TEXTO PARA DISCUSSÃO

No. 538

Estimation and asymptotic theory for a new class of mixture models

Eduardo F. Mendes
Alvaro Veiga
Marcelo C. Medeiros

DEPARTAMENTO DE ECONOMIA
www.econ.puc-rio.br

ESTIMATION AND ASYMPTOTIC THEORY FOR A NEW CLASS OF MIXTURE MODELS

Eduardo F. Mendes

Department of Statistics, Northwestern University, Evanston, IL

E-mail: eduardo.mendes@northwestern.edu

Alvaro Veiga

Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro

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

Marcelo C. Medeiros

Department of Economics, Pontifical Catholic University of Rio de Janeiro

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

ABSTRACT

In this paper a new model of mixture of distributions is proposed, where the mixing structure is determined by a smooth transition tree architecture. The tree structure yields a model that is simpler, and in some cases more interpretable, than previous proposals in the literature. Based on the Expectation-Maximization (EM) algorithm a quasi-maximum likelihood estimator is derived and its asymptotic properties are derived under mild regularity conditions. In addition, a specific-to-general model building strategy is proposed in order to avoid possible identification problems. Both the estimation procedure and the model building strategy are evaluated in a Monte Carlo experiment. The approximation capabilities of the model is also analyzed in a simulation experiment. Finally, applications with real datasets are considered.

KEYWORDS: Mixture models, smooth transition, regression tree, conditional distribution.

1. Introduction

Recent years have witnessed a vast development of nonlinear time series techniques (Tong 1990, Granger and Teräsvirta 1993). From a parametric point of view, the Smooth Transition (Auto-)Regression, ST(A)R, proposed by Chan and Tong (1986)[1] and further developed by Luukkonen, Saikkonen and Teräsvirta (1988) and Teräsvirta (1994), has found a number of successful applications; see van Dijk, Teräsvirta and Franses (2002) for a review. In the time series literature, the STAR model is a natural generalization of the Threshold Autoregressive (TAR) models pioneered by Tong (1978) and Tong and Lim (1980).

On the other hand, nonparametric models that do not make assumptions about the parametric form of the functional relationship between the variables to be modeled have become widely applicable due to computational advances (Härdle 1990, Härdle, Lütkepohl and Chen 1997, Fan and Yao 2003). Another class of models, the flexible functional forms, offers an alternative that leaves the functional form of the relationship partially unspecified. While these models do contain parameters, often a large number of them, the parameters are not globally identified. Identification, if achieved, is local at best without imposing restrictions on the parameters. Usually, the parameters are not interpretable as they often are in parametric models. In most cases, these models are interpreted as nonparametric sieve (or series) approximations (Chen and Shen 1998).

The neural network (NN) model is a prominent example of such a flexible functional form. Although the NN model can be interpreted as a parametric alternative (Kuan and White 1994, Trapletti, Leisch and Hornik 2000, Medeiros, Teräsvirta and Rech 2006), its use in applied work is generally motivated by the mathematical result stating that, under mild regularity conditions, a NN model is capable of approximating any Borel-measurable function to any given degree of accuracy; see, for instance, Hornik, Stinchombe and White (1990), Gallant and White (1992), and Chen and White (1998).

[1] Chan and Tong (1986) called the model Smooth Threshold Auto-regression.

The above mentioned models aim to describe the conditional mean of the series. In terms of the conditional variance, Engle's (1982) Autoregressive Conditional Heteroskedasticity (ARCH) model, Bollerslev's (1986) Generalized ARCH (GARCH) specification, and Taylor´s (1986) Stochastic Volatility (SV) model are the most popular alternatives for capturing time-varying volatility, and have motivated a myriad of extensions (Poon and Granger 2003, McAleer 2005, Andersen, Bollerslev and Diebold 2006).

However, when the attempt is to model the entire conditional distribution, the mixture-of-experts (ME) proposed by Jacobs, Jordan, Nowlan and Hinton (1991) becomes a viable alternative. The core idea is to have a family of models, which is flexible enough to capture not only the nonlinearities in the conditional mean, but also to capture other complexities in the conditional distribution. The model is based on the ideas of Nowlan (1990), viewing competitive adaptation in unsupervised learning as an attempt to fit a mixture of simple probability distributions. Jordan and Jacobs (1994) proposed the hierarchical mixture-of-experts (HME). Applications of ME and HME models in time series are given by Weigend, Mangeas and Srivastava (1995) and Huerta, Jiang, and Tanner (2001,2003). Recently, Carvalho and Tanner (2005a) proposed the mixture of generalized linear time series models and derived several asymptotic results. It would worth mentioning the Mixture Autoregressive (MAR) model proposed by Wong and Li (2000,2001).

In this paper we contribute to the literature by proposing a new class of mixture of models that is based on regression-trees with smooth splits. Our proposal has the advantage of being flexible but less complex than the HME specification, providing possible interpretation for the final estimated model. Furthermore, a simple model building strategy has been developed and Monte Carlo simulations show that it works well in small samples. A quasi-maximum likelihood estimator (QMLE) is described and its asymptotic properties are analyzed.

The paper proceeds as follows. In Section 2 a brief review of the literature on mixture of models for time series is presented. Our proposal is presented in Section 3. In Section 4, parameter estimation and the asymptotic theory are considered. The modeling cycle is described in Section 5. Simulations are shown in Section 6, and Section 7 presents some examples with actual data. Finally, Section 8 concludes. All technical proofs are relegated to the appendix.

2. Mixture of Models: A Brief Review of the Literature

In this section we present the class of models considered in this paper.

DEFINITION 1. *The conditional probability density function (p.d.f.),* $f(y_t|\mathbf{x}_t;\boldsymbol{\theta})$*, of a random variable* y_t *is a Mixture of Models with* K *basis distributions if*

$$f(y_t|\mathbf{x}_t;\boldsymbol{\theta}) = \sum_{i=1}^{K} g_i(\mathbf{x}_t;\boldsymbol{\theta}_i)\pi_i(y_t|\mathbf{x}_t;\boldsymbol{\psi}_i), \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^q$ *is a vector of covariates,* $\boldsymbol{\theta} = [\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K', \boldsymbol{\psi}_1', \ldots, \boldsymbol{\psi}_K']'$ *is a parameter vector,* $\pi_i(y_t|\mathbf{x}_t;\boldsymbol{\psi}_i)$ *is some known parametric family of distributions (basis distributions), indexed by the vector of parameters* $\boldsymbol{\psi}_i$*, and* $g_i(\mathbf{x}_t;\boldsymbol{\theta}_i) \in [0,1]$ *is the weight function.*

If y_t is distributed as in (1), then

$$\mathbb{E}[y_t|\mathbf{x}_t] = \sum_{i=1}^{K} g_i(\mathbf{x}_t;\boldsymbol{\theta}_i)\mathbb{E}_{\pi_i}[y_t|\mathbf{x}_t;\boldsymbol{\psi}_i] \quad \text{and} \quad \mathbb{V}[y_t|\mathbf{x}_t] = \sum_{i=1}^{K} g_i^2(\mathbf{x}_t;\boldsymbol{\theta}_i)\mathbb{V}_{\pi_i}[y_t|\mathbf{x}_t;\boldsymbol{\psi}_i],$$

where $\mathbb{E}_{\pi_i}$ and $\mathbb{V}_{\pi_i}$ are the expected value and the variance taken with respect to π_i.

The simplest model belonging to this class is the TAR model, where a threshold variable controls the switching between different local Gaussian linear models. An indicator variable defines which local model is active and only one model is active each time. The conditional p.d.f. remains Gaussian and the conditional moments do not depend on the covariates. Many models have been proposed to overcome these limitations. The MAR model of Wong and Li (2000) uses a mixture of Gaussian distributions with static weights. However, this model is still limited because the weights do not vary across time (or with the covariate vector). Wong and Li (1999) suggest a generalization called a generalized mixture of autoregressive model (GMARX). This generalization considers only two Gaussian local models and the weights are given by a logistic equation. The GMARX model has a limited number of local models. The ME model of Jacobs et al. (1991) describes the conditional distribution using gated NNs to switch between local nonlinear models. This specification though very flexible, has a high number of parameters and is very hard to interpret, specify, and estimate. On the other hand, the HME is a tree-structured mixture of generalized linear models, where the weights are given by a product of multinomial logit functions. Each node of the tree can have any number of splits, hence the specification and estimation of the model are also demanding. Furthermore, for the most general model there are no results that guarantee consistency of the estimators. Finally, the model is not completely interpretable once the subdivisions of the space are done by hyperplanes which, in turn, are not necessarily interpretable.

To overcome some of the drawbacks caused by a profligate parametrization, Zeevi, Meir and Adler (1998) proposed the mixture autoregressive (MixAR) and Carvalho and Tanner (2005b) considered the mixture of generalized experts, which are simplifications of the HME model. In both cases the weights are given by a multinomial logit function. Probabilistic properties and approximation results were proved for both models; see Zeevi et al. (1998), Carvalho and Tanner (2005a) and Carvalho and Skoulakis (2004).

The model proposed in this paper combines the simplicity of the decision trees with the flexibility of the mixture of models. However, our model is simpler, is less parameterized, is more easily interpretable and the model building strategy is well defined. The tree-structured mixture of models has a binary tree as the decision structure and the decision frontier is not a linear combination of the covariates, just one of the covariates each time.

3. MODEL PRESENTATION

The core idea is to model the weight functions g in (1) as a smooth transition regression-tree model, as in da Rosa, Veiga and Medeiros (2008).

To represent a 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 $s_j \in \mathbb{S} = \{1, 2, \ldots, q\}$. Let $\mathbb{J}$ and $\mathbb{T}$ be the sets of indexes of the parent and terminal nodes, respectively. Then, any tree $\mathbb{JT}$ can be fully determined by $\mathbb{J}$ and $\mathbb{T}$.

