Volatility Prediction Using Kernel Regression

Jussi Klemelä

October 16, 2017

Abstract

We study the prediction of a squared return of a financial asset in a discrete time setting. The predictor is a kernel regression estimator whose explanatory variables are a moving average of squared returns and a moving average of returns. The predictor performs better than the GARCH(1, 1) predictor and some related predictors, in the sense of the mean squared prediction error, and when applied in the estimation of conditional quantiles. The kernel predictor is able to cope with the leverage effect.

Keywords: Leverage effect, Nadaraya–Watson estimator, news impact curve, quantile estimation, risk management.
Journal of Economic Literature classification: C14 (Semiparametric and Nonparametric Methods: General), C53 (Forecasting and Prediction Methods, Simulation Methods).

1 Introduction

We are interested in predicting the squared return

$$R_{t+\eta}^2, \tag{1}$$

where t is the current time, $\eta \geq 1$ is the prediction horizon, and $R_i = S_i/S_{i-1} - 1$ is the net return of an asset with prices S_i. The prediction is done using the observed past returns $R_1, \ldots, R_t$. We use the term "volatility prediction" to mean the prediction of a squared return.

We consider the prediction of $R_{t+\eta}^2$ as an regression problem where $R_{t+\eta}^2$ is the response variable and the explanatory variable is $X_t = (X_{1t}, X_{2t})$, where X_{1t} is the square root of a moving average of past squared returns and X_{2t} is

a moving average of past returns. We apply the kernel regression predictor, defined as

$$\hat{f}(t,\eta) = \sum_{i=k}^{t-\eta} p_i(t) R_{i+\eta}^2, \tag{2}$$

where $1 < k < t - \eta$, and $p_i(t) \geq 1$ are the weights satisfying $\sum_{i=k}^{t-\eta} p_i(t) = 1$. The kernel regression predictor is a weighted average of the past squared returns. The weights $p_i(t)$ are such that they are large for those time points i for which the value X_i of the predictor is close to the current value X_t of the predictor. Thus, the predictor gives a higher weight to those time points which were similar to the current time point. The kernel regression estimator is called also the Nadaraya-Watson estimator.

The exponentially weighted moving average of squared returns is in itself a good predictor for volatility. In fact, the exponentially weighted moving average is close to the one-step GARCH$(1,1)$ predictor; see (8) and (18). The kernel predictor adds an additional layer on the top of the moving average of squared returns. Including a moving average of returns as a predictive variable makes it possible to take the leverage effect into account. The leverage effect means that previous negative returns are followed by a higher volatility than the volatility which is followed by the previous positive returns.[1]

An important ingredient of the procedure is the transformation of the values of the predictive variables before applying the kernel regression. Namely, we change the design distribution so that the marginal distributions of X_{1t} and X_{2t} are the standard normal distributions. Without this transformation the application of the kernel regression would be difficult, because it would be necessary to apply a spatially adaptive smoothing parameter. With the transformation we obtain a design distribution which has a sufficiently smooth density so that the basic kernel regression will perform well.

We compare the kernel regression predictor with the GARCH$(1,1)$ predictor using a time series of differences of cumulative sums of squared prediction errors. This graphical tool is able to reveal the relative performance of predictors over all time intervals. Thus we avoid the possible problem that the choice of the testing period would affect the conclusions. In particular, a single financial crisis can dominate the performance measurement of volatility predictors. The crises of autumn 1987 and 2008 are the two events which tend to dominate the performance measures. We avoid the problem of

[1]The leverage effect was noted in Nelson (1991). The name "leverage effect" comes from the fact that a possible explanation for this effect is that a drop in a stock price increases the financial leverage of the firm. and thus the risk is increased. An other explanation for the leverage effect is that a higher volatility requires a higher expected return in order to compensate for the increased risk (volatility feedback).

these dominating events by using the time series of differences of cumulative sums. We show that the kernel regression predictor performs better than the GARCH$(1,1)$ predictor over almost all time periods.

Testing the statistical significance of the superior performance is done by computing p-values over all time periods. In the case of p-values we do not have a convenient tool which allows to summarize results using a single time series, as in the case of the differences of cumulative sums of squared prediction errors. We use a graphical tool which shows level sets of a two-dimensional function which assigns p-values to the testing periods $[t_1, t_2]$. Time t_1 is the first argument of the function and t_2 is the second argument of the function.

There exists dozens or even hundreds of volatility predictors which are alternative to the GARCH$(1,1)$ predictor. We make additional comparisons between the GARCH$(1,1)$ predictor and an asymmetric GARCH predictor, and an asymmetric exponentially weighted moving average predictor. These asymmetric predictors are designed to take the leverage effect into account, but they are not able to consistently beat the GARCH$(1,1)$ predictor.

We compare also the performance of the kernel regression predictor and the GARCH$(1,1)$ predictor in the estimation of conditional quantiles. We show that the kernel regression predictor leads to a better quantile estimator than the GARCH$(1,1)$ predictor, over almost all time periods.

It is important, that the kernel regression predictor is able to give insight into the way how the past squared returns and the past returns affect the future volatility- We are able to show versions of the "news impact curve".

Volatility prediction can be applied in variance and volatility trading, covariance trading, quantile estimation, portfolio selection, and option pricing. See Klemelä (2018, Chapter 7.1) for details about the applications of volatility prediction.[2]

Often it is of interest to predict the realized variance

$$\sum_{i=1}^{\eta} R_{t+i}^2. \tag{3}$$

[2]The prediction of $R_{t+\eta}^2$ can be considered as the estimation of the conditional expectation $E_t(R_{t+\eta}^2)$, because the conditional expectation of $R_{t+\eta}^2$ is the best prediction of $R_{t+\eta}^2$ in the mean squared error sense. A closely related concept is the estimation of the conditional variance

$$\mathrm{Var}_t(R_{t+\eta}) = E_t\left(R_{t+\eta}^2\right) - (E_t R_{t+\eta})^2.$$

Since the squared conditional expectation $(E_t R_{t+\eta})^2$ is often negligible as compared to $E_t(R_{t+\eta}^2)$, the estimation of the conditional variance is close to the estimation of the conditional expectation of the squared return. In our case the conditional expectation E_t is taken with respect to the sigma-algebra generated by the past returns $R_1, \ldots, R_t$. It is also possible to consider larger sigma-algebras.

The predictor of the realized variance is obtained as

$$\sum_{i=1}^{\eta} \hat{f}(t,i), \tag{4}$$

where $\hat{f}(t,i)$ is the predictor in (2). The realized variance of S&P 500 (with daily logarithmic returns) is the underlying of a volatility futures contract in CBOE.

We evaluate the methods using daily S&P 500 data. It is useful to predict volatility also for lower frequencies. Define the s-period net return as

$$R_{t,s} = S_{t+s}/S_t - 1,$$

where $s \geq 1$. It is of interest to predict $R_{t,s}^2$. For example, in risk management it is important to estimate the conditional quantile for $s = 20$ day return, which is roughly one month return. An estimator of the conditional quantile can be constructed from a volatility predictor (see Section 7). We construct the predictor of $R_{t,s}^2$ by lengthening the frequency of the data from one day to s days.

We consider only the prediction of squared net returns. The prediction of the squares of logarithmic returns $\log(S_i/S_{i-1})$ leads to almost similar but not to completely identical results. The GARCH models are typically postulated for logarithmic returns. A motivation for postulating a GARCH model for logarithmic returns comes from the fact that then the price S_i is positive with probability one. The use of net returns can be motivated by the fact that the s-period loss can be written in terms of the net return:

$$L_{t,s} = -(S_{t+s} - S_t) = -S_t(S_{t+s}/S_t - 1) = -S_t R_{t,s},$$

where $s \geq 1$. In risk management it is of interest to estimate the upper quantiles of the loss distribution (value-at-risk). Volatility prediction can be applied in estimation of conditional quantiles and thus net returns are relevant in volatility prediction.

We study the performance using the daily S&P 500 data, which consists of the daily closing prices

$$\text{starting at January 4th 1950 and ending at April 2nd 2014,} \tag{5}$$

which gives 16 046 daily observations. The data is obtained from web page of Yahoo (http://finance.yahoo.com/) with ticker ^GSPC.

The computations of the article can be reproduced by the R-programs and instructions given in page http://jklm.fi/art/volapred/. The page contains also supplementary material.

