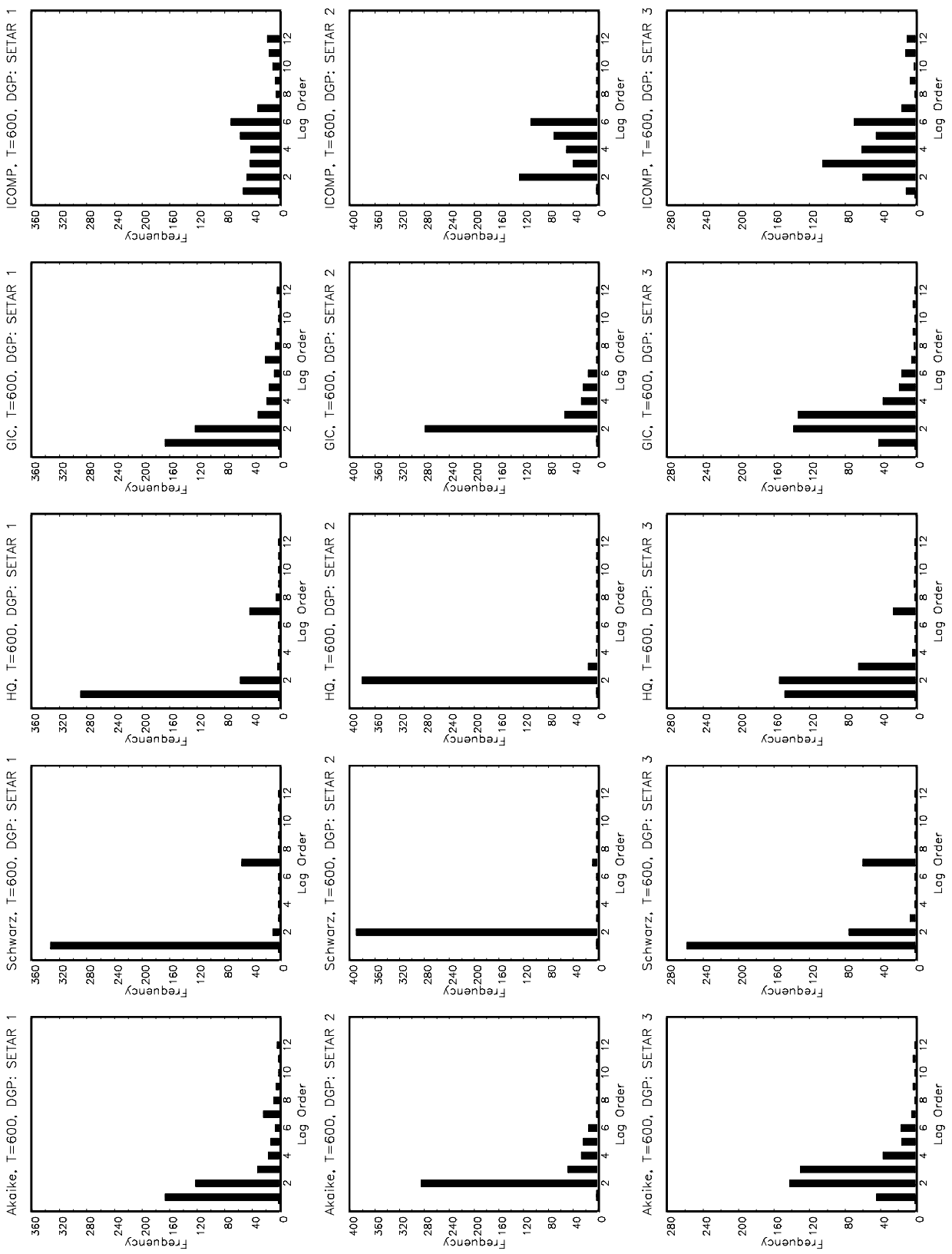

Figure 5: Lag selection in SETAR models. DGP 1,2,3: T=600

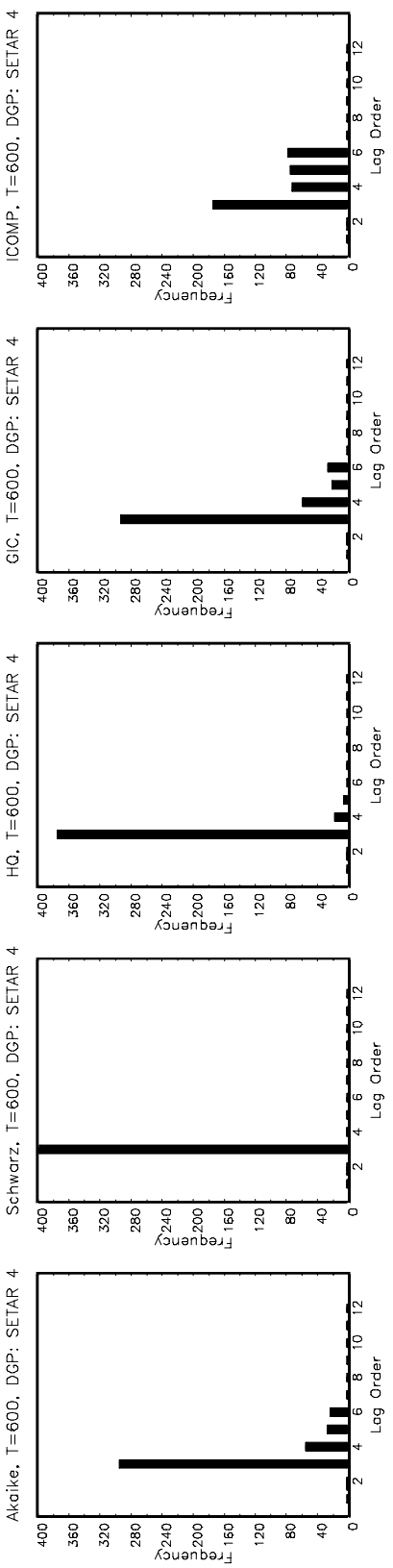

