A TREE-STRUCTURED SMOOTH TRANSITION REGRESSION MODELS

MARCELO C. MEDEIROS, JOEL CORRÊA DA ROSA, AND ALVARO VEIGA

ABSTRACT. The goal of this paper is to introduce a class of tree-structured models that combines aspects of regression trees and smooth transition regression models. The model is called the Smooth Transition Regression Tree (STR-Tree). The main idea relies on specifying a multiple-regime parametric model through a tree-growing procedure with smooth transitions among different regimes. Decisions about splits are entirely based on a sequence of Lagrange Multiplier (LM) tests of hypotheses.

KEYWORDS. Semi-parametric models, regression-trees, smooth transitions, nonlinear models, time series, regression, modeling cycle.

Very Preliminary and Incomplete

1. INTRODUCTION

IN RECENT YEARS much attention has been devoted to nonlinear modeling. Techniques such as artificial neural networks, nonparametric regression and recursive partitioning methods are frequently used to approximate unknown functional forms. In spite of their success in various applications, frequently these approaches lack interpretability due to the complexity of the final model. Some cases in which the fitted model can be given a reasonable interpretation, there are no inferential procedures that guarantee the statistical significance of the parameters. The proposal of the present paper is the construction of a nonlinear regression model that combines aspects of two well-known methodologies: Regression Trees as in Breiman, Friedman, Olshen, and Stone (1984) and the Smooth Transition Regression (STR) presented in Granger and Teräsvirta (1993), by taking advantages of their main capabilities. Our proposal inherits from tree-structured models the simplicity and interpretability of the tree-based models while the STR framework provides tools for inference-based decisions. The proposed model is called the Smooth Transition Regression Tree (STR-Tree). In our proposal, by allowing smooth splits on the tree nodes instead of sharp ones, we associate each tree architecture with a smooth transition regression model and thus it turns possible to formulate a splitting criteria that are entirely based on statistical tests of hypotheses. The Lagrange Multiplier (LM) test in the context presented by Luukkonen, Saikkonen, and Teräsvirta (1988) is adapted for deciding if a node should be split or not [1]. Here, the tree growing procedure is used as a tool for specifying a parametric model that can be analyzed either as STR model. In the former case, we can obtain confidence intervals for the parameters estimates in the tree leaves and predicted values. Decisions based on statistical inference also lessen the importance of post-pruning techniques to reduce the model complexity.

[Include more material here]

Date: November 1, 2005.

[1]See Teräsvirta (1994), van Dijk, Teräsvirta, and Franses (2002), and the references therein for successful applications of similar testing procedures.

2. Model Definition

2.1. **A Brief Introduction to Regression Trees.** Let $\mathbf{x}_t = (x_{1t}, \ldots, x_{qt})' \in \mathbb{X} \subseteq \mathbb{R}^q$ be a vector which contains q explanatory variables (covariates or predictor variables) for a continuous univariate response $y_t \in \mathbb{R}$, $t = 1, \ldots, T$. The vector $\mathbf{x}_t$ may contain lags of y_t (in case of time-series data) as well as a pre-determined or an exogenous group of variables. Suppose that the relationship between y_t and $\mathbf{x}_t$ follows a regression model of the form

$$y_t = f(\mathbf{x}_t) + \varepsilon_t, \tag{1}$$

where the function $f(\cdot)$ is unknown and, in principle, there are no assumptions about the distribution of the random term ε_t. A regression tree is a nonparametric model based on the recursive partitioning of the covariate space $\mathbb{X}$, which approximates the function $f(\cdot)$ as a sum of local models, each of which is determined in $K \in \mathbb{N}$ different regions (partitions) of $\mathbb{X}$. The model is usually displayed in a graph which has the format of a binary decision tree with $N \in \mathbb{N}$ parent (or split) nodes and $K \in \mathbb{N}$ terminal nodes (also called leaves), and which grows from the root node to the terminal nodes. Usually, the paritions are defined by a set of hyperplanes, each of which is orthogonal to the axis of a given predictor variable, called the *split variable*; see Examples 1 and 2 bellow.

The most important reference in regression tree models is the Classification and Regression Trees (CART) approach put forward by Breiman, Friedman, Olshen, and Stone (1984). In this context, the local models are just constants. However, in this paper we follow Friedman (1979) and Chaudhuri, Huang, Loh, and Yao (1994), considering linear models in each leaf [2]. Hence, conditionally to the knowledge of the subregions, the relationship between y_t and $\mathbf{x}_t$ in (1) is approximated by a piecewise linear regression, where each leaf (or terminal node) represents a distinct regime.

To mathematically represent a complex regression-tree model, we introduce the following notation. The root node is at position 0 and a parent node at position j generates left- and right-child nodes at positions $2j + 1$ and $2j + 2$, respectively. Every parent node has an associated split variable $x_{s_j t} \in \mathbf{x}_t$, where $s_j \in \mathbb{S} = \{1, 2, \ldots, q\}$. Furthermore, let $\mathbb{J}$ and $\mathbb{T}$ be the sets of indexes of the parent and terminal nodes, respectively. Then, a tree architecture can be fully determined by $\mathbb{J}$ and $\mathbb{T}$; see Examples 1 and 2.

Although tree-structured regression models are a very popular tool to nonparametric regression in biostatistics, medicine, ecology, and related areas, in Economics the applications are rather limited; see Cooper (1998), Durlauf and Johnson (1995), and Garcia and Johnson (2000) for some of the few exceptions. However, some nonlinear econometric models previously proposed in the literature are special cases of regression-trees.

[2] Other choices of approximation functions have been proposed in the literature. For example, Chaudhuri, Huang, Loh, and Yao (1994) also considered polynomial functions in each leaf. Chaudhuri, Lo, and Yang (1995) discussed the tree-structured Poisson and logistic regressions. Audrino and Bühlmann (2001) put forward a new tree-structured model, where a GARCH process is estimated in each leaf.

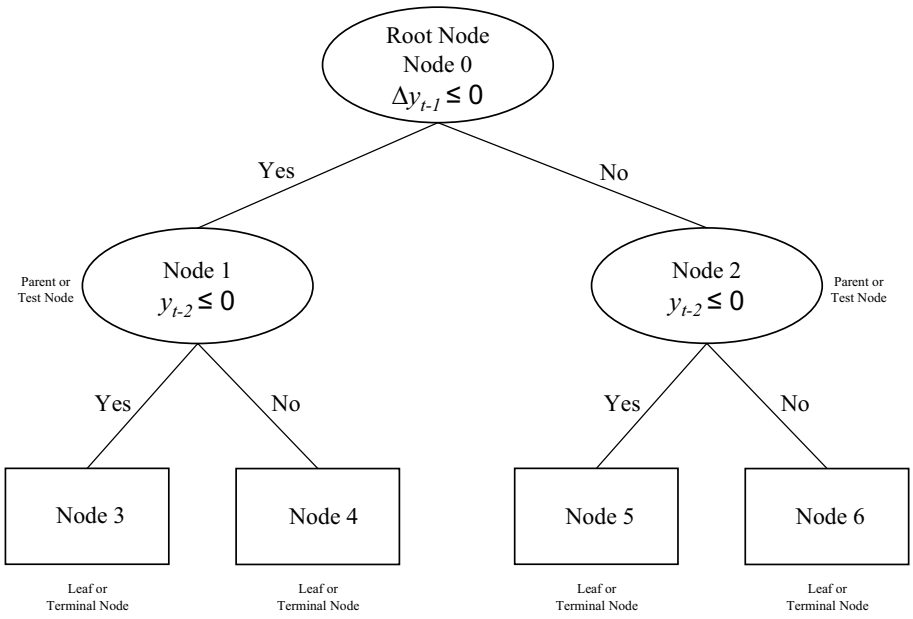

FIGURE 1. Regression tree with four terminal nodes representing (2).

EXAMPLE 1. *Consider the four-regime SETAR model for the quarterly US real GNP growth rates (in 1982 dollars), estimated in Tiao and Tsay (1994):*

$$\widehat{y}_t = \begin{cases} -0.015 - 1.076y_{t-1} & \text{if } \Delta y_{t-1} \leq 0 \text{ and } y_{t-2} \leq 0; \\ 0.006 + 0.438y_{t-1} & \text{if } \Delta y_{t-1} \leq 0 \text{ and } y_{t-2} > 0; \\ -0.006 + 0.630y_{t-1} & \text{if } \Delta y_{t-1} > 0 \text{ and } y_{t-2} \leq 0; \\ 0.004 + 0.443y_{t-1} & \text{if } \Delta y_{t-1} > 0 \text{ and } y_{t-2} > 0. \end{cases} \tag{2}$$

In (2), y_t is the quarterly change in the logarithm of the GNP series.