The article is organized as follows. Section 2 defines the kernel predictor. Section 3 explains how the differences between cumulative sums of squared prediction errors can be used to compare predictors. Section 4 reviews previous work on volatility prediction. Section 5 compares squared prediction errors between the kernel predictor and the GARCH(1, 1) predictor. Section 6 shows news impact curves of the kernel predictor. Section 7 compares the performance of the kernel predictor and the GARCH(1, 1) predictor in the estimation of conditional quantiles. Section 8 contains a summary.

2 Definition of the Kernel Predictor

2.1 Predictive Variables

We use an exponentially weighted moving average of past squared returns and past returns as predictive variables. We consider also the case where the moving average of squared returns is replaced by the GARCH(1, 1) volatility. A transformation of the variables is needed before applying kernel regression.

Denote $Z_t = (Z_{t1}, Z_{t2})$, where

$$Z_{t1} = \left(\sum_{i=1}^{t} q_i^1(t) R_i^2 \right)^{1/2}, \qquad Z_{t2} = \sum_{i=1}^{t} q_i^2(t) R_i, \tag{6}$$

where the weights are

$$q_i^l(t) = \frac{L((t-i)/g_l)}{\sum_{j=1}^{t} L((t-j)/g_l)}, \tag{7}$$

$l = 1, 2$, $g_l > 0$ are the smoothing parameters, and $L : [0, \infty) \to \mathbf{R}$ is the kernel function. We choose a different smoothing parameter g_1 for the moving average of squared returns and g_2 for the moving average of returns. We choose $L(x) = \exp(-x) I_{[0,\infty)}$, which leads to the exponentially weighted moving average.[3]

[3]When $g = -1/\log\gamma$, $0 < \gamma < 1$, then

$$\sum_{i=1}^{t} q_i(t)\, R_i = \frac{1-\gamma}{1-\gamma^t} \sum_{i=1}^{t} \gamma^{t-i} R_i = \frac{1-\gamma}{1-\gamma^t} \sum_{i=0}^{t-1} \gamma^i R_{t-i}. \tag{8}$$

Note that the exponentially weighted moving average is often defined as

$$(1-\gamma) \sum_{i=1}^{t} \gamma^{t-i} R_i, \qquad \text{or} \qquad (1-\gamma) \sum_{i=0}^{\infty} \gamma^i R_{t-i}. \tag{9}$$

We consider also the case where the moving average $\sum_{i=1}^{t} q_i^1(t) R_i^2$ is replaced by the GARCH$(1,1)$ predictor $\hat{\sigma}_{t+1}^2$ of the volatility, so that

$$Z_{t1} = \hat{\sigma}_{t+1}; \tag{10}$$

see (18) for the formula of GARCH$(1,1)$ volatility. The replacement will lead to quite similar results.

The sample Z_t, $t = 1, \ldots, T$, of the observed values of the predictive variables is defined in (6). Let the rank of observation Z_{tl}, $t = 1, \ldots, T$, $l = 1, 2$, be the number of observations of the lth variable smaller or equal to Z_{tl}:

$$\text{rank}(Z_{tl}) = \#\left\{Z_{jl} : Z_{jl} \leq Z_{tl}, j = 1, \ldots, T\right\}.$$

We define

$$X_t = \left(\Phi^{-1}\left(\frac{\text{rank}(Z_{t1})}{T+1}\right), \Phi^{-1}\left(\frac{\text{rank}(Z_{t2})}{T+1}\right)\right), \tag{11}$$

for $t = 1, \ldots, T$, where Φ is the distribution function of the standard normal distribution. Now $X_1, \ldots, X_T$ is a sample from a distribution whose marginals are approximately standard normal, but the copula is the same as the copula of the distribution of Z_t.

Figure 1 shows (a) a scatter plot of (Z_{t1}, Z_{t2}) and (b) a scatter plot of (X_{t1}, X_{t2}). Panel (a) shows that the original data of explanatory variables is spatially inhomogeneous, and in panel (b) we have a more spatially homogeneous data. Thus, kernel regression with the original data would require spatially adaptive smoothing parameter selection, whereas with the transformed data we can apply kernel regression with a single scalar spatially nonadaptive smoothing parameter h. This simplifies the procedure considerable. While there are available good spatially adaptive regression methods, their use requires considerable care and the methods add some computational complexity.

Note that leaving Φ^{-1} out of (11) would make the marginals approximately uniformly distributed on $[0, 1]$. With the transformation of the marginals to have approximately the uniform distribution many observations would be near the boundaries of the square $[0, 1]^2$. Especially the corners would be filled with observations. Thus we would need to apply boundary kernels in kernel regression. Choosing the marginals to be approximately standard normal we obtain a design density whose tails decrease smoothly.[4]

[4]The transformation of the marginals to follow the standard normal distribution might have some kinship with winsorizing, which has been used in volatility prediction to cope with financial crises; see Bekaert and Hoerova (2014). Winsorizing is a transformation of data with puts the extreme observations equal to a specified percentile, so that the

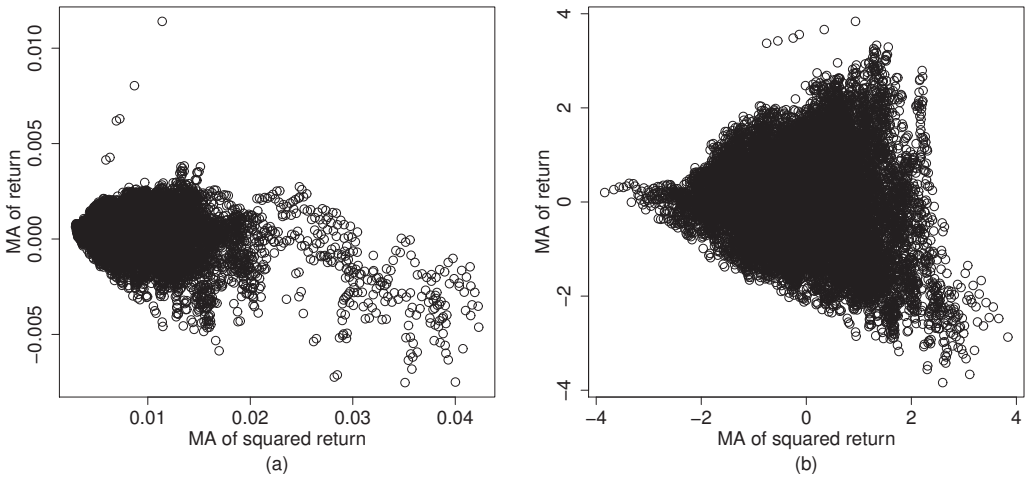

Figure 1: *Scatter plots of predictive variables.* (a) Without the transform; (b) with the transform.

2.2 Kernel Regression

The kernel regression predictor is defined as

$$\hat{f}(t) = \hat{f}_h(t, \eta) = \sum_{i=k}^{t-\eta} p_i(t) R_{i+\eta}^2, \tag{12}$$

where $\eta \geq 1$ is the prediction horizon,

$$p_i(t) = \frac{K_h(X_t - X_i)}{\sum_{j=k}^{t-\eta} K_h(X_t - X_j)}, \tag{13}$$

$K_h(x) = K(x/h)/h^d$ is the scaled kernel function, $K : \mathbf{R}^d \to \mathbf{R}$ is the kernel function, and $h > 0$ is the smoothing parameter. We choose K to be the density of the standard normal distribution. The predictor is computed from returns $R_1, \ldots, R_t$. Since X_i is computed from $R_1, \ldots, R_i$, we choose $k \geq 1$ to be a time point where the predictors can be computed to have a reasonable accuracy.

In cross-validation the smoothing parameter h is chosen at time t as the minimizer of

$$\mathrm{CV}_t(h) = \sum_{i=t_0}^{t-\eta} \left(R_{i+\eta}^2 - \hat{f}_h(i, \eta) \right)^2, \tag{14}$$

observations smaller than the 1% empirical quantile are set equal to the 1% quantile, for example. (Winsorizing is different from trimming, which removes the extreme observations altogether.)

where $t_0 + \eta \leq t \leq T$. This gives a different smoothing parameter h_t at each time t.

