Tree-Structured GARCH Models

Francesco Audrino and Peter Bühlmann*
ETH Zürich, Switzerland

Revised Version
February 2001

Abstract

We propose a new GARCH model with tree-structured multiple thresholds for volatility estimation in financial time series. The approach relies on the idea of a binary tree where every terminal node parameterizes a (local) GARCH model for a partition cell of the predictor space. Fitting of such trees is constructed within the likelihood framework for non-Gaussian observations: it is very different from the well-known CART procedure for regression which is based on residual sum of squares. Our strategy includes the classical GARCH model as a special case and allows one to increase model-complexity in a systematic and flexible way. We derive a consistency result and conclude by simulation and real data analysis that the new method has better predictive potential than other approaches.

Keywords: Conditional variance; Financial time series; GARCH model; Maximum likelihood; Threshold model; Tree model; Volatility.

1 Introduction

We propose a new method for estimating volatility in stationary financial time series. The real data examples of interest are daily log-returns $X_t = \log(P_t/P_{t-1})$, where P_t denotes the price of an asset at day t. Our modeling technique is parametric and potentially high-dimensional: it relies on estimating thresholds by using the idea of binary tree construction for partitioning a predictor space.

As a starting point, consider a nonparametric GARCH(1,1) model,

$$\begin{aligned} &X_t = \sigma_t Z_t \ (t \in \mathbb{Z}), \\ &\sigma_t^2 = f(X_{t-1}, \sigma_{t-1}^2), \ f : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+, \end{aligned} \tag{1.1}$$

where $(Z_t)_{t\in\mathbb{Z}}$ is a sequence of i.i.d. innovation variables with $\mathbb{E}[Z_t] = 0$, $\mathrm{Var}(Z_t) = 1$ and Z_t independent from $\{X_s; s < t\}$. The so-called volatility σ_t is then defined as

$$\sigma_t^2 = \mathrm{Var}(X_t|\mathcal{F}_{t-1}) = f(X_{t-1}, \sigma_{t-1}^2),$$

where $\mathcal{F}_{t-1}$ denotes the σ-algebra (the information) of the variables $\{X_s; s \leq t-1\}$. The restriction to model the squared volatility σ_t^2 as a function of the previous values X_{t-1} and σ_{t-1}^2 only is natural in finance. Note that it still generates a dependence of σ_t^2 from *all previous* observations $\{X_s; s < t\}$ due to the recursive definition of σ_{t-1}^2: this is the important difference

*Corresponding author. Address: Seminar für Statistik, LEO D12, ETH Zentrum, CH-8092 Zürich, Switzerland. E-mail: buhlmann@stat.math.ethz.ch.

between ARCH and GARCH models. Also, the implicit assumption in (1.1) that $\mathbb{E}[X_t|\mathcal{F}_{t-1}] \equiv 0$ is a reasonable approximation for many financial time series: the substantial modeling effort goes into the dominant volatility, although for real data, we usually subtract first a linear AR(1) estimate for the conditional mean. The simplest but often used example for (1.1) is the classical GARCH(1,1) model (Bollerslev, 1986),

$$f(x, \sigma^2) = \alpha_0 + \alpha_1 x^2 + \beta\sigma^2, \ \alpha_0, \alpha_1, \beta > 0. \tag{1.2}$$

Generally, the unknown function $f(\cdot, \cdot)$ in (1.1) may be nonlinear and even not smooth; for example, an asymmetry in financial trading with positive and negative return values from the past implies asymmetric or even discontinuous behavior for $f(x, y)$ around $x = 0$. Estimation of $f(\cdot, \cdot)$ in general is very difficult due to the non-observable volatility in the second argument. An iterative nonparametric estimation procedure has been proposed in Bühlmann and McNeil (1999). Its advantage is the general form for $f(\cdot, \cdot)$, but quite a few theoretical issues are not rigorously settled yet. Its disadvantages are mainly poor performance at edges, including the high values of volatility which are of particular interest in practice; lack of ability to deal with non-Gaussian observations and sensitivity to the choice of smoothing parameters. Our approach here is more in the spirit of a *sieve approximation* with parametric models for the nonparametric function $f(\cdot, \cdot)$. The approximation builds on the following principles:

(1) it includes the classical GARCH(1,1) model as a simple special case;

(2) it uses a binary tree type selection strategy to estimate thresholds (splits) for building up an approximating multiple threshold GARCH model. The binary tree construction, where every terminal node represents a (local) three-dimensional GARCH model, is based on the likelihood in model (1.1).

Item (1) has an important link to practice: there is a relatively strong belief that the classical GARCH(1,1) model is appropriate, despite its simplicity with only three parameters. Our tree-structured nested modeling strategy allows to verify this by using known selection techniques for nested models: as we will see, there is potential to improve upon the classical GARCH(1,1) for real data and we will quantify such gains in terms of prediction accuracy for volatility, rather than testing structural properties of $f(\cdot, \cdot)$.

The likelihood-driven tree method mentioned in item (2) marks an essential difference to CART (Breiman et al., 1984). Underlying an approximate normality assumption for observations, CART uses residual sum of squares. For financial data, the normality assumption for observations and the corresponding techniques is not appropriate and can result in very poor performance. Our approach resembles more the general tree fitting with the deviance criterion used by Clark and Pregibon (1993). Another difference to CART (or more general versions driven by deviance) is that our tree structured scheme for conditional variance estimation employs a three-dimensional (local) GARCH model in every terminal node from the binary tree; CART uses only one location parameter per node. Finally, our tree GARCH procedure models the function $f(\cdot, \cdot)$ in (1.1) and hence the *infinite past* in terms of the observations, whereas CART (or versions thereof) in autoregressive modeling deals with a p-dimensional predictor space ($p < \infty$) from finitely many lagged observations. Our methodology is also markedly different from autoregressive threshold models (SETAR) for conditional expectations (Tong, 1990), as we focus on the conditional variance of very non-Gaussian observations, and because we allow for non-Markovian models.

Extending the GARCH(1,1) model in (1.2) in the direction of adding potentially high parametric complexity has not yet been considered. Other versions of GARCH(1,1) with three or

four parameters and a GARCH model with one or two thresholds at *fixed* locations (Rabemananjara and Zakoian, 1993) have been proposed. But there seems to be no systematic flexible way to build up a class of models from the classical GARCH(1,1) in (1.2) to a potentially high dimensional approximation of the general model in (1.1). The paper here deals mainly with this latter task: we describe the methodology in Section 2, present a consistency result in Section 3 and demonstrate the new procedure on simulated and real data in Section 4.

2 Tree-structured GARCH estimation

We describe here our methodology for approximating $f(\cdot,\cdot)$ in (1.1) by a piecewise linear function. The novel part is the estimation of thresholds in $\mathbb{R}\times\mathbb{R}^+$, the start- and end-points of piecewise approximating functions. The working model is

$$X_t = \phi X_{t-1} + \sigma_t(\theta)Z_t,\ \sigma_t^2(\theta) = f_\theta(X_{t-1},\sigma_{t-1}^2(\theta))\ (t\in\mathbb{Z}), \tag{2.1}$$

with $\sigma_t(\theta)Z_t$ as in model (1.1), but the functional form $f(\cdot,\cdot)=f_\theta(\cdot,\cdot)$ now parameterized by a threshold function, see formula (2.2) below. We also add here a linear autoregressive term for estimating a conditional mean (being of minor importance for many financial time series). The function $f_\theta(\cdot,\cdot)$ is parameterized as binary tree structured GARCH(1,1). It involves a partition

$$\mathcal{P} = \{\mathcal{R}_1,\dots\mathcal{R}_k\},\ \cup_{j=1}^k\mathcal{R}_j = \mathbb{R}\times\mathbb{R}^+,\ \mathcal{R}_i\cap\mathcal{R}_j=\emptyset\ (i\neq j),$$

for the predictor space. For every partition cell $\mathcal{R}_j$, we employ a GARCH(1,1) model: the parametric form of the function then depends on $\mathcal{P}$,

$$f_\theta(x,\sigma^2) = f_\theta^{\mathcal{P}}(x,\sigma^2) = \sum_{j=1}^k \left(\alpha_{0,j}+\alpha_{1,j}x^2+\beta_j\sigma^2\right) I_{[(x,\sigma^2)\in\mathcal{R}_j]}, \tag{2.2}$$

where θ denotes the parameter set $\{\alpha_{0,j},\alpha_{1,j},\beta_j;\ j=1,\dots,k\}$. For $k=1$, we have the classical GARCH(1,1) model from (1.2). As we will discuss in section 2.1, the partition $\mathcal{P}=\{\mathcal{R}_1,\dots,\mathcal{R}_k\}$ is constructed from a binary tree: every terminal node represents a rectangular partition cell $\mathcal{R}_j$ whose edges are determined by thresholds.

FIGURE 1 ABOUT HERE.

Figure 1 represents an example of a binary tree partition of the predictor space $\mathbb{R}\times\mathbb{R}^+ = \{(x,\sigma^2); x\in\mathbb{R},\sigma^2\in\mathbb{R}^+\}$: the partition cells, corresponding to terminal nodes in the tree, are $\mathcal{R}_1=\{(x,\sigma^2); x\le d_1\}$, $\mathcal{R}_2=\{(x,\sigma^2); x>d_1 \text{ and } \sigma^2\le d_2\}$, $\mathcal{R}_3=\{(x,\sigma^2); x>d_1 \text{ and } \sigma^2>d_2\}$.

