Universität St.Gallen

Splines for Financial Volatility

Francesco Audrino and Peter Bühlmann

April 2007 Discussion Paper no. 2007-11

Department of Economics University of St. Gallen

Editor: Prof. Jörg Baumberger
University of St. Gallen
Department of Economics
Bodanstr. 1
CH-9000 St. Gallen
Phone +41 71 224 22 41
Fax +41 71 224 28 85
Email joerg.baumberger@unisg.ch

Publisher: Department of Economics
University of St. Gallen
Bodanstrasse 8
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 22 98

Electronic Publication: http://www.vwa.unisg.ch

Splines for Financial Volatility

Francesco Audrino and Peter Bühlmann

Author's address:

Prof. Dr. Francesco Audrino
Institute of Mathematics and Statistics
University of St. Gallen
Bodanstrasse 6
9000 St. Gallen
Tel. +41 71 224 2431
Fax +41 71 224 2894
Email francesco.audrino@unisg.ch
Website http://www.people.lu.unisi.ch/audrinof/

Author's address:

Prof. Dr. Peter Bühlmann
Seminar für Statistik
ETH Zentrum
LEO C17
8092 Zürich
Tel. +41 1 632 7338
Fax +41 1 632 1228
Email buhlmann@stat.math.ethz.ch
Website http://stat.ethz.ch/~buhlmann/

Abstract

We propose a flexible GARCH-type model for the prediction of volatility in financial time series. The approach relies on the idea of using multivariate B-splines of lagged observations and volatilities. Estimation of such a B-spline basis expansion is constructed within the likelihood framework for non-Gaussian observations. As the dimension of the B-spline basis is large, i.e. many parameters, we use regularized and sparse model fitting with a boosting algorithm. Our method is computationally attractive and feasible for large dimensions. We demonstrate its strong predictive potential for financial volatility on simulated and real data, also in comparison to other approaches, and we present some supporting asymptotic arguments.

Keywords

Boosting, B-splines; Conditional variance; Financial time series; GARCH model; Volatility.

JEL Classification

C13; C14; C22; C51; C53; C63

1 Introduction

In the last 30 years there has been a growing literature on financial volatility with a huge number of new models proposed to predict volatility. The reason why researchers have devoted such an attention to this particular topic can be explained by the central role that volatility plays in most financial applications in practice. Most of the models that have been proposed are simple with a small number of parameters only. In general, we are confronted with finding a good trade-off between parameter parsimony and model flexibility. The main research stream on financial volatility has focused more on the former, also by the desire for econometric interpretation. More flexible approaches can be found in the non-parametric setting: see, for example, Gourieroux and Monfort (1992), Härdle and Tsybakov (1997), Hafner (1998), Yang et al. (1999), Audrino (2005), and Andersen et al. (2005) for a survey of methods for nonparametric volatility modeling.

We propose a flexible model based on a high-dimensional parameterization from a B-spline basis expansion. So far, to our knowledge, the only other study that used splines to estimate financial volatility is from Engle and Rangel (2005) who introduced the Spline GARCH model. However, the use of splines in their work is completely different from ours: they find that an exponential spline is a convenient non-negative parameterization for the slow changes over time of the unconditional variance whereas we use B-spline basis functions for approximating the general conditional variance function. One of the novelties of our approach is to bring regularized and sparse model fitting into the field of volatility estimation: even when having over-parameterized the model a-priori, our estimation method will regularize by selecting the relevant basis functions only and shrinking all others exactly or close to zero. B-splines have been mathematically justified for function approximation, see for example de Boor (2001). In fact, B-splines represent piecewise polynomial functions and consequently, they can approximate any given continuous function of interest. Moreover, B-splines also give rise to an easy interpretation of the model. For example, if we construct the additive expansion for the conditional variance with B-splines of order one (i.e. constant functions equal to one in different regions of the predictor variables), the model can be interpreted as a threshold-regime model for the volatility, where regimes are associated with different regions of the predictor space and the conditional variance is locally constant. Another nice feature of our approach is that it is computationally feasible despite that the number of parameters to be estimated can be large. The computations rely on fitting a possibly over-complete dictionary of basis functions, in our case from B-splines, using a greedy boosting algorithm: the approach is related to the work by Bühlmann (2006) but with a loss function tailored for volatility estimation.

We validate the goodness of our model on simulated and real data. We collect strong empirical evidence for superiority of our model in comparison with two other approaches: the first one being the standard, widely used parametric GARCH(1,1) model and the second one being the univariate nonparametric functional gradient descent method in Audrino and Bühlmann (2003). The use of the former as a benchmark model is motivated by the remarkable consensus that it is appropriate

to describe the dynamics of financial volatility, despite its simplicity, and by the empirical evidence that it is very difficult to beat the GARCH(1,1) model with more sophisticated methods (Lunde and Hansen, 2005). The choice of the latter approach has been motivated by comparing with a very competitive nonparametric estimator. Our proposed B-splines method outperforms the competitors with respect to different performance statistics, both for simulated and real data.

2 The model

As a starting point, we consider a non-parametric GARCH(1,1) model for the dynamics of the time series of interest, for example from the log-returns $X_t = \log(P_t) - \log(P_{t-1}) \approx (P_t - P_{t-1})/P_{t-1}$ of a financial instrument with prices P_t:

$$\begin{aligned} X_t &= \mu_t + \sigma_t Z_t \quad (t \in \mathbb{Z}), \\ \sigma_t^2 &= f(X_{t-1}, \sigma_{t-1}^2), \quad f : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+, \end{aligned} \tag{2.1}$$

were $(Z_t)_{t\in\mathbb{Z}}$ is a sequence of independent identically distributed innovation variables with zero mean and variance equal to one, independent from $\{X_s;\ s < t\}$. Therefore, $\mu_t = \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$ and $\sigma_t^2 = \mathrm{Var}(X_t \mid \mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1}$ is the σ-algebra generated from the random variables $\{X_s;\ s \le t-1\}$. Generally, in financial applications, there is no need to allow for a large degree of flexibility in the dynamics of the conditional mean. We assume that

$$\mu_t = \alpha_0 + \alpha_1 X_{t-1} \tag{2.2}$$

follows a simple AR(1) equation. Much more attention must be devoted to the modeling of the time-varying dynamics of the so-called volatility $\sigma_t = \sqrt{\mathrm{Var}(X_t \mid \mathcal{F}_{t-1})}$. The estimation and prediction of volatility is a central task in the financial field because of its primary importance in many practical applications. Models for financial volatility must be as accurate as possible due to the major effects of volatility prediction on the computation of risk measures and portfolio choices. Therefore, we first consider the general (squared) volatility function in a nonparametric GARCH(1,1) model,

$$\sigma_t^2 = \mathrm{Var}(X_t \mid \mathcal{F}_{t-1}) = f(X_{t-1}, \sigma_{t-1}^2). \tag{2.3}$$

The unknown function $f(\cdot,\cdot) \in \mathbb{R}^+$ above may be non-linear or even not smooth. Nonparametric techniques can be used for the estimation of $f(\cdot,\cdot)$. Their advantages include generality which is often discounted by decreased or non-improved average prediction performance. Even worse, nonparametric methods exhibit poor performance at edges which represent the periods of high volatility that are of major interest in practical applications. Additional difficulties are due to the strong sensitivity of choosing smoothing parameters.