Model (2) can be analyzed as a regression tree as illustrated in Figure 1. The tree induced by (2) has three parent nodes (including the root node), four terminal nodes (leaves), and the depth[3] *is equal to two. In the present case, $\mathbb{J} = \{0, 1, 2\}$ and $\mathbb{T} = \{3, 4, 5, 6\}$. The split variables are $x_{s_0 t} = \Delta y_{t-1}$ and $x_{s_1 t} = x_{s_2 t} = y_{t-2}$. Model (2) maybe also written as sum of local linear models defined in regions determined by a product of indicator functions, such as,*

$$\begin{aligned} \widehat{y}_t = &(-0.015 - 1.076y_{t-1})\, I(\Delta y_{t-1}; 0) I(y_{t-2}; 0) + \\ &(0.006 + 0.438y_{t-1})\, I(\Delta y_{t-1}; 0)\, [1 - I(y_{t-2}; 0)] + \\ &(-0.006 + 0.630y_{t-1})\, [1 - I(\Delta y_{t-1}; 0)]\, I(y_{t-2}; 0) + \\ &(0.004 + 0.443y_{t-1})\, [1 - I(\Delta y_{t-1}; 0)]\, [1 - I(y_{t-2}; 0)]\,, \end{aligned}$$

[3]The depth of a tree os defined as the length of the path to the deepest leaf, i.e., the number of parent nodes (including the root node) between the root and the deepest leaf.

where

$$I(x;c) = \begin{cases} 1 & \text{if } x \leq c; \\ 0 & \text{otherwise.} \end{cases}$$

EXAMPLE 2. *Consider the estimated NeTAR model for the Jökulsá eystri (the eastern glacier river) in north-west Iceland as in Astatkie, Watts, and Watt (1997). Daily data on flow (Q_t), precipitation (P_t), and temperature (T_t) were used to estimate the following model:*

$$\widehat{Q}_t = \begin{cases} 4.82 + 0.82Q_{t-1} & \text{if } Q_{t-2} \leq 92 \text{ and} \\ & T_t \leq -2; \\ 1.32Q_{t-1} - 0.32Q_{t-2} + 0.20P_{t-1} + 0.52T_t & \text{if } Q_{t-2} \leq 92 \text{ and} \\ & -2 < T_t \leq 1.8; \\ 1.15Q_{t-1} - 0.18Q_{t-2} + 0.014P_{t-1}^2 + 1.22T_t - 0.89T_{t-3} & \text{if } Q_{t-2} \leq 92 \text{ and} \\ & T_t > 1.8; \\ 49 + 0.45Q_{t-1} + 3.47T_t + 3.75T_{t-1} - 6.08T_{t-3} & \text{if } Q_{t-2} > 92 \end{cases} \tag{3}$$

A graphical representation of (3) is illustrated in Figure 2. Model (3) has three parent nodes (including the root node), four leaves, and depth three. In this case $\mathbb{J} = \{0, 1, 4\}$ and $\mathbb{T} = \{2, 3, 9, 10\}$. The split variables are $x_{s_0 t} = Q_t$ and $x_{s_1 t} = x_{s_4 t} = T_t$. As in the previous example, model (3) can be represented by a sum of local linear models in regions determined by a product of indicator functions, such as

$$\begin{aligned} \widehat{Q}_t = {} & (4.82 + 0.82Q_{t-1})\, I(Q_{t-2}; 92) I(T_t; -2) + \\ & (1.32Q_{t-1} - 0.32Q_{t-2} + 0.20P_{t-1} + 0.52T_t)\, I(Q_{t-2}; 92)\, [1 - I(T_t; -2)]\, I(T_t; -1.8) + \\ & \left(1.15Q_{t-1} - 0.18Q_{t-2} + 0.014P_{t-1}^2 + 1.22T_t - 0.89T_{t-3}\right) I(Q_{t-2}; 92) \times \\ & \qquad [1 - I(T_t; -2)]\, [1 - I(T_t; -1.8)] + \\ & (49 + 0.45Q_{t-1} + 3.47T_t + 3.75T_{t-1} - 6.08T_{t-3})\, [1 - I(Q_{t-2}; 92)]\,. \end{aligned}$$

2.2. **Tree-Structured Smooth Transition Regression (STR-Tree).** Consider the following assumptions about the data generating process (DGP).

ASSUMPTION 1. *The observed sequence of real-valued dependent variable $\{y_t\}_{t=1}^T$ is a realization of a stationary and ergodic stochastic process on a complete probability space generated as*

$$y_t = f(\mathbf{x}_t) + \varepsilon_t, \quad t = 1, \ldots, T,$$

where $f(\mathbf{x}_t)$ is a unknown measurable function of the real-valued random vector $\mathbf{x}_t \in \mathbb{X} \subseteq \mathbb{R}^q$, which has distribution function F on Ω, an Euclidean space. The sequence $\{\varepsilon_t\}_{t=1}^T$ is formed by random variables drawn from an absolutely continuous (with respect to a Lebesgue measure on the real line), positive everywhere distribution such that $\mathsf{E}[\varepsilon_t] = 0$ and $\mathsf{E}[\varepsilon_t^2] = \sigma_t^2 < \infty, \forall\, t$. In addition, assume that $\mathsf{E}[\varepsilon_t | \mathbf{x}_t] = 0$ and $\mathsf{E}[\varepsilon_t f(\mathbf{x}_t)] = 0$.

Assumption 1 imposes some mild restrictions on the true DGP. It is important to notice that, in principle, there are no strict assumptions neither on the functional form of the function $f(\mathbf{x}_t)$ nor on the distribution of ε_t.

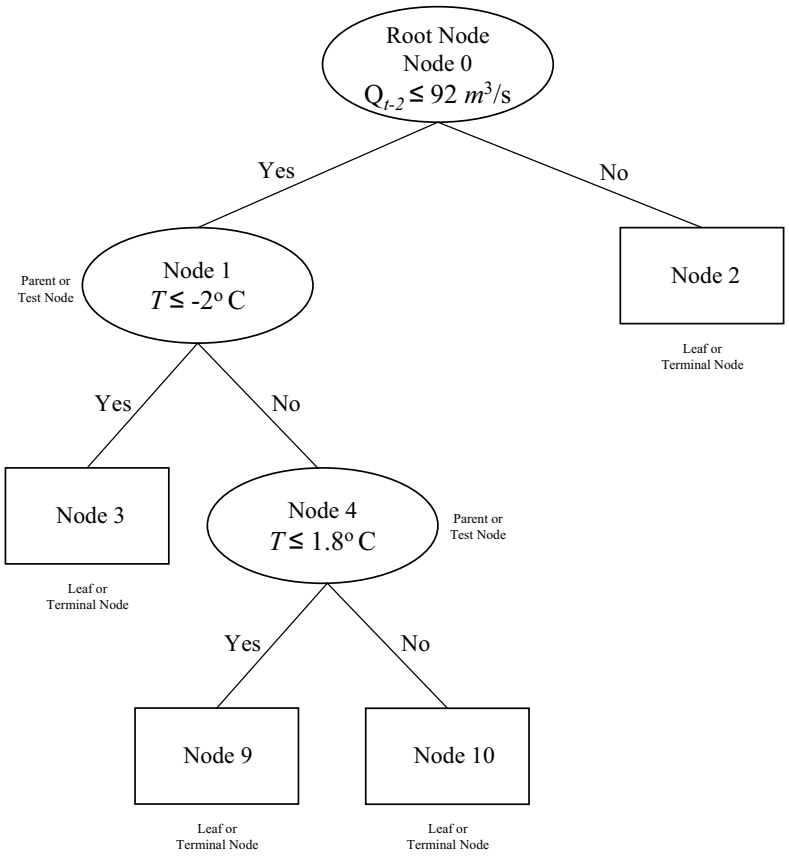

FIGURE 2. Regression tree model with four terminal nodes representing (3).

In this paper we propose a tree-structured model to best approximate the conditional mean $\mathsf{E}[y_t|\mathbf{x}_t] = f(\mathbf{x}_t)$. Extending the results presented in da Rosa, Veiga, and Medeiros (2003), we consider a linear model in each leaf (regime) as well as applications to dependent observations, taking advantage of much of the regression-tree structure presented in Section 2.1, but also introducing elements which make it feasible to use standard inferential procedures. It is possible to interpret our proposal as a fully parametric, semiparametric, or nonparametric model.

