Model Selection Uncertainty and Detection of Threshold Effects

Jean-Yves Pitarakis
University of Southampton
J.Pitarakis@soton.ac.uk

July 1, 2004
First Draft

Abstract

Inferences about the presence or absence of threshold type nonlinearities in TAR models are conducted within models whose lag length has been estimated in a preliminary stage. Typically the null hypothesis of linearity is then tested against a threshold alternative on which the same lag length is imposed on each regime. In this paper we evaluate the properties of test statistics for detecting the presence of threshold effects in autoregressive models when this model uncertainty is taken into account. We show that this approach may lead to important distortions when the underlying model has truly threshold effects by establishing the limiting properties of the estimated lag length in the mispecified linear autoregressive fit and assessing the impact of this model uncertainty on the power of the tests. We subsequently propose a full model selection based approach designed to jointly detect the presence of threshold effects and optimally specify its dynamics and compare its performance with the traditional test based approach.

Keywords: Threshold models, SETAR, Model Selection.

JEL: C22, C50.

[1]Address for Correspondence: Jean-Yves Pitarakis, University of Southampton, School of Social Sciences, Division of Economics. Highfield, Southampton, SO17 1BJ, United-Kingdom. Tel: +44-23-80592631, Fax: +44-23-80593858.

1 Introduction

A vast body of the recent theoretical and applied econometrics literature has focused on techniques for modelling economic time series within a nonlinear framework with the aim of explicitly capturing regime specific behaviour and general types of asymmetries for which linear models are inappropriate.

Although economic theory is often silent about the specific type of nonlinearities characterising an economic variable it frequently points to models with switching regimes for capturing changing dynamics across the business cycle for instance (see Potter (1995), Koop and Potter (1999), Altissimo and Violante (2001), Hansen (1997, 1999, 2000), Caner and Hansen (2001) among numerous others). In this context a popular family of models that has attracted considerable recent attention is the class of threshold autoregressive models originally introduced by Tong (1983). Such models aim to model nonlinear dynamics via piecewise linear specifications separated according to the magnitude of a threshold variable. Despite being introduced in the early 70s it is only recently that sufficiently general and formal estimation and inference tools have been proposed and continue to being developed for such models. A sampling theory for testing for the presence of threshold effects within general threshold models has for instance been proposed in Hansen (1996, 1997, 1999) and subsequently extended to the case where the underlying series of interest might be characterised by a unit root in its autoregressive polynomial in Caner and Hansen (2001). The asymptotic properties of estimators obtained from such models have been investigated in Hansen (2000), extending earlier work of Chan (1990, 1993). Additional theoretical results related to testing for the presence of threshold effects and the limiting properties of the resulting estimators have also been introduced in Gonzalo and Pitarakis (2002) for the multiple regime case.

In the context of threshold models where the regimes involve linear autoregressions (SETAR models) the common approach to inference and specification involves first fitting an appropriate linear AR(p) model to the data using some standard model selection criterion such as the AIC, BIC or HQ in order to select an appropriate lag length, say $\hat{p}$. This linear model is subsequently tested against a threshold specification that imposes the lag order $\hat{p}$ in each regime. Although the theoretical properties of tests for detecting the presence of threshold effects are now well understood little is known about their behaviour in finite samples and more importantly about the influence of the preliminary model selection stage on their large and finite sample behaviour. How does the

use of an estimated lag length required in practice prior to implementing the tests of threshold nonlinearity for instance affects the properties of the tests?

Our objectives in this paper are twofold. We will initially investigate the properties of the lag length estimate obtained from a linear fit when the true underlying process is a threshold autoregression. In a related paper, Yang (2002) investigated a similar issue in the context of a stationary VAR model with a structural break in its constant term and established that typically the lag length estimated from a linear VAR will systematically overfit the true lag length. Highlighting the theoretical properties of $\hat{p}$ obtained in this fashion will then allow us to infer the consequences that this preliminary estimation stage will have on the subsequent SupLM type tests for the presence of threshold effects. We are particularly interested in the ability of the tests to detect the presence of threshold effects (i.e. power) when the test statistic is constructed using $\hat{p}$. Our next and key objective is then to evaluate the properties of a full model selection based approach for assessing the presence of SETAR type nonlinearities. This will then allow us to compare the relative merits and shortcomings of both approaches for applied work.

The plan of the paper is as follows. Section 2 introduces the general model and assumptions under which we will operate. Section 3 establishes the limiting behaviour of $\hat{p}$ when the underlying DGP is a SETAR model and subsequently explores the impact of the preliminary lag length estimation stage on the commonly used tests for testing the null hypothesis of linearity againts a threshold alernative. Section 4 introduces our model selection approach and compares its behaviour with the standard test based approach. Section 5 concludes. All proofs are relegated to the appendix.

2 The Model and Assumptions

We consider the following two-regime threshold autoregression also commonly referred to as a $SETAR(2;p,p)$ model

$$y_t = \begin{cases} \phi_{10} + \phi_{11} y_{t-1} + \ldots + \phi_{1p} y_{t-p} + \epsilon_t & \text{if } y_{t-d} \leq \gamma \\ \phi_{20} + \phi_{21} y_{t-1} + \ldots + \phi_{2p} y_{t-p} + \epsilon_t & \text{if } y_{t-d} > \gamma, \end{cases} \tag{1}$$

where $d \in \mathcal{D} = \{1, \ldots, p\}$ denotes the delay parameter, y_{t-d} the threshold variable trigerring the regime switches and γ the threshold parameter. The lag length p is such that $p \leq p_{max}$ for some known upperbound p_{max}.

In what follows we assume that the lag polynomials characterising each regime have their

roots lying strictly outside the unit circle and the threshold parameter is such that $\gamma \in \Gamma$ with $\Gamma = \{\gamma : -\infty < \underline{\gamma} < \gamma < \bar{\gamma} < \infty\}$. The random disturbance term ϵ_t is taken to be a real valued martingale difference sequence with respect to some increasing sequence of sigma fields $\mathcal{F}_t$ generated by $\{(y_{j+1}, \epsilon_{j+1}), j \leq t\}$ with $E|\epsilon|^{4r} < \infty$ for some $r > 1$.

Letting $X = [1\ y_{t-1} \dots y_{t-p}]$ denote the $(T-p) \times (p+1)$ regressor matrix characterising each regime, $y = [y_{p+1}, \dots, y_T]$ the $(T-p) \times 1$ vector of observations on the dependent variable and defining $X_1(\gamma, d) = X * I(y_{-d} \leq \gamma)$ and $X_2(\gamma, d) = X * I(y_{-d} > \gamma)$ with $I(y_{-d} \leq \gamma)$ and $I(y_{-d} > \gamma)$ denoting the stacked vectors of indicator functions and $*$ the Hadamard product, we can reformulate the model in (1) in matrix form as

$$y = X_1(\gamma, d)\phi_1 + X_2(\gamma, d)\phi_2 + \epsilon \tag{2}$$

where $\phi_1 = (\phi_{10}, \phi_{11}, \dots, \phi_{1p})'$, $\phi_2 = (\phi_{20}, \phi_{21}, \dots, \phi_{2p})'$ are $(p \times 1)$ parameter vectors. Noting that given γ and d the model is linear in $\phi = (\phi_1', \phi_2')'$ the concentrated sum of squared errors function can be written as

$$S_T(\gamma, d) = y'y - \sum_{j=1}^{2} y'X_j(\gamma, d)(X_j(\gamma, d)'X_j(\gamma, d))^{-1}X_j(\gamma, d)'y \tag{3}$$

from which the least squares estimators of γ and d can be obtained as $(\hat{\gamma}, \hat{d}) = \arg\min_{\gamma,d} S_T(\gamma, d)$ and the slope parameter estimates are then obtained as $\hat{\phi} = \hat{\phi}(\hat{\gamma}, \hat{d})$. For later use we let $\hat{\sigma}^2(p)$ denote the residual variance from the least squares estimation of the linear model $y = X\phi_1 + u$ (an AR(p) here) fitted to SETAR data. Similarly we let $\hat{\sigma}^2(\gamma, d|p) = S_T(\gamma, d|p)/T$ denote the residual variance obtained from fitting the $SETAR(2; p, p)$ model. Throughout the rest of the paper we will be operating under the following set of assumptions.