DEFINITION 2. *The random variable $y_t \in \mathbb{R}$ follows a tree-structured mixture of models (Tree-MM) if its conditional probability density function (p.d.f.) $f(y_t|\mathbf{x}_t;\boldsymbol{\theta})$ is written as*

$$f(y_t|\mathbf{x}_t;\boldsymbol{\theta}) = \sum_{i\in\mathbb{T}} B_i(\mathbf{x}_t;\boldsymbol{\theta}_i)\pi(y_t|\mathbf{x}_t;\boldsymbol{\beta}_i'\widetilde{\mathbf{x}}_t,\sigma_i^2), \tag{2}$$

where $\mathbf{x}_t \in \mathbb{R}^q$ is a vector of explanatory variables, $\boldsymbol{\theta}$ is the conditional p.d.f. parameter vector, $\pi(\cdot)$ is the Gaussian p.d.f. with parameter vector $\boldsymbol{\psi}_i = (\boldsymbol{\beta}', \sigma_i)'$, $\widetilde{\mathbf{x}}_t = (1, \mathbf{x}_t')'$,

$$B_i\left(\mathbf{x}_t;\boldsymbol{\theta}_i\right) = \prod_{j\in\mathbb{J}} g(x_{s_{j,t}};\gamma_j,c_j)^{\frac{n_{i,j}(1+n_{i,j})}{2}}\left[1-g(x_{s_{j,t}};\gamma_j,c_j)\right]^{(1-n_{i,j})(1+n_{i,j})}, \tag{3}$$

and

$$n_{i,j} = \begin{cases} -1 & \textit{if the path to leaf } i \textit{ does not include the parent node } j; \\ 0 & \textit{if the path to leaf } i \textit{ includes the right-child node of the parent node } j; \\ 1 & \textit{if the path to leaf } i \textit{ includes the left-child node of the parent node } j. \end{cases} \tag{4}$$

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 $\boldsymbol{\nu}_k = (\gamma_k, c_k)$ such that $k \in \mathbb{J}_i$, $i \in \mathbb{T}$. Furthermore, $g(x_{s_k,t};\gamma_k,c_k) = \left[1+e^{-\gamma_k(x_{s_k,t}-c_k)}\right]^{-1}$.

4. Parameter Estimation

The estimation of $\boldsymbol{\theta}$ is carried out by maximizing the quasi-likelihood of the density function in (2). In a more general framework we cannot suppose that our probability model is correctly specified, so we use the Quasi-Maximum Likelihood Estimator (QMLE), which is the same as the Maximum Likelihood Estimator under the correct specification. Thus, we can write the conditional quasi-likelihood based on a sample $\{y_t\}_{t=1}^{T}$ as

$$\mathcal{L}_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log \left[\sum_{i \in \mathbb{T}} B_i(\mathbf{x}_t; \boldsymbol{\theta}_i) \pi(y_t | \mathbf{x}_t; \boldsymbol{\psi}_i) \right]. \tag{5}$$

Numerical optimization is carried out using the EM algorithm of Dempster, Laird and Rubin (1977). The idea behind the EM algorithm is to maximize a sequence of simple functions which leads to the same solution as maximizing a complex function. This technique were also used by Jordan and Jacobs (1994), Le, Martin and Raftery (1996), Wong and Li (1999,2000), Huerta, Jiang and Tanner (2001) and Carvalho and Tanner (2005b).

4.1. Asymptotic Theory.

In this section we present a set of asymptotic results with respect to the estimator. First, we present a set of assumptions about the (unknown) true probability model.

Assumption 1. *The observed data are a realization of a stochastic process* $\{(y_t, \mathbf{x}_t)\}_{t=1}^{T}$, *where the unknown true probability model* $\mathcal{G}_t \equiv \mathcal{G}[(y_t, \mathbf{x}_t); \cdot]$ *is a continuous density on* $\mathbb{R}$, *and the true likelihood function is identifiable and has a unique maximum at* $\boldsymbol{\theta}_0$.

We define $\boldsymbol{\theta}^*$ as the parameter vector that minimize the Kullback-Leibler divergence criterion between the true probability model, $\mathcal{G}_t$, and the estimated probability model, $f(\cdot; \boldsymbol{\theta})$. Hence, the QMLE $\widehat{\boldsymbol{\theta}}_T$ of $\boldsymbol{\theta}^*$, is defined as:

$$\widehat{\boldsymbol{\theta}}_T = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\mathsf{argmax}} \mathcal{L}_T(\boldsymbol{\theta}). \tag{6}$$

ASSUMPTION 2. *The parameter vector $\boldsymbol{\theta}^*$ is interior to a compact parameter space $\boldsymbol{\Theta} \in \mathbb{R}^{r_1} \times \mathbb{R}_+^{r_2}$, where $r_1 = 2(\#\mathbb{J}) + (p+1)(\#\mathbb{T})$, $r_2 = \#\mathbb{T}$, and $\#$ is the cardinality operator.*

The identifiability of mixture of experts models was shown in Jiang and Tanner (1999) for the case where the gating functions are multinomial logits. Since our gating function is different, the conditions presented there are not adequate. We show in Appendix A that under mild conditions, the model is identifiable such that the following assumption holds.

ASSUMPTION 3. *The tree mixture-of-expert structure, as presented in (2), is identifiable, in the sense that, for a sample $\{y_t; \mathbf{x}_t\}_{t=1}^T$, and for $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$,*

$$\prod_{t=1}^{T} f(y_t|\mathbf{x}_t; \boldsymbol{\theta}_1) = \prod_{t=1}^{T} f(y_t|\mathbf{x}_t; \boldsymbol{\theta}_2) \quad , a.s.$$

is equivalent to $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

The following theorem establishes the existence of the QMLE.

THEOREM 1 (Existence). *Under Assumptions 1 – 3, the QMLE exists and $\mathbb{E}[\mathcal{L}_T(\boldsymbol{\theta})]$ has a unique maximum at $\boldsymbol{\theta}^*$.*

To ensure the consistency of the QMLE, we state additional conditions.

ASSUMPTION 4. *The process $\{(y_t, \mathbf{x}_t)\}_{t=1}^T$ is strictly stationary and strong mixing.*

ASSUMPTION 5. *Let $\mathbf{Y}_t = (y_t,\, \mathbf{x}_t')'$, then $\mathbb{E}\left[\mathbf{Y}_t \mathbf{Y}_t'\right] < \infty$.*

THEOREM 2. *Under Assumptions 1–5, $\widehat{\boldsymbol{\theta}}_T \overset{a.s.}{\rightarrow} \boldsymbol{\theta}^*$.*

For asymptotic normality we need the following additional assumption:

ASSUMPTION 6. $\mathbb{E}\left[\mathbf{Y}_t \otimes \mathbf{Y}_t \otimes \mathbf{Y}_t \otimes \mathbf{Y}_t\right] < \infty$.

THEOREM 3 (Asymptotic Normality). *Under Assumptions 1–6,*

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

where

$$\mathbf{A}(\boldsymbol{\theta}^*) = \mathbb{E}\left[-\left.\frac{\partial^2 \mathcal{L}_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}^*}\right] \textit{ and } \mathbf{B}(\boldsymbol{\theta}^*) = \mathbb{E}\left[\left.\frac{\partial \mathcal{L}_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}^*} \left.\frac{\partial \mathcal{L}_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}^*}\right].$$

5. MODELING CYCLE

There are two approaches for the model selection problem: hypothesis testing and the use of a model information criterion (IC). As shown in Quinn, McLachlan and Hjort (1987), the likelihood ratio tests are not applicable because, under the null hypothesis of fewer basis distributions, the model is non-identified and the test does not have the standard chi-square asymptotic distribution.

There is also an alternative hypothesis testing methodology following, for example, Medeiros and Veiga (2005) and Medeiros et al. (2006). This methodology is based on a sequence of Lagrange multiplier tests applied to a linearized version of the model. However, adapting this approach to mixture of models is not trivial.

We will introduce a specification algorithm based on two ICs: Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). Both criteria have been used to select the number of experts; see, among others, Carvalho and Tanner (2005a) and Wong and Li (1999,2000). The two criteria are defined as

$$BIC = -2\sum_{t=1}^{T}\log f(y_t|\mathbf{x}_t;\widehat{\theta}) + M\log T \tag{7}$$

$$AIC = -2\sum_{t=1}^{T}\log f(y_t|\mathbf{x}_t;\widehat{\theta}) + 2M, \tag{8}$$

where $M = 3\#\mathbb{J} + (p+2)\#\mathbb{T}$ is the number of estimated parameters.

It is known that, for well behaved models, BIC is consistent for model selection. Furthermore, when the sample size goes to the infinity, the true model will be selected because it has the smallest BIC with probability tending to one. However, when the model is overidentified, the usual regularity conditions to support this result fail, but Wood, Jiang and Tanner (2001) present some evidence that, even when we have overidentified models, the BIC may still be consistent for model selection.

$\mathbb{J}$ and $\mathbb{T}$ define the tree $\mathbb{JT}$ with $\#\mathbb{T}$ local models and let $k \in \mathbb{T}$ be a node to be split. When we split the node k we have the new tree $\mathbb{JT}^{(k)}$ defined by $\mathbb{J}^{(k)}$ and $\mathbb{T}^{(k)}$, where

$$\mathbb{J}^{(k)} = \mathbb{J} \cup \{k\} \tag{9}$$

$$\mathbb{T}^{(k)} = \{2k+1, 2k+2\} \cup (\mathbb{T} \setminus \{k\}), \tag{10}$$

where $\mathbb{T} \setminus \{k\}$ is the complement of $\{k\}$ in $\mathbb{T}$. The new parameter vector $\boldsymbol{\theta}^{(k)}$ is defined as

$$\boldsymbol{\theta}^{(k)} = \left[\boldsymbol{\nu}'_{j_1}, \ldots, \boldsymbol{\nu}'_{j_{\#\mathbb{J}^{(k)}}}, \psi'_{t_1}, \ldots, \psi'_{t_{\#\mathbb{T}^{(k)}}}\right]', \tag{11}$$

where $j_i \in \mathbb{J}^{(k)}$ and $t_i \in \mathbb{T}^{(k)}$.

The growing algorithm for the first split is the following: (1) set the number of covariates as p and estimate a linear model with all p regressors and compute the value of the IC; (2) for each covariate $x_{s_0} \in \mathbf{x}_t$ with $s_0 = 1, \ldots, p$, estimate the model $\mathbb{JT}^{(0)}$, where each terminal node is a linear model with all p regressors, and compute the IC; and (3) select the model with the smallest IC.