The key idea of the paper is to follow da Rosa, Veiga, and Medeiros (2003), replacing the sharp splits in the regression-tree model by smooth splits. For instance, consider the tree model in Examples 1 and 2 and replace the indicator function $I(x;c)$ by a logistic function defined as

$$G(x;\gamma,c) = \frac{1}{1+e^{-\gamma(x-c)}}. \tag{4}$$

Now we have the additional parameter γ, called the *slope parameter*, which controls the smoothness of the logistic function. Note that the regression tree model is nested in the smooth transition specification as a special case obtained when the slope parameter approaches infinity. The parameter c is called the *location parameter*.

DEFINITION 1. *Let $\mathbf{z}_t \subseteq \mathbf{x}_t$ such that $\mathbf{z}_t \in \mathbb{R}^p$, $p \leq q$. The sequence of of real-valued vectors $\{\mathbf{z}_t\}_{t=1}^T$ is stationary and ergodic. Set $\widetilde{\mathbf{z}}_t = (1, \mathbf{z}_t)'$. A parametric model $\mathcal{M}$ defined by the function $H_{\mathbb{JT}}(\mathbf{x}_t; \boldsymbol{\psi}) : \mathbb{R}^{q+1} \to \mathbb{R}$, indexed by the vector of parameters $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, a compact subset of the Euclidean space, is called Smooth Transition Regression Tree model, STR-Tree, if*

$$H_{\mathbb{JT}}(\mathbf{x}_t; \boldsymbol{\psi}) = \sum_{i \in \mathbb{T}} \boldsymbol{\beta}_i' \widetilde{\mathbf{z}}_t B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_i) \tag{5}$$

where

$$B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_i) = \prod_{j \in \mathbb{J}} G(x_{s_j,t}; \gamma_j, c_j)^{\frac{n_{i,j}(1+n_{i,j})}{2}} \left[1 - G(x_{s_j,t}; \gamma_j, c_j)\right]^{(1-n_{i,j})(1+n_{i,j})} \tag{6}$$

and

$$n_{i,j} = \begin{cases} -1 & \text{if the path to leaf } i \text{ does not include the parent node } j; \\ 0 & \text{if the path to leaf } i \text{ includes the right-child node of the parent node } j; \\ 1 & \text{if the path to leaf } i \text{ includes the left-child node of the parent node } j. \end{cases} \tag{7}$$

Let $\mathbb{J}_i$ be the subset of $\mathbb{J}$ containing the indexes of the parent nodes that form the path to leaf i. Then, $\boldsymbol{\theta}_i$ is the vector containing all the parameters (γ_k, c_k) such that $k \in \mathbb{J}_i$, $i \in \mathbb{T}$.

The functions $B_{\mathbb{J}i}$, $0 < B_{\mathbb{J}i} < 1$, are known as the membership functions. Note that $\sum_{j \in \mathbb{J}} B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_j) = 1$, $\forall\, \mathbf{x}_t \in \mathbb{R}^{q+1}$.

The reason for considering $\mathbf{z}_t$ as a subset of $\mathbf{x}_t$ is to avoid nonstationary regressors in the local linear models. For example, $\mathbf{x}_t$ may contain a linear trend that is an interesting split variable when there are possible structural breaks in the series.

3. COMPARISON WITH OTHER NONLINEAR MODELS

3.1. **Parametric View.** First, consider the case where $\mathsf{E}[y_t|\mathbf{x}_t] = H_{\mathbb{JT}}(\mathbf{x}_t; \boldsymbol{\psi})$ and where the number of limiting regimes (or terminal nodes) is bounded. Then, the true DGP is a piecewise linear model with possibly smooth transitions among the regimes and the STR-Tree specification is a model of fixed complexity. Several nonlinear parametric models fall into this category. For example, setting $K = 2$ in (5), the model becomes the Logistic STR (LSTR) model. If the slope parameter tends to infinity, then the resulting model is the well-known Threshold Regression. The multiple regime STR (MRSTR) model of van Dijk and Franses (1999), the additive LSTAR model of Öcal and Osborn (2000), and the Time-Varying STR (TV-STR) model of Lundbergh, Teräsvirta, and van Dijk (2003) are also special cases of our specification. Thus, the STR-Tree model can be an interesting parametrization to describe asymmetries and multiple regimes in the dynamics of macroeconomic variables over the business cycles.

On the other, neural network based models of fixed complexity are, in principle, not nested into the STR-Tree framework. In neural networks and related models the transition (or split) variables are linear combinations of all the elements of $\mathbf{x}_t$, that are estimated together with all the other parameters. This allows that the hyperplanes that determine the division of the input space to be nonorthogonal among themselves. In that sense, one may argument that such models are more flexible than our tree-structured specification. However, we have three arguments in favor of our model. First, estimating the linear combination of input variables is rather difficult, specially when the dimension of the input

space increases and the number of observations is limited. Furthermore, interpretation of the linear combination is, in most cases, not available. As shown before, tree-structured models are more easily interpreted. Second, to overcome the lost in flexibility due to the restriction of orthogonal hyperplanes, the nested structure of tree models allows that only a subregion of the input space is split, allowing for genuinely different regimes. Finally, all the theoretical results of the paper are still valid if we consider a linear combination of variables as split variables. In that case, it is straightforward to generalize the STR-Tree specification in order to include Neural Network based models as special cases.

3.2. **Nonparametric View.** Suppose now that the true DGP is not nested into our tree-structured specification and we want to best approximate the conditional mean $\mathsf{E}[y_t|\mathbf{x}_t] = f(\mathbf{x}_t)$. In that situation the STR-Tree model is a misspecified parametric model and is more convenient to interpret the model as a nonparametric specification. As shown in Section 4, the STR-Tree model is capable of simultaneously approximating the unknown function $f(\mathbf{x}_t)$ and its derivatives as far as the complexity of the model is allowed to grow as the sample size increases. In such way, we can interpret the STR-Tree specification as a nonparametric model, where instead of traditional kernel functions we use the membership functions $B_{\mathbb{J}i}(\mathbf{x}_t, \boldsymbol{\theta}_i)$, $i = 1 \in \mathbb{T}$; see Equation 5.

3.3. **Semiparametric View.** Consider a given terminal node $i^* \in \mathbb{T}$. Model 5 can be easily written as

$$H_{\mathbb{J}\mathbb{T}}(\mathbf{x}_t; \boldsymbol{\psi}) = \boldsymbol{\beta}'_{i^*}\widetilde{\mathbf{z}}_t B_{\mathbb{J}i^*}(\mathbf{x}_t; \boldsymbol{\theta}_{i^*}) + \sum_{i \in \mathbb{T}-\{i^*\}} \boldsymbol{\beta}'_i \widetilde{\mathbf{z}}_t B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_i) \equiv \boldsymbol{\lambda}'\widetilde{\mathbf{z}}_t + \eta(\mathbf{x}_t), \tag{8}$$

where $\eta(\mathbf{x}_t)$ is a nonlinear function of $\mathbf{x}_t$. If our interest relies only on the parameter vector $\boldsymbol{\beta}_{i^*}$ than the STR-Tree model can be seen as a semiparametric specification.

4. Approximation Capabilities

Define

$$\mathbb{H}_{q,K,\boldsymbol{\psi}} = \left\{ \mathbb{R}^q \to \mathbb{R} \middle| H_{\mathbb{J}\mathbb{T}}(\mathbf{x}_t; \boldsymbol{\psi}) = \sum_{i \in \mathbb{T}} \boldsymbol{\beta}'_i \widetilde{\mathbf{z}}_t B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_i) \right\}, \tag{9}$$

as the set of all functions implemented by the STR-Tree specification with K terminal nodes. The goal of this section is first to show that any function $f(\mathbf{x}_t) \in \mathcal{C}(\mathbb{X})$, where $\mathcal{C}(\mathbb{X})$ is the space of all continuous functions on $\mathbb{X}$, a compact subspace of $\mathbb{R}^q$, can be arbitrarily accurate approximated by a function that belongs to $\mathbb{H}_{q,K,\boldsymbol{\psi}}$, as far sufficiently many terminal nodes are available. In addition, we also show that if the function f has continuous derivatives, than there exist a function $h \in \mathbb{H}_{q,K,\boldsymbol{\psi}}$ that can simultaneously approximate the function f and its derivatives. This section is strongly based on developments presented in Hornik (1991).