Assumptions *As $T \to \infty$, uniformly over $\gamma \in \Re$*

(i) $\dfrac{X_1(\gamma)'X_1(\gamma)}{T} \xrightarrow{p} G(\gamma)$ and $\dfrac{X'X}{T} \xrightarrow{p} G$,

(ii) $\dfrac{X'\epsilon}{T} \xrightarrow{p} 0$,

(iii) $\dfrac{X'\epsilon}{\sqrt{T}} = O_p(1)$,

where G and $G(\gamma)$ are finite symmetric positive definite matrices. $G(\gamma, d)$ is an absolutely continuous and strictly increasing function of γ.

Note that for notational parsimony we have omitted the dependence of the above matrices on $d \in \mathcal{D}$. Since $\mathcal{D}$ is finite convergence over $d \in \mathcal{D}$ is uniform. For later use, we also introduce the following partitioned versions of X together with the limiting counterparts of the corresponding sample moments. Letting p_0 denote the true lag length of the SETAR model in (1), for $p < p_0$ we let $X = [1\ y_{t-1}, \ldots, y_{t-p}, y_{t-(p+1)}, \ldots, y_{t-p_0}]$ and the corresponding partitions of the limiting matrices defined in (i) are written as $G(\gamma) = [G_1(\gamma)\ G_2(\gamma)]$ and $G = [G_1\ G_2]$. The dimensions of G_1, G_2, $G_1(\gamma)$ and $G_2(\gamma)$ are $(p_0+1) \times (p+1)$, $(p_0+1) \times (p_0-p)$, $(p_0+1) \times (p+1)$ and $(p_0+1) \times (p_0-p)$ respectively. We also write $G_1 = (G_{11}\ G_{21})'$ with G_{11} and G_{21} denoting $(p+1) \times (p+1)$ and $(p_0-p) \times (p+1)$ dimensional matrices. For $p > p_0$ we maintain $X = [1, y_{t-1}, \ldots, y_{t-p_0}]$ and define $Z = [y_{t-(p_0+1)}, \ldots, y_{t-p}]$. Within this senario we formulate our assumptions as $Z'Z/T \xrightarrow{p} Q$, $X'Z/T \xrightarrow{p} L$. Also, uniformly over $\gamma \in \Re$, $X_1(\gamma)'Z/T \xrightarrow{p} L(\gamma)$ also implying that $X_2(\gamma)'Z/T \xrightarrow{p} L - L(\gamma)$. Here Q and L are finite symmetric positive definite matrices. Matrix $L(\gamma)$ is an abolutely continuous and strictly increasing function of γ. Similarly, assumptions (ii)-(iii) specialises into $Z'\epsilon/T \xrightarrow{p} 0$ and $Z'\epsilon/\sqrt{T} = O_p(1)$.

Assumptions $(i)-(ii)$ above are law of large number type of conditions. They exclude integrated processes and hold for instance if y_t is strictly stationary and ergodic (see Hansen (1996, Lemma 1)). In the context of the SETAR specification in (1) they will hold provided that the lag polynomials characterising each regime have their roots outside the complex unit circle and the random error term ϵ_t has a bounded and continuous density (see Hansen (1996, Lemma 1)). Assumption (iii) is a central limit theorem type of result. It holds for instance under strict stationarity and ergodicity of the sequence $\{y_t, \epsilon_t\}$ combined with the requirement that ϵ_t is a martingale difference sequence and finite fourth order moment conditions $E|\epsilon_t|^4 < \infty$ and $E|y_t\epsilon_t|^4 < \infty$. In the context of model (1) the stochastic boundedness requirement in (iii) holds provided that the two lag polynomials have all their roots outside the complex unit circle and an m.d.s error sequence with a continuous and bounded pdf.

3 Detecting Threshold Effects: Model Selection Followed by Testing

The practical implementation of a test for the presence of threshold effects as in the specification presented in (1) first involves selecting an appropriate linear autoregression, say $AR(\hat{p})$. The latter

is then tested against the $SETAR(2;\hat{p},\hat{p})$ alternative via the null hypothesis $H_0 : \phi_1 = \phi_2$. Since the parameters γ and d are unidentified under this null hypothesis the test is conducted using a functional such as $\max_{\gamma,d} J_T(\gamma, d)$ where $J_T(\gamma, d) = T(\hat{\sigma}^2(\hat{p}) - \hat{\sigma}^2(\gamma, d|\hat{p}))/\hat{\sigma}^2(\gamma, d|\hat{p})$. Hansen (1996, 1999) obtained the limiting distribution of this test statistic assuming correct specification (i.e. $\hat{p} = p_0$) and showed that the limiting behaviour of $\max_{\gamma,d} J_T(\gamma, d)$ depends on the population moments of the regressors and threshold variable and thus cannot be universally tabulated. Instead a bootstrap model based approach has been proposed. In Hansen (1996) the author also provided a limited Monte-Carlo study evaluating the finite sample behaviour of the above tests. From our reading of the literature however it appears that little is known about the behaviour of the tests for detecting threshold nonlinearity when model selection uncertainty is taken into account.

3.1 Large Sample Behaviour of $\hat{p}$ under a SETAR DGP

If the true model is a linear autoregression, say $AR(p_0)$ and $\hat{p}$ is a consistent estimator of p_0 then large sample inferences about the null hypothesis of linearity based on $J_T(\gamma, d)$ can naturally be used by proceeding as if we knew the true lag length. This obviously does not preclude the possibility of serious finite sample distortions due to the use of a contaminated $\hat{p}$ in the computation of the test statistic. The picture could be very different however if the true model has threshold effects and we test the null hypothesis using $\hat{p}$ obtained from a linear AR fit. Indeed if the true model is a $SETAR(2; p_0, p_0)$ for instance then estimating an optimal lag length within a linear AR(p) specification may lead to estimated lag lengths that are far off the true p_0 characterising each regime of the underlying SETAR even asymptotically. If $\hat{p}$ turns out to be substantially higher than p_0 for instance then the null hypothesis of linearity will be tested within an overfitted model allowing more parameters than necessary to shift under the alternative, with potentially serious consequences for the power properties of the tests. If $\hat{p}$ undershoots the true lag length p_0 on the other hand then the null of linearity will be tested within a model with residual serial correlation using inappropriate distributional results.

