TEXTO PARA DISCUSSÃO

No. 570

Modeling and Forecasting short-term Interest Rates: The Benefits of Smooth Regimes, Macroeconomic Variables, and Bagging

Francesco Audrino
Marcelo C. Medeiros

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

Modeling and Forecasting short-term Interest Rates: The Benefits of Smooth Regimes, Macroeconomic Variables, and Bagging

Francesco Audrino[a*] and Marcelo C. Medeiros[b]

[a]University of St. Gallen

[b]Pontifical Catholic University of Rio de Janeiro

Revised: November 2009

Abstract

In this paper we propose a smooth transition tree model for both the conditional mean and variance of the short-term interest rate process. The estimation of such models is addressed and the asymptotic properties of the quasi-maximum likelihood estimator are derived. Model specification is also discussed. When the model is applied to the US short-term interest rate we find (1) leading indicators for inflation and real activity are the most relevant predictors in characterizing the multiple regimes' structure; (2) the optimal model has three limiting regimes. Moreover, we provide empirical evidence of the power of the model in forecasting the first two conditional moments when it is used in connection with bootstrap aggregation (bagging).

Keywords: short-term interest rate, regression tree, smooth transition, conditional variance, bagging, asymptotic theory.

[*]Address for correspondence: University of St. Gallen, Institute of Mathematics and Statistics, Bodanstrasse 6, CH-9000 - St. Gallen, Switzerland. Tel: +41 71 224 2431. Fax: +41 71 224 2894. Email: francesco.audrino@unisg.ch.

1 Introduction

The relevance of the short-term interest rate is directly related to the fact that, from a macroeconomic point of view, the rate is a policy instrument under the control of the central banks to maintain economic stability. Moreover, from a finance perspective, the short rate is the essential quantity needed to construct the whole yield curve, given that yields at other maturities are just risk adjusted averages of expected future short rates. Therefore, it is not surprising that in the last two decades a number of different models have been proposed for the conditional dynamics of the short-term interest rate process.

One important stylized fact that must be taken into account when constructing a model for the short rate dynamics is that the short rate is subject to regime-shifts; see, for example, Gray (1996), Hansen and Poulsen (2000) and Audrino (2006). The empirical studies of Gray (1996) and Audrino (2006), in particular, confirmed that regime-switching models for the conditional mean and variance dynamics of the short rate process yield more accurate short rate forecasts. As a direct consequence, regime-switching models also yield more accurate predictions of the whole yield curve, with important implications for the pricing of interest-rate-sensitive instruments and for risk management; see, among others, Bansal and Zhou (2002), Bansal et al. (2004), and Audrino and De Giorgi (2007).

Besides the statistical properties of a proposed model for the short rate (that is, asymptotic results, in- and out-of-sample performances), the model must also offer some reduced-form insight into the nature of the underlying economic forces that drive the short rate movements. In several studies published in the last five years, researchers incorporated macroeconomic variables as predictors or latent factors in models for the short rate and, more generally, the whole yield curve. For example, Diebold et al. (2006) used three observable macroeconomic variables (that is, real activity, inflation, and a monetary-policy instrument). In Ang and Piazzesi (2003) and Ang et al. (2007) the macroeconomic variables used are measures for inflation and real activity. In particular, Ang and Piazzesi (2003) constructed the measures for inflation and real activity as the first principal component of a large set of candidate macroeconomic series for inflation and real activity, respectively. Rudebusch and Wu (2004) provided an example of a macro-finance model that employs more macroeconomic structure and includes both rational expectations and inertial elements. Finally, a whole set of macroeconomic variables for real activity and inflation were used in Audrino (2006). In his model, Audrino (2006) chose the most important macroeconomic series for the estimation and prediction of the short rate process dynamics via information criteria.

We propose a generalization of the Audrino (2006) tree-structured model that is able to take into account regime-shifts in the conditional dynamics of the short rate process, and to exploit all possible information coming from macroeconomic and other relevant exogenous variables for estimation and interpretation as well as for prediction. The most important difference between the Audrino (2006) model and the model we propose here is that we allow regime-shifts to be smooth. Our model is a compromise between the Markovian regime-switching model introduced by Gray (1996), where regime-shifts are driven by an unobservable state variable with associated transition probabilities and a consequent loss of interpretation, and the Audrino (2006) tree model, where regime-shifts are drastic: at a given time, the short rate process is driven exactly by the local dynamics of one limiting regime (that is, the probabilities associated with the regimes are of the type 0-1). The degree of the smoothness is determined endogenously when estimating the model.

The model we propose is also a generalization of the smooth transition regression tree (STR-tree) model introduced by da Rosa et al. (2008). In this study, we expand the STR-tree model to allow not only the conditional mean dynamics, but also the conditional variance dynamics to be non-linear and regime-dependent as in Audrino and Bühlmann (2001) and Medeiros and Veiga (2009). We derive the asymptotic theory for our model based on the assumption that the model structure is correctly specified apart from the error distribution, which is left unspecified. Our specification differs in many aspects from the above mentioned papers. First, contrary to Audrino and Bühlmann (2001) we consider smooth transitions among regimes instead of sharp ones. Second, the model proposed in Medeiros and Veiga (2009) allows for only one transition variable and the conditional mean is assumed to be zero[1]. We relax these two restrictions and allow for multiple transition variables and also a nonlinear conditional mean. The purpose of modelling and forecasting the conditional variance is threefold. First, in terms of understanding and modelling the dynamics of the short-term interest rates, it is important to check if the regime switches are also present in the conditional variance. Second, the conditional variance forecasts are essential for the construction of prediction intervals. Finally, the dynamics of the conditional volatility is crucial for the understanding of the interest-rate risk.

Since one of our goals is to investigate the appropriateness of our model for forecasting the short rate process, as with Inoue and Kilian (2008) and Hillebrand and Medeiros (2007) we use bootstrap aggregating (bagging, introduced by Breiman, 1996) to improve predictions. In fact, tree-based procedures based on hard decisions with indicator functions are known to be highly unstable. As Bühlmann and Yu (2002) have shown, bagging is a statistical procedure effective in the case of regression trees in alleviating such a problem.

We test the estimation and forecasting ability of our model on the time series of the US short-term interest rate process. First, similarly to previous studies, we find that leading indicators for inflation and real activity are the most relevant predictors in characterizing the regimes' structure. The optimal model has three limiting regimes, with significantly different local conditional mean and variance dynamics. We also find some correspondence between NBER expansions/recessions and our limiting regimes.

Second, we provide empirical evidence that our model is the one yielding the most

[1] The theoretical results in Medeiros and Veiga (2009) are heavily dependent on these two restrictions and are not applicable to the present case. Our asymptotic results are not a straightforward application of Ling and McAleer (2003) as the later considered only linear specifications.

accurate predictions, in particular when used in connection with bagging, and also when compared with several competitors introduced in the literature. By performing a series of superior predictive ability (SPA) tests (Hansen, 2005), we conclude that such improvements are in most cases statistically significant.

The remainder of the paper is organized as follows: In Section 2 we introduce the double smooth transition tree (DST-Tree) model. Estimation and asymptotic properties are discussed in Section 3. Bagging is discussed in Section 4. Section 5 presents the empirical application to the US short-term interest rate series. Section 6 concludes.

2 Model

In this paper we consider a general version of the Smooth Transition Regression Tree (STR-Tree) model of da Rosa et al. (2008). The novelty of our model is to allow a similar tree-structured nonlinearity in conditional variance of the model. First, consider the following assumption regarding the data generating process (DGP):

ASSUMPTION 1. *The observed sequence of real-valued vector of variables* $\mathbf{Y}_t = \{y_t, \mathbf{x}_t\}_{t=1}^{T}$ *is a realization of a stationary and ergodic stochastic process on a complete probability space generated as*

$$y_t = f\left(\mathbf{x}_t; \boldsymbol{\psi}_0\right) + \varepsilon_t, \quad t = 1, \ldots, T, \tag{1}$$

where $f\left(\mathbf{x}_t; \boldsymbol{\psi}_0\right)$ *is a (nonlinear) function of the real-valued random vector* $\mathbf{x}_t \in \mathbb{X} \subseteq \mathbb{R}^q$*, which has distribution function* F *on* Ω*, a Euclidean space.* $\boldsymbol{\psi}_0$ *is a vector of unknown (true) parameters. 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 and symmetric distribution such that* $\mathbb{E}[\varepsilon_t] = 0$ *and* $\mathbb{E}[\varepsilon_t^2] = \sigma^2$, $0 < \sigma^2 < \infty$, $\forall\, t$*. In addition,* $\mathbb{E}\left[\varepsilon_t | \mathbf{x}_t, \mathcal{F}_{t-1}\right] = 0$*, where* $\mathcal{F}_{t-1}$ *is the filtration with respect to all past information. Finally, we allow the conditional variance to be time-varying, such that* $\mathbb{E}\left[\varepsilon_t^2 | \mathbf{x}_t, \mathcal{F}_{t-1}\right] = h_t(\boldsymbol{\psi}_0) < \infty$*, and* $h_t(\boldsymbol{\psi}_0) > 0$*,* $\forall\, t$*.*

In the practical application of Section 5, $y_t \equiv \Delta r_t = r_t - r_{t-1}$ is the first difference of the short rate process at time t, r_t is the short rate process at time t, and $\mathbf{x}_t = (\Delta r_{t-1}, r_{t-1}, (\mathbf{x}_{t-1}^{\text{ex}})')'$ is the vector of all relevant information for prediction at time t, with $\mathbf{x}_{t-1}^{\text{ex}}$ denoting the vector of exogenous variables, like indices for inflation and real activity.

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 $\mathbb{T}$. The proposed model follows from the following definition.

Definition 1. *Set $\widetilde{\mathbf{x}}_t = (1, \mathbf{x}_t)'$. A parametric model $\mathcal{M}$ defined by the function $H_{\mathbb{J}\mathbb{T}}(\mathbf{x}_t; \boldsymbol{\psi}_0) : \mathbb{R}^{q+1} \to \mathbb{R}$, indexed by the vector of parameters $\boldsymbol{\psi}_0 \in \boldsymbol{\Psi}$, a compact subset of the Euclidean space, is called a double smooth transition tree model (DST-Tree), if*

$$y_t = H_{\mathbb{J}\mathbb{T}}(\mathbf{x}_t; \boldsymbol{\psi}_0) + \varepsilon_t = \sum_{i \in \mathbb{T}} \boldsymbol{\beta}_i' \widetilde{\mathbf{x}}_t B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_i) + h_t(\boldsymbol{\psi}_0)^{1/2} u_t, \tag{2}$$

where

$$h_t(\boldsymbol{\psi}_0) \equiv h_t = \sum_{i \in \mathbb{T}} \left(a_i \varepsilon_{t-1}^2 + b_i h_{t-1} + \boldsymbol{\lambda}_i' \widetilde{\mathbf{x}}_t \right) B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_i), \tag{3}$$

$$B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_i) = \prod_{j \in \mathbb{J}} G(x_{s_j,t}; \gamma_j, c_j)^{\frac{n_{i,j}(1+n_{i,j})}{2}} \left[1 - G(x_{s_j,t}; \gamma_j, c_j)\right]^{(1-n_{i,j})(1+n_{i,j})}, \tag{4}$$

$$G(x_{s_j,t}; \gamma_j, c_j) = \frac{1}{1 + e^{-\gamma_j \left(x_{s_j,t} - c_j\right)}}, \tag{5}$$

and

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

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* $k \in \mathbb{J}_i$, $i \in \mathbb{T}$. *Finally,* $\{u_t\}$ *is a sequence of independent and identically distributed zero-mean random variables with unit variance,* $u_t \sim \mathsf{IID}(0,1)$.