The negative log-likelihood in the working model (2.1) is

$$-\ell(\phi,\theta;X_2^n) = -\sum_{t=2}^n \log\left[\sigma_t^{-1}(\theta)p_Z\left\{\frac{X_t-\phi X_{t-1}}{\sigma_t(\theta)}\right\}\right], \tag{2.3}$$

where $p_Z(\cdot)$ denotes the density of the innovation Z_t. The log-likelihood is always considered conditional on X_1 and some reasonable starting value $\sigma_1^2(\theta)$, e.g. $\sigma_1^2(\theta)=\mathrm{Var}(X_1)$.

The strategy of our flexible tree-structured GARCH estimation is as follows.

(a) The minimizing criterion is always the negative log-likelihood in (2.3) with innovation density $p_Z(\cdot)$, either specified (e.g. standard normal) or of parametric form such as scaled Student's t with unknown degrees of freedom. The parametric form of $\sigma_t(\theta)$ in (2.3) is always of the form of a threshold model in (2.2). As we will see in Section 3, the model and innovation density in (2.1) do not need to be true for good approximating properties.

(b) Optimization with threshold functions in item (a) becomes an estimation and model selection problem. The former is done by maximum likelihood. For the latter, an exhaustive search is computationally prohibitive and we propose a tree-structured partial search: within a data-determined tree structure, the optimal model is estimated using the AIC criterion.

2.1 Forward entering of thresholds: growing the binary tree

Forward entering of threshold variables, using a binary tree construction, induces a partition for $\mathbb{R} \times \mathbb{R}^+$ as follows. A first threshold $d_1 \in \mathbb{R}$ or $\mathbb{R}^+$ together with a component index $\iota_1 \in \{1, 2\}$ partitions

$$\mathbb{R} \times \mathbb{R}^+ = \mathcal{R}_{left} \cup \mathcal{R}_{right},$$

where $\mathcal{R}_{left} = \{(x, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+;\ (x, \sigma^2)_{\iota_1} \leq d_1\}$ and $\mathcal{R}_{right}$ analogously but with the relation '$>$' instead. Then, one of the partition cells $\mathcal{R}_{left}$ or $\mathcal{R}_{right}$ is again partitioned with a threshold d_2 and a component index ι_2 in the same fashion. We iterate this procedure: for the mth iteration step, we specify a pair (d_m, ι_m) and an existing partition cell for further refinement by splitting it into two cells. The refinement of an existing partition $\mathcal{P}^{(old)}$ is always constructed according to the following rule:

$$\begin{aligned}
&\mathcal{P}^{(old)} = \cup_j \mathcal{R}_j \text{ an existing partition } \rightarrow \text{ pick an element } \mathcal{R}_{j^*} \in \mathcal{P}^{(old)} \\
&\rightarrow \text{ split } \mathcal{R}_{j^*} = \mathcal{R}_{j^*,left} \cup \mathcal{R}_{j^*,right} \text{ according to a pair } (d, \iota) \in \mathbb{R} \times \{1, 2\} \\
&\rightarrow \ \mathcal{P}^{(new)} = \cup_{j \neq j^*} \mathcal{R}_j \cup (\mathcal{R}_{j^*,left} \cup \mathcal{R}_{j^*,right}),
\end{aligned} \tag{2.4}$$

with (d, ι) describing a threshold and component index, $\mathcal{R}_{j^*,left} = \{(x, \sigma^2) \in \mathcal{R}_{j^*} \subset \mathbb{R} \times \mathbb{R}^+; (x, \sigma^2)_\iota \leq d\}$ and analogously for $\mathcal{R}_{j^*,right}$ with the relation '$>$'. The whole procedure then produces a partition $\mathcal{P} = \{\mathcal{R}_1, \ldots, \mathcal{R}_k\}$ which can be described as a binary tree whose terminal nodes represent the partition cells, see also Figure 1. This is conceptually as in CART (Breiman et al., 1984). But as discussed below, the data-driven construction of a tree-structured partition and estimation of parameters are very different.

Our algorithm below constructs a binary tree, corresponding to a partition of $\mathbb{R} \times \mathbb{R}^+$, by optimizing reduction of a conditional negative log-likelihood.

Step 1. Compute the negative log-likelihood from (2.3) without partitioning (i.e. in the partition $\mathcal{P}_{opt}^{(0)} = \mathbb{R} \times \mathbb{R}^+$) with

$$f_{\theta^{(0)}}^{\mathcal{P}_{opt}^{(0)}}(x, \sigma^2) = \alpha_0 + \alpha_1 x^2 + \beta\sigma^2,\ \theta^{(0)} = (\alpha_0, \alpha_1, \beta),$$

and derive from it the maximum likelihood estimates $\hat{\phi}^{(0)}, \hat{\theta}^{(0)}$ using a quasi-Newton method, cf. Nocedal and Wright (1999). Set $m = 0$.

Step 2. Increment m by one. Search for the best refined partition $\mathcal{P}_{opt}^{(m)}$ by binary splitting of a cell from $\mathcal{P}_{opt}^{(m-1)}$ as described in (2.4). The details are as follows:
(I) Given $\mathcal{P}_{opt}^{(m-1)} = \{\mathcal{R}_1, \ldots, \mathcal{R}_m\}$, consider a new partition $\mathcal{P}^{(m)}$, where one partition cell $\mathcal{R}_{j^*} \in \mathcal{P}^{(m-1)}$ is split into $\mathcal{R}_{j^*} = \mathcal{R}_{j^*,left} \cup \mathcal{R}_{j^*,right}$ as in (2.4). The function associated with $\mathcal{P}^{(m)}$ is

$$\begin{aligned}
f_{(\theta^{(m-1)\setminus *}, \theta^*)}^{\mathcal{P}^{(m)}}(x, \sigma^2) \ &= \ \sum_{j \neq j^*} (\alpha_{0,j} + \alpha_{1,j} x^2 + \beta_j \sigma^2) I_{[(x,\sigma^2) \in \mathcal{R}_j]} \\
&+ \sum_{i \in \{j^*_{left}, j^*_{right}\}} (\alpha^*_{0,i} + \alpha^*_{1,i} x^2 + \beta^*_i \sigma^2) I_{[(x,\sigma^2) \in \mathcal{R}_i]},
\end{aligned} \tag{2.5}$$

where

$$\theta^{(m-1)\setminus *} = \{\alpha_{0,j}, \alpha_{1,j}, \beta_j;\ j = 1, \ldots, m,\ j \neq j^*\} \in (\mathbb{R}^+)^{3(m-1)},$$
$$\theta^* = \{\alpha^*_{0,i}, \alpha^*_{1,i}, \beta^*_i;\ i \in \{j^*_{left}, j^*_{right}\}\} \in (\mathbb{R}^+)^6.$$

(II) Compute the minimal negative conditional log-likelihood in the refined partition $\mathcal{P}^{(m)}$, holding the parameter vector $\hat{\phi}^{(m-1)}, \hat{\theta}^{(m-1)\setminus *}$ fixed,

$$\min_{\theta^*} \left(-\ell^{\mathcal{P}^{(m)}}(\hat{\phi}^{(0)}, (\hat{\theta}^{(m-1)\setminus *}, \theta^*); X_2^n) \right), \tag{2.6}$$

by numerical minimization over θ^* using a quasi-Newton method. Here, $-\ell^{\mathcal{P}^{(m)}}$ is as in (2.3) with the function $f^{\mathcal{P}^{(m)}}(\cdot,\cdot)$ from (2.5). For this numerical minimization over θ^*, use as starting values for the parameters θ^* in both new cells $\mathcal{R}_{j^*,left}, \mathcal{R}_{j^*,right}$ the components of $\hat{\theta}^{(m-1)}$ corresponding to cell $\mathcal{R}_{j^*}$.
(III) Optimize (2.6) by varying $\mathcal{P}^{(m)}$ in (I) and recomputing (II). Denote the optimal refined partition by $\mathcal{P}^{(m)}_{opt}$.

Step 3. Compute the maximum likelihood in partition $\mathcal{P}^{(m)}_{opt}$. Minimize with a quasi-Newton method the negative log-likelihood in (2.3) with $f^{\mathcal{P}^{(m)}_{opt}}(\cdot,\cdot)$ from (2.2) to obtain $(\hat{\phi}^{(m)}, \hat{\theta}^{(m)})$. For that purpose, use the starting values $\hat{\phi}^{(m-1)}$, $\hat{\theta}^{(m-1)\setminus *}$ and the minimizer $\hat{\theta}^*$ in (2.6) which is computed in Step 2.

Step 4. Repeat Steps 2 and 3 until $m = M$. This yields a partition $\mathcal{P}^{(M)}_{opt}$ corresponding to a large binary tree equipped with parameter estimates $(\hat{\phi}^{(M)}, \hat{\theta}^{(M)})$.