Our approach is in the spirit of a sieve approximation with a potentially high-dimensional parametric model (i.e. several dozens up to hundreds of parameters) for the non-parametric function $f(\cdot,\cdot)$. As we will describe in Section 3, our estimation technique is computationally efficient and addresses in an elegant way a

major obstacle of estimating many parameters in a non-linear model. We model the dynamics of the logarithm of the squared volatility σ_t^2 as an additive expansion of simple bivariate B-spline basis functions on a predictor space $\mathbb{R} \times \mathbb{R}^+$ arising from the lagged values $(X_{t-1}, \sigma_{t-1}^2)$. Using the log-transform allows to get rid of positivity restrictions and enables the use of a convex loss function $\lambda(\cdot, \cdot)$ in formula (3.2). In details, we model

$$\begin{aligned}\log(\sigma_t^2(\theta)) &= \log(f_\theta(X_{t-1}, \sigma_{t-1}^2(\theta)) = \\ &= g_{\theta_0}(X_{t-1}, \sigma_{t-1}^2(\theta)) + \sum_{j_1=1}^{k_1} \sum_{j_2=1}^{k_2} \beta_{j_1,j_2} B_{j_1,j_2}(X_{t-1}, \sigma_{t-1}^2(\theta)), \end{aligned} \tag{2.4}$$

where $g_{\theta_0}(\cdot, \cdot)$ is a simple, parametric starting function and θ denotes the parameter set composed by $\{\theta_0,\ \beta_{j_1,j_2}, j_1 = 1, \ldots, k_1,\ j_2 = 1, \ldots, k_2\}$. We propose to take $g_{\theta_0}(\cdot, \cdot)$ from a parametric GARCH(1,1) process, see Bollerslev (1986). We may view our specification in (2.4) as a sieve approximation which is parametrically guided by $g_{\theta_0}(\cdot, \cdot)$. If all $\beta_{j_1,j_2} \equiv 0$, which may arise in our sparse estimation procedure from Section 3, we obtain the classical parametric GARCH(1,1) model; in general, we try to improve using the second term $\sum_{j_1=1}^{k_1} \sum_{j_2=1}^{k_2} \beta_{j_1,j_2} B_{j_1,j_2}(X_{t-1}, \sigma_{t-1}^2(\theta))$ with the bivariate B-spline basis functions $B_{j_1,j_2}(X_{t-1}, \sigma_{t-1}^2(\theta))$.

Multivariate B-splines can be written as products of univariate B-splines and, therefore, can be computed in an easy way. In our particular case, we have

$$B_{j_1,j_2}(X_{t-1}, \sigma_{t-1}^2(\theta)) = B_{j_1}(X_{t-1}) B_{j_2}(\sigma_{t-1}^2(\theta)),\ j_1 = 1, \ldots, k_1, \text{ and } j_2 = 1, \ldots, k_2. \tag{2.5}$$

The definition of univariate B-splines and some of their nice mathematical properties are described in Appendix B. In fact, B-splines represent piecewise polynomial functions and consequently, they can be used to approximate a general continuous, nonparametric volatility function in (2.3). B-splines allow for a large flexibility in the shape of the volatility function, depending on how we choose the following two tuning parameters: the degree and the number of breaks (or the knots) of each univariate B-spline basis function. In our particular case, we have two predictors given by past lagged returns and past lagged squared volatilities. We allow that the squared volatility function can be quadratic in X_{t-1} and thus, we fix the degree of the $B_{j_1}(X_{t-1})$-splines to be equal to 3. Furthermore, we choose a piecewise linear relation in σ_{t-1}^2 and thus, we fix the degree of the $B_{j_2}(\sigma_{t-1}^2)$-splines to be equal to 2. The number of breaks is a measure for the approximation accuracy: with a larger number of breaks, we obtain a better approximation but a higher variability due to larger complexity. In our empirical analysis, we always choose as break points the empirical α-quantiles of the corresponding predictor variables with $\alpha = i/\text{mesh}$, $i = 1, \ldots, \text{mesh} - 1$, and mesh$\in \mathbb{N}$.[1] A concrete example of univariate B-splines for the predictor variable X_{t-1} $(t = 1, \ldots, n-1)$ is shown in Figure 1. The data

[1] In general, one can also use a third tuning parameter to control the smoothness of the approximation at each break, i.e. the knot's multiplicity. We impose our approximation to be continuous and smooth at each break. This means that we set the knot's multiplicity to be equal to 1 for all knots except for the first and last one; for more details see Appendix B.

are annualized daily log-returns of the S&P500 index for the time period between January 1990 and December 1998 (2212 observations). Results are reported for degree= 3 (i.e. quadratic splines) and mesh= 4, resulting in $k_1 = 6$ basis functions.

FIGURE 1 ABOUT HERE.

We see from Figure 1 that each B_j-spline, $j = 1, \ldots, 6$ is piecewise parabolic. The breaks (or knots) are clearly visible as places of discontinuity in the derivatives of the B-spline. In particular, in our example, the three breaks are $\{-7.664, 0.701, 8.848\}$. Over the whole range of the data, the sum of the B-splines at every possible value x for X_{t-1} is equal to one. Last, for every point x, there are always exactly three B_j-splines that are different from zero. A more detailed description about the properties of B-splines that are of interest in our context are presented in Appendix B.

3 The estimation algorithm and its properties

We estimate the model specified in (2.1)-(2.5) by pseudo-maximum-likelihood, using a Gaussian assumption for the innovations. Due to the potentially large number of parameters, we employ additional regularization in terms of a boosting algorithm. This will lead to improved prediction performance but also ensures computational feasibility in high dimensions. Assuming that the innovations Z_t in (2.1) are standard normally distributed, the negative log-likelihood in the model is given by

$$\begin{aligned} -\log L(\alpha, \theta; X_2^T) &= \sum_{t=1}^{T} \frac{1}{2}\Big(\log(2\pi) + \log(\sigma_t^2(\theta)) + \frac{(X_t - \mu_t(\alpha))^2}{\sigma_t^2(\theta)} \Big) \\ &= \sum_{t=1}^{T} \frac{1}{2}\Big(\log(2\pi) + g_\theta(X_{t-1}, \sigma_{t-1}^2(\theta)) + \frac{(X_t - \mu_t(\alpha))^2}{\exp(g_\theta(X_{t-1}, \sigma_{t-1}^2(\theta)))} \Big), \end{aligned} \tag{3.1}$$

where $g_\theta(X_{t-1}, \sigma_{t-1}^2(\theta)) = \log(\sigma_t^2(\theta))$. The log-likelihood is always considered conditional on X_1 and some reasonable starting value $\sigma_1^2(\theta)$, e.g. $\sigma_1^2(\theta) = \mathrm{Var}(X_1)$.

We estimate the (many) parameters in the model using essentially the functional gradient descent algorithm from Friedman (2001) which belongs to the class of boosting procedures. Three ingredients are required: a loss function and its partial derivative, a base procedure or weak learner and an initial starting estimate. We choose the loss function from the likelihood framework above, i.e.

$$\lambda(y, g) = \frac{1}{2}\big(\log(2\pi) + g + \frac{y^2}{e^g} \big), \tag{3.2}$$

where $y = (x - \mu)$, see also Audrino and Bühlmann (2003). Note that when summing the values of the loss function (3.2) over the data sample, i.e. the empirical risk, we get the negative log-likelihood in (3.1). To proceed with the minimization, we need the partial derivative of the loss function with respect to the log squared volatility g.

This is the direction of g that yields the best improvements in the pseudo-maximum-likelihood optimization:

$$\frac{\partial \lambda(y,g)}{\partial g} = \frac{1}{2}\big(1 - \frac{y^2}{e^g}\big). \tag{3.3}$$

As a weak learner or base procedure, we propose the use of a componentwise least squares method, which fits one B-spline basis function at a time. Finally, as an initial starting estimate $g_0(\theta)$, we propose the use of the estimates from the simple parametric GARCH(1,1) model.

In more details, our estimation algorithm is as follows.

Coordinatewise gradient descent algorithm

Step 1 (initialization). Choose the starting parameters $\hat{\alpha}$ and $\hat{\theta}_0$ from a simple parametric AR(1) or GARCH(1,1) model, respectively. Denote by

$$\hat{\mu}(t) = \hat{\alpha}_1 + \hat{\alpha}_2 X_{t-1}$$

and by

$$\hat{g}_0(t) = \hat{\theta}_{0,1} + \hat{\theta}_{0,2} X_{t-1}^2 + \hat{\theta}_{0,3} \exp(\hat{g}_0(t-1)).$$

Set $m = 1$.