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

Remark 2. *Note that the same tree structure is considered in the conditional mean and conditional variance. This simplifies estimation, avoids possible "curse of dimensionality", and facilitates the final interpretation of the model.*

Remark 3. *Although the notation in (2) may seem a bit complicated at first sight, it has the main advantage of being capable of mathematically representing any tree-structure. For a simple example of a smooth transition tree structured model, we refer to da Rosa et al. (2008).*

For simplicity, and to be consistent with other models introduced in the literature (see, for example, Gray, 1996, or Audrino, 2006), in our real data investigation of Section 5 on the short rate process $\{r_t\}_{t\in\mathbb{N}}$, we restrict the general local conditional mean and variance dynamics given in (2) and (3) to follow:

$$y_t = \Delta r_t = \mu_t(\boldsymbol{\psi}_0) + \varepsilon_t = \sum_{i\in\mathbb{T}} (\alpha_i + \beta_i r_{t-1}) B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_i) + h_t(\boldsymbol{\psi}_0)^{1/2} u_t, \tag{6}$$

and

$$h_t(\boldsymbol{\psi}_0) = \sum_{i \in \mathbb{T}} \left(a_i \varepsilon_{t-1}^2 + b_i h_{t-1} + \sigma_i^2 r_{t-1} \right) B_{\mathbb{J}i} \left(\mathbf{x}_t; \boldsymbol{\theta}_i \right). \tag{7}$$

Note that there are no constant terms in the variance equation (7). According to Gray (1996), the lower bound on the variance equation, such that variance is strictly positive, is given by the level effects of interest rates.

3 Estimation and asymptotic theory

In this section we discuss the estimation of the DST-Tree model and the corresponding asymptotic theory. As the true distribution of u_t is unknown, the parameters of model (2) are estimated by a quasi-maximum likelihood estimator (QMLE). The quasi-maximum likelihood function of (2) is

$$\mathcal{L}_T(\boldsymbol{\psi}) = \frac{1}{T} \sum_{t=1}^{T} \ell_t(\boldsymbol{\psi}) = \frac{1}{T} \sum_{t=1}^{T} \left[-\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(h_t) - \frac{\varepsilon_t^2}{2h_t} \right]. \tag{8}$$

Note that the processes y_t, $\mathbf{x}_t$, and h_t, $t \leq 0$, are unobserved, and hence are only arbitrary constants. Thus, $\mathcal{L}_T(\boldsymbol{\psi})$ is a quasi-log-likelihood function that is not conditional on the true $(y_0, \mathbf{x}_0, h_0)$, making it suitable for practical applications. However, to prove the asymptotic properties of the QMLE, it is more convenient to work with the unobserved process $\{(\varepsilon_{u,t}, h_{u,t}) : \, t = 0, \pm 1, \pm 2, \ldots\}$.

Conditional on $\mathcal{F}_0 = (y_0, \mathbf{x}_0, y_{-1}, \mathbf{x}_{-1}, y_{-2}, \mathbf{x}_{-2}, \ldots)$, the unobserved quasi-log-likelihood function is given by

$$\mathcal{L}_{u,T}(\boldsymbol{\psi}) = \frac{1}{T} \sum_{t=1}^{T} \ell_{u,t}(\boldsymbol{\psi}) = \frac{1}{T} \sum_{t=1}^{T} \left[-\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(h_{u,t}) - \frac{\varepsilon_{u,t}^2}{2h_{u,t}} \right]. \tag{9}$$

The main difference between $\mathcal{L}_T(\boldsymbol{\psi})$ and $\mathcal{L}_{u,T}(\boldsymbol{\psi})$ is that the former is conditional on any initial values, whereas the latter is conditional on an infinite series of past observations. In practice, the use of (9) is not possible.

3.1 Asymptotic theory

Let

$$\widehat{\boldsymbol{\psi}}_T = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\mathsf{argmax}} \mathcal{L}_T(\boldsymbol{\psi}) = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\mathsf{argmax}} \left[\frac{1}{T}\sum_{t=1}^{T} \ell_t(\boldsymbol{\psi})\right],$$

and

$$\widehat{\boldsymbol{\psi}}_{u,T} = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\mathsf{argmax}} \mathcal{L}_{u,T}(\boldsymbol{\psi}) = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\mathsf{argmax}} \left[\frac{1}{T}\sum_{t=1}^{T} \ell_{u,t}(\boldsymbol{\psi})\right].$$

Define $\mathcal{L}(\boldsymbol{\psi}) = \mathbb{E}\left[\ell_{u,t}(\boldsymbol{\psi})\right]$. We proceed to discuss the existence of $\mathcal{L}(\boldsymbol{\psi})$ and prove the consistency of $\widehat{\boldsymbol{\psi}}_T$ and $\widehat{\boldsymbol{\psi}}_{u,T}$. We first prove the strong consistency of $\widehat{\boldsymbol{\psi}}_{u,T}$, and then show that $\sup\limits_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} |\mathcal{L}_{u,T}(\boldsymbol{\psi}) - \mathcal{L}_T(\boldsymbol{\psi})| \overset{a.s.}{\to} 0$, so that the consistency of $\widehat{\boldsymbol{\psi}}_T$ follows. Asymptotic normality of both estimators is considered in sequence. We prove the asymptotic normality of $\widehat{\boldsymbol{\psi}}_{u,T}$. The proof of $\widehat{\boldsymbol{\psi}}_T$ is straightforward. Detailed proofs of the following theorems are given in Appendix A.

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

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

Remark 4. *In Assumption 1 we restrict the process to be stationary and ergodic. Finding necessary and sufficient stationary conditions for nonlinear models is, in general, a very difficult task. In most cases, only sufficient and overly restrictive conditions are available. The case of the model considered in this paper is not different. Considering equations (6) and (7), one possible sufficient condition is to impose that the model in each limiting regime is stationary: $|\beta_i| < 1$ and $|a_i + b_i| < 1$, $\forall\, i$. However, as pointed out in Medeiros and Veiga (2009), this set of restrictions may be too restrictive. In practical terms, one can always simulated the paths generated from an estimated model and check whether or not it has stationary dynamics.*

Remark 5. *For example, a set of sufficient conditions for almost sure positivity of the conditional variance of the DST-Tree model is:*

1. $a_i \geq 0$*,* $b_i \geq 0$*, and* $\sigma_i^2 > 0$*,* $\forall\, i \in \mathbb{T}$*;*

2. $r_t > 0$*, a.s.,* $t = 1, \ldots, T$*.*

Consider the following assumption.

Assumption 2. *The true and unique parameter vector* $\boldsymbol{\psi}_0 \in \boldsymbol{\Psi}$ *is in the interior of* $\boldsymbol{\Psi}$*, a compact subset of finite dimensional Euclidean space.*

Assumption 3. *The DST-Tree model is identifiable, in the sense that, for a sample* $\{y_t, \mathbf{x}_t\}_{t=1}^T$ *and for* $\boldsymbol{\psi}_1$*,* $\boldsymbol{\psi}_2 \in \boldsymbol{\Psi}$*,* $\mathcal{L}_T(\boldsymbol{\psi}_1) = \mathcal{L}_T(\boldsymbol{\psi}_2)$ *with probability 1 is equivalent to* $\boldsymbol{\psi}_1 = \boldsymbol{\psi}_1$*.*

Assumption 2 is standard while Assumption 3 guarantees the identification of the model. The consistency result is given in the following theorem.

Theorem 2. *Under the Assumptions 1–3 the QMLE* $\widehat{\boldsymbol{\psi}}_T$ *is weak consistent for* $\boldsymbol{\psi}_0$*, i.e.,* $\widehat{\boldsymbol{\psi}_T} \xrightarrow{p} \boldsymbol{\psi}_0$*.*

We introduce the following matrices:

$$\mathbf{A}(\boldsymbol{\psi}_0) = \mathbb{E}\left[-\left.\frac{\partial^2 \ell_{u,t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'}\right|_{\boldsymbol{\psi}_0} \right], \quad \mathbf{B}(\boldsymbol{\psi}_0) = \mathbb{E}\left[\left.\frac{\partial \ell_{u,t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}_0} \left.\frac{\partial \ell_{u,t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\right|_{\boldsymbol{\psi}_0} \right],$$

and

$$\begin{aligned}\mathbf{A}_T(\boldsymbol{\psi}) =& \frac{1}{T}\sum_{t=1}^{T} \left[\frac{1}{2h_t}\left(\frac{\varepsilon_t^2}{h_t} - 1\right)\frac{\partial^2 h_t}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} - \frac{1}{2h_t^2}\left(2\frac{\varepsilon_t^2}{h_t} - 1\right)\frac{\partial h_t}{\partial \boldsymbol{\psi}}\frac{\partial h_t}{\partial \boldsymbol{\psi}'} \right. \\ & \left. + \left(\frac{\varepsilon_t}{h_t^2}\right)\left(\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}}\frac{\partial h_t}{\partial \boldsymbol{\psi}'} + \frac{\partial h_t}{\partial \boldsymbol{\psi}}\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}'}\right) + \frac{1}{h_t}\left(\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}}\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}'} + \varepsilon_t \frac{\partial^2 \varepsilon_t}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'}\right) \right],\end{aligned} \tag{10}$$

$$
\begin{aligned}
\mathbf{B}_T(\boldsymbol{\psi}) &= \frac{1}{T}\sum_{t=1}^{T}\frac{\partial \ell_t(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\frac{\partial \ell_t(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'} \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{4h_t^2}\left(\frac{\varepsilon_t^2}{h_t}-1\right)^2\frac{\partial h_t}{\partial \boldsymbol{\psi}}\frac{\partial h_t}{\partial \boldsymbol{\psi}'}+\frac{\varepsilon_t^2}{h_t}\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}}\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}'}\right. \\
&\qquad\left. -\frac{\varepsilon_t}{2h_t^2}\left(\frac{\varepsilon_t^2}{h_t}-1\right)\left(\frac{\partial h_t}{\partial \boldsymbol{\psi}}\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}'}+\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}}\frac{\partial h_t}{\partial \boldsymbol{\psi}'}\right)\right]
\end{aligned}
\tag{11}
$$

Consider the following assumption:

Assumption 4. $\mathbb{E}\left[\varepsilon_t^4\right] = \mu_4 < \infty$.

The following theorem states the asymptotic normality result.