The normal reference rule for the choice of the smoothing parameter h in (13) puts

$$h_t = \left(\frac{4}{d+2}\right)^{1/(d+4)} t^{-1/(d+4)}, \tag{15}$$

where $d = 2$. It turns out that the normal reference rule gives smaller smoothing parameters than the cross-validation.[5]

3 The Sum of Squared Prediction Errors

We have used the sum of squared prediction errors in (14) to choose the smoothing parameter. The sum of squared prediction errors can also be used to compare the performance of the predictors. Note that we make always out-of-sample evaluations, which means that the prediction error is computed using observations which have not been used to construct the predictor.

The sum of squared prediction errors over the whole sample is not an informative tool to compare predictors, because it is very sensitive to the choice of the time period. The total sum of squared prediction errors is dominated by few financial crises. The two main events are the crises of autumn 1987 and the autumn 2008. However, the difference between the cumulative sums of squared prediction errors gives a useful tool to compare predictors. Looking at the time series of the cumulative sums of squared prediction errors reveals the comparative behavior of the predictors over all time periods, and thus we do not need to remove any data points as "outliers" from the consideration.

We assume to have returns $R_1, \ldots, R_T$. To compare two predictors $\hat{f}_1$ and $\hat{f}_2$ we compute the difference of the cumulative sums of squared prediction errors. Denote

$$D_t = \mathrm{SSPE}_t(\hat{f}_1) - \mathrm{SSPE}_t(\hat{f}_2), \tag{16}$$

where

$$\mathrm{SSPE}_t(\hat{f}) = \sum_{i=t_0}^{t-\eta} \left(R_{i+\eta}^2 - \hat{f}(i)\right)^2,$$

[5]The normal reference rule can be found in Silverman (1986, page 45), for the case of kernel density estimation. The rule can be used for kernel regression, because the kernel regression estimator can be obtained as the conditional expectation of a kernel density estimator of the distribution of (X, Y). More generally, the coordinatewise smoothing parameter is $h_i = (4/(d+2))^{1/(d+4)} t^{-1/(d+4)} \hat{\sigma}_i$, where $\hat{\sigma}_i$ is the sample standard deviation for the ith variable. In our case $\hat{\sigma}_i = 1$.

where $t_0 + \eta \leq t \leq T$. Time series $\{D_t\}$ reveals useful information about the time periods where the one predictor outperforms the other. When $D_t - D_u < 0$, then predictor $\hat{f}_1$ performs better than $\hat{f}_2$ over time period $[u, t]$, where $t > u$. When $D_t - D_u > 0$, then $\hat{f}_2$ is better over time period $[u, t]$. Indeed,

$$D_t - D_u = \sum_{i=u-\eta+1}^{t-\eta} \left(R_{i+\eta}^2 - \hat{f}_1(i)\right)^2 - \sum_{i=u-\eta+1}^{t-\eta} \left(R_{i+\eta}^2 - \hat{f}_2(i)\right)^2, \qquad (17)$$

where $t > u$. This graphical diagnostics has been applied in Goyal and Welch (2003) and Goyal and Welch (2008).

Below we choose $\hat{f}_2$ always to be the GARCH$(1, 1)$ predictor.

The supplementary material contains a discussion about the distinction between the sequential (or recursive) sum of squares of prediction errors, the rolling sum of squares of prediction errors, and the sum of squares of prediction errors which uses a division to the estimation and testing set.

4 Previous Work

A large part of literature related to volatility prediction focuses on modeling a time series of financial returns. Then a predictor is obtained as a byproduct, by deriving a formula for the conditional expectation of the squared return. In contrast, we are not focusing on modeling but our focus is on the prediction. Andersen et al. (2006) contains a review of volatility prediction, which includes a comprehensive list of references.

4.1 GARCH$(1, 1)$ Model

The GARCH$(1, 1)$ model is defined as

$$R_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 R_{t-1}^2 + \beta \sigma_{t-1}^2,$$

where $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta \geq 0$, and $\{\epsilon_t\}$ is an IID$(0, 1)$ process. The condition $\alpha_1 + \beta < 1$ implies strict stationarity. Now

$$E(R_{t+\eta}^2 \,|\, R_t, R_{t-1}, \ldots) = \bar{\sigma}^2 + (\alpha_1 + \beta)^{\eta-1} \left(\sigma_{t+1}^2 - \bar{\sigma}^2\right),$$

where

$$\bar{\sigma}^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta}.$$

We can write

$$\sigma_{t+1}^2 = \frac{\alpha_0}{1 - \beta} + \alpha_1 \sum_{k=0}^{\infty} \beta^k R_{t-k}^2. \qquad (18)$$

These formulas give the best prediction in the sense of the mean squared error.[6] We obtain a usable predictor by replacing the unknown parameters by their estimates, and by truncating the infinite sum in the formula of σ^2_{t+1}.

Note that for the one step prediction ($\eta = 1$) the GARCH$(1,1)$ predictor is close to the exponentially weighted moving average, as can be seen by comparing (8) and (18). An exponentially weighted moving average in (6)–(7) can be used to predict squared returns $R^2_{t+\eta}$ for all horizons $\eta \geq 1$, When η increases, then smoothing parameter g has to be chosen larger. When g increases, then the moving average approaches the sample mean. This means that for a large prediction horizon we use the sample mean of squared returns as a predictor.

4.2 Asymmetric GARCH$(1,1)$ Models

The leverage effect is taken into account in the model

$$\begin{aligned} \sigma_t^2 &= \alpha_0 + \alpha_1(\epsilon_{t-1} - \lambda\sigma_{t-1})^2 + \beta\sigma^2_{t-1} \\ &= \alpha_0 + \alpha_1 \frac{(R_{t-1} - \lambda\sigma^2_{t-1})^2}{\sigma^2_{t-1}} + \beta\sigma^2_{t-1}, \end{aligned} \tag{19}$$

where $\lambda \in \mathbf{R}$ is the skewness parameter. The model was applied in Heston and Nandi (2000) to price options. There are many other asymmetric GARCH models but we make comparisons with the Heston-Nandi model because under this model it is possible to derive almost closed form expressions for the prices of European call and put options. Andersen et al. (2006) identify the three most common GARCH formulations for describing the leverage effect being (1) asymmetric GARCH models, (2) threshold GARCH models, and (3) exponential GARCH models. The Heston-Nandi model in (19) is an example of an asymmetric GARCH model. See the supplementary material for a discussion about the different GARCH formulations which take the leverage effect into account.

Now

$$E(R^2_{t+\eta} \,|\, R_t, R_{t-1}, \ldots) = \bar{\sigma}^2 + (\alpha_1\lambda^2 + \beta)^{\eta-1}\left(\sigma^2_{t+1} - \bar{\sigma}^2\right),$$

where

$$\bar{\sigma}^2 = \frac{\alpha_0 + \alpha_1}{1 - \alpha_1\lambda^2 - \beta}.$$

[6] Let us denote $E_t(\,\cdot\,) = E(\,\cdot \,|\, R_t, R_{t-1}, \ldots)$. Using the fact that for $\eta \geq 1$, $E_t\sigma^2_{t+\eta} = E_t R^2_{t+\eta}$, we obtain the recursive formula

$$E_t R^2_{t+\eta} = E_t(\alpha_0 + \alpha_1 R^2_{t+\eta-1} + \beta\sigma^2_{t+\eta-1}) = \alpha_0 + (\alpha_1 + \beta)E_t R^2_{t+\eta-1}.$$

The recursion starts with $E_t R^2_{t+1} = \sigma^2_{t+1}$.

