Estimation and Model Selection Based Inference in Single and Multiple Threshold Models

Jesus Gonzalo
Universidad Carlos III de Madrid
jgonzalo@elrond.uc3m.es

Jean-Yves Pitarakis
University of Reading
J.Pitarakis@reading.ac.uk

First Version: July 2000
This Version: April 2001

Abstract

This paper evaluates the properties of a joint and sequential estimation procedure for estimating the parameters of single and multiple threshold models. We initially proceed under the assumption that the number of regimes is known à priori but subsequently relax this assumption via the introduction of a model selection based procedure that allows the estimation of both the unknown parameters and their number to be performed jointly. Theoretical properties of the resulting estimators are derived and their finite sample properties investigated.

Keywords: Threshold Models, Non-Linear Models, Information Criteria, Model Selection.

JEL: C22, C50.

[1]We thank two anonymous referees, our discussant Shakeeb Khan at the North American Winter Meetings of the Econometric Society, New-Orleans 2001, and seminar participants at Queen Mary and Westfield College, CORE and the Cardiff Conference on Long Memory and Nonlinear Time Series for useful comments and suggestions. The first author gratefully acknowledges the financial support of the Spanish Ministry of Education (DGYCIT PB98-0026) and the second author thanks the British Academy. Address for Correspondence: Jean-Yves Pitarakis, Department of Economics, University of Reading, PO Box 218 Whiteknights, Reading RG6 6AA, United-Kingdom. Tel:+44-118-9866328, Fax:+44-118-9750236.

1 Introduction

The recent applied and theoretical econometrics literature has witnessed a growing interest in the class of threshold models characterized by piecewise linear processes separated according to the magnitude of a threshold variable. When each linear regime follows an autoregressive process for instance we have the well known threshold autoregressive family of models, the statistical properties of which have been investigated in early work by Tong and Lim (1980), Tong (1983, 1990), and more recently reconsidered and extended in Hansen (1996, 1997, 1999a, 1999b, 2000), Caner and Hansen (2000), Gonzalez and Gonzalo (1997) among others. Given their rich dynamic structure and their ability to capture nonlinearities and asymmetries within an intuitive mathematical framework, this class of nonlinear models has also generated a growing interest among economists interested in capturing economically meaningful nonlinearities. Examples include the analyis of asymmetries in persistence in the US output growth (Beaudry and Koop (1993), Potter (1995)), nonlinearities in unemployment rates (Hansen (1997), Koop and Potter (1999)), threshold effects in cross-country growth regressions (Durlauf and Johnson (1995)) and in international relative prices (Obstfeld and Taylor (1997), O'Connell and Wei (1997)) among numerous others.

Although economic theory is often silent about the specific type of nonlinearities, it frequently suggests models with switching behavior as in the case of the speculative storage model recently analyzed in Michaelides and Ng (2000) or situations where macroeconomic variables such as output or employment present different dynamics according to the stage of the business cycle (see Koop and Potter (1999), Altissimo and Violante (1999)). It is also important to point out that the threshold family of models is only one among a multitude of other possible specifications able to capture nonlinearities in economic variables. The choice is typically dictated by the particular stylized facts the model is designed to capture as well as the availability of statistical tools for conducting inferences. Alternative formulations include Hamilton's regime switching model (Hamilton (1989)), the standard change-point model, bilinear processes, among numerous others (see Carrasco (1999) for an encompassing testing strategy covering a wide range of nonlinear specifications). Although the multitude of potential specifications may suggest that the threshold family of models is only a narrow subset, recently Petruccelli (1992) has shown that the latter may also be viewed as an approximation to a more general class of nonlinear processes.

Despite their ability to capture interesting asymmetric features and jump phenomena observed

in economic and financial time series the use of threshold models in the applied economics literature has been quite limited when compared with specifications such as Hamilton's regime switching model. Among the significant problems encountered when modelling data with threshold type of models are the prohibitive computational costs when estimating specifications with more than two regimes and on the theoretical side the difficulties in tabulating the limiting distributions of LR type statistics for detecting single or multiple threshold effects. For the latter case for instance, inferences are nonstandard due to the well known unidentified nuisance parameters problem together with the fact that the relevant limiting distributions tend to depend on model specific moments, thus ruling out any general tabulation. Tsay (1989) proposed a very interesting graphical approach for detecting the number and location of the thresholds and more recently, Hansen (1996) has developed a general methodology for the treatment of the at most two regime case which to our knowledge is the only technique that can handle very general threshold models including SETAR's of any order, but its applicability to models with possibly more than two regimes is unclear.

In this paper our aim is to focus on some of the above mentioned computational and theoretical difficulties by first formally establishing the large sample properties of a sequential estimation approach that makes the estimation of multiple-threshold models computationally feasible. We subsequently concentrate on the possibility of using an alternative approach to testing for a data based determination of the unknown number of regimes. The plan of the paper is as follows. Section II focuses on the sequential estimation of the parameters of a multiple threshold model under the assumption that the number of regimes is fixed and known. Section III extends the results to the case of an unknown number of regimes by investigating the properties of a model selection based approach for the joint determination of the threshold parameters and their number. Section III concludes. All proofs are relegated to the appendix.

2 Joint and Sequential Estimation under a known number of thresholds

We consider the following multiple threshold model with $m+1$ regimes

$$y_t = \sum_{j=1}^{m+1} \boldsymbol{\beta}_j' \mathbf{x}_t I(\gamma_{j-1} < z_t \leq \gamma_j) + \epsilon_t \tag{1}$$

where y_t is the dependent variable, $x_t I(\gamma_{j-1} < z_t \leq \gamma_j)$ is a $K \times 1$ vector of regressors with $I(.)$ denoting the indicator function, $\boldsymbol{\beta}_j$ the corresponding $K \times 1$ vector of coefficients and z_t the

threshold variable that triggers the regime switches. The random error term ϵ_t is a real valued martingale difference sequence with respect to some increasing sequence of sigma-fields $\mathcal{F}_t$ generated by $\{(x_{j+1}, z_{j+1}, \epsilon_j), j \leq t\}$ with $E|\epsilon_t|^{4r} < \infty$ for some $r > 1$. The threshold parameters denoted $(\gamma_1, \ldots, \gamma_m)$ with $\gamma_0 = -\infty$, $\gamma_{m+1} = \infty$ are such that $\gamma_i \in \Gamma_m \ \forall i = 1, \ldots, m$ with $\Gamma_m = \{(\gamma_1, \ldots, \gamma_m) : -\infty < \underline{\gamma} < \gamma_1 < \ldots < \gamma_m < \bar{\gamma} < \infty\}$. Thus we require all threshold parameters to lie in the bounded subset $[\underline{\gamma}, \bar{\gamma}]$ of the threshold variable sample space.

The multiple threshold model (1) can also be expressed in matrix form as

$$\mathbf{y} = \sum_{j=1}^{m+1} \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon} \tag{2}$$

where $\mathbf{y}$ and $\boldsymbol{\epsilon}$ are $T \times 1$ vectors obtained by stacking y_t and ϵ_t, $\mathbf{X}_j \equiv \mathbf{X} * \mathbf{I}(\gamma_{j-1} < z \leq \gamma_j)$ is the $T \times K$ matrix obtained by stacking the regressor vectors. The dependence of the $\mathbf{X}_j's$ on the threshold parameters is omitted for notational parsimony. Here the symbol $*$ denotes the Hadamard product operator that multiplies on an element by element basis, $\mathbf{I}(\gamma_{j-1} < z \leq \gamma_j)$ is the stacked $T \times 1$ vector of indicator variables and throughout this paper we require $\text{rank}(\mathbf{X}_j) = K$ for all $T_j \geq K$ with T_j denoting the number of observations present in regime j. Note also that the threshold variable z_t could be a component of the regressor matrix which may contain lagged values of y_t or a variable that is external to the system. Given data collected in $\mathbf{y}$, $\mathbf{X}$ and z, and assuming that the number of regimes is known, our objective is to estimate the regression coefficients together with the threshold parameters. Specifically the unknown $(m+1)K + m$ dimensional parameter vector is given by $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{m+1}, \gamma_1, \ldots, \gamma_m)$. It is also worth noting that within the specification in (2) we have $\mathbf{X} = \sum_{j=1}^{m+1} \mathbf{X}_j$ and the regressors are such that $\mathbf{X}_i' \mathbf{X}_j = \mathbf{0} \ \forall i \neq j$. Before proceeding with the estimation of $\boldsymbol{\theta}$ we introduce a set of preliminary assumptions ensuring the identification of the unknown parameter vector. We define $\mathbf{X}_\zeta = \mathbf{X} * \mathbf{I}(\gamma_j^0 - \zeta < z \leq \gamma_j^0)$ and $\overline{\mathbf{X}}_\zeta = \mathbf{X} * \mathbf{I}(\gamma_j^0 < z \leq \gamma_j^0 + \zeta)$ for a small $\zeta - \textit{neighborhood}$ of each of the m true threshold parameters and $\forall j$.

Assumption A1 *(i) The minimum eigenvalues of* $\mathbf{X}_\zeta' \mathbf{X}_\zeta / T$ *and* $\overline{\mathbf{X}}_\zeta' \overline{\mathbf{X}}_\zeta / T$ *are bounded away from zero in probability for large* T *and (ii) the threshold variable* z_t *has a positive density on* $[\underline{\gamma}, \bar{\gamma}]$.

Part (i) of the above assumption ensures that there are enough observations around each true threshold parameter so that they can be identified. It implies that $\mathbf{X}_\zeta$ and $\overline{\mathbf{X}}_\zeta$ have full column rank for T sufficiently large. Part (ii) rules out the possibility that two distinct threshold values produce the same fit. In practice the estimation procedure is conducted by imposing an ad-hoc lower bound for the number of observations present in each regime by requiring $T_j / T \geq \lambda$, with

λ typically set to 10% or 15% (see Andrews (1993), Hansen (1996, 1999a), Bai and Perron (1998, 2000a, 2000b)).

Conditional on $(\gamma_1, \ldots, \gamma_m)$ the model in (2) is linear in the $\boldsymbol{\beta}_j's$ and thus the application of the least squares principle leads to the concentrated sum of squared errors function

$$S_T(\gamma_1, \ldots, \gamma_m) = \mathbf{y}'\mathbf{y} - \sum_{j=1}^{m+1} \mathbf{y}'\mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{y} \tag{3}$$

from which the threshold parameters can be jointly estimated through the following optimization program

$$(\hat{\gamma}_1, \ldots, \hat{\gamma}_m) = \arg \min_{(\gamma_1, \ldots, \gamma_m) \in \Gamma_m} S_T(\gamma_1, \ldots, \gamma_m). \tag{4}$$

The slope parameter estimates can then be computed as $\hat{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\beta}}_j(\hat{\gamma}_1, \ldots, \hat{\gamma}_m)$. We next introduce a set of high level assumptions which will allow us to establish the limiting properties of both the joint and sequential threshold parameter estimators. We let $(\gamma_1^0, \ldots, \gamma_m^0)$ denote the true configuration of threshold parameters and $\mathbf{X}_j^0 = \mathbf{X} * \mathbf{I}(\gamma_{j-1}^0 < z \leq \gamma_j^0)$ $\forall j = 1, \ldots, m+1$ refers to the corresponding regressor matrix.

Assumption A2 *As $T \to \infty$, uniformly over $\gamma_j \in \Re$*

(i) $\dfrac{\mathbf{X}_j'\mathbf{X}_j^0}{T} \xrightarrow{p} \Big[\mathbf{G}(\gamma_j \wedge \gamma_j^0) - \mathbf{G}(\gamma_{j-1} \wedge \gamma_j^0)\Big] - \Big[\mathbf{G}(\gamma_j \wedge \gamma_{j-1}^0) - \mathbf{G}(\gamma_{j-1} \wedge \gamma_{j-1}^0)\Big]$,

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

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

where $\mathbf{G}(\gamma_j^0)$ are finite symmetric positive definite matrices $\forall j$ and the $\mathbf{G}(\gamma_j)'s$ are finite symmetric positive definite matrices, absolutely continuous and strictly increasing functions of γ_j, $\forall j = 1, \ldots, m+1$.

In what follows it will also be understood that $\mathbf{G}(\gamma_0^0 \wedge .) \equiv \mathbf{0}$, $\mathbf{G}(\gamma_0 \wedge .) \equiv \mathbf{0}$, $\mathbf{G}(\gamma_{m+1} \wedge \gamma_m^0) \equiv \mathbf{G}(\gamma_m^0)$, and $\mathbf{G}(\gamma_m \wedge \gamma_{m+1}^0) \equiv \mathbf{G}(\gamma_m)$. Within our notational conventions it is also implicit that $\mathbf{G}(\gamma_{m+1} \wedge \gamma_{m+1}^0) = \mathbf{G} \succ 0$ together with $\mathbf{G}(\gamma_{m+1}) \equiv \mathbf{G}(\gamma_{m+1}^0) \equiv \mathbf{G} \succ 0$. Thus an immediate consequence of assumption **A2**(i) is that $\mathbf{X}'\mathbf{X}/T \xrightarrow{p} \mathbf{G} \succ 0$.

Assumptions **A2**(i)-(ii) are law of large number type of conditions. They exclude integrated processes and hold if for instance (x_t, z_t, ϵ_t) is strictly stationary and ergodic and the threshold variable

z_t has a continuous distribution (see Lemma 1, Hansen (1996)). Assumption **A2**(iii) is a central limit theorem type of assumption. It again excludes integrated processes. Sufficient conditions for this boundedness in probability to hold involve requiring an appropriate mixing decay rate for the above sequence as in assumption 1 of Hansen (2000) (see also Tsay (1998)) combined with finite fourth order moment conditions $E|x_t^4| < \infty$, $E|x_t\epsilon_t|^4 < \infty$ and a bounded density for z_t (see Lemma A.4 in Hansen (2000)). The above assumptions hold under a wide range of specifications considered in applied work. If y_t is generated by a SETAR process for instance then from Chan (1990, 1993), **A2**(i)-(iii) hold provided that the relevant characteristic polynomials have roots that lie outside the unit circle, the error process is iid with a bounded and continuous pdf (see also Hansen (1996, pp. 420-422) for a more general discussion on specifications under which **A2** holds). Assumptions **A2**(i)-(iii) will also hold under the framework of the threshold unit root model considered in Gonzalez and Gonzalo (1997) but will not hold for the threshold stochastic unit root model (TSTUR) considered in Gonzalo and Montesinos (2000) since in general the model will not be either weakly stationary or ergodic.

The limiting behaviour of the jointly estimated threshold parameters is summarized in the following proposition

Proposition 2.1 *As $T \to \infty$ and under A1 and A2(i)-(ii) we have $\hat{\gamma}_i \xrightarrow{p} \gamma_i^0$, $i = 1, \ldots, m$.*

The above joint estimators are straightforward to compute when the model is characterized by two regimes ($m = 1$) since the optimization program in (4) requires a one-dimensional grid search only. When $m > 1$ however, the computational burden becomes substantial, requiring multi-parameter grid based simulations over all possible values of all threshold parameters taken together. The problem in hand is analogous to the computational problems that arise when dealing with multiple change-point models, recently investigated by Bai (1997), Bai and Perron (1998, 2000a, 2000b) and in the earlier work of Hawkins (1976) and Vostrikova (1981). In that literature it has been suggested that one may proceed sequentially by estimating the change-points one at a time since the change-point estimator obtained as an optimizer of a misspecified single parameter based objective function (derived from a fitted model with a single break while the true model contains more than one) maintains its consistency property for one of the true change-points. Given the similarities between threshold and change-point models, Hansen (1999b) also conjectured that a similar feature should hold when fitting threshold models. To our knowledge however the recent literature does

not provide any formal proof of the above result in the context of general threshold models such as the specification considered in (2) and even in the context of standard change-point models, the properties of the sequential estimation approach have only been established for simple mean shift models with no other included regressors (see Bai (1997), Bai and Perron (1998), Altissimo and Corradi (1999)).