Theorem 3. *Under Assumptions 1–4,* $\sqrt{T}(\widehat{\boldsymbol{\psi}}_T - \boldsymbol{\psi}_0) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \mathbf{A}(\boldsymbol{\psi}_0)^{-1}\mathbf{B}(\boldsymbol{\psi}_0)\mathbf{A}(\boldsymbol{\psi}_0)^{-1}\right)$*. Furthermore, the matrices* $\mathbf{A}(\boldsymbol{\psi}_0)$ *and* $\mathbf{B}(\boldsymbol{\psi}_0)$ *are consistently estimated by* $\mathbf{A}_T(\widehat{\boldsymbol{\psi}})$ *and* $\mathbf{B}_T(\widehat{\boldsymbol{\psi}})$*, respectively.*

3.2 Modeling Cycle

In this section we briefly present the modeling cycle adopted in this paper. 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 are carried out by the use of a information criterium, such as the BIC. An alternative procedure, which has not been used in this paper, is to use a a sequence of Lagrange Multiplier (LM) tests following the ideas originally presented in Luukkonen et al. (1988) and widely used in the literature; see, for example, da Rosa et al. (2008). Our choice to use the BIC is motivated by the empirical evidence that such an approach works well in practice with regression-tree models; see, for example, Audrino (2006).

As pointed out by one of the referees, the use of a information criterium to specify the DST-Tree model inevitably means estimating a number of unidentified models, which may cause numerical problems and instabilities in the modelling procedure. This is of course true, however, the use of Bagging as described in Section 4 can attenuate such problems. Furthermore, it is not clear if sequence of LM tests advocated by da Rosa et al. (2008) is a consistent procedure to specify the structure of the DST-Tree model due the nested nature of such models.

Consider that y_t follows a DST-Tree model with K leaves and we want to decide whether or not the terminal node $i^* \in \mathbb{T}$ should be split.

The approach adopted here is closely related to the one advocated in Audrino and Bühlmann (2001). First, a growing algorithm is used until a maximum number of limiting regimes is achieved. At each step, the idea is to select the node to be split and the respective transition variable such that the log-likelihood is maximized. Of course, such procedure can lead to an over-parametrized specification. The second step is to prune the model. This is carried out by the use of information criterium: We search for a best subtree with respect to the BIC which is often computationally feasible since the number of regimes is not very big. For example, in our empirical analysis we found three limiting regimes. For more details, see Audrino and Bühlmann (2001) or Audrino (2006).

4 Forecasting: The role of bagging

It is well known that instability (that is, the variance of the estimator is high) often occurs when hard decisions with indicator functions are involved as in the case of regression or classification trees; see, for example, Hastie et al. (2001) or Berk (2008). One way to reduce such an instability is bootstrap aggregating (bagging, for short) introduced by Breiman (1996). Bagging is a statistical procedure designed to improve forecast accuracy of models selected by unstable decision rules. Bagging has been shown to be a useful technique to improve the accuracy of final forecasts based on the predictive power of potentially many relevant predictors that, individually, have only weak explanatory power. In essence, bagging involves (i) fitting a given model to the original sample, considering as predictors in the estimation all potentially relevant variables; (ii) generating a large number of bootstrap resamples from this approximation of the data; (iii) applying the decision rule to each of the resamples; and (iv) averaging across bootstrap resamples the forecasts obtained from the models selected by the decision rule when estimated on the different resamples. By averaging across resamples, bagging effectively removes the instability of the decision rule. Improvements are relevant in particular when the variance of the estimator is high, as in the case of tree-based procedures.

Bühlmann and Yu (2002) showed that bagging has the potential to achieve dramatic reductions in forecast mean squared errors for a wide range of unstable procedures. In particular, bagging turns out to be advantageous when aiming to improve the predictive performance of regression and classification trees. In case of regression trees, the theory developed in Bühlmann and Yu (2002) confirms Breiman's intuition that bagging is a variance reduction technique, reducing also the mean squared error. Recently, Inoue and Kilian (2004) extended the use of bagging to the time series framework, presented the theoretical arguments in favor of bagging, and characterized the conditions under which one would expect bagging to work. In two succeeding applications, Inoue and Kilian (2008) (bagging applied to the forecast of US CPI inflation) and Hillebrand and Medeiros (2007) (bagging applied to the forecast realized volatility) found good and encouraging results. Therefore, we propose bagging to alleviate the instability problem directly related to the use of tree-based procedures, and to improve the forecasts of short-term interest rate process dynamics obtained from the smooth-transition tree-structured model.

Based on the bagging procedure proposed by Inoue and Kilian (2004) for the linear regression model, the bagged DST-Tree model for the short-term interest rate dynamics is constructed as follows.

1. Arrange the set of response and predictor variables in the form of a matrix of

dimension $T \times K$, where $K = 1+$ the number of predictor variables considered:

$$\left\{\Delta r_t, \mathbf{x}_t'\right\}, \quad t = 1, \ldots, T$$

where $\mathbf{x}_t = \left(\Delta r_{t-1}, r_{t-1}, (\mathbf{x}_{t-1}^{\text{ex}})'\right)'$.

Construct B bootstrap samples of the form

$$\left(\Delta r_{(i)1}^*, (\mathbf{x}_{(i)1}^*)'\right), \ldots, \left(\Delta r_{(i)T}^*, (\mathbf{x}_{(i)T}^*)'\right),\ i = 1, \ldots, B$$

by drawing with replacement blocks of m rows of this matrix, where the block size m is chosen to capture the dependence in the error term.

2. For each bootstrap sample, estimate the DST-Tree model with three limiting regimes[2] following the procedure proposed in Section 3. Note that for each bootstrap sample the optimal selection of predictor variables and splitting points, as well as the optimal local parameters will be different. Compute the forecasts of the conditional mean and variance of the short-rate process for the out-of-sample period by using the optimal parameters estimated from the i-th bootstrap sample, and call them

$$(\mu_{(i)T+t}^*, h_{(i)T+t}^*),\ t = 1, \ldots, T_{\text{out}}.$$

3. Compute the average forecasts of the conditional mean and variance of the short-rate process for the out-of-sample period:

$$\left(\hat{\mu}_{T+t} = \frac{1}{B}\sum_{i=1}^{B} \mu_{(i)T+t}^*, \hat{h}_{T+t} = \frac{1}{B}\sum_{i=1}^{B} h_{(i)T+t}^*\right),\ t = 1, \ldots, T_{\text{out}}.$$

In most cases, bagging trees will not be a prohibitive computational burden; see, for

[2] We fixed the depth of the tree to be the same as the optimal tree estimated from the original data; see Section 5.2.

example, the results illustrated in Berk (2008). Nevertheless, if the bagging procedure becomes too computationally expensive, the same properties holding for bagging trees (and discussed above) also hold for subsample aggregating (subagging) trees, that is a computationally cheaper version of bagging, given that it implies re-estimation of the models on resamples with smaller sizes than the original one; for more details, see Bühlmann and Yu (2002) or Buya and Stuetzle (2006).

5 Real Data Investigation

5.1 Data

The data used in this study are one-month U.S. Treasury bill rates downloaded from the Fama CRSP Treasury bill files. The data span the time period between January 1960 and December 2006, for a total of 564 monthly observations. We split the data sample in two parts; Consistent with the literature, we use the period between January 1960 and December 2001 (504 observations) as in-sample estimation period. The remaining 60 observations are left to test the prediction accuracy of the different model specifications. Figure 1 plots the data as well as the monthly changes in short-term interest rates. Table 1 presents some sample statistics.

Figure 1 illustrates well the dramatic changes in the short-term interest rates that occurred during the OPEC oil crises in the 1973-75 period and the Fed experiment in the 1979-82 period. The volatility of the monthly changes associated with the Fed experiment is striking. Volatility is also noticeably higher than average during the 1973-75 period and immediately after the October 1987 stock market crash. As expected, Table 1 shows that the mean change in the short-term interest rates is close to zero, that there is significant excess kurtosis, and that the correlation between Δr_t and r_{t-1} is negative. All these stylized facts have been documented in the literature and justify the introduction of regime-switching models (of Markovian or threshold type) as reasonable and simple processes for the short-term interest rate dynamics.

Similarly to Ang and Piazzesi (2003), Audrino (2006) and Diebold et al. (2006), we consider a number of term structure and macroeconomic factors as predictors in our smooth transition tree structured model. This is done to exploit the additional information of the yield curve, real activity, and inflation, for estimation and prediction purposes. In greater detail, we consider the 60-month zero coupon bond rates from the Fama CRSP discount bond files, as well as the spread between the 60-month and the 1-month yields, the CPI and the PPI of finished goods as measures of inflation, and the index of Help Wanted Advertising in Newspapers (HELP), unemployment (UE) and the growth rate of industrial production (IP), and GDP to capture real activity. All the macroeconomic data have been downloaded from *Datastream International* for the time period under investigation. This list of variables includes most that have been used in the macro literature. Among these variables, HELP is traditionally considered a leading indicator of real activity. Summary statistics of these variables are reported in Table 1.

5.2 Estimation results

We analyze the optimal regimes' structure, transition functions, and parameter estimates of the local conditional mean and variance of the short-term interest rate obtained using the DST-Tree model introduced in Section 2. Local parameter estimates and optimal limiting regimes are summarized in Table 2. They are computed for the in-sample period beginning January 1960 and ending December 2001, for a total of 504 monthly observations. The detailed specification of the model is noted under Table 2.

We find that the estimated DST-Tree model has three limiting regimes. Similar to the findings of Audrino (2006), such limiting regimes are fully characterized by the two main indices for real activity and inflation. The first limiting regime is characterized by a low real activity, the implied long-run mean is relatively low (3.6%), and there is strong statistical evidence of a moderate mean reversion. Individual shocks have a negligible immediate effect on the conditional variance, but are strongly persistent. The conditional variance is also significantly related to the level of the short rate, although the small value of the CIR parameter renders it economically insignificant.

The second and third limiting regimes are both characterized by high real activity, but by a different level of inflation. In the second limiting regime, inflation is low. The implied long-run mean is large and negative (approximately -26%). Individual shocks have neither immediate nor persistent effect on the conditional variance. On the contrary, conditional variance is significantly related to the level of the short rate.

In the third limiting regime, both real activity and inflation are high. There is strong evidence of mean reversion around a high implied long-run mean (approximately 13%). The local GARCH process is not weakly stationary ($a_3 + b_3 > 1$) [3]; individual shocks have a large (but not statistically significant) immediate impact on the conditional variance and are strongly persistent. Although, the t-statistic for $\widehat{a}_3$ is low (1.4551), $\widehat{a}_3$ is quite high, in particular when compared to $\widehat{a}_1$ and $\widehat{a}_2$.

To complete this section, we now analyze the optimal functions $B_{\mathbb{J}i}(\cdot)$, that is the probability functions associated with the three different local specifications given in Table 2. The shape of the functions is shown in Figure 2.