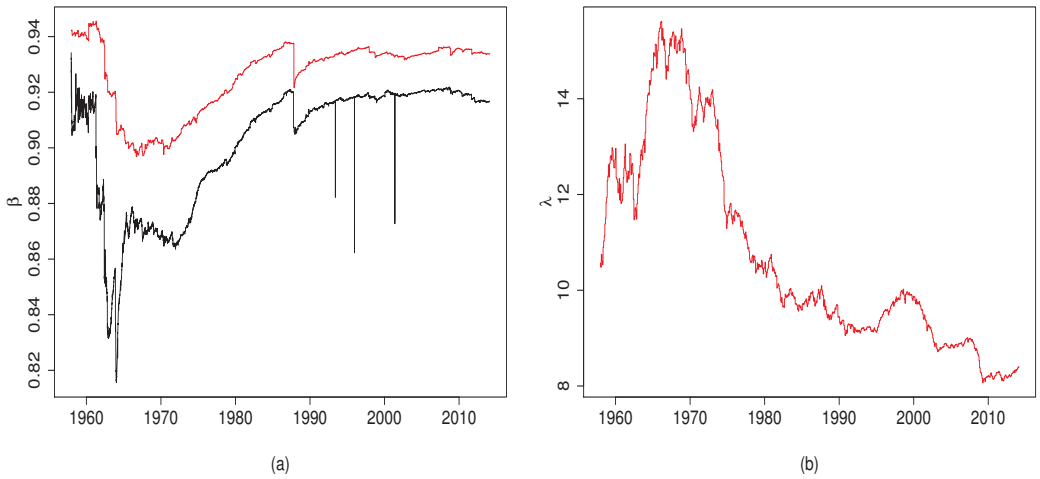

Figure 2: *Time series of GARCH parameter estimates.* (a) Estimates of β in GARCH$(1,1)$ model (black) and in the asymmetric model (red). (b) Estimates of λ in the asymmetric model.

The formula gives the best prediction in the sense of the mean squared error.[7] We obtain a usable predictor by replacing the unknown parameters by their estimates, and computing σ^2_{t+1} using recursion in (19), the starting value being $\bar{\sigma}^2$ or the sample variance of the initial observations.

Figure 2 shows the times series of parameter estimates. Panel (a) shows the sequentially estimated β. The black time series shows the estimates in the GARCH$(1,1)$ model and the red time series shows the estimates in the Heston-Nandi model (19). Panel (b) shows estimates of λ in model (19). The parameters of the standard GARCH$(1,1)$ model are estimated using R-package "tseries" and the parameters of the Heston-Nandi model are estimated using R-package "fOptions".

It is also of interest to apply an asymmetric moving average. The exponentially weighted moving average in (9) is obtained by the recursive definition $\hat{f}(t) = (1-\gamma)R_t^2 + \gamma\hat{f}(t-1)$ where $0 \leq \gamma \leq 1$. The recursive definition

[7]Let us denote $E_t(\,\cdot\,) = E(\,\cdot \mid R_t, R_{t-1}, \ldots)$. It holds that for $\eta \geq 1$, $E_t\sigma^2_{t+\eta} = E_t R^2_{t+\eta}$. Using the fact

$$E_t(\epsilon_{t+\eta-1} - \lambda\sigma_{t+\eta-1})^2 = 1 + \lambda^2 E_t\sigma^2_{t+\eta-1}$$

we obtain the recursive formula

$$E_t R^2_{t+\eta} = \alpha_0 + \alpha_1(1 + \lambda^2 E_t\sigma^2_{t+\eta-1}) + \beta E_t\sigma^2_{t+\eta-1} = \alpha_0 + \alpha_1 + (\alpha_1\lambda^2 + \beta)E_t R^2_{t+\eta-1}.$$

The recursion starts with $E_t R^2_{t+1} = \sigma^2_{t+1}$.

can be modified to take the leverage effect into account:

$$\hat{f}_\gamma(t) = (1-\gamma)\hat{f}_\gamma(t-1)\left(\frac{R_t}{\sqrt{\hat{f}(t-1)}} - \lambda\right)^2 + \gamma\hat{f}_\gamma(t-1), \qquad (20)$$

where $\lambda \in \mathbf{R}$ is the skewness parameter. This is analogous to the GARCH model in Engle and Ng (1993). The smoothing parameter g in (7) and γ are related by $g = -1/\log\gamma$.

The smoothing parameter γ of the moving average can be chosen similarily as in (14) by minimizing the criterion

$$\mathrm{CV}_t(\gamma) = \sum_{i=t_0}^{t-\eta}\left(R_{i+\eta}^2 - \hat{f}_\gamma(i)\right)^2,$$

where $t_0 + \eta \le t \le T$. Aggregation of predictors provides more stable predictions than cross-validation, because $\mathrm{CV}_t(\gamma)$ jumps during financial crises. The aggregated predictor is

$$\hat{f}(t) = \sum_{m=1}^{M} w_{t,m}\hat{f}_{\gamma_m}(t),$$

where $\gamma_1, \ldots, \gamma_M$ is a finite collection of smoothing parameters, and

$$w_{t,m} = \frac{\exp\{-\mathrm{CV}_t(\gamma_m)\}}{\sum_{l=1}^{M}\exp\{-\mathrm{CV}_t(\gamma_l)\}}.$$

Aggregation has been used in machine learning; see Györfi et al. (2006) and Györfi et al. (2012) for financial applications.

Figure 3 shows time series of

$$D_t = \mathrm{SSPE}_t(\hat{f}) - \mathrm{SSPE}_t(\hat{f}_{garch}),$$

where $\hat{f}_{garch}$ is the GARCH$(1,1)$ predictor. In panel (a) the prediction horizon is $\eta = 1$ and in panel (b) $\eta = 10$. The predictor $\hat{f}$ is the asymmetric GARCH$(1,1)$ of (19) (black with "a-GARCH"), the exponentially weighted moving average with $\lambda = 0$ and cross-validation (red with "EWMA"), the exponentially weighted moving average with $\lambda = 0.2$ and cross-validation (blue with "a-EWMA"), the exponentially weighted moving average with $\lambda = 0$ and aggregation (orange with "EWMA-agg"), and the exponentially

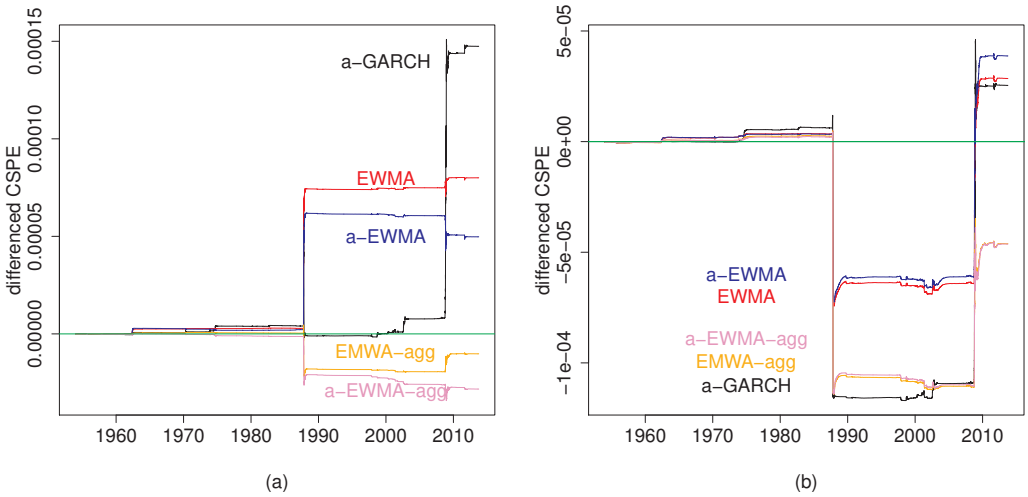

Figure 3: *Differences of cumulative sums of squared prediction errors.* (a) The prediction horizon is $\eta = 1$; (b) $\eta = 10$. Time series $D_t = \text{SSPE}_t(\hat{f}) - \text{SSPE}_t(\hat{f}_{garch})$, where $\hat{f}$ is the asymmetric GARCH (black with "a-GARCH"), the cross-validated moving average with $\lambda = 0$ (red with "EWMA") and $\lambda = 0.2$ (blue with "a-EWMA"), the aggregated moving average with $\lambda = 0$ (orange with "EWMA-agg") and $\lambda = 0.2$ (violet with "a-EWMA-agg").

weighted moving average with $\lambda = 0.2$ and aggregation (violet with "a-EWMA-agg").[8] Figure 4 has the same setting as Figure 3(a), but now panel (a) shows the first half of the time series D_t, and panel (b) the second half. Figure 5 has the same setting as Figure 3(b), but now panel (a) shows the first half of the time series D_t, and panel (b) the second half.

We see that the sums of squared prediction errors are dominated by the autumns 1987 and 2008. The basic GARCH$(1, 1)$ has the best performance until 1987, and after that the moving averages have a better performance. The asymmetric GARCH$(1, 1)$ performs worse than the moving averages during almost all time periods. The asymmetric moving average performs slightly better than the symmetric moving average. Aggregation of moving averages seems to outperform the cross-validation.

