Splines for Financial Volatility

Francesco Audrino[a*] and Peter Bühlmann[b]

[a]University of St. Gallen

[b]ETH Zürich

Revised version: October 2008

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

*Corresponding address: Fachbereich für Mathematik und Statistik, University of St. Gallen, Bodanstrasse 6, CH-9000 St. Gallen, e-mail: francesco.audrino@unisg.ch, Phone: 0041 71 2242431.

1 Introduction

In the last 25 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. Our approach is in the spirit of a sieve approximation of the conditional variance function, where the relevant B-spline basis functions are estimated using a regularization procedure.

Nowadays there is a large literature about adaptive, nonparametric estimation. Powerful results about model selection have been developed in a series of papers by Birgé and Massart (1997, 1998), Barron et al. (1999) and Laurent and Massart (2000), with applications to the adaptive estimation of density or regression functions. More recently, Comte and Rozenholc (2002) studied adaptive estimation of the mean and volatility functions in a penalized regression framework, with possible

applications to financial volatility. However, there are several differences from the model and methodology we propose. In particular, the volatility dynamics considered by Comte and Rozenholc (2002) followed an ARCH(1)-type process, whereas we generalize to the more realistic GARCH(1,1)-type of financial volatility dynamics. Moreover, both the empirical risk criterion and the penalty function used in the adaptive procedure applied in Comte and Rozenholc (2002) are very different from those we propose here which are computationally feasible for high-dimensional parameterizations with tensor-product B-spline basis functions. Our approach bears some similarities with Lin (2000) for function estimation in high-dimensional nonparametric regression. However, the setting with GARCH(1,1)-models exhibiting infinite memory and our estimation algorithm are entirely different.

One aim of this paper 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 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. piecewise constant functions 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 (Friedman, 2001): 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 in terms of volatility forecasting accuracy

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.

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 the dynamics of 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, and, for each t, the random variable Z_t is independent from $\{X_s;\ s < t\}$.[1] 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

[1] Note that if f satisfies a contraction property of the type

$$\sup_{x\in\mathbb{R}} |f(x,\sigma^2) - f(x,\tau^2)| \le D|\sigma^2 - \tau^2| \text{ for some } 0 < D < 1, \text{ and for all } \sigma^2,\ \tau^2 \in \mathbb{R}^+,$$

and assuming moment conditions $\mathbb{E}|\sigma_t|^4 \le C_1 < \infty$ and $\mathbb{E}|X_{t-1}|^4 \le C_2 < \infty$ for all $t \in \mathbb{Z}$, there exists an expansion $h(X_{t-1}, X_{t-2}, \ldots, X_{t-m})$ which converges in the L_2-sense to σ_t^2. An explicit construction is given in Bühlmann and McNeil (2002) (Theorem 1 using $\sigma_{t,0}^2 = X_{t-1}^2$).

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: finding a methodology that yields accurate volatility predictions is one of the main goals in both academic research and practice. We first consider the general conditional variance 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 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 the logarithm of 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 $B_{j_1,j_2}(\cdot,\cdot)$.

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

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

In fact, B-splines represent piecewise polynomial functions and consequently, they can be used to approximate a general continuous, nonparametric conditional variance function in (2.3). B-splines allow for a large flexibility in the shape of the conditional variance function, depending on how we choose the following two tuning parameters: the degree and the number of breaks (or 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$\in \mathbb{N}$.[2]

[2]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

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 conditional 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. $\sigma_1^2(\theta) = \mathrm{Var}(X_1)$. Note that the influence of the starting value decays exponentially fast under e.g. the contraction assumption stated in footnote 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 $y = (x - \mu)$ is mean-centered, 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

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 we refer to de Boor (2001).

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 log-transformed 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

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

Set $m = 1$.

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

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

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.[3]

[3] k_1 (k_2) is the number of univariate B-spline basis functions for X_{t-1} (σ^2_{t-1}), and in our case can be computed as $k_1 = (\text{mesh}_{X_{t-1}} - 1) + 3$ ($k_2 = (\text{mesh}_{\sigma^2_{t-1}} - 1) + 2$).

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

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

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 using sample-splitting, that is the optimal model structure is estimated on the first 70% of the data (estimation sample), and then fitted to the remaining 30% of the data (validation sample). The optimal stopping value M is the one that minimizes the empirical risk of the validation sample. Note that the 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. Regarding the choice of the breaks (or the knots) in the two predictors of the bivariate B-splines, 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. The optimal values of the tuning

parameters differ from application to application and can be found using cross validation or similar techniques.