The value M, corresponding to $M+1$ partition cells or terminal nodes in the binary tree, is pre-specified in advance such that the binary tree is sufficiently large. With financial return data, choosing M around 6 is often appropriate. The search for the splitting value in Step 2 is done on a grid: we propose grid-points that are empirical α-quantiles of the data with $\alpha = i/\text{mesh},\ i = 1, \ldots, \text{mesh} - 1$. We typically choose mesh = 8 or 16.

The conditional log-likelihood in (2.6) yields a substantial computational short-cut compared to a full likelihood approach. For every given partition $\mathcal{P}^{(m)}$, the numerical nonlinear minimization in (2.6) involves only the 6-dimensional parameter θ^*. Since our algorithm searches over many candidate partitions $\mathcal{P}^{(m)}$ in every iteration step m, a relatively fast nonlinear minimization is important. Finding the best split is thus determined by maximal reduction of the negative conditional log-likelihood. The parameter estimates in Step 3 are computed from the full likelihood. For $(\hat{\phi}^{(m)}, \hat{\theta}^{(m)})$ we take advantage of the fact that the starting values specified in Step 3 are very reasonable for obtaining a reliable and fast maximum likelihood estimate in a possibly high-dimensional parameter space.

2.2 Pruning the tree

The binary tree, or the partition $\mathcal{P}^{(M)}_{opt}$, constructed in Section 2.1 is too large, or too fine, respectively. We correct by pruning: we search for a best subtree (with respect to AIC in (2.7) below), which is often computationally feasible since M around 6 is usually large enough for financial time series. Denote by τ the set of all binary subtrees from $\mathcal{P}^{(M)}_{opt}$: its elements are denoted by $\mathcal{P}_i$. Note that τ is generally larger than the set $\{\mathcal{P}^{(0)}_{opt}, \mathcal{P}^{(1)}_{opt}, \ldots, \mathcal{P}^{(M)}_{opt}\}$.

For every $\mathcal{P}_i$, we compute the maximum likelihood estimates $(\hat{\phi}^{\mathcal{P}_i}, \hat{\theta}^{\mathcal{P}_i})$ with a quasi-Newton method, according to (2.3) with $f^{\mathcal{P}_i}(\cdot,\cdot)$ of the form (2.2). Note that reasonable starting values

are again at hand by going backwards from $(\hat{\phi}^{(M)}, \hat{\theta}^{(M)})$ in a stage-wise manner. We then consider the penalized negative log-likelihood or AIC statistic

$$\mathrm{AIC}(\mathcal{P}_i) = -2\ell(\hat{\phi}^{\mathcal{P}_i}, \hat{\theta}^{\mathcal{P}_i}; X_2^n) + 2(\dim(\hat{\theta}^{\mathcal{P}_i}) + 1) \tag{2.7}$$

as a measure for predictive performance. The additional contribution 1 in the penalty term arises whenever the conditional expectation parameter ϕ in (2.1) is estimated. Choose the binary tree, or the partition $\hat{\mathcal{P}}$, minimizing (2.7). The final tree-structured GARCH model is thus given by (2.1) with $\hat{\phi}^{\hat{\mathcal{P}}}$, and $f^{\hat{\mathcal{P}}}_{\hat{\theta}^{\hat{\mathcal{P}}}}(\cdot,\cdot)$ as in (2.2) based on partition $\hat{\mathcal{P}}$.

The AIC statistic in (2.7) can be replaced by any other sensible model selection criterion. We have experimented with two other versions but found that overall performance with AIC is satisfactory.

3 Consistency

We give here some supporting asymptotics for a threshold model specified by (2.1) with Gaussian innovations and (2.2). Extensions to non-Gaussian innovations with suitably nice densities will be analogous. We exclude the effect of model selection and assume that the model structure is fixed, i.e. the structure of a binary tree partition $\mathcal{P}$ in (2.2) is specified. The unknown parameters of the model are then

$$\xi = (\phi, \theta, d)$$

for the autoregressive, GARCH (within a partition cell) and threshold parameters, respectively. The volatility can thus be written as

$$\begin{aligned} &\sigma_t^2(\theta, d) = f_\theta^{\mathcal{P}_d}(X_{t-1}, \sigma_{t-1}^2(\theta, d)), \\ &f_\theta^{\mathcal{P}_d} \text{ as in (2.2) but with } \mathcal{P} = \mathcal{P}_d \text{ parameterized with unknown thresholds } d. \end{aligned} \tag{3.1}$$

Since we view the tree GARCH model only as a suitable approximation for the data generating process, we argue here that our procedure yields consistent estimates for the best parameters ξ_0, projected on the model specified by (2.1) with Gaussian innovations and (3.1),

$$\xi_0 = \mathrm{argmax}_\xi h(\xi),\ h(\xi) = \mathbb{E}\left[\log\left\{\sigma_t(\theta, d)^{-1}\varphi\left(\frac{X_t - \phi X_{t-1}}{\sigma_t(\theta, d)}\right)\right\}\right]. \tag{3.2}$$

We assume that ξ_0 is unique, e.g. $h(\cdot)$ is strictly concave. If the model is true, ξ_0 will be the true parameter. Although ξ_0 is unique, the maximum likelihood estimator (based on Gaussian innovations) $\hat{\xi}$ is not: given data, the log-likelihood is constant between observations with respect to the threshold parameters d. For theoretical purposes, we define $\hat{d}$ as the smallest values representing a maximum likelihood estimator.

Theorem 1. *Assume that the data generating process $(X_t)_{t\in\mathbb{Z}}$ is stationary. Under the regularity conditions (A1)-(A5) given in the Appendix, the maximum likelihood estimator in the model specified by (2.1) with Gaussian innovations and (3.1) satisfies $\hat{\xi} \to \xi_0$ in probability, as $n \to \infty$.*

A proof is given in the Appendix. Theorem 1 describes consistency in mis-specified models. The true model isn't necessarily of GARCH type as in (1.1); or the correct model is of the form (1.1) but with non-normal innovations $(Z_t)_{t\in\mathbb{Z}}$. The interpretation of the consistency result is then as follows: the model with parameter ξ_0 is the closest element in the class of threshold

models, described by (2.1) and (3.1), to the true underlying stationary process with respect to the Kullback-Leibler divergence, see for example Shibata (1997). In particular, if the fitted model is true, the true parameters are consistently estimated.

Asymptotic normality with convergence rate $n^{-1/2}$ for the parameters ϕ, θ usually holds due to partial differentiability with respect to ϕ, θ of the log-likelihood function; this is outlined in Pollard (1984, Ch. VII). Under the realistic assumption that the data generating model is not exactly of threshold-type form, the estimated thresholds $\hat{d}$ will generally have a different limiting distribution and, provided that the true model has a smooth distribution, its convergence rate is slower than $n^{-1/2}$, due to a non-continuous likelihood function with respect to the split point parameters. A general relevant theory for this phenomenon is given by Kim and Pollard (1990); see also Hansen (2000).

4 Numerical results

We consider here the performance of tree-structured GARCH models for simulated and real data. We compare the results with the GARCH(1,1) estimate in (1.2) and a nonparametric generalized additive model for log-transformed squared data which is described in the Appendix. We always report with the use of $M = 5$ in Step 4 from Section 2.1 and with grid search using mesh $= 8$ as described in Section 2.1 (except in Table 3.2 where mesh $= 16$ is also used): these specifications lead to good tree-structured model fits, despite their simplicity.

4.1 Simulations

The models that we use for simulating data are as in (1.1) with various $f(\cdot, \cdot)$. One is a threshold model

$$f(x, \sigma^2) = \begin{cases} 0.1 + 0.5x^2 \ , & \text{if } x \le d_1 = 0, \\ 0.2 + 0.2x^2 + 0.75\sigma^2 \ , & \text{if } x > d_1 = 0 \text{ and } \sigma^2 \le d_2 = 0.5, \\ 0.8 + 0.5\sigma^2 \ , & \text{if } x > d_1 = 0 \text{ and } \sigma^2 > d_2 = 0.5, \end{cases} \tag{4.1}$$

The parameters are chosen to mimic time series of real log-returns. Also, the first threshold $d_1 = 0$ allows for an asymmetry natural in finance. Another model is a classical GARCH(1,1)

$$f(x, \sigma^2) = 0.05 + 0.1x^2 + 0.85\sigma^2. \tag{4.2}$$

A third model is neither GARCH nor threshold GARCH

$$f(x, \sigma^2) = (0.1 + 0.2\,|x| + 0.9x^2) \cdot (0.8 \exp(-1.5\,|x|\,|\sigma|)) + (0.4x^2 + 0.5\sigma^2)^{3/4} \tag{4.3}$$

Finally, we consider a multiplicative model for which the GAM fitting procedure, described in the Appendix, is consistent,

$$\begin{aligned} &X_t = \sigma_t Z_t \ (t \in \mathbb{Z}), \\ &\log(\sigma_t^2) = \log(0.4 + 0.28|X_{t-1}|^3) - 0.15X_{t-2}^2, \end{aligned} \tag{4.4}$$

where $(Z_t)_{t \in \mathbb{Z}}$ is as in (1.1). The distribution of innovations is either standard normal $Z_t \sim \mathcal{N}(0,1)$ or scaled t_6, $\sqrt{6/4}Z_t \sim t_6$, so that Z_t has variance one. We always take sample size $n = 1000$: for real daily data, this would correspond to about four years and stationarity would be expected to be approximately true.