Our next objective therefore is to formally establish the properties of threshold estimators obtained via a sequential estimation approach, requiring solely a single parameter based grid search in each sequence. We initially concentrate on the limiting behaviour of a single threshold parameter estimate obtained from a fitted two-regime specification when the true model is given by (2). This will subsequently allow us to formally establish the properties of a sequential algorithm for estimating all threshold parameters one at a time. Specifically, the fitted model is now given by

$$\mathbf{y} = \mathbf{Z}_1\boldsymbol{\delta}_1 + \mathbf{Z}_2\boldsymbol{\delta}_2 + \mathbf{u}, \tag{5}$$

where $\mathbf{Z}_1 = \mathbf{X} * \mathbf{I}(z \leq r)$ and $\mathbf{Z}_2 = \mathbf{X} * \mathbf{I}(z > r)$ while the true model is specified as in (2). Note that $\mathbf{Z}_1 + \mathbf{Z}_2 = \mathbf{X}$ and $\mathbf{Z}_1'\mathbf{Z}_2 = \mathbf{0}$. Applying the conditional least squares approach outlined above to (5) leads to the following optimization program for the threshold parameter estimator

$$\hat{r} = \arg\min_{r \in \Gamma_1} S_T(r) \tag{6}$$

where

$$S_T(r) = \mathbf{y}'\mathbf{y} - \sum_{j=1}^{2} \mathbf{y}'\mathbf{Z}_j(\mathbf{Z}_j'\mathbf{Z}_j)^{-1}\mathbf{Z}_j'\mathbf{y} \tag{7}$$

and Γ_1 is the sample space of the threshold variable given by the "merged" version of Γ_m, i.e. $\Gamma_1 = [\underline{\gamma}, \bar{\gamma}]$. For greater technical convenience it is useful to define an alternative objective function $J_T(r) = S_T - S_T(r)$, with $S_T = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ denoting the sum of squared errors obtained under the restriction $\boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_{m+1}$ imposed on (2). More specifically, recalling that $\mathbf{X} = \mathbf{Z}_1 + \mathbf{Z}_2$ together with $\mathbf{Z}_j'\mathbf{y} = (\mathbf{Z}_j'\mathbf{Z}_j)\hat{\boldsymbol{\delta}}_j$ for $j = 1, 2$ which follows from the least squares formula applied to (5), and using (7) we obtain

$$J_T(r) = (\hat{\boldsymbol{\delta}}_2 - \hat{\boldsymbol{\delta}}_1)'\mathbf{Z}_2'\mathbf{Z}_2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{Z}_1'\mathbf{Z}_1(\hat{\boldsymbol{\delta}}_2 - \hat{\boldsymbol{\delta}}_1). \tag{8}$$

The optimization program in (6) is now reformulated as

$$\hat{r} = \arg\max_{r \in \Gamma_1} J_T(r). \tag{9}$$

The limiting behaviour of a properly normalized version of $J_T(r)$ is established in the following lemma

Lemma 2.1 *As $T \to \infty$ and under A1 and A2(i)-(ii) we have*

$$\sup_{r \in \Gamma_1} \left| \frac{J_T(r)}{T} - J_\infty(r) \right| \xrightarrow{p} 0$$

where $J_\infty(r)$ is a nonstochastic continuous function given by

$$\begin{aligned} J_\infty(r) &= \left[\sum_{\ell=1}^{m} \boldsymbol{\rho}_\ell' \mathbf{G}(r \wedge \gamma_\ell^0) \mathbf{G}(r)^{-1} + \sum_{\ell=1}^{m} \boldsymbol{\rho}_\ell' (\mathbf{G}(r \wedge \gamma_\ell^0) - \mathbf{G}(\gamma_\ell^0))(\mathbf{G} - \mathbf{G}(r))^{-1} \right] \\ &\quad (\mathbf{G} - \mathbf{G}(r)) \mathbf{G}^{-1} \mathbf{G}(r) \\ &\quad \left[\mathbf{G}(r)^{-1} \sum_{\ell=1}^{m} \mathbf{G}(r \wedge \gamma_\ell^0) \boldsymbol{\rho}_\ell + (\mathbf{G} - \mathbf{G}(r))^{-1} \sum_{\ell=1}^{m} (\mathbf{G}(r \wedge \gamma_\ell^0) - \mathbf{G}(\gamma_\ell^0)) \boldsymbol{\rho}_\ell \right] \end{aligned} \tag{10}$$

with $\boldsymbol{\rho}_\ell = (\boldsymbol{\beta}_\ell - \boldsymbol{\beta}_{\ell+1})$.

The above limit function $J_\infty(r)$ will have different expressions over the $m+1$ regimes. For $r = \gamma_k^0$ and $k = 1, \ldots, m$ we have

$$\begin{aligned} J_\infty(r = \gamma_k^0) &= \left[\sum_{\ell=1}^{k} \boldsymbol{\rho}_\ell' \mathbf{G}(\gamma_\ell^0) \mathbf{G}(\gamma_k^0)^{-1} + \sum_{\ell=k+1}^{m} \boldsymbol{\rho}_\ell' (\mathbf{G} - \mathbf{G}(\gamma_\ell^0))(\mathbf{G} - \mathbf{G}(\gamma_k^0))^{-1} \right] \\ &\quad (\mathbf{G} - \mathbf{G}(\gamma_k^0)) \mathbf{G}^{-1} \mathbf{G}(\gamma_k^0) \\ &\quad \left[\mathbf{G}(\gamma_k^0)^{-1} \sum_{\ell=1}^{k} \mathbf{G}(\gamma_\ell^0) \boldsymbol{\rho}_\ell + (\mathbf{G} - \mathbf{G}(\gamma_k^0))^{-1} \sum_{\ell=k+1}^{m} (\mathbf{G} - \mathbf{G}(\gamma_\ell^0) \boldsymbol{\rho}_\ell \right], \end{aligned} \tag{11}$$

and for $r \in (\gamma_k^0, \gamma_{k+1}^0)$ with $k = 1, \ldots, m-1$ we have

$$\begin{aligned} J_\infty(r \in (\gamma_k^0, \gamma_{k+1}^0)) &= \left[\sum_{\ell=1}^{k} \boldsymbol{\rho}_\ell' \mathbf{G}(\gamma_\ell^0) \mathbf{G}(r)^{-1} + \sum_{\ell=k+1}^{m} \boldsymbol{\rho}_\ell' (\mathbf{G} - \mathbf{G}(\gamma_\ell^0))(\mathbf{G} - \mathbf{G}(r))^{-1} \right] \\ &\quad (\mathbf{G} - \mathbf{G}(r)) \mathbf{G}^{-1} \mathbf{G}(r) \\ &\quad \left[\mathbf{G}(r)^{-1} \sum_{\ell=1}^{k} \mathbf{G}(\gamma_\ell^0) \boldsymbol{\rho}_\ell + (\mathbf{G} - \mathbf{G}(r))^{-1} \sum_{\ell=k+1}^{m} (\mathbf{G} - \mathbf{G}(\gamma_\ell^0) \boldsymbol{\rho}_\ell \right]. \end{aligned} \tag{12}$$

Following the derivation of the uniform limit in (10), the most important subsequent step in the evaluation of the asymptotic properties of the extremum estimator defined in (9) involves establishing the existence of a unique maximum of $J_\infty(r)$. Since $J_\infty(r)$ may have multiple local maxima we initially introduce an assumption ensuring that one of the true thresholds dominates in the data, in the sense that among the m true threshold parameters there is one that most contributes to the maximization of $J_\infty(r)$. We subsequently establish that $J_\infty(r)$ has a unique maximum that occurs at that dominant threshold parameter.

Assumption A3 *There exists a single threshold parameter say* $\gamma^0_{(1)} \in \{\gamma^0_1, \ldots, \gamma^0_m\}$ *such that* $J_\infty(r = \gamma^0_{(1)}) > J_\infty(r = \gamma^0_k)$ $\forall \gamma^0_k \neq \gamma^0_{(1)}$ *and* $k = 1, \ldots, m$.

According to the above assumption $\gamma^0_{(1)}$ strictly dominates all the remaining $m-1$ threshold parameters in terms of their contribution to the maximization of $J_\infty(r)$. Note also that $\gamma^0_{(1)}$ could correspond to any of the $\gamma^{0\prime}_i s$ $\forall i = 1, \ldots, m$ and is not necessarily equal to γ^0_1. To better highlight the meaning of a dominant threshold parameter as described in assumption **A3** we consider the case of a three regime model with $\gamma^0_{(1)} = \gamma^0_1$ for instance (i.e. assuming that in a three regime model the first true threshold parameter *dominates*). Given the expression of $J_\infty(r = \gamma^0_k)$ in (11) the above assumption then translates into the following requirement on the limiting objective function

$$
\begin{aligned}
J_\infty(\gamma^0_1) - J_\infty(\gamma^0_2) &= \boldsymbol{\rho}_1' \mathbf{G}(\gamma^0_1)\mathbf{G}(\gamma^0_2)^{-1}(\mathbf{G}(\gamma^0_2) - \mathbf{G}(\gamma^0_1))\boldsymbol{\rho}_1 \\
&- \boldsymbol{\rho}_2'(\mathbf{G}(\gamma^0_2) - \mathbf{G}(\gamma^0_1))(\mathbf{G} - \mathbf{G}(\gamma^0_1))^{-1}(\mathbf{G} - \mathbf{G}(\gamma^0_2))\boldsymbol{\rho}_2 > 0. \qquad (13)
\end{aligned}
$$

Since $\boldsymbol{\rho}_1 = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$ the above will be true for instance if the slopes corresponding to the first and second regimes are sufficiently far apart and/or a large proportion of the observations belongs to the first regime. Note also that (13) can be seen as analogous to condition (6) of Bai (1997, p. 319) in the context of a multiple change-point framework. The next lemma establishes the existence of a unique maximum of the limiting objective function.

Lemma 2.2 *Under A3 the limiting functional* $J_\infty(r)$ *in (10) is uniquely maximized at* $r = \gamma^0_{(1)}$.

The following two propositions next focus on the consistency and rate of convergence of the threshold parameter estimator defined in (6) or (9).

Proposition 2.2 *As* $T \to \infty$ *and under A1, A2(i)-(ii) and A3 we have* $\hat{r} \xrightarrow{p} \gamma^0_{(1)}$.

Proposition 2.3 *As* $T \to \infty$ *and under A1, A2(i)-(ii) and A3 we have* $T|\hat{r} - \gamma^0_{(1)}| = O_p(1)$.

Propositions 2.1 and 2.3 establish that the single threshold parameter estimator obtained from a misspecified two regime model is T-consistent for one of the m true threshold parameters. More specifically it is consistent for the threshold parameter $\gamma^0_{(1)} \in \{\gamma^0_1, \ldots, \gamma^0_m\}$ that most contributes to the maximization of the objective function. Although assumption **A3** is not restrictive from a practical perspective we conjecture that it would still be possible to establish results analogous to our propositions 2.1 and 2.3 while maintaining the possibility that $J_\infty(r)$ has m local maxima. This would require the use of different technical tools and an analysis along the lines of Bai (1997) who focused on a three regimes mean-shift framework and established the convergence in distribution

of the single change-point estimator to a random variable with equal mass at the two local optima of the limiting objective function.

The above results provide a rationale for a sequential estimation algorithm of the m true threshold parameters by proceeding one at a time via a sequence of m one-dimensional optimization programs as in (9) over appropriately defined search domains. Once the first step estimate, say $\hat{r}^{(1)}$ has been obtained for instance we can proceed conditional on $\hat{r}^{(1)}$ and estimate the second threshold parameter by evaluating a second stage objective function analogous to (8), say $J_T(r|\hat{r}^{(1)})$ over $r \in (\underline{\gamma}, \hat{r}^{(1)}) \cup (\hat{r}^{(1)}, \bar{\gamma})$. This is due to the fact that although $\hat{r}^{(1)}$ is T-consistent for one of the m true threshold parameters, in practice it is not known to which true threshold parameter $\gamma^0_{(1)}$ corresponds to. Thus in the second stage we need to consider search regions that lie to the left as well as to the right of $\hat{r}^{(1)}$. More generally, suppose that we have estimated $h-1$ threshold parameters $(\hat{r}^{(1)}, \ldots, \hat{r}^{(h-1)})$ by proceeding as described above, and let $(\hat{r}_{(1)}, \ldots, \hat{r}_{(h-1)})$ denote their ordered counterpart (note that the ordering of the first $h-1$ sequentially obtained estimates is known when proceeding with the estimation of the h^{th} threshold parameter estimator). The estimation of the h^{th} threshold parameter estimator will then involve maximizing $J_T(r|\hat{r}_{(1)}, \ldots, \hat{r}_{(h-1)})$ over $r \in (\underline{\gamma}, \hat{r}_{(1)}) \cup \ldots \cup (\hat{r}_{(h-2)}, \hat{r}_{(h-1)}) \cup (\hat{r}_{(h-1)}, \bar{\gamma})$. More specifically, letting $\widehat{\mathbf{Z}}_i = \mathbf{X} * \mathbf{I}(\hat{r}_{(i-1)} \leq z \leq \hat{r}_{(i)})$ for $i = 1, \ldots, h$ with the convention that $\hat{r}_{(0)} = \underline{\gamma}$ and $\hat{r}_{(h)} = \bar{\gamma}$ and introducing the corresponding projection matrices

$$\mathbf{Q}_\ell = \mathbf{I} - \sum_{\substack{i=1 \\ i \neq \ell}}^{h} \widehat{\mathbf{Z}}_i (\widehat{\mathbf{Z}}_i' \widehat{\mathbf{Z}}_i)^{-1} \widehat{\mathbf{Z}}_i' \qquad \ell = 1, \ldots, h, \tag{14}$$

the estimator of the h^{th} threshold parameter can then be defined as

$$\hat{r}^{(h)} = \arg\max_r J_T(r|\hat{r}_{(1)}, \ldots, \hat{r}_{(h-1)}) \tag{15}$$

with

$$J_T(r|\hat{r}_{(1)}, \ldots, \hat{r}_{(h-1)}) = \sum_{\ell=1}^{h} J_{\ell T}(r|\hat{r}_{(1)}, \ldots, \hat{r}_{(h-1)}) I(\hat{r}_{(\ell-1)} < r < \hat{r}_{(\ell)}) \tag{16}$$

and where $J_{\ell T}(r|\hat{r}_{(1)}, \ldots, \hat{r}_{(h-1)})$ corresponds to an objective function analogous to (8) but derived from each of the following h canonical forms of (5) instead,

$$\mathbf{Q}_\ell \mathbf{y} = \mathbf{Z}_{1,\ell} \boldsymbol{\delta}_{1,\ell} + \mathbf{Z}_{2,\ell} \boldsymbol{\delta}_{2,\ell} + \mathbf{u}_\ell \qquad \ell = 1, \ldots, h, \tag{17}$$

where $\mathbf{Z}_{1,\ell}$ and $\mathbf{Z}_{2,\ell}$ are defined as in (5) but with $r \in (\hat{r}_{(\ell-1)}, \hat{r}_{(\ell)})$, i.e. $\mathbf{Z}_{1,\ell} = \mathbf{X} * \mathbf{I}(\hat{r}_{(\ell-1)} < z \leq r)$,

$\mathbf{Z}_{2,\ell} = \mathbf{X} * \mathbf{I}(r < z < \hat{r}_{(\ell)})$ and $\mathbf{Z}_{1,\ell} + \mathbf{Z}_{2,\ell} = \widehat{\mathbf{Z}}_\ell$. Specifically,

$$J_{\ell T}(r|\hat{r}_{(1)}, \ldots, \hat{r}_{(h-1)}) = (\hat{\boldsymbol{\delta}}_{2,\ell} - \hat{\boldsymbol{\delta}}_{1,\ell})' \mathbf{Z}'_{2,\ell} \mathbf{Z}_{2,\ell} (\widehat{\mathbf{Z}}'_\ell \widehat{\mathbf{Z}}_\ell)^{-1} \mathbf{Z}'_{1,\ell} \mathbf{Z}_{1,\ell} (\hat{\boldsymbol{\delta}}_{2,\ell} - \hat{\boldsymbol{\delta}}_{1,\ell}). \tag{18}$$

Note that since $\widehat{\mathbf{Z}}'_i \mathbf{Z}_{1,\ell} = \mathbf{0}$ and $\widehat{\mathbf{Z}}'_i \mathbf{Z}_{2,\ell} = \mathbf{0}$ $\forall i \neq \ell$ and $i = 1, \ldots, h$ it follows that $\mathbf{Z}'_{1,\ell} \mathbf{Q}_\ell = \mathbf{Z}'_{1,\ell}$ and $\mathbf{Z}'_{2,\ell} \mathbf{Q}_\ell = \mathbf{Z}'_{2,\ell}$ and the least squares estimators in (18) are defined as $\hat{\boldsymbol{\delta}}_{1,\ell} = (\mathbf{Z}'_{1,\ell} \mathbf{Z}_{1,\ell})^{-1} \mathbf{Z}_{1,\ell} \mathbf{y}$ and $\hat{\boldsymbol{\delta}}_{2,\ell} = (\mathbf{Z}'_{2,\ell} \mathbf{Z}_{2,\ell})^{-1} \mathbf{Z}_{2,\ell} \mathbf{y}$.