Step 2 (projection of the gradient to the B-splines). Compute the negative gradient vector

$$U_t = -\frac{1}{2}\Big(1 - \frac{X_t - \hat{\mu}_t}{e^{\hat{g}_{m-1}(t)}}\Big),\ t = 2, \ldots, T.$$

Then, fit the negative gradient vector with individual bivariate B-spline basis functions. Here, we will exclusively consider the componentwise linear least-squares base procedure

$$\hat{S}_m = argmin_{\mathbf{1} \le \mathbf{d} \le \mathbf{k}} \sum_{t=2}^{T} \Big[U_t - \hat{\beta}_{\mathbf{d}} B_{\mathbf{d}}\big(X_{t-1}, e^{\hat{g}_{m-1}(t-1)}\big)\Big]^2,$$

where $\mathbf{d} = (d_1, d_2)$ is a bivariate index, $\hat{\beta}_{\mathbf{d}}$ is the least-squares estimated coefficient when regressing U_t versus the spline basis function $B_{\mathbf{d}}(X_{t-1}, e^{\hat{g}_{m-1}(t-1)})$ $(t = 2, \ldots, T)$ and $\mathbf{k} = (k_1, k_2)$ is the bivariate order of the B-splines.

Step 3 (line search). Perform a one-dimensional optimization for the step-length when up-dating $\hat{g}_{m-1}$:

$$\hat{\beta}_{\hat{S}_m} = argmin_w \sum_{t=2}^{T} \lambda\big(X_t - \hat{\mu}_t, \hat{g}_{m-1}(t) + w B_{\hat{S}_m}(X_{t-1}, e^{\hat{g}_{m-1}(t-1)}\big).$$

Up-date

$$\hat{g}_m(t) = \hat{g}_{m-1}(t) + \hat{\beta}_{\hat{S}_m} B_{\hat{S}_m}(X_{t-1}, \exp(\hat{g}_{m-1}(t-1))).$$

Step 4 (iteration and stopping). Increase m by one and iterate Steps 2 and 3 until stopping with $m = M$. This produces the estimate

$$\hat{g}_M(t) = \hat{g}_0(t) + \sum_{m=1}^{M} \hat{\beta}_{\hat{S}_m} B_{\hat{S}_m}(X_{t-1}, \exp(\hat{g}_{m-1}(t-1)))$$

for the log (squared) volatility function in (2.4).

Analogously to Audrino and Bühlmann (2003), the stopping value M is chosen to optimize a cross-validated empirical risk with the first 70% of the data as training and the remaining 30% as test data. Note that this stopping parameter M is of fundamental importance to avoid overfitting and to obtain reliable results in an out-of-sample analysis.

Furthermore, it is often desirable to introduce shrinkage to zero in Step 3, to reduce the variance of the estimated B-spline components. The up-date $\hat{\beta}_{\hat{S}_m} B_{\hat{S}_m}$ in Step 3 of the algorithm above is then replaced by

$$\kappa \hat{\beta}_{\hat{S}_m} B_{\hat{S}_m}, \text{ with } 0 < \kappa \leq 1.$$

In our empirical analysis, we find that values $\kappa \in \{0.1; 0.2\}$ are very reasonable.

A final remark on the algorithm is about the choice of the breaks (or the knots) in the two predictors of the bivariate B-splines. In our empirical analysis we choose break points corresponding to empirical quantiles of the predictor variables. Since volatility is not observable, we fix the structure (i.e. the break sequence) of the B-splines for $\sigma^2_{t-1}(\theta)$ as the quantiles of the estimates $\exp(\hat{g}_0(t))$ from the simple GARCH(1,1) starting model.

3.1 Connections to penalized maximum likelihood

The estimation algorithm from Section 3 above yields sparse solutions and a regularized maximum likelihood estimate, depending on the stopping iteration M. The sparsity is induced by the nature of the coordinatewise procedure: it fits only one parameter (i.e. $\hat{\beta}_{\hat{S}_m}$ in the mth iteration) at a time. Due to early stopping (i.e. a "small" M), the estimated parameter vector $\hat{\beta}$ will be sparse, in terms of the number of non-zero elements or also in terms of the ℓ^1-norm $\|\hat{\beta}\|_1 = \sum_j |\hat{\beta}_j|$.

In case of the squared error loss function with $\lambda(y, g) = (y-g)^2$, there is a striking similarity of a coordinatewise gradient descent and the ℓ^1-penalized squared error regression, i.e. the Lasso (Tibshirani, 1996), see Efron et al. (2004). An extension of this result for more general cases than squared error loss has been given by Zhao and Yu (2005). It is argued that under some conditions on the design matrix, the solutions from the coordinatewise gradient descent algorithm approximate, as $\kappa \to 0$, the solutions from the Lasso which is defined as

$$\hat{\theta}(\xi) = \text{ argmin}_\beta(-2L(\beta) + \xi\|\beta\|_1), \tag{3.4}$$

where $L(\beta)$ denotes the log-likelihood function, $\xi \geq 0$ a penalty parameter and $\|\beta\|_1 = \sum_j |\beta_j|$. The whole range of Lasso solutions when varying the penalty parameter ξ can be approximated by the coordinatewise gradient descent method when varying the stopping iteration M over a large range of values. This is in the spirit of a path-following algorithm (Rosset and Zhu, 2007).

It is worth emphasizing that our algorithm proceeds with a computationally efficient up-dating rule in Step 4 (using the notation θ for the entire parameter vector):

$$\sigma_t(\theta_{new}) = \sigma_t(\theta_{old}) \cdot h(X_{t-1}, \sigma_{t-1}(\theta_{old})), \tag{3.5}$$

where $h(X_{t-1}, \sigma_{t-1}(\theta_{old})) = B_{\hat{S}_m}(X_{t-1}, \exp(\hat{g}_{m-1}(t-1))$ using the notation from Step 4 in iteration m. That is, the up-date is very fast and does not require $O(t)$ operation counts for recursive computation of $\sigma_t(\theta_{new})$ in the parameterization (2.4).

3.2 Supporting asymptotics

We will argue that the estimation algorithm from Section 3 is approximating a general volatility process $\sigma_t^2 = Var(X_t \mid \mathcal{F}_{t-1}) = f(X_{t-1}, \sigma_{t-1}^2)$ as in formula (2.3).

We assume that the process $\log(\sigma_t^2)_{t\in\mathbb{Z}}$ can be approximated by a stationary process $S_{t;p} = g_p(X_{t-1}, X_{t-2}, \ldots, X_{t-p})$ $(t \in \mathbb{Z})$ such that

$$\mathbb{E}|S_{t;p} - \log(\sigma_t^2)|^2 \to 0 \ (p \to \infty). \tag{3.6}$$

That is, we make the mild regularity condition that the true nonparametric GARCH(1,1) volatility process can be approximated by a nonparametric ARCH(p) model.

Furthermore, we assume that $g_p(\cdot, \ldots, \cdot)$ is sufficiently smooth and can be approximated by a B-spline basis. We can parameterize $\mathbb{R}^p$ with tensor-product B-spline basis functions as in Section 2,

$$B_{j_1,j_2,\ldots,j_p}(x_1, \ldots, x_p) = B_{j_1}(x_1)B_{j_2}(x_2)\cdots B_{j_p}(x_p). \tag{3.7}$$

Then, our assumption becomes

$$\mathbb{E}|S_{t;p} - \sum_{j_1=1,\ldots,j_p=1}^{\mathbf{k}} \beta_{j_1,j_2,\ldots,j_p} B_{j_1}(X_{t-1})B_{j_2}(X_{t-2})\cdots B_{j_p}(X_{t-p})|^2 \to 0 \ (k \to \infty). \tag{3.8}$$

Due to the approximations in (3.6) and (3.8), we will base our reasoning on an ARCH(p) model which is parameterized by a B-spline basis:

$$X_t = \sigma_t Z_t, \quad \log(\sigma_t^2) = g_p(\beta; X_{t-1}, \ldots, X_{t-p}) \quad (t \in \mathbb{Z}),$$
$$g_p(\beta; x_1, \ldots, x_p) = \sum_{j_1=1,\ldots,j_p=1}^{\mathbf{k}} \beta_{j_1,j_2,\ldots,j_p} B_{j_1}(x_1)B_{j_2}(x_2)\cdots B_{j_p}(x_p), \tag{3.9}$$

where $(Z_t)_{t\in\mathbb{Z}}$ is as in the model (2.1).