Our approximation results are based on the $L^p(\mu)$ performance criteria, where μ is a finite input environment measure and $L^p(\mu)$ is the space of all functions f such that $\|f\|_{p,\mu} < \infty$, where

$$\|f\|_{p,\mu} = \left[\int_{\mathbb{R}^q} |f(\mathbf{x}_t)|^p \right]^{1/p},$$

$1 \leq p < \infty$.

Define $\rho_{p,\mu}\left[f(\mathbf{x}_t) - h(\mathbf{x}_t)\right] = \left\|f(\mathbf{x}_t) - h(\mathbf{x}_t)\right\|_{p,\mu}$. A subset $\mathcal{S}$ of $L^p(\mu)$ is dense in $L^p(\mu)$ if for arbitrary $f \in L^p(\mu)$ and $\delta > 0$ there is a function $h \in \mathcal{S}$ such that $\rho_{p,\mu}(f,h) < \delta$. In other words, h can approximate f to any desired degree of accuracy.

Our first approximation theorem states the main result concerning continuous functions.

THEOREM 1. *The set $\mathbb{H}_{q,K,\boldsymbol{\psi}}$ is dense in $\mathcal{C}(\mathbb{X})$ for all compact subsets $\mathbb{X}$ of $\mathbb{R}^q$.*

The following developments concerns the simultaneous approximation of the unknown function and its derivatives. First, define as a multi-index the q-tuple $\alpha = (\alpha_1, \ldots, \alpha_q)$ of nonnegative integers. Let $|\alpha| = \alpha_1 + \cdots + \alpha_q$ be the order of the multi-index α and set

$$D^\alpha f(\mathbf{x}_t) = \frac{\partial^{|\alpha|}}{\partial x_{1t}^{\alpha_1} \cdots \partial x_{qt}^{\alpha_q}} f(\mathbf{x}_t)$$

as the corresponding partial derivative of a sufficiently smooth function $f(\mathbf{x}_t)$. In addition, define $\mathcal{C}^m(\mathbb{R}^q)$ as the space of all functions f which, together with all their partial derivatives $D^\alpha g$ of order $|\alpha| \leq m$, are continuous on $\mathbb{R}^q$. Finally, for all subsets $\mathbb{X}$ of $\mathbb{R}^q$ and $f \in \mathcal{C}^m(\mathbb{R}^q)$, let

$$\|f\|_{m,\mu,\mathbb{X}} \equiv \max_{|\alpha| \leq m} \sup_{\mathbf{x}_t \in \mathbb{X}} \left|D^\alpha f(\mathbf{x}_t)\right|.$$

A subset $\mathcal{S}$ of $\mathcal{C}^m(\mathbb{R}^q)$ is uniformly m-dense on compacta in $\mathcal{C}^m(\mathbb{R}^q)$ if for all $f \in \mathcal{C}^m(\mathbb{R}^q)$, for all compact subsets $\mathbb{X}$ of $\mathbb{R}^q$, and for all $\delta > 0$ there is a function $h = h(f, \mathbb{X}, \delta) \in \mathcal{S}$ such that $\|f - h\|_{m,\mu,\mathbb{X}} < \delta$.

For $f \in \mathcal{C}^m(\mathbb{R}^q)$, μ a finite measure on $\mathbb{R}^q$ and $1 \leq p < \infty$, let

$$\|f\|_{m,p,\mu} \equiv \left[\sum_{|\alpha| \leq m} \int_{\mathbb{R}^q} |D^\alpha g|^p\right]^{1/p},$$

and the weighted Sobolev space $\mathcal{C}^{m,p}(\mu)$ be defined by

$$\mathcal{C}^{m,p}(\mu) = \left\{f \in \mathcal{C}^m(\mathbb{R}^q) : \|f\|_{m,p,\mu} < \infty\right\}.$$

Observe that $\mathcal{C}^{m,p}(\mu) = \mathcal{C}^m(\mathbb{R}^q)$ if μ has compact support. A subset $\mathcal{S}$ of $\mathcal{C}^{m,p}(\mu)$ is dense in $\mathcal{C}^{m,p}(\mu)$, if for all $f \in \mathcal{C}^{m,p}(\mu)$ and $\delta > 0$ there is a function $h = h(f, \delta) \in \mathcal{S}$ such that $\|f - h\|_{m,p,\mu} < \delta$.

We then have the following result.

THEOREM 2. *$\mathbb{H}_{q,K,\boldsymbol{\psi}}$ is uniformly m-dense on compacta in $\mathcal{C}^m(\mathbb{R}^q)$ and dense in $\mathcal{C}^{m,p}(\mu)$ for all finite measures μ on $\mathbb{R}^q$.*

5. PROBABILISTIC PROPERTIES

Consider the case where $\mathbf{x}_t = (y_{t-1}, \ldots, y_{t-p})'$. In that situation, it is interesting to know under which conditions the STR-Tree model is stationary. The following theorem gives sufficient conditions for second-order stationarity of the STR-Tree model.

THEOREM 3. *Let $\mathbf{x}_t = (y_{t-1}, \ldots, y_{t-p})'$. Consider $\mathbb{T}_0$ the set of all terminal nodes (regions) where $B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_i) \longrightarrow 0$, as $\|\mathbf{x}_t\| \longrightarrow \infty$. Denote $\overline{\mathbb{T}}_0$ as its complement and $\boldsymbol{\beta}_i = (\beta_{1i}, \beta_{2i}, \ldots, \beta_{pi})'$. The*

STR-Tree model is second-order stationary if all the roots of the polynomial $1 - \beta_{1i}w - \beta_{2i}w^2 - \cdots - \beta_{pi}w^p$, $i \in \overline{\mathbb{T}}_0$, *are outside the unit circle.*

6. Model Estimation and Asymptotic Theory

In this section we discuss the estimation of the STR-Tree model and the corresponding asymptotic theory. As the true DGP is unknown, the STR-Tree model is just an approximation to $f(\mathbf{x}_t)$. The parameters of model (5) are estimated by nonlinear least-squares (NLS) which is equivalent to quasi-maximum likelihood (QML) estimation. Let $\widehat{\boldsymbol{\psi}}$ be the quasi-maximum likelihood estimator (QMLE) of $\boldsymbol{\psi}$ given by

$$\widehat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmin}} \mathcal{Q}_T(\boldsymbol{\psi}) = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmin}} \frac{1}{T}\sum_{t=1}^{T} q_t(\boldsymbol{\psi}) = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmin}} \left\{ \frac{1}{T}\sum_{t=1}^{T} \left[y_t - H_{\mathbb{JT}}(\mathbf{x}_t;\boldsymbol{\psi})\right]^2 \right\}. \tag{10}$$

Following Domowitz and White (1982) and White (1994), define $\boldsymbol{\psi}^*$ as the parameter vector that minimizes the average prediction (or approximation) mean squared error [4],

$$\overline{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} \int \left[f(\mathbf{x}_t) - H_{\mathbb{JT}}(\mathbf{x}_t;\boldsymbol{\psi}) + \varepsilon_t\right]^2 dF. \tag{11}$$

As the data-generating and approximation functions are assumed to be time-invariant, $\boldsymbol{\psi}^*$ can be considered as the parameter vector of a locally unique best approximation $H_{\mathbb{JT}}(\mathbf{x}_t;\boldsymbol{\psi})$ to $f(\mathbf{x}_t)$ over some compact set $\mathbb{C}$ with a nonempty interior contained in the range of $\mathbf{x}_t$.

6.1. **Parametric Estimation.**

Define $\mathcal{Q}(\boldsymbol{\psi}) = \mathsf{E}[q_t(\boldsymbol{\psi})]$. In the following two subsections, we discuss the existence of $\mathcal{Q}(\boldsymbol{\psi})$ and the identifiability of the STR-Tree model. Then, in Subsection 6.1.3, we prove the strong consistency of $\widehat{\boldsymbol{\psi}}_T$ to $\boldsymbol{\psi}^*$. Asymptotic normality of the QMLE is considered in Subsection 6.1.4.

6.1.1. *Existence of the QMLE.* The following theorem proves the existence of $\mathcal{Q}(\boldsymbol{\psi})$. It is based on Theorem 2.12 in White (1994), which establishes that under certain conditions of continuity and measurability of the quasi-loglikelihood function, $\mathcal{Q}(\boldsymbol{\psi})$ exists.