The growing algorithm for the k-th split is: (1) for each $k \in \mathbb{T}$ and for each $x_{s_i} \in \mathbf{x}_t$ with $s_i = 1, \ldots, p$, (a) split the node k following (9) and (10), (b) estimate the new parameter vector $\boldsymbol{\theta}^{(k)}$ as in (11), and (c) compute the IC; (2) select the tree $\mathbb{JT}^{(k)}$ with the smallest IC; (3) if the smallest IC for the tree $\mathbb{JT}^{(k)}$ is greater than the IC of the tree $\mathbb{JT}$, then we stop growing the tree. Case contrary, repeat the steps above setting $\mathbb{JT} = \mathbb{JT}^{(k)}$.

6. MONTE-CARLO STUDY

Consider the following models:

- Model 1: a linear AR(2) model.

$$y_t = 1.0 + 0.5y_{t-1} - 0.2y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{NID}(0, 1.0).$$

- Model 2: a Tree-MM model with two AR(4) local models, $\mathbb{J} = \{0\}$, $\mathbb{T} = \{1, 2\}$, and $g_0(\cdot; \gamma_0, c_0) = g(y_{t-4}; 2, 3)$. The local models are:

$$\begin{aligned} y_{1t} &= 2.0 - 0.1y_{t-1} + 0.7y_{t-2} + 0.2y_{t-4} + \varepsilon_{1t}, \quad \varepsilon_{1t} \sim \mathsf{NID}(0, 1.0) \\ y_{2t} &= 2.0 + 0.2y_{t-1} - 0.6y_{t-2} + 0.3y_{t-3} - 0.3y_{t-4} + \varepsilon_{2t}, \quad \varepsilon_{2t} \sim \mathsf{NID}(0, 0.6). \end{aligned}$$

- Model 3: a Tree-MM model with three local AR(2) models, $\mathbb{J} = \{0, 2\}$, $\mathbb{T} = \{1, 5, 6\}$, $g_0(\cdot; \gamma_0, c_0) = g(y_{t-2}; 2, 1)$, $g_2(\cdot; \gamma_2, c_2) = g(y_{t-1}; 2, 4)$, and

$$\begin{aligned} y_{1t} &= 0.5 - 0.4y_{t-1} + 0.7y_{t-2} + \varepsilon_{1t}, \quad \varepsilon_{1t} \sim \mathsf{NID}(0, 1.0) \\ y_{5t} &= 4.0 + 0.8y_{t-1} - 0.5y_{t-2} + \varepsilon_{5t}, \quad \varepsilon_{5t} \sim \mathsf{NID}(0, 0.6) \\ y_{6t} &= 8.0 - 0.9y_{t-1} + 0.2y_{t-2} + \varepsilon_{6t}, \quad \varepsilon_{6t} \sim \mathsf{NID}(0, 1.1). \end{aligned}$$

- Model 4: a Tree-MM model with four local AR(2) models, $\mathbb{J} = \{0, 1, 2\}$, $\mathbb{T} = \{3, 4, 5, 6\}$, $g_0(\cdot; \gamma_0, c_0) = g(y_{t-2}; 1, 1)$, $g_1(\cdot; \gamma_1, c_1) = g(y_{t-1}; 3, 0)$, $g_2(\cdot; \gamma_2, c_2) = g(y_{t-1}; 2, 4)$, and

$$\begin{aligned} y_{3t} &= 0.7y_{t-1} - 0.3y_{t-2} + \varepsilon_{3t}, \quad \varepsilon_{3t} \sim \mathsf{NID}(0, 0.7) \\ y_{4t} &= -0.5 - 0.4y_{t-1} + 0.7y_{t-2} + \varepsilon_{4t}, \quad \varepsilon_{4t} \sim \mathsf{NID}(0, 1.0) \\ y_{5t} &= 4.0 + 0.8y_{t-1} - 0.5y_{t-2} + \varepsilon_{5t}, \quad \varepsilon_{5t} \sim \mathsf{NID}(0, 0.6) \\ y_{6t} &= 8.0 - 0.9y_{t-1} + 0.2y_{t-2} + \varepsilon_{6t}, \quad \varepsilon_{6t} \sim \mathsf{NID}(0, 1.1). \end{aligned}$$

Models 1–4 are used to evaluate the small sample properties of the QMLE and the modeling cycle strategy. The results are presented in the following subsections.

6.1. **Parameter estimation.**

We present the empirical results of the estimation of the parameters of Models 2–4. We report the mean, the median, the standard deviation, and the median absolute deviation around the median (MAD) across 2000 replications. The MAD is defined as $MAD(\widehat{\theta}) = \text{median}\left(\left|\theta - \text{median}(\widehat{\theta})\right|\right)$.

We have simulated Models 2–4 with two different sample sizes: 150 and 500 observations. Tables 1–3 show estimation results for each model. From the tables, it is clear that the estimation turns to be rather precise, with the only exception of the slope parameter γ, which is usually overestimated. This overestimation were noticed in Medeiros and Veiga (2005), and it is caused by the lack of observations around the transition location.

TABLE 1. SIMULATED MODEL 2: DESCRIPTIVE STATISTICS OF THE ESTIMATES.

The table shows the mean, the standard deviation (SD), the median, and the median absolute deviation (MAD) of the estimates of the parameters of Model 2 over 2000 simulations. 150 and 500 observations are considered.

		150				500			
	Actual	Mean	SD	Median	MAD	Mean	SD	Median	MAD
γ_0	2.00	5.06	1.75	4.68	0.89	5.01	1.65	4.66	0.83
c_0	3.00	3.01	0.23	3.02	0.15	3.00	0.22	3.00	0.15
σ_1^2	1.00	0.95	0.14	0.94	0.09	0.95	0.13	0.95	0.09
β_{01}	2.00	2.00	0.11	2.00	0.07	2.00	0.10	2.00	0.07
β_{11}	-0.10	-0.10	0.04	-0.10	0.03	-0.10	0.04	-0.10	0.03
β_{21}	0.70	0.70	0.04	0.70	0.02	0.70	0.04	0.70	0.02
β_{31}	0	0.00	0.05	0.00	0.03	0.00	0.04	0.00	0.03
β_{41}	0.20	0.20	0.06	0.20	0.04	0.20	0.06	0.20	0.04
σ_2^2	0.60	0.31	0.08	0.31	0.05	0.30	0.08	0.31	0.05
β_{02}	2.00	1.99	0.36	1.99	0.23	2.01	0.34	2.01	0.20
β_{12}	0.20	0.20	0.05	0.20	0.03	0.02	0.72	0.22	0.08
β_{22}	-0.60	-0.60	0.06	-0.60	0.04	-0.60	0.06	-0.60	0.03
β_{32}	0.30	0.03	0.05	0.30	0.03	0.03	0.04	0.30	0.03
β_{42}	-0.30	-0.30	0.06	-0.30	0.04	-0.30	0.04	-0.30	0.04

TABLE 2. SIMULATED MODEL 3: DESCRIPTIVE STATISTICS OF THE ESTIMATES.

The table shows the mean, the standard deviation (SD), the median, and the median absolute deviation (MAD) of the estimates of the parameters of Model 3 over 2000 simulations. 150 and 500 observations are considered.

		150				500			
	Actual	Mean	SD	Median	MAD	Mean	SD	Median	MAD
γ_0	2.00	5.41	2.25	4.98	1.25	4.99	1.17	4.86	0.66
c_0	1.00	0.97	0.35	0.97	0.18	0.94	0.20	0.94	0.12
σ_1^2	1.00	0.87	0.44	0.87	0.18	0.97	0.15	0.96	0.10
β_{01}	0.50	0.63	0.98	0.51	0.19	0.51	0.15	0.40	0.09
β_{11}	-0.40	-0.39	0.27	-0.41	0.08	-0.40	0.06	-0.40	0.04
β_{21}	0.70	0.62	0.48	0.66	0.10	0.68	0.08	0.69	0.05
γ_2	2.00	5.47	2.48	4.89	1.36	5.21	1.64	4.85	0.88
c_2	4.00	3.92	0.42	3.95	0.14	3.92	0.2	3.93	0.13
σ_5^2	0.60	0.34	0.09	0.33	0.05	0.35	0.04	0.35	0.03
β_{05}	4.00	3.99	0.24	4.00	0.14	4.00	0.11	4.00	0.07
β_{15}	0.80	0.80	0.06	0.80	0.04	0.80	0.03	0.80	0.02
β_{25}	-0.50	-0.50	0.05	-0.50	0.04	-0.50	0.03	-0.50	0.02
σ_6^2	1.10	1.13	0.34	1.11	0.18	1.21	0.15	1.20	0.10
β_{06}	8.00	8.09	1.48	1.09	0.79	8.07	0.62	8.06	0.39
β_{16}	-0.90	-0.90	0.21	-0.90	0.11	-0.90	0.09	-0.90	0.05
β_{26}	0.20	0.17	0.22	0.18	0.11	0.18	0.09	0.18	0.06

6.2. **Specification Algorithm.** We simulate 200 replications of Models 1–4 with two sample sizes: 150 and 500 observations. Table 4 presents the number of times a model is correctly (C)/incorrectly (I) specified. We define the model to be correctly specified if the sets $\mathbb{J}$, $\mathbb{T}$ and $\mathbb{S} = \{s_0, \ldots, s_{\#\mathbb{J}}\}$ are equal to the true sets $\mathbb{J}_0$, $\mathbb{T}_0$ and $\mathbb{S}_0$. The tree is incorrectly specified if any of these sets are different.

The BIC has a better performance then AIC in small and large samples. As expected, the performance of the modeling strategy improves as the sample sizes increases.

6.3. **Approximation Capabilities.** We illustrate the ability of the Tree-MM model to approximate unknown conditional probability density functions. We simulate two AR(1)-GARCH(1,1) models and two NN models. We generate 2000 observations, where the first 1000 are used for estimation and the remaining 1000 for out-of sample evaluation.

TABLE 3. SIMULATED MODEL 4: DESCRIPTIVE STATISTICS OF THE ESTIMATES.

The table shows the mean, the standard deviation (SD), the median, and the median absolute deviation (MAD) of the estimates of the parameters of Model 4 over 2000 simulations. 150 and 500 observations are considered.