Here, our initial aim is to establish the large sample behaviour of $\hat{p}$ estimated using a model selection based approach within a linear autoregression when the true underlying model is in fact a $SETAR(2; p_0, p_0)$. Specifically, we assume that the lag length is estimated from a linear autoregression, say $y_t = \phi_0 + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + u_t$ with $p \in [1, p_{max}]$ and $p_0 \leq p_{max}$. The model

selection criteria used for the estimation of p in the linear autoregression take the general form $IC(p) = \log\hat{\sigma}^2(p) + \frac{c_T}{T}(p+1)$ where c_T is a deterministic penalty term and $\hat{\sigma}^2(p) = \sum_{t=1}^{T}\hat{u}_t^2/T$ denotes the residual variance of the estimated $AR(p)$ model. The lag length estimator is then defined as $\hat{p} = \arg\min_{1\leq p\leq p_{max}} IC(p)$.

Before establishing the large sample behaviour of $\hat{p}$ we initially investigate the properties of the residual variance $\hat{\sigma}^2(p)$ across $p \in [1,\ldots,p_0,\ldots,p_{max}]$ when the true DGP is a $SETAR(2;p_0,p_0)$. The result is summarised in the following lemma.

Lemma 1: *Under assumptions (i)-(ii) and letting $\hat{\sigma}^2(p)$ denote the residual variance from fitting a linear $AR(p)$ to a $SETAR(2;p_0,p_0)$ DGP, we have as $T\to\infty$*

$$\hat{\sigma}^2(p=p_0) \xrightarrow{p} \sigma_\epsilon^2 + (\phi_2-\phi_1)'(G-G(\gamma))G^{-1}G(\gamma)(\phi_2-\phi_1), \tag{4}$$

$$\hat{\sigma}^2(p<p_0) - \hat{\sigma}^2(p_0) \xrightarrow{p} f(\gamma)'G_{22.1}^{-1}f(\gamma), \tag{5}$$

and

$$\hat{\sigma}^2(p>p_0) - \hat{\sigma}^2(p_0) \xrightarrow{p} -(\phi_2-\phi_1)'H(\gamma)'(Q-L'G^{-1}L)^{-1}H(\gamma)(\phi_2-\phi_1). \tag{6}$$

where $f(\gamma) = \phi_1'[G_2(\gamma) - G_1(\gamma)G_{11}^{-1}G_{12}] + \phi_2'[(G_2-G_2(\gamma)) - (G_1-G_1(\gamma))G_{11}^{-1}G_{12}]$, $H(\gamma) = (L - L(\gamma)) - (G-G(\gamma))G^{-1}L$ and $G_{22.1} = G_{22} - G_{21}G_{11}^{-1}G_{12}$.

From the above lemma we note that the large sample behaviour of $\hat{\sigma}^2(p)$ presented in (5) is conventional in the sense that it is qualitatively similar to the behaviour one would observe even within a purely linear framework in which an underparameterised AR is fitted to the data (e.g. fitting an AR(1) to AR(2) data). The result in (6) on the other hand indicates that increasing the linear AR lag order beyond p_0 may lead to a reduction in residual variance asymptotically. This would clearly not have been the case within a purely linear framework in which we would have $\hat{\sigma}^2(p>p_0) - \hat{\sigma}^2(p_0) = o_p(1)$. The behaviour of $\hat{p} = \arg\min_{1\leq p\leq p_{max}} IC(p)$ in this framework is now summarised in the following proposition.

Proposition 1 *Under assumptions (i)-(iii) and the DGP in (1) we have as $T\to\infty$, (a) $P[\hat{p} < p_0]\to 0$ if $c_T/T\to 0$, (b) $P[\hat{p}>p_0]\to 1$ if $c_T = constant$ or $c_T\to\infty$.*

From the above proposition it is clear that when the true process is a $SETAR(2;p_0,p_0)$ on which we attempt to fit a linear $AR(p)$ model, none of the conventional model selection criteria (i.e. the

AIC under $c_T = 2$, the BIC under $c_T = \ln T$ and the HQ under $c_T = 2 \ln \ln T$) will point to a lag length smaller than p_0 since they all satisfy the requirement that $c_T/T \to 0$ as $T \to \infty$. In the present context of selecting an optimal lag length within a misspecified linear model and analogous to its behaviour documented in the conventional lag length selection literature it is also clear from Proposition 1b that an AIC type criterion with $c_T = 2$ will point to lag lengths greater than p_0 asymptotically. The behaviour of the BIC or HQ type criteria is clearly unusual. Indeed, the result in part (b) of Proposition 1 indicates that both the BIC and HQ criteria will point to lag lengths greater than the true lag length of p_0 asymptotically since their penalty terms is such that $c_T \to \infty$.

At this stage it is important to note that the above results are valid in large samples. In practice, when dealing with finite samples it is natural to expect for instance that the decision frequencies across the different model selection criteria will depend on the magnitudes of the true parameters and in particular on the closeness of the true SETAR to a linear model. To shed further light on this point we also explore the limiting properties of $\hat{p}$ by considering the following local to linear parameterisation of (1)

$$y \quad = \quad X\phi_1 + X_2(\gamma, d)\lambda_T + \epsilon \tag{7}$$

where $\lambda_T = (\phi_2 - \phi_1)/\sqrt{T}$. Proceeding as before we initially establish the limiting behaviour of the residual variance obtained from a linear AR(p) fit to data generated from (7) across the different relevant magnitudes of p.

Lemma 2: *Under assumptions (i)-(ii) and letting $\hat{\sigma}^2(p)$ denote the residual variance obtained from fitting a linear AR(p) to data generated from the $SETAR(2; p_0, p_0)$ in (7) we have as $T \to \infty$*

$$\hat{\sigma}^2(p = p_0) \quad \xrightarrow{p} \quad \sigma_\epsilon^2, \tag{8}$$

$$\hat{\sigma}^2(p < p_0) - \hat{\sigma}^2(p_0) \quad \xrightarrow{p} \quad \phi_1'(G_2 - G_1 G_{11}^{-1} G_{12}) G_{22.1}^{-1} (G_2 - G_1 G_{11}^{-1} G_{12})\phi_1, \tag{9}$$

and

$$\hat{\sigma}^2(p > p_0) - \hat{\sigma}^2(p_0) \quad \xrightarrow{p} \quad 0. \tag{10}$$

Unlike in the fixed parameter case the above lemma suggests that when the SETAR DGP is close to a linear autoregression due to small shifts across the two regimes the residual variance from the

misspecified linear AR fit will behave in a conventional manner, converging to its true counterpart for both $p = p_0$ and $p > p_0$. Our subsequent result about the large sample behaviour of $\hat{p}$ when the DGP is given by (7) is summarised in the following Proposition.

Proposition 2 *Under assumptions (i)-(iii), the* $SETAR(2; p_0, p_0)$ *DGP in (7) and as* $T \to \infty$ *we have* $P[\hat{p} = p_0] \to 1$ *if* $c_T \to \infty$ *and* $c_T/T \to 0$. *Specifically* $P[\hat{p} < p_0] \to 0$ *if* $c_T/T \to 0$ *and* $P[\hat{p} > p_0] \to 0$ *if* $c_T \to \infty$.