From the above notation it is clear that the consistency of the second stage threshold parameter estimator $\hat{r}^{(2)} = \arg\max_r J_T(r|\hat{r}^{(1)})$ and that of the subsequent ones can be established in exactly the same manner as for $\hat{r}^{(1)}$. For this purpose we need to introduce a generalization of assumption **A3** requiring that in each of the m estimation sequences there is an ordering among the true threshold parameters in terms of their contribution to the maximization of the limiting objective function evaluated at that sequence. Specifically we let $(\gamma^0_{(1)}, \gamma^0_{(2)}, \ldots, \gamma^0_{(m)})$ denote a particular configuration of the m true threshold parameters appearing not necessarily in the same order as the true configuration $(\gamma^0_1, \gamma^0_2, \ldots, \gamma^0_m)$ (i.e. $\gamma^0_{(i)} = \gamma^0_j$, $\forall i, j = 1, \ldots, m$ but with i not necessarily equal to j) and also let $J_\infty(r|\gamma^0_{(1)}, \ldots, \gamma^0_{(h-1)})$ denote the limiting objective function associated with the $(h-1)^{th}$ estimation sequence. We assume the following

Assumption A4 *There exists a configuration $(\gamma^0_{(1)}, \gamma^0_{(2)}, \ldots, \gamma^0_{(m)})$ of the m true threshold parameters such that $J_\infty(\gamma^0_{(h)}|\gamma^0_{(1)}, \ldots, \gamma^0_{(h-1)}) > J_\infty(\gamma^0_k|\gamma^0_{(1)}, \ldots, \gamma^0_{(h-1)})$ $\forall \gamma^0_k \in \{\gamma^0_{(h+1)}, \ldots, \gamma^0_{(m)}\}$ and $h = 1, \ldots, m$.*

The above assumption is a generalization of **A3** in the sense that we now require that in each of the h estimation sequences a single true threshold parameter dominates the remaining $m - h$ in terms of its contribution to the maximization of the corresponding limiting objective function. Given **A4** we can now generalize our two previous propositions to the entire configuration of sequentially estimated threshold parameters.

Proposition 2.4 *As $T \to \infty$ and under A1, A2(i)-(ii) and A4 we have (a) $\hat{r}^{(h)} \xrightarrow{p} \gamma^0_{(h)}$ and (b) $T|\hat{r}^{(h)} - \gamma^0_{(h)}| = O_p(1)$, $\forall h = 1, \ldots, m$.*

Although it is beyond our scope to concentrate on the limiting distributions of the threshold parameter estimators it is also important to mention that analogous to the change-point framework of Bai (1997), the first $m - 1$ sequentially obtained threshold parameter estimators will not have the same limiting distribution as their jointly estimated counterparts since the former have been

estimated using misspecified objective functions contaminated by the wrongly omitted thresholds and as a result will be less efficient regardless of the sample size. Note that this will not be the case for $\hat{r}^{(m)}$ the threshold parameter estimator obtained in the last sequence. It is however possible to refine the sequentially obtained estimates so as to make them have the same asymptotic distribution as their jointly estimated counterparts. This is achieved by adapting the technique referred to as *repartition* in Bai (1997) to this multiple threshold framework. The approach is straightforward to implement in practice and involves reestimating the threshold parameters conditionally on the initially estimated ones so that each refined estimate is obtained without an underlying neglected regime. Under $m = 2$ for instance this can be achieved by reestimating $r^{(1)}$ taking $\hat{r}^{(2)}$ as given and subsequently reestimating $\hat{r}^{(2)}$ taking the refined first stage estimate as given. This is the principle adopted in the analysis that follows.

2.1 Empirical Properties

Having established the consistency of the joint and sequential estimators, our next objective is to evaluate their relative behaviour in finite samples, viewing the joint estimation as the benchmark case. Our empirical results will also provide an overall picture of the finite sample behaviour and quality of estimators derived from threshold type specifications, features that to our knowledge have not been investigated in the recent time series literature and that are crucial for applied research. Given the computational burden that arises when dealing with models having more than three regimes we limit our analysis of the properties of the jointly estimated threshold parameters to models with at most two threshold parameters (three regimes).

Before proceeding with the empirical performance of the threshold parameter estimators however, it is important to highlight some difficulties that arise when designing a threshold type data generating process. The problem is related to the sensitivity of the variance of the estimators of the slopes (and implicitly that of the threshold parameter estimators) to the choice of the true threshold level. In a two regime (single threshold parameter) setup for instance one would expect to obtain more accurate estimates of both the threshold parameter and slopes if the true threshold parameter is set equal to the median or mean of the distribution of the threshold variable. In practice however it is often impossible to evaluate the moments of the threshold variable appearing in the DGP analytically making the interpretation of the resulting estimators (empirical bias, variance etc) extremely sensitive to the choice of the true threshold parameter. It is this latter

aspect that we wish to initially illustrate by concentrating on a very simple DGP that lends itself to analytically tractable results. This will then allow us to achieve a fairer interpretation of our subsequent simulations based on richer dynamic structures.

We initially consider the following two regime model

$$y_t = \beta_1 I(y_{t-1} \leq \gamma_1) + \beta_2 I(y_{t-1} > \gamma_1) + \epsilon_t \tag{19}$$

where $\epsilon_t \equiv NID(0, \sigma_\epsilon^2)$ with σ_ϵ^2 set equal to 1 with no loss of generality. We also let γ_1^0 denote the true value of the threshold parameter and $\beta_1(\gamma_1)$, $\beta_2(\gamma_1)$ and $\sigma_\epsilon^2(\gamma_1)$ refer to the probability limits of $\hat{\beta}_1(\gamma_1)$, $\hat{\beta}_2(\gamma_1)$ and $\hat{\sigma}_\epsilon^2(\gamma_1)$ respectively. Letting $\Phi()$ denote the c.d.f. of a standard normal random variable and noting that $I(y_{t-1} \leq \gamma_1)$ is a Markov Chain, standard calculations using its transition matrix lead to $P(y_t \leq \gamma_1) = \Phi(\gamma_1 - \beta_2)/(1 - \Phi(\gamma_1 - \beta_1) + \Phi(\gamma_1 - \beta_2)) \equiv \pi(\gamma_1)$ from which it is straightforward to obtain

$$\beta_2(\gamma_1) - \beta_1(\gamma_1) = (\beta_2 - \beta_1)\frac{\pi(\gamma_1 \wedge \gamma_1^0) - \pi(\gamma_1)\pi(\gamma_1^0)}{\pi(\gamma_1)(1 - \pi(\gamma_1))} \tag{20}$$

and

$$\begin{aligned} \sigma_\epsilon^2(\gamma_1) &= \sigma_\epsilon^2 + (\beta_2 - \beta_1)^2 \pi(\gamma_1^0)(1 - \pi(\gamma_1^0)) \\ &\quad -(\beta_2 - \beta_1)^2 \frac{[\pi(\gamma_1 \wedge \gamma_1^0) - \pi(\gamma_1)\pi(\gamma_1^0)]^2}{\pi(\gamma_1)(1 - \pi(\gamma_1))} \end{aligned} \tag{21}$$

where $\pi(\gamma_1 \wedge \gamma_1^0) = \pi(\gamma_1) I(\gamma_1 \leq \gamma_1^0) + \pi(\gamma_1^0) I(\gamma_1 > \gamma_1^0)$. From the expression of $\pi(\gamma_1)$ it is clear that under the above DGP we will have $\pi(\gamma_1) = 0.5$ when $\gamma_1 = 0.5(\beta_1 + \beta_2)$ also implying that $\gamma_1 = E(y_t)$. In other words choosing a true threshold parameter equal to the average of the parameters appearing in each regime ensures that it will also equal to the mean and median of the threshold variable, thus leaving an equal number of observations in both regimes. Our next objective is to evaluate the limiting behaviour of the variance of $\hat{\beta}_2(\gamma_1) - \hat{\beta}_1(\gamma_1)$. The latter should provide valuable information about the impact of the location of the true threshold parameter γ_1^0 on the estimators of the parameters. Standard calculations lead to

$$V_T(\hat{\beta}_2(\gamma_1) - \hat{\beta}_1(\gamma_1)) \rightarrow \frac{\sigma_\epsilon^2(\gamma_1)}{\pi(\gamma_1)(1 - \pi(\gamma_1))} \tag{22}$$

with $\sigma_\epsilon^2(\gamma_1)$ defined in (21). Under $\beta_1 = 1$ and $\beta_2 = 2$ for instance and using the expression of $\pi(\gamma_1)$ derived above we can establish that the value of γ_1 corresponding to the first quartile (25% upper regime, 75% lower regime) is 1.022 and that corresponding to the third quartile is 1.979 with the

median located at $\gamma_1 = 1.5$. More importantly the limiting variance of $\hat{\beta}_2(\gamma_1) - \hat{\beta}_1(\gamma_1)$ draws like a U-shaped curve, centered at $\gamma_1 = 0.5(\beta_1 + \beta_2)\ \forall \gamma_1^0$ and increasing rapidly when we move outside the flat horizontal region. This suggests that choosing γ_1^0 in an improper range will lead to estimators with an extremely high variance, relative to the most favourable mean (or median) location. Under $\beta_1 = 1$ and $\beta_2 = 2$ for instance, the parabola is centered at $\gamma_1 = 1.5\ \forall \gamma_1^0$ with the corresponding variance equal to 4 while the variance corresponding to $\gamma_1^0 = 0$ for instance is close to 40, a ten-fold increase. In order to illustrate the usefulness of the above points we conducted a simulation experiment using the DGP in (19) and evaluated the empirical bias and variance of $\hat{\gamma}_1$ for different values of γ_1^0 together with the corresponding magnitudes for the slope estimates. Specifically, we chose $\gamma_1^0 \in \{0.75, 1.00, 1.50, 2.40\}$ corresponding to first regime proportions of 15.0%, 24.0%, 50.0% and 89.0% respectively. Results are displayed in Table 1.

Table 1 about here

It is immediately clear that the threshold parameter estimate becomes highly imprecise for values of γ_1^0 that fall outside the [1,2] range, with a typically greater than three-fold increase in its empirical standard deviation. Note that the corresponding empirical first regime proportions were 16.1%, 24.7%, 50.3% and 86.6% respectively, remarkably close to their theoretical counterparts. The third and fourth columns of Table 1 display the empirical means and standard deviations of the resulting estimated slope parameters $\hat{\beta}_1(\hat{\gamma}_1)$ and $\hat{\beta}_2(\hat{\gamma}_1)$. It is interesting to note that the latter display a substantially smaller bias and a more stable variability when compared with that of the threshold parameter estimates. In summary the purpose of this preliminary exercise was to highlight the importance of experiment design when considering threshold type DGPs and that extreme caution should be taken when selecting the magnitude of γ_0. Ideally for results to give a sufficiently global picture it is an important imperative to scan across a wide range of possible true threshold parameter values since for models with richer dynamics, many of our analytical results would be unfeasible to obtain.

We next concentrate on a similar specification with three regimes given by

$$y_t = \beta_1 I(y_{t-1} \leq \gamma_1) + \beta_2 I(\gamma_1 < y_{t-1} \leq \gamma_2) + \beta_3 I(y_{t-1} > \gamma_2) + \epsilon_t. \tag{23}$$

Under the above true model and using standard but lengthy algebra we have

$$P(y_t \leq \gamma_1) = \frac{\Phi(\gamma_2 - \beta_3)\Phi(\gamma_1 - \beta_2) + \Phi(\gamma_1 - \beta_3)\Phi(\beta_2 - \gamma_2)}{\Delta(\gamma_1, \gamma_2)} \tag{24}$$

and

$$P(\gamma_1 < y_t \leq \gamma_2) = \frac{\Phi(\gamma_2 - \beta_3)\Phi(\beta_1 - \gamma_1) - \Phi(\gamma_1 - \beta_3)\Phi(\beta_1 - \gamma_2)}{\Delta(\gamma_1, \gamma_2)} \tag{25}$$

where

$$\begin{aligned} \Delta(\gamma_1, \gamma_2) &= [\Phi(\gamma_1 - \beta_2) - \Phi(\gamma_1 - \beta_3)][\Phi(\beta_1 - \gamma_2) + \Phi(\gamma_2 - \beta_3)] \\ &+ [\Phi(\beta_2 - \gamma_2) + \Phi(\gamma_2 - \beta_3)][\Phi(\beta_1 - \gamma_1) + \Phi(\gamma_1 - \beta_3)] \end{aligned} \tag{26}$$

and $P(y_t > \gamma_2) = 1 - P(y_t \leq \gamma_1) - P(\gamma_1 < y_t \leq \gamma_2)$.

Our next objective therefore involves comparing the finite sample properties of the joint and sequential estimation approaches when applied to (23). We concentrate on DGPs given by (23) with $\beta_1 = 1$, $\beta_2 = 2$, $\beta_3 = 3$ and $\epsilon_t \equiv NID(0,1)$. The chosen threshold parameter structure encompasses a wide range of configurations leading to models with approximately equally divided regime proportions as well as models in which a single regime dominates. Specifically we consider $(\gamma_1^0, \gamma_2^0) = (1,2), (1.5,2.5), (1,3)$ and $(2,3)$ which using (23)-(26) imply regime proportions of approximately $(10\%, 20\%, 70\%)$, $(35\%, 30\%, 35\%)$, $(20\%, 60\%, 20\%)$ and $(70\%, 20\%, 10\%)$ respectively. All our experiments are performed using $T = 200$ across $N = 2000$ replications. The empirical means and corresponding standard deviations of the sequentially and jointly estimated threshold parameters together with the implied $\hat{\beta}'s$ are displayed in Table 2a.

Table 2a about here

As expected the precision of the estimates for both the joint and sequential approaches are highly sensitive to the location of the true threshold parameters with the most favourable scenario occuring when all three regimes have an approximately equal amount of observations. The increase in the variability of the threshold parameter estimators also translates into more imprecise estimated slopes with a quantitatively similar shift in magnitudes. When comparing both methods of estimation it is immediately apparent that the figures corresponding to the sequential and joint approaches are remarkably close, even for the moderately small sample size used in the experiment. Both the point estimates and their corresponding standard errors are virtually identical across all configurations of the true threshold parameters. Table 2b displays the results of a similar exercise using a SETAR(3;1,1,1) model given by $y_t = 0.2y_{t-1}I(y_{t-1} \leq -0.5) + 0.8y_{t-1}I(-0.5 < y_{t-1} \leq 0.5) - 0.5y_{t-1}I(y_{t-1} > 0.5) + \epsilon_t$. The choice of the true parameters is such that the regime proportions are approximately (40%, 35%, 25%).

Table 2b about here

For this scenario, results based on both T=200 and T=400 are presented. Focusing first on the relative behaviour of both estimation techniques it is again clear that they lead to estimates that remain very similar in terms of their finite sample variability and bias even in the context of models with richer dynamic structures. When evaluating the overall quality of the resulting estimators however and regardless of the estimation technique it is important to note the drastic deterioration (in terms of loss of precision and finite sample bias) of both the threshold and slope estimates when moving from the simple threshold model with no conditional mean dynamics in each regime towards a more general SETAR process. In the latter case despite small finite sample biases the threshold parameter estimators display a very high degree of variability which persists even as we move from T=200 to T=400.