Estimation always uses the knowledge that $\mu_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}] \equiv 0$, i.e. $\phi = 0$ in (2.1). Figures 2 and 3 display some results from the tree-structured GARCH model, in comparison with the

classical fit of a GARCH(1,1) in (1.2) and with an estimated generalized additive model (GAM) described in the Appendix.

FIGURE 2 ABOUT HERE.

We observe the following in Figure 2. The tree structured GARCH fit using $\mathcal{N}(0,1)$ innovations overestimates the number of thresholds: the first two thresholds are approximately correct. An improvement is given by the tree-structured GARCH fit with scaled t_ν-distributed innovations: the thresholds and also the maximum likelihood estimated degrees of freedom $\hat{\nu} = 5.12$ for the true $\nu = 6$ are very satisfactory. The classical GARCH(1,1) model with scaled t_ν-distributed innovations yields $\hat{\nu} = 4.37$ and of course, it doesn't exhibit any thresholds in the volatility surface. There is no surprise that the tree structured GARCH with scaled t_ν-distributed innovations is best, since the true model is of this form. But the tree GARCH estimate with $\mathcal{N}(0,1)$ misspecified innovations fits, as a quasi-maximum-likelihood, still reasonably well.

FIGURE 3 ABOUT HERE

Figure 3 highlights a desirable important feature: the tree-structured GARCH with scaled t_ν-distributed innovations performs very well in regions of high conditional variance. This is not the case with the classical GARCH(1,1) fit; and the GAM estimate described in the Appendix is very poor in regions of high volatility. Note the different scales for the various parts of Figure 3.

For quantifying the goodness of fit, we consider various measures:

$$\text{IS-L}_p = \sum_{t=1}^{n} \mid \sigma_t^2 - \widehat{\sigma}_t^2 \mid^p, \ p = 1, 2, \ \text{(in-sample loss)}$$

the AIC statistic from (2.7),

$$\text{OS-L}_p = \sum_{t=1}^{n} \mid \sigma_t^2 - \widehat{\sigma}_t^2(Y_1^{t-1}) \mid^p, \ p = 1, 2, \ Y_1^n \text{ a new test set (out-sample loss)},$$

where for OS-L, $\widehat{\sigma}_t^2(Y_1^{t-1})$ uses the model estimated from the data X_1^n but evaluates it on new test data Y_1^n that is another independent realization of the data. Both out-sample OS-L- and AIC-statistic are measures for predictive performance. The IS- and even more the OS-L-statistics are interesting measures for our simulations, but we can't calculate them for real data. The ratio $\hat{\sigma}_t^2/\sigma_t^2 = 1 + (\hat{\sigma}_t^2 - \sigma_t^2)/\sigma_t^2$ would measure a relative performance: in financial applications, the regions with large σ_t^2 are of particular interest and we therefore prefer the absolute scale which does not down-weight errors where volatility is high. Detailed results for eight independent realizations from model (1.1) with various $f(\cdot,\cdot)$ and from model (4.4) are reported in Tables 1–3.

TABLES 1–3 ABOUT HERE.

As an out-sample statistic, we view OS-L_2 as the most important measure for simulations. It gives more weight to large deviations than the OS-L_1 criterion: therefore, it is more appropriate when focusing on regions of high volatility.

The tree-structured GARCH procedure consistently outperforms the classical GARCH(1,1): for data 5 (GARCH(1,1) realization), it doesn't fit any thresholds and coincides with the GARCH(1,1) fit. Our new tree GARCH procedure compares favorably with the GAM estimate: it is much better for data 5 (GARCH(1,1) realization), data 1–4 (threshold GARCH realizations) and even for data 6 (GAM model), but a bit worse for data 7 and 8 (realizations with (4.3)). The GAM procedure has notorious difficulties in regions of high volatility which is

exploited most by its poor OS-L_2 performance. The reason for this may be that least squares (back-) fitting is applied to strongly non-Gaussian data. The classical GARCH(1,1) can exhibit huge variation in out-sample accuracy OS-L, for example a poor performance in data 2. This may be due to a very flat log-likelihood at the observed data: see Zumbach (1999) who also proposes a remedy. Interestingly, the tree-structured GARCH model exhibits much more stability here.

4.2 Two real data examples

We consider two financial instruments with 1000 daily negative log-returns $X_t = -100 \log(P_t/P_{t-1})$ (in percentages): from the German DAX index between January 18, 1994 and November 17, 1997; and from the BMW stock price between September 23, 1992 and July 23, 1996. We consider the tree-structured GARCH model, again in comparison with the GARCH(1,1) from (1.2) (both with the additional model term ϕX_{t-1} from (2.1) for $\mathbb{E}[X_t|\mathcal{F}_{t-1}]$), and with the GAM model described in the Appendix.

FIGURE 3 ABOUT HERE.

Figure 3 shows the result for the DAX index from a tree GARCH fit with $\mathcal{N}(0,1)$-distributed innovations. Three thresholds are fitted in the volatility surface. Graphical diagnostics for the residuals are satisfactory, with a tendency for heavier tails than standard normal.

For these real-data examples, we measure goodness of fit with the AIC statistic from (2.7) and with in- and out-sample losses for predicting centered second moments,

$$\text{IS-PL}_2 = \sum_{t=1}^{n} \left(\widehat{\sigma}_t^2 - (X_t - \widehat{\mu}_t)^2 \right)^2 \text{ (in-sample prediction loss)},$$

$$\text{OS-PL}_2 = \sum_{t=1}^{n} \left(\widehat{\sigma}_t^2(Y_1^{t-1}) - (Y_t - \widehat{\mu}_t(Y_1^{t-1}))^2 \right)^2, \; Y_1^n \text{ a new test set (out-sample prediction loss)}$$

with $\widehat{\mu}_t = \hat{\phi} X_{t-1}$ and $\widehat{\mu}_t(Y_1^{t-1})$, $\widehat{\sigma}_t^2(Y_1^{t-1})$ using the estimated model from the data X_1^n but evaluating at test data $Y_1^n = X_{n+1}^{2n}$ with $n = 1000$. A straightforward calculation shows that any stationary process $(Y_t)_{t\in\mathbb{Z}}$, having finite second moments and conditional first and second moments, can be decomposed as

$$\begin{aligned} Y_t &= \mu_t + \sigma_t U_t, \\ \mu_t &= \mathbb{E}[Y_t|Y_{t-1}, Y_{t-2}, \dots], \; \sigma_t^2 = \text{Var}(Y_t|Y_{t-1}, Y_{t-2}, \dots) \end{aligned}$$

with uncorrelated mean zero variables U_t with $\mathbb{E}[U_t^2|Y_{t-1}, Y_{t-2}, \dots] = 1$. Hence,

$$(Y_t - \mu_t)^2 = \sigma_t^2 + V_t, \; V_t = \sigma_t^2(U_t^2 - 1),$$

so that $\mathbb{E}[V_t|Y_{t-1}, Y_{t-2}, \dots] = 0$. When ignoring the estimation effect of $\hat{\mu}_t$, the quantity n^{-1}OS-PL_2 approximates

$$\mathbb{E}_Y[(\hat{\sigma}_t^2(Y_{-\infty}^{t-1}) - \sigma_t^2)^2] + \mathbb{E}[V_t^2].$$

The left term is our target, whereas the right term represents the noise variance in the problem. For real data, the variance $\mathbb{E}[V_t^2]$ seems often dominant and differences of various volatility estimates in terms of $\mathbb{E}_Y[(\hat{\sigma}_t^2(Y_{-\infty}^{t-1}) - \sigma_t^2)^2]$ will be masked. It indicates that the evaluation of volatility forecasts for real data is difficult. Thus, the criterion OS-PL_2 (and also others) for

real data allows only to discriminate between volatility forecasts whose performance is different in large orders of magnitude.

TABLE 4 ABOUT HERE.

We view OS-PL_2 as the most important measure, followed by AIC: as an in-sample loss, IS-PL_2 is not very relevant. One of the tree-structured models improves upon the classical GARCH(1,1) (both with $\mathcal{N}(0,1)$-distributed innovations) for the DAX index; for the BMW data, the two procedures have about the same performance. This is consistent with a common belief that classical GARCH(1,1) is better for individual prices than indices. The GAM estimate has the poorest performance in both data sets: for the BMW series, its behavior is extremely poor.

4.3 Summarizing numerical results

The tree-structured GARCH procedure often outperforms the estimate from the GARCH(1,1) model in (1.2) and the nonparametric GAM fit described in the Appendix: such better performance is with respect to many goodness of fit and graphical criteria. More specifically:

(i) For real data, the tree-structured GARCH procedure was slightly best for log-returns from an index (DAX) and equally good as the classical GARCH(1,1) fit for an individual share (BMW). The GAM estimate was poorest on real data. Despite the difficulties with discriminating performances for real data (see Section 4.2), we found strong evidence that the GAM method is not competitive.

(ii) For simulated data, the tree-structured GARCH is clearly better than classical GARCH(1,1). In one case (data 7), the GAM fit was found best; but overall, GAM was poorer and it may be very unstable, resulting in extremely low performance.