Proposition 2 establishes the result that under a local alternative to the linear AR(p) model the lag length estimated from a misspecified linear autoregression using either the BIC or HQ criterion will be consistent for the true lag length characterising each regime of the true $SETAR(2; p_0, p_0)$ model. A direct consequence of the above result is that *asymptotically* the use of $\hat{p}$ instead of p_0 will not affect the local power properties of the test of the null of linearity against a $SETAR(2; p_0, p_0)$.

Having established the large sample properties of $\hat{p}$ when the true DGP is given by a threshold model we next focus on evaluating the properties of $\hat{p}$ presented in Proposition 1 in small to moderately sized samples. This is achieved through a set of Monte-Carlo experiments in which SETAR specifications are used to generate the data. All our experiments are conducted using samples of size $T = 200$, $T = 400$ and $T = 1000$ across N=2000 replications. The random error term is taken as a standard normal random variable throughout.

We initially consider a $SETAR(2; 2, 2)$ DGP taking the maximum allowed lag order as $p_{max} = 6$. Results across the different lag lengths and the three commonly used model selection criteria are presented in Table 1 which displays the empirical frequencies of selecting a specific lag order ranging from 1 to 6. Across all model selection criteria and sample sizes $\hat{p}$ is clearly seen to point to lag orders much greater than the one characterising each regime of the SETAR DGP (here $p_0 = 2$). Although this would have been expected from a criterion such as the AIC it turns out that both the BIC and HQ criteria also display a strong tendency to overfit in this context as suggested by the result in Proposition 1.

Table 1 about here

In fact all three criteria appear to display a behaviour that is quantitatively very similar across the different sample sizes. Under $T = 400$ for instance we note that close to 99% of the AIC, BIC and HQ based decision frequencies are concentrated at orders greater than or equal to 4. It is also

worth noting that across all sample sizes none of the three criteria display any tendency to underfit. Even under $T = 400$ for instance the frequencies of selecting lag lengths smaller than $p_0 = 2$ are virtually zero for the AIC as well as the BIC and HQ.

Although under this DGP the finite sample behaviour of $\hat{p}$ conforms with our large sample analysis it is important to emphasise that the chosen parameterisation is such that both regimes are far apart (if we take the mean of each AR regime as a distance metric for instance) and the AR parameter corresponding to y_{t-2} is sufficiently large in at least one regime for its order to be picked up by a statistical criterion sufficiently often. Our next concern therefore is to evaluate the finite sample behaviour of the alternative criteria when the two regimes of the SETAR are"closer" and/or the parameter configuration is such that the lagged right hand side variables enter the specification with coefficients that are nearer to zero.

Our second set of DGPs is again given by a SETAR(2;2,2) with all its parameters allowed to switch across the two regimes. This experiment is designed to explore the sensitivity of the previously documented features of $\hat{p}$ to alternative parameterisations that allow the parameters of the two regimes to be closer to each other and closer to zero individually. The specific DGPs and the corresponding finite sample behaviour of $\hat{p}$ are presented in Table 2. From the first panel of Table 2 it is again clear that a criterion such as the BIC will continue to overfit provided that the AR parameters are sufficiently far away from zero and the two regimes sufficiently distant. In this case we note that for both the BIC and HQ criteria the bulk of the frequencies are concentrated around $p = p_0 + 1 = 3$ across all sample sizes. All three model selection criteria appear to display remarkably stable decision frequencies across all considered sample sizes. On average across all sample sizes, approximately 62% of the AIC's frequencies are concentrated ar $p = p_0 + 1 = 3$ while the figure is approximately 96% and 83% for the BIC and HQ respectively. For all three criteria the bulk of the remaining frequencies is spread across lag lengths $p > p_0 = 3$.

Looking at the second and third panels of Table 2 it becomes clear that the previous picture changes drastically as the parameters characterising the two regimes are allowed to be closer. Here we note that the BIC might display a significant tendency to underfit, pointing very often to lag lengths that are smaller than $p_0 = 2$.

Table 2 about here

Although this tendency declines as the sample size is allowed to increase (see Proposition 2), impractically large sample sizes might be needed for the BIC to move away from the smallest possible lag length. The most drastic pattern can be seen from the bottom panel of Table 2. In this latter case more than 90% of the BIC's frequencies remain clustered at $p = p_0 - 1 = 1$ for both $T = 200$ and $T = 400$. Under this scenario even the AIC's based decision frequencies are clustered below $p_0 = 2$ close to 55% of the times under both $T = 200$ and $T = 400$. Overall for the AIC criterion we observe a clear decline of the frequency to underfit as $T \to \infty$ across all parameter configurations characterising models B to E. Under Model F for instance the AIC points to $p = 1 < p_0 = 2$ about 50% of the times when $T = 200$ but only 26% of the times under $T = 1000$. The same is not true for either the BIC and HQ which appear to have much greater difficulty moving away from the lowest possible lag length $p = 1$. Within the same model for instance the BIC points to $p = 1$ close to 87% of the times when T=200 and this high frequency of underfitting tends to persist as T increases equalling 76.95% under T=1000.

Based on the finite sample properties of the model selection criteria documented in Tables 1-2 it is difficult to conjecture which model selection criterion might be most appropriate for lag length selection prior to linearity testing. Despite the documented large sample overfitting feature of all criteria our simulation based results indicate that this feature might be materialising across all sample sizes solely under the presence of "strong" threshold effects. When the latter are "weak" and the parameters entering each regime kept small it appears that all three criteria might be pointing to lag lengths smaller than p_0 relatively often with potentially severe consequences for the properties of the subsequent tests about the presence or absence of threshold effects. Overall however if we take the natural view that underfitting will lead to greater distortions in any subsequent analysis the choice of using the AIC criterion is clearly more appropriate than using either the BIC or HQ.

3.2 Impact of $\hat{p}$ on Power

Our next objective is to evaluate how the contamination of $\hat{p}$ documented above affects the behaviour of the commonly used test statistics for testing the null of AR type linearity against the SETAR alternative. Based on our results in Proposition 2 we can infer the fact that the use of the pre-estimation stage for selecting the optimal linear AR fit before implementing the test for threshold type nonlinearity will have asymptotically no influence on the local power properties of

the tests. At the same time however Proposition 1 and our empirical results presented in Tables 1-2 point to the fact that the finite sample power properties of the tests could be substantially different relative to a scenario under which the tests are implemented on correctly specificed models Our results in Table 2 also suggest that regardless of the model selection criterion used we might end up with an underfitted specification if the two regimes charactering the SETAR model are close. As a result inferences based on the limiting distribution that assumes a serially uncorrelated error process will be misleading.

Here our aim is to understand the impact that the distortions about $\hat{p}$ will have on the subsequent tests of the null hypothesis $H_0 : \phi_1 = \phi_2$ against the SETAR alternative. For this purpose we evaluate the finite sample properties of the SupLM test across the DGPs considered in Tables 1-2. Table 3 below presents the frequencies of rejection of the null hypothesis of linearity against SETAR across the eight parameter configurations of a SETAR(2;2,2) DGP (coded A to E) using a 2.5% nominal significance level. The empirical power has been computed using the true lag length (here $p_0 = 2$) in the implementation of the test as well as the three estimated lag lengths obtained via the AIC, BIC and HQ criteria.