Theorem 4. *Under Assumption 1,* $\mathcal{Q}(\boldsymbol{\psi})$ *exists and is finite.*

6.1.2. *Identifiability of the Model.* A fundamental problem for statistical inference that haunts nonlinear econometric models is the unidentifiability of the quasi-loglikelihood function. In order to carry statistical inference we need to show that $\boldsymbol{\psi}^*$ is the unique minimizer of $\mathcal{Q}(\boldsymbol{\psi})$. However, the unconstrained STR-Tree model is neither locally nor globally identified. Three properties of the STR-Tree model cause unidentifiability of the models:

(P.1) The property of interchangeability of the regimes (terminal nodes). If in (5), a group of functions $B_{\mathbb{J}i}(\mathbf{x}_t;\boldsymbol{\theta}_i)$, $i \in \mathbb{T}$, is permuted, than we can also permute the parameters $\boldsymbol{\beta}_i$ yielding the same value of the quasi-loglikelihood function. This results in several different models that are

[4] $\boldsymbol{\psi}^*$ also minimizes the Kullback-Leibler Information Criterion (KLIC).

indistinct among themselves. As a consequence, in the estimation of parameters, we will have several equal local minima for $\mathcal{Q}(\boldsymbol{\psi})$.

(P.2) The fact that $G(x_{s_j};\gamma_j,c_j) = 1 - G(x_{s_j};-\gamma_j,c_j)$, $j \in \mathbb{J}$.

(P.3) The presence of irrelevant regimes (leaves). For example, consider a parent node j and its child-nodes $2j+1$ and $2j+2$. If $\boldsymbol{\beta}_{2j+1} = \boldsymbol{\beta}_{2j+2}$, then parameters γ_j and c_j remain unidentified, for some $j \in \mathbb{J}$. On the other hand, if $\gamma_j = 0$, then parameters $\boldsymbol{\beta}_{2j+1}$ and $\boldsymbol{\beta}_{2j+2}$ may take on any value without affecting the value of the quasi-loglikelihood function.

Hence, establishing restrictions on the parameters of (5) that simultaneously avoid any permutation of regimes (property (P.1)), and symmetries in the logistic function (property (P.2)), and model reducibility (property (P.3)), we guarantee the identifiability of the model.

The problem of interchangeability (property (P.1)) can be prevented by considering the following restriction:

RESTRICTION 1. *Consider the two adjacent split nodes $2j+1$ and $2j+2$ generated from the same parent node $j \in \mathbb{J}$. If $x_{s_{2j+1},t} = x_{s_{2j+2},t}$, then $c_{2j+1} \leq c_{2j+2}$ (or $\gamma_{2j+1} \leq \gamma_{2j+2}$). Equality is resolved by arranging the split nodes such that there is at least one $k \in \{0,1,2,\ldots,p\}$ such that $\beta_{k,2j+1} < \beta_{k,2j+2}$.*

Now the consequences due to the symmetry of the logistic function (property (P.2)) and the presence of irrelevant regimes, property (P.3), can be solved if we impose the following restrictions:

RESTRICTION 2. *The parameters of the STR-Tree model satisfy the following restrictions:*

(1) *Consider two terminal nodes $2j+1$ and $2j+2$ generated from the same parent node $j \in \mathbb{J}$. The parameter vectors $\boldsymbol{\beta}_{2j+1}$ and $\boldsymbol{\beta}_{2j+2}$ are such that $\boldsymbol{\beta}_{2j+1} \neq \boldsymbol{\beta}_{2j+2}$.*
(2) *The parameters $\boldsymbol{\beta}_i \neq \mathbf{0}$, $\forall\, i \in \mathbb{T}$.*
(3) *The parameters $\gamma_j > 0$, $\forall\, j \in \mathbb{J}$.*

THEOREM 5. *Under Assumption 1 and Restrictions 1 and 2, the STR-Tree model is globally identifiable. Furthermore, $\mathcal{Q}(\boldsymbol{\psi})$ is uniquely maximized at $\boldsymbol{\psi}^*$.*

In practice, the presence of irrelevant regimes will be circumvented by applying a "specific-to-general" model building strategy as discussed in Section 7.1.

6.1.3. *Consistency.* Consider the following assumption.

ASSUMPTION 2. *The parameter vector $\boldsymbol{\psi}^* \in \boldsymbol{\Psi}$ with maximizes the quasi-loglikelihood function (or minimizes the KLIC) is in the interior of $\boldsymbol{\Psi}$, a compact subset of finite dimensional Euclidean space.*

THEOREM 6. *Under the Assumptions 1 and 2 and the additional assumption that $\mathsf{E}\left[|\varepsilon_t|^2\right] < \infty$, $\forall\, t$, the QMLE $\widehat{\boldsymbol{\psi}}$ is strong consistent for $\boldsymbol{\psi}^*$, i.e., $\widehat{\boldsymbol{\psi}} \xrightarrow{a.s.} \boldsymbol{\psi}^*$.*

6.1.4. *Asymptotic Normality*. First, we introduce the following matrices:

$$\mathbf{A}(\boldsymbol{\psi}^*) = \mathsf{E}\left[-\frac{\partial^2 q_t(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\bigg|_{\boldsymbol{\psi}^*}\right] \text{ and}$$

$$\mathbf{B}(\boldsymbol{\psi}^*) = \mathsf{E}\left[\frac{\partial q_t(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\bigg|_{\boldsymbol{\psi}^*}\frac{\partial q_t(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}'}\bigg|_{\boldsymbol{\psi}^*}\right].$$

Consider the additional matrices:

$$\begin{aligned}\mathbf{A}_T(\boldsymbol{\psi}) &= \frac{2}{T}\sum_{t=1}^{T}\left\{\frac{\partial H_{\mathbb{JT}}(\mathbf{x}_t;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\frac{\partial H_{\mathbb{JT}}(\mathbf{x}_t;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}'} - \frac{\partial^2 H_{\mathbb{JT}}(\mathbf{x}_t;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\left[y_t - H_{\mathbb{JT}}(\mathbf{x}_t;\boldsymbol{\psi})\right]\right\} \\ \mathbf{B}_T(\boldsymbol{\psi}) &= \frac{4}{T}\sum_{t=1}^{T}\left\{\left[y_t - H_{\mathbb{JT}}(\mathbf{x}_t;\boldsymbol{\psi})\right]\frac{\partial H_{\mathbb{JT}}(\mathbf{x}_t;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\frac{\partial H_{\mathbb{JT}}(\mathbf{x}_t;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}'}\right\}.\end{aligned} \tag{12}$$

The following theorem states the asymptotic normality result.

THEOREM 7. *Under Assumptions 1–2, then*

$$\sqrt{T}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) \xrightarrow{D} \mathsf{N}\left(\mathbf{0}, \mathbf{A}(\boldsymbol{\psi}^*)^{-1}\mathbf{B}(\boldsymbol{\psi}^*)\mathbf{A}(\boldsymbol{\psi}^*)^{-1}\right), \tag{13}$$

where $\mathbf{A}(\boldsymbol{\psi}^*)$ *and* $\mathbf{B}(\boldsymbol{\psi}^*)$ *are consistently estimated by* $\mathbf{A}_T(\widehat{\boldsymbol{\psi}})$ *and* $\mathbf{B}_T(\widehat{\boldsymbol{\psi}})$*, respectively.*

7. MODEL BUILDING

The modeling cycle of the STR-Tree model involves three steps, namely: Specification, estimation, and model evaluation. The specification consists of three decisions: The choice of relevant variables, the selection of the node to be split (if this is the case), and the selection of the splitting (or transition) variable.

The possible candidate variables maybe a set of (weakly) exogenous variables determined from some underlying theory and/or a set of lags of the dependent variable. When the number of exogenous (or predetermined) variables are too high it is useful to first select a subset of the candidate variables (the relevant ones) before continuing the building of the STR-Tree model. The same thing happens when lagged values of y_t are considered as candidate variables. The maximum lag order should be determined first. The reasons for doing this are twofold: First, the model building strategy will be accelerated and second, a parsimonious model will be estimated in each leaf. Several ways of selecting the relevant variables exist. In the STAR literature is common to select the set of relevant variables with information criteria, making use a linear approximation to the true DGP. This is also a possibility here. However, as pointed out by Pitarakis (2004) this may have an adverse effect on the final model specification. An alternative approach is to adopt a polynomial approximation to the true DGP as proposed in Rech, Teräsvirta, and Tschernig (2001) and applied with success in Medeiros, Teräsvirta, and Rech (in press), Medeiros and Veiga (2005), Suarez-Fariñas, Pedreira, and Medeiros (2004), and Medeiros, Veiga, and Pedreira (2001). A third possibility is to use nonparametric techniques. Several nonparametric variable selection techniques are available (Tschernig and Yang 2000, Vieu 1995, Tjøstheim and Auestad 1994, Yao and

Tong 1994, Auestad and Tjøstheim 1990), but they are computationally very demanding, in particular when the number of observations is not small.