Awartani and Corradi (2005) compare the various asymmetric GARCH models to the standard GARCH$(1, 1)$ in predicting squared returns, and they find evidence that the asymmetric models improve the prediction.

[8]The cross-validation and aggregation are done using the grid $\gamma = \exp\{-1/g\}$, where $g = 5, 10, 20, 30, 40$ when $\eta = 1$ and $g = 20, 30, 40, 60, 80, 100, 120, 140, 200, 300$ when $\eta = 10$.

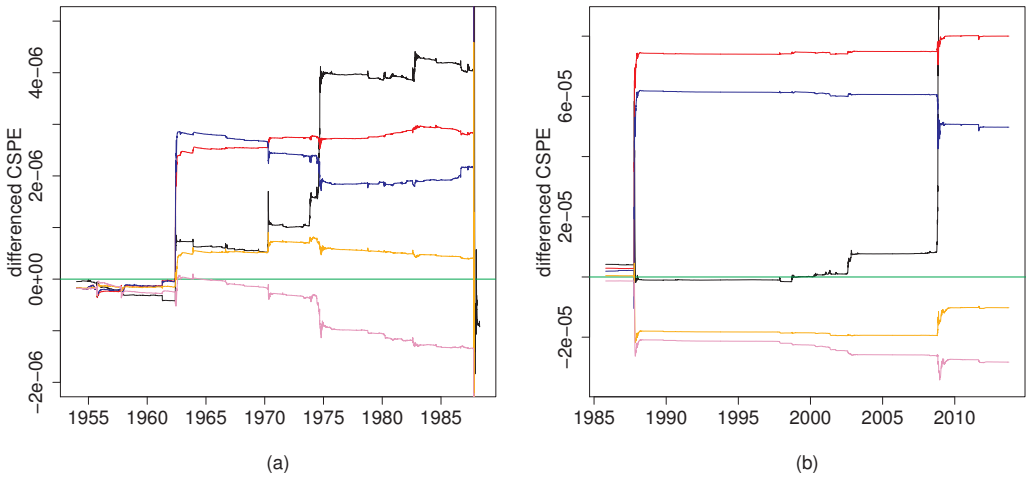

Figure 4: *Differences of cumulative sums of squared prediction errors:* $\eta = 1$. (a) The first half of time series D_t and (b) the second half, where D_t is the same as in Figure 3(a).

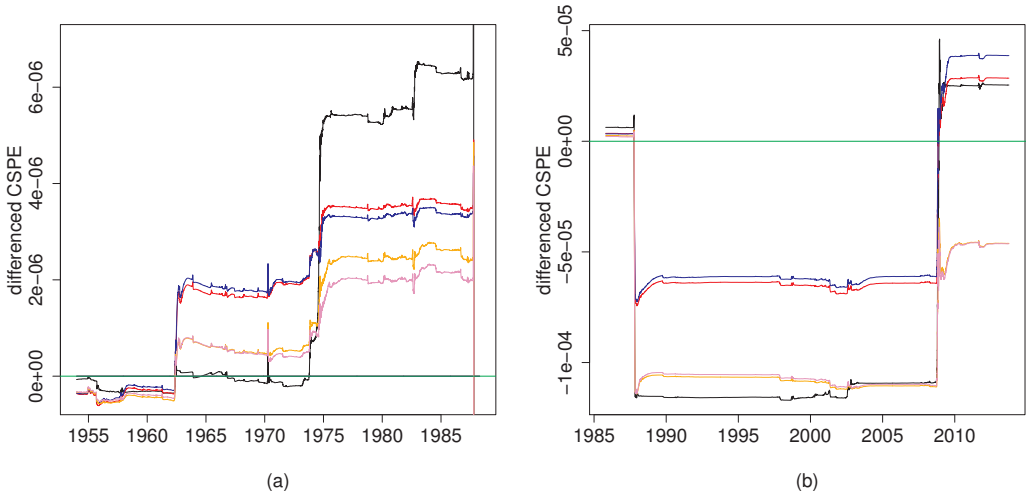

Figure 5: *Differences of cumulative sums of squared prediction errors:* $\eta = 10$. (a) The first half of time series D_t and (b) the second half, where D_t is the same as in Figure 3(b).

4.3 Nonparametric Volatility Prediction

Pagan and Schwert (1990) and Pagan and Hong (1991) consider a nonparametric volatility function $\sigma_t(R_{t-1}, \dots R_{t-p})$, which is a function of p previous returns. Härdle and Tsybakov (1997) apply a local linear estimator for the volatility function $\sigma_t(R_{t-1})$ and Härdle et al. (1998) consider the multivariate case of p previous returns. Masry and Tjøstheim (1995) consider the kernel regression estimator in a nonparametric ARCH model.

Audrino and Bühlmann (2001) considers nonparametric model

$$R_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = f(R_{t-1}, \sigma_{t-1}^2),$$

where $f : \mathbf{R}^2 \to \mathbf{R}$ is the unknown function to be estimated, and proposes tree-structured iterative estimation algorithms. Linton and Mammen (2005) consider a general ARCH(∞) model defined by

$$\sigma_t^2 = \alpha + \sum_{k=1}^{\infty} \psi_k(\theta) m(R_{t-k}), \tag{21}$$

where $\alpha \in \mathbf{R}$, $\theta \in \mathbf{R}^p$, and $m : \mathbf{R} \to \mathbf{R}$ is called a news impact curve. In the special case $\psi_j(\theta) = \theta^{j-1}$, $0 < \theta < 1$, it holds that

$$\sigma_t^2 = \theta \sigma_{t-1}^2 + m(R_{t-1}),$$

$t = 1, 2, \dots$.

Eberlein et al. (2003) consider uniformly weighted moving averages in volatility modeling. Mercurio and Spokoiny (2004) consider uniformly weighted moving averages of squared returns, whose window width is adaptively chosen, and Chen et al. (2008) applies their method to volatility prediction. Mishra et al. (2010) study a semiparametric estimator of conditional variance, which first computes residuals by dividing the observed returns with parameteric estimates of the conditional standard deviation, and then applies a kernel regression to the residuals.

Note that we mention in (30) a semiparametric model for stock prices which could be used to analyze the kernel predictor.

4.4 Intraday Data and Realized Volatility

Let us define the realized volatility as the sum of squares of five minute returns over one trading day:

$$\mathrm{RV}(t, t+1) = \sum_{i=1}^{m} R_{t+(i-1)\Delta, t+i\Delta}^2, \qquad R_{t,s} = \log(S_s / S_t),$$

where Δ is five minutes, and m is the number of five minute periods during the day $[t, t+1]$. This realized volatility can be used to approximate integral $\int_t^{t+1} \sigma_u^2 \, du$ in the continuous time model $dP_t = \mu_t dt + \sigma_t dW_t$, where $P_t = \log S_t$, μ_t is the drift, σ_t is the volatility process, and W_t is the Brownian motion; see Andersen et al. (2006, p. 818) and Chen and Ghysels (2012). Besides predicting the realized volatility $\mathrm{RV}(t, t+1)$ several articles consider predicting the longer horizon realized volatility

$$\mathrm{RV}(t, t+\eta) = \sum_{i=t}^{t+\eta-1} \mathrm{RV}(i, i+1),$$

where $\eta \geq 1$.

Andersen et al. (2007) consider linear prediction in the model where predictors are past realized volatilities over horizons of one day, one week, and one month:

$$\mathrm{RV}(t, t+\eta) = \alpha + \beta_1 \mathrm{RV}(t-1, t) + \beta_2 \mathrm{RV}(t-5, t) + \beta_3 \mathrm{RV}(t-22, t) + \epsilon_{t+\eta}. \quad (22)$$

They consider also including the indicator of a jump as a predictive variable, and replacing the realized volatility by bi-power variation. Chen and Ghysels (2012) take the asymmetry into account by introducing semi-variance as an explanatory variable, and also apply model (21) with intraday data for modeling the realized volatility. Bekaert and Hoerova (2014) consider linear model (22), and include the VIX index as an explanatory variable. Ghysels et al. (2006), Forsberg and Ghysels (2006), and Corsi (2009) contain further results of these type of models.