The optimal parameters are $\gamma_1 = 0.2882$ and $\gamma_2 = 0.1488$.[4] As Figure 2 clearly shows, the three logistic functions are non-linear in the predictors and considerably smoother than the identity (that is 0-1) functions used by classical trees. This renders a clear interpretation of the regimes in terms of contractions/expansions periods difficult. Nevertheless, time periods characterized by values of the HELP index smaller than 80 can be reasonably associated with regime 1 (the probability of being in such a regime is very high; see again Figure 2). On the contrary, time periods characterized by values of the

[3] Note that, even with a nonstationary regime, the global model can be still stationary; see Medeiros and Veiga (2009) for a discussion

[4] We computed values of the t-statistics of the optimal γ parameters based on heteroskedastic-consistent standard errors. Such values are not significant. This is not surprising, since under the null-hypothesis the parameters are not identified and the distribution of the statistic is not correctly specified.

HELP index larger than 100 can be associated with regimes 2 and 3. A clear distinction between regimes 2 and 3 is more difficult and can lead to wrong conclusions. In Figure 3 we overlay shaded NBER recessions to the time series of the HELP index to illustrate recessions/expansions correspondence.

Not surprisingly, Figure 3 shows that during most NBER recessions between 1960 and 2001, the conditional dynamics of the short-term interest rate followed closely those described under regime 1. This is consistent with the results found in Audrino (2006).

5.3 Forecasting results

Here we investigate the accuracy of the proposed models for the prediction of first and second conditional moments of the one-month-ahead short-term interest rate process. The out-of-sample period goes from January 2002 to December 2006, for a total of 60 monthly observations. To reduce computational costs, we adopt here a split-sample procedure.

We compare goodness-of-fit results of the smooth transition tree-structured (ST-tree) model with those from: (1) a global CIR-GARCH-type model with level effects in conditional variances (single-regime ST-tree model); (2) a global CIR-GARCH-type model with level effects in conditional variances and all relevant macro-variables in the conditional mean equation. The significant macro-variables in the conditional mean are chosen using subset selection (see Hastie et al., 2001, pages 55-57). We found that the relevant macro-variables are HELP, PPI and GDP; (3) the Markovian regime-switching (RS) model with two regimes proposed by Gray (1996); (4) a modification of the RS model proposed by Gray (1996), where probabilities are also allowed to depend on macro-variables (see Audrino, 2006). We found that the most relevant macro-variable is the HELP index; and (5) the standard tree-structured model proposed by Audrino (2006).

For each model specification, we also consider the bagged version of it. We quantify the goodness-of-fit of the different models for predicting monthly first and second conditional moments by means of three different measures: the out-of-sample negative log-likelihood

(Loglik), and the mean squared errors (MSE) for the conditional mean and variance. Mathematically speaking, the last two performance measures are given by:

$$\text{MSE-mean} = \frac{1}{60}\sum_{t=1}^{60}\left(\Delta r_t - \hat{\mu}_t\right)^2 \tag{12}$$

$$\text{MSE-variance} = \frac{1}{60}\sum_{t=1}^{60}\left(\hat{h}_t - (\Delta r_t - \hat{\mu}_t)^2\right)^2 \tag{13}$$

where $\hat{\mu}_t$ and $\hat{h}_t$ are computed using the optimal parameters estimated with the in-sample data (from January 1960 to December 2001). We performed a series of the superior predictive ability (SPA) tests for forecasting one-month ahead first and second conditional moments introduced by Hansen (2005) to quantify statistical differences among the models. In the SPA tests, we test the null-hypothesis that each particular model is not outperformed by any of the alternative specifications.

The performance results are summarized in Table 3. In the bagging procedure using the block-bootstrap of Künsch (1989), we use $B = 50$ replications and a block size of $m = 20$. p-values of the SPA tests are reported in parentheses (Panel A).

Without considering bagging, the DST-Tree model yields the best result with respect to the out-of-sample negative log-likelihood and is also competitive for forecasting conditional variance. It shows some problems when the focus is the prediction of the conditional mean. As argued in Section 4, such difficulties may be a consequence of the instability of tree-based models. Results showed in Table 3 support this thesis. The usefulness of bagging is particularly evident. The bagged DST-Tree yields the best results with respect to all out-of-sample performance measures considered. It clearly outperforms all other model specifications. Such differences are in most cases statistically significant at the 5 percent or 10 percent confidence levels, as the results of the SPA tests show.

To end the analysis, we also perform a series of generalized Diebold and Mariano tests to take into account serial correlation (see Diebold and Mariano, 1995). We perform pairwise comparisons of the bagged version of the DST-Tree model (benchmark model) against the bagged alternative specifications. Results are shown in Panel B of Table 3. Negative values of the statistic are in favor of the bagged DST-Tree model. Once again, the superior forecasting power of the DST-Tree model is particularly evident.

6 Conclusions

In this paper we propose a novel smooth transition conditional heteroskedastic model that combines regression trees and GARCH models. Our model uses the interpretability of regression trees and the flexibility of smooth transition models. We have applied our new model to describe regime switches in the short-term interest rate series. We carefully address the estimation of such models, we derive the asymptotic properties of the quasi-maximum likelihood estimator, and we discuss the different modeling cycle strategies. When the model was applied to the US short-term interest rate we reached several interesting conclusions. First, the leading indicators for inflation and real activity are the most relevant predictors in characterizing the multiple regimes' structure. Second, the optimal model has three limiting regimes, with significantly different local conditional mean and variance dynamics. Third, there is some correspondence between NBER recessions/expansions and our limiting regimes. Finally, we investigate the forecasting accuracy of the new model's conditional mean and variance predictions, concluding that the new model in most cases significantly outperforms existing alternatives introduced in the literature.

A Proofs

Before proceeding to the proofs, we define our notation, as follows. First, set $\boldsymbol{\psi} = (\boldsymbol{\psi}_M', \boldsymbol{\psi}_V')'$, where $\boldsymbol{\psi}_M$ and $\boldsymbol{\psi}_V$ are the parameters of the conditional mean and variance, respectively and define, as in Section 3, $\mathbf{z}_t = (\varepsilon_{t-1}^2, h_{t-1}, \widetilde{\mathbf{x}}_t')'$. In addition, let model (2)–(3) be written as

$$y_t = g(\mathbf{x}_t; \boldsymbol{\psi}_M) + h(\mathbf{z}_t; \boldsymbol{\psi}_V)^{1/2} u_t \tag{14}$$

and set $g_t \equiv g(\mathbf{x}_t; \boldsymbol{\psi}_M)$ and $h_t \equiv h(\mathbf{z}_t; \boldsymbol{\psi}_V)$. Furthermore, write $\varepsilon_t \equiv \varepsilon_t(\boldsymbol{\psi}_M) = y_t - g_t$, let J and K be the number of parent and terminal nodes, respectively, and define $\boldsymbol{\pi}_i = (a_i, b_i, \boldsymbol{\lambda}_i')'$, $i = 1, \ldots, K$. Finally, to simplify notation define $B_{i,t} \equiv B_{\mathbb{J}i}(\mathbf{x}_t; \boldsymbol{\theta}_i)$, $i = 1, \ldots, K$ and $G_{j,t} \equiv G\left(x_{j,t}; \gamma_j, c_j\right)$, $j = 1, \ldots, J$.

Derivatives of the Log-likelihood Function

The first-order derivative of the log-likelihood function is given by

$$\frac{\partial \mathcal{L}_T(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \frac{1}{T} \sum_{t=1}^{T} \left[\frac{1}{2h_t} \left(\frac{\varepsilon_t^2}{h_t} - 1 \right) \frac{\partial h_t}{\partial \boldsymbol{\psi}_V} + \frac{\varepsilon_t}{h_t} \frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}_M} \right], \tag{15}$$

where

$$
\begin{aligned}
\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}_M} &= -\left[\widetilde{\mathbf{x}}_t' B_{1,t}, \ldots, \widetilde{\mathbf{x}}_t' B_{K,t}, \boldsymbol{\beta}_1' \widetilde{\mathbf{x}}_t \frac{\partial B_{1,t}}{\partial \boldsymbol{\theta}_1'}, \ldots, \boldsymbol{\beta}_K' \widetilde{\mathbf{x}}_t \frac{\partial B_{K,t}}{\partial \boldsymbol{\theta}_K'}\right]', \\
\frac{\partial h_t}{\partial \boldsymbol{\psi}_V} &= \sum_{k=1}^{t}\left[\prod_{j=k+1}^{t}\left(\sum_{i=1}^{K} b_i B_{i,t}\right)\right]\mathbf{w}_k + \left[\prod_{j=1}^{t}\left(\sum_{i=1}^{K} b_i B_{i,t}\right)\right]\frac{\partial h_0}{\partial \boldsymbol{\psi}_V'}, \\
\mathbf{w}_t &= \left[\mathbf{z}_t' B_{1,t}, \ldots, \mathbf{z}_t' B_{K,t}, \boldsymbol{\pi}_1' \mathbf{z}_t \frac{B_{1,t}}{\partial \boldsymbol{\theta}_1'}, \ldots, \boldsymbol{\pi}_K' \mathbf{z}_t \frac{\partial B_{K,t}}{\partial \boldsymbol{\theta}_K'}\right]', \text{ and} \\
\frac{\partial B_{i,t}}{\partial \boldsymbol{\theta}_i'} &= \left\{\sum_{j \in \mathbb{J}_i}\left[\frac{n_{i,j}\left(1+n_{i,j}\right)}{2} G_{j,t}^{\frac{n_{i,j}\left(1+n_{i,j}\right)}{2}-1} \times \left(1-G_{j,t}\right)^{(1-n_{i,j})(1+n_{i,j})}\right.\right. \\
&\quad \left. - \left(1-n_{i,j}\right)\left(1+n_{i,j}\right) G_{j,t}^{\frac{n_{i,j}\left(1+n_{i,j}\right)}{2}} \times \left(1-G_{j,t}\right)^{(1-n_{i,j})(1+n_{i,j})-1}\right]\frac{\partial G_{j,t}}{\partial \boldsymbol{\theta}_i'} \\
&\quad \left. \times \prod_{k \in \mathbb{J}_i, k \neq j} G_{j,t}^{\frac{n_{i,j}\left(1+n_{i,j}\right)}{2}}\left(1-G_{j,t}\right)^{(1-n_{i,j})(1+n_{i,j})}\right\} \\
&\quad \times \left[\prod_{j \notin \mathbb{J}_i} G_{j,t}^{\frac{n_{i,j}\left(1+n_{i,j}\right)}{2}}\left(1-G_{j,t}\right)^{(1-n_{i,j})(1+n_{i,j})}\right].
\end{aligned}
$$

The second order derivative is given by

$$
\begin{aligned}
\frac{\partial^2 \mathcal{L}_T(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} &= \left(\frac{\varepsilon_t^2}{h_t} - 1\right)\frac{1}{2h_t}\frac{\partial^2 h_t}{\partial \boldsymbol{\psi}_V \boldsymbol{\psi}_V'} - \frac{1}{2h_t^2}\left(2\frac{\varepsilon_t^2}{h_t} - 1\right)\frac{\partial h_t}{\partial \boldsymbol{\psi}_V}\frac{\partial h_t}{\partial \boldsymbol{\psi}_V'} \\
&\quad + \left(\frac{\varepsilon_t}{h_t^2}\right)\left(\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}_M}\frac{\partial h_t}{\partial \boldsymbol{\psi}_V'} + \frac{\partial h_t}{\partial \boldsymbol{\psi}_V}\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}_M'}\right) + \frac{1}{h_t}\left(\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}_M}\frac{\partial \varepsilon_t}{\partial \boldsymbol{\psi}_M'} + \varepsilon_t \frac{\partial^2 \varepsilon_t}{\partial \boldsymbol{\psi}_M \partial \boldsymbol{\psi}_M'}\right).
\end{aligned}
$$

Proof of Theorem 1

It is easy to see that model (14) is a continuous function in the parameter vector $\boldsymbol{\psi}$. Similarly, we can see that (14) is continuous in $\mathbf{x}_t$ and $\mathbf{z}_t$, and therefore is measurable, for each fixed value of $\boldsymbol{\psi}$.