(iii) The tree-structured GARCH estimate may be much better in the interesting regions where the true volatility is high, see Figure 3. The nonparametric GAM fit can be extremely poor in regions of high volatility (this problem doesn't disappear when trying other smoothing parameters).

(iv) The tree-structured fitting procedure is sometimes slightly improved by assuming scaled t_ν-distributed innovations Z_t for the likelihood in (2.3), provided that the underlying innovations are heavier tailed. This is weakly evident in Figure 3 for real data. See also Table 1.

(v) The AIC-statistic is an indicator for ranking out-sample performance; and pruning with AIC in (2.7) works reasonably well.

Items (i), (ii) and (iii) indicate a strong advantage of parametric, likelihood based methods over nonparametric least squares smoothing techniques such as the GAM specification used here or multiplicative nonparametric ARCH models in Hafner (1998) or Yang et al. (1999). As pointed out in (iv), the likelihood approach can be easily modified to heavier tailed innovations inducing then even more heavy tails for the observations in the model. If performance is judged with a criterion putting emphasis on accurate prediction in high volatility regions, our tree based GARCH model seems clearly best among all alternative methods considered here.

4.4 How appropriate is GARCH(1,1) for daily returns?

The GARCH(1,1) model in (1.2) is very popular for analyzing daily log-returns of financial assets: it is often argued that it performs well despite that it has only three parameters describing a very low-dimensional model for sample size in the range of 1000. We quantify here the possible gains by using the flexible tree-structured GARCH model. In examples of daily log-return stock data, the new tree GARCH procedure was never found to be significantly worse than GARCH(1,1). A better performance for volatility forecasting may be masked by the high noise variance in evaluating predictions, see Section 4.2. In simulations, where evaluation of procedures is not obscured, we found substantial gains with tree-structured GARCH over GARCH(1,1). Particularly, the tree-structured GARCH seems to have an advantage in periods of high volatility. With real data, the first split has always been found in the x-axis (for the first lagged log-return) around zero: this is compatible with the interpretation that there is an asymmetric behavior depending on the sign of the previous log-return.

5 Concluding remarks

We propose a tree-structured GARCH model which is more flexible and accurate for prediction of volatility in financial time series than classical GARCH(1,1). The modeling strategy includes the classical GARCH(1,1) as a special case (no thresholds) and allows to increase complexity in a *systematic* way. On finite data, the new method compares favorably with a nonparametric technique based on additive models: especially in the interesting regions where the true volatility is moderate or large.

As supporting asymptotics, we present a consistency result about estimation of a best tree-structured GARCH model for approximating a general stationary process, see Section 3. More refined statistical inference is difficult due to the non-continuous nature of thresholds or trees: limiting distributions of estimated thresholds are typically non-Gaussian, see the discussion following Theorem 1. If the primary goal is construction of better volatility forecasts, which in turn can be used for dynamic risk management (cf. McNeil and Frey, 2000), we choose the route to optimize an information or complexity criterion rather than the somewhat inappropriate structural tool of testing for a model structure. As a simple solution, we use the AIC criterion: we have gained evidence on numerical examples that it can be used as a reasonable guideline.

Our univariate tree-structured GARCH procedure has a straightforward application in multivariate models where the conditional variance of an individual series is modeled as a function of the individual lagged values and individual lagged volatility. The multivariate cross-dependence is then modeled with cross-dependent innovations. For example, individual tree GARCH models for volatilities lead to an attractive version of the multivariate, constant conditional correlation model (Bollerslev, 1990). Another straightforward extension of our methodology is tree structured GARCH(p,q) modeling with $p > 1$ or $q > 1$. As already mentioned in Section 1, this may be of minor importance since the general model in (1.1) is in vogue and believed to capture the most important aspects of the underlying mechanism.

Acknowledgments: Comments by the Editor and two referees greatly helped to improve the presentation of our results.

References

[1] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. J. of Econometrics **31**, 307–327.

[2] Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model. The Review of Economics and Statistics **72**, 498–505.

[3] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont (CA).

[4] Bühlmann, P. and McNeil, A.J. (1999). Nonparametric GARCH models. Preprint, ETH Zürich.

[5] Clark, L.A. and Pregibon, D. (1993). Tree-based models. In *Statistical Models in S*, pp.377–419, auths. J.M. Chambers and T.J. Hastie. Chapman & Hall, London.

[6] Hafner, C.M. (1998). Estimating high-frequency foreign exchange rate volatility with nonparametric ARCH models. J. of Statistical Planning and Inference **68**, 247–269.

[7] Hansen, B.E. (2000). Splitting and threshold estimation. Econometrica **68**, 575–603.

[8] Hastie, T.J. and Tibshirani, R.J. (1990). Generalized Additive Models. Chapman & Hall, London.

[9] Kim, J. and Pollard, D. (1990). Cube root asymptotics. Annals of Statistics **18**, 191–219.

[10] McNeil, A.J. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. J. of Empirical Finance **7**, 271–300.

[11] Nocedal, J. and Wright, S.J. (1999). *Numerical Optimization.* Springer, New York.

[12] Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

[13] Rabemananjara, R. and Zakoian, J.M. (1993). Threshold ARCH models and asymmetries in volatility. J. of Applied Econometrics **8**, 31–49.

[14] Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. Statistica Sinica **7**, 375–394.

[15] Tong, H. (1990). *Non-linear Time Series. A Dynamical System Approach.* Oxford University Press.

[16] Yang, L., Härdle, W. and Nielsen, J.P. (1999). Nonparametric autoregression with multiplicative volatility and additive mean. J. of Time Series Analysis **20**, 579–604.

[17] Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. Ann. Probab. **22**, 94–116.

[18] Zumbach, G. (1999). The pitfalls in fitting GARCH(1,1) processes. Preprint. Olsen & Associates, Zürich.

Appendix

Assumptions and proof of Theorem 1

We first give and discuss a set of regularity conditions for Theorem 1.

(A1) The data generating process $(X_t)_{t\in\mathbb{Z}}$ is stationary, β-mixing (cf. Yu, 1994) with $\beta(k) \leq Ck^{-\delta}$ for some $C, \delta > 0$. Moreover, $\mathbb{E}|X_t|^2 < \infty$.

(A2) Consider the parameter set $\mathcal{L} = \{(\theta, d);\ \sup_t \sigma_t(\theta, d) = O_P(1)\}$, i.e. the set where the volatility doesn't 'explode'. Denote by $\Xi = \{\xi;\ (\theta, d) \in \mathcal{L}, |\phi| < 1\}$. Assume that the best projected parameter vector ξ_0 is an interior point of Ξ.

(A3) The function $f_\theta^{\mathcal{P}_d}(\cdot,\cdot)$ in (3.1) and $(X_t)_{t\in\mathbb{Z}}$ are such that $h(\cdot)$ in (3.2) has a unique maximizer and is continuous for $\xi \in \Xi$.

(A4) Define the truncated squared volatility as

$$\{\sigma_t^{trunc(p)}(\theta, d)\}^2 = f_{X_{t-1}} \circ f_{X_{t-2}} \circ \ldots \circ f_{X_{t-p}}(X_{t-p}^2),$$
$$f_x(\tau^2) = f_\theta^{\mathcal{P}_d}(x, \tau^2),$$

i.e. an approximation with p (instead of infinitely many) lagged observed values for the volatility. Assume that the function $f_\theta^{\mathcal{P}_d}(\cdot,\cdot)$ in (3.1) satisfies:

$$\sup_{t,(\theta,d)\in\mathcal{L}} |\sigma_t(\theta, d) - \sigma_t^{trunc(p)}(\theta, d)| = O_P(\gamma(p)),\ \text{for some } \gamma(p) \to 0\ (p \to \infty).$$

(A5) The function $f_\theta^{\mathcal{P}_d}(\cdot,\cdot)$ in (3.1) is such that

$$\inf_{x,\sigma^2,(\theta,d)\in\mathcal{L}} f_\theta^{\mathcal{P}_d}(x, \sigma^2) > 0.$$

Assumption (A1) indicates that we do not require the data being generated from a GARCH-type model as in (1.1) or (2.1) and (2.2). (A2) is related to stationarity of the model. Continuity of $h(\cdot)$ in (A3) is implied by smoothness of the distribution of $(X_t)_{t\in\mathbb{Z}}$. (A4) requires that the non-Markovian volatility process can be approximated by the truncated Markovian model; this holds for example, if all $|\beta_j| < 1$ in (2.2) and $(X_t)_{t\in\mathbb{Z}}$ having a smooth distribution. The assumption is related to the Shannon-McMillan-Breiman Theorem. (A5) implies strict positivity of the volatility function (being reasonable).

Proof of Theorem 1. For n large enough, the maximizer of the likelihood is in Ξ (otherwise, the values $\sigma_t(\hat{\theta}, \hat{d})$ would explode implying a small likelihood).