We have predicted random variables $R^2_{t+\eta}$, but we conjecture that a similar kernel regression could be used to predict $\mathrm{RV}(t+\eta-1, t+\eta)$. Note that using the explanatory variables $\mathrm{RV}(t-1, t)$, $\mathrm{RV}(t-5, t)$, and $\mathrm{RV}(t-22, t)$ in (22) is similar to using uniform moving averages of squared returns as predictors.

We have defined the realized volatility from daily returns in (3), and proposed an estimator in (4). We use an approach where the individual squared returns are response variables, and a predictor for the realized volatility in (4) is obtained by summing the predictors for the individual squared returns. We believe that our approach has advantages compared to (22), where a long horizon realized variance is taken as the response variable, because a different procedure seems to be optimal depending on the return horizon of the squared return. In particular, the window width of the moving average of returns, which is used as a predictive variable (X_2 is our notation), should be larger when predicting squared returns with a longer prediction horizon.

The kernel regression predictor is easy to extend by adding more explanatory variables, like the VIX index. Kernel regression suffers from the curse of dimension when the number of explanatory variables is large, but using four to five explanatory variables is possible, when the sample size is several thousands of observations.

5 Evaluation of the Kernel Prediction

We compare the kernel regression predictor to the GARCH$(1,1)$ predictor. GARCH$(1,1)$ can be considered as a workhorse of current risk management. The comparisons of Section 4.2 indicate that the asymmetric versions of GARCH$(1,1)$ do not bring obvious improvements to the standard GARCH$(1,1)$.

We compare the performance over all subperiods $[t_1, t_2]$, where $T_1 \leq t_1 < t_2 \leq T_2$, and $[T_1, T_2]$ is the total time period, given in (5). Note however, that the predictors are computed sequentially using data on the period $[T_1, t]$, for $t \in [t_1, t_2]$. In most other studies only a couple of time periods are analyzed (one time period containing a crisis and an other time period which does not contain a crisis.)

5.1 Squared Prediction Error

We compare the differences of cumulative sums of squared prediction errors, defined in (16). That is, we study time series $\mathrm{SSPE}_t(\hat{f}) - \mathrm{SSPE}_t(\hat{f}_{garch})$, where $\hat{f}$ is a kernel regression predictor and $\hat{f}_{garch}$ is the GARCH$(1,1)$ predictor.

We use two versions of kernel regression. The first version uses the moving average of squared returns as an explanatory variable, as in (6). The second version uses the GARCH$(1,1)$ volatility as an explanatory variable, as in (10).

Figure 6 considers the prediction horizon $\eta = 1$. The orange time series shows the case where $\hat{f}$ is the kernel predictor with exponentially weighted moving average (EWMA) volatility as the first explanatory variable as in (6), with smoothing parameter $g_1 = 20$. The black time series shows the case where $\hat{f}$ is the kernel predictor with GARCH$(1,1)$ volatility as the first explanatory variable, as in (10). The smoothing parameter of the EWMA of the returns is $g_2 = 10$. Panel (b) shows the time series of the cross-validation smoothing parameters for the kernel predictor with the GARCH$(1,1)$ volatility as the first explanatory variable (black), and the time series of the smoothing parameter chosen by the normal reference rule in (15) (blue). Figure 7

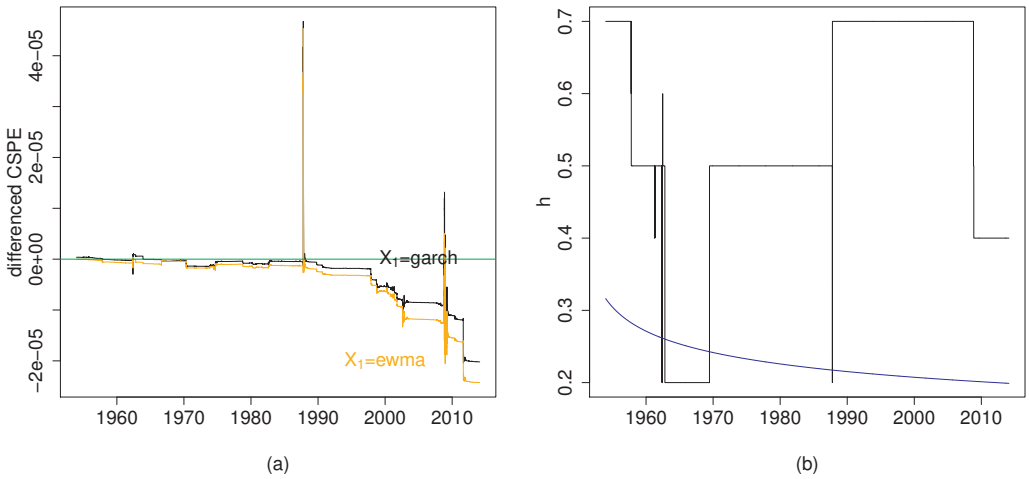

Figure 6: *Prediction horizon* $\eta = 1$. (a) Time series $\mathrm{SSPE}_t(\hat{f}) - \mathrm{SSPE}_t(\hat{f}_{garch})$. The explanatory variable is GARCH$(1,1)$ volatility (black) and the exponentially weighted moving average (orange). (b) The time series of chosen smoothing parameters h (black) and the time series of h for the normal reference rule (blue).

zooms into the time series of Figure 6(a). Panel (a) shows the beginning of the time series and panel (b) shows the end of the time series.

We see that the kernel regression beats GARCH$(1,1)$ rather constantly, because the time series is almost monotonically decreasing. There are some jumps in the time series. It seems that the exponentially weighted moving average of squared returns works slightly better than the GARCH$(1,1)$ volatility as a predictive variable of the kernel regression.

Figure 8 considers the prediction horizon $\eta = 10$. Otherwise it is similar to Figure 6. Figure 9 zooms into the time series of Figure 8(a). Panel (a) shows the beginning of the time series and panel (b) shows the end of the time series.

We see that the autumn 1987 favors the kernel predictor, and autumn 2008 favors the GARCH$(1,1)$ predictor. Otherwise, the time series is mostly decreasing, which means that the kernel predictor has an advantage. The period about 1975-1985 slightly favors GARCH$(1,1)$.

Figure 10 considers return horizon $s = 10$. Otherwise it is similar to Figure 6. We predict $R_{t,s}^2$, where $R_{t,s} = S_{t+s}/S_t - 1$. We see that the kernel predictor performs mostly better, because the time series is almost monotonically decreasing.

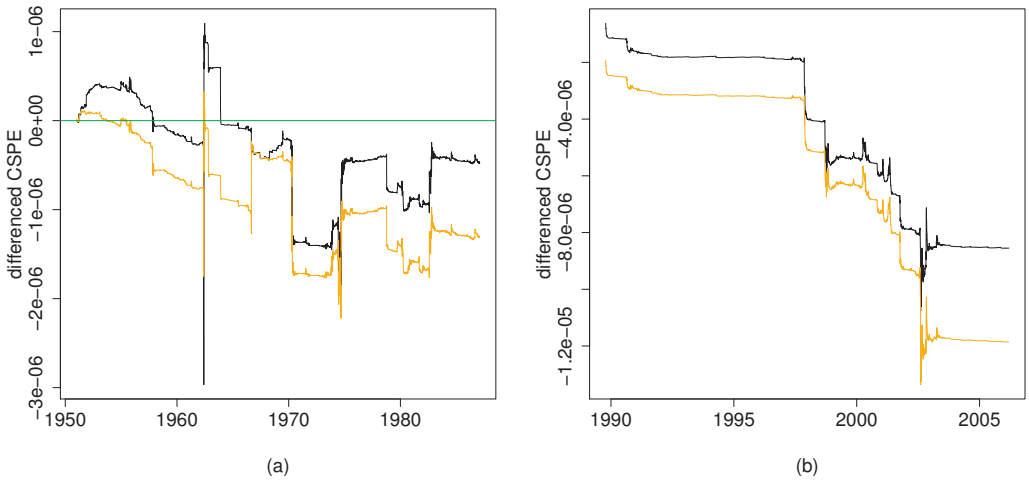

Figure 7: *Prediction horizon* $\eta = 1$. Differences of cumulative sums of squared prediction errors. (a) The beginning of the time series and (b) the end of the time series. The explanatory variable is the GARCH$(1,1)$ volatility (black) and the exponentially weighted moving average (orange).