Furthermore, under the stationarity requirement in Assumption 1 and the restrictions in Assumption 3, $\mathbb{E}\left[\sup_{\boldsymbol{\psi} \in \boldsymbol{\Psi}} |h_{u,t}|\right] < \infty$ and $\mathbb{E}\left[\sup_{\boldsymbol{\psi} \in \boldsymbol{\Psi}} |y_{u,t}|\right] < \infty$. By Jensen's inequality, it is clear that $\mathbb{E}\left[\sup_{\boldsymbol{\psi} \in \boldsymbol{\Psi}} |\ln |h_{u,t}||\right] < \infty$. Thus, $\mathbb{E}\left[|\ell_{u,t}(\boldsymbol{\psi})|\right] < \infty \; \forall \, \boldsymbol{\psi} \in \boldsymbol{\Psi}$.

Let $h_{0,t}$ be the true conditional variance and $\varepsilon_{0,t} = h_{0,t}^{1/2} u_t$. In order to show that $\mathcal{L}(\boldsymbol{\psi})$

is uniquely maximized at $\boldsymbol{\psi}_0$, rewrite the maximization problem as

$$\max_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left[\mathcal{L}(\boldsymbol{\psi}) - \mathcal{L}(\boldsymbol{\psi}_0)\right] = \max_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left\{\mathbb{E}\left[\ln\left(\frac{h_{0,t}}{h_{u,t}}\right) - \frac{\varepsilon_{u,t}^2}{h_{u,t}} + 1\right]\right\}. \tag{16}$$

Writing $\varepsilon_{u,t} = \varepsilon_{u,t} - \varepsilon_{0,t} + \varepsilon_{0,t}$, equation (16) becomes

$$\begin{aligned}\max_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left[\mathcal{L}(\boldsymbol{\psi}) - \mathcal{L}(\boldsymbol{\psi}_0)\right] &= \max_{\boldsymbol{\psi}\in\boldsymbol{\Psi}}\Bigg\{\mathbb{E}\left[\ln\left(\frac{h_{0,t}}{h_{u,t}}\right) - \frac{h_{0,t}}{h_{u,t}} + 1\right] - \mathbb{E}\left[\frac{\left[\varepsilon_{u,t} - \varepsilon_{0,t}\right]^2}{h_{u,t}}\right] \\ &\qquad - \mathbb{E}\left[\frac{2u_t h_{0,t}^{1/2}\left(\varepsilon_{u,t} - \varepsilon_{0,t}\right)}{h_{u,t}}\right]\Bigg\} \qquad (17) \\ &= \max_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left\{\mathbb{E}\left[\ln\left(\frac{h_{0,t}}{h_{u,t}}\right) - \frac{h_{0,t}}{h_{u,t}} + 1\right] - \mathbb{E}\left[\frac{\left[\varepsilon_{u,t} - \varepsilon_{0,t}\right]^2}{h_{u,t}}\right]\right\},\end{aligned}$$

where $\mathbb{E}\left[\frac{2u_t h_{0,t}^{1/2}(\varepsilon_{u,t}-\varepsilon_{0,t})}{h_{u,t}}\right] = 0$ by the Law of Iterated Expectations.

Note that, for any $x > 0$, $m(x) = \ln(x) - x \leq 0$, so that

$$\mathbb{E}\left[\ln\left(\frac{h_{0,t}}{h_{u,t}}\right) - \frac{h_{0,t}}{h_{u,t}}\right] \leq 0.$$

Furthermore, $m(x)$ is maximized at $x = 1$. If $x \neq 1$, $m(x) < m(1)$, implying that $\mathbb{E}[m(x)] \leq \mathbb{E}[m(1)]$, with equality only if $x = 1$ a.s.. However, this will occur only if $\frac{h_{0,t}}{h_{u,t}} = 1$, a.s.. In addition,

$$\mathbb{E}\left[\frac{\left[\varepsilon_{u,t} - \varepsilon_{0,t}\right]^2}{h_{u,t}}\right] = 0$$

if and only if $\varepsilon_{u,t} = \varepsilon_{0,t}$. Hence, $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. This completes the proof. ■

Proof of Theorem 2

Following White (1994), Theorem 3.5, $\widehat{\boldsymbol{\psi}}_{u,T} \overset{a.s.}{\to} \boldsymbol{\psi}_0$ if the following conditions hold: (1) The parameter space $\boldsymbol{\Psi}$ is compact; (2) $\mathcal{L}_{u,T}(\boldsymbol{\psi})$ is continuous in $\boldsymbol{\psi} \in \boldsymbol{\Psi}$. Furthermore, $\mathcal{L}_{u,T}(\boldsymbol{\psi})$ is a measurable function of y_t, $\mathbf{x}_t$, and $\mathbf{z}_t$, $t = 1, \ldots, T$, for all $\boldsymbol{\psi} \in \boldsymbol{\Psi}$; (3)$\mathcal{L}(\boldsymbol{\psi})$

has a unique maximum at $\boldsymbol{\psi}_0$; and (4) $\lim\limits_{T\to\infty} \sup\limits_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} |\mathcal{L}_{u,T}(\boldsymbol{\psi}) - \mathcal{L}(\boldsymbol{\psi})| = 0$, *a.s.*.

Condition (1) holds by assumption. Theorem 1 shows that Conditions (2) and (3) are satisfied. By Lemma 1, Condition (4) is also satisfied. Thus, $\widehat{\boldsymbol{\psi}}_{u,T} \overset{a.s.}{\to} \boldsymbol{\psi}_0$. Lemma 2 shows that $\lim\limits_{T\to\infty} \sup\limits_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} |\mathcal{L}_{u,T}(\boldsymbol{\psi}) - \mathcal{L}_T(\boldsymbol{\psi})| = 0\, a.s.$, implying that $\widehat{\boldsymbol{\psi}}_T \overset{a.s.}{\to} \boldsymbol{\psi}_0$. This completes the proof. ■

Proof of Theorem 3

We start by proving asymptotic normality of the QMLE using the unobserved log-likelihood. When this is shown, the proof using the observed log-likelihood is immediate by Lemmas 2 and 4. According to Theorem 6.4 in White (1994), to prove the asymptotic normality of the QMLE we need the following conditions in addition to those stated in the proof of Theorem 2: (5) The true parameter vector $\boldsymbol{\psi}_0$ is interior to $\boldsymbol{\Psi}$; (6)the matrix

$$\mathbf{A}_T(\boldsymbol{\psi}) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial^2 \ell_t(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\right)$$

exists *a.s.* and is continuous in $\boldsymbol{\Psi}$; (7) the matrix $\mathbf{A}_T(\boldsymbol{\psi}) \overset{a.s.}{\to} \mathbf{A}(\boldsymbol{\psi}_0)$, for any sequence $\boldsymbol{\psi}_T$, such that $\boldsymbol{\psi}_T \overset{a.s.}{\to} \boldsymbol{\psi}_0$; and (8) the score vector satisfies

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\frac{\partial \ell_t(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right) \overset{d}{\to} \mathsf{N}(\mathbf{0}, \mathbf{B}(\boldsymbol{\psi}_0)).$$

Condition (5) is satisfied by assumption. Condition (6) follows from the fact that $\ell_t(\boldsymbol{\psi})$ is differentiable of order two on $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, and the stationarity of the DST-Tree model. The non-singularity of $\mathbf{A}(\boldsymbol{\psi}_0)$ and $\mathbf{B}(\boldsymbol{\psi}_0)$ follows from Lemma 4. Furthermore, Lemmas 3 and 5 implies that Condition (7) is satisfied. In Lemma 6 below, we prove that condition (8) is also satisfied. This completes the proof. ■

B Lemmas

Lemma 1. *Suppose that y_t follows a DST-Tree model satisfying the restrictions in Assumptions 1 and 3, and stationarity holds. Then,* $\lim_{T\to\infty}\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} |\mathcal{L}_{u,T}(\boldsymbol{\psi}) - \mathcal{L}(\boldsymbol{\psi})| = 0$, *a.s.*.

Proof. Set $g(\mathbf{Y}_t, \boldsymbol{\psi}) = \ell_{u,t}(\boldsymbol{\psi}) - \mathbb{E}\left[\ell_{u,t}(\boldsymbol{\psi})\right]$, where $\mathbf{Y}_t = \left[y_t, \mathbf{x}_t', \mathbf{x}_{t-1}', \ldots\right]'$. Hence, $\mathbb{E}\left[g(\mathbf{Y}_t, \boldsymbol{\psi})\right] = 0$. Under stationarity, it is clear that $\mathbb{E}\left[\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} |g(\mathbf{Y}_t, \boldsymbol{\psi})|\right] < \infty$. Furthermore, as $g(\mathbf{Y}_t, \boldsymbol{\psi})$ is strictly stationary and ergodic, then, by Theorem 3.1 in Ling and McAleer (2003), it follows that $\lim_{T\to\infty}\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left|T^{-1}\sum_{t=1}^{T} g(\mathbf{Y}_t, \boldsymbol{\psi})\right| = 0$, *a.s.*. This completes the proof. ■

Lemma 2. *Under the assumptions of Lemma 1,* $\lim_{T\to\infty}\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} |\mathcal{L}_{u,T}(\boldsymbol{\psi}) - \mathcal{L}_T(\boldsymbol{\psi})| = 0, a.s.$.