The estimation algorithm from Section 3 can be adapted in a straightforward way to the model in (3.9). The coordinatewise gradient descent method is an approximation of the following prototype Gauss-Southwell algorithm which has been formulated by Bickel et al. (2006). Consider the empirical risk

$$w(g_p(\beta)) = n^{-1} \sum_{t=p+1}^{n} \lambda(X_t, g_p(\beta; X_{t-1}, \ldots, X_{t-p})), \tag{3.10}$$

where $\lambda(\cdot, \cdot)$ is as in (3.2). The prototype algorithm up-dates the parameter vector $\hat{\beta}_m$ as follows:

$$\hat{\beta}_{m,\hat{S}_m} = \hat{\beta}_{m-1,\hat{S}_m} + \kappa_m \ (\kappa_m \in \mathbb{R}),$$
$$\hat{\beta}_{m,\mathbf{d}} = \hat{\beta}_{m-1,\mathbf{d}} \text{ for } \mathbf{d} \neq \hat{S}_m,$$
$$\text{such that } w(g_p(\hat{\beta}_m)) \leq \min_{\kappa\in\mathbb{R},\mathbf{d}} w(g_p(\hat{\beta}_{m-1} + \kappa\delta_{\mathbf{d}})). \tag{3.11}$$

Here, $\delta_{\mathbf{d}}$ denotes a vector whose entries are 1 for index $\mathbf{d}$ and zero elsewhere. The prototype estimation procedure is a greedy algorithm striving for maximal reduction of the empirical risk when up-dating $\hat{\beta}_m$ linearly with a (selected) B-spline basis function.

We make the following assumptions for the model in (3.9).

(A1) The process $(X_t)_{t\in\mathbb{Z}}$ is strictly stationary and α-mixing with geometrically decaying mixing coefficients $\alpha(j) \leq C\rho^j$ for some $0 < C < \infty$ and some $0 < \rho < 1$.

(A2) The innovations satisfy $\mathbb{E}|Z_t|^2 < \infty$.

(A3) The knots of the B-spline basis functions are in a compact sub-space of $\mathbb{R}^p$ and the parameter-space $\mathcal{C}$ with $\beta \in \mathcal{C}$, is a compact sub-space of $\mathbb{R}^{kp}$.

Then, the following holds.

Theorem 1. *Consider the prototype estimation algorithm as described in formula (3.11). Assume that $(X_t)_{t\in\mathbb{Z}}$ is as in model (3.9) and conditions (A1)-(A2) hold. Then, for any $0 < p < \infty$, there exists a stopping iteration M such that*

$$\begin{aligned}\mathbb{E}_Y[\lambda(Y_t, g_p(\hat{\beta}_M; Y_{t-1}, \ldots, Y_{t-p}))] &= \omega_0 + o_P(1) \ (n \to \infty), \\ \omega_0 &= \inf_{\beta\in\mathcal{C}} \mathbb{E}[\lambda(Y_t, g_p(\beta; Y_{t-1}, \ldots, Y_{t-p}))], \end{aligned} \tag{3.12}$$

where $\mathcal{C}$ is as in (A3), $\hat{\beta}_M$ is based on the observed sample $X_1, \ldots, X_n$ and $(Y_t)_{t\in\mathbb{Z}}$ is an independent copy from $(X_t)_{t\in\mathbb{Z}}$.

A proof is given in Appendix A. Theorem 1 says that the out-of-sample loss of the estimated model converges to the minimal achievable loss; note that the risk is a convex function of the parameters β and the minimal risk is unique. The result can be extended to include growing dimensionality of the β-vector as sample size increases, corresponding to a growing number of knots for the B-spline basis and a growing dimension p as $n \to \infty$, corresponding to the approximations in (3.6) and (3.8).

4 Numerical results

We consider the spline-GARCH(1,1) model, introduced in (2.1)-(2.5), on simulated and real data. We compare performance measures with those obtained from a simple, parametric GARCH(1,1) fit (Bollerslev, 1986) and from an univariate functional gradient descent (FGD) estimation as proposed by Audrino and Bühlmann (2003). The first comparison is important, since the classical GARCH(1,1) model is recognized to be a benchmark model for financial volatility which is difficult to outperform significantly, see for example Lunde and Hansen (2005). Further more, the FGD method is an excellent competitor using a nonparametric estimation methods. We always report with the use of mesh $\in \{4, 8\}$ as described in Section 2 and a shrinkage factor $\kappa \in \{0.1, 0.2\}$ as introduced in Section 3: these specifications have lead to very reasonable spline-GARCH(1,1) forecasts.

4.1 A simulation exercise

We report here goodness-of-fit results for synthetic data. We generate 2000 observations generated from a model which is able to mimic well stylized facts of financial daily return data. We always use the first 1000 simulated data as in-sample period to estimate the model and the successive 1000 values as out-of-sample testing period. This is repeated for 100 independent model simulations.

The data generating process for the volatility dynamics is a two-regime process with the first lagged return as a threshold variable being and a threshold value fixed at 0. The local time-varying volatility dynamics in the two regimes evolve according to a FIGARCH(1,d,1) model (see Baillie et al., 1996) and the model from Audrino and Bühlmann (2001) which is not of GARCH-type form. In detail, we consider a squared volatility function $\sigma_t^2 = f(X_{t-1}, X_{t-2}, \sigma_{t-1}^2)$ (which we use instead of $f(X_{t-1}, \sigma_{t-1}^2)$ in model (2.1)) given by

$$f(x_1, x_2, \sigma^2) = \begin{cases} 0.12 + 0.3\sigma^2 + [1 - 0.3L - (1 - 10^{-6}L)(1 - L)^d]x_1^2 \ , & \text{if } x_1 \leq d_1 = 0, \\ (0.4 + 0.28|x_1|^3) \cdot \exp(-0.15x_2^2) \ , & \text{if } x_1 > d_1 = 0. \end{cases} \tag{4.1}$$

Here, in the first expression, L denotes the lag or backshift operator and the fractional differencing operator $(1 - L)^d$ has a binomial expansion which is most conveniently expressed in terms of the hypergeometric function F: $(1 - L)^d = F(-d, 1, 1; L)$; for more details, see Baillie et al. (1996). In our simulations, we fix $d = 0.4$. Therefore, the resulting process ia a nonparametric GARCH(2,1) and it allows for long memory in second moments and for asymmetric (leverage) effects in volatility in response to past positive and negative returns. These are all stylized facts exhibited by real financial return time series. Note that our spline-GARCH(1,1) is misspecified in terms of the order of the ARCH part.

The distribution of innovations is chosen as standard normal, i.e. $Z_t \sim \mathcal{N}(0, 1)$ and we set $\mu_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}] \equiv 0$ in (2.1). In Figure 2 we show the above mentioned features of the volatility dynamics for one simulated sample from model (4.1).

FIGURE 2 ABOUT HERE.

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

$$\text{IS-L}_p = \frac{1}{T}\sum_{t=1}^{T} \mid \sigma_t^2 - \widehat{\sigma}_t^2 \mid^p, \ p = 1, 2, \ \text{(in-sample loss)} \tag{4.2}$$

$$\text{the in-sample and out-of-sample log-likelihood given in (3.1)}, \tag{4.3}$$

$$\text{OS-L}_p = \frac{1}{T}\sum_{t=1}^{T} \mid \sigma_t^2 - \widehat{\sigma}_t^2(X_{T+1}^{2T}) \mid^p, \ p = 1, 2, \ \ \text{(out-sample loss)}, \tag{4.4}$$

where for the out-of-sample measures, $\widehat{\sigma}_t^2(X_{T+1}^{2T})$ uses the model estimated from the data X_1^T but evaluates it on the successive test data X_{T+1}^{2T}, $T = 1000$. Both, the out-of-sample OS-L$_p$ and the out-of-sample log-likelihood statistic are measures for predictive performance. The IS- and even more the OS-L$_p$-statistic are interesting

measures for our simulations, but note that we cannot calculate them for real data since the true volatility σ_t is unknown. In the real data analysis shown in the next Section 4.2, we will overcome this problem by substituting realized volatility for the true volatility, where the former is constructed exploiting the information from high frequency data.