3 Estimation under an unknown number of thresholds: A Sequential Model Selection Approach

In the preceding section our analysis was conducted under the assumption that the number of regimes of the threshold models is known. In practice however economic theory rarely offers an intuitive rationale for an à priori imposition of a specific number of regimes in the data. Numerous empirical applications aiming to describe the dynamics of macroeconomic variables have taken the ad-hoc view that two regimes may be appropriate for describing alternative dynamics for expansions and recessions. Others (e.g. Koop and Potter (1999)) have argued that perhaps three regimes, encompassing bad times, good times *and* normal times should be modelled. Given this uncertainty it is then natural to inquire about data-based methods for the determination of the number of regimes.

The literature on threshold models does not seem to offer any formal methodology for detecting the number of regimes in threshold type specifications, beyond the case involving testing single threshold versus linear models. In Chan (1990) for instance, the author obtained the limiting distribution of an LR type test statistic in the context of a general two regime SETAR model, but with the exception of a few special cases the limiting distribution does not lend itself to conventional tabulations due to its dependence on a large number of unknown parameters (e.g. moments of the regressors). More recently Hansen (1996), developed a bootstrap based procedure that allows

the construction of asymptotically valid p-values for a large number of test statistics for the null of linearity versus two regimes. To our knowledge, Hansen's (1996) asymptotic p-value based approach is the only technique that allows the treatment of general threshold type models such as SETAR's of any order since its implementation is not restricted to models with simple dynamics. Although its validity is established for the treatment of the at most two regimes case it is not clear whether Hansen's (1996) approach can be legitimately extended to a framework that allows the sequential determination of the number of regimes when the latter could be greater than two (see Hansen (1999a)). Given the numerous unresolved difficulties arising in this context our objective here is to propose an alternative to sequential testing.

We propose to view the problem of specifying the number of regimes from a model selection perspective in which our main task is to select the optimal model among a portfolio of nested specifications and where the selection is made via the optimization of a penalized objective function. The objective function is such that one of its component is a monotonic function of the model dimension (e.g. the residual variance) and its other component penalizes the increase or decrease of the first component caused by the increase in the model dimension. Within our threshold framework the purpose of the penalty term is to penalize over-segmentation as m is allowed to increase. Formally, letting $S_T(\gamma_1, \ldots, \gamma_m)$ denote the concentrated sum of squared errors defined in (3), then in the spirit of the traditional model selection literature (see Akaike (1973), Hannan and Deistler (1988) and references therein) we introduce the following criterion

$$IC_T(\gamma_1, \ldots, \gamma_m) = \log S_T(\gamma_1, \ldots, \gamma_m) + \frac{\lambda_T}{T}[K(m+1)] \tag{27}$$

where λ_T is a deterministic function of the sample size (or a constant independent of T) that is in turn multiplied by the number of free parameters. Clearly an increase in m will lead to a reduction in $S_T(\gamma_1, \ldots, \gamma_m)$, a reduction that will be penalized due to the resulting increase in the number of estimated parameters. It is also important to observe that the minimization of the above objective function for given m will lead to the same estimates of the threshold parameters as in (4) since the penalty term does not depend on the magnitude of the threshold parameters. In a related study, Liu, Wu and Zidek (1997) also considered a criterion similar to (27) for the estimation of the number of threshold parameters. They used simulation based evidence to introduce a penalty term playing the role of λ_T. Their analysis however is based on a direct joint estimation of the concentrated sum of squared errors function $S_T(\gamma_1, \ldots, \gamma_m)$ and differs from ours in its implementation and

probabilistic framework. The use of a model selection approach to inferences with a criterion analogous to (27) has also been advocated in numerous other areas of the econometric literature, 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.

Noting that under the linear specification the objective function in (27), say $IC_T(0) = \log S_T + \frac{\lambda_T}{T}K$, does not depend on the threshold parameters we can introduce a modified criterion defined as

$$Q_T(m) \quad = \quad IC_T(0) - \min_{\gamma_1,..,\gamma_m} IC_T(\gamma_1,\ldots,\gamma_m)$$

or more specifically as

$$Q_T(m) \quad = \quad \max_{\gamma_1,...,\gamma_m} \log\left[\frac{\hat{\sigma}^2}{\hat{\sigma}^2(\gamma_1,\ldots,\gamma_m)}\right] - \frac{\lambda_T}{T}K\ m, \tag{28}$$

with $\hat{\sigma}^2 = S_T/T$ and $\hat{\sigma}^2(\gamma_1,\ldots,\gamma_m) = S_T(\gamma_1,\ldots,\gamma_m)/T$. The model selection based estimator of the number of unknown threshold parameters can then be formally defined as

$$\hat{m} \quad = \quad \arg\max_{0\leq m\leq M} Q_T(m) \tag{29}$$

for some upperbound $M \geq m_0$. Note that the threshold parameter estimates are implicitely obtained as a by-product of the above regime determination procedure. It is also useful to observe that T times the first component in the right hand side of (28) is the likelihood ratio statistic for testing linearity against $m+1$ regimes. Thus if we let $F_T(\gamma)$ denote any of the conventional LR, Score or Wald type test statistics we can also consider alternative versions of the objective function in (28) by introducing

$$\overline{Q}_T(m) \quad = \quad \max_{\gamma_1,...,\gamma_m} F_T(\gamma_1,\ldots,\gamma_m) - \lambda_T K m, \tag{30}$$

as a more general version of $Q_T(m)$ in (28). This also suggests that the approach can accomodate the presence of heteroscedasticity via the use of heteroscedasticity robust versions of $F_T(.)$ in (30). We next concentrate on the theoretical and empirical properties of the model selection based estimates obtained as a solution to (29).

3.1 m=0 versus m=1 case

When our objective is to select between a linear and a two-regime specification we have $\hat{m} = \arg\max_{0\leq m\leq 1} Q_T(m)$. Recalling that $Q_T(0) = 0$ by construction the model selection procedure

involves accepting the linear specification (m=0) if $Q_T(1) < Q_T(0)$ or equivalently if

$$IC_T(0) \leq \min_{\gamma_1 \in \Gamma_1} IC_T(\gamma_1) \tag{31}$$

and decide for the threshold model when

$$IC_T(0) > IC_T(\gamma_1) \tag{32}$$

for some $\gamma_1 \in \Gamma_1$. Using the expressions of $IC_T(0)$ and $IC_T(\gamma_1)$ given above it is useful to note that the selection rule in (31) can be reformulated as

$$\max_{\gamma_1 \in \Gamma_1} T \log\left[\frac{\hat{\sigma}^2}{\hat{\sigma}^2(\gamma_1)}\right] \leq \lambda_T K \tag{33}$$

or equivalently as

$$\max_{\gamma_1 \in \Gamma_1} \frac{T(\hat{\sigma}^2 - \hat{\sigma}^2(\gamma_1))}{\hat{\sigma}^2(\gamma_1)} \leq T(\exp(\frac{\lambda_T K}{T}) - 1) \tag{34}$$

where $T[\exp\left(\frac{\lambda_T K}{T}\right) - 1] \approx \lambda_T K$. At this stage it is again interesting to note that the quantities appearing on the left hand side of (33) and (34) are conventional likelihood ratio and Wald type test statistics for the hypothesis of linearity versus a two regime threshold model. Their limiting distributions typically depend on unknown and model specific moments and cannot be tabulated. An important advantage of the model selection approach is that it does not rely on the critical values of the test statistics for deciding between the linear and threshold specifications. Instead the decision rule is based on the deterministic penalty term, solely function of the sample size multiplied by the number of free parameters. Equivalently when seen from a conventional testing perspective the above decision rule can be interpreted as using a test statistic in which the significance level is allowed to converge to zero as the sample size increases. Such a strategy has often been advocated when one performs a sequence of nested tests so as to avoid a build up of Type I errors or more generally to make the testing strategy lead to consistent estimates of the number of thresholds.

We next show that the above model selection procedure leads to an estimator of m_0 that is weakly consistent. The result is summarized in the following proposition

Proposition 3.1 *Letting m_0 denote the true number of threshold parameters with $m_0 \in \{0, 1\}$, $\hat{m}$ defined as in (29) with λ_T such that (i) $\lambda_T \to \infty$ and (ii) $\frac{\lambda_T}{T} \to 0$ then under A1-A2(i)-(iii) we have $P(\hat{m} = m_0) \to 1$ as $T \to \infty$.*

The above proposition establishes that with probability tending to one and assuming that $m_0 \in \{0, 1\}$, the model selection procedure leads to an estimated number of threshold parameters that

coincides with the true number provided that the penalty term satisfies conditions (i) and (ii). A possible candidate for the choice of the penalty term is $\lambda_T = \log T$ corresponding to a Schwarz type criterion but clearly the set of possible choices is extremely wide making it difficult to argue for an optimal penalty choice. To our knowledge theoretical guidelines about specific choices of λ_T remain an open question in most frameworks that advocate the use of model selection criteria.

Our next objective is to evaluate the finite sample performance of the alternative criteria across a wider range of DGPs. We initially concentrate on linear models (i.e. $m_0 = 0$) and evaluate the performance (correct decision frequencies) of the various criteria when used for distinguishing between linearity and single threshold type nonlinearity. We initially consider an AR(1) model given by $y_t = \rho y_{t-1} + \epsilon_t$ as our linear DGP and a corresponding fitted threshold model given by $y_t = \rho_1^{(1)} y_{t-1} I(y_{t-1} \leq \gamma_1) + \rho_1^{(2)} y_{t-1} I(y_{t-1} > \gamma_1) + \epsilon_t$. Table 3a presents the correct decision frequencies (i.e. choosing $m = 0$ over $m = 1$) across three sample sizes ($T = 200, 400$ and $T = 600$) and where BIC, AIC, HQ, BIC2 and BIC3 refer to the model selection criteria with penalty terms $\lambda_T = \log T$, $\lambda_T = 2$, $\lambda_T = 2 \log\log T$, $\lambda_T = 2 \log T$ and $\lambda_T = 3 \log T$ respectively. The main motivation for the inclusion of the less familiar penalty terms labeled as BIC2 and BIC3 is to provide a sufficiently general description of the sensitivity of the model selection based decision frequencies to the magnitude of λ_T.

Table 3a about here

The frequencies corresponding to the AIC clearly highlight its inadequacy in this framework, with the criterion shown to point spuriously to the threshold model more than 50% of the times. This empirical frequency further deteriorates as the autoregressive parameter ρ approaches the unit root region. Similarly the HQ criterion, despite its ability to point to the true model asymptotically, is also performing poorly in moderately large samples by wrongly selecting the threshold model close to 30% of the times. As expected from proposition 3.1 the criterion improves its ability to point to the true model as the sample size grows but this latter improvement occurs very slowly reflecting the weakness of the HQ penalty in this context. Among all model selection criteria the best performance is displayed by the BIC and its variants, denoted BIC2 and BIC3. Under $|\rho| < 1$ for instance and for reasonably large sample sizes the BIC is able to point to the linear model more than 93% of the times with a deterioration occuring only under the random walk model. Also, contrary to the linear regression framework the BIC does not appear to lead to

spurious parsimonious choices. Both the BIC2 and BIC3 are pointing to the correct model with an empirical probability close to 1. At this stage however the BIC2 and BIC3 based frequencies must be interpreted with caution since a close to 100% correct decision frequency might be due to a spurious choice of the most parsimonious structure due to the strength of the penalty terms characterizing both criteria.

We next consider a threhold DGP (i.e. $m_0 = 1$) of the form $y_t = \rho y_{t-1} I(y_{t-1} \leq 0) - \rho y_{t-1} I(y_{t-1} > 0) + \epsilon_t$ with $\rho \in \{-0.40, -0.25, -0.15, -0.10, -0.05\}$. Note that as the magnitude of $|\rho|$ decreases, the existence of a two regime process will become more and more difficult to detect. The empirical correct decision frequencies corresponding to this experiment are presented in Table 3b.

Table 3b about here

Table 3b suggests that the BIC and to a lesser extent the BIC2 display the best overall performance, with an excellent ability to point to the true model even for moderately small sample sizes. As expected, the ability of all criteria to point to the correct threshold model decreases with $|\rho|$ but even under $|\rho| = 0.15$ and T=600 the BIC is still able to select the true specification close to 99% of the times, compared with 60% for the BIC2.

3.2 General Case

Here we consider the case where there may be more than one threshold parameter (i.e. more than two regimes) in the set of possible models. Taking advantage of our general result on the consistency of the threshold parameter estimators in underspecified models we propose a sequential model selection based strategy for the estimation of the unknown number of threshold parameters, regardless of their number. Specifically the idea involves first proceeding as in the above section, deciding between a linear model (m=0) and a two regime threshold specification (m=1). If $Q_T(0) > Q_T(1)$ the procedure stops and we decide that the data support the linear model. If $Q_T(0) < Q_T(1)$ we obtain the estimate of the first threshold parameter, say $\hat{r}^{(1)}$ and conditional on this first stage threshold parameter estimator we proceed with a second stage $m = 0$ versus $m = 1$ decision process conducted on both subsamples in order to detect the eventual presence of a second threshold. The procedure continues until the model selection procedure leads to the choice $m = 0$ on all subsamples. More formally, letting $Q_T^{(i,j)}(1)$ denote the magnitude of (28) or (30) obtained in step

i and subsample j, the stopping rule involves concluding for the presence of $m+1$ regimes (or m threshold parameters) when $Q_T^{(m+1,j)}(1) < 0$ for all $j = 1, \ldots, m+1$. The following proposition summarizes the asymptotic properties of the sequentially estimated number of thresholds

Proposition 3.2 *Letting $\hat{m}_{seq}$ denote the number of threshold parameters estimated via the sequential procedure with (i) $\lambda_T \to \infty$ and (ii) $\frac{\lambda_T}{T} \to 0$ then under A1-A2(i)-(iii) and A4 we have $P(\hat{m}_{seq} = m_0) \to 1$ as $T \to \infty$.*

At this stage it is also important to relate our analysis to the recent work on multiple structural breaks developed in a series of recent papers by Bai (1997) and Bai and Perron (1998). Taking advantage of the T-consistency of the sequentially estimated change-points for instance Bai and Perron (1998) also proposed a sequential change-point estimation/detection scheme under an unknown number of breaks. Rather than relying on a model selection based approach however the authors considered a stopping rule based on a sequence of supremum F type tests, the limiting distribution of which was shown to depend solely on the dimension of the parameter vector the stability of which is being tested and the set of all possible values for the break fractions. In a related set of companion papers (Bai and Perron (2000a, 2000b)) the authors also focused on the computational aspects that arise when estimating multiple change-point models. This has allowed them to consider threshold models by reformulating the latter in the form of a multiple change-point specification via an appropriate change in the time scale (see also Tsay (1998)). Besides difficulties that may arise when tied values of the threshold variable are present it is important to note that despite their similarities threshold and change-point models have fundamentally different probabilistic properties. As pointed out in Hansen (2000) for instance the sorting operation when the threshold variable is one of the regressors will induce a trend in the regressors of the change-point counterpart, a framework for which distributional results are not readily available. From a practical modeling perspective it is also not clear how the change-point reparameterization may accomodate frameworks where the threshold variable is composite as for instance in a SETAR model where the regime switches are driven by the first and second lags of the threshold variable. Regarding alternative methods for the determination of the number of regimes a promising new approach in the context of structural breaks is also presented in Altissimo and Corradi (1999) where the authors focused on the estimation of the number of shifts in the mean of a stationary process and designed a procedure that leads to a strongly consistent estimator of the unknown number of breaks. Their procedure is also characterized by both Type I and Type II errors that converge to

zero asymptotically.

In order to evaluate the finite sample behaviour of the sequential model selection based approach described above we next conducted two sets of experiments using models with $m_0 = 1$ (two regimes) and $m_0 = 2$ (three regimes) respectively. We concentrate solely on the properties of the BIC and its two variants since our previous analysis demonstrated the unreliability of alternative criteria such as the AIC or HQ. Results corresponding to the two-regime specification are presented in Table 4a. Note first that the convergence of $\hat{m}$ to its true value $m_0 = 1$ is clearly visible across the increasing sample sizes, with the BIC detecting the true number of threshold parameters more than 90% of the times under T=600 and close to 95% of the times under T=800.