Finally, it is worth emphasizing that our algorithm proceeds with a computationally efficient up-dating rule in Step 3 (using the notation θ for the entire parameter vector):

$$\sigma_t^2(\theta_{new}) = \sigma_t^2(\theta_{old}) \cdot h(X_{t-1}, \sigma_{t-1}^2(\theta_{old})), \tag{3.4}$$

where $h(X_{t-1}, \sigma_{t-1}^2(\theta_{old})) = B_{\hat{S}_m}(X_{t-1}, \exp(\hat{g}_{m-1}(t-1))$ using the notation from Step 3 in iteration m. The up-date is very fast and does not require $O(t)$ operation counts for recursive computation of $\sigma_t^2(\theta_{new})$ in the parameterization (2.4).

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

$$\hat{\theta}(\xi) = \ \text{argmin}_\beta(-2\log(L(\beta)) + \xi\|\beta\|_1), \tag{3.5}$$

where $L(\beta)$ denotes the likelihood function, $\xi \geq 0$ a penalty parameter and $\|\beta\|_1 = \sum_j |\beta_j|$. Or in more practical terms, the whole range of Lasso solutions in (3.5) when varying the penalty parameter ξ is similar to the solutions from the coordinatewise gradient descent method when varying the stopping iteration M over a large range

of values. This is in the spirit of an approximate path-following algorithm (Rosset and Zhu, 2007).

3.2 Supporting asymptotics

We will provide some asymptotic theory for fitting a nonparametric ARCH(p) model. A rigorous asymptotic analysis of our estimation method for fitting a nonparametric GARCH(1,1) model seems very difficult. However, we assume that the data generating process $(X_t)_{t\in\mathbb{Z}}$ is rather general and in particular, it does not need to be a nonparametric ARCH(p) or GARCH(1,1) process.

For the model to be fitted (which is not assumed to be the data-generating model), we consider an ARCH(p) model which is parameterized by a B-spline basis:

$$U_t = \sigma_t Z_t, \quad \log(\sigma_t^2) = g_p(\beta; U_{t-1}, \ldots, U_{t-p}) \quad (t \in \mathbb{Z}),$$
$$g_p(\beta; u_1, \ldots, u_p) = \sum_{j_1=1,\ldots,j_p=1}^{k} \beta_{j_1,j_2,\ldots,j_p} B_{j_1}(u_1) B_{j_2}(u_2) \cdots B_{j_p}(u_p), \tag{3.6}$$

where $(Z_t)_{t\in\mathbb{Z}}$ is as in the model (2.1). Note that as k and p increase, we can approximate (some sub-class of) nonparametric GARCH(1,1) processes. We use the notation $(U_t)_{t\in\mathbb{Z}}$ to distinguish the model process from the data-generating process $(X_t)_{t\in\mathbb{Z}}$.

The estimation algorithm from Section 3 can be adapted in a straightforward way to the model in (3.6). 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

$$w(g_p(\beta)) = n^{-1} \sum_{t=p+1}^{n} \lambda(X_t, g_p(\beta; X_{t-1}, \ldots, X_{t-p})), \tag{3.7}$$

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

$$\hat{\beta}_{m,\hat{S}_m} = \hat{\beta}_{m-1,\hat{S}_m} + \kappa_m \ (\kappa_m \in \mathbb{R}),$$
$$\hat{\beta}_{m,\mathbf{d}} = \hat{\beta}_{m-1,\mathbf{d}} \text{ for } \mathbf{d} \neq \hat{S}_m,$$
$$\text{such that } w(g_p(\hat{\beta}_m)) \leq \min_{\kappa\in\mathbb{R},\mathbf{d}} w(g_p(\hat{\beta}_{m-1} + \kappa\delta_{\mathbf{d}})). \tag{3.8}$$

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 data-generating process $(X_t)_{t\in\mathbb{Z}}$ (which may be different from the model in (3.6)).

(A1) The data-generating 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 data-generating process $(X_t)_{t\in\mathbb{Z}}$ satisfies $\mathbb{E}|X_t|^{2+\delta} < \infty$ for some $\delta > 0$.

(A3) For the model in (3.6) to be fitted, 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 $\mathbb{R}^{kp}$.

Assumption (A1) has been shown to hold for certain classes of ARCH and GARCH processes, see for example Carrasco and Chen (2002), Ango Nze and Doukhan (2004) and Francq and Zakoian (2006). However, we emphasize again that we do not assume that the data-generating process $(X_t)_{t\in\mathbb{Z}}$ is from a GARCH-type model.