Table 3 about here

Under both $T = 200$ and $T = 400$ we note substantial differences in empirical power between the case where the test is implemented on a correctly specified model (setting $p_0 = 2$) without the use of a pre-estimated lag length and the case where p is estimated with a model selection criterion prior to implementing the test. Across all parameter configurations power declines by as much as 50 to 60% and occasionally by more when the lag length has been preestimated using a model selection criterion. Although less pronounced, these substantial differences remain present even under T=400. The worst power performance is displayed when the lag length is estimated via the BIC. Under Model D and T=400 for instance the BIC based SupLM test leads to an empirical power of only 14% compared with 73% when the true lag has been used and 39% for the AIC based SupLM. The corresponding figures based on the AIC and HQ are 39% and 25% respectively. Looking at the power estimates corresponding to a sample size of T=1000 it is again interesting to note the substantial differences in power between the cases where the test is implemented imposing $p = p_0$ and the cases where p has been estimated using the three criteria. Under model E for instance the estimated power of the test when $p = p_0 = 2$ was used was 67.90%. The corresponding power

when computed using $\hat{p}_{aic}$ and $\hat{p}_{BIC}$ were 34.65% and 10.90% respectively. These figures suggest that a test for threshold effects implemented on a model whose lag length has been estimated via the BIC criterion will have a very strong tendency to fail to reject the null of linearity if false.

Based on the results presented in Table 3 we thus also note substantial differences in test power across the use of the three alternative criteria. Comparing the selection frequencies in Table 2 with the empirical power figures of Table 3 we can see that the substantial reduction in power due to the use of the BIC criterion is mainly due to its pointing to lag lengths smaller than the true order characterising each regime of the SETAR.

In summary our results in this section have highlighted the severe distortions that will arise in practice when the researcher's goal is to specify a SETAR type of model following the traditional approach of first selecting an optimal linear autoregression and subsequently testing the latter against a SETAR with the same dynamics in each of its regimes. If the true model is a SETAR for instance then the first stage involving the estimation of an appropriate lag length via some model selection criteria may severely contaminate the properties of the subsequent test of the null hypothesis of linearity. Overall our results indicate that the AIC criterion and to a lesser extent the HQ are to be favoured in practice since they track the "true" power most closely.

4 A Model Selection Based Approach

As an alternative to the above standard testing procedure we now propose to view the problem of detecting the potential presence of a SETAR type nonlinearity as a model selection problem. The problem involves selecting an optimal model among a portfolio of specifications. The selection is made via the optimisation of a penalised objective function. The objective function is such that one of its components is a monotonic function of the model dimension (e.g. the residual variance) and its other component penalises the increase or decrease of the first component caused by the increase in the model dimension. Unlike the previous two stage based approach, in our model selection based inferences the p_{max} linear autoregressive specifications are included in the portfolio of models to select from so that our purpose is not solely that of detecting the presence of linearity against threshold effects as in Gonzalo and Pitarakis (2002) where the dynamics of the models were assumed to be correctly specified and the goal was to estimate the number of regimes.

More formally, the model selection procedure will be based on the optimisation of the following objective functions

$$IC(p) \quad = \quad \log \hat{\sigma}^2(p) + \frac{c_T}{T}(p+1), \tag{11}$$

$$IC(p,d;\gamma) \quad = \quad \log \hat{\sigma}^2(p,d;\gamma) + \frac{c_T}{T}(2p+2), \tag{12}$$

where $\hat{\sigma}^2(p)$ is the residual variance from an AR(p) model and $\hat{\sigma}^2(p,d;\gamma)$ denotes the residual variance obtained from a $SETAR(2;p,p)$ as in (1). Our objective is to select an optimal model among a portfolio of models via the optimisation of the above penalised objective function. The model selection procedure will lead to the choice of a linear autoregression if

$$\min_{p} IC(p) \quad < \quad \min_{p,d,\gamma} IC(p,d;\gamma)$$

with $1 \leq p \leq p_{max}$, $d \leq p$ and $\gamma \in \Gamma$. If the above inequality is reversed for some configuration $\{p,d,\gamma\}$ it will then follow that the model selection rule points to a SETAR model with $\hat{p}$, $\hat{d}$ and $\hat{\gamma}$ obtained as minimisers of $IC(p,d;\gamma)$. The implementation of the above approach is intuitively simple. We use the objective function in (11) to determine the best linear model that minimises $IC(p)$ and the objective function in (12) to determine the optimal nonlinear specification amongst all possible nonlinear specifications as indexed by the quantities $\{p,d,\gamma\}$. This then allows us to decide between the optimal linear fit and the optimal nonlinear fit.

Before proceeding with the practical implementation of the model selection approach it is important to highlight some of its advantages relative to the previously analysed test based approach. First recall that the limiting distributions of test statistics such as the SupLM depend on a large number of unknown parameters (e.g. moments of the regressors and threshold variable) and can therefore not be tabulated. Inferences are instead conducted using a bootstrap based approach that allows the construction of asymptotically valid p-values for testing the null of linearity against the threshold alternative (see Hansen (1996)). The model selection approach described above on the other hand does not require a simulation based approach in its implementation since the decision rules rely solely on the magnitudes of the penalty term c_T. The merits of this penalty based approach to inference in the context of nonlinear models has been established in Gonzalo and Pitarakis (2002) in the context of determining the number of regimes characterising a multiple threshold model. The use of a model selection approach to inference with criteria analogous to (11)-(12) has also been advocated in numerous other areas of the econometric litearature, including the detection of

the number of breaks in the mean of a stationary series (Yao (1988)), the estimation of the rank of a matrix (Cragg and Donald (1997)), the estimation of the cointegrating rank (Gonzalo and Pitarakis (1998, 1999)) among numerous others. In the context of the model under study it is also important to note that the full model selection procedure naturally accomodates the case where the regimes characterising the SETAR model might have different dynamics.

We implement the model selection approach on the SETAR DGPs of Table 3. In the implementation of the model selection approach we let $p \in [1, 6]$ and $d \leq p$. As in the test based approach we also let the threshold parameter $\gamma \in \Gamma$. The total number of competing models is given by $p_{max}(p_{max}+1)/2$ nonlinear specifications and p_{max} linear ones. Thus under our choice of $p_{max} = 6$ we have a portfolio of $21+6$ models to select from. Note that within our model selection framework we require both regimes of the SETAR specification to be equal to p. Our key concern is that of distinguishing between a linear AR and a nonlinear SETAR specification rather than achieving a detailed specification of a SETAR model in case the latter turns out to be selected by our procedure.

Before proceeding with the interpretation of the empirical correct decision frequencies of the model selection criteria when the DGPs are given by SETAR models it is important to be aware of their behaviour under linear specifications. Indeed a strong ability of a criterion to detect SETAR type nonlinearity could be due to a spurious tendency to systematically point to the nonlinear model even when the DGP is a linear autoregression for instance. In the terminology of the traditional testing approach it is important to evaluate the "size" properties of the model selection approach prior to interpreting their ability to detect SETAR type nonlinearity. For this purpose we focused on the individual regimes of some of our previous models coded A-E as linear models and evaluated the number of times the three model selection criteria pointed to linear as opposed to nonlinear models. Results for this set of experiments are presented in Table 4. Overall it is clear that both the AIC and HQ criteria will be inappropriate for distinguishing between AR and SETAR models since they display a very strong tendency to point to the SETAR model even when the true model is linear. Under all sample sizes for instance the AIC criterion's frequency of selection of a linear AR rarely exceeds 2%. Similarly, that of the HQ criterion is typically in the 55%-65% range. The inappropriateness of the AIC and HQ penalties was also documented in Gonzalo and Pitarakis (2002) in the context of selecting the number of regimes of a multiple threshold model.