Tables 4a and 4b about here

It is also important to note that the procedure does not display any tendency to under-segment in the sense that the wrong decisions are mostly clustered at $\hat{m} = m_0 + 1$. An overall similar picture also arises from the results corresponding to a true model with three regimes (see Table 4b). Note that here although the chosen specification is globally stationary its corridor regime is characterized by a characteristic polynominal with roots that lie inside the unit circle. For this model, the BIC and its variants do not display any tendency to under-segment and the wrong decisions are again clustered at $m_0 + 1$. Overall the BIC displays desirable large sample properties and a reasonably good finite sample behaviour. Obviously for the latter case one should interpret any experimental result with caution since finite sample simulation based performance can be highly DGP specific. Under our DGP in Table 4a for instance, our choice of true parameter values is such that each regime has an approximately equal number of observations (50%). If we were to modify the magnitude of the slope and/or threshold parameters in such a way that one regime strongly dominates then it is natural to expect a deterioration in performance of the model selection criteria in small samples.

4 Conclusion

In this paper our objective was to provide a model selection based framework for estimating and conducting inferences in the context of multiple threshold models. We formally established that estimating the threshold parameters one at a time leads to T-consistent estimates of their true counterparts and subsequently investigated the asymptotic and finite sample properties of a se-

quentially implemented model selection based approach for the determination of the number of regimes.

Table 1: Empirical Mean and Standard Deviation of Estimators

DGP: $y_t = \beta_1 I(y_{t-1} \leq \gamma_1^0) + \beta_2 I(y_{t-1} > \gamma_1^0) + \epsilon_t$

$\beta_1 = 1,\ \beta_2 = 2,\ T = 250.$

γ_1^0	$\hat{\gamma}_1$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.75	0.758	1.003	2.005
	(0.142)	(0.184)	(0.071)
1.00	0.996	0.993	2.006
	(0.133)	(0.156)	(0.086)
1.50	1.491	0.995	2.004
	(0.096)	(0.105)	(0.105)
2.40	2.285	0.992	1.993
	(0.286)	(0.078)	(0.261)

Table 2a: Empirical Mean and Standard Deviation of Estimators

DGP: $y_t = \beta_1 I(y_{t-1} \leq \gamma_1^0) + \beta_2 I(\gamma_1^0 < y_{t-1} \leq \gamma_2^0) + \beta_3 I(y_{t-1} > \gamma_2^0) + \epsilon_t$

$\beta_1 = 1,\ \beta_2 = 2,\ \beta_3 = 3,\ T = 200.$

	Sequential Estimation				
(γ_1^0, γ_2^0)	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
(1,2)	1.244	2.230	1.183	2.211	3.002
	(0.364)	(0.539)	(0.379)	(0.445)	(0.136)
(1.5,2.5)	1.480	2.495	0.995	2.003	3.005
	(0.127)	(0.132)	(0.125)	(0.163)	(0.128)
(1,3)	0.996	2.973	0.998	2.001	3.002
	(0.138)	(0.146)	(0.174)	(0.104)	(0.175)
(2,3)	1.682	2.655	1.003	1.718	2.763
	(0.581)	(0.381)	(0.144)	(0.479)	(0.396)
	Joint Estimation				
(1,2)	1.247	2.213	1.186	2.106	3.003
	(0.360)	(0.512)	(0.372)	(0.434)	(0.143)
(1.5,2.5)	1.479	2.493	0.992	1.997	3.011
	(0.126)	(0.124)	(0.126)	(0.152)	(0.123)
(1,3)	0.993	2.972	0.986	1.997	3.006
	(0.144)	(0.169)	(0.175)	(0.116)	(0.181)
(2,3)	1.686	2.711	1.013	1.715	2.699
	(0.579)	(0.391)	(0.132)	(0.472)	(0.400)

Table 2b: Empirical Mean and Standard Deviation of Estimators

DGP: $y_t = \beta_1 y_{t-1} I(y_{t-1} \leq \gamma_1^0) + \beta_2 y_{t-1} I(\gamma_1^0 < y_{t-1} \leq \gamma_2^0) + \beta_3 y_{t-1} I(y_{t-1} > \gamma_2^0) + \epsilon_t$

$\beta_1 = 0.2$, $\beta_2 = 0.8$, $\beta_3 = -0.5$.

	Sequential Estimation				
$(\gamma_1^0 = -0.5, \gamma_2^0 = 0.5)$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
T=200	-0.582	0.483	0.179	1.028	-0.512
	(0.519)	(0.344)	(0.096)	(1.209)	(0.125)
T=400	-0.536	0.507	0.187	0.974	-0.509
	(0.418)	(0.207)	(0.066)	(0.683)	(0.086)
	Joint Estimation				
T=200	-0.511	0.415	0.180	1.250	-0.508
	(0.452)	(0.340)	(0.090)	(2.037)	(0.126)
T=400	-0.502	0.492	0.188	0.985	-0.501
	(0.409)	(0.205)	(0.071)	(0.675)	(0.082)

Table 3a: Correct Decision Frequencies: Linear Model

DGP: $y_t = \rho y_{t-1} + \epsilon_t$

	T=200				
ρ	BIC	AIC	HQ	BIC2	BIC3
0.5	88.4	44.9	70.9	99.3	99.8
0.7	87.8	45.1	69.7	98.8	99.8
0.9	85.6	41.2	66.3	98.7	99.9
1.0	50.0	9.7	24.9	89.8	98.7
	T=400				
0.5	92.4	44.9	72.5	99.2	100.0
0.7	91.1	44.7	71.9	99.5	100.0
0.9	90.9	42.0	70.7	99.4	100.0
1.0	56.9	9.7	27.4	93.2	99.7
	T=600				
0.5	93.5	45.7	74.1	99.7	100.0
0.7	92.1	45.8	74.4	99.7	100.0
0.9	91.9	42.6	73.5	99.6	100.0
1.0	60.1	9.8	29.2	94.2	99.7

Table 3b: Correct Decision Frequencies: Threshold Model

DGP: $y_t = \rho y_{t-1} I(y_{t-1} \leq 0) - \rho y_{t-1} I(y_{t-1} > 0) + \epsilon_t$

	T=200				
ρ	BIC	AIC	HQ	BIC2	BIC3
-0.40	100.0	100.0	100.0	99.3	95.1
-0.25	94.2	99.7	98.9	72.4	43.4
-0.15	63.3	91.3	80.4	25.2	7.4
-0.10	38.7	78.2	58.9	9.9	1.8
-0.05	19.0	61.5	38.2	2.7	0.2
	T=400				
-0.40	100.0	100.0	100.0	100.0	100.0
-0.25	99.9	100.0	100.0	97.0	84.2
-0.15	84.5	98.0	94.6	46.6	18.2
-0.10	53.2	88.6	75.2	16.5	3.5
-0.05	21.6	66.6	42.4	2.5	0.2
	T=600				
-0.40	100.0	100.0	100.0	100.0	100.0
-0.25	100.0	100.0	100.0	99.6	97.1
-0.15	93.5	99.6	98.8	60.5	36.3
-0.10	66.5	95.2	86.3	25.4	6.0
-0.05	22.7	72.1	46.9	2.8	0.2

Table 4a: Correct Decision Frequencies: Threshold Model ($m_0 = 1$)

Sequential Model Selection

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

	T=400			
	$\hat{m} = 0$	$\hat{m} = 1$	$\hat{m} = 2$	$\hat{m} \geq 3$
BIC	0.8	80.5	18.7	0.0
BIC2	1.3	91.1	7.6	0.0
BIC3	1.3	91.5	7.1	0.0
	T=600			
	$\hat{m} = 0$	$\hat{m} = 1$	$\hat{m} = 2$	$\hat{m} \geq 3$
BIC	0.1	90.0	10.0	0.0
BIC2	0.2	96.3	3.5	0.0
BIC3	0.4	96.4	3.3	0.0
	T=800			
	$\hat{m} = 0$	$\hat{m} = 1$	$\hat{m} = 2$	$\hat{m} \geq 3$
BIC	0.1	94.4	5.4	0.0
BIC2	0.1	98.4	1.5	0.0
BIC3	0.2	98.5	1.3	0.0

Table 4b: Correct Decision Frequencies: Threshold Model ($m_0 = 2$)

Sequential Model Selection

$$y_t = \begin{cases} 2.7 + 0.8y_{t-1} - 0.2y_{t-2} + \epsilon_t & y_{t-2} \leq 5 \\ 6 + 1.9y_{t-1} - 1.2y_{t-2} + \epsilon_t & 5 < y_{t-2} \leq 12 \\ 1 + 0.7y_{t-1} - 0.3y_{t-2} + \epsilon_t & y_{t-2} > 12 \end{cases}$$

	T=400			
	$\hat{m} \leq 1$	$\hat{m} = 2$	$\hat{m} = 3$	$\hat{m} \geq 4$
BIC	0.0	79.7	20.3	0.0
BIC2	0.0	98.1	1.9	0.0
BIC3	0.0	99.1	0.9	0.0
	T=600			
	$\hat{m} \leq 1$	$\hat{m} = 2$	$\hat{m} = 3$	$\hat{m} \geq 4$
BIC	0.0	85.4	14.6	0.0
BIC2	0.0	99.0	1.0	0.0
BIC3	0.0	99.3	0.7	0.0
	T=800			
	$\hat{m} \leq 1$	$\hat{m} = 2$	$\hat{m} = 3$	$\hat{m} \geq 4$
BIC	0.0	88.1	11.9	0.0
BIC2	0.0	99.0	1.0	0.0
BIC3	0.0	99.8	0.2	0.0

APPENDIX

We refer to the fact that a symmetric matrix $\mathbf{A}$ is positive (semi) definite by writing $\mathbf{A} \succ (\succeq) 0$. More specifically matrix $\mathbf{A}$ is said to be larger than another symmetric matrix $\mathbf{B}$ if $\mathbf{A} - \mathbf{B} \succeq 0$. Equivalently, $\mathbf{A} \succeq \mathbf{B} \Leftrightarrow \mathbf{A} - \mathbf{B} \succeq 0 \Leftrightarrow x'\mathbf{A}x \geq x'\mathbf{B}x$ together with $\mathbf{A} \succ \mathbf{B} \Leftrightarrow \mathbf{A} - \mathbf{B} \succ 0 \Leftrightarrow x'\mathbf{A}x > x'\mathbf{B}x$. By $\lambda^{min}(A)$ and $\lambda^{max}(A)$ we denote the smallest and largest eigenvalue of matrix A.

PROOF OF PROPOSITION 2.1: We reparameterize the true specification in (2) as

$$\mathbf{y} = \mathbf{W}\boldsymbol{\rho} + \mathbf{X}\boldsymbol{\beta}_{m+1} + \boldsymbol{\epsilon} \tag{A.1}$$

with $\mathbf{W} = [\mathbf{X}_{\gamma_1^0}, \mathbf{X}_{\gamma_2^0}, \ldots, \mathbf{X}_{\gamma_m^0}]$, $\boldsymbol{\rho} = (\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \ldots, \boldsymbol{\rho}_m)'$, $\mathbf{X}_{\gamma_i^0} = \mathbf{X} * \mathbf{I}(z \leq \gamma_i^0)$ and $\boldsymbol{\rho}_i = \boldsymbol{\beta}_i - \boldsymbol{\beta}_{i+1}$, $\forall i = 1, \ldots, m$. Next, defining $\mathbf{M} = I - \sum_{i=1}^{m+1} \mathbf{P}_i$ with $\mathbf{P}_i = \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'$, the concentrated sum of squared errors function (3) obtained from the fitted model (2) can be written as $S_T(\gamma_1, \ldots, \gamma_m) = \mathbf{y}'\mathbf{M}\mathbf{y}$. Using (A.1) and noting that $\mathbf{M}\mathbf{X} = \mathbf{0}$ we reformulate $S_T(\gamma_1, \ldots, \gamma_m)$ as

$$S_T(\gamma_1, \ldots, \gamma_m) = \boldsymbol{\rho}'\mathbf{W}'\mathbf{M}\mathbf{W}\boldsymbol{\rho} + \boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon} + 2\boldsymbol{\rho}'\mathbf{W}'\mathbf{M}\boldsymbol{\epsilon}. \tag{A.2}$$

From assumptions **A1-A2**(i)-(ii) we have $\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}/T = \boldsymbol{\epsilon}'\boldsymbol{\epsilon}/T + o_p(1)$ and $\boldsymbol{\rho}'\mathbf{W}'\mathbf{M}\boldsymbol{\epsilon}/T = o_p(1)$ uniformly over $\gamma_i \in \Gamma_m$, leading to

$$\frac{S_T(\gamma_1, \ldots, \gamma_m)}{T} - \frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{T} = \frac{\boldsymbol{\rho}'\mathbf{W}'\mathbf{M}\mathbf{W}\boldsymbol{\rho}}{T} + o_p(1). \tag{A.3}$$

Letting $R_T(\gamma_1, \ldots, \gamma_m) = S_T(\gamma_1, \ldots, \gamma_m) - S_T(\gamma_1^0, \ldots, \gamma_m^0)$ and using

$$\frac{S_T(\gamma_1^0, \ldots, \gamma_m^0)}{T} = \frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{T} + o_p(1), \tag{A.4}$$

we write

$$\frac{R_T(\gamma_1, \ldots, \gamma_m)}{T} = \frac{\boldsymbol{\rho}'\mathbf{W}'\mathbf{M}\mathbf{W}\boldsymbol{\rho}}{T} + o_p(1). \tag{A.5}$$

Next, matrix $\mathbf{M}$ is symmetric idempotent and therefore positive semi-definite, $\mathbf{W}$ having full column rank $\mathbf{W}\boldsymbol{\rho} \neq 0$ and $(\mathbf{W}\boldsymbol{\rho})'\mathbf{M}(\mathbf{W}\boldsymbol{\rho}) \succeq 0$. We next write $\boldsymbol{\rho}' R_\infty(\gamma_1, \ldots, \gamma_m)\boldsymbol{\rho}$ for the nonstochastic continuous uniform probability limit of the left hand side of (A.5) and establish that it reaches its minimum value of zero uniquely if and only if $\gamma_i = \gamma_i^0$ $\forall i = 1, \ldots, m$. Letting $R_\infty^{\ell\ell}(\gamma_1, \ldots, \gamma_m)$ denote the diagonal components of $R_\infty(\gamma_1, \ldots, \gamma_m)$ and using assumption **A2**(i) we obtain

$$\begin{aligned} R_\infty^{\ell\ell}(\gamma_1, \ldots, \gamma_m) = \mathbf{G}(\gamma_\ell^0) - \sum_{i=1}^{m+1} &[\mathbf{G}(\gamma_\ell^0 \wedge \gamma_i) - \mathbf{G}(\gamma_\ell^0 \wedge \gamma_{i-1})][\mathbf{G}(\gamma_i) - \mathbf{G}(\gamma_{i-1})]^{-1} \\ &[\mathbf{G}(\gamma_\ell^0 \wedge \gamma_i) - \mathbf{G}(\gamma_\ell^0 \wedge \gamma_{i-1})] \end{aligned} \tag{A.6}$$

for $\ell = 1, \ldots, m$ and

$$\begin{aligned} R_\infty^{\ell k}(\gamma_1, \ldots, \gamma_m) = \mathbf{G}(\gamma_\ell^0 \wedge \gamma_k^0) - \sum_{i=1}^{m+1} &[\mathbf{G}(\gamma_\ell^0 \wedge \gamma_i) - \mathbf{G}(\gamma_\ell^0 \wedge \gamma_{i-1})][\mathbf{G}(\gamma_i) - \mathbf{G}(\gamma_{i-1})]^{-1} \\ &[\mathbf{G}(\gamma_k^0 \wedge \gamma_i) - \mathbf{G}(\gamma_k^0 \wedge \gamma_{i-1})] \end{aligned} \tag{A.7}$$

for $\ell \neq k$, $k = 1, \ldots, m$ and where $\mathbf{G}(\gamma_i^0 \wedge \gamma_0) \equiv \mathbf{0}$, $\mathbf{G}(\gamma_i^0 \wedge \gamma_{m+1}) \equiv \mathbf{G}(\gamma_i^0)$ and $\mathbf{G}(\gamma_{m+1}) \equiv \mathbf{G}$. As a direct consequence of (A.6) and (A.7) we have $R_\infty^{\ell\ell}(\gamma_1, \ldots, \gamma_m) \succeq R_\infty^{\ell k}(\gamma_1, \ldots, \gamma_m)$ $\forall \ell \neq k$ and

$$
\begin{aligned}
R_\infty^{\ell\ell}(\gamma_1, \gamma_2, \ldots, \gamma_{\ell-1}, \gamma_\ell^0, \gamma_{\ell+1}, \ldots, \gamma_m) &= 0, \\
R_\infty^{\ell\ell}(\gamma_1^0, \gamma_2^0, \ldots, \gamma_{\ell-1}^0, \gamma_\ell, \gamma_{\ell+1}^0, \ldots, \gamma_m^0) &\succ 0 \quad \forall \gamma_\ell \neq \gamma_\ell^0,
\end{aligned} \tag{A.8}
$$

implying that $R_\infty(\gamma_1, \ldots, \gamma_m) = 0$ iff $\gamma_i = \gamma_i^0$ $\forall i = 1, \ldots, m$. Since $(\hat{\gamma}_1, \ldots, \hat{\gamma}_m) = \arg\min[R_T(\gamma_1, \ldots, \gamma_m)/T]$ and (A.5) converges uniformly in probability to the nonstochastic continuous functional $\boldsymbol{\rho}' R_\infty(\gamma_1, \ldots, \gamma_m)\boldsymbol{\rho}$ that is uniquely minimized at $(\gamma_1^0, \ldots, \gamma_m^0)$ it follows from Theorem 2.1 in Newey and McFadden (1994) that $(\hat{\gamma}_1, \ldots, \hat{\gamma}_m) \xrightarrow{p} (\gamma_1^0, \ldots, \gamma_m^0)$.