The following consistency result holds.

Theorem 1. *Consider the prototype estimation algorithm as described in formula (3.8). Assume that the data-generating process $(X_t)_{t\in\mathbb{Z}}$ in (3.6) satisfies (A1)-(A2) and assume (A3) for the model to be fitted. Then, for any $0 < p < \infty$, there exists a stopping iteration $M = M_p$ such that*

$$\mathbb{E}_V[\lambda(V_t, g_p(\hat{\beta}_{M,T}; V_{t-1}, \ldots, V_{t-p}))] = \omega_0 + o_P(1) \ (T \to \infty),$$
$$\omega_0 = \inf_{\beta\in\mathcal{C}} \mathbb{E}[\lambda(V_t, g_p(\beta; V_{t-1}, \ldots, V_{t-p}))], \tag{3.9}$$

where $\mathcal{C}$ is as in (A3), $\hat{\beta}_{M,T}$ is based on the observed sample $X_1, \ldots, X_T$ and $(V_t)_{t\in\mathbb{Z}}$ 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 risk, achievable when fitting a nonparametric ARCH(p) model to a general data-generating process. Note that the risk is a strictly convex function of the parameters β and hence, the minimizer

$$\beta^* = \text{argmin}_{\beta \in \mathcal{C}} \mathbb{E}[\lambda(V_t, g_p(\beta; V_{t-1}, \ldots, V_{t-p}))]$$

is unique. This fact and Theorem 1 then imply consistency for the parameter vector (because of uniform integrability of the loss function for any fixed parameter vector β),

$$\hat{\beta}_{M,T} - \beta^* = o_P(1) \ (T \to \infty).$$

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). Furthermore, 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 Simulated data

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 squared volatility dynamics is a two-regime process with the first lagged return as a threshold variable with a threshold value fixed at 0. The local time-varying conditional variance 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 is 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 stylized facts exhibited by real financial return time series. Note that our spline-GARCH(1,1) is misspecified in terms of the order for 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).

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

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

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

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

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 so 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 conditional variance 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 1. Qualitatively similar figures could be plotted for the other performance measures, too.

FIGURE 1 ABOUT HERE.

In the left panel of Figure 1, 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 1, 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. In particular, we employ the multi-scale DST realized volatility estimator proposed by Curci and Corsi (2003) which consists in a multi-frequency regression based approach robustified by a Discrete Sine Transform filter that optimally de-correlates the price signal from microstructure noise.[4] We then compute the same performance statistics (4.2)-(4.4) introduced in Section 4.1 by substituting underlying true volatilities with realized volatilities. Using realized volatilities we are less exposed to the danger of getting wrong rankings due to noisy proxies for volatilities and biased performance measures; see, among others, Hansen and Lunde (2006) and Patton (2006).

To begin the analysis, Figure 2 shows the optimal conditional variance estimates (in-sample) obtained using our spline-GARCH(1,1) approach.

FIGURE 2 ABOUT HERE.

Both estimated conditional variance functions (for the S&P500 and Treasury-bond returns) are highly non-linear, asymmetric in past lagged returns of the series. As expected, a sort of leverage effect is visible in both series, in particular for high values

[4] A similar estimator has been proposed by Zhang et al. (2005).

of past lagged returns: negative past shocks increase conditional variance more than positive past shocks of the same size. A simple, additive structure of the logarithm of the conditional variance function in terms of past volatilities and lagged returns is clearly not supported from our spline-GARCH(1,1) model: in fact, almost all terms in the additive expansions are products of functions of the two predictor variables.

Performance results where squared 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 in particular 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.

The reported gains could be considered as marginal and too small only. However, such small differences are to be expected for real data because of the high noise component introduced by the estimation of the "true" unknown conditional variances using realized conditional variances needed to measure the performance of the competing approaches. To verify whether the gains are statistically relevant, we perform a series of classical superior predictive ability (SPA) tests, firstly introduced by Diebold and Mariano (1995).[5] The results are summarized in Table 3. Positive values of the statistic are always in favor of our spline-GARCH(1,1) model.

TABLE 3 ABOUT HERE.

Table 3 confirms the higher predictive power of our approach in terms of conditional

[5]Note that in this analysis we are not interested in building model confidence sets (see Hansen et al., 2003, for all details), but only in pairwise comparisons.

variance prediction over the competitors. Only in the case of the S&P500 returns, the FGD and the spline-GARCH(1,1) approach yield similar results.