Denote the log-likelihood (conditioned on the first observation and using any $\sigma_1 \neq 0$) by

$$\ell_n(\xi) = n^{-1} \sum_{t=2}^{n} \log\left\{\sigma_t^{-1} \varphi\left(\frac{X_t - \phi X_{t-1}}{\sigma_t}\right)\right\},$$

abbreviating $\sigma_t(\theta, d)$ by σ_t. Analogously, define the truncated log-likelihood as

$$\ell_n^{trunc(p)}(\xi) = n^{-1} \sum_{t=p+1}^{n} \log\left\{(\sigma_t^{trunc(p)})^{-1} \varphi\left(\frac{X_t - \phi X_{t-1}}{\sigma_t^{trunc(p)}}\right)\right\}.$$

We will show that

$$\sup_{\xi\in\Xi}|\ell_n(\xi)-\ell_n^{trunc(p)}(\xi)|=O_P(\gamma(p)) \tag{5.1}$$

with $\gamma(p)$ as in assumption (A4). Similarly, denoting by

$$h^{trunc(p)}(\xi)=\mathbb{E}\left[\log\left\{\frac{1}{\sigma_t^{trunc(p)}(\theta,d)}\varphi\left(\frac{X_t-\phi X_{t-1}}{\sigma_t^{trunc(p)}(\theta,d)}\right)\right\}\right],$$

we get for the population version,

$$\sup_{\xi\in\Xi}|h(\xi)-h^{trunc(p)}(\xi)|=O(\gamma(p)). \tag{5.2}$$

We give the proofs for (5.1) and (5.2) at the end. Now, the idea is to work with the truncated log-likelihood which is of the form

$$\ell_n^{trunc(p)}(\xi)=n^{-1}\sum_{t=p+1}^{n}g(X_t,\ldots,X_{t-p+1};\xi),$$

$$g(X_t,\ldots,X_{t-p+1};\xi)=\log\left\{(\sigma_t^{trunc(p)})^{-1}\varphi\left(\frac{X_t-\phi X_{t-1}}{\sigma_t^{trunc(p)}}\right)\right\}.$$

By the mixing property of $(X_t)_{t\in\mathbb{Z}}$,

$$\sup_{\xi\in\Xi}|\ell_n^{trunc(p)}(\xi)-h^{trunc(p)}(\xi)|=o_P(1), \tag{5.3}$$

for every p, see Yu (1994). More details proving (5.3) are given at the end. By (5.1)–(5.3) it follows that

$$\sup_{\xi\in\Xi}|\ell_n(\xi)-h(\xi)|=o_P(1). \tag{5.4}$$

Now, use a sandwich argument:

$$\sup_{\xi\in\Xi}|\ell_n(\xi)-h(\xi)|+h(\xi_0)\geq|\ell_n(\hat{\xi})-h(\hat{\xi})|+h(\hat{\xi})\geq\ell_n(\hat{\xi})\geq\ell_n(\xi_0)=(\ell_n(\xi_0)-h(\xi_0))+h(\xi_0),$$

where we have used that ξ_0 and $\hat{\xi}$ are the maximizers of $h(\cdot)$ and $\ell_n(\cdot)$, respectively. Due to (5.4), the left and right hand side are asymptotically equal to $h(\xi_0)$, implying that $h(\hat{\xi})=h(\xi_0)+o_P(1)$. Since $h(\cdot)$ is continuous and ξ_0 its unique maximizer, it follows that $\hat{\xi}=\xi_0+o_P(1)$, i.e. consistency.

Proof of (5.1) and (5.2): using a first order Taylor expansion,

$$\log\{\sigma^{-1}\varphi(u/\sigma)\}=\log\{\tau^{-1}\varphi(u/\tau)\}+h(u,\tilde{\sigma})(\sigma-\tau),\ |\sigma-\tilde{\sigma}|<|\sigma-\tau|,$$
$$h(u,\tilde{\sigma})=-\tilde{\sigma}^{-1}+\tilde{\sigma}^{-4}u^2\varphi(u/\tilde{\sigma}).$$

Use this with $u=X_t-\phi X_{t-1}$, $\sigma=\sigma_t$, $\tau=\sigma_t^{trunc(p)}$. Assumption (A5), implying that σ_t, $\sigma_t^{trunc(p)}$ are bounded away from zero, yields that $\Delta_t(\xi)=h(X_t-\phi X_{t-1},\tilde{\sigma}_t(\theta,d))$ is uniformly bounded over t and ξ. Thus,

$$\ell_n(\xi)=\ell_n^{trunc(p)}(\xi)+n^{-1}\sum_{t=p+1}^{n}\Delta_t(\xi)(\sigma_t-\sigma_t^{trunc(p)}),$$
$$\sup_{t,\xi\in\Xi}|\Delta_t(\xi)|=O(1).$$

Now use (A4): since $\gamma(p)$ is uniform with respect to t and (θ, d), we obtain (5.1). Formula (5.2) follows similarly.

Proof of (5.3): apply Theorem 3.4 in Yu (1994). Consider the permissible class $\mathcal{G} = \{g(\cdot;\xi);\ \xi \in \Xi\}$. An envelope function $G(\cdot)$ can be constructed straightforwardly using (A5), and integrability of $G(\cdot)$ is implied by the moment assumption in (A1). Furthermore, the class $\mathcal{G}$ consists of functions which are compositions of a smooth, with a piecewise (involving indicators) quadratic function (from GARCH model in a partition cell): a good bound for the metric entropy condition then follows. □

Estimation with generalized additive model (GAM)

A nonparametric estimate for σ_t^2 can be derived as follows.
1. Estimate the conditional mean $\mu_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}]$ by an AR(1) model,

$$\hat{\mu}_t = \hat{\phi}\ X_{t-1}$$

with parametric estimate $\hat{\phi}$ obtained from least squares fitting.
2. Compute $Y_t = \log\{(X_t - \hat{\mu}_t)^2\},\ t = 2, \ldots, n$.
3. In model (2.1), we have $Y_t \approx \beta + \log(\sigma_t^2) + (\log(Z_t^2) - \beta)$ with $\beta = \mathbb{E}[\log(Z_t^2)]$. Denote by $\gamma_t = \beta + \log(\sigma_t^2)$. Fit a GAM model with the transformed data Y_2^n,

$$\hat{\gamma}_t = \hat{h}_1(X_{t-1}) + \hat{h}_2(X_{t-2}) + \ldots + \hat{h}_k(X_{t-k}),\ t = k+1, \ldots, n,$$

with nonparametric estimates $\hat{h}_i(\cdot)$ obtained from a least squares backfitting algorithm with response variables Y_t, cf. Hastie and Tibshirani (1990). The optimal value of k is chosen by minimizing the AIC statistic for Gaussian additive modeling of Y_t.
4. Back-transform $\delta_t = \exp(\hat{\gamma}_t) \approx \exp(\beta)\sigma_t^2 = \frac{1}{c}\sigma_t^2$ and build $R_t^2 = (X_t - \hat{\mu}_t)^2/\delta_t \approx c\ Z_t^2$. Thus, set

$$\hat{c} = (n)^{-1} \sum_{t=1}^{n} R_t^2.$$

5. Then, set

$$\hat{\sigma}_t^2 = \hat{c}\delta_t.$$

6. Iterate steps 1.-5. Thereby use weighted estimation in step 1 with weights $w_t = \frac{1}{\hat{\sigma}_t}$, where $\hat{\sigma}_t^2$ is the estimate from the previous iteration. Stop iterating by monitoring convergence of $\hat{\sigma}_t^2$ and $\hat{\mu}_t$.

A related technique is given in Yang et al. (1999): they don't use the log-transform but work with squared observations and dependent, but uncorrelated innovations.

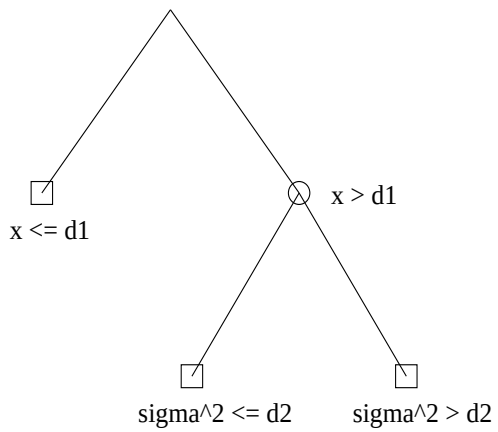

Figure 1: Binary tree partition $\mathcal{P} = \{\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3\}$ of predictor space $\{(x, \sigma^2); x \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$.