The the selection of the node to be split (if this is the case), and the selection of the splitting (or transition) variable will be carried out by sequence of Lagrange Multiplier (LM) tests following the ideas originally presented in Luukkonen, Saikkonen, and Teräsvirta (1988) and vastly used in the literature; see Teräsvirta (1994), van Dijk, Teräsvirta, and Franses (2002), Medeiros, Teräsvirta, and Rech (in press), da Rosa, Veiga, and Medeiros (2003), and the references therein. As mentioned in the introduction, our goal is to build a coherent strategy to select the variables/lags and to grow the STR-Tree model using statistical inference. As the STR-Tree is a nonlinear approximation of the true model, the results of Domowitz and White (1982) and White (1994) will be advocated. An alternative approach based on 10-fold cross-validation is also possible; however, the computational burden involved is dramatically high as shown in da Rosa, Veiga, and Medeiros (2003).

The estimation of the parameters of the model will be carried out by nonlinear least-squares which is equivalent to quasi-maximum likelihood estimation as discussed in Section 6.

What follows thereafter is *evaluation* of the final estimated model. Tree-structured models are usually evaluated by their out-of-sample performance (predictive ability). However, the tree-grown procedure is carried out throw a sequence of neglected nonlinearity tests which can be interpreted also as an evaluation test. The construction of tests for serial autocorrelation in the same spirit of Eitrheim and Teräsvirta (1996) and Medeiros and Veiga (2003) is also possible but is beyond the scope of the paper.

Following the "specific-to-general" principle, we start the cycle from the root node (depth 0) and the general steps are:

(1) Selection of the relevant variables.
(2) Specification of the model by selecting in the depth d, using the LM test, a node to be split (if not in the root node) and a splitting variable.
(3) Estimation of the parameters.
(4) Evaluation of the estimated model by checking if it is necessary to:
    (a) Change the node to be split.
    (b) Change the splitting variable.
    (c) Remove the split.
(5) Use the final tree model for prediction or descriptive purposes.

7.1. **Growing the Tree.** Consider a STR-Tree model with K leaves and we want to test if the terminal node $i^* \in \mathbb{T}$ should be split or not. Write the model as

$$\begin{aligned} y_t = & \sum_{i \in \mathbb{T}-\{i^*\}} \boldsymbol{\beta}_i' \widetilde{\mathbf{z}}_t B_{\mathbb{J}i}\left(\mathbf{x}_t; \boldsymbol{\theta}_i\right) + \\ & \boldsymbol{\beta}_{2i^*+1}' \widetilde{\mathbf{z}}_t B_{\mathbb{J}2i^*+1}\left(\mathbf{x}_t; \boldsymbol{\theta}_{2i^*+1}\right) + \boldsymbol{\beta}_{2i^*+2}' \widetilde{\mathbf{z}}_t B_{\mathbb{J}2i^*+2}\left(\mathbf{x}_t; \boldsymbol{\theta}_{2i^*+2}\right) + \varepsilon_t, \end{aligned} \tag{14}$$

where

$$\begin{aligned} B_{\mathbb{J}2i^*+1}\left(\mathbf{x}_t; \boldsymbol{\theta}_{2i^*+1}\right) &= B_{\mathbb{J}i^*}\left(\mathbf{x}_t; \boldsymbol{\theta}_{i^*}\right) G(x_{i^*t}; \gamma_{i^*}, c_{i^*}) \\ B_{\mathbb{J}2i^*+2}\left(\mathbf{x}_t; \boldsymbol{\theta}_{2i^*+2}\right) &= B_{\mathbb{J}i^*}\left(\mathbf{x}_t; \boldsymbol{\theta}_{i^*}\right) \left[1 - G(x_{i^*t}; \gamma_{i^*}, c_{i^*})\right]. \end{aligned}$$

In a more compact form, Equation (14) maybe written as

$$\begin{aligned} y_t = & \sum_{i \in \mathbb{T}-\{i^*\}} \boldsymbol{\beta}_i' \widetilde{\mathbf{z}}_t B_{\mathbb{J}i}\left(\mathbf{x}_t; \boldsymbol{\theta}_i\right) + \\ & \boldsymbol{\phi}' \widetilde{\mathbf{z}}_t B_{\mathbb{J}i^*}\left(\mathbf{x}_t; \boldsymbol{\theta}_{i^*}\right) + \boldsymbol{\lambda}' \widetilde{\mathbf{z}}_t B_{\mathbb{J}i^*}\left(\mathbf{x}_t; \boldsymbol{\theta}_{i^*}\right) G(x_{i^*t}; \gamma_{i^*}, c_{i^*}) + \varepsilon_t, \end{aligned} \tag{15}$$

where $\boldsymbol{\phi} = \boldsymbol{\beta}_{2i^*+2}$ and $\boldsymbol{\lambda} = \boldsymbol{\beta}_{2i^*+1} - \boldsymbol{\beta}_{2i^*+2}$.

In order to test the statistical significance of the split, a convenient null hypothesis is $\mathsf{H}_0 : \gamma_{i^*} = 0$ against the alternative $\mathsf{H}_a : \gamma_{i^*} > 0$. An alternative null hypothesis is $\mathsf{H}_0' : \boldsymbol{\lambda} = \mathsf{0}$. However, it is clear in (15) that under H_0, the nuisance parameters $\boldsymbol{\lambda}$ and c_{i^*} can assume different values without changing the quasi-loglikelihood function, posing an identification problem (Davies 1977, Davies 1987).

We adopt as a solution for this problem the one proposed in Luukkonen, Saikkonen, and Teräsvirta (1988), that is to approximate the logistic function by a third-order Taylor expansion around $\gamma_{i^*} = 0$. After some algebra we get

$$\begin{aligned} y_t = & \sum_{i \in \mathbb{T}-\{i^*\}} \boldsymbol{\beta}_i' \widetilde{\mathbf{z}}_t B_{\mathbb{J}i}\left(\mathbf{x}_t; \boldsymbol{\theta}_i\right) + \boldsymbol{\alpha}_0' \widetilde{\mathbf{z}}_t B_{\mathbb{J}i^*}\left(\mathbf{x}_t; \boldsymbol{\theta}_{i^*}\right) + \\ & \boldsymbol{\alpha}_1' \widetilde{\mathbf{z}}_t B_{\mathbb{J}i^*}\left(\mathbf{x}_t; \boldsymbol{\theta}_{i^*}\right) x_{i^*t} + \boldsymbol{\alpha}_2' \widetilde{\mathbf{z}}_t B_{\mathbb{J}i^*}\left(\mathbf{x}_t; \boldsymbol{\theta}_{i^*}\right) x_{i^*t}^2 + \boldsymbol{\alpha}_3' \widetilde{\mathbf{z}}_t B_{\mathbb{J}i^*}\left(\mathbf{x}_t; \boldsymbol{\theta}_{i^*}\right) x_{i^*t}^3 + e_t, \end{aligned} \tag{16}$$

where $e_t = \varepsilon_t + \boldsymbol{\lambda}' \widetilde{\mathbf{z}}_t B_{\mathbb{J}i^*}\left(\mathbf{x}_t; \boldsymbol{\theta}_{i^*}\right) R(x_{i^*t}; \gamma_{i^*}, c_{i^*})$ and $R(x_{i^*t}; \gamma_{i^*}, c_{i^*})$ is the remainder. The parameter vectors $\boldsymbol{\alpha}_k$, $k = 0, \ldots, 3$ are function of the original parameters of the model.