Figure 6: Lag selection in SETAR models. DGP 4: T=600

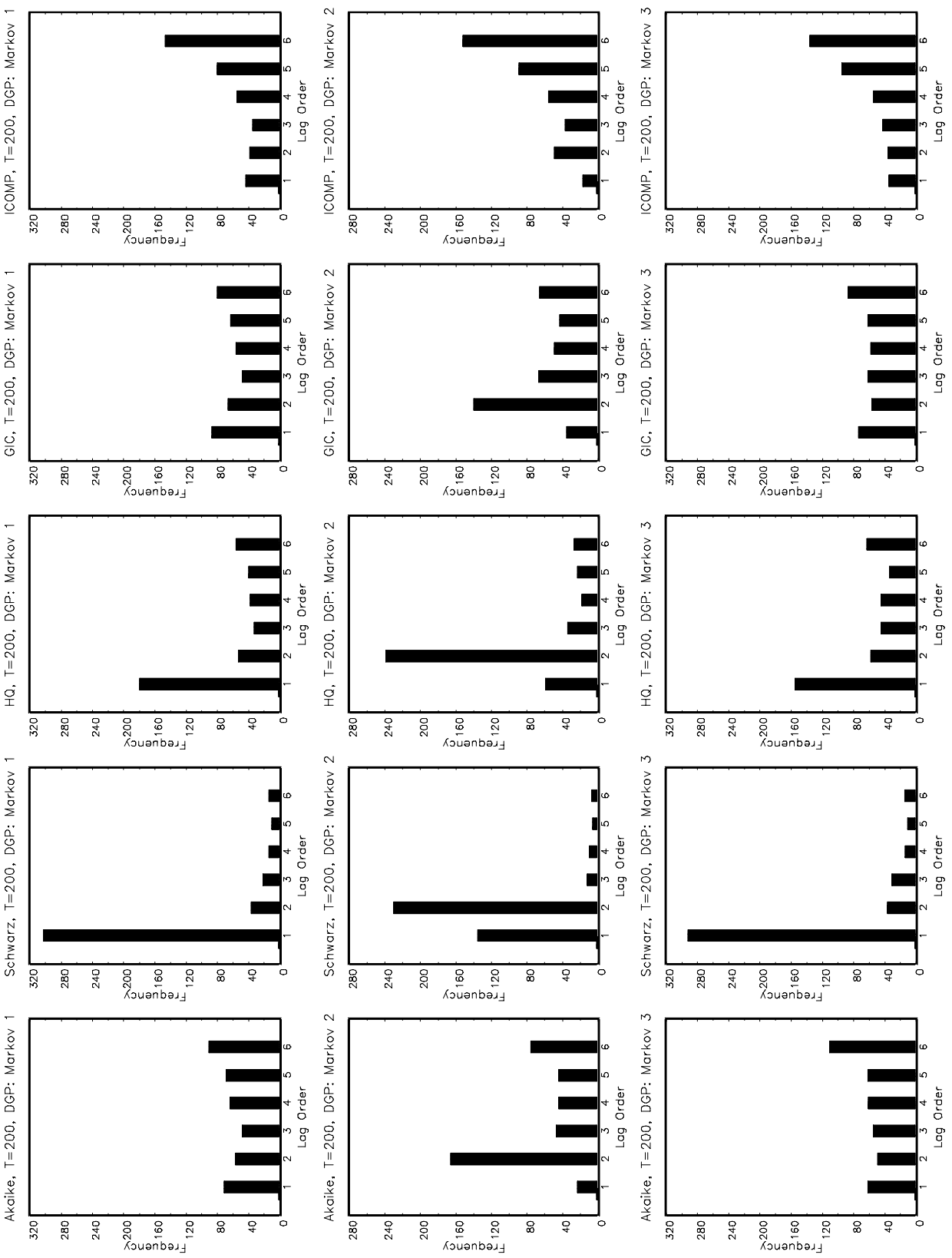

Figure 7: Lag selection in