Table 4 about here

The BIC on the other hand appears to display good finite sample properties in the sense that even under moderately small sample sizes it is pointing to the linear models most of the time. At the same time it does not appear to be artificially clustering its frequencies at linear models. Throughout all our DGPs it displayed an ability to select the true linear specification about 90% of the times under T=200, and more than 95% of the times under T=400 with the frequency tending to 100% as T increases.

We next focus on the ability of the model selection criteria to detect SETAR nonlinearity and compare their behaviour with the traditional SupLM based testing approach. Table 5 presents the frequencies of selection of SETAR models as opposed to the linear AR specification. Comparing the empirical correct decision frequencies based on the BIC criterion with the empirical power of the SupLM test obtained either using estimated lag lengths or the true one we note substantial gains in power in favour of the BIC based model selection approach.

Table 5 about here

Under T=200 for instance the model selection approach based on the BIC criterion led to correct decision frequencies on average 10% higher than the ones obtained with the SupLM implemented on the true model. For Model B for instance the SupLM based power of 45.90% (see Table 3) can be compared with a BIC based correct decision frequency of 58.90%. More importantly when we compare the model selection based decision frequencies with the empirical power of the SupLM statistic implemented using estimated lag length we note gains of 50% or more in favour of the BIC based full model selection based approach. This improvement occurs unanimously across all DGPs. Under T=200 and Model C for instance, the SupLM based statistic implemented on a model whose lag length has been estimated via the AIC and BIC led to an empirical power of 13.60% and 11.15% respectively. These figures can be compared with a correct decision frequency of 31.20% when inferences are conducted with the BIC based full model selection approach. Although these power advantages tend to narrow down as the sample size increases they continue to persist even under T=400 and T=1000.

5 Conclusions

In this paper we highlighted the limitation underlying the practical implementation of the tests of the null hypothesis of linearity against a SETAR alternative. More specifically, we showed that the uncertainty induced by the use of a pre-estimated lag length within a linear autoregression when implementing the SupLM type tests can have drastic negative consequences on the power properties of the test. We then introduced a full model selection procedure designed to jointly detect nonlinearity and at the same type establish the optimal specification in terms of its dynamics. Our simulation experiments strongly confirm the advantages of this approach relative to the traditional test based inferences. Based on our simulation results our analysis also indicates that when specifying a linear autoregression for the purpose of testing the model against a SETAR alternative, the use of the AIC model selection criterion is to be favoured. On the other hand when adopting a full model selection based approach the BIC criterion appears to lead to the most accurate results, offering an excellent trade off between wrongly overfitting and wrongly underfitting.

Table 1. Linear Model Selection Under a Threshold DGP

$$y_t = \begin{cases} -3 + 0.5y_{t-1} - 0.9y_{t-2} + \epsilon_t & y_{t-2} \leq 1.5 \\ 2 + 0.3y_{t-1} + 0.2y_{t-2} + \epsilon_t & y_{t-2} > 1.5 \end{cases}$$

	$T = 200$			$T = 400$			$T = 1000$			$T = 10000$		
	AIC	*BIC*	*HQ*	*AIC*	*BIC*	*HQ*	*AIC*	*BIC*	*HQ*	*AIC*	*BIC*	*HQ*
$p = 1$	0.20	1.30	0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$p = 2$	3.95	6.95	5.20	0.45	1.15	1.30	0.05	0.10	0.10	0.00	0.00	0.00
$p = 3$	0.80	0.55	0.70	0.15	0.30	0.15	0.00	0.00	0.00	0.00	0.00	0.00
$p = 4$	52.65	73.60	64.15	36.60	72.50	54.70	13.80	53.20	29.95	0.00	0.00	0.00
$p = 5$	22.05	11.75	17.55	27.90	16.70	24.15	27.25	26.60	29.85	0.00	0.10	0.00
$p = 6$	20.35	5.85	11.75	34.90	9.35	20.25	58.90	20.10	40.10	100.00	99.00	100.00

Table 2. Linear Model Selection Under a Threshold DGP

$$y_t = \begin{cases} \phi_{01} + \phi_{11} y_{t-1} - \phi_{21} y_{t-2} + \epsilon_t & y_{t-2} \leq 0 \\ \phi_{02} - \phi_{11} y_{t-1} + \phi_{21} y_{t-2} + \epsilon_t & y_{t-2} > 0 \end{cases}$$

	$T = 200$			$T = 400$			$T = 1000$		
	Model A: $\phi_{01} = 0.5, \phi_{02} = 0.1, \phi_{11} = 0.7, \phi_{21} = 0.3$								
	AIC	*BIC*	*HQ*	*AIC*	*BIC*	*HQ*	*AIC*	*BIC*	*HQ*
$p = 1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$p = 2$	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$p = 3$	65.25	94.35	82.75	62.05	95.95	83.30	57.20	96.40	82.90
$p = 4$	13.95	3.75	9.60	15.35	3.50	9.80	18.20	3.25	10.80
$p = 5$	7.55	0.90	3.30	9.20	0.30	3.10	11.55	0.25	3.85
$p = 6$	13.25	0.85	4.35	13.40	0.25	3.80	13.05	0.10	2.45
	Model B: $\phi_{01} = 0.5, \phi_{02} = 0.1, \phi_{11} = 0.2, \phi_{21} = -0.1$								
$p = 1$	6.95	27.70	14.20	0.35	7.00	1.70	0.00	0.05	0.00
$p = 2$	58.10	68.50	70.80	61.50	89.70	83.90	57.40	97.00	84.95
$p = 3$	13.95	2.90	8.80	16.15	2.75	9.30	19.85	2.60	11.05
$p = 4$	8.55	0.80	3.50	8.85	0.40	2.85	9.40	0.35	2.55
$p = 5$	6.35	0.05	1.60	6.60	0.15	1.45	7.30	0.00	1.05
$p = 6$	6.10	0.05	1.10	6.55	0.00	0.80	6.05	0.00	0.40
	Model C: $\phi_{01} = 0.5, \phi_{02} = 0.1, \phi_{11} = 0.1, \phi_{21} = 0.1$								
$p = 1$	17.35	49.80	31.25	4.70	28.85	11.85	0.05	1.80	0.40
$p = 2$	50.00	46.85	56.10	60.10	68.80	75.95	63.15	96.00	88.30
$p = 3$	12.60	2.60	7.30	14.05	1.95	7.45	15.65	2.05	7.60
$p = 4$	7.85	0.60	3.00	8.55	0.20	2.65	8.20	0.15	2.30
$p = 5$	7.00	0.10	1.55	6.50	0.20	1.50	7.20	0.00	0.95
$p = 6$	5.20	0.05	0.80	6.10	0.00	0.60	5.75	0.00	0.45
	Model D: $\phi_{01} = 0.2, \phi_{02} = 0.1, \phi_{11} = 0.2, \phi_{21} = -0.1$								
$p = 1$	49.90	87.10	72.15	42.95	86.45	68.60	26.05	76.95	52.10
$p = 2$	22.30	10.90	19.45	27.70	12.35	23.15	34.85	21.30	35.90
$p = 3$	10.50	1.55	4.75	11.95	1.15	5.15	17.85	1.70	8.60
$p = 4$	6.50	0.35	1.90	6.85	0.00	1.85	8.25	0.05	2.25
$p = 5$	5.45	0.05	0.85	5.10	0.05	0.95	7.25	0.00	0.80
$p = 6$	5.35	0.05	0.90	5.45	0.00	0.30	5.75	0.00	0.35
	Model E: $\phi_{01} = 0.2, \phi_{02} = 0.1, \phi_{11} = 0.1, \phi_{21} = 0.1$								
$p = 1$	55.70	91.15	76.50	52.65	91.85	76.65	42.85	89.15	70.15
$p = 2$	20.20	7.75	15.90	21.55	7.45	17.80	28.90	10.30	24.30
$p = 3$	8.50	0.80	3.80	8.75	0.60	3.25	10.55	0.50	3.40
$p = 4$	5.65	0.20	2.00	6.60	0.05	1.30	6.25	0.00	1.30
$p = 5$	5.70	0.05	1.15	5.60	0.05	0.75	6.50	0.05	0.70
$p = 6$	4.25	0.05	0.65	4.85	0.00	0.25	4.95	0.00	0.15