Detailed results averaged over 100 independent realizations from model (2.1) with volatility function f given in (4.1) are reported in Table 1.

TABLE 1 ABOUT HERE.

The spline-GARCH(1,1) method consistently outperforms both competitor approaches. In particular, the out-of-sample gains over the standard GARCH(1,1) model are about 10% with respect to both OS-L_p statistics. The reason for this may be assigned to the lack of ability of the (symmetric) GARCH(1,1) model for estimating an asymmetric volatility process. However, more or less the same out-of-sample gains occur over the nonparametric (not-symmetric) FGD model. In addition, the spline-GARCH(1,1) model fitting needs about 30% less computing time than the FGD.

Detailed results for the OS-L_1 statistic across the 100 simulations are shown in Figure 3. Qualitatively identical figures could be plotted for the other performance measures, too.

FIGURE 3 ABOUT HERE.

In the left panel of Figure 3, the OS-L_1 results are plotted against the relative gains over the classical GARCH(1,1) model. The better forecasting accuracy of the spline-GARCH(1,1) model across the simulations is clearly evident: only in one case (out of 100), the spline-GARCH(1,1) method performs worse than the GARCH(1,1) model. Gains over the GARCH(1,1) model range up to 30%. In the right panel of Figure 3, the same plot is made for the relative gains over the FGD method. Also in this case, the better forecasting potential of the spline-GARCH(1,1) method is easily seen, although the number of times that the FGD method yields better OS-L_1 results raises to 8 (out of 100). Gains over the FGD model are again up to 30%, as before when comparing with a GARCH(1,1) model.

4.2 Two real data examples

We consider two financial instruments with 3376 daily log-returns (in percentages, annualized): from the US S&P500 index and from the 30-years US Treasury Bonds between January 1990 and October 2003. Note that we consider here annualized returns whereas the simulation model in Section 4.1 is on the scale of daily returns. We use the first 2212 observations (i.e. January 1990 to December 1998) as in-sample estimation period and the successive remaining 1164 observations as out-of-sample test data. For this data, some additional high-frequency tick-by-tick observations are available to construct realized volatilities which we use as a highly accurate measure for the unknown underlying true volatility. We then compute the same performance statistics (4.2)-(4.4) introduced in Section 4.1 by substituting underlying true volatilities with realized volatilities. For details about the construction of

realized volatilities from tick-by-tick data see for example Andersen et al. (2001), (2003) and (2005), or Curci and Corsi (2003).

Performance results where volatility estimates and forecasts are obtained from a standard GARCH(1,1) fit, the univariate FGD fit (Audrino and Bühlmann, 2003) and the spline-GARCH(1,1) model are summarized in Table 2.

TABLE 2 ABOUT HERE.

As for simulated data, the spline-GARCH(1,1) model consistently outperforms both competitors. In both real data analyses under investigation, the predictive gains over the classical GARCH(1,1) model and the univariate FGD procedure range from 1 to 6%, depending on the performance measure. Note that when fitting the models on 30-years US Treasury Bond returns, the FGD approach is not able to improve the out-of-sample results obtained from a GARCH(1,1) fit, in contrast to the spline-GARCH(1,1) model which again improves upon the classical GARCH(1,1) fit. When comparing the computational costs of the FGD approach with the spline-GARCH(1,1) method, we find similar results as reported for the simulation exercise: the spline-GARCH(1,1) model is about a factor 1.5 faster. It is also important to remark that among the large number of parameters used in the general description of the spline model, only few of them are estimated to be different from zero.

5 Conclusions

We propose the use of B-splines for approximating a general nonparametric GARCH(1,1)-type volatility process of a financial time series. Our model is flexible and involves a relatively large dimension of the unknown parameters, e.g. in the dozens or even in the hundreds. For accurate prediction and estimation, regularization is essential: we advocate the use of a coordinatewise functional gradient descent algorithm, in the spirit of boosting methods which are very popular in the area of machine learning.

We present some supporting asymptotics of our estimation algorithm and we demonstrate, using simulated and real data, the excellent prediction capacity of our method.

Our modeling and computational framework can be easily extended to the case of multivariate time series or a non-stationary model with time-varying parameters (and hence time-varying volatility function). Exemplifying the latter, which would be in the spirit of Engle and Rangel (2005), we could easily replace the parameter vector β in (2.4) (and also the parameter vector θ_0) by a slowly changing function which is again parameterized by a B-spline basis: that is,

$$\beta_{j_1,j_2}(t) = \sum_r \alpha_{r;j_1,j_2} B_r(t), \tag{5.5}$$

where $B_r(\cdot)$ is a B-spline basis function for the time point t. Plugging this into (2.4), we would get a trivariate B-spline basis (product of three B-spline basis functions) and a larger parameter vector whose estimation would be pursued with the same methodology as described in Section 3.

A Proof of Theorem 1

We first argue that the population version of the prototype estimation algorithm (i.e. with $T = \infty$) converges to the minimizer

$$\omega_0 = \inf_{\beta \in \mathcal{C}} \mathbb{E}[\lambda(Y_t, g_p(\beta; Y_{t-1}, \ldots, Y_{t-p}))], \tag{A.1}$$

where $\mathcal{C}$ is a compact set. This claim follows from verifying in a straightforward way the condition (GS1) from Bickel et al. (2006). Thereby, we use that the B-spline basis is bounded by placing the knots in a compact subset of $\mathbb{R}^p$.

Thus, for $\epsilon > 0$, there exists a stopping iteration $M = M(\epsilon)$ for the population algorithm such that

$$\mathbb{E}[\lambda(Y_t, g_p(\beta_M; Y_{t-1}, \ldots, Y_{t-p}))] \leq \omega_0 + \epsilon. \tag{A.2}$$

Here, the Mth iterate of the population algorithm is denoted by β_M.

Hence, we only need to control the errors due to finite sample size n for the first $M(\epsilon)$ iterations. Since there are only finitely many B-spline basis functions and due to the finite iteration number $M(\epsilon)$, a uniform law of large numbers

$$\sup_{\beta \in \mathcal{C}} |(T-p)^{-1} \sum_{t=p+1}^{T} \lambda(Y_t; g_p(\beta; Y_{t-1}, \ldots, Y_{t-p})) - \mathbb{E}[\lambda(Y_t; g_p(\beta; Y_{t-1}, \ldots, Y_{t-p}))]| = o_P(1) \tag{A.3}$$

is sufficient to complete the proof. To show that (A.3) holds, note that

$$\begin{aligned} & (T-p)^{-1} \sum_{t=p+1}^{T} \lambda(Y_t; g_p(\beta; Y_{t-1}, \ldots, Y_{t-p})) - \mathbb{E}[\lambda(Y_t; g_p(\beta; Y_{t-1}, \ldots, Y_{t-p}))] \\ = & \quad \frac{1}{2} \sum_{j_1=1,\ldots,j_p=1}^{\mathbf{k}} \beta_{j_1,\ldots j_p} \left(B_{j_1}(Y_{t-1}) \ldots B_{j_p}(Y_{t-p}) - \mathbb{E}[B_{j_1}(Y_{t-1}) \ldots B_{j_p}(Y_{t-p})] \right) \\ + & \quad \frac{1}{2(T-p)} \sum_{t=p+1}^{T} (Z_t^2 - \mathbb{E}[Z_t^2]). \end{aligned} \tag{A.4}$$

For the first part, we can invoke the uniform law of large numbers, as implied by Theorem 2.2 and Corollary 2.3 from Andrews and Pollard (1994), using our assumption (A1) and the fact that we have a Lipschitz-continuous (i.e. linear in the parameters β) family of functions. For the second part, which is independent of the β-parameter, a standard law of large numbers yields convergence to zero. Hence, formula (A.3) holds and the proof of Theorem 1 is complete. □

B A brief description of B-splines

We first give a formal definition.