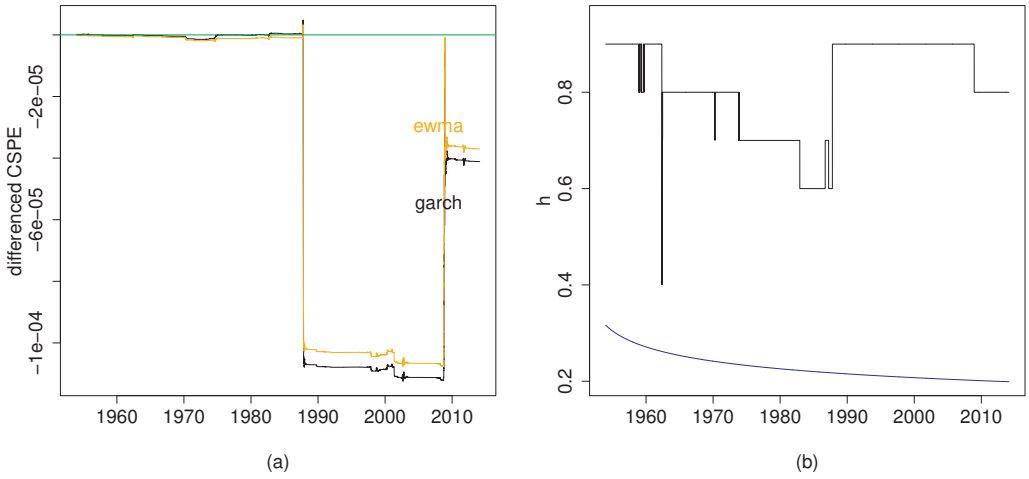

Figure 8: *Prediction horizon* $\eta = 10$. (a) Time series $\mathrm{SSPE}_t(\hat{f}) - \mathrm{SSPE}_t(\hat{f}_{garch})$. The explanatory variable is GARCH$(1,1)$ volatility (black) and the exponentially weighted moving average (orange). (b) The time series of chosen smoothing parameters h (black) and the time series of h for the normal reference rule (blue).

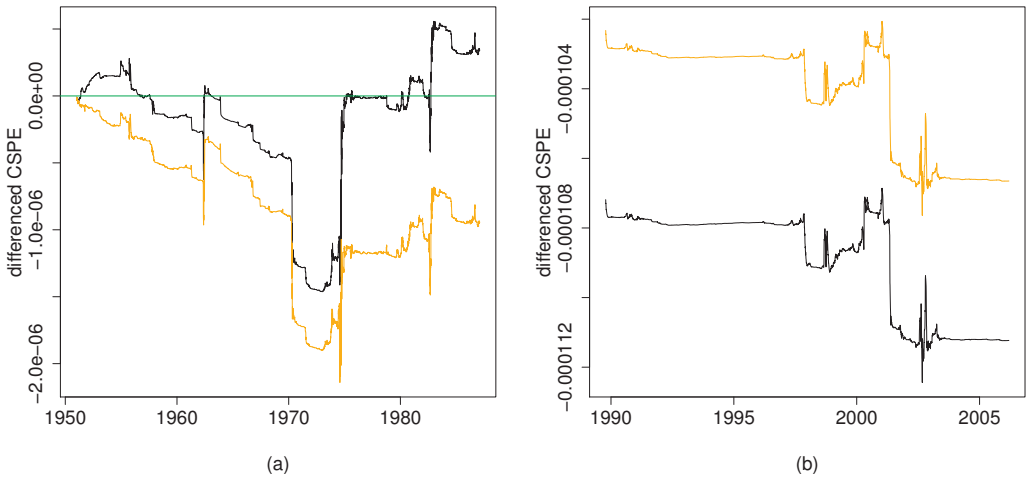

Figure 9: *Prediction horizon* $\eta = 10$. Differences of cumulative sums of squared prediction errors. (a) The beginning of the time series and (b) the end of the time series. The explanatory variable is the GARCH$(1,1)$ volatility (black) and the exponentially weighted moving average (orange).

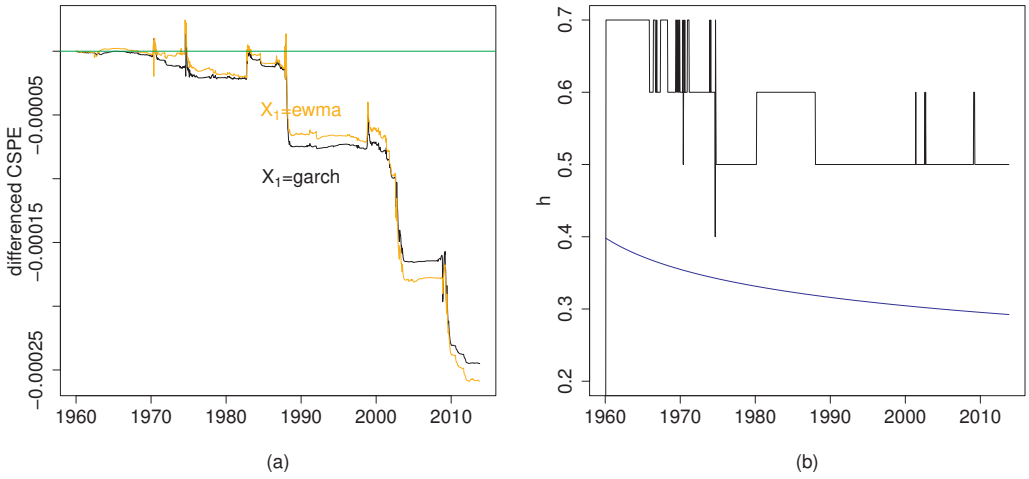

Figure 10: *Return horizon* $s = 10$ *and prediction horizon* $\eta = 1$. (a) Time series $\mathrm{SSPE}_t(\hat{f}) - \mathrm{SSPE}_t(\hat{f}_{garch})$. The explanatory variable is the GARCH$(1,1)$ volatility (black) and the exponentially weighted moving average (orange). (b) The time series of chosen smoothing parameters h (black) and the time series of h for the normal reference rule (blue).

5.2 Statistical Significance

It is of interest to find p-values for testing the hypothesis

$$H_0 : E(D_t) = 0, \qquad H_1 : E(D_t) < 0.$$

where

$$D_t = \mathrm{SSPE}_t(\hat{f}) - \mathrm{SSPE}_t(\hat{f}^{garch}) = \sum_{i=t_0+\eta}^{t} d_i,$$

with

$$d_i = (R_i - \hat{f}(i-\eta))^2 - (R_i - \hat{f}^{garch}(i-\eta))^2,$$

where $\hat{f}$ is the kernel regression predictor and $\hat{f}^{garch}$ is the GARCH$(1,1)$ predictor.

Random variables d_i are not identically distributed, because the sample size used to construct the predictors $\hat{f}$ and $\hat{f}^{garch}$ increases with i. Random variables d_i are not independent. However, we can assume that d_i are approximately identically distributed, and that the dependence is weak. Then it is possible to apply a central limit theorem to approximate the distribution of D_i.

Let us assume that d_i are identically distributed, so that

$$H_0 : E(d_i) = 0, \qquad H_1 : E(d_i) < 0.$$

Assume that $E|d_i|^\delta < \infty$ and $\sum_{j=1}^{\infty} \alpha_j^{1-2/\delta} < \infty$ for some constant $\delta > 2$, where α_j are the α-mixing coefficients. For a notational simplicity, let us take $t_0 + \eta = 1$ (which abuses the notation). Then,

$$t^{-1/2} \sum_{i=1}^{t} (d_i - Ed_i) \xrightarrow{d} N\left(0, \sigma^2\right), \tag{23}$$

as $t \to \infty$, where

$$\sigma^2 = \sum_{j=-\infty}^{\infty} \gamma(j) = \gamma(0) + 2\sum_{j=1}^{\infty} \gamma(j), \tag{24}$$

$\gamma(j) = \mathrm{Cov}(d_i, d_{i+j})$, and we assume that $\sigma^2 > 0$. Ibragimov and Linnik (1971, Theorem 18.4.1) gave necessary and sufficient conditions for a central limit theorem under α-mixing conditions. A proof for our statement of the central limit theorem in (23) can be found in Peligrad (1986); see also Fan

and Yao (2005, Theorem 2.21) and Billingsley (2005, Theorem 27.4). To estimate σ^2 in (24) we use the sample variance[9]

$$\hat{\sigma}^2 = \hat{\gamma}(0) = t^{-1} \sum_{i=1}^{t} (d_i - \bar{d})^2,$$

where $\bar{d}$ is the sequential sample mean.