Table 3. Power Properties of SupLM with True and Estimated Lag Lengths

$$y_t = \begin{cases} \phi_{01} + \phi_{11} y_{t-1} - \phi_{21} y_{t-2} + \epsilon_t & y_{t-2} \leq 0 \\ \phi_{02} - \phi_{11} y_{t-1} + \phi_{21} y_{t-2} + \epsilon_t & y_{t-2} > 0 \end{cases}$$

$T = 200$				$T = 400$				$T = 1000$			
TRUE	AIC	BIC	HQ	TRUE	AIC	BIC	HQ	TRUE	AIC	BIC	HQ
Model A: $\phi_{01} = 0.5, \phi_{02} = 0.1, \phi_{11} = 0.7, \phi_{21} = 0.3$											
100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Model B: $\phi_{01} = 0.5, \phi_{02} = 0.1, \phi_{11} = 0.2, \phi_{21} = -0.1$											
45.90	33.30	32.70	36.30	89.00	82.00	82.00	86.00	99.95	100.00	99.90	100.00
Model C: $\phi_{01} = 0.5, \phi_{02} = 0.1, \phi_{11} = 0.1, \phi_{21} = 0.1$											
21.40	13.60	11.15	13.90	57.00	47.00	41.00	49.00	99.05	98.05	97.15	98.45
Model D: $\phi_{01} = 0.2, \phi_{02} = 0.1, \phi_{11} = 0.2, \phi_{21} = -0.1$											
32.55	12.60	7.45	10.30	73.00	39.00	14.00	25.00	99.85	74.30	26.30	49.30
Model E: $\phi_{01} = 0.2, \phi_{02} = 0.1, \phi_{11} = 0.1, \phi_{21} = 0.1$											
8.30	4.75	3.90	4.15	22.00	10.00	5.00	7.00	67.90	34.65	10.90	22.05

Table 4. Model Selection Based Correct Decision Frequencies under Linear DGPs

$$y_t = \phi_{01} + \phi_{11} y_{t-1} + \phi_{21} y_{t-2} + \epsilon_t$$

$T = 200$			$T = 400$			$T = 1000$		
AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
$\phi_{01} = 0.5, \phi_{11} = 0.7, \phi_{21} = -0.3$								
1.80	90.60	48.60	1.20	97.65	61.75	0.00	0.00	0.00
$\phi_{01} = 0.5, \phi_{11} = 0.2, \phi_{21} = 0.1$								
1.60	90.00	47.00	1.70	93.80	55.80	0.00	0.00	0.00
$\phi_{01} = 0.2, \phi_{11} = 0.1, \phi_{21} = -0.1$								
1.80	89.75	45.30	1.35	94.15	54.80	0.00	0.00	0.00

Table 5. Model Selection Based Correct Decision Frequencies under SETAR DGPs

$$y_t = \begin{cases} \phi_{01} + \phi_{11} y_{t-1} - \phi_{21} y_{t-2} + \epsilon_t & y_{t-2} \leq 0 \\ \phi_{02} - \phi_{11} y_{t-1} + \phi_{21} y_{t-2} + \epsilon_t & y_{t-2} > 0 \end{cases}$$

$T = 200$			$T = 400$			$T = 1000$		
AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
Model A: $\phi_{01} = 0.5, \phi_{02} = 0.1, \phi_{11} = 0.7, \phi_{21} = 0.3$								
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Model B: $\phi_{01} = 0.5, \phi_{02} = 0.1, \phi_{11} = 0.2, \phi_{21} = -0.1$								
99.95	58.90	93.50	100.00	89.95	99.50	0.00	0.00	0.00
Model C: $\phi_{01} = 0.5, \phi_{02} = 0.1, \phi_{11} = 0.1, \phi_{21} = 0.1$								
99.80	31.20	80.20	99.95	57.50	93.90	0.00	0.00	0.00
Model D: $\phi_{01} = 0.2, \phi_{02} = 0.1, \phi_{11} = 0.2, \phi_{21} = -0.1$								
99.80	33.75	84.00	99.90	58.30	95.75	0.00	0.00	0.00
Model E: $\phi_{01} = 0.2, \phi_{02} = 0.1, \phi_{11} = 0.1, \phi_{21} = 0.1$								
99.00	15.80	64.40	99.80	14.15	70.70	0.00	0.00	0.00

APPENDIX

Proof of Lemma 1. We treat the case $p > p_0$. With $X = [1\ y_{t-1}, \ldots, y_{t-p_0}]$ and $Z = [y_{t-(p_0+1)}, \ldots, y_{t-p}]$ we let W denote the $(T-p) \times (p+1)$ regressor matrix $W = [X\ Z]$ and the fitted $AR(p)$ model is $y = W\phi + u$ from which we have $\hat{\sigma}^2(p > p_0) = (y'y - y'W(W'W)^{-1}W'y)/T$. Using standard least squares algebra we next note that we can reformulate $\hat{\sigma}^2(p > p_0)$ as

$$\hat{\sigma}^2(p > p_0) \quad = \quad \frac{1}{T}(y'y - y'X(X'X)^{-1}X'y - y'M(M'M)^{-1}M'y) \tag{13}$$

where $M = Z - X(X'X)^{-1}X'Z$. Next observing that $\hat{\sigma}^2(p_0) = (y'y - y'X(X'X)^{-1}X'y)/T$ and applying appropriate normalisations we can write

$$\hat{\sigma}^2(p > p_0) - \hat{\sigma}^2(p_0) \quad = \quad -\frac{y'M}{T}\left(\frac{M'M}{T}\right)^{-1}\frac{M'y}{T}. \tag{14}$$