		150				500			
	Actual	Mean	SD	Median	MAD	Mean	SD	Median	MAD
γ_0	1.00	3.41	1.62	3.07	0.67	3.38	1.14	3.16	0.52
c_0	1.00	0.98	0.42	0.94	0.19	0.96	0.23	0.94	0.12
γ_1	3.00	6.07	3.29	5.51	2.31	6.39	3.02	6.12	2.27
c_1	0.00	0.04	0.49	0.04	0.20	0.01	0.34	0.02	0.15
γ_2	2.00	5.44	3.04	4.69	2.02	5.13	2.86	4.43	1.91
c_2	4.00	3.48	0.87	3.68	0.40	3.59	0.67	3.74	0.29
σ_3^2	0.70	0.45	0.29	0.42	0.08	0.47	0.17	0.46	0.05
β_{03}	0.00	-0.02	0.35	-0.02	0.17	-0.02	0.17	-0.01	0.09
β_{13}	0.70	0.68	0.16	0.69	0.08	0.69	0.08	0.69	0.04
β_{23}	-0.30	-0.31	0.10	-0.31	0.05	-0.31	0.05	-0.31	0.03
σ_4^2	1.00	0.87	0.37	0.85	0.18	0.95	0.16	0.94	0.10
β_{04}	-0.50	-0.56	0.46	-0.57	0.28	-0.53	0.22	-0.53	0.15
β_{14}	-0.40	-0.40	0.17	-0.40	0.10	-0.40	0.08	-0.40	0.05
β_{24}	0.70	0.67	0.17	0.67	0.10	0.68	0.09	0.68	0.05
σ_5^2	0.60	0.33	0.14	0.31	0.06	0.34	0.05	0.34	0.03
β_{05}	4.00	4.00	0.19	4.00	0.11	4.00	0.08	4.00	0.05
β_{15}	0.80	0.80	0.05	0.80	0.03	0.80	0.02	0.80	0.02
β_{25}	-0.50	-0.50	0.05	-0.50	0.02	-0.50	0.02	-0.50	0.01
σ_6^2	1.10	1.14	0.67	1.04	0.31	1.28	0.37	1.23	0.18
β_{06}	8.00	7.94	2.17	8.00	1.08	7.81	1.08	7.91	0.57
β_{16}	-0.90	-0.86	0.37	-0.88	0.17	-0.84	0.18	-0.86	0.09
β_{26}	0.20	0.12	0.29	0.13	0.15	0.16	0.12	0.16	0.07

TABLE 4. SPECIFICATION ALGORITHM.

This table shows the number of cases where the each model is correctly (C)/incorrectly (I) specified. We consider two different samples: 150 and 500 observations. Both the AIC and BIC are used to select the structure of the models.

	150				500			
	AIC		BIC		AIC		BIC	
Model	C	I	C	I	C	I	C	I
1	163	37	172	28	197	3	200	0
2	107	93	134	66	193	7	196	4
3	83	117	96	104	150	50	166	34
4	57	143	81	119	123	77	135	65

We use the coverage test Christoffersen (1998) over a set of percentiles to evaluate the coverage. The test is applied to the out-of-sample period. The correlation and the mean squared error (MSE) of the one-step-ahead predictions are also used to compare the Tree-MM models with the true data generation process. The Christoffersen´s (1998) test consists in two likelihood ratio (LR) tests. The first one is the LR test of unconditional coverage and the second one the LR test of independence.

All the AR(1)-GARCH(1,1) models have the same linear part and distinct GARCH(1,1) conditional variances. The linear model is:

$$y_t = 0.7y_{t-1} + u_t,$$

where $u_t = h_t^{1/2}\epsilon_t$, $\epsilon_t \sim \mathsf{NID}(0,1)$, and

- Model 5: $h_t = 10^{-5} + 0.85h_{t-1} + 0.05u_{t-1}^2$;
- Model 6: $h_t = 10^{-5} + 0.90h_{t-1} + 0.085u_{t-1}^2$.

The simulated NN models are the following:

$$\begin{aligned} y_t &= 0.1 + 0.75y_{t-1} - 0.05y_{t-4} + 0.8g(0.45y_{t-1} - 0.89y_{t-4}; 2.24, -0.09) \\ &\quad -0.7g(0.44y_{t-1} + 0.89y_{t-4}; 1.12, -0.35) + u_t, \end{aligned}$$

where $u_t = h_t^{1/2}\epsilon_t$, $\epsilon_t \sim \mathsf{NID}(0,1)$, and

- Model 7: $h_t = 1$;
- Model 8: $h_t = 10^{-5} + 0.85h_{t-1} + 0.05\epsilon_{t-1}^2$.

We evaluate the conditional coverage over the following tail percentiles: 90%, 95%, 97.5% and 99%. Table 5 shows empirical coverage and the results of the Christoffersen´s (1998) test. LR_{uc} is the p-value of the unconditional coverage test and LR_{cc} is the p-value of the conditional coverage test. From the results, it is clear that the Tree-MM is able to model the tail of conditional distribution.

TABLE 5. EMPIRICAL COVERAGE.

The table shows the empirical coverage and the p-value of the Christoffersen test for the estimated AR(1)-GARCH(1,1) and NN-GARCH(1,1) models. LR_{uc} and LR_{cc} are the p-values of the unconditional conditional coverage tests, respectively.

		Empirical Coverage			
		90.0	95.0	97.5	99.0
Model 5	Est. Percentile	90.39	95.60	97.60	99.20
	LR_{uc}	0.695	0.389	1.000	0.534
	LR_{cc}	0.223	0.184	0.596	0.202
Model 6	Est. Percentile	90.76	95.51	97.10	99.00
	LR_{uc}	0.457	0.918	0.215	0.273
	LR_{cc}	0.542	0.228	0.118	0.273
Model 7	Est. Percentile	89.30	95.00	97.40	99.20
	LR_{uc}	0.3344	.6878	1.000	.5297
	LR_{cc}	0.2939	.6664	.3583	.2349
Model 8	Est. Percentile	90.60	95.40	97.10	99.00
	LR_{uc}	0.5352	0.5735	0.4385	1.000
	LR_{cc}	0.6199	0.6018	0.5317	0.000

Table 6 compares the out-of-sample performance of the estimated Tree-MM model with the true NN specification. The correlation row shows the average correlation between the estimates, MSE_{NN} and $\text{MSE}_{\text{Tree-MM}}$ are the average out-of-sample MSE for the NN and Tree-MM models, respectively. From the results in the tables we can see that the correlation between the estimates are high and the MSEs are very close for both models, showing the approximation capabilities of the Tree-MM models.

TABLE 6. FORECASTING PERFORMANCE RESULTS.

The table shows the forecasting results and the correlation between the true data generating process and the estimated Tree-MM model.

Model	Correlation	MSE_{NN}	$\text{MSE}_{\text{Tree-MM}}$
Model 7	0.88	0.0223	0.0242
Model 8	0.74	2.25×10^{-3}	2.81×10^{-3}

7. EXAMPLES

7.1. **Example 1: Canadian Lynx.** The first set analyzed is the 10-based logarithm of the annual record of the numbers of Canadian Lynx trapped in the Mackenzie River district of north-west Canada for the period 1821–1934 (114 observations). For further details and background history, see Tong (1990).We report only the results for in-sample fitting because the number of observations is rather small and most of the previous studies in literature have only considered the in-sample analysis. It is commonly accepted that the data are cyclical, with a period of 9-10 years and multimodality.

The variables are selected following the methodology of Rech, Teräsvirta and Tschernig (2001). The final model using either AIC or BIC is a 1-split tree, where the transition variable is the y_{t-2} and

$$\begin{aligned} g_0(\cdot) &= g(y_{t-2}; 9.9826, 2, 3.2655), \\ y_{1t} &= 0.5465 + 1.319 y_{t-1} - 0.4655 y_{t-2} + \hat{\varepsilon}_{1t} \quad \hat{\varepsilon}_{1t} \sim N(0, 0.0325), \\ y_{2t} &= 0.9892 + 1.5173 y_{t-1} - 0.8832 y_{t-2} + \hat{\varepsilon}_{2t} \quad \hat{\varepsilon}_{2t} \sim N(0, 0.0493). \end{aligned}$$

We compare the Tree-MM model with the following alternatives: an AR(2) model; the SETAR model of Tong (1990); the MAR model of Wong and Li (2000); and the GMAR model of Wong and Li (1999). The models MAR and GMAR have a mixture of Gaussian models as the conditional density and the others have a Gaussian conditional density. The final Tree-MM model has the same number of regimes and the same transition variable as the models SETAR and GMAR.

All the models have similar empirical coverage. However, it terms of the conditional mean fit, the Tree-MM model attains the lowest mean absolute error (MAE).

7.2. **Example 2: Brazilian Financial Dataset.** In this section we apply the Tree-MM model to automatic trading using data from the Brazilian stock exchange. We compare the

TABLE 7. EXAMPLE 1: EMPIRICAL COVERAGE.

The table shows the empirical coverage as well as the mean absolute error (MAE) for a set of different models.

	Empirical Coverage						
Model	50	60	70	80	90	95	MAE
AR(2)	50.00	58.93	68.75	75.89	88.61	92.86	1.99
SETAR	44.86	56.07	69.16	81.31	90.65	95.33	2.27
MAR	52.68	63.39	70.54	82.14	88.39	96.43	2.36
GMAR	47.32	58.93	68.75	82.14	92.86	93.75	2.25
Tree-MM	48.21	57.14	67.86	79.46	89.29	96.43	1.89

results with an NN model estimated with Bayesian regularization (MacKay 1992), with the ARMA model, and the naïve method (the forecast for any period equals the previous period's actual value). We choose an asset which tracks the BOVESPA Index (IBOVESPA). IBOVESPA is an index of the 50 most liquid stocks traded at the São Paulo Stock Exchange. The selected asset is the Petrobras PN (PETR4) (Brazilian Oil Company). The observations cover the period from 01/20/1999 to 12/30/2004 (1476 observations). The sample is divided into two groups. The first one consists of 1227 observations (from 01/20/1999 to 12/30/2003) and is used to estimate the model. The second group consists of 249 observations (from 02/02/2004 to 12/30/2004) and it is used for out-of-sample evaluation.