PROOF OF LEMMA 2.1: Using (8) we write

$$
\frac{J_T(r)}{T} = (\hat{\boldsymbol{\delta}}_2 - \hat{\boldsymbol{\delta}}_1)' \left(\frac{\mathbf{Z}_2'\mathbf{Z}_2}{T}\right) \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1} \left(\frac{\mathbf{Z}_1'\mathbf{Z}_1}{T}\right) (\hat{\boldsymbol{\delta}}_2 - \hat{\boldsymbol{\delta}}_1).
$$

Noting that $\mathbf{Z}_2'\mathbf{X} = \mathbf{Z}_2'\mathbf{Z}_2$, $\mathbf{Z}_1'\mathbf{X} = \mathbf{Z}_1'\mathbf{Z}_1$ together with $(\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\boldsymbol{\epsilon} = o_p(1)$ for $i = 1, 2$ which follows from assumptions **A2**(i)-(ii) we can express $\hat{\boldsymbol{\delta}}_2 - \hat{\boldsymbol{\delta}}_1$ obtained from (5) as

$$
\begin{aligned}
\hat{\boldsymbol{\delta}}_2 - \hat{\boldsymbol{\delta}}_1 &= \left[\frac{\mathbf{Z}_2'\mathbf{Z}_2}{T}\right]^{-1} \left[\frac{\mathbf{Z}_2'\mathbf{W}}{T}\right] \boldsymbol{\rho} - \left[\frac{\mathbf{Z}_1'\mathbf{Z}_1}{T}\right]^{-1} \left[\frac{\mathbf{Z}_1'\mathbf{W}}{T}\right] \boldsymbol{\rho} + o_p(1) \\
&= \left[\frac{\mathbf{Z}_2'\mathbf{Z}_2}{T}\right]^{-1} \sum_{\ell=1}^{m} \frac{\mathbf{Z}_2'\mathbf{X}_{\gamma_\ell^0}}{T} \boldsymbol{\rho}_i - \left[\frac{\mathbf{Z}_1'\mathbf{Z}_1}{T}\right]^{-1} \sum_{\ell=1}^{m} \frac{\mathbf{Z}_1'\mathbf{X}_{\gamma_\ell^0}}{T} \boldsymbol{\rho}_i + o_p(1),
\end{aligned} \tag{A.9}
$$

where we used the same parameterization for $\mathbf{y}$ as in the proof of Proposition 2.1. Next, from assumption **A2**(i) we have

$$
\sup_r \left| \left(\frac{\mathbf{Z}_1'\mathbf{Z}_1}{T}\right)^{-1} \sum_{\ell=1}^{m} \frac{\mathbf{Z}_1'\mathbf{X}_{\gamma_\ell^0}}{T} \boldsymbol{\rho}_\ell - \mathbf{G}(r)^{-1} \sum_{\ell=1}^{m} \mathbf{G}(\gamma_\ell^0 \wedge r)\boldsymbol{\rho}_\ell \right| \xrightarrow{p} 0, \tag{A.10}
$$

and

$$
\sup_r \left| \left(\frac{\mathbf{Z}_2'\mathbf{Z}_2}{T}\right)^{-1} \sum_{\ell=1}^{m} \frac{\mathbf{Z}_2'\mathbf{X}_{\gamma_\ell^0}}{T} - (\mathbf{G} - \mathbf{G}(r))^{-1} \sum_{\ell=1}^{m} (\mathbf{G}(\gamma_\ell^0) - \mathbf{G}(\gamma_\ell^0 \wedge r))\boldsymbol{\rho}_\ell \right| \xrightarrow{p} 0, \tag{A.11}
$$

which together with

$$
\sup_r \left| \frac{\mathbf{Z}_2'\mathbf{Z}_2}{T} \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1} \frac{\mathbf{Z}_1'\mathbf{Z}_1}{T} - (\mathbf{G} - \mathbf{G}(r))\mathbf{G}^{-1}\mathbf{G}(r)^{-1} \right| \xrightarrow{p} 0 \tag{A.12}
$$

lead to the desired result in (10).

LEMMA A.1: For $r \in (\gamma_1^0, \gamma_2^0)$ and letting

$$
\begin{aligned}
K_1 &= \mathbf{G}(\gamma_1^0)\mathbf{G}(r)^{-1}(\mathbf{G}(r) - \mathbf{G}(\gamma_1^0)) \\
K_2 &= \mathbf{G}(\gamma_2^0)\mathbf{G}(r)^{-1}(\mathbf{G}(r) - \mathbf{G}(\gamma_1^0))(\mathbf{G} - \mathbf{G}(\gamma_1^0))^{-1}(\mathbf{G} - \mathbf{G}(\gamma_2^0)) \\
M_1 &= \mathbf{G}(\gamma_1^0)\mathbf{G}(\gamma_2^0)^{-1}(\mathbf{G}(\gamma_2^0) - \mathbf{G}(\gamma_1^0)) \\
M_2 &= (\mathbf{G}(\gamma_2^0) - \mathbf{G}(\gamma_1^0))(\mathbf{G} - \mathbf{G}(\gamma_1^0))^{-1}(\mathbf{G} - \mathbf{G}(\gamma_2^0)),
\end{aligned}
$$

we have

(i) $M_1 \succ K_1$ and $M_2 \succ K_2$

(ii) $\forall x \neq 0,\ 0 < \frac{x'K_1x}{x'M_1x} < 1$

(iii) $\forall x \neq 0,\ z \neq 0$ and $x \neq z,\ \left[\frac{x'M_1x}{x'K_1x}\right]\left[\frac{z'K_2z}{z'M_2z}\right] \leq 1.$

PROOF OF LEMMA A.1: (i) From assumption **A2**(i), $\mathbf{G}(r)$ is a continuous strictly increasing function of r. Since $r \in (\gamma_1^0, \gamma_2^0)$ it follows that $\mathbf{G}(\gamma_2^0) \succ \mathbf{G}(r)$ directly implying that $M_1 \succ K_1$. The result $M_2 \succ K_2$ follows using the same argument. (ii) We use that fact that $\lambda^{min}(M_1^{-1}K_1) \leq (x'K_1x/x'M_1x) \leq \lambda^{max}(M_1^{-1}K_1)\ \forall x \neq 0$. Since $K_1 \prec M_1$ we have $K_1M_1^{-1} \prec I$ and therefore $\lambda^{max}(M_1^{-1}K_1) < 1$ which together with $\lambda^{min}(M_1^{-1}K_1) > 0$ implies the desired result. (iii) First note that $K_2M_2^{-1} = (\mathbf{G}(\gamma_2^0) - \mathbf{G}(\gamma_1^0))M_1^{-1}K_1(\mathbf{G}(\gamma_2^0) - \mathbf{G}(\gamma_1^0))^{-1}$, implying that $K_2M_2^{-1}$ and $M_1^{-1}K_1$ have the same characteristic roots. Next we have

$$\frac{z'K_2z}{z'M_2z} \leq \lambda^{max}(M_2^{-1}K_2)$$

and

$$\frac{x'M_1x}{x'K_1x} \leq \lambda^{max}(K_1^{-1}M_1) = \lambda^{max}((M_1^{-1}K_1)^{-1}) = [\lambda^{max}(M_1^{-1}K_1)]^{-1}$$

which implies the desired result.

PROOF OF LEMMA 2.2: With no loss of generality we provide the proof assuming $m = 2$ and setting $\gamma_{(1)}^0 = \gamma_1^0$ in the context of the requirements of assumption **A3**. The proof is in three parts. Since $J_\infty(r)$ takes different expressions over the three regions given by $(\underline{\gamma}, \gamma_1^0)$, (γ_1^0, γ_2^0) and $(\gamma_2^0, \bar{\gamma})$, the result will follow by showing that the maximum of $J_\infty(r)$ cannot occur in any of the three regions in the sense that $J_\infty(\gamma_1^0) > J_\infty(r)$, $J_\infty(\gamma_2^0) > J_\infty(r)$, and the requirement that $J_\infty(\gamma_1^0) > J_\infty(\gamma_2^0)$. We start by treating the case $r \in (\gamma_1^0, \gamma_2^0)$. Using the expression of $J_\infty(r)$ in (10)-(11) and setting $m = 2$ we have

$$\begin{aligned} J_\infty(\gamma_1^0) - J_\infty(r) &= \boldsymbol{\rho}_1'\mathbf{G}(\gamma_1^0)\mathbf{G}(r)^{-1}(\mathbf{G}(r) - \mathbf{G}(\gamma_1^0))\boldsymbol{\rho}_1 \\ &- \boldsymbol{\rho}_2'(\mathbf{G} - \mathbf{G}(\gamma_2^0))(\mathbf{G} - \mathbf{G}(r))^{-1}(\mathbf{G}(r) - \mathbf{G}(\gamma_1^0)) \\ &\quad (\mathbf{G} - \mathbf{G}(\gamma_1^0))^{-1}(\mathbf{G} - \mathbf{G}(\gamma_2^0))\boldsymbol{\rho}_2. \end{aligned} \tag{A.13}$$

Next, using the expressions of K_1 and K_2 defined in Lemma A.1 we can rewrite (A.13) as

$$\begin{aligned} J_\infty(\gamma_1^0) - J_\infty(r) &= \boldsymbol{\rho}_1'K_1\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2'(\mathbf{G} - \mathbf{G}(\gamma_2^0))(\mathbf{G} - \mathbf{G}(r))^{-1}\mathbf{G}(r)\mathbf{G}(\gamma_2^0)^{-1}K_2\boldsymbol{\rho}_2 \\ &= \boldsymbol{\rho}_1'K_1\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2'K_2\boldsymbol{\rho}_2 + \\ &\quad \boldsymbol{\rho}_2'[I - (\mathbf{G} - \mathbf{G}(\gamma_2^0))(\mathbf{G} - \mathbf{G}(r))^{-1}\mathbf{G}(r)\mathbf{G}(\gamma_2^0)^{-1}]K_2\boldsymbol{\rho}_2 \end{aligned}$$

and observing that $\mathbf{G}(\gamma_2^0)\mathbf{G}(r)^{-1} \prec (\mathbf{G} - \mathbf{G}(\gamma_2^0))(\mathbf{G} - \mathbf{G}(r))^{-1}$ since $r < \gamma_2^0$, we have

$$J_\infty(\gamma_1^0) - J_\infty(r) > \boldsymbol{\rho}_1'K_1\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2'K_2\boldsymbol{\rho}_2$$

$$= \frac{\boldsymbol{\rho}_1' K_1 \boldsymbol{\rho}_1}{\boldsymbol{\rho}_1' M_1 \boldsymbol{\rho}_1} \left[\boldsymbol{\rho}_1' M_1 \boldsymbol{\rho}_1 - \frac{\boldsymbol{\rho}_1' M_1 \boldsymbol{\rho}_1}{\boldsymbol{\rho}_1' K_1 \boldsymbol{\rho}_1} \frac{\boldsymbol{\rho}_2' K_2 \boldsymbol{\rho}_2}{\boldsymbol{\rho}_2' M_2 \boldsymbol{\rho}_2} \boldsymbol{\rho}_2' M_2 \boldsymbol{\rho}_2 \right]$$
$$> \frac{\boldsymbol{\rho}_1' K_1 \boldsymbol{\rho}_1}{\boldsymbol{\rho}_1' M_1 \boldsymbol{\rho}_1} \left[J_\infty(\gamma_1^0) - J_\infty(\gamma_2^0) \right] > 0,$$

where the last inequality follows from Lemma A.1 (iii) and the fact that $J_\infty(\gamma_1^0) - J_\infty(\gamma_2^0) = \boldsymbol{\rho}_1' M_1 \boldsymbol{\rho}_1 - \boldsymbol{\rho}_2' M_1 \boldsymbol{\rho}_2$ as established in (13). Thus, the maximum of $J_\infty(r)$ cannot occur in (γ_1^0, γ_2^0). We next concentrate on the case $r < \gamma_1^0$. Using (10)-(11) and standard algebra we can write

$$J_\infty(\gamma_1^0) - J_\infty(r) = \mathbf{w}'(\mathbf{G} - \mathbf{G}(\gamma_1^0))^{-1}(\mathbf{G}(\gamma_1^0) - \mathbf{G}(r))(\mathbf{G} - \mathbf{G}(r))^{-1}\mathbf{w} \tag{A.14}$$

with $\mathbf{w} = [(\mathbf{G} - \mathbf{G}(\gamma_1^0))\boldsymbol{\rho}_1 + (\mathbf{G} - \mathbf{G}(\gamma_2^0))\boldsymbol{\rho}_2]$. Next note that $\mathbf{w} = 0$ implies $J_\infty(\gamma_1^0) < J_\infty(\gamma_2^0)$ which is ruled out by assumption, thus $\mathbf{w} \neq 0$ and therefore the above quadratic form is strictly positive, implying that the maximum of $J_\infty(r)$ cannot occur for $r < \gamma_1^0$. The treatment of the case $r > \gamma_2^0$ is identical.

PROOF OF PROPOSITION 2.2: The result follows from Lemmas 2.1, 2.2 and using Theorem 2.1 in Newey and McFadden (1994).