Definition. *Let $\xi = \{\xi_i\}_1^{l+1}$ be a strictly increasing sequence of points, and let k be a positive integer. If $P_1, \ldots, P_l$ is any sequence of l polynomials, each of order k (that is of degree $< k$), we define the corresponding* piecewise polynomial function (pp function) f of order k *by*

$$f(x) = P_i(x) \quad \text{if } \xi_i \le x < \xi_{i+1};\ i = 1, \ldots, l. \tag{B.1}$$

The points ξ_i are called breaks *of f. Whenever convenient, we think of such a function f as defined on the whole real line $\mathbb{R}$ by extending the first and the last piece.*

We say that two pp functions agree if and only if they consist of the same polynomial pieces and the same brakes. We denote the collection of all pp functions of order k (or degree $< k$) with break sequence $\xi = \{\xi_i\}_1^{l+1}$ by $\Pi_{k,\xi}$. In addition, the function $f \in \Pi_{k,\xi}$ is assumed to have a certain number of continuous derivatives at the break points ξ. Formalization of such *homogeneity conditions* leads to a new subspace

$$\Pi_{k,\xi,\nu} \tag{B.2}$$

for some vector $\nu = \{\nu_i\}_2^l$ of nonnegative integers. Here, ν_i denotes the number of continuity conditions required at ξ_i. In particular, $\nu_i = 0$ means that no continuity condition is imposed at ξ_i ("pure jump"). In our study we require at least continuity of the function.

We can construct a basis for $\Pi_{k,\xi,\nu}$ using *B-splines*, defined by the following recursion. Consider a nondecreasing sequence of knots $\mathbf{t} = \{t_j\}$(which may be infinite). Let

$$B_{j,1,\mathbf{t}} = \begin{cases} 1, & \text{if } t_j \le x < t_{j+1}, \\ 0, & \text{otherwise} \end{cases}$$

be the characteristic function of the jth knot interval. Note that the B_j functions form a *partition of the unity* (see Property 3 below). In particular, if $t_{j+1} = t_j$ then $B_{j,1,\mathbf{t}} = 0$. Starting with these first-order B-splines, we can construct higher-order B-splines by using the recurrence relation: for $k > 1$,

$$B_{j,k,\mathbf{t}} = w_{j,k,\mathbf{t}} B_{j,k-1,\mathbf{t}} + (1 - w_{j+1,k,\mathbf{t}}) B_{j+1,k-1,\mathbf{t}}, \text{ with } w_{j,k,\mathbf{t}}(x) = \frac{x - t_j}{t_{j+k-1} - t_j}. \tag{B.3}$$

Thus, the second-order B-spline is given by

$$B_{j,2,\mathbf{t}} = w_{j,2,\mathbf{t}} B_{j,1,\mathbf{t}} + (1 - w_{j+1,2,\mathbf{t}}) B_{j+1,1,\mathbf{t}},$$

and hence consists, in general, of two nontrivial linear pieces which join continuously to form a piecewise linear function that vanishes outside the interval $[t_j, t_{j+1})$. For this reason, $B_{j,2,\mathbf{t}}$ is called a *linear* B-spline. In general,the following holds.

Property 1. *(Support and positivity). The B-spline $B_{j,k,\mathbf{t}}$ is a pp function of order k with breaks $t_j, \ldots, t_{j+k}$. Hence, it is made up of at most k nontrivial polynomial pieces, vanishes outside the interval $[t_j, t_{j+k})$, and is positive on the interior of that interval.*

Another interesting property of B-splines is given below. Note that we already described that property for first-order B-splines.

Property 2. *(Partition of unity). The sequence* $\{B_{j,k,\mathbf{t}}\}$ *with* $\mathbf{t} = (t_1, \ldots, t_{n+k})$ *provides a positive and local partition of unity, that is, each* $B_{j,k,\mathbf{t}}$ *is positive on* (t_j, t_{j+k})*, is zero outside* $[t_j, t_{j+k}]$*, and* $\sum_j B_{j,k,\mathbf{t}} = 1$ *on* $[t_k, t_{n+1}]$*.*

We assume (as usual) that the first and last knot in the sequence have multiplicity equal to k. The actual smoothness of $B_{j,k,\mathbf{t}}$ depends on the multiplicity of the break ξ_i appearing in the knot sequence $(t_j, \ldots, t_{j+k})$. A general result, proved first by Curry and Schoenberg (1966), is as follows.

Property 3. *For a given strictly increasing sequence* $\xi = \{\xi_i\}_1^{l+1}$*, and a given nonnegative integer sequence* $\nu = \{\nu_i\}_2^l$ *with* $\nu_i < k$ *for all* i*, set*

$$n = k + \sum_{i=2}^{l}(k - \nu_i) = kl - \sum_{i=2}^{l}\nu_i = dim(\Pi_{k,\xi,\nu})$$

and let $\mathbf{t} = \{t_i\}_1^{n+k}$ *be the nondecreasing sequence obtained from* ξ *by the following two requirements:*

(i) for $i = 2, \ldots, l$*, the number* ξ_i *occurs exactly* $k - \nu_i$ *times in* $\mathbf{t}$*;*

(ii) $t_1 \leq t_2 \leq \ldots \leq t_k \leq \xi_1$ *and* $\xi_{l+1} \leq t_{n+1} \leq \ldots \leq t_{n+k}$*.*

Then, the sequence $B_1, \ldots, B_n$ *of B-splines of order* k *for the knot sequence* $\mathbf{t}$ *is a basis for* $\Pi_{k,\xi,\nu}$*, considered as functions on* $[t_k, t_{n+1}]$*.*

Property 3 enables the construction of a B-spline basis for any particular pp space $\Pi_{k,\xi,\nu}$, by providing a recipe for an appropriate knot sequence $\mathbf{t}$. This choice of $\mathbf{t}$ translates the desired amount of smoothness at a break (as specified by ν) into a corresponding number of knots at that site, with fewer knots corresponding to more continuity conditions. In our empirical analysis, we always take each break only once in the knot sequence, therefore allowing for maximal smoothness conditions at each break.

Finally, Property 3 allows us to represent pp functions (and therefore to approximate any continuous function) in terms of B-splines, the so-called *B-form* for a pp function.

Definition. *The* B-form *for* $f \in \Pi_{k,\xi,\nu}$ *consists of*

(i) the integers k *and* n*, giving the order of* f *and the number of linear parameters, respectively;*

(ii) the vector $\mathbf{t} = \{t_i\}_1^{n+k}$ *containing the knots (constructed from* ξ *and* ν *as in Property 3) in increasing order;*

(iii) the vector $\beta = \{\beta_i\}_1^n$ *of the coefficients of* f *with respect to the B-spline basis* $\{B_i\}_1^n$ *(with the knot sequence* $\mathbf{t}$*) for* $\Pi_{k,\xi,\nu}$*.*

In terms of these quantities, the value of f at x in $[t_k, t_{n+1}]$ is given by

$$f(x) = \sum_{i=1}^{n} \beta_i B_i(x). \tag{B.4}$$

For more details, proofs and the multivariate generalization of these results we refer the reader to de Boor (2001), Chapters 7-9 and 17.

References

Andersen, T.G., Bollerslev, T. and Diebold, F.X. (2005). Parametric and nonparametric volatility measurement, in *Handbook of Financial Econometrics*, eds. Y. Aït-Sahalia and L.P. Hansen, Amsterdam: Elsevier Science B.V.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2001). The distribution of exchange rate volatility, *Journal of the American Statistical Association* **96**, 42-55.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2003). Modeling and forecasting realized volatility, *Econometrica* **71**, 579-625.

Andrews, D.W.K. and Pollard, D. (1994). An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review* **62**, 119-132.

Audrino, F. (2005). Local Likelihood for non parametric ARCH(1) models, *Journal of Time Series Analysis* **26**, 251-278.

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

Audrino, F. and Bühlmann, P. (2003). Volatility Estimation with Functional Gradient Descent for Very High-Dimensional Financial Time Series, *Journal of Computational Finance* **6**, 65-89.