The set of possible covariates is composed by the first 10 lags of the log-return of the asset, the first 10 lags of the volatility, the first 10 lags of the traded volume between 2 days, the first difference of the 10- and 20-days moving averages of the return (MA10 and MA20, respectively), and the first difference of the following 10 exogenous variables: IBOVESPA, S&P 500 Index (S&P), US Dollar exchange rate (DOL), 10-year Treasury bill (T10), C-Bond (C-BOND), the spread between C-Bond and T10 (SOT), Oil price (OIL), Swap 360 (SW360), a set of commodities (CRY) and the Developing Countries Stock Index (BINDEX).

The statistical measures used to evaluate the model are the mean absolute error (MAE), the root mean square error ($RMSE$), and the correct direction of change (CDC). The financial measures are the average return ($\bar{R}$), the annual return (R^A), the accumulate return (R^C), the annual volatility (σ^A), the Sharpe index (SR), the number of trades ($\#T$), and the percentage of winning trades (WT). Furthermore, we present the coverage of the model and the statistics of the coverage for the NN and Tree-MM models. The trading strategy is the following. We sell the stock every time the forecasted return is negative and we buy when the forecasted return is positive. Table 8 shows the results.

We first select the set of regressors using the procedure proposed by Rech et al. (2001). The final Tree-MM model is given by:

$$\begin{aligned}
g_0(\cdot) &= g(v_{t-1}; 5.2572, 2, 0.0318), \\
y_{1t} &= -7.9906 \times 10^{-4} - 0.0542 y_{t-1} + 0.0775 v_{t-1} + 5.0897 \times 10^{-4} q_{t-1} \\
&\quad -3.6653 \times 10^{-6} MA10 + 0.0180 CRY + \varepsilon_{1t} \quad \varepsilon_{1t} \sim N(0, 1.9374 \times 10^{-4}), \\
y_{2t} &= -6.956 \times 10^{-3} + 0.3382 y_{t-1} + 0.1673 v_{t-1} + 1.5762 \times 10^{-3} q_{t-1} \\
&\quad -1.6057 \times 10^{-5} MA10 + 0.0167 CRY + \varepsilon_{2t} \quad \varepsilon_{2t} \sim N(0, 6.5584 \times 10^{-4}).
\end{aligned}$$

The estimated NN model has two hidden units and uses the whole set of variables. The Tree-MM model, the NN model and the linear model have similar performance accordind to the statistical measures. However, the financial measures indicate that the Tree-MM model has the best performance among the competing models.

8. Conclusions

In this paper we proposed a new mixture of models based on smooth transition regression trees. A quasi-maximum likelihood estimator was developed and its asymptotic properties were derived under mild regularity conditions. A model building strategy was also considered. Monte Carlo simulations gave strong support for the theory developed here, even in small samples. Two applications with real data were used to illustrate the model.

TABLE 8. STATISTICAL AND FINANCIAL RESULTS

This table shows the mean absolute error (MAE), the root mean square error ($RMSE$), the correct direction of change (CDC), and the average $\bar{R}$), the annual (R^A), and the accumulated (R^C) returns, respectively. The table also shows the annual volatility (σ^A), the Sharpe index (SR), the number of trades ($\#T$), and the percentage of winning trades (WT).

	ARMA	Naïve	NN	TREE-MM
MAE	0.012	0.016	0.012	0.012
RMSE	0.017	0.022	0.017	0.017
CDC	60.48	58.07	65.73	62.50
$\bar{R}$	0.68	0.50	1.65	1.45
R^A	41.64	26.38	60.47	61.69
R^C	40.98	25.96	59.51	60.71
σ^A	24.45	22.59	18.47	18.31
SR	1.70	1.17	3.27	3.37
$\#T$	60	52	36	42
WT	55.00 %	46.67 %	75.00 %	76.19%

APPENDIX A. IDENTIFIABILITY

Let $\mathbb{JT}$ be a tree with sets $\mathbb{J}$, $\mathbb{T}$ and $\mathbb{S}$, where $\mathbb{S}$ is the set of indexes s_j, $\forall j \in \mathbb{J}$ and parameter vector $\boldsymbol{\theta}$. We define a subtree $\mathbb{JT}^k$ as the tree beginning at node k, with the sets $\mathbb{J}^k \subseteq \mathbb{J}$, $\mathbb{T}^k \subseteq \mathbb{T}$ and $\mathbb{S}^k \subseteq \mathbb{S}$, where $i \in \mathbb{JT}^k \Leftrightarrow k \in \mathbb{J}_i \cup \{i\}$ and parameter vector $\boldsymbol{\theta}^k$. For example, assume the tree $\mathbb{JT} = \{0, 1, 2, 3, 4, 5, 6, 11, 12\}$ then $\mathbb{JT}^2 = \{2, 5, 6, 11, 12\}$.

ASSUMPTION 7. *Let $f_k(y_t|\mathbf{x}_t; \boldsymbol{\theta}_k)$ be the conditional p.d.f. of the subtree $\mathbb{JT}^k$. Then $\forall k \in \mathbb{J}$, $f_{2k+1}(y_t|\mathbf{x}_t; \boldsymbol{\theta}^{2k+1}) \neq f_{2k+2}(y_t|\mathbf{x}_t; \boldsymbol{\theta}^{2k+2})$.*

This assumption guarantees that our tree is irreducible in the sense that any split cannot be changed by a subtree or by a local model.

ASSUMPTION 8. *We assume that for any tree $\mathbb{JT}$ and all sub-trees $\mathbb{JT}^k$: (1) $\gamma_j > 0, \forall j \in \mathbb{J}$; (2) $\forall j \in \mathbb{J}^{2k+1}$, if $s_j = s_k$ then $c_j < c_k$; (3) $\forall j \in \mathbb{J}^{2k+2}$, if $s_j = s_k$ then $c_j \geq c_k$.*

These assumptions together ensure that the sets $\mathbb{J}$, $\mathbb{T}$ and $\mathbb{S}$ uniquely specify any tree.

LEMMA 1. *Under Assumptions (7) and (8), a tree $\mathbb{JT}$ is uniquely specified and the parameter vector $\boldsymbol{\theta}$ has a unique representation.*

PROOF. Suppose that for any node $k \in \mathbb{J}$, $f_{2k+1}(y_t|\mathbf{x}_t; \boldsymbol{\theta}^{2k+1}) = f_{2k+2}(y_t|\mathbf{x}_t; \boldsymbol{\theta}^{2k+2})$ such that $f_k = g_k(\cdot)f_{2k+1} + (1 - g_k(\cdot))f_{2k+2} = f_{2k+1} = f_{2k+2}$. Hence, we can change the node k by the node $2k + 1$ or $2k + 2$. If $f_{2k+1}(\cdot) \neq f_{2k+2}(\cdot)$, $\forall k \in \mathbb{J}$, then the tree cannot be reduced, so it is irreducible.

Now, suppose there is an irreducible tree $\mathbb{JT}$. On the first split at s_0, c_0 can assume any value in $\mathbb{R}$. Now consider the sub-trees $\mathbb{JT}^1$ and $\mathbb{JT}^2$. Following the condition (8), on the next split at $s_k = s_0, k \in \mathbb{J}^1$, c_k can assume any value in $(-\infty, c_0)$ and on the next split at $s_l = s_0, l \in \mathbb{J}^2$, c_l can assume any value in $[c_0, \infty)$. So, the values of c_k and c_l cannot be interchanged. Repeating this argument for all splits, and considering that the transition has the same shape (which is guaranteed by the constraint over the γs), we show that any irreducible tree under Assumption (8) is uniquely specified.

Q.E.D.

The next theorem gives the conditions under which the Tree-MM model is identifiable.

THEOREM 4. *Under Assumptions (7) and (8), and assuming that $\pi(y_t|\mathbf{x}_t; \boldsymbol{\psi})$ is uniquely identified by a parameter vector $\boldsymbol{\psi}$, model (2) is identifiable, in the sense that, for a sample $\{y_t; \mathbf{x}_t\}_{t=1}^T$, and for $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2 \in \Theta$*

$$\prod_{t=1}^{T} f(y_t|\mathbf{x}_t; \boldsymbol{\theta}_1) = \prod_{t=1}^{T} f(y_t|\mathbf{x}_t; \boldsymbol{\theta}_2) \quad , a.s.$$

is equivalent to $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

PROOF. Suppose that $f(y_t|\mathrm{x}_t;\boldsymbol{\theta}_1) = f(y_t|\mathrm{x}_t;\boldsymbol{\theta}_2)$, for any sequence $\{y_t;\mathrm{x}_t\}_{t=1}^{T}$. Therefore,

$$\sum_{i\in\mathbb{T}_1} B_i(x_{s_i};\boldsymbol{\theta}_{1i})\pi(y_t|\mathrm{x}_t;\boldsymbol{\psi}_{1i}) = \sum_{i\in\mathbb{T}_2} B_i(x_{s_i};\boldsymbol{\theta}_{2i})\pi(y_t|\mathrm{x}_t;\boldsymbol{\psi}_{2i}). \tag{A.1}$$

According to Lemma 1, $\mathbb{T}_1 = \mathbb{T}_2 = \mathbb{T}$. Furthermore, if $\boldsymbol{\psi}_{1i} = \boldsymbol{\psi}_{2i}$ then $\pi(y_t|\mathrm{x}_t;\boldsymbol{\psi}_{1i}) = \pi(y_t|\mathrm{x}_t;\boldsymbol{\psi}_{2i})$ and $\sum_{i\in\mathbb{T}}(B_i(\cdot;\boldsymbol{\theta}_{1i}) - B_i(\cdot;\boldsymbol{\theta}_{2i}))\pi(y_t|\mathrm{x}_t;\boldsymbol{\psi}_i) = 0$, where $\boldsymbol{\psi}_i = \boldsymbol{\psi}_{1i} = \boldsymbol{\psi}_{2i}$.