PROOF OF PROPOSITION 2.3: We proceed using the same simplifications as in the proof of Lemma 2.2, setting $m = 2$ and $\gamma_{(1)}^0 = \gamma_1^0$ with the true model given by $\mathbf{y} = \mathbf{X}_1^0\beta_1 + \mathbf{X}_2^0\beta_2 + \mathbf{X}_3^0\beta_3 + \boldsymbol{\epsilon}$. To establish the T-consistency of $\hat{r}$ it suffices to show that $S_T(r) - S_T(\gamma_1^0) > 0$ for $T|r - \gamma_1^0|$ sufficiently large (see Chan (1993)). From Proposition 2.2 we operate in a small neighborhood of γ_1^0 and treat the case $r < \gamma_1^0$. Formally we establish that for every $\nu > 0$, there exists an $0 < M < \infty$ such that for all T large we have

$$P\left[\min_{\frac{M}{T} < (\gamma_1^0 - r)} S_T(r) - S_T(\gamma_1^0) \leq 0 \right] < \nu. \tag{A.15}$$

We initially write $S_T(r) - S_T(\gamma_1^0) = (S_T(r) - S_T(r, \gamma_1^0)) - (S_T(\gamma_1^0) - S_T(r, \gamma_1^0))$ where $S_T(r, \gamma_1^0)$ denotes the concentrated sum of squared errors function from the following auxiliary specification

$$\mathbf{y} = \mathbf{Z}_1\boldsymbol{\alpha}_1 + \mathbf{X}_{r\gamma_1^0}\boldsymbol{\alpha}_2 + \overline{\mathbf{X}}_{\gamma_1^0}\boldsymbol{\alpha}_3 + \mathbf{u} \tag{A.16}$$

with $\mathbf{X}_{r\gamma_1^0} = \mathbf{X} * I(r < z \leq \gamma_1^0)$ and $\overline{\mathbf{X}}_{\gamma_1^0} = \mathbf{X} * I(z > \gamma_1^0)$. Here $S_T(r) - S_T(r, \gamma_1^0)$ corresponds to the difference in the sum of squared errors obtained from model (A.16) on which the restriction $\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_3$ has been imposed and the unrestricted counterpart. Making use of $\mathbf{X}_{r\gamma_1^0} + \overline{\mathbf{X}}_{\gamma_1^0} = \mathbf{Z}_2$ and standard algebra gives

$$S_T(r) - S_T(r, \gamma_1^0) = (\hat{\boldsymbol{\alpha}}_3 - \hat{\boldsymbol{\alpha}}_2)'\mathbf{X}_{r\gamma_1^0}'\mathbf{X}_{r\gamma_1^0}(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\overline{\mathbf{X}}_{\gamma_1^0}'\overline{\mathbf{X}}_{\gamma_1^0}(\hat{\boldsymbol{\alpha}}_3 - \hat{\boldsymbol{\alpha}}_2). \tag{A.17}$$

Similarly, $S_T(\gamma_1^0) - S_T(r, \gamma_1^0)$ corresponds to the difference in the sum of squared errors from (A.16) on which the restriction $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2$ has been imposed and the unrestricted counterpart and since $\mathbf{Z}_1 + \mathbf{X}_{r\gamma_1^0} = \mathbf{X}_1^0$ we also have

$$S_T(\gamma_1^0) - S_T(r, \gamma_1^0) = (\hat{\boldsymbol{\alpha}}_2 - \hat{\boldsymbol{\alpha}}_1)'\mathbf{Z}_1'\mathbf{Z}_1(\mathbf{X}_1^{0\prime}\mathbf{X}_1^0)^{-1}\mathbf{X}_{r\gamma_1^0}'\mathbf{X}_{r\gamma_1^0}(\hat{\boldsymbol{\alpha}}_2 - \hat{\boldsymbol{\alpha}}_1). \tag{A.18}$$

Note that by assumption **A1**(i) $\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}$ is positive definite for large T implying that $\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\overline{\mathbf{X}}'_{\gamma_1^0}\overline{\mathbf{X}}_{\gamma_1^0} \equiv [(\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0})^{-1}+(\overline{\mathbf{X}}'_{\gamma_1^0}\overline{\mathbf{X}}_{\gamma_1^0})^{-1}]^{-1}$ and $\mathbf{Z}_1'\mathbf{Z}_1(\mathbf{X}_1^{0\prime}\mathbf{X}_1^0)^{-1}\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0} \equiv [(\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0})^{-1}+(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}]^{-1}$ are positive definite. Using (A.17) and (A.18) we write

$$\begin{aligned}\frac{S_T(r)-S_T(\gamma_1^0)}{T(\gamma_1^0-r)} &= (\hat{\boldsymbol{\alpha}}_3-\hat{\boldsymbol{\alpha}}_2)'\left[\frac{\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\overline{\mathbf{X}}'_{\gamma_1^0}\overline{\mathbf{X}}_{\gamma_1^0}}{T(\gamma_1^0-r)}\right](\hat{\boldsymbol{\alpha}}_3-\hat{\boldsymbol{\alpha}}_2) - \\ &\quad (\hat{\boldsymbol{\alpha}}_2-\hat{\boldsymbol{\alpha}}_1)'\left[\frac{\mathbf{Z}_1'\mathbf{Z}_1(\mathbf{X}_1^{0\prime}\mathbf{X}_1^0)^{-1}\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}}{T(\gamma_1^0-r)}\right](\hat{\boldsymbol{\alpha}}_2-\hat{\boldsymbol{\alpha}}_1). \qquad \text{(A.19)}\end{aligned}$$

Since $(\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}) \succ [(\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0})^{-1}+(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}]^{-1}$ and noting that $\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\overline{\mathbf{X}}'_{\gamma_1^0}\overline{\mathbf{X}}_{\gamma_1^0} = \mathbf{X}_{r\gamma_1^0}(\mathbf{I}-\mathbf{P}_{Z_2})\mathbf{X}_{r\gamma_1^0}$ which follows from $\overline{\mathbf{X}}_{\gamma_1^0} = \mathbf{Z}_2 - \mathbf{X}_{r\gamma_1^0}$ and $\mathbf{X}'_{r\gamma_1^0}\mathbf{Z}_2 = \mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}$ when $r<\gamma_1^0$ we have

$$\begin{aligned}\frac{S_T(r)-S_T(\gamma_1^0)}{T(\gamma_1^0-r)} &\geq (\hat{\boldsymbol{\alpha}}_3-\hat{\boldsymbol{\alpha}}_2)'\left[\frac{\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}}{T(\gamma_1^0-r)}\right](\hat{\boldsymbol{\alpha}}_3-\hat{\boldsymbol{\alpha}}_2) - (\hat{\boldsymbol{\alpha}}_3-\hat{\boldsymbol{\alpha}}_2)'\left[\frac{\mathbf{X}'_{r\gamma_1^0}\mathbf{P}_{Z_2}\mathbf{X}_{r\gamma_1^0}}{T(\gamma_1^0-r)}\right](\hat{\boldsymbol{\alpha}}_3-\hat{\boldsymbol{\alpha}}_2) - \\ &\quad (\hat{\boldsymbol{\alpha}}_2-\hat{\boldsymbol{\alpha}}_1)'\left[\frac{\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}}{T(\gamma_1^0-r)}\right](\hat{\boldsymbol{\alpha}}_2-\hat{\boldsymbol{\alpha}}_1). \qquad \text{(A.20)}\end{aligned}$$

From (A.16) and noting the orthogonality of the relevant regressors we write

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_1 &= \boldsymbol{\beta}_1+\left(\frac{\mathbf{Z}_1'\mathbf{Z}_1}{T}\right)^{-1}\frac{\mathbf{Z}_1'\boldsymbol{\epsilon}}{T}, \\ \hat{\boldsymbol{\alpha}}_2 &= \boldsymbol{\beta}_1+\left(\frac{\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}}{T}\right)^{-1}\frac{\mathbf{X}'_{r\gamma_1^0}\boldsymbol{\epsilon}}{T}, \\ \hat{\boldsymbol{\alpha}}_3 &= \left(\frac{\overline{\mathbf{X}}'_{\gamma_1^0}\overline{\mathbf{X}}_{\gamma_1^0}}{T}\right)^{-1}\left(\frac{\mathbf{X}_2^{0\prime}\mathbf{X}_2^0}{T}\boldsymbol{\beta}_2+\frac{\mathbf{X}_3^{0\prime}\mathbf{X}_3^0}{T}\boldsymbol{\beta}_3+\frac{\overline{\mathbf{X}}'_{\gamma_1^0}\boldsymbol{\epsilon}}{T}\right)\end{aligned}$$

and using assumptions **A2**(i)-(ii) it follows that

$$\hat{\boldsymbol{\alpha}}_2-\hat{\boldsymbol{\alpha}}_1 = o_p(1), \qquad \text{(A.21)}$$

and

$$\hat{\boldsymbol{\alpha}}_3-\hat{\boldsymbol{\alpha}}_2 = (\boldsymbol{\beta}_2-\boldsymbol{\beta}_1)+(\mathbf{G}-\mathbf{G}(\gamma_1^0))^{-1}(\mathbf{G}-\mathbf{G}(\gamma_2^0))(\boldsymbol{\beta}_3-\boldsymbol{\beta}_2)+o_p(1) \qquad \text{(A.22)}$$

uniformly over $r<\gamma_1^0$. From (A.21) and since under our assumptions $\left\|\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}/T(\gamma_1^0-r)\right\| = O_p(1)$, the third term in the right hand side of (A.20) can be made arbitrarily small. Similarly since we are operating with r in a small neighborhood of γ_1^0, the second term on the right hand side of (A.20) can also be made arbitrarily small. This follows by writing

$$\frac{\mathbf{X}'_{r\gamma_1^0}\mathbf{P}_{Z_2}\mathbf{X}_{r\gamma_1^0}}{T(\gamma_1^0-r)} = \left(\frac{\mathbf{X}'_{r\gamma_1^0}\mathbf{Z}_2}{T(\gamma_1^0-r)}\right)\left(\frac{\mathbf{Z}_2'\mathbf{Z}_2}{T}\right)^{-1}\left(\frac{\mathbf{Z}_2'\mathbf{X}_{r\gamma_1^0}}{T(\gamma_1^0-r)}\right)(\gamma_1^0-r),$$

from which we have

$$\left\|\frac{\mathbf{X}'_{r\gamma_1^0}\mathbf{P}_{Z_2}\mathbf{X}_{r\gamma_1^0}}{T(\gamma_1^0-r)}\right\| \leq \left\|\frac{\mathbf{X}'_{r\gamma_1^0}\mathbf{Z}_2}{T(\gamma_1^0-r)}\left(\frac{\mathbf{Z}_2'\mathbf{Z}_2}{T}\right)^{-1}\right\|\left\|\frac{\mathbf{Z}_2'\mathbf{X}_{r\gamma_1^0}}{T(\gamma_1^0-r)}\right\|(\gamma_1^0-r).$$

Finally for the first term on the right hand side of (A.20) we write

$$(\hat{\boldsymbol{\alpha}}_3 - \hat{\boldsymbol{\alpha}}_2)' \left[\frac{\mathbf{X}'_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}}{T(\gamma_1^0 - r)}\right](\hat{\boldsymbol{\alpha}}_3 - \hat{\boldsymbol{\alpha}}_2) \quad \geq \quad \lambda^{min}\left[\frac{\mathbf{X}_{r\gamma_1^0}\mathbf{X}_{r\gamma_1^0}}{T(\gamma_1^0 - r)}\right]\|\hat{\boldsymbol{\alpha}}_3 - \hat{\boldsymbol{\alpha}}_2\|^2. \tag{A.23}$$

From (A.22), $\|\hat{\boldsymbol{\alpha}}_3 - \hat{\boldsymbol{\alpha}}_2\|$ converges to a non-zero limit, thus $\|\hat{\boldsymbol{\alpha}}_3 - \hat{\boldsymbol{\alpha}}_2\|^2$ is no less than $C\|(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1) + (\mathbf{G} - \mathbf{G}(\gamma_1^0))^{-1}(\mathbf{G} - \mathbf{G}(\gamma_2^0))(\boldsymbol{\beta}_3 - \boldsymbol{\beta}_2)\|^2$ for some positive constant C and T sufficiently large. Since by assumption **A1**(i) the minimum eigenvalue of the normalized moment matrix taken in the neighborhood of γ_1^0 is also *strictly* positive for T sufficiently large it follows that $S_T(r) - S_T(\gamma_1^0) > 0$ on the relevant set, thus establishing the required result.

LEMMA A.2: Consider the objective function in (16) but conditioned on $h-1$ true threshold parameters, say $J_T(r|\gamma_{s_1}^0, \ldots, \gamma_{s_{h-1}}^0)$, with $(\gamma_{s_1}^0, \gamma_{s_2}^0, \ldots, \gamma_{s_{h-1}}^0)$ denoting a configuration of $h-1$ true threshold parameters ranked in ascending but not necessarily consecutive order. Also, let $J_\infty^{a:b}(r)$ denote a truncated version of $J_\infty(r)$ in (10) with $\ell = a, \ldots, b$ and the convention $J_\infty^{a:b}(r) \equiv 0$ for $a < b$, then as $T \to \infty$ and under assumptions **A1-A2**(i)-(ii) we have

$$\sup_r \left| \frac{J_T(r|\gamma_{s_1}^0, \ldots, \gamma_{s_{h-1}}^0)}{T} - J_\infty(r|\gamma_{s_1}^0, \ldots, \gamma_{s_{h-1}}^0) \right| \quad \xrightarrow{p} \quad 0 \tag{A.24}$$

where $J_\infty(r|\gamma_{s_1}^0, \ldots, \gamma_{s_{h-1}}^0)$ is a nonstochastic continuous function given by

$$J_\infty(r|\gamma_{s_1}^0, \ldots, \gamma_{s_{h-1}}^0) \quad = \quad \sum_{\ell=1}^{h} J_\infty^{s_{\ell-1}+1:s_\ell-1}(r) I(\gamma_{s_{\ell-1}}^0 < r < \gamma_{s_\ell}^0) \tag{A.25}$$

with the notational conventions $\gamma_{s_0}^0 \equiv \underline{\gamma}$, $\gamma_{s_h}^0 \equiv \bar{\gamma}$, $s_0 \equiv 0$ and $s_h - 1 \equiv m$.

PROOF OF LEMMA A.2: Here we have $\mathbf{Z}_{1,\ell} = \mathbf{X} * \mathbf{I}(\gamma_{s_{\ell-1}}^0 < z \leq r)$ and $\mathbf{Z}_{2,\ell} = \mathbf{X} * \mathbf{I}(r < z \leq \gamma_{s_\ell})$ and write $\widehat{\mathbf{Z}}_\ell = \mathbf{Z}_{1,\ell} + \mathbf{Z}_{2,\ell} = \mathbf{X} * \mathbf{I}(\gamma_{s_{\ell-1}}^0 \leq z \leq \gamma_{s_\ell}^0)\ \forall \ell = 1, \ldots, h$. Since $\widehat{\mathbf{Z}}_i'\mathbf{Z}_{1,\ell} = \mathbf{0}$ and $\widehat{\mathbf{Z}}_i'\mathbf{Z}_{2,\ell} = \mathbf{0}\ \forall i \neq \ell$ we have $\mathbf{Q}_\ell\mathbf{Z}_{1,\ell} = \mathbf{Z}_{1,\ell}$ and $\mathbf{Q}_\ell\mathbf{Z}_{2,\ell} = \mathbf{Z}_{2,\ell}$ and from (17) we can write $\hat{\boldsymbol{\delta}}_{2,\ell} - \hat{\boldsymbol{\delta}}_{1,\ell} = (\mathbf{Z}_{2,\ell}'\mathbf{Z}_{2,\ell})^{-1}\mathbf{Z}_{2,\ell}'\mathbf{y} - (\mathbf{Z}_{1,\ell}'\mathbf{Z}_{1,\ell})^{-1}\mathbf{Z}_{1,\ell}'\mathbf{y}$. Using the same parameterization of the true specification for $\mathbf{y}$ given in (A.1) together with assumptions **A2**(i)-(ii) we can also write

$$\begin{aligned}
\hat{\boldsymbol{\delta}}_{2,\ell} - \hat{\boldsymbol{\delta}}_{1,\ell} \quad &= \quad (\mathbf{Z}_{2,\ell}'\mathbf{Z}_{2,\ell})^{-1}\mathbf{Z}_{2,\ell}'\mathbf{W}\boldsymbol{\rho} - (\mathbf{Z}_{1,\ell}'\mathbf{Z}_{1,\ell})^{-1}\mathbf{Z}_{1,\ell}'\mathbf{W}\boldsymbol{\rho} + o_p(1) \\
&= \quad (\mathbf{Z}_{2,\ell}'\mathbf{Z}_{2,\ell})^{-1}\sum_{i=1}^{m}\mathbf{Z}_{2,\ell}'\mathbf{X}_{\gamma_i^0}\boldsymbol{\rho}_i - (\mathbf{Z}_{1,\ell}'\mathbf{Z}_{1,\ell})^{-1}\sum_{i=1}^{m}\mathbf{Z}_{1,\ell}'\mathbf{X}_{\gamma_i^0}\boldsymbol{\rho}_i + o_p(1).
\end{aligned} \tag{A.26}$$

Using the orthogonality of indicator functions for disjoint sets we next note that $\mathbf{Z}_{j,\ell}'\mathbf{X}_{\gamma_i^0} = 0\ \forall i \leq s_{\ell-1}$, $\mathbf{Z}_{j,\ell}'\mathbf{X}_{\gamma_i^0} = \mathbf{Z}_{j,\ell}'\mathbf{Z}_{j,\ell}\ \forall i \geq s_\ell$ and $j = 1, 2$. This leads to

$$\hat{\boldsymbol{\delta}}_{2,\ell} - \hat{\boldsymbol{\delta}}_{1,\ell} \quad = \quad \left(\frac{\mathbf{Z}_{2,\ell}'\mathbf{Z}_{2,\ell}}{T}\right)^{-1}\sum_{i=s_{\ell-1}+1}^{s_\ell-1}\frac{\mathbf{Z}_{2,\ell}'\mathbf{X}_{\gamma_i^0}}{T}\boldsymbol{\rho}_i - \left(\frac{\mathbf{Z}_{1,\ell}'\mathbf{Z}_{1,\ell}}{T}\right)^{-1}\sum_{i=s_{\ell-1}+1}^{s_\ell-1}\frac{\mathbf{Z}_{1,\ell}'\mathbf{X}_{\gamma_i^0}}{T}\boldsymbol{\rho}_i, \tag{A.27}$$

and the required result follows by proceeding exactly as in the proof of Lemma 2.1. Note that here $\widehat{\mathbf{Z}}_\ell'\widehat{\mathbf{Z}}_\ell$ is analogous to the term $\mathbf{X}'\mathbf{X}$ appearing in the context of Lemma 2.1 (see (A.12)), both corresponding to the

entire coverage of the relevant region.