5 Conclusions

We propose the use of B-splines for approximating a general nonparametric GARCH(1,1)-type squared 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 extended to the case of multivariate time series, although most financial institutions still use univariate models for their applications; see, for example, the study by Berkowitz and O'Brien (2001). Nevertheless, we can easily incorporate our spline-GARCH(1,1) procedure for univariate conditional variances in a standard DCC-GARCH setting (see Engle, 2002). For a N-dimensional time series (N can be also very large), first estimate N univariate spline-GARCH models for the individual conditional variances. Then estimate conditional correlations in the classical way. Another extension is for non-stationary models with time-varying parameters (and hence time-varying volatility function). Exemplifying this approach, 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 of the starting function) 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

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

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

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

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

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

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

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

For both parts of the right hand side, we can invoke a uniform law of large numbers where uniformity is with respect to β. More precisely, for the first term, the function is linear and hence Lipschitz with bounded Lipschitz constant since the B-spline basis functions are bounded (moreover, the function is bounded). Hence, using our assumptions (A1) and (A2), a uniform law of large numbers follows by Theorem 2.2 and Corollary 2.3 from Andrews and Pollard (1994). For the second term, the functions

$$\frac{X_t^2}{\exp(g_p(\beta; X_{t-1}, \ldots, X_{t-p}))} \tag{A.4}$$

are not bounded. We invoke Theorem 3 from de Jong (1998): note that the α-mixing property of $(X_t)_{t\in\mathbb{Z}}$ implies the α-mixing property of the vector-valued process $(X_{t-p}, \ldots, X_t)_{t\in\mathbb{Z}}$ (with fixed p) and the latter process is also near-epoch dependent on itself with trivial coefficients $v(q) \equiv 0$ for all $q \geq 1$, de Jong (1998, p. 249). The Lipschitz-conditions in de Jong (1998) follow by using differentiability of (A.4) with respect to β and with respect to $X_{t-p}, \ldots, X_t$ and using boundedness of $g_p(\beta; X_{t-1}, \ldots, X_{t-p})$ with respect to $\beta \in \mathcal{C}$ and with respect to $X_{t-1}, \ldots, X_{t-p}$. Therefore, formula (A.3) holds and the proof of Theorem 1 is complete. □

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.

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.

Ango Nze, P. and Doukhan, P. (2004). Weak dependence: models and applications to econometrics, *Econometric Theory* **20**, 995-1045.

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.

Barron, A., Birgé, L. and Massart, P. (1999). Risks bounds for model selection via penalization, *Probability Theory and Related Fields* **113**, 301-413.

Berkowitz, J. and O'Brien, J. (2001). How accurate are Value-at-Risk models at commercial banks? Finance and Economic Discussion Series, Board of Governors of the Federal Reserve System, U.S.

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

Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In: Pollard, D., Torgensen, E., Yangs, G. (Eds.), Festschrift for Lucien Le

Cam: research paper in Probability and Statistics, 55-87. Springer-Verlag, New York.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence, *Bernoulli* **4**, 329-375.

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.

Bühlmann, P. and McNeil, A.J. (2002). An algorithm for nonparametric GARCH modelling. *Computational Statistics and Data Analysis* **40**, 665-683.

Carrasco, M. and Chen, X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models, *Econometric Theory* **18**, 17-39.

Comte, F. and Rozenholc, Y. (2002). Adaptive estimation of mean and volatility functions in (auto-)regressive models, *Stochastic Processes and their Applications* **97**, 111-145.

Curci, G. and Corsi, F. (2003). A discrete sine transform approach for realized volatility measurement. NCCR FINRISK Working Paper No. 44, available under *http://www.nccr-finrisk.uzh.ch/media/pdf/wp/WP044_6.pdf.*

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.

de Jong, R.M. (1998). Uniform laws of large numbers and stochastic Lipschitz-continuity, *Journal of Econometrics* **86**, 243-268.

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

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. (2002). Dynamic conditional correlation - a simple class of multivariate GARCH models. *Journal of Business and Economic Statistics* **20**, 339-350.

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.

Francq, C. and Zakoian, J.-M. (2006). Mixing properties of a general class of GARCH(1,1) models without moment assumptions on the observed process, *Econometric Theory* **22**, 815-834.

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

Laurent, B. and Massart, P. (2000). Adaptive estimation of quadratic functionals via model selection, *Annals of Statistics* **28**, 1302-1338.