We have to show that $B_i(\cdot;\boldsymbol{\theta}_{1i}) - B_i(\cdot;\boldsymbol{\theta}_{2i}) = 0$. Following the definition of $B_i(\cdot;\cdot)$ in (3) and the definition of the logistic function, we can write $B_i(\cdot;\cdot)$ as a product of logistic functions. Hence, $g_0(\cdot;\boldsymbol{\nu}_{10})\prod_{k\in\mathbb{J}_i} g_k(\cdot;\boldsymbol{\nu}_{1k}) = g_0(\cdot;\boldsymbol{\nu}_{20})\prod_{k\in\mathbb{J}_i} g_k(\cdot;\boldsymbol{\nu}_{2k})$.

If we show $g_0(\cdot;\boldsymbol{\nu}_{10}) = g_0(\cdot;\boldsymbol{\nu}_{20})$, then we can show iteratively that $B_i(\cdot;\boldsymbol{\theta}_{1i}) = B_i(\cdot;\boldsymbol{\theta}_{2i})$, $g_0(\cdot;\boldsymbol{\nu}_{10}) = g_0(\cdot;\boldsymbol{\nu}_{20})$, and $\frac{1}{1+e^{-\gamma_{10}(x_{s_0,t}-c_{10})}} = \frac{1}{1+e^{-\gamma_{20}(x_{s_0,t}-c_{20})}}$, which is true only if $(\gamma_{10}, c_{10}) = (\gamma_{20}, c_{20})$.

Concluding, we have shown that $f(y_t|\mathrm{x}_t;\boldsymbol{\theta}_1) = f(y_t|\mathrm{x}_t;\boldsymbol{\theta}_2)$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

Q.E.D.

APPENDIX B. STATIONARITY AND GEOMETRIC ERGODICITY

It is important to know under which conditions a Tree-MM process with only autoregressive local models is stationary. Some results on the stability of mixture of experts were shown in Zeevi et al. (1998) for the case of a MixAR$(m;d)$ model. We can use similar results because the behavior of the multinomial logistics and the $B(\cdot)$ functions are equivalent.

Set $\alpha_k \equiv \max_{i\in\mathbb{T}}|\beta_{ik}|$, $k = 1,\ldots,p$, where β_{ik} is the k-th component of $\boldsymbol{\beta}_i$.

THEOREM 5. *Let* $\{y_t\}_{t\geq 0}$ *follow a Tree-MM model (2) with AR(*p*) local models. Assume that the polynomial*

$$\mathcal{P}(z) = z^d - \sum_{k=1}^{p}\alpha_k z^{d-k} \quad ; \; z \in C$$

has all its zeros in the open unit disk, $z < 1$. Then the vector process y_t has a unique stationary probability measure, and is geometrically ergodic.

PROOF. To use the results of Zeevi et al. (1998), we need to show some similarities between the multinomial logit and the $B(\cdot)$ functions. Set $B^{(1)}$ as the left most expert of the tree and $B^{(J)}$ as the right most expert. $B^{(1)}$ is a product of $1 - g(\cdot)$ functions and $B^{(J)}$ is a product of $g(\cdot)$ functions. Any $B^{(j)}$ for $j = 2, \dots, J-1$ has at least one term $g(\cdot)$ and one term $1 - g(\cdot)$. We can show the equivalence of the proofs if we satisfy the following conditions: (i) $B^{(1)} \to 1$ for $y_{s^{(1)}} \to -\infty$; (ii) $B^{(1)} \to 0$ for $x_{s^{(1)}} \to \infty$; (iii) $B^{(J)} \to 1$ for $x_{s^{(J)}} \to \infty$; (iv) $B^{(J)} \to 0$ for $x_{s^{(J)}} \to -\infty$; and (v) $B^{(j)} \to 0$ for $x_{s^{(j)}} \to \pm\infty$.

We know that $g(\mathbf{x}_t, \boldsymbol{\nu}_k) \to 1$ for $x_{s_k} \to \infty$ and $g(\mathbf{x}_t, \boldsymbol{\nu}_k) \to 0$ for $x_{s_k} \to -\infty$. Thus, $[1 - g(\mathbf{x}_t, \boldsymbol{\nu}_k)] \to 0$ for $x_{s_k} \to \infty$, $[1 - g(\mathbf{x}_t, \nu_k)] \to 1$ for $x_{s_k} \to -\infty$, and

$$\lim_{x_{s^{(1)}} \to -\infty} B^{(1)}(\cdot) = \lim_{x_{s^{(1)}} \to -\infty} \prod [1 - g(\cdot)] = \prod \lim_{x_{s^{(1)}} \to -\infty} [1 - g(\cdot)] = 1,$$

such that Condition (i) holds. Conditions (ii)–(v) can be verified using the same steps.

Q.E.D.

APPENDIX C. PROOFS OF THEOREMS

We follow White (1992), to prove the existence, consistency and asymptotic normality of the QMLE. Besides, we define some notation to make the proofs clearer.

Define $f_t \equiv f(y_t|\mathbf{x}_t; \theta)$, $f_t^* \equiv f(y_t|\mathbf{x}_t; \boldsymbol{\theta}^*)$, $\pi_{it} \equiv \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_i)$, $\pi_{it}^* \equiv \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_i^*)$, $B_{it} \equiv B_i(\mathbf{x}_t; \boldsymbol{\theta}_i)$, and $B_{it}^* \equiv B_i(\mathbf{x}_t; \boldsymbol{\theta}_i^*)$. Furthermore, define recursively $f_{k,t} = (1 - g(x_{s_k}; \boldsymbol{\nu}_k)) f_{2k+1,t} + g(x_{s_k}; \boldsymbol{\nu}_k) f_{2k+2,t}$, for all k in $\mathbb{J}$, and $f_{k,t} = \pi_{kt}$, for all k in $\mathbb{T}$.

C.1. **Proof of Theorem 1.** We need to satisfy Assumptions 2.1, 2.3 and 2.4 of Theorem 2.13 in White (1992) and show that $|\mathcal{L}_T(\boldsymbol{\theta})| < \infty$ with $\boldsymbol{\theta}^*$ being the unique maximum of $\mathcal{L}_T(\boldsymbol{\theta})$. Assumption 2.1 is satisfied by Assumption 1, and Assumption 2.3 is satisfied by

Assumption 2 and Lemma 2. Assumption 2.4 and $|\mathcal{L}(\boldsymbol{\theta})| < \infty$ are satisfied by Lemma 2. So we need to show that $\mathcal{L}_T(\boldsymbol{\theta})$ has a unique maximum at $\boldsymbol{\theta}^*$.

First, we write the maximization problem as follows:

$$\max_{\boldsymbol{\theta}\in\Theta} [\mathcal{L}_T(\boldsymbol{\theta}) - \mathcal{L}_T(\boldsymbol{\theta}^*)] = \max_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\left[\log\frac{f_t^*}{f_t} - \frac{f_t}{f_t^*} - 1\right].$$

Furthermore, for any $x > 0$, $m(x) = x - log(x) \leq 0$, then

$$\mathbb{E}\left[\log\frac{f_t^*}{f_t} - \frac{f_t}{f_t^*}\right] \leq 0.$$

Given that $m(x)$ archives its maximum at $x = 1$, $\mathbb{E}[m(x)] \leq \mathbb{E}[m(1)]$ with equality holding almost surely only if $f_t^* = f_t$ with probability one. By the mean value theorem, it is equivalent to show that

$$(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'\frac{\partial \log f_t}{\partial \boldsymbol{\theta}} = 0 \tag{C.2}$$

almost surely. A straightforward application of Lemma 3 shows that it happens if, and only if, $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ with probability one, which completes the proof.

Q.E.D.

C.2. **Proof of Theorem 2.** We must satisfy Assumptions 2.1, 2.3, 2.4, 3.1, and 3.2' of Theorem 3.5 in White (1992). Assumptions 2.1, 2.3, 2.4 and 3.2' are satisfied by Assumptions 1–3. Assumption 3.1 states that: (a) $\mathbb{E}_{\mathcal{G}}(\log f_t) < \infty, \forall t$; (b) $\mathbb{E}_{\mathcal{G}}(\log f_t)$ is continuous in **Θ**; and (c) $\{\log f_t\}$ obeys the uniform law of large numbers (ULLN).

It is clear that $\mathbb{E}_{\mathcal{G}}(\log f_t) \leq \log \mathbb{E}_{\mathcal{G}}(f_t) \leq \log \mathbb{E}_{\mathcal{G}}(\sup_t f_t)$. But $\sup_t f_t = \Delta < \infty$, then $\log\left[\mathbb{E}_{\mathcal{G}}(\sup_t f_t)\right] = \log\Delta < \infty$ and (a) is satisfied. In addition, $\log(\cdot)$, $\mathcal{G}_t$ and f_t are continuous, measurable, and integrable functions, so $h_t = \mathcal{G}_t \log f_t$ is also continuous, measurable, and integrable. Then, $\int h_t dy$ is continuous and (b) is satisfied. Finally, (c) is satisfied by Lemma 8.

Q.E.D.

C.3. **Proof of Theorem 3.** We must satisfy Assumptions 2.1, 2.3, 3.1, 3.2', 3.6, 3.7(a), 3.8, 3.9 and 6.1 in White (1992). Assumptions 2.1, 2.3, 3.1, 3.2' are satisfied by Assumptions 1–6 (see proof of Theorem 2). Assumption 3.6 is satisfied by Lemma 2, Assumption 3.7(a) is satisfied by Lemma 5, Assumption 3.8 by Lemmas 6 and 8, Assumption 3.9 by Lemma 7, and Assumption 6.1 is shown here.

Assumption 6.1 requires that $\left\{T^{-1/2}\partial_\theta f_t|_{\boldsymbol{\theta}^*}\right\}$ obeys a central limit theorem with covariance matrix $B(\boldsymbol{\theta}^*)$, where $B(\boldsymbol{\theta}^*)$ is $O(1)$ and uniformly positive definite. We must show the following to satisfy Assumption 6.1: (a) $T^{-1}\sum_{t=1}^{T}\partial_\theta f_t|_{\theta^*}\partial_{\theta'}f_t|_{\theta^*} \overset{a.s.}{\rightarrow} \mathbb{E}(\partial_\theta f_t|_{\theta^*}\partial_{\theta'}f_t|_{\theta^*})$; (b) the sequence is strictly stationary.