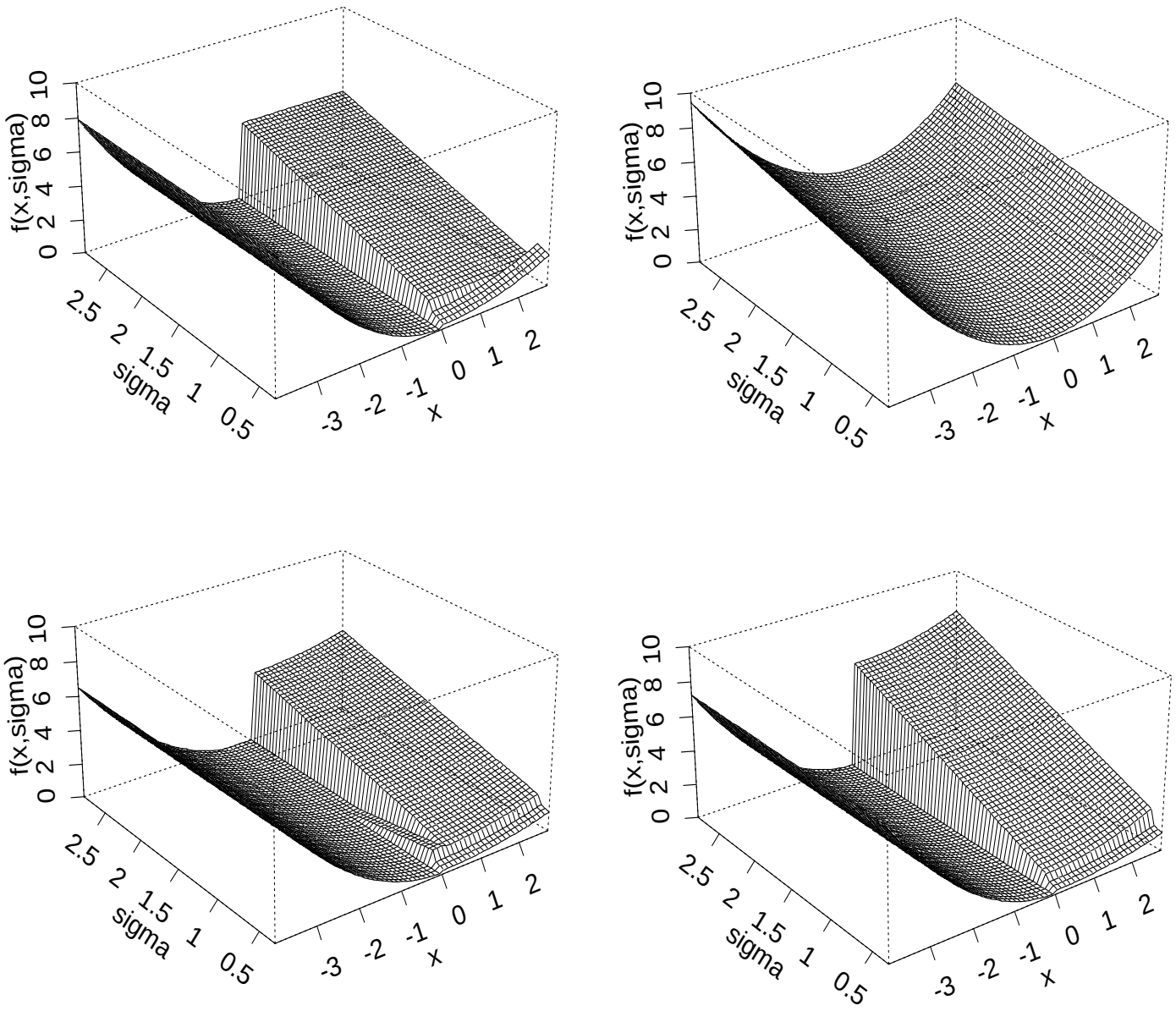

Figure 2: Top left: true conditional variance $f(x, \sigma)$ given by (4.1), plotted against x and σ. This is used in model (2.1) with scaled t_6-distributed innovations for a data realization (data 4 from Table 1) which is the basis for the other panels. Top right: estimated conditional variance from classical GARCH(1,1) model with scaled t_ν-distributed innovations (ν unknown). Bottom left: estimated conditional variance from tree GARCH model with standard normal innovations. Bottom right: estimated conditional variance from tree GARCH model with scaled t_ν-distributed innovations (ν unknown).

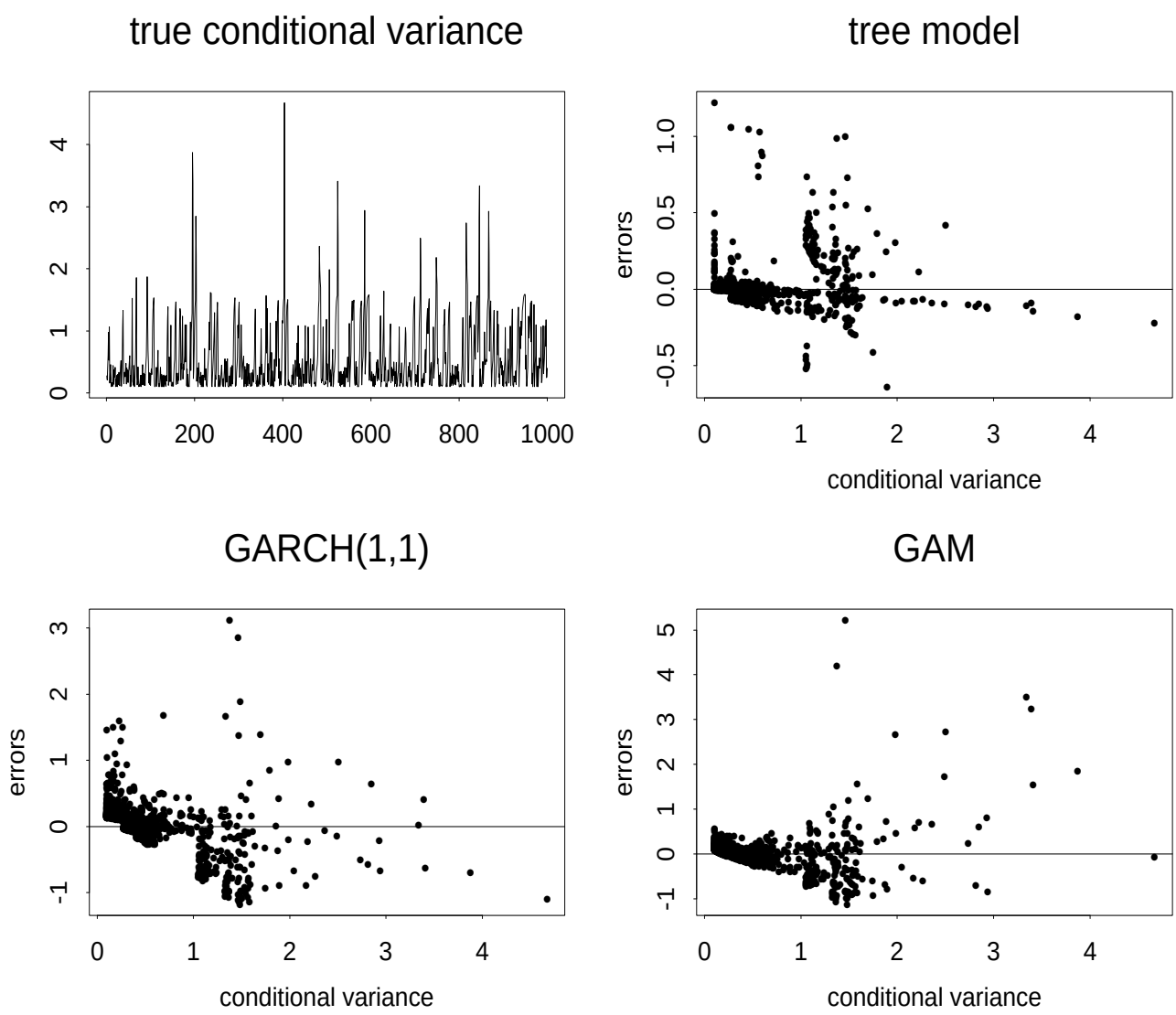

Figure 3: Top left: realization of volatility from model (2.1) with (4.1) and $\mathcal{N}(0,1)$-distributed innovations (data 2 from Table 1). Top right: errors $\hat{E}_t = \hat{\sigma}_t^2 - \sigma_t^2$ from tree GARCH fit with scaled t_ν-distributed innovations (ν unknown) versus true conditional variance σ_t^2. Bottom left: errors $\hat{E}_t$ from classical GARCH(1,1) fit with scaled t_ν-distributed innovations (ν unknown) versus σ_t^2. Bottom right: errors $\hat{E}_t$ from GAM estimate (as described in the Appendix) versus σ_t^2.