Given that $M'M = Z'Z - Z'X(X'X)^{-1}X'Z$, using assumption (i) and the corresponding partitioned versions we have $(M'M/T)^{-1} \xrightarrow{p} (Q - L'G^{-1}L)^{-1}$. Next we write the true model as $y = X\phi_1 + X_2(\gamma)\lambda + \epsilon$ with $\lambda = (\phi_2 - \phi_1)$. We have

$$\frac{M'y}{T} \quad = \quad \left[\frac{Z'X_2(\gamma)}{T} - \frac{Z'X}{T}\left(\frac{X'X}{T}\right)^{-1}\frac{X'X_2(\gamma)}{T}\right]\lambda + \left[\frac{Z'\epsilon}{T} - \frac{Z'X}{T}\left(\frac{X'X}{T}\right)^{-1}\frac{X'\epsilon}{T}\right].$$

From assumption (ii) we have $X'\epsilon/T = o_p(1)$ and $Z'\epsilon/T = o_p(1)$ leading to

$$\frac{M'y}{T} \quad = \quad \left[\frac{Z'X_2(\gamma)}{T} - \frac{Z'X}{T}\left(\frac{X'X}{T}\right)^{-1}\frac{X'X_2(\gamma)}{T}\right]\lambda + o_p(1). \tag{15}$$

Since $X'X_2(\gamma) = X_2(\gamma)'X_2(\gamma)$ assumption (i) and its specialised versions lead to the desired result in (6). The proofs for the cases $p = p_0$ and $p < p_0$ follow identical lines and are omitted.

Proof of Proposition 1 We initially treat the underfitting case by showing that when the penalty term is such that $c_T/T \to 0$ the corresponding model selection criteria used for choosing an optimal p within the linear AR(p) family of models will not point to a lag length below the true p_0 characterising the SETAR model in (2). This is achieved by establishing that for $p < p_0$, $P[IC(p) < IC(p_0)] \to 0$ as $T \to \infty$. We have

$$\begin{aligned} P[IC(p) < IC(p_0)] \quad &= \quad P\left[\log\frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p_0)} < \frac{c_T}{T}(p_0 - p)\right] \\ &= \quad P\left[\frac{\hat{\sigma}^2(p) - \hat{\sigma}^2(p_0)}{\hat{\sigma}^2(p_0)} < e^{\frac{c_T}{T}(p_0-p)} - 1\right]. \end{aligned} \tag{16}$$

Next, from Lemma 1, $\hat{\sigma}^2(p < p_0) - \hat{\sigma}^2(p_0) \xrightarrow{p} \Delta > 0$ with Δ given by the right hand side of (5). We thus have that $(\hat{\sigma}^2(p < p_0) - \hat{\sigma}^2(p_0))/\hat{\sigma}^2(p_0)$ converges to a strictly positive constant and since when $c_T/T \to 0$ we have $\left[e^{\frac{c_T}{T}(p_0-p)} - 1\right] \to 0$ the required result follows.

We next consider the case $p > p_0$. We have

$$\begin{aligned} P[IC(p) < IC(p_0)] &= P\left[\log\frac{\hat{\sigma}^2(p_0)}{\hat{\sigma}^2(p)} > \frac{c_T}{T}(p-p_0)\right] \\ &= P\left[\frac{T(\hat{\sigma}^2(p_0)-\hat{\sigma}^2(p))}{\hat{\sigma}^2(p)} > T\left(e^{\frac{c_T}{T}(p-p_0)}-1\right)\right] \end{aligned} \tag{17}$$

Using (14) above and the fact that $y = X\phi_1 + X_2(\gamma)\lambda + \epsilon$ we can write

$$\begin{aligned} T(\hat{\sigma}^2(p_0)-\hat{\sigma}^2(p)) &= y'M(M'M)^{-1}M'y \\ &= T\ (A_T B_T^{-1} A_T) \end{aligned} \tag{18}$$

with

$$A_T = \left(\lambda'\frac{X_2'Z}{T} - \lambda'\frac{X_2'X}{T}\left(\frac{X'X}{T}\right)^{-1}\frac{X'Z}{T}\right) + \left(\frac{\epsilon'Z}{\sqrt{T}} - \frac{\epsilon'X}{\sqrt{T}}\left(\frac{X'X}{T}\right)^{-1}\frac{X'Z}{T}\right)$$

and

$$B_T = \left(\frac{Z'Z}{T} - \frac{Z'X}{T}\left(\frac{X'X}{T}\right)^{-1}\frac{X'Z}{T}\right)^{-1}.$$

From assumptions (i) we have $B_T \xrightarrow{p} (Q - L'GL)^{-1} > 0$. Also, using assumptions (i)-(iii) we have that $A_T B_T^{-1} A_T = O_p(1)$ and it therefore follows that $T(\hat{\sigma}^2(p_0) - \hat{\sigma}^2(p)) = O_p(T)$. Since the right hand side in (17) is $O(c_T)$ the required result follows.

Proof of Lemma 2. The result in (10) follows by noting that $M'y/T = O_p(T^{-\frac{1}{2}})$ in (14) when λ is replaced by $\lambda_T \equiv (\phi_2 - \phi_1)/\sqrt{T}$.

Proof of Proposition 2. Follows by applying the result in Lemma 2 to the proof of Proposition 1.

REFERENCES

Altissimo, F. and G. L. Violante (2001) The Nonlinear Dynamics of Output and Unemployment in the US. Journal of Applied Econometrics 16, 461-486.

Caner, M. and B. E. Hansen (2001) Threshold Autoregression with a Unit Root. Econometrica 69, 1555-1596.

Chan, K. S. (1990) Testing for Threshold Autoregression. Annals of Statistics 18, 1886-1894.

Chan, K. S. (1993) Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model. Annals of Statistics 21, 520-553.

Cragg, J. G. and S. G. Donald (1997) Inferring the Rank of a Matrix. Journal of Econometrics 76, 223-250.

Gonzalo, J. and J-Y. Pitarakis (2002) Estimation and Model Selection Based Inference in Single and Multiple Threshold Models. Journal of Econometrics 2002, 319-352.

Gonzalo, J. and J-Y. Pitarakis (1998) Specification via Model Selection in Vector Error Correction Models. Economics Letters 60, 321-328.

Hansen, B. E. (1996) Inference when a Nuisance Parameter is not identified under the Null Hypothesis. Econometrica 64, 413-430.

Hansen, B. E. (1997) Inference in TAR Models. Studies in Nonlinear Dynamics and Econometrics 2, 1-14.

Hansen, B. E. (1999) Testing for Linearity. Journal of Economic Surveys 13, 551-576.

Hansen, B. E. (2000) Sample Splitting and Threshold Estimations. Econometrica 68, 575-603.

Koop, G. and S. M. Potter (1999) Dynamic Asymmetries in US Unemployment. Journal of Business and Economic Statistics 17, 298-312.

Potter, S. M. (1995) A Nonlinear Approach to US GNP. Journal of Applied Econometrics 2, 109-125.

Tong, H. (1983) Threshold Models in Nonlinear Time Series. Lecture Notes in Statistics, Vol. 21, Springer, Berlin.

Tong, H. (1990) Nonlinear Time Series: A Dynamical Systems Approach. Oxford University Press, Oxford.

Yang, M. (2002) Lag Length and Mean Break in Stationary VAR Models. Econometrics Journal, Vol. 5, pp. 374-386.

Yao, Y. C. (1988) Estimating the number of Changepoints via Schwarz' criterion. Statistics and Probability Letters 6, 181-189.