Baillie, R.T., Bollerslev,T. and Mikkelsen, H.O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **74**, 3-30.

Bickel, P.J., Ritov, Y. and Zakai, A. (2006). Some theory for generalized boosting algorithms. *Journal of Machine Learning Research* **7**, 705-732.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307-327.

Bühlmann, P. (2006). Boosting for high-dimensional linear models, *Annals of Statistics* **34**, 559-583.

Curci, G. and Corsi, F. (2003). A discrete sine transform approach for realized volatility measurement. NCCR FINRISK Working Paper No. 44.

Curry, H.B. and Schoenberg, I.J. (1947). On spline distributions and their limits: The Polya distribution function, *Bulletin of the American Mathematical Society* **53**, 1114.

Curry, H.B. and Schoenberg, I.J. (1966). On Polya frequency functions IV: the fundamental spline functions and their limits, *Journal d'Analyse Mathematique* **17**, 71-107.

de Boor, C. (2001). *A practical guide to splines*, Revised Edition, Springer Series in Applied Mathematical Sciences 27, New York.

Efron, B. and Hastie, T. and Johnstone, I. and Tibshirani, R. (2004). Least Angle Regression (with discussion), *Annals of Statistics* **32**, 407-451.

Engle, R.F. and Rangel, J.G. (2005). The spline GARCH model for unconditional variance and its global macroeconomic causes, Working paper, Stern School of Business, New York University.

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics* **29**, 1189-1232.

Lunde, A. and Hansen, P.R. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)?, *Journal of Applied Econometrics* **20**, 873-889.

Gourieroux, C. and Monfort, A. (1992). Qualitative threshold ARCH models, *Journal of Econometrics* **52**, 159-199.

Härdle, W. and Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression, *Journal of Econometrics* **81**, 223-242.

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

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths, *Annals of Statistics* **35**, in print.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

Yang, L., Härdle, W. and Nielsen, J. (1999). Nonparametric autoregression with multiplicative volatility and additive mean, *Journal of Time Series Analysis* **20**, 579-604.

Zhao, P. and Yu, B. (2005). Boosted Lasso, Working paper, Department of Statistics, University of California, Berkeley.

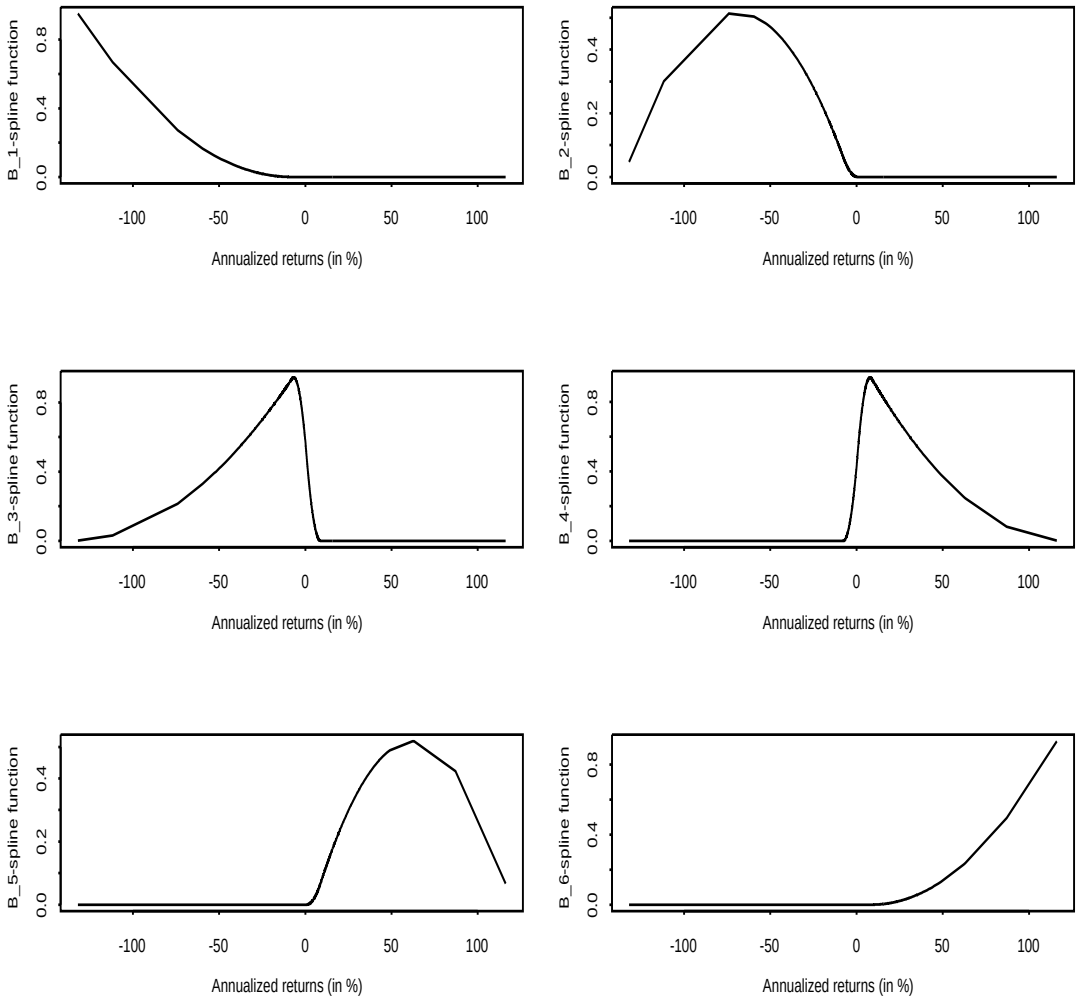

Figure 1: Quadratic B-splines for the predictor variable X_{t-1}, i.e. the lagged return. The data are annualized log-returns of the S&P500 index for the time period between January 1990 and December 1998 (2212 daily observations). Break points are empirical α-quantiles of the predictor variables with $\alpha = i/\text{mesh}$, $i = 1, \ldots, \text{mesh} - 1$, and mesh= 4. Explicitly, the breaks are $\{-7.664, 0.701, 8.848\}$.