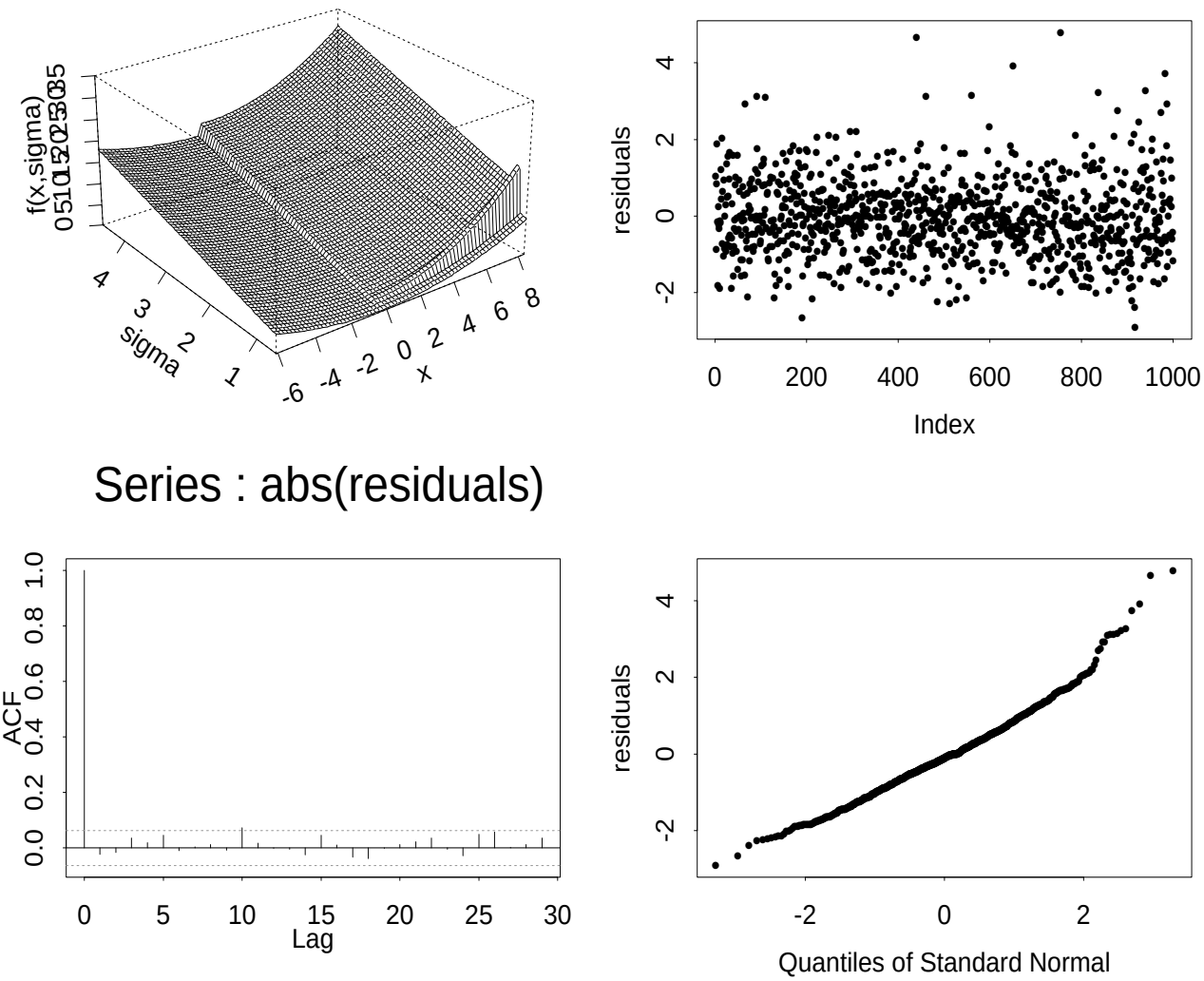

Figure 4: Results for negative log-returns of the DAX index from tree-structured GARCH model with $\mathcal{N}(0,1)$-distributed innovations. Top left: estimated function $\hat{f}(x,\sigma)$ for the conditional variance, plotted against x and σ. Top right: residuals $\hat{Z}_t = (X_t - \widehat{\mu}_t)/\widehat{\sigma}_t$ from the tree GARCH fit versus time t. Bottom left and right: autocorrelation function of the absolute residuals $|\hat{Z}_t|$ and normal-plot for the residuals $\hat{Z}_t$, respectively.

	tree-structured GARCH					
	Thresholds	AIC	L_1	L_2	OS-L_1	OS-L_2
data 1	$\widehat{d}_1 =$ -0.04 in x $\widehat{d}_2 =$ 0.77 in σ^2 $\widehat{d}_3 =$ 0.56 in σ^2	1842.2	102.9	43.5	110.7 121.0	44.4 61.4
data 2	$\widehat{d}_1 =$ -0.02 in x $\widehat{d}_2 =$ 0.47 in σ^2 $\widehat{d}_3 =$ 0.72 in x	1862.7	74.5	25.2	70.8 79.8	28.2 43.0
data 3	$\widehat{d}_1 =$ -0.01 in x $\widehat{d}_2 =$ 0.19 in x $\widehat{d}_3 =$ 0.98 in σ^2	1886.0	101.6	47.1	86.5 96.4	32.5 31.1
average	–	1863.6	93.0	38.6	94.2	40.1
data 4	$\widehat{d}_1 =$ -0.01 in x $\widehat{d}_2 =$ 0.39 in σ^2 $\widehat{d}_3 =$ 0.31 in x	1744.0	114.0	40.5	131.1 122.5	65.7 55.9
	$\widehat{d}_1 =$ -0.01 in x $\widehat{d}_2 =$ 0.41 in σ^2	1709.9	99.5	51.0	123.0 101.5	117.3 70.7
data 5	no thresholds	2935.8	85.1	20.5	67.4 82.6	9.2 16.5
data 6	$\widehat{d}_1 =$ -0.23 in x $\widehat{d}_2 =$ 0.24 in x	2114.0	74.9	150.2	66.0 61.0	36.8 20.8
data 7	$\widehat{d}_1 =$ -0.87 in x $\widehat{d}_2 =$ 0.91 in x $\widehat{d}_3 =$ 1.45 in σ^2 $\widehat{d}_4 =$ 1.71 in σ^2	3350.4	327.9	301.1	330.1 354.1	321.6 297.1
data 8	$\widehat{d}_1 =$ -0.49 in x $\widehat{d}_2 =$ 0.85 in x	3372.0	252.4	117.2	265.8 273.0	128.4 129.7

Table 1: Estimated thresholds $\hat{d}_i$ and goodness of fit measures for eight simulations. The data 1-5 and 7–8 are from model (1.1): data 1–3 with (4.1) and $\mathcal{N}(0,1)$ innovations, data 4 with (4.1) and scaled innovations $\sqrt{6/4}Z_t \sim t_6$, data 5 with (4.2) and $\mathcal{N}(0,1)$ innovations, data 7 and 8 with (4.3) and $\mathcal{N}(0,1)$ innovations. Data 6 is from model (4.4) with $\mathcal{N}(0,1)$ innovations. The likelihood for estimation is based on standard normal innovations (data 1–3, data 5–8); for data 4, we use standard normal innovation likelihood (upper part) and scaled t_ν innovation likelihood with ν unknown (lower part). Out-sample performances OS-L are evaluated at two independent test-sets.

	GARCH(1,1)				
	AIC	L_1	L_2	OS-L_1	OS-L_2
data 1	2009.4	225.7	116.2	248.5 265.0	150.1 175.5
data 2	2078.1	261.6	159.4	233.4 273.3	119.9 178.9
data 3	2073.8	284.4	227.2	243.6 265.5	140.5 181.0
average	2053.7	257.2	167.6	254.9	157.7
data 4	1897.2	257.9	154.1	284.7 280.5	281.7 276.4
	1818.4	265.7	165.3	287.8 290.6	348.4 350.3
data 5	2935.8	85.1	20.5	67.4 82.6	9.2 16.5
data 6	2129.0	81.7	179.9	69.4 66.2	47.1 24.9
data 7	3375.2	236.4	98.2	476.4 495.4	515.8 566.1
data 8	3393.1	265.6	136.6	270.6 267.3	137.0 138.8

Table 2: Goodness of fit measures for the same eight simulations from Table 1 using the GARCH(1,1) in (1.2) with $\mathcal{N}(0,1)$-distributed innovations, except data 4 (lower part) where scaled t_ν-distributed innovations, with ν unknown, are used.

	GAM			
	L_1	L_2	OS-L_1	OS-L_2
data 1	200.0	138.7	225.3 227.4	152.2 206.6
data 2	208.8	171.3	197.0 245.7	168.2 496.3
data 3	250.1	377.7	231.0 282.7	348.6 549.1
average	219.7	229.2	234.8	320.2
data 4	220.0	428.0	275.0 281.0	459.5 869.4
data 5	549.6	4353.0	416.4 451.2	987.4 1028.2
data 6	100.7	77.1	95.6 100.9	147.2 82.1
data 7	315.8	388.2	259.9 347.8	152.1 325.6
data 8	227.8	124.6	229.6 231.5	133.0 136.9

Table 3: Goodness of fit measures for the same eight simulations from Table 1 using the GAM model described in the Appendix.

	tree-structured GARCH			
	Thresholds	AIC	IS-PL_2	OS-PL_2
DAX	$\widehat{d}_1$ = -0.35 in x $\widehat{d}_2$ = 1.24 in σ^2 $\widehat{d}_3$ = 1.66 in σ^2	2776.2	8555.1	20001.3
	$\widehat{d}_1$ = -0.35 in x $\widehat{d}_2$ = -0.89 in x $\widehat{d}_3$ = 1.66 in σ^2 $\widehat{d}_4$ = -0.51 in x	2764.4	8507.6	20535.6
BMW	$\widehat{d}_1$ = -0.32 in x $\widehat{d}_2$ = 1.11 in σ^2	3155.0	12059.2	15111.6

	GARCH(1,1)			GAM	
	AIC	IS-PL_2	OS-PL_2	IS-PL_2	OS-PL_2
DAX	2785.3	9190.6	20387.9	12588.2	24787.8
BMW	3165.1	12063.9	15111.0	10964.8	50789.1

Table 4: Estimated thresholds $\hat{d}_i$ and goodness of fit measures for negative log-returns from the DAX index and the BMW stock price. The tree GARCH model (with $\mathcal{N}(0,1)$-distributed innovations) is fitted with mesh = 8 (DAX, upper part; BMW) and mesh = 16 (DAX, lower part). The GARCH(1,1) model is with $\mathcal{N}(0,1)$-distributed innovations.