Thus the null hypothesis becomes

$$\mathsf{H}_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_3 = \mathbf{0}. \tag{17}$$

Note that under H_0 the remainder of the Taylor expansion vanishes and $e_t = \varepsilon_t$, so that the properties of the error process remain unchanged under the null and thus asymptotic inference can be used, yielding the following result

THEOREM 8. *If, under the null, the conditions of Theorem 8 are met,* $\mathsf{E}(|x_{lt}|^8) < \infty$, $l = 1, \ldots, q$ *, and* $\varepsilon_t \sim \mathsf{NID}(0, \sigma^2)$*, then*

$$LM = \frac{1}{\widehat{\sigma}^2} \sum_{t=1}^{T} \widehat{u}_t \widehat{\boldsymbol{\nu}}_t' \left\{ \sum_{t=1}^{T} \widehat{\boldsymbol{\nu}}_t \widehat{\boldsymbol{\nu}}_t' - \sum_{t=1}^{T} \widehat{\boldsymbol{\nu}}_t \widehat{\mathbf{h}}_t' \left(\sum_{t=1}^{T} \widehat{\mathbf{h}}_t \widehat{\mathbf{h}}_t' \right)^{-1} \sum_{t=1}^{T} \widehat{\mathbf{h}}_t \widehat{\boldsymbol{\nu}}_t' \right\}^{-1} \sum_{t=1}^{T} \widehat{\boldsymbol{\nu}}_t \widehat{u}_t \tag{18}$$

where $\widehat{u}_t = y_t - H_{\mathbb{J}\mathbb{T}}(\mathbf{x}_t; \widehat{\boldsymbol{\psi}})$*,*

$$\widehat{\mathbf{h}}_t = \left. \frac{\partial H_{\mathbb{J}\mathbb{T}}(\mathbf{x}_t; \boldsymbol{\psi})'}{\partial \boldsymbol{\psi}} \right|_{\boldsymbol{\psi}=\widehat{\boldsymbol{\psi}}} = \Bigg[\widetilde{\mathbf{z}}_t B_{\mathbb{J}i_1}(\mathbf{x}_t; \boldsymbol{\psi}), \ldots, \widetilde{\mathbf{z}}_t B_{\mathbb{J}i_K}(\mathbf{x}_t; \boldsymbol{\psi}), \\ \boldsymbol{\beta}_{i_1}' \widetilde{\mathbf{z}}_t \left. \frac{\partial B_{\mathbb{J}i_1}(\mathbf{x}_t; \boldsymbol{\theta}_{i_1})}{\partial \boldsymbol{\theta}_{i_1}} \right|_{\boldsymbol{\theta}_{i_1}=\widehat{\boldsymbol{\theta}}_{i_1}}, \ldots, \left. \frac{\partial B_{\mathbb{J}i_K}(\mathbf{x}_t; \boldsymbol{\theta}_{i_K})}{\partial \boldsymbol{\theta}_{i_K}} \right|_{\boldsymbol{\theta}_{i_K}=\widehat{\boldsymbol{\theta}}_{i_K}} \Bigg]'$$

and $\boldsymbol{\nu}_t = \left[\widetilde{\mathbf{z}}_t B_{\mathbb{J}i^*} x_{i^*t}, \widetilde{\mathbf{z}}_t B_{\mathbb{J}i^*} x_{i^*t}^2, \widetilde{\mathbf{z}}_t B_{\mathbb{J}i^*} x_{i^*t}^3 \right]$*, has an asymptotic* χ^2 *distribution with* $m = 3(p+1)$ *degrees of freedom.*

The test can also be carried out in stages as follows:

(1) Estimate model (5) with K regimes. If the sample size is small and the model is thus difficult to estimate, numerical problems in applying the maximum likelihood algorithm may lead to a solution such that the residual vector is not precisely orthogonal to the gradient matrix of $H_{\mathbb{JT}}(\mathbf{x}_t; \widehat{\boldsymbol{\psi}})$. This has an adverse effect on the empirical size of the test. To circumvent this problem, we regress the residuals $\widehat{u}_t$ on $\widehat{\mathbf{h}}_t$ and compute the sum of squared residuals $SSR_0 = \sum_{t=1}^{T} \tilde{u}_t^2$. The new residuals $\tilde{u}_t$ are orthogonal to $\widehat{\mathbf{h}}_t$.
(2) Regress $\tilde{u}_t$ on $\widehat{\mathbf{h}}_t$ and $\widehat{\boldsymbol{\nu}}_t$. Compute the sum of squared residuals $SSR_1 = \sum_{t=1}^{T} \widehat{v}_t^2$.
(3) Compute the χ^2 statistic

$$LM_{\chi^2}^{hn} = T\frac{SSR_0 - SSR_1}{SSR_0}, \tag{19}$$

or the F version of the test

$$LM_F^{hn} = \frac{(SSR_0 - SSR_1)/m}{SSR_1/(T-n-m)}, \tag{20}$$

where $n = (q+2)h + p + 1$. Under H_0, $LM_{\chi^2}^{hn}$ has an asymptotic χ^2 distribution with m degrees of freedom and LM_F^{hn} is approximately F-distributed with m and $T - n - m$ degrees of freedom.

In case of the assumption of normal and homoskedastic errors is violated, than a robust version of the test is available following the results inWooldridge (1990). A robust version of the LM type test, can be carried out as follows:

(1) As before.
(2) Regress $\widehat{\boldsymbol{\nu}}_t$ on $\widehat{\mathbf{h}}_t$ and compute the residuals $\mathbf{r}_t$.
(3) Regress 1 on $\tilde{\varepsilon}_t \mathbf{r}_t$ and compute the sum of squared residuals SSR_1.
(4) Compute the value of the test statistic

$$LM_{\chi^2}^{r} = T - SSR_1. \tag{21}$$

The test statistic has the same asymptotic χ^2 null distribution as before.

7.2. **Sequential Tests.** To achieve the final tree model, we perform a sequence of n correlated LM-type tests of hypothesis in which n is a random variable. During this sequence, the harmful decision to be taken, according to the principle of tree-complexity as function of the number of terminal nodes, is to decide erroneously for splitting a node. Due to multiplicity from repeated significance testing, we have to control the overall type I error under the risk of an overstatement of the significance of the results (more splits are reported to be significant than it should be). To remedy this situation, we adopt the following procedure. For the nth test in the sequence, if it is performed in the dth depth the significance level is $\alpha(d, n) = \frac{\alpha}{n^d}$.

In the root node $(d = 0)$ and we apply the first test $(n = 1)$ for splitting the node at a significance level α, if the null is rejected than we the second $(n = 2)$ test is applied in the 1st depth $(d = 1)$ and the significance level is $\alpha/2$. Then, if the tree grows by completing all depths, the significance level evolves like $\alpha/3$, $\alpha/4^2$, $\alpha/5^2$, $\alpha/6^3$, $\alpha/7^3$, $\alpha/8^4$, $\alpha/9^4$, etc. By forcing the test to be more rigorous in deeper depths, we create a procedure that diminishes the importance of using post-pruning techniques.

There are several alternatives to control the overall size of the sequence of tests (Hochberg 1988, Benjamini and Hochberg 1995, Benjamini and Hochberg 1997, Benjamini and Yekutieli 2000, Benjamini

and Yekutieli 2001, Benjamini and Liu 1999). However, by our experiments, the simple methodology described above seems to work quite well and the comparison between different techniques to reduce the nominal size of each test is beyond the scope of the paper. In practice, different methodologies can be tested and possible different architectures may be compared by their out-of-sample performance.

8. Monte Carlo Experiment

9. Real Examples

10. Conclusions

References

Ahn, H. (1996): "Log-Gamma Regression Modeling Through Regression Trees," *Communications in Statistics – Theory and Methods*, 25, 295–311.

Astatkie, T., D. G. Watts, and W. E. Watt (1997): "Nested Threshold Autoregressive (NeTAR) Models," *International Journal of Forecasting*, 13, 105–116.

Audrino, F., and P. Bühlmann (2001): "Tree-Structured GARCH Models," *Journal of the Royal Statistical Society, Series B*, 63, 727–744.

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

Benjamini, Y., and Y. Hochberg (1995): "Controlling the False Dicovery Rate - A practical and Powerful Approach to Multiple Testing," *Journal of teh Royal Statistical Society – Series B*, 57, 289–300.

——— (1997): "Multiple Hypotheses Testing with Weights," *Scandinavian Journal of Statistics*, 24, 407–418.

Benjamini, Y., and W. Liu (1999): "A Step-Down Multiple Hypothesis Testing Procedures that Controls the False Discovery Rate Under Independence," *Journal of Statistical Inference and Planning*, 82, 163–170.

Benjamini, Y., and D. Yekutieli (2000): "On the Adaptive Control of the Discovery Fate in Multiple Testing with Independent Statistics," *Journal of Educational and Behavioral Statistics*, 25, 60–83.

——— (2001): "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *Annals of Statistics*, 29, 1165–1188.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984): *Classification and Regression Trees*. Belmont Wadsworth Int. Group, New York.

Öcal, N., and D. Osborn (2000): "Business Cycle Nonlinearities in UK Consumption and Production," *Journal of Applied Econometrics*, 15, 27–43.