Proof. Set $\overline{a}(\mathbf{x}_t) = \sum_{i=1}^{K} a_i B_{i,t}$, $\overline{b}(\mathbf{x}_t) = \sum_{i=1}^{K} b_i B_{i,t}$, $\overline{\boldsymbol{\lambda}}(\mathbf{x}_t) = \sum_{i=1}^{K} \boldsymbol{\lambda}_i B_{i,t}$, and write

$$\begin{aligned}
h_t &= \overline{a}(\mathbf{x}_t)\varepsilon_{t-1}^2 + \overline{b}(\mathbf{x}_t)h_{t-1} + \overline{\boldsymbol{\lambda}}(\mathbf{x}_t)'\widetilde{\mathbf{x}}_t \\
&= \sum_{i=1}^{t}\left\{\left[\overline{a}(\mathbf{x}_i)\varepsilon_{i-1}^2 + \overline{\boldsymbol{\lambda}}(\mathbf{x}_t)'\widetilde{\mathbf{x}}_t\right]\left[\prod_{j=i+1}^{t}\overline{b}(\mathbf{x}_j)\right]\right\} + \left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right]h_0, \text{ and} \\
h_{u,t} &= \overline{a}(\mathbf{x}_t)\varepsilon_{t-1}^2 + \overline{b}(\mathbf{x}_t)h_{t-1} + \overline{\boldsymbol{\lambda}}(\mathbf{x}_t)'\widetilde{\mathbf{x}}_t \\
&= \sum_{i=1}^{t}\left\{\left[\overline{a}(\mathbf{x}_i)\varepsilon_{u,i-1}^2 + \overline{\boldsymbol{\lambda}}(\mathbf{x}_t)'\widetilde{\mathbf{x}}_t\right]\left[\prod_{j=i+1}^{t}\overline{b}(\mathbf{x}_j)\right]\right\} + \left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right]h_{u,0}
\end{aligned} \tag{18}$$

Hence,

$$h_{u,t} - h_t = \overline{a}(\mathbf{x}_1)\left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right]\left(\varepsilon_{u,0}^2 - \varepsilon_0^2\right) + \left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right](h_{u,0} - h_0)$$

and

$$|h_{u,t} - h_t| \leq \overline{a}(\mathbf{x}_1)\left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right]\left|\left(\varepsilon_{u,0}^2 - \varepsilon_0^2\right)\right| + \left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right]\left|(h_{u,0} - h_0)\right|,$$

as $\overline{a}(\mathbf{x}_t) > 0$ and $\overline{b}(\mathbf{x}_t) > 0$, $\forall\, t$ by assumption and, under the stationarity of the process,

$$\left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right] \overset{a.s.}{\to} 0.$$

Furthermore, $h_{u,0}$ and $\varepsilon^2_{0,u}$ are well defined, as

$$\mathsf{Pr}\left[\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}}\left(h_{u,0} > K_1\right)\right] \rightarrow 0 \, \text{as} \, K_1 \rightarrow \infty, \text{ and } \mathsf{Pr}\left[\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}}\left(\varepsilon^2_{u,0} > K_2\right)\right] \rightarrow 0 \, \text{as} \, K_2 \rightarrow \infty.$$

Thus,

$$\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} |h_t - h_{u,t}| \leq K_h \rho_1^t, \; a.s., \text{and}$$

$$\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left|\varepsilon_0^2 - \varepsilon^2_{u,0}\right| \leq K_\varepsilon \rho_2^t, \; a.s.,$$

where K_h and K_ε are positive and finite constants, $0 < \rho_1 < 1$, and $0 < \rho_2 < 1$. Hence, as $h_t > \delta$,δ a positive and finite constant, and $\log(x) \leq x - 1$,

$$\begin{aligned}\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} |\ell_t - \ell_{u,t}| &\leq \sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left[\varepsilon_t^2 \left|\frac{h_{u,t} - h_t}{h_t h_{u,t}}\right| + \left|\log\left(1 + \frac{h_t - h_{u,t}}{h_{u,t}}\right)\right|\right] \\ &\leq \sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left(\frac{1}{\delta^2}\right) K_h \rho_1^t \varepsilon_t^2 + \sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left(\frac{1}{\delta}\right) K_h \rho_1^t, \; a.s..\end{aligned}$$

Following the same arguments as in the proof of Theorems 2.1 and 3.1 in Francq and Zakoïan (2004), it can be shown that $\lim_{T\rightarrow\infty} \sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} |\mathcal{L}_{u,T}(\boldsymbol{\psi}) - \mathcal{L}_T(\boldsymbol{\psi})| = 0, a.s..$ This completes the proof. ■

Lemma 3. *Under the conditions of Theorem 3,*

$$\mathbb{E}\left[\left|\left.\frac{\partial \ell_t(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}_0}\right|\right] < \infty, \tag{19}$$

$$\mathbb{E}\left[\left|\left.\frac{\partial \ell_t(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}_0} \left.\frac{\partial \ell_t(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\right|_{\boldsymbol{\psi}_0}\right|\right] < \infty, \; \textit{and} \tag{20}$$

$$\mathbb{E}\left[\left|\left.\frac{\partial^2 \ell_t(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'}\right|_{\boldsymbol{\psi}_0}\right|\right] < \infty. \tag{21}$$

Proof. As the derivatives of the transition function are bounded, if stationarity holds, the derivatives of the likelihood function are clearly bounded. Hence, the remainder of the proof follows from the proof of Theorem 3.2 (part (i)) in Francq and Zakoïan (2004). This completes the proof. ■

LEMMA 4. *Under the conditions of Theorem 3,* $\mathbf{A}(\boldsymbol{\psi}_0)$ *and* $\mathbf{B}(\boldsymbol{\psi}_0)$ *are nonsingular and, when* u_t *has a symmetric distribution, are block-diagonal.*

PROOF. First, note that the restrictions in Assumption 3 guarantee the minimality (identifiability) of the DST-Tree model considered in this paper. Therefore, the results follow from the proof of Theorem 3.2 (part (*ii*)) in Francq and Zakoïan (2004). This completes the proof. ■

LEMMA 5. *Under the conditions of Theorem 3,*

$$\begin{aligned}
&\text{(a)} \qquad \lim_{T\to\infty} \sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left\| \frac{1}{T}\sum_{t=1}^{T} \left[\frac{\partial \ell_{u,t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} - \frac{\partial \ell_{t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \right] \right\| = \mathbf{0}, \ a.s., \\
&\text{(b)} \qquad \lim_{T\to\infty} \sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left\| \frac{1}{T}\sum_{t=1}^{T} \left[\frac{\partial^2 \ell_{u,t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}'} - \frac{\partial^2 \ell_{t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}'} \right] \right\| = \mathbf{0}, \ a.s, \ \text{and} \\
&\text{(c)} \qquad \lim_{T\to\infty} \sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \left\| \frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 \ell_{u,t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}'} - \mathbb{E}\left[\frac{\partial^2 \ell_{u,t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}'} \right] \right\| = \mathbf{0}, \ a.s..
\end{aligned}$$

PROOF. First, assume that h_0 and $h_{u,0}$ are fixed constants and write

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\psi}}\left(h_{u,t}-h_t\right) = \Bigg[&\frac{\partial}{\partial \boldsymbol{\beta}_1'}\left(h_{u,t}-h_t\right), \ldots, \frac{\partial}{\partial \boldsymbol{\beta}_K'}\left(h_{u,t}-h_t\right), \frac{\partial}{\partial \boldsymbol{\pi}_1'}\left(h_{u,t}-h_t\right), \ldots, \frac{\partial}{\partial \boldsymbol{\pi}_K'}\left(h_{u,t}-h_t\right), \\
&\frac{\partial}{\partial \boldsymbol{\theta}_1'}\left(h_{u,t}-h_t\right), \ldots, \frac{\partial}{\partial \boldsymbol{\theta}_J'}\left(h_{u,t}-h_t\right) \Bigg]',
\end{aligned}$$

where

$$\frac{\partial}{\partial \boldsymbol{\beta}_i'}\left(h_{u,t}-h_t\right) = 2\overline{a}(\mathbf{x}_1)\left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right]\left(\varepsilon_{u,0}\frac{\partial \varepsilon_{u,0}}{\partial \boldsymbol{\beta}_i} - \varepsilon_0\frac{\partial \varepsilon_0}{\partial \boldsymbol{\beta}_i}\right),$$

$$\frac{\partial}{\partial \boldsymbol{\pi}_i'}\left(h_{u,t}-h_t\right) = \left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right]\left(\frac{\partial h_{u,0}}{\partial \boldsymbol{\pi}_i} - \frac{\partial h_0}{\partial \boldsymbol{\pi}_i}\right),$$

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}_i'}\left(h_{u,t}-h_t\right) = {} & \left\{\frac{\partial \overline{a}(\mathbf{x}_1)}{\partial \boldsymbol{\theta}_i'}\left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right] + \overline{a}(\mathbf{x}_1)\frac{\partial}{\partial \boldsymbol{\theta}_i'}\left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right]\right\}\left(\varepsilon_{u,0}^2 - \varepsilon_0^2\right) \\ & + 2\overline{a}(\mathbf{x}_1)\left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right]\left(\varepsilon_{u,0}\frac{\partial \varepsilon_{u,0}}{\partial \boldsymbol{\theta}_i} - \varepsilon_0\frac{\partial \varepsilon_0}{\partial \boldsymbol{\theta}_i}\right) \\ & + \frac{\partial}{\partial \boldsymbol{\theta}_i}\left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right]\left(h_{u,0}-h_0\right) + \left[\prod_{j=1}^{t}\overline{b}(\mathbf{x}_j)\right]\left(\frac{\partial h_{u,0}}{\partial \boldsymbol{\pi}_i} - \frac{\partial h_0}{\partial \boldsymbol{\pi}_i}\right).\end{aligned}$$

It is clear that, under stationarity of the process, all the derivatives above are bounded. Hence, as in Francq and Zakoïan (2004), part (a) follows trivially. The proof of part (b) follows along similar lines. The proof of part (c) follows the same arguments as in the proof of Theorem 3.2 (part (v)) in Francq and Zakoïan (2004). This completes the proof. ■

Lemma 6. *Under the conditions of Theorem 3,*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left.\frac{\partial \ell_t(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}_0} \xrightarrow{d} \mathsf{N}(\mathbf{0}, \mathbf{B}(\boldsymbol{\psi}_0)).$$

Proof. Let $S_T = \sum_{t=1}^{T}\mathbf{c}'\nabla_0\ell_{u,t}$, where $\mathbf{c}$ is a constant vector. Then S_T is a martingale with respect to $\mathcal{F}_t$, the filtration generated by all past observations of y_t. By the given assumptions, $\mathbb{E}\left[S_T\right] > 0$. Using the central limit theorem of Stout (1974), $\sqrt{T}S_T \xrightarrow{d} \mathsf{N}\left(0, \mathbf{c}'\mathbf{B}(\boldsymbol{\psi}_0)\mathbf{c}\right)$. By the Cramer-Wold device, $\sqrt{T}\sum_{t=1}^{T}\left.\frac{\partial \ell_{u,t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}_0} \xrightarrow{d} \mathsf{N}\left(0, \mathbf{B}(\boldsymbol{\psi}_0)\right)$.

By Lemma 5, $\sqrt{T}\sum_{t=1}^{T}\left\|\left.\frac{\partial \ell_{u,t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}_0} - \left.\frac{\partial \ell_t(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}_0}\right\| \xrightarrow{a.s.} \mathbf{0}$. Thus, $\sqrt{T}\sum_{t=1}^{T}\left.\frac{\partial \ell_t(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}_0} \xrightarrow{d} \mathsf{N}(0, \boldsymbol{B}_0)$. This completes the proof. ■

References

Ang, A. and Piazzesi, M. (2003), A no-arbitrage vector autoregression of the term structure dynamics with macroeconomic and latent variables. *Journal of Monetary Economics* **50**, 745–787.