Markov-switching models. DGP 1,2,3: T=200

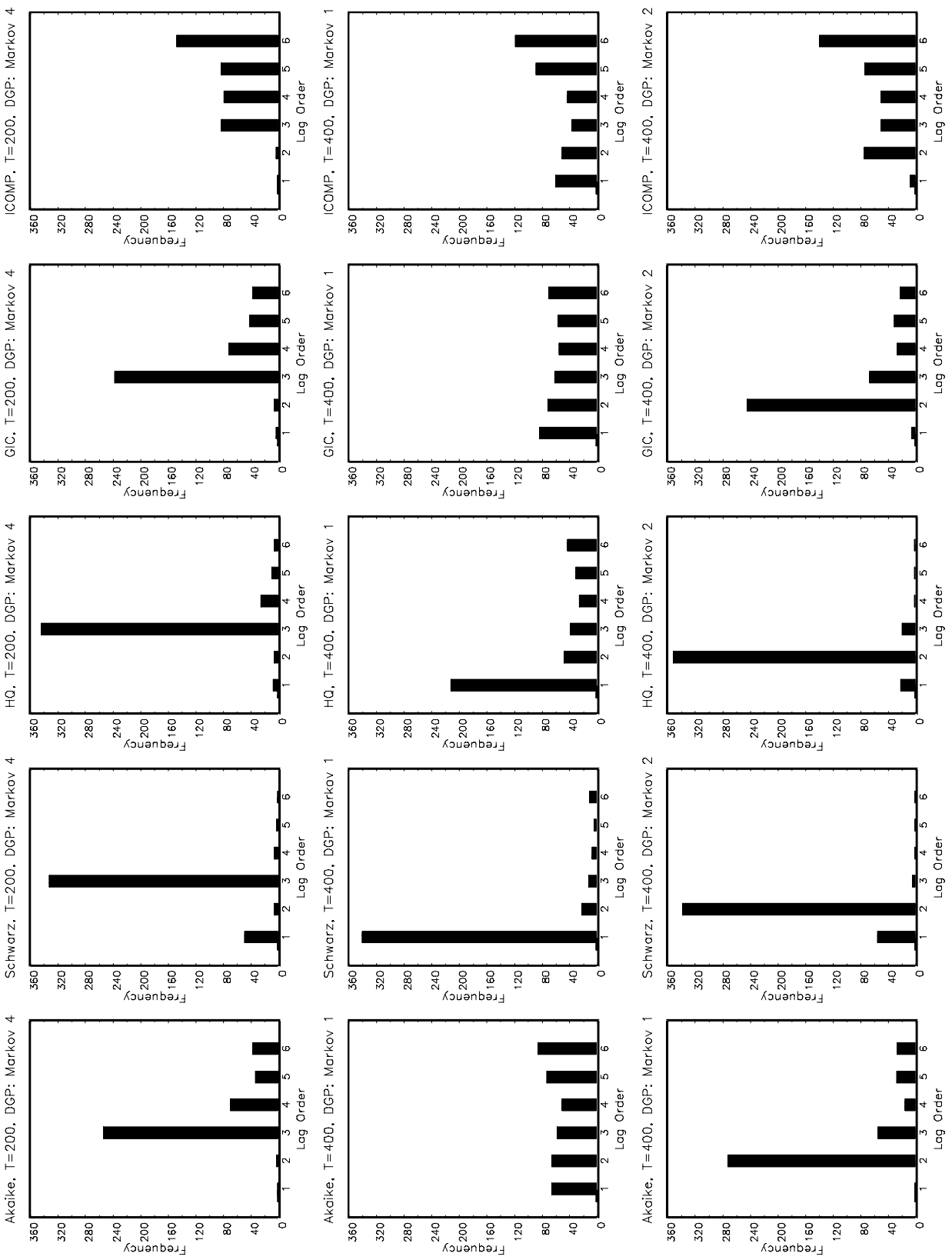

Figure 8: Lag selection in Markov-switching models. DGP 