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.

Hansen, P.R. and Lunde, A. (2006). Consistent ranking of volatility models, *Journal of Econometrics* **131**, 97-121.

Hansen, P.R., Lunde, A. and Nason, J.M. (2003). Choosing the best volatility models: The model confidence set approach, *Oxford Bulletin of Economics and Statistics* **65**, 839-861.

Lin, Y. (2000). Tensor product space ANOVA models, *The Annals of Statistics* **28**, 734-755.

Patton, A.J. (2006). Volatility forecast evaluation and comparison using imperfect volatility proxies. Working paper, London School of Economics.

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.

Zhang, L., Aït-Sahalia, Y. and Mykland, P.A. (2005). A tale of two time scales: determining integrated volatility with noisy high frequency data. *Journal of the American Statistical Association* **100**, 1394-1411.

Zhao, P. and Yu, B. (2007). Stagewise Lasso, *Journal of Machine Learning Research* **8**, 2701-2726.

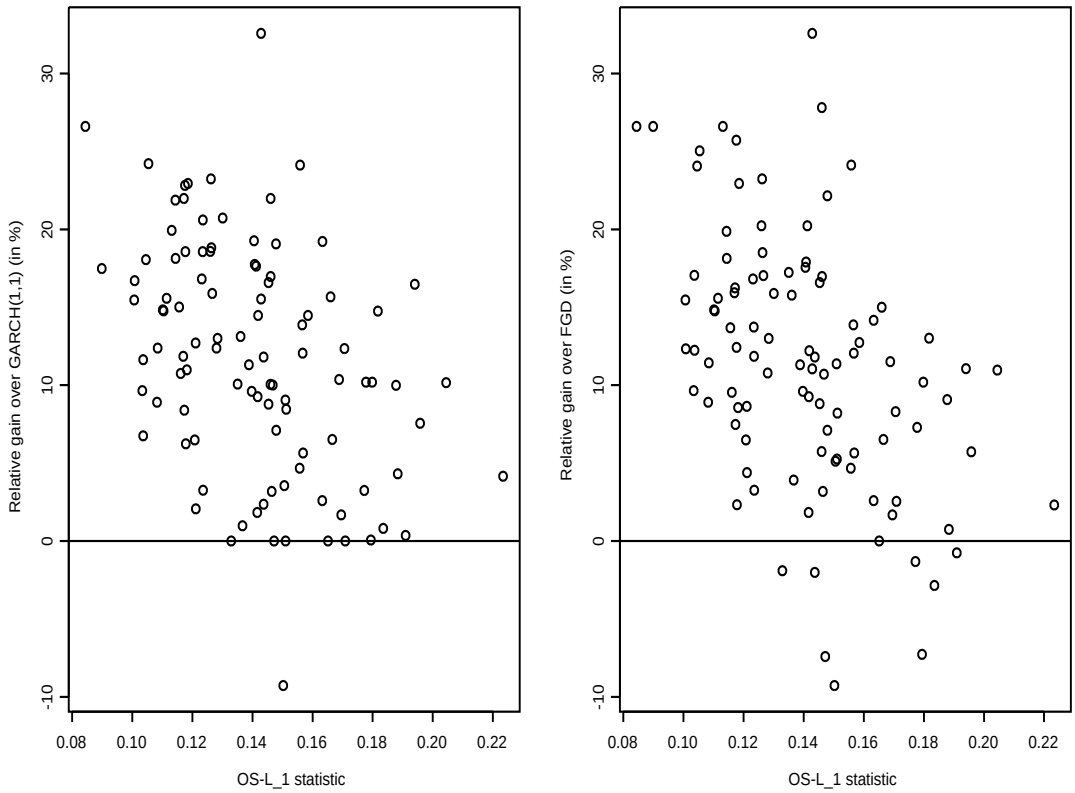

Figure 1: Plot of mean absolute errors (OS-L_1 statistic) for the (squared) volatilities estimated using the spline-GARCH(1,1) model against relative gains of mean absolute errors over the classical GARCH(1,1) model (left panel) and the FGD approach (right panel). Results are reported for 100 independent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