PROOF OF PROPOSITION 2.4: We first consider the case $h = 2$ in (14)-(18). Since from proposition 2.3, $\hat{r}^{(1)}$ is T consistent for $\gamma^0_{(1)}$, it follows that as $T \to \infty$, the uniform probability limits of $J_T(r|\hat{r}^{(1)})/T$ as defined in (16) and $J_T(r|\gamma^0_{(1)})/T$ will be the same, say $J_\infty(r|\gamma^0_{(1)})$. It therefore suffices to show that $J_\infty(r|\gamma_{(1)})$ is uniquely maximized at $r = \gamma^0_{(2)}$. Letting $\gamma^0_{(1)} \equiv \gamma^0_{s_1}$ for any $s_1 \in \{1, 2, \ldots, m\}$, Lemma A.2 applies and $J_\infty(r|\gamma_{s_1})$ is given by expression (A.25). Since from assumption **A4** $\gamma^0_{(2)}$ is the dominant threshold among $\{\gamma^0_1, \ldots, \gamma^0_m\}\backslash\{\gamma^0_{(1)}\}$, the result then follows by proceeding as in Lemma 2.2. Using the reparameterizations presented in (14)-(18) and proceeding conditional on $\gamma^0_{(1)}$ it is also clear that the T-consistency of $\hat{r}_{(2)}$ can be established following steps that are virtually identical to the proof of proposition 2.3. The arguments extends to any h via a repeated use of Lemma A.2.

PROOF OF PROPOSITION 3.1: We first consider the case $m_0 = 0$ and prove that $P(\hat{m} = 1) \to 0$ as $T \to \infty$, which by (32) is equivalent to $P[IC_T(0) > IC_T(\gamma_1)] \to 0$ for some $\gamma_1 \in \Gamma_1$, thus implying that the procedure does not oversegment asymptotically. Using (31)-(34) we write

$$\begin{aligned} P[IC_T(0) > IC_T(\gamma_1)] &\leq P[IC_T(0) > \min_{\gamma_1 \in \Gamma_1} IC_T(\gamma_1)] \\ &= P\left[\max_{\gamma_1 \in \Gamma_1} T \log\left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2(\gamma_1)}\right) > \lambda_T K\right] \\ &= P\left[\max_{\gamma_1 \in \Gamma_1} \frac{T(\hat{\sigma}^2 - \hat{\sigma}^2(\gamma_1))}{\hat{\sigma}^2(\gamma_1)} > T(e^{\frac{\lambda_T K}{T}} - 1)\right]. \end{aligned} \tag{A.28}$$

Next note that when $m_0 = 0$ (i.e. the true model is linear, say $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$) and the fitted model is given by (2) with $m = 1$ (i.e. two regimes), (3) together with $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$ give

$$\hat{\sigma}^2(\gamma_1) = \frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{T} - \frac{\boldsymbol{\epsilon}'\mathbf{X}_1}{T}\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{T}\right)^{-1}\frac{\mathbf{X}_1'\boldsymbol{\epsilon}}{T} - \frac{\boldsymbol{\epsilon}'\mathbf{X}_2}{T}\left(\frac{\mathbf{X}_2'\mathbf{X}_2}{T}\right)^{-1}\frac{\mathbf{X}_2'\boldsymbol{\epsilon}}{T} \tag{A.29}$$

from which it follows that $\hat{\sigma}^2(\gamma_1) \xrightarrow{p} \sigma^2_\epsilon \equiv E(\epsilon^2_t)$ using assumptions **A2**(i)-(ii). Under $m_0 = 0$ and when the fitted model is given by (2) with two regimes (i.e $m = 1$) we also have

$$T(\hat{\sigma}^2_\epsilon - \hat{\sigma}^2_\epsilon(\gamma_1)) = \mathbf{H}_T(\gamma_1)'\left[\frac{\mathbf{X}_1'\mathbf{X}_1}{T} - \frac{\mathbf{X}_1'\mathbf{X}_1}{T}\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\frac{\mathbf{X}_1'\mathbf{X}_1}{T}\right]^{-1}\mathbf{H}_T(\gamma_1) \tag{A.30}$$

with

$$\mathbf{H}_T(\gamma_1) = \frac{\mathbf{X}_1'\boldsymbol{\epsilon}}{\sqrt{T}} - \frac{\mathbf{X}_1'\mathbf{X}_1}{T}\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\frac{\mathbf{X}_1'\boldsymbol{\epsilon}}{\sqrt{T}}. \tag{A.31}$$

Using assumptions **A1-A2**(i)-(ii) we obtain

$$\sup_{\gamma_1 \in \Gamma_1}\left|\left[\frac{\mathbf{X}_1'\mathbf{X}_1}{T} - \frac{\mathbf{X}_1'\mathbf{X}_1}{T}\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\frac{\mathbf{X}_1'\mathbf{X}_1}{T}\right]^{-1} - [\mathbf{G}(\gamma_1) - \mathbf{G}(\gamma_1)\mathbf{G}^{-1}\mathbf{G}(\gamma_1)]^{-1}\right| \xrightarrow{p} 0$$

with $[\mathbf{G}(\gamma_1) - \mathbf{G}(\gamma_1)\mathbf{G}^{-1}\mathbf{G}(\gamma_1)]^{-1} \succ 0$. Assumptions **A2**(i)-(iii) applied to (A.31) further imply that $\max_{\gamma_1 \in \Gamma_1} |\mathbf{H}_T(\gamma_1)| = O_p(1)$ leading to $\max_{\gamma_1 \in \Gamma_1} F_T(\gamma_1) = O_p(1)$ and thus $P[\max_{\gamma_1 \in \Gamma_1} F_T(\gamma_1) > \lambda_T K] \to 0$ since $\lambda_T \to \infty$.

Next, we concentrate on the case $m_0 = 1$ and show that $P(\hat{m} = 0) \to 0$ as $T \to \infty$, implying that the procedure does not undersegment asymptotically. We have

$$\begin{aligned} P[\hat{m} = 0] &= P[IC_T(0) < \min_{\gamma_1 \in \Gamma_1} IC_T(\gamma_1)] \\ &\leq P[IC_T(0) < IC_T(\gamma_1^0)] \\ &= P\left[\frac{\hat{\sigma}^2 - \hat{\sigma}^2(\gamma_1^0)}{\hat{\sigma}^2(\gamma_1^0)} < (e^{\frac{\lambda_T K}{T}} - 1)\right]. \end{aligned} \tag{A.32}$$

Using **A2**(i)-(ii) and standard algebra leads to

$$\hat{\sigma}^2 = \frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{T} + \boldsymbol{\rho}_1' \frac{\mathbf{X}_1^{0\prime}\mathbf{X}_1^0}{T} \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1} \frac{\mathbf{X}_2^{0\prime}\mathbf{X}_2^0}{T} \boldsymbol{\rho}_1 + o_p(1).$$

Since $\hat{\sigma}^2(\gamma_1^0) = \boldsymbol{\epsilon}'\boldsymbol{\epsilon}/T + o_p(1)$, it is then straightforward to establish that

$$\hat{\sigma}^2 - \hat{\sigma}^2(\gamma_1^0) \xrightarrow{p} \boldsymbol{\rho}_1' \mathbf{G}(\gamma_1^0)\mathbf{G}^{-1}(\mathbf{G} - \mathbf{G}(\gamma_1^0))\boldsymbol{\rho}_1 \succ 0.$$

Since when $\frac{\lambda}{T} \to 0$ we have $(e^{\frac{\lambda_T K}{T}} - 1) \to 0$ and given that the left hand side in (A.32) converges to a strictly positive constant it follows that $P[\hat{m} = 0] \to 0$ when $m_0 = 1$, as required.

PROOF OF PROPOSITION 3.2 We first show that the event $\{\hat{m} > m_0\}$ cannot occur as $T \to \infty$. Let $Q_T^{(i,j)}(1)$ denote the value of (28) evaluated in step i for subsample j. For the sequential model selection procedure to stop at m_0 (assuming all previous decisions to be correct since $\hat{m} > m_0$) it is required that $Q_T^{(m_0+1,j)}(1) < 0 \ \forall j = 1, 2, \ldots, m_0 + 1$. Thus the occurence of the event $\{\hat{m} > m_0\}$ implies the existence of at least one $j \in \{1, 2, \ldots, m_0 + 1\}$ for which $Q_T^{(m_0+1,j)}(1) > 0$. We can therefore write

$$P[\hat{m} > m_0] \leq \sum_{j=1}^{m_0+1} P[Q_T^{(m_0+1,j)}(1) > 0]$$

and

$$P[\hat{m} > m_0] \leq \sum_{j=1}^{m_0+1} P[\max_r \in (\hat{r}_{(j-1)}, \hat{r}_{(j)}) F_T^{(j)}(r) > \lambda_T K] \to 0$$

(A.34)

provided that $\lambda_T \to \infty$ and where it is understood that $\hat{r}_{(0)} \equiv \underline{\gamma}$ and $\hat{r}_{(m_0+1)} \equiv \bar{\gamma}$. The case $\{\hat{m} < m_0\}$ follows in exactly the same manner as in Proposition 3.1 since in any subsample that has at least one ignored threshold say γ we have $\hat{\sigma}^2 - \hat{\sigma}^2(\gamma) \xrightarrow{p} C > 0$.

REFERENCES

AKAIKE, H. (1973) "Information Theory and an Extension of the Maximum Likelihood Principle," in 2^{nd} *Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Budapest: Akademiai Kiado.

ALTISSIMO, F. AND G. L. VIOLANTE (1999) "The nonlinear dynamics of output and unemployment in the US," Unpublished Manuscript. University College London, Department of Economics.

ALTISSIMO, F. AND V. CORRADI (1999) "Strong Rules for Detecting the Number of Breaks in a Time Series," Unpublished Manuscript. Exeter University, Department of Economics.

ANDREWS, D. W. K. (1993) "Test for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica,* 61, 821-856.

BAI, J. (1997) "Estimating multiple breaks one at a time," *Econometric Theory,* 13, 315-352.

BAI, J. AND P. PERRON (1998) "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica,* 66, pp. 47-78.

BAI, J. AND P. PERRON (2000a) "Computation and analysis of multiple structural change models," Unpublished Manuscript. Boston University, Department of Economics.

BAI, J. AND P. PERRON (2000b) "Multiple Structural Change Models: A Simulation Analysis," Unpublished Manuscript. Boston University, Department of Economics.

BEAUDRY, P. AND G. KOOP (1993) "Do recessions permanently change output?" *Journal of Monetary Economics,* 31, 149-164.

CANER, M. AND B. E. HANSEN (2000) "Threshold autoregression with a unit root", *Econometrica,* Forthcoming.

CARRASCO, M. (1999) "Misspecified Structural Change, Threshold and Markov Switching Models," Unpublished Manuscript. University of Rochester, Department of Economics.

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, 520-533.

CRAGG, J. G. and S. G. DONALD (1997) "Inferring the Rank of a Matrix," *Journal of Econometrics,* 76, pp. 223-250.

DURLAUF, S. N., AND P. A. JOHNSON (1995) "Multiple Regimes and Cross-Country Growth Behavior," *Journal of Applied Econometrics,* 10, 365-384.

GONZALEZ, M. and J. GONZALO (1997) "Threshold Unit Root Models," Unpublished Manuscript. Universidad Carlos III de Madrid, Department of Statistics and Econometrics.

GONZALO, J. and J-Y. PITARAKIS (1998) "Specification via Model Selection in Vector Error Correction Models," *Economics Letters,* 60, pp. 321-328.

GONZALO, J. and J-Y. PITARAKIS (1999) "Dimensionality Effect in Cointegration Analysis," in *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W. J. Granger*, eds. R. F. Engle and H. White, Oxford University Press.

GONZALO, J. and R. MONTESINOS (2000) "Threshold Stochastic Unit Root Models," Unpublished Manuscript. Universidad Carlos III de Madrid, Department of Statistics and Econometrics

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

HANNAN, E. J. and M. DEISTLER (1988) *The Statistical Theory of Linear Systems,* New-York: John Wiley.

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. (1999a) "Testing for linearity," *Journal of Economic Surveys,* 13, 551-576.

HANSEN, B. E. (1999b) "Threshold effect in non-dynamic panels: Estimation, Testing, and Inference," *Journal of Econometrics,* 93, pp. 345-368.

HANSEN, B. E. (2000) "Sample Splitting and Threshold Estimation," *Econometrica,* 68, 575-603.

HAWKINS, D. M. (1976) "Point estimation of the parameters of piecewise regression models," *Applied Statistics,* 25, 51-57.

KOOP, G. AND S. M. POTTER (1999) "Dynamic asymmetries in U.S. Unemployment," *Journal of Business and Economic Statistics,* 17, 298-312.

LIU, J., S. WU AND J. W. ZIDEK (1997) "On Segmented Multivariate Regression," *Statistica Sinica,* 7, 497-525.

MICHAELIDES, A. and S. NG (2000) "Estimating the Rational Expectations Model of Speculative Storage: A Monte-Carlo Comparison of three Simulation Estimators," *Journal of Econometrics,* 96, 231-266.

NEWEY, W. K., AND D. L. MCFADDEN (1994) "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics, Vol. IV,* ed. by R. F. Engle and D. L. McFadden. New-York: Elsevier, pp. 2113-2245.

OBSTFELD, M. AND A. TAYLOR (1997) "Nonlinear Aspects of Goods Market Arbitrage and Adjustment," *Journal of Japanese and International Economics,* 11, 441-79.

O'CONNELL, P. G. J. AND S. WEI (1997) "The bigger they are the harder they fall: How price differences across U.S. cities are arbitraged," *NBER Working Paper, No. W6089.*

PETRUCCELLI, J. D. (1992) "On the approximation of time series by threshold autoregressive models," *Sankhya, Series B, 54.*

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

TONG, H. AND K. S. LIM (1980) "Threshold Autoregression, Limit Cycles and Cyclical Data," *Journal of The Royal Statistical Society, Series B,* 4, 245-292.

TONG, H. (1983) *Threshold Models in Non-Linear Time Series Analysis: Lecture Notes in Statistics,* 21, Berlin, Springer-Verlag.

TONG, H. (1990) *Non-Linear Time Series: A Dynamical System Approach,* Oxford: Oxford University Press.

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

TSAY, R. S. (1998) "Testing and Modeling Multivariate Threshold Models," *Journal of the American Statistical Association,* Vol. 93, 1188-1202.

VOSTRIKOVA, L. J. (1981) "Detecting disorder in multidimensional random processes," *Soviet. Math. Dokl.*, 24, 55-59.

YAO, Y. C. (1988) "Estimating the Number of Change-Points via Schwarz' Criterion," *Statistics and Probability Letters,* 6, 181-189.