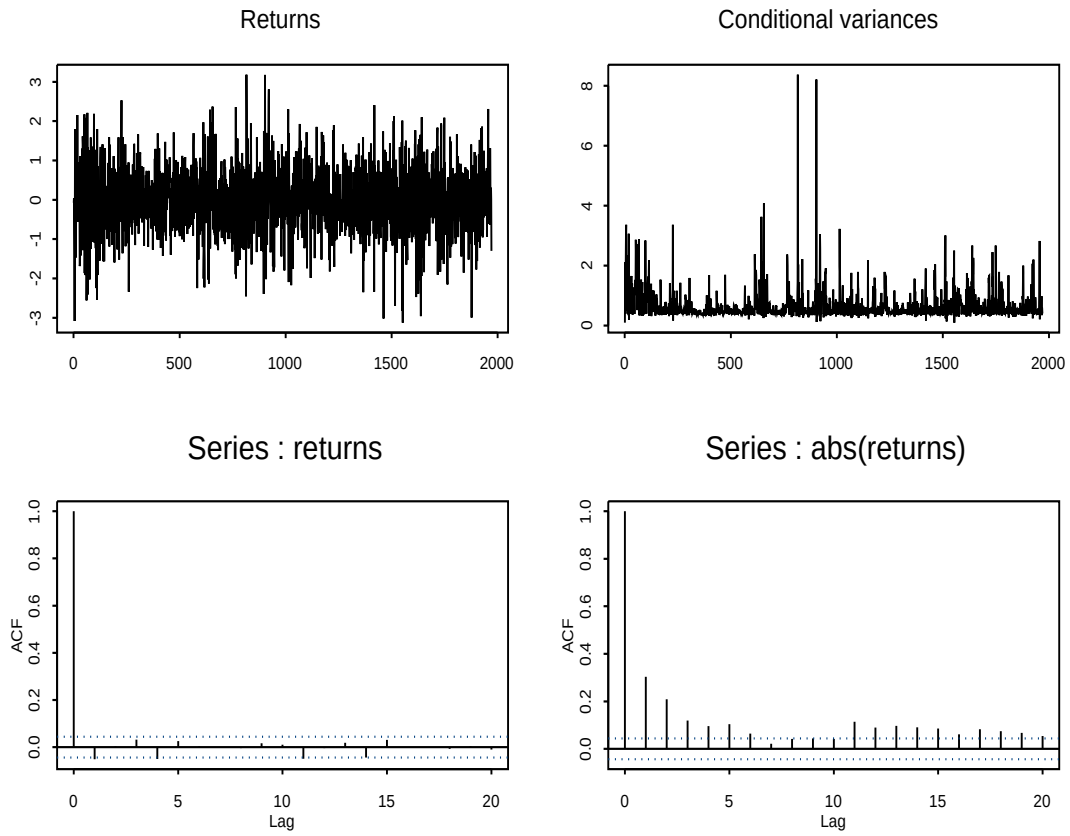

Figure 2: Returns, conditional variances, autocorrelation function of returns and absolute returns for one simulated sample of 2000 observations from the general non-parametric GARCH(2,1)-model with volatilities generated according to the threshold model (4.1). The local volatility dynamics follows a FIGARCH(1,d,1) model if the past lagged return is non-positive, and a model which is not of GARCH-type form if the past lagged return is positive.

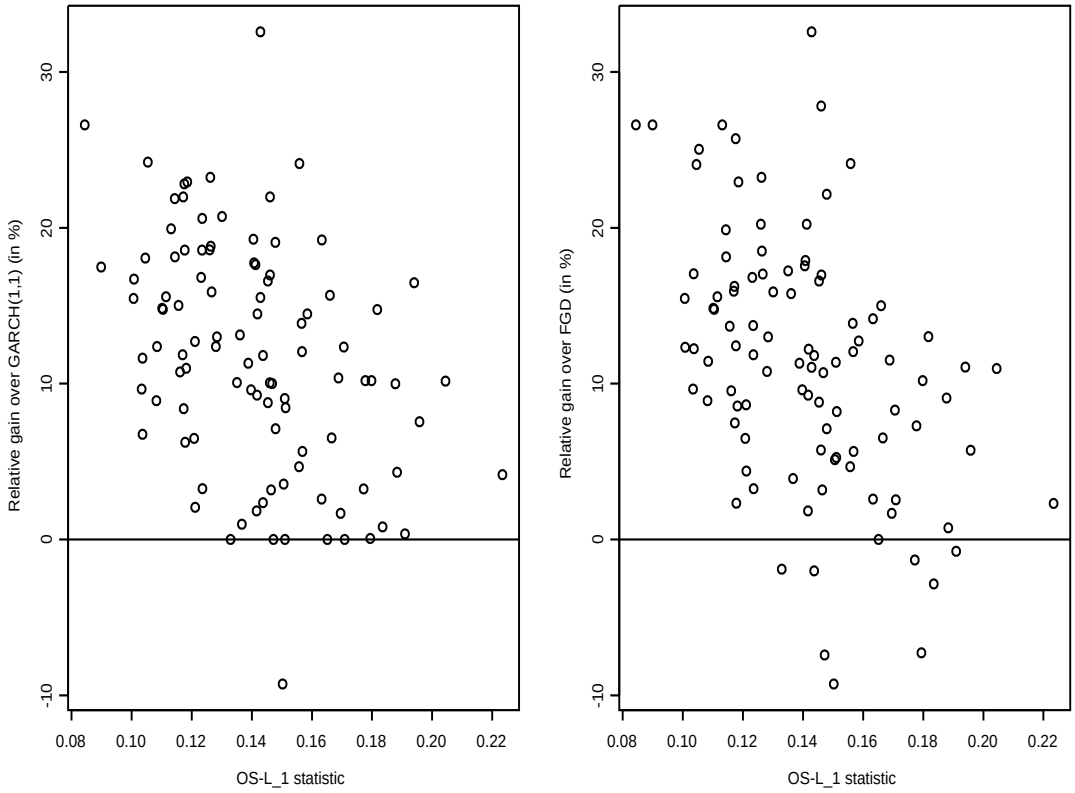

Figure 3: Plot of mean absolute errors (OS-L_1 statistic) for the (squared) volatilities estimated using the spline-GARCH(1,1) model against relative gains of mean absolute errors over the classical GARCH(1,1) model (left panel) and the FGD approach (right panel). Results are reported for 100 independent simulations from the general nonparametric GARCH(2,1) model with volatility function specified in (4.1).

Model	$\hat{M}_{\text{opt}}$	Averaged IS-			Averaged OS-		
		log-lik.	L_1	L_2	log-lik.	L_1	L_2
GARCH(1,1)		1132.782	0.16025	0.19954	1143.809	0.15955	0.13663
FGD	13.29	1126.902	0.15846	0.19806	1143.426	0.15876	0.13573
Spline-GARCH(1,1)	30.32	1120.341	0.13865	0.17196	1138.879	0.14074	0.12308

Table 1: Performance results averaged over 100 independent simulations from the general nonparametric GARCH(2,1) model with volatility dynamics specified in (4.1). In-sample (IS) and out-of-sample (OS) mean absolute errors (L_1), mean squared errors (L_2) and log-likelihood statistic. $\hat{M}_{\text{opt}}$ is the optimal stopping parameter averaged over the 100 simulations in the functional gradient descent (FGD) methodology and the spline-GARCH(1,1) model introduced in Section 3. The FGD algorithm is estimated using regression trees with three end-nodes as base learners, shrinkage factor $\kappa = 0.1$ and the correct number of predictor variables (two) given by the last two-lagged past returns. The tuning parameters in the spline-GARCH(1,1) estimation procedure are chosen as mesh= 8 for univariate splines constructed on past lagged returns, mesh= 4 for those constructed on past (squared) volatilities, and shrinkage $\kappa = 0.1$.

Panel A: S&P500 returns

Model	$\hat{M}_{\text{opt}}$	# par.	Averaged IS-			Averaged OS-		
			log-lik.	L_1	L_2	log-lik.	L_1	L_2
GARCH(1,1)		5	8661.73	90.4795	40569.5	5053.23	148.882	80281.4
FGD	23	118	8588.21	90.5723	39611.5	5047.80	144.161	76931.9
Splines	45	95	8606.69	85.7587	34316.7	5047.69	143.427	75644.3

Panel B: 30-years US Treasury Bond returns

Model	$\hat{M}_{\text{opt}}$	# par.	Averaged IS-			Averaged OS-		
			log-lik.	L_1	L_2	log-lik.	L_1	L_2
GARCH(1,1)		5	7760.16	36.6546	2985.50	4189.61	34.9955	3102.79
FGD	1	10	7754.80	36.9716	2915.56	4198.67	35.7989	3159.56
Splines	13	35	7743.19	34.7944	2890.44	4186.44	33.8643	3046.64

Table 2: Performance results for the S&P500 annualized returns (panel A) and the 30-years US Treasury Bond annualized returns (panel B) between January 1990 and October 2003 for a total of 3376 daily observations (in-sample until December 1998, 2212 observations). In-sample (IS) and out-of-sample (OS) mean absolute errors (L_1), mean squared errors (L_2) and log-likelihood statistic. $\hat{M}_{\text{opt}}$ denotes the optimal stopping parameter in the functional gradient descent (FGD) and spline-GARCH(1,1) estimation procedures, and # par reports the total number of parameters. The L-statistics are computed using realized volatilities as a proxy for the "true" unknown volatilities. The FGD algorithm is estimated using regression trees with three end-nodes as base learners, shrinkage factor $\kappa = 1$ and the last five-lagged past returns as predictor variables. The tuning parameters in the spline-GARCH(1,1) estimation procedure are mesh= 8 for both univariate splines constructed on past lagged returns and past (squared) volatilities for the S&P500 data, and we use mesh= 4 for the US Bond examples. The shrinkage factor is for both data-sets $\kappa = 0.2$.