Ang, A., Dong, S. and Piazzesi, M. (2007). No-arbitrage Taylor rules. Working Paper, University of Chicago.

Audrino, F. (2006). Tree-structured multiple regimes in interest rates. *Journal of Business & Economic Statistics* **24**(3), 338–353.

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

Audrino, F. and De Giorgi, E. (2007). Beta regimes for the yield curve. *Journal of Financial Econometrics* **5** (3), 456–490.

Bansal, R., Tauchen, G., and Zhou, H. (2004). Regime shifts, risk premiums in the term structure, and the business cycle. *Journal of Business and Economic Statistics* **22** (4), 396-409.

Bansal, R. and Zhou, H. (2002). Term structure of interest rates with regime shifts. *Journal of Finance* **57** (5), 1997–2043.

Berk, R.A. (2008). *Statistical learning from a regression perspective*, Springer Series in Statistics, Springer, New York.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **36**, 105–139.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Annals of Statistics* **30**, 927–961.

Buya, A. and Stuetzle, W. (2006). Observations on Bagging. *Statistica Sinica* **16**, 323–351.

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

Davies, R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247-254.

Davies, R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33-43.

Diebold, F.X. and Mariano, R.S. (1995), Comparing predictive accuracy, *Journal of Business and Economic Statistics* **13**, 253–263.

Diebold, F.X., Rudebusch, G.D., and Aruoba, S.B. (2006). The macroeconomy and the yield curve: a dynamic latent factor approach. *Journal of Econometrics* **131**, 309–338.

Francq, C. and Zakoïan J.-M. (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH Processes. *Bernoulli* **10**, 605-637.

Gray, S.F. (1996), Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* **42**, 27–62.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics* **23**, 365–380.

Hansen, A.T. and Poulsen, R. (2000), A simple regime switching term structure model. *Finance and Stochastics* **4**, 409–429.

Hastie T., Tibshirani, R. and, Friedman, J. (2001). *The elements of statistical learning: data mining, inference and prediction.* Springer Series in Statistics, Springer, Canada.

Hillebrand, E., and Medeiros, M.C. (2007). Forecasting realized volatility models: the benefits of bagging and nonlinear specifications. Working paper series **547**, PUC-Rio (Brazil).

Inoue, A., and Kilian, L. (2004). Bagging time series models. Discussion paper **4333**, Centre for Economic Policy Research (CEPR).

Inoue, A., and Kilian, L. (2008). How useful is bagging in forecasting time series? A case study of US CPI inflation. *Journal of the American Statistical Association* **103**, 511–522.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* **17**, 1217–1241.

Ling, S. and McAleer, M. (2003). Asymptotic theory for a vector ARMA-GARCH model. *Econometric Theory* **19**, 280–310.

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

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. (2009). Modeling multiple regimes in financial volatility with a flexible coefficient GARCH(1,1) model. *Econometric Theory* **25**, 117–161.

Rudebusch, G.D. and Wu, T. (2004). A macro-finance model of the term structure, monetary policy, and the economy. Forthcoming in *Economic Journal.*

Stout, W.F. (1974). *Almost Sure Convergence.* Academic Press, New York.

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

Van Dijk, D., Teräsvirta, T. and Franses, P. (2002). Smooth transition autoregressive models A survey of recent developments. *Econometric Reviews* **21**, 1-47.

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.

Summary statistics of data

	Central moments				Autocorrelations		
	Mean	Stdev	Skew	Kurt	Lag 1	Lag 2	Lag 3
1 mth rates	5.2462	2.6496	1.1198	4.9472	0.9652	0.9376	0.9120
1 mth changes	0.0023	0.6953	1.0930	16.721	-0.1028	-0.0361	-0.0589
60 mth rates	6.6416	2.5365	0.9179	3.5773	0.9878	0.9739	0.9615
Spread	1.3944	1.1878	0.0353	3.8250	0.8449	0.7607	0.6774
CPI	4.0881	2.7835	1.3625	4.5441	0.9902	0.9761	0.9606
PPI	3.5130	4.4441	1.0462	4.5846	0.9761	0.9449	0.9159
HELP	83.169	25.369	0.1720	2.1040	0.9892	0.9786	0.9653
IP	3.0453	4.3952	0.7951	3.9041	0.9684	0.9178	0.8537
UE	1.3869	15.616	1.0880	4.2022	0.9550	0.9149	0.8566
GDP	6.8332	2.7445	0.0191	3.3684	0.9661	0.9324	0.8986

Table 1: The one-month yield is from the Fama CRSP treasury bill files. The 60 month yield is the annual zero coupon bond yield from the Fama CRSP bond files. Spread refers to the difference between long and short-term interest rates. The inflation measures CPI and PPI refer to CPI inflation and PPI (finished goods) inflation, respectively. We calculate the inflation measure at time t using $\log(P_t/P_{t-12})$ where P_t is the (seasonally adjusted) inflation index. The real activity measures HELP, IP, UE and GDP refer to the index of help wanted advertising in newspapers, the (seasonally adjusted) growth rate in industrial production, the unemployment rate, and the US gross domestic product, respectively. The growth rate in industrial production is calculated using $\log(I_t/I_{t-12})$ where I_t is the (seasonally adjusted) industrial production index. The sample period is January 1960 to December 2006, for a total of 564 observations.

DST-Tree local parameter estimates

		Number of regimes: 3 regimes	
Limiting Regimes	Parameter	Estimate	t |(p-value)
	α_1	0.2109	3.1297*
	β_1	-0.0586	-3.3020*
$\text{HELP}_{t-1} \leq 90.91$	a_1	≈ 0	≈ 0
	b_1	0.8977	2.8193*
	σ_1^2	0.0013	1.8126*
	α_2	-2.1159	-1.3761
$\text{HELP}_{t-1} > 90.91$,	β_2	-0.0807	-0.4259
$\text{CPI}_{t-1} \leq 1.467$	a_2	≈ 0	0.0001
	b_2	≈ 0	≈ 0
	σ_2^2	0.0369	2.1224*
	α_3	3.5026	2.6878*
$\text{HELP}_{t-1} > 90.91$,	β_3	-0.2703	-2.3732*
$\text{CPI}_{t-1} > 1.467$	a_3	0.2748	1.4551
	b_3	1.0015	3.6891*
	σ_3^2	0.0029	0.1766
Log-likelihood		-358.703	
LB_5^2		3.8051	(0.5778)
LB_{10}^2		9.6482	(0.4719)
LB_{15}^2		10.892	(0.7602)

Table 2: Local parameter estimates, limiting regimes' structure (that is, when the slope parameters $\gamma_k = \infty,\ k \in \mathbb{J}_i,\ i \in \mathbb{T}$), and related statistics for the double smooth transition tree (DST-Tree) model which uses the additional information included in the term structure and in other macroeconomic variables for prediction ($\mathbf{x}_t = (\Delta r_{t-1}, r_{t-1}, (\mathbf{x}_{t-1}^{\text{ex}})')'$). The sample period is January 1960 to December 2001, for a total of 504 monthly observations. t-statistics are based on heteroskedastic-consistent standard errors. Asterisks denote significance at the 5% level. LB_i^2 denotes the Ljung-Box statistic for serial correlation of the squared residuals out to i lags. p-values are reported in parentheses.
In the double smooth transition tree (DST-Tree) model: $y_t \mid \mathcal{F}_{t-1} = \Delta r_t \mid \mathcal{F}_{t-1} \sim \text{N}(\mu_t, h_t)$, with

$$\mu_t = \sum_{i \in \mathbb{T}} (\alpha_i + \beta_i r_{t-1}) B_{\mathbb{J}i}\left(\mathbf{x}_t; \boldsymbol{\theta}_i\right),$$

$$h_t = \sum_{i \in \mathbb{T}} \left(a_i \varepsilon_{t-1}^2 + b_i h_{t-1} + \sigma_i^2 r_{t-1}\right) B_{\mathbb{J}i}\left(\mathbf{x}_t; \boldsymbol{\theta}_i\right),$$

where the (probability) functions $B_{\mathbb{J}i}\left(\mathbf{x}_t; \boldsymbol{\theta}_i\right), i \in \mathbb{T}$, are given in (4).

Panel A: Forecasting performances: SPA tests

Model	Loglik	MSE-mean	MSE-variance
Global	−5.4947 (0.0001)	0.0464 (0.0001)	0.0071 (0.0471)
Global with macro	4.7607 (0.0089)	0.0680 (0.0158)	0.0095 (0.0228)
Bagged Global	7.3379 (0.0000)	0.0432 (0.0019)	0.0081 (0.0103)
Gray's RS	−4.1150 (0.0075)	0.0456 (0.0835)	0.0064 (0.0606)
RS with macro	−4.3733 (0.0098)	0.0451 (0.0521)	0.0055 (0.0463)
Bagged RS with macro	−4.1298 (0.0054)	0.0412 (0.3662)	0.0054 (0.0885)
Audrino's tree	−7.3686 (0.0401)	0.0475 (0.0039)	0.0057 (0.3241)
Bagged Audrino's tree	−14.756 (0.0166)	0.0399 (0.5889)	0.0049 (0.4608)
DST-Tree	−8.8808 (0.0109)	0.0517 (0.0631)	0.0056 (0.1871)
Bagged DST-Tree	−18.320 (0.6846)	0.0389 (0.6259)	0.0045 (0.8834)

Panel B: Forecasting performances: Diebold and Mariano tests

Alternative Model	Loglik	MSE-mean	MSE-variance
Bagged Global	−11.929 (0)	−2.3475 (0.0094)	−5.1461 (0)
Bagged RS with macro	−11.445 (0)	−1.1047 (0.1346)	−2.3462 (0.0095)
Bagged Audrino's tree	−4.0838 (0)	−0.8376 (0.2011)	−3.2600 (0.0005)

Table 3: The models considered in the analysis are: the classical global CIR-GARCH-type model, also including macro-variables as linear predictors in the conditional mean equation; the Markovian regime-switching (RS) model with and without macro-variables used to specify the transition probabilities; the tree-structured model proposed by Audrino (2006); the double smooth transition tree (DST-Tree) model; and the bagged versions of the best performing different model specifications. In Panel B, we consider pairwise comparisons of the bagged alternative specifications against the bagged DST-Tree model. Negative statistic values are in favor of the bagged DST-Tree model. Loglik refers to the out-of-sample negative log-likelihood, and MSE-mean and MSE-variance are the mean squared errors computed for predicting first and second conditional moments, respectively. *p*-values of superior predictive ability (SPA) tests (Panel A) and pairwise generalized Diebold and Mariano tests (Panel B) are reported in parentheses.