Condition (a) is readily verified by Lemmas 8 and 5. Condition (b) is satisfied by Assumptions 4 and 6. Hence, satisfying these assumptions, the result follows.

Q.E.D.

APPENDIX D. LEMMAS

LEMMA 2. *Under Assumptions (2)–(3), $f(y_t|\mathrm{x}_t;\boldsymbol{\theta})$ is a measurable, limited, positive and continuously differentiable function of $\boldsymbol{Y}_t = [y_t, \mathrm{x}_t]'$ on $\boldsymbol{\Theta}$.*

PROOF. Trivially, $\pi(y_t|\mathrm{x}_t;\boldsymbol{\psi}_i)$ and $g(x_{s_jt};\boldsymbol{\nu}_j)$ are continuous, mensurable, finite, positive and differentiable functions of $\boldsymbol{Y}_t$. The function $f(y_t|\mathrm{x}_t;\boldsymbol{\theta})$ is a sequence of sums and products of these functions. As a result, $f(y_t|\mathrm{x}_t;\boldsymbol{\theta})$ is a continuous, mensurable, finite, positive and differentiable function of $\boldsymbol{Y}_t$.

Q.E.D.

LEMMA 3. *Let $\boldsymbol{d}$ be a constant vector with the same dimension of $\boldsymbol{\theta}$. Then, it follows that*

$$\boldsymbol{d}'\left(\frac{\partial \log f_t}{\partial \boldsymbol{\theta}}\right) = 0 \quad a.s.$$

if, and only if, $\boldsymbol{d} = \boldsymbol{0}$.

PROOF. First, write $\boldsymbol{d}'\left(\dfrac{\partial \log f_t}{\partial \boldsymbol{\theta}}\right) = \boldsymbol{d}'\dfrac{1}{f_t}\dfrac{\partial f_t}{\partial \boldsymbol{\theta}} = 0$. From Lemma 2, we know that $f_t > 0$. Hence,

$$\boldsymbol{d}'\frac{\partial \pi_{it}}{\partial \boldsymbol{\psi}_i} = 0 \quad \text{and} \quad [f_{2k+1,t}g_{kt} - f_{2k+2,t}(1-g_{kt})]\boldsymbol{d}'\frac{\partial[-\gamma_k(y_t - c_k)]}{\partial \boldsymbol{\nu}_k} = 0,$$

which are both functions of y_t. By the non-degeneracy condition, and supposing that y_t is not null for all t, $\boldsymbol{d}'\frac{\partial f_t}{\partial \boldsymbol{\theta}} = 0$ if, and only if, $\boldsymbol{d} = \boldsymbol{0}$.

Q.E.D.

LEMMA 4. *Under Assumptions 2, 5 and 6,* $\mathbb{E}(\log f_t) < \infty$.

PROOF. Write $\log f_t = \log \sum_{i \in \mathbb{T}} B_{it}\pi_{it} < \log \sum_{i \in \mathbb{T}} \pi_{it} < \log \#\mathbb{T} + \log \sup_{i \in \mathbb{T}} \pi_{it}$. Let $\pi_{It} = \pi(y_t|\mathbf{x}_t;\mathbf{x}_t'\boldsymbol{\beta}_I, \sigma_I^2) = \sup_{i \in \mathbb{T}} \pi_{it}$. Then, $\log \pi_{It} = -\frac{1}{2}\log 2\pi\sigma_I^2 - \frac{1}{2\sigma_I^2}(\boldsymbol{x}_t'\boldsymbol{\beta}_I - y_t)^2$. Under Assumptions 2 and 5, $\mathbb{E}\left[\log \pi_{It}\right] = -\frac{1}{2}\log 2\pi\sigma_I^2 - \frac{1}{2\sigma_I^2}\mathbb{E}\left[(\boldsymbol{x}_t'\boldsymbol{\beta}_I - y_t)^2\right] < \infty$.

Q.E.D.

LEMMA 5. *Under Assumptions 2, 4, 5 and 6,*

$$\mathbb{E}\left(\frac{\partial \log f_t}{\partial \boldsymbol{\theta}}\right) < \infty \quad \textit{and} \quad \mathbb{E}\left(\frac{\partial \log f_t}{\partial \boldsymbol{\theta}}\frac{\partial \log f_t}{\partial \boldsymbol{\theta}'}\right) < \infty.$$

PROOF. Let $\partial_\theta \equiv \frac{\partial}{\partial \boldsymbol{\theta}}$. It is clear that

$$\partial_\theta \log f_t = \frac{1}{f_t}\partial_\theta f_t = \frac{1}{f_t}\sum_{i \in \mathbb{T}} \pi_{it}\partial_\theta B_{it} + B_{it}\partial_\theta \pi_{it} \leq \Delta_{\pi,f}\sum_{i \in \mathbb{T}} \partial_\theta B_{it} + \Delta_B \sum_{i \in \mathbb{T}} \partial_\theta \pi_{it}, \tag{D.3}$$

where $\Delta_{\pi,f} = \sup_i(f_t^{-1}\pi_{it}) < \infty$ and $\Delta_B = \sup_i f_t^{-1} < \infty$. Set $\partial_{\psi_i} \equiv \partial/\partial\psi_i$, $\partial_{\nu_j} \equiv \partial/\partial\nu_j$, and$\Delta_\pi = \sup_i \pi_{it}$. Hence,

$$\partial_{\psi_i}\pi_{it} = \pi_{it}\partial_{\psi_i}\log \pi_{it} \leq \Delta_\pi \partial_{\psi_i}\log \pi_{it}, \tag{D.4}$$

$$\partial_{\nu_j} B_{it} = B_{it}(-g_{jt})(1-g_{jt})\partial_{\nu_j}[-\gamma_j(x_{s_j} - c_j)] \leq \left|\partial_{\nu_j}[-\gamma_j(x_{s_j} - c_j)]\right|. \tag{D.5}$$

As $\boldsymbol{\psi}_i = [\beta_{0i}, \ldots, \beta_{pi}, \sigma_i^2]'$, the right size of (D.4) can be written as

$$\Delta_\pi \partial_{\beta_{ki}} \log \pi_{it} = -\Delta_\pi \frac{\tilde{x}_{kt}(\tilde{\mathbf{x}}_t'\boldsymbol{\beta} - y_t)}{\sigma_i^2}, \tag{D.6}$$

$$\Delta_\pi \partial_{\sigma_i^2} \log \pi_{it} = \Delta_\pi \left(-\frac{1}{2\sigma_i^2} + \frac{(\tilde{\mathbf{x}}_t'\boldsymbol{\beta} - y_t)^2}{2\sigma_i^4} \right), \tag{D.7}$$

where $\tilde{x}_{kt}$ is the k-th element of the vector $\tilde{\mathbf{x}}_t$.

Using the same argument, we can write the right side of equation (D.5) as

$$\left|\partial_{\gamma_j}[-\gamma_j(x_{s_j} - c_j)]\right| = \left|-(x_{s_j} - c_j)\right|, \tag{D.8}$$

$$\left|\partial_{c_j}[-\gamma_j(x_{s_j} - c_j)]\right| = \left|\gamma_j\right|. \tag{D.9}$$

It is readily verified that, under Assumptions 2, 4 and 5, the expected values of (D.6) – (D.9) are finite. Furthermore, under Assumption 6, the expected value of any product between these equations is also finite.

Q.E.D.

LEMMA 6. *Under Assumptions 2, 4, 5 and 6,* $\mathbb{E}(\partial^2 \log f_t / \partial\boldsymbol{\theta}\partial\boldsymbol{\theta}') < \infty$.

PROOF. Set $\partial_{\theta\theta'} \equiv \frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$. It is clear that

$$\partial_{\theta\theta'} \log f_t = -\partial_\theta \log f_t \partial_{\theta'} \log f_t + f_t^{-1}\partial_{\theta\theta'} f_t. \tag{D.10}$$

Using the product law of differentiation, we can write $\partial_{\boldsymbol{\theta}\boldsymbol{\theta}'} f_t$ as a sum of products of $\partial_{\boldsymbol{\theta}} B_{it}$ and $\partial_{\boldsymbol{\theta}} \pi_{it}$ with $\partial_{\boldsymbol{\theta}\boldsymbol{\theta}'} B_{it}$ and $\partial_{\boldsymbol{\theta}\boldsymbol{\theta}'} \log \pi_{it}$. Using the results of Lemma 5, the expected value of the product of any two of these derivatives is finite. Therefore, we must show that $\mathbb{E}[\partial_{\boldsymbol{\theta}\boldsymbol{\theta}'} B_{it}] < \infty$ and $\mathbb{E}[\partial_{\boldsymbol{\theta}\boldsymbol{\theta}'} \log \pi_{it}] < \infty$. Considering that $\boldsymbol{\psi}_i$ and $\boldsymbol{\psi}_j$ do not have elements in common, and that B_{it} depends only on the vectors $\boldsymbol{\nu}_j$, $j \in \mathbb{J}_i$, we can write these derivatives in terms of $\partial_{\boldsymbol{\psi}_i\boldsymbol{\psi}_i'}$ and $\partial_{\boldsymbol{\nu}_j\boldsymbol{\nu}_k'}$. But $\boldsymbol{\psi}_i = [\beta_{0i}, \ldots, \beta_{pi}, \sigma_i^2]'$ and $\boldsymbol{\nu}_j = [\gamma_j, c_j]'$.

Then,

$$\partial_{\beta_{li}\beta_{ki}} log\pi_{it} = -\sigma_i^{-2}\tilde{x}_{kt}\tilde{x}_{lt}, \tag{D.11}$$

$$\partial_{\beta_{li}\sigma_i^2} log\pi_{it} = \sigma_i^{-4}\tilde{x}_{lt}(\tilde{\mathbf{x}}_t'\boldsymbol{\beta} - y_t), \tag{D.12}$$

$$\partial_{\sigma_i^2\sigma_i^2} log\pi_{it} = (2\sigma_i^4)^{-1}\sigma_i^{-8}(\tilde{\mathbf{x}}_t'\boldsymbol{\beta} - y_t)^2, \tag{D.13}$$

$$\left|\partial_{\nu_k\nu_j'} B_{it}\right| < \left|\partial_{\nu_k}[-\gamma_k(x_{s_k} - c_k)]\partial_{\nu_j'}[-\gamma_j(x_{s_j} - c_j)]\right| . \tag{D.14}$$

Under Assumptions 2, 4 and 5, the expected values of (D.11)–(D.14) are finite.