4: T=200; DGP 1,2: T=400

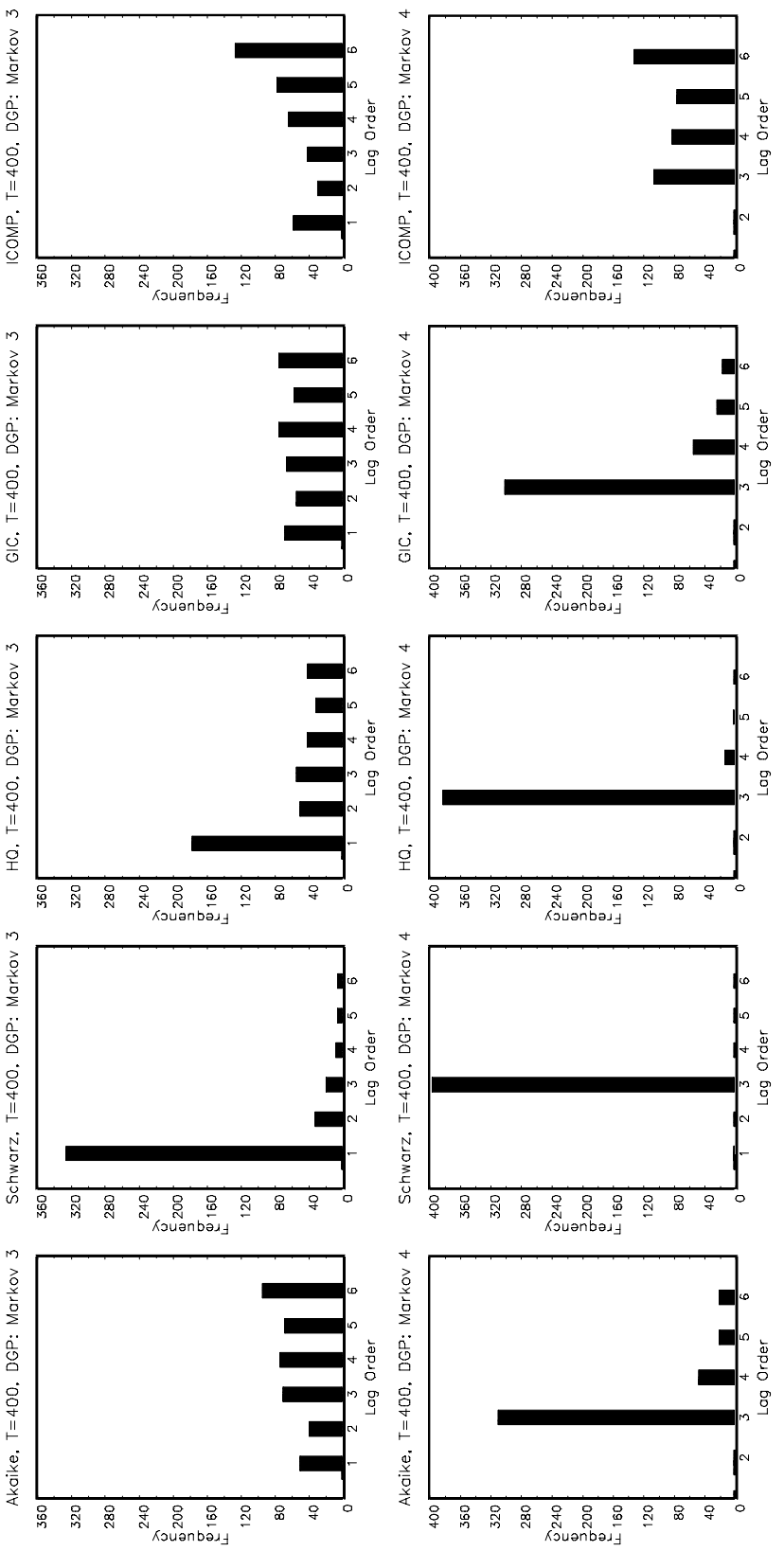

Figure 9: Lag selection in Markov-switching models. DGP 3,4: T=400