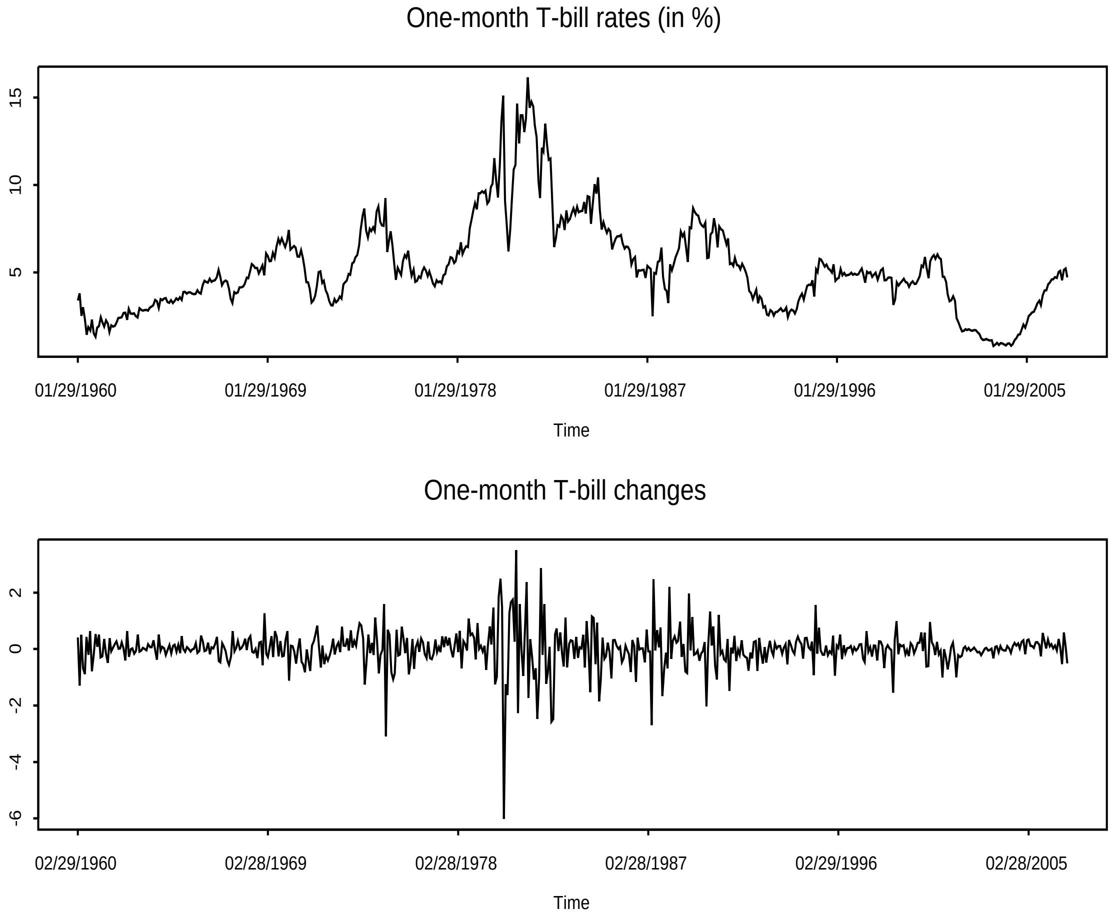

Figure 1: The top panel contains a time series of monthly one-month treasury-bill rates (in percentages). The first differences of this series are shown in the bottom panel. The sample period is January 1960 to December 2006, for a total of 564 observations.

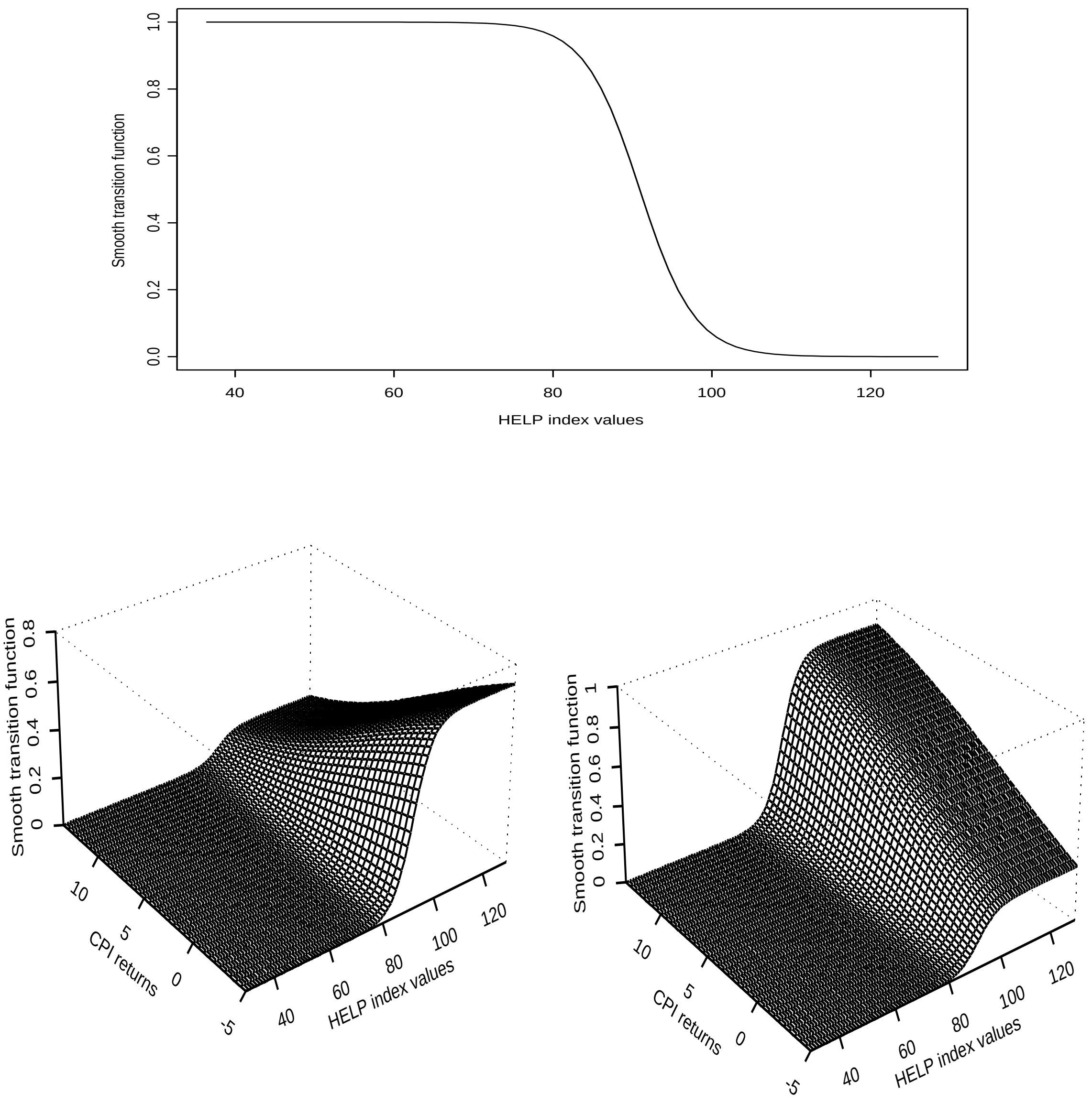

Figure 2: Probability functions associated with the three optimal limiting regimes (first regime top, second and third regimes bottom left and right, respectively) of the double smooth transition tree (DST-Tree) model. The in-sample period goes from January 1960 to December 2001, for a total of 504 observations.

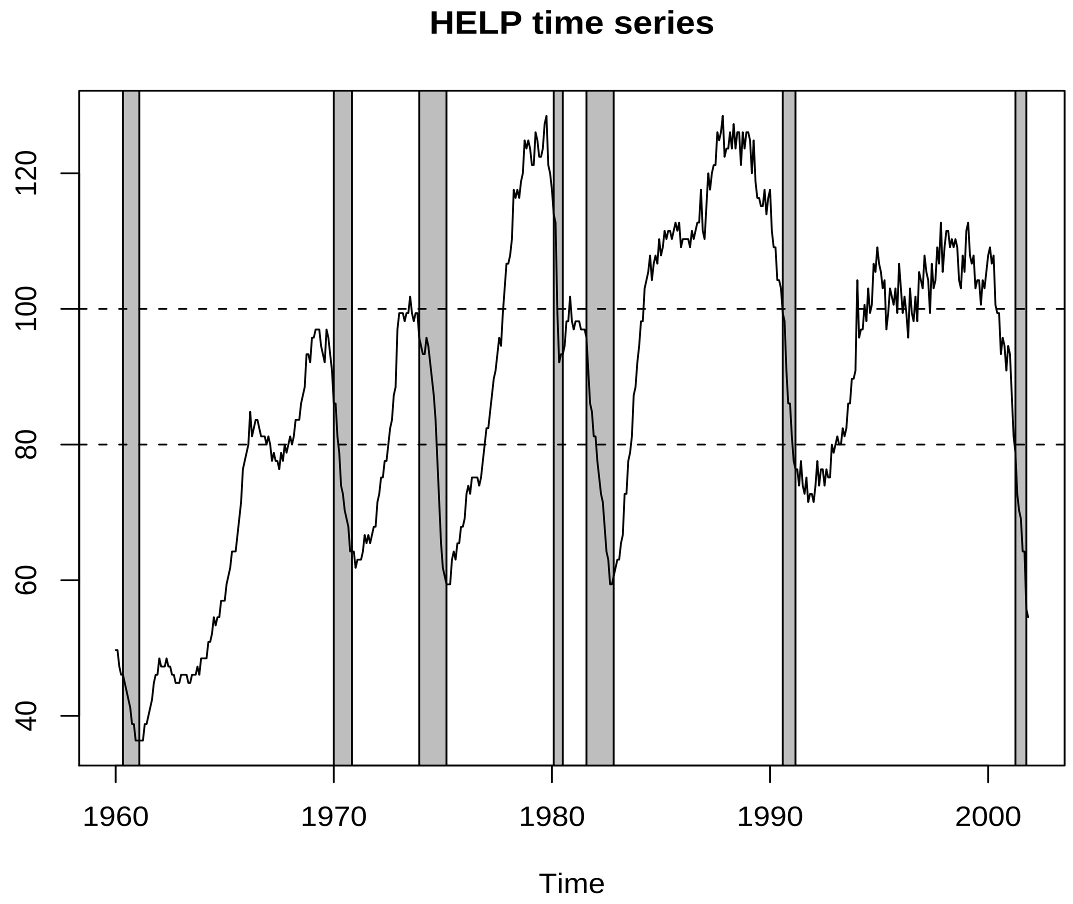

Figure 3: Help Wanted Advertising in Newspaper (HELP) time series for the period January 1960 to December 2001. Shaded NBER recession periods are overlaid to show regime correspondence with recessions/expansions. For values of the HELP index smaller (larger) than 80 (100) the dynamics of the short-term interest rate closely follow the local processes under regime 1 (regimes 2 and 3) given in Table 2.

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