Q.E.D.

LEMMA 7. *Under Assumptions 2, 4, 5 and 6,* $\mathbb{E}(\partial^2 \log f_t/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'|_{\theta^*})$ *is negative definite.*

PROOF. If $\mathbb{E}(\partial^2 \log f_t/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'|_{\theta^*})$ is negative definite, then $\log f_t$ has a maximum in $\boldsymbol{\Theta}$. We know by Lemma 3 that $\log f_t$ has only one maximum or minimum in $\boldsymbol{\Theta}$; thus we only have to show that f_t must have a maximum.

Trivially, the Gaussian functions π_{it} have a maximum. If we multiply by a constant or monotone functions or add functions with a maximum, the function still has a maximum. The logistic function is a monotone function (in relation to its parameters and the variable). Hence, $B_{it}\pi_{it}$ has a maximum and f_t has a maximum, and $\mathbb{E}(\partial^2 \log f_t/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'|_{\theta^*})$ is negative definite.

Q.E.D.

LEMMA 8. *Under Assumptions 2, 4, 5 and 6, it follows that: (a)*$T^{-1}\sum_{t=1}^{T} f_t \overset{a.s.}{\to} \mathbb{E}(f_t)$*; (b)*$T^{-1}\sum_{t=1}^{T} \partial_\theta f_t \overset{a.s.}{\to} \mathbb{E}(\partial_\theta f_t)$*; and (c)*$T^{-1}\sum_{t=1}^{T} \partial_{\theta\theta'} f_t \overset{a.s.}{\to} \mathbb{E}(\partial_{\theta\theta'} f_t)$.

PROOF. First we must show that $T^{-1}\sum_{t=1}^{T} y_t \overset{a.s.}{\to} \mathbb{E}(y_t)$. Once y_t is a mixing process, we just need to show that (i) $\mathbb{E}\left(T^{-1}\sum_{t=1}^{T} y_t\right) = \mathbb{E}(y_t)$ and (ii) $\mathbb{V}\left(T^{-1}\sum_{t=1}^{T} y_t\right) < \infty$. As y_t is stationary, (i) is trivially satisfied and as $\sum \mathbb{E}(y_t y_{t-k}) < \Delta < \infty$, (ii) is satisfied.

Lemma 2 ensures that f_t, $\partial_\theta f_t$ and $\partial_{\theta\theta'} f_t$ are continuous functions of y_t given $\boldsymbol{\theta}$. Besides, Lemmas 4, 5 and 6 guarantee that the expected value is also finite. Once the functions are continuous and the expected value is finite, we can extend the results of y_t for f_t, $\partial_\theta f_t$ and $\partial_{\theta\theta'} f_t$, thereby completing the proof.

Q.E.D.

References

Andersen, T., Bollerslev, T. and Diebold, F.: 2006, Parametric and nonparametric measurement of volatility, *in* G. Elliott, C. Granger and A. Timmermann (eds), *Handbook of Financial Econometrics*, North-Holland, Amsterdam.

Bollerslev, T.: 1986, Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **21**, 307–328.

Carvalho, A. and Skoulakis, G.: 2004, Ergodicity and existence of moments for local mixture of linear autoregressions, *Technical report*, Northwestern University.

Carvalho, A. and Tanner, M.: 2005a, Mixture-of-experts of autoregressive time series: asymptotic normality and model specification, *IEEE Transactions on Neural Networks* **16**, 39–56.

Carvalho, A. and Tanner, M.: 2005b, Modeling nonlinear time series with mixture-of-experts of generalized linear models, *The Canadian Journal of Statistics* **33**, 1–17.

Chan, K. S. and Tong, H.: 1986, On estimating thresholds in autoregressive models, *Journal of Time Series Analysis* **7**, 179–190.

Chen, X. and Shen, X.: 1998, Sieve Extremum Estimates for Weakly Dependent Data, *Econometrica* **66**, 289–314.

Chen, X. and White, H.: 1998, Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators, *IEEE Transactions on Information Theory* **18**, 17–39.

Christoffersen, P. F.: 1998, Evaluating interval forecasts, *International Economic Review* **39**, 841–862.

da Rosa, J. C., Veiga, A. and Medeiros, M. C.: 2008, Tree-structured smooth transition regression models, *Computational Statistics and Data Analysis* **52**, 2469–2488.

Dempster, A. P., Laird, N. M. and Rubin, D. B.: 1977, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Engle, R. F.: 1982, Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation, *Econometrica* **50**, 987–1007.

Fan, J. and Yao, Q.: 2003, *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer-Verlag, New York, NY.

Gallant, A. R. and White, H.: 1992, On learning the derivatives of an unknown mapping with multilayer feedforward networks, *Neural Networks* **5**, 129–138.

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

Hornik, K., Stinchombe, M. and White, H.: 1990, Universal approximation of an unknown mapping and its derivatives using multi-layer feedforward networks, *Neural Networks* **3**, 551–560.

Härdle, W.: 1990, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Härdle, W., Lütkepohl, H. and Chen, R.: 1997, A review of nonparametric time series analysis, *International Statistical Review* **65**, 49–72.

Huerta, G., Jiang, W. and Tanner, M.: 2001, Mixtures of time series models, *Journal of Computational and Graphical Statistics* **10**, 82–89.

Huerta, G., Jiang, W. and Tanner, M.: 2003, Time series modeling via hierachical mixtures, *Statistica Sinica* **13**, 1097–1118.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E.: 1991, Adaptive mixtures of local experts, *Neural Computation* **3**, 79–87.

Jiang, W. and Tanner, M.: 1999, On the identifiability of mixtures-of-experts, *Neural Networks* **12**, 1253–1258.

Jordan, M. I. and Jacobs, R. A.: 1994, Hierarchical mixtures of experts and the EM algorithm, *Neural Computation* **6**, 181–214.

Kuan, C. M. and White, H.: 1994, Artificial neural networks: An econometric perspective, *Econometric Reviews* **13**, 1–91.

Le, N., Martin, R. and Raftery, A.: 1996, Modeling flat streches, bursts, and outliers in time series using mixture transition distribution models, *Journal of the American Statistical Association* **91**, 1504–1515.

Luukkonen, R., Saikkonen, P. and Teräsvirta, T.: 1988, Testing linearity against smooth transition autoregressive models, *Biometrika* **75**, 491–499.

MacKay, D. J. C.: 1992, Bayesian interpolation, *Neural Computation* **4**, 415–447.

McAleer, M.: 2005, Automated inference and learning in modeling financial volatility, *Econometric Theory* **21**, 232–261.

Medeiros, M., Teräsvirta, T. and Rech, G.: 2006, Building neural network models for time series: A statistical approach, *Journal of Forecasting* **25**, 49–75.

Medeiros, M. and Veiga, A.: 2005, Flexible coeficient smooth transition time series model, *IEEE Transactions on Neural Networks* **16**, 97–113.

Nowlan, S. J.: 1990, Maximum likelihood competitive learning, *Advances in Neural Information Processing Systems*, Vol. 2, Morgan Kaufmann, pp. 574–582.

Poon, S. and Granger, C.: 2003, Forecasting volatility in financial markets, *Journal of Economic Literature* **41**, 478–539.

Quinn, B., McLachlan, G. and Hjort, L.: 1987, A note on the Aitkin-Rubin approach to hypothesis testing in mixture models, *Journal of the Royal Statistal Society, Series B* **49**, 311–314.

Rech, G., Teräsvirta, T. and Tschernig, R.: 2001, A simple variable selection technique for nonlinear models, *Communications in Statistics, Theory and Methods* **30**, 1227–1241.

Taylor, S.: 1986, *Modelling Financial Time Series*, Wiley, Chichester.

Teräsvirta, T.: 1994, Specification, estimation, and evaluation of smooth transition autoregressive models, *Journal of the American Statistical Association* **89**, 208–218.

Tong, H.: 1978, On a threshold model, *in* C. H. Chen (ed.), *Pattern Recognition and Signal Processing*, Sijthoff and Noordhoff, Amsterdam.

Tong, H.: 1990, *Non-linear Time Series: A Dynamical Systems Approach*, Vol. 6 of *Oxford Statistical Science Series*, Oxford University Press, Oxford.

Tong, H. and Lim, K.: 1980, Threshold autoregression, limit cycles and cyclical data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 245–292.

Trapletti, A., Leisch, F. and Hornik, K.: 2000, Stationary and integrated autoregressive neural network processes, *Neural Computation* **12**, 2427–2450.

van Dijk, D., Teräsvirta, T. and Franses, P. H.: 2002, Smooth transition autoregressive models - a survey of recent developments, *Econometric Reviews* **21**, 1–47.

Weigend, A. S., Mangeas, M. and Srivastava, A. N.: 1995, Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting, *International Journal of Neural Systems* **6**, 373–399.

White, H.: 1992, *Estimation, Inference and Specification Analysis*, Cambridge University Press, New York, NY.

Wong, C. S. and Li, W. K.: 1999, On a generalized mixture autoregressive model, *Research Report 242*, Department of Statistics and Actuarial Science, University of Hong Kong.

Wong, C. S. and Li, W. K.: 2000, On a mixture autoregressive model, *Journal of the Royal Statistical Society, Series B* **62**, 91–115.

Wong, C. S. and Li, W. K.: 2001, On a mixture autoregressive conditional heterocedastic model, *Journal of the American Statistical Association* **96**, 982–995.

Wood, S., Jiang, W. and Tanner, M.: 2001, Bayesian mixture of splines for spatially adaptative nonparametric regression, *Biometrika* **89**, 513–528.

Zeevi, A., Meir, R. and Adler, R.: 1998, Non-linear models for time series using mixtures of autoregressive models, *Technical report*, Technion.

Departamento de Economia PUC-Rio
Pontifícia Universidade Católica do Rio de Janeiro
Rua Marques de Sâo Vicente 225 - Rio de Janeiro 22453-900, RJ
Tel.(21) 35271078 Fax (21) 35271084
www.econ.puc-rio.br
flavia@econ.puc-rio.br