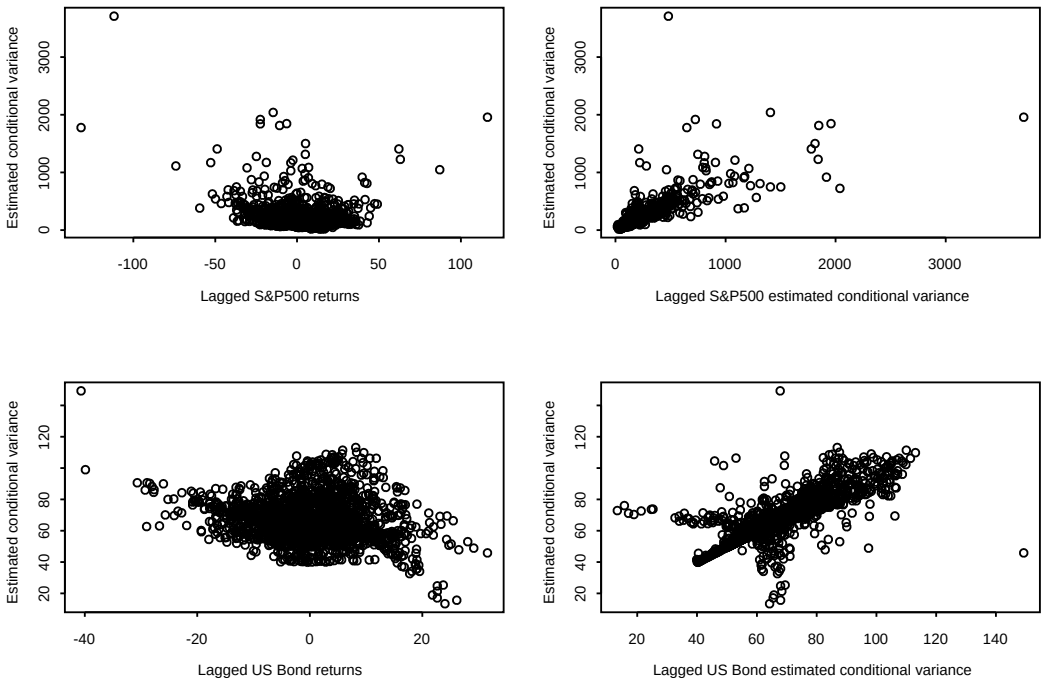

Figure 2: Conditional variance estimates of the S&P500 returns (upper panels) and 30-years US Treasury-bond returns (lower panels) obtained using the spline-GARCH(1,1) model for the in-sample time period between January 1990 and December 1998. The conditional variance estimates are plotted against the lagged returns (left panels) and against lagged estimated conditional variances (right panels).

Model	$\hat{M}_{\text{opt}}$	Averaged IS-			Averaged OS-		
		− log-lik.	L_1	L_2	− log-lik.	L_1	L_2
GARCH(1,1)		1132.78	0.1602	0.1995	1143.81	0.1596	0.1366
FGD	13.29	1126.90	0.1585	0.1981	1143.43	0.1588	0.1357
Spline-GARCH(1,1)	30.32	1120.34	0.1387	0.1720	1138.88	0.1407	0.1231

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 negative 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 terminal 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	2895.50	4189.61	34.9955	3102.79
FGD	1	10	7754.80	36.9716	2915.56	4198.67	35.7989	3159.56
Splines	11	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 negative 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 terminal 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 $\kappa = 0.2$.

Panel A: Tests for superior predictive ability: S&P500

Models	OS- − log-lik.	OS-L_1	OS-L_2
GARCH(1,1) vs. Splines	0.7381 (0.2302)	1.7411 (0.0408**)	1.9931 (0.0231**)
FGD vs. Splines	0.0655 (0.4740)	0.3201 (0.3744)	0.4973 (0.3095)

Panel B: Tests for superior predictive ability: 30-years US T-Bond

Models	OS- − log-lik.	OS-L_1	OS-L_2
GARCH(1,1) vs. Splines	1.3453 (0.0892*)	0.9995 (0.1588)	1.4604 (0.0721*)
FGD vs. Splines	5.0553 ($\approx 0^{***}$)	2.3210 (0.0101**)	1.9231 (0.0272**)

Table 3: Tests on differences of out-of-sample negative log-likelihood (− log-lik.), L_1 and L_2 performance terms for the S&P500 annualized returns (Panel A) and the 30-years US Treasury Bond annualized returns (Panel B). The out-of-sample period goes from January 1999 to October 2003, for a total of 1164 daily observations. Positive values of the statistic are always in favor of the spline-GARCH(1,1) model. p-values are reported in parentheses, with *, **, *** denoting significance at the $\leq 10\%$, 5% and 1% level, respectively.