Chaudhuri, P., M.-C. Huang, W.-Y. Loh, and R. Yao (1994): "Piecewise-Polynomial Regression Trees," *Statistica Sinica*, 4, 143–167.

Chaudhuri, P., W.-D. Lo, and W.-Y. L. C.-C. Yang (1995): "Generalized Regression Trees," *Statistica Sinica*, 5, 641–666.

Ciampi, A. (1991): "Generalized Regression Trees," *Computational Statistics and Data Analysis*, 12, 57–78.

Cooper, S. J. (1998): "Multiple Regimes in US Output Fluctuations," *Journal of Business and Economic Statistics*, 16, 92–100.

Crowley, J., and M. L. Blanc (1993): "Survival Trees by Goodness of Split," *Journal of the American Statistical Association*, 88, 457–467.

da Rosa, J. C., A. Veiga, and M. C. Medeiros (2003): "Tree-Structured Smooth Transition Regression Models Based on CART Algorithm," Textos para Discussão 469, Pontifical Catholic University of Rio de Janeiro.

Davies, R. B. (1977): "Hypothesis Testing When the Nuisance Parameter in Present Only Under the Alternative," *Biometrika*, 64, 247–254.

——— (1987): "Hypothesis Testing When the Nuisance Parameter in Present Only Under the Alternative," *Biometrika*, 74, 33–44.

DENISON, T., B. K. MALLIK, AND A. F. M. SMITH (1998): "A Bayesian CART Algorithm," *Biometrika*, 85, 363–377.

DOMOWITZ, I., AND H. WHITE (1982): "Misspecified Models with Dependent Observations," *Journal of Econometrics*, 20, 35–58.

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

EITRHEIM, ., AND T. TERÄSVIRTA (1996): "Testing the Adequacy of Smooth Transition Autoregressive Models," *Journal of Econometrics*, 74, 59–75.

FRIEDMAN, J. H. (1979): "A Tree-Structured Approach to Nonparametric Multiple Regression," in *Smoothing Techniques for Curve Estimation*, ed. by T. G. anf M. Rosenblatt, pp. 5–22. Springer-Verlag.

GARCIA, M. G. P., AND P. A. JOHNSON (2000): "A Regression Tree Analysis of Real Interest Rate Regime Changes," *Applied Financial Economics*, 10, 171–176.

GRANGER, C. W. J., AND T. TERÄSVIRTA (1993): *Modelling Nonlinear Economic Relationships*. Oxford University Press, Oxford.

HOCHBERG, Y. (1988): "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 75, 800–802.

HORNIK, K. (1991): "Approximation Capabilities of Multilayer Feedforward Networks," *Neural Networks*, 4, 251.

JAJUGA, K. (1986): "Linear Fuzzy Regression," *Fuzzy Sets and Systems*, 20, 343–353.

LUNDBERGH, S., T. TERÄSVIRTA, AND D. VAN DIJK (2003): "Time-Varying Smooth Transition Autoregressive Models," *Journal of Business and Economic Statistics*, 21, 104–121.

LUUKKONEN, R., P. SAIKKONEN, AND T. TERÄSVIRTA (1988): "Testing Linearity Against Smooth Transition Autoregressive Models," *Biometrika*, 75, 491–499.

MEDEIROS, M. C., T. TERÄSVIRTA, AND G. RECH (in press): "Building Neural Network Models for Time Series: A Statistical Approach," *Journal of Forecasting*.

MEDEIROS, M. C., AND A. VEIGA (2003): "Diagnostic Checking in a Flexible Nonlinear Time Series Model," *Journal of Time Series Analysis*, 24, 461–482.

——— (2005): "A Flexible Coefficient Smooth Transition Time Series Model," *IEEE Transactions on Neural Networks*, 16.

MEDEIROS, M. C., A. VEIGA, AND C. E. PEDREIRA (2001): "Modelling Exchange Rates: Smooth Transitions, Neural Networks, and Linear Models," *IEEE Transactions on Neural Networks*, 12, 755–764.

MORGAN, J., AND J. SONQUIST (1963): "Problems in The Analysis of Survey Data and a Proposal," *Journal of the American Statistical Association*, 58, 415–434.

PITARAKIS, J. Y. (2004): "Model Selection Uncertainty and Detection of Threshold Effects," Discussion Papers in Economics and Econometrics 409, University of Southampton.

RECH, G., T. TERÄSVIRTA, AND R. TSCHERNIG (2001): "A Simple Variable Selection Technique for Nonlinear Models," *Communications in Statistics, Theory and Methods*, 30.

SEGAL, M. R. (1992): "Tree-Structured Methods for Longitudinal Data," *Journal of the American Statistical Association*, 87, 407–418.

SUAREZ-FARIÑAS, M., C. E. PEDREIRA, AND M. C. MEDEIROS (2004): "Local Global Neural Networks: A New Approach for Nonlinear Time Series Modeling," *Journal of the American Statistical Association*, 99, 1092–1107.

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

TIAO, G. C., AND R. S. TSAY (1994): "Some Advances in Non-linear and Adaptive Modelling in Time-series," *Journal of Forecasting*, 13, 109–131.

TJØSTHEIM, D., AND B. AUESTAD (1994): "Nonparametric Identification of Nonlinear Time Series – Selecting Significant Lags," *Journal of the American Statistical Association*, 89(428), 1410–1419.

TSCHERNIG, R., AND L. YANG (2000): "Nonparametric Lag Selection for Time Series," *Journal of Time Series Analysis*, 21, 457–487.

VAN DIJK, D., AND P. H. FRANSES (1999): "Modelling Multiple Regimes in the Business Cycle," *Macroeconomic Dynamics*, 3, 311–340.

VAN DIJK, D., T. TERÄSVIRTA, AND P. H. FRANSES (2002): "Smooth Transition Autoregressive Models - A Survey of Recent Developments," *Econometric Reviews*, 21, 1–47.

VIEU, P. (1995): "Order Choice in Nonlinear Autoregressive Models," *Statistics*, 26, 307–328.

WHITE, H. (1994): *Estimation, Inference and Specification Analysis*. Cambridge University Press, New York, NY.

WOOLDRIDGE, J. M. (1990): "A unified approach to robust, regression-based specification tests," *Econometric Theory*, 6, 17–43.

YAO, Q., AND H. TONG (1994): "On Subset Selection in Non-Parametric Stochastic Regression," *Statistica Sinica*, 4, 51–70.

Appendix A. PROOFS

Appendix B. COMPUTATIONAL ISSUES

Appendix B.1. **Concentrated Least-Squares.** Conditional on the knowledge of the parameters $\boldsymbol{\theta}_i$ in (5), $i \in \mathbb{T}$, model (5) is just a linear regression and the vector of parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_i)_{i\in\mathbb{T}}$ can be estimated by ordinary least-squares (OLS) as

$$\widehat{\boldsymbol{\beta}} = \left[\mathbf{B}(\boldsymbol{\theta})'\mathbf{B}(\boldsymbol{\theta})\right]^{-1}\mathbf{B}(\boldsymbol{\theta})'\mathbf{y}, \tag{B.1}$$

where $\mathbf{y} = (y_1, \ldots, y_T)'$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i\in\mathbb{T}}$, and $\mathbf{B}(\boldsymbol{\theta}) = (B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_i)\mathbf{z}_t)_{i\in\mathbb{T}, t=1,\ldots,T}$.

The parameters $\boldsymbol{\theta}_i$ are estimated conditionally on $\boldsymbol{\beta}$ by applying the Levenberg-Marquadt algorithm which completes the ith iteration. As the NLS algorithm is sensitive to the choice of starting-values, we suggest the use of a grid of possible starting-values.

Appendix B.2. **Starting Values.**

(M. C. Medeiros) DEPARTMENT OF ECONOMICS, PONTIFICAL CATHOLIC UNIVERSITY OF RIO DE JANEIRO, RIO DE JANEIRO, RJ, BRAZIL.

E-mail address: mcm@econ.puc-rio.br

(J. C. da Rosa) DEPARTMENT OF STATISTICS, FEDERAL UNIVERSITY OF PARANÁ, CURITIBA, PR, BRAZIL.

E-mail address: joelm@est.ufpr.br

(A. Veiga) DEPARTMENT OF ELECTRICAL ENGINEERING, PONTIFICAL CATHOLIC UNIVERSITY OF RIO DE JANEIRO, RIO DE JANEIRO, RJ, BRAZIL.

E-mail address: alvf@ele.puc-rio.br