Let us choose the test statistics

$$\mathrm{TS} = t^{-1/2} \hat{\sigma}^{-1} \sum_{i=1}^{t} d_i.$$

Under H_0,

$$\mathrm{TS} \xrightarrow{d} N(0,1),$$

as $t \to \infty$. Small values of the test statistics lead to the rejection of the null hypothesis. Let the observed value of TS be equal to y. The p-value is equal to $P(\mathrm{TS} < y) \approx \Phi(y)$, where Φ is the distribution function of the standard normal distribution. The test statistics is along the lines of Diebold and Mariano (1995) and West (1996). Giacomini and White (2006) proposes a generalization of the test, which is discussed in the supplementary material.

Figure 11 shows level sets of function $p(t_1, t_2)$, which assigns the p-value to the period $[t_1, t_2]$, where t_1 is the beginning of the period, t_2 is the end of the period, and the length of the period is at least one year (250 trading days). The red region shows the periods where $p(t_1, t_2) \leq 0.01$. The orange region adds the periods where $p(t_1, t_2) \leq 0.05$. Panel (a) shows the case of prediction horizon $\eta = 1$ day. Panel (b) shows the case of prediction horizon $\eta = 10$ days. When $\eta = 1$, then the smallest p-values are obtained during periods $[t_1, t_2]$, where t_1 is in the range 1990-2000 and t_2 is in the range 1990-2010. When $\eta = 10$, then smallest p-values are obtained during a larger collection of time periods: the smallest p-values are obtained during periods $[t_1, t_2]$, where t_1 is in the range 1955-1990 and t_2 is in the range 1990-2010.

[9]The sample variance leads to almost same results as

$$\hat{\sigma}^2 = \hat{\gamma}(0) + 2 \sum_{j=1}^{t-1} w(j) \hat{\gamma}(j), \tag{25}$$

where $\hat{\gamma}(j) = \frac{1}{t} \sum_{i=1}^{t-j} \left(d_i - \bar{d}\right) \left(d_{i+j} - \bar{d}\right)$, for $j = 0, \ldots, t-1$, $\bar{d}$ is the sample mean, and the weights are defined as $w(j) = L(j/g)$, where $L : [0, \infty) \to [0, 1]$ is a kernel function satisfying $L(0) = 1$ and $|L(x)| \leq 1$ for all $x > 0$. For example, $L(x) = \max\{1 - x, 0\}$ and $1 \leq g \leq t$. The sample variance is obtained by choosing $L(x) = I_{[0,1)}(x)$ and $g = 1$, because then $w(j) = 0$ for $j \geq 1$. The idea of using weights in asymptotic covariance estimation can be found in Newey and West (1987).

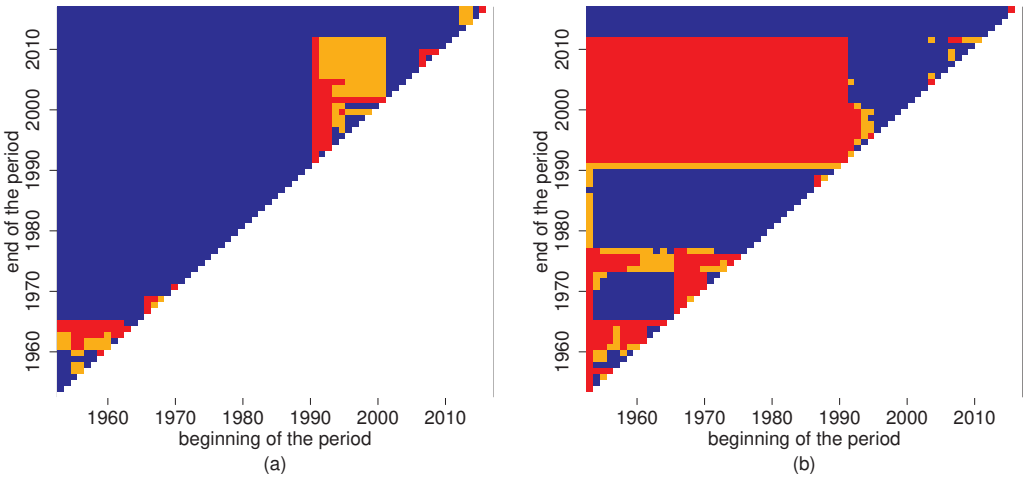

Figure 11: *Image of p-values* (a) Prediction horizon $\eta = 1$; (b) $\eta = 10$. The red region shows the periods where $p(t_1, t_2) \leq 0.01$. The orange region adds the periods where $p(t_1, t_2) \leq 0.05$.

6 The News Impact Curve

The news impact curve was introduced in Engle and Ng (1993). Linton (2009) defines the news impact curve as the relationship between σ_t^2 and $y_{t-1} = y$ holding past values σ_{t-1}^2 constant at some level σ^2. For example, in the GARCH(1, 1) model the news impact curve is

$$m(y, \sigma^2) = \alpha_0 + \alpha_1 y^2 + \beta\sigma^2.$$

We study estimates of the regression function

$$f(x_1, x_2) = E(R_{t+\eta}^2 \,|\, X_1 = x_1, X_2 = x_2),$$

where x_1 is the square root of the exponentially weighted moving average of squared returns, and x_2 is the exponentially weighted moving average of the returns. A close relative to the news impact curve is a slice

$$x_2 \mapsto f(x_1, x_2),$$

where x_1 is fixed. In this function the argument x_2 is not a one day return but a moving average of past returns.

Figure 12 shows plots of a regression function estimate for prediction horizon $\eta = 1$. Panel (a) shows a contour plot and panel (b) shows a perspective

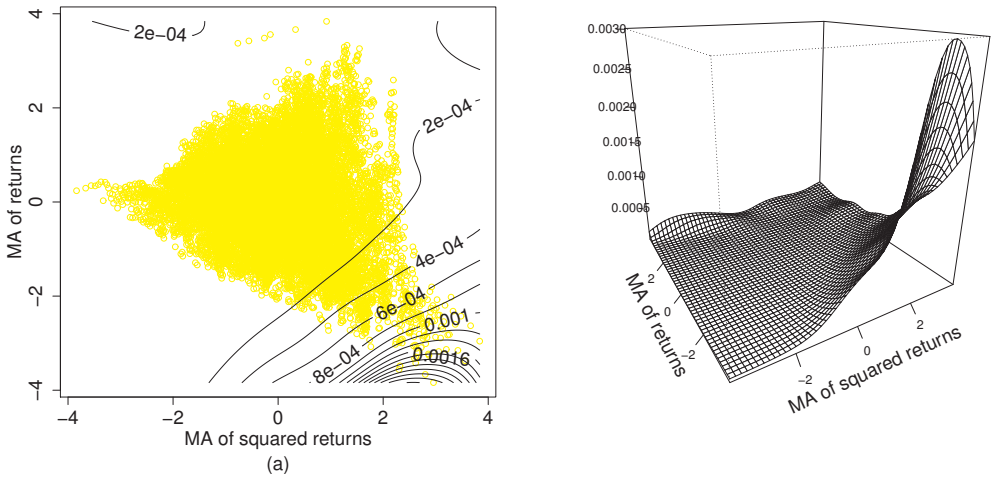

Figure 12: *Plots of the regression function estimate.* (a) A contour plot and (b) a perspective plot,

plot. The smoothing parameter of the moving average of squared returns is $g_1 = 20$ and the smoothing parameter of the moving average of returns is $g_2 = 10$. The smoothing parameter of the kernel regression is $h = 0.3$.

Figure 13 shows slices of the regression estimate of Figure 12. Panel (a) shows slices $x_1 \mapsto \hat{f}(x_1, x_2)$ for several values of x_2. Panel (b) shows slices $x_2 \mapsto \hat{f}(x_1, x_2)$ for several values of x_1. Now we have scaled the arguments from the range of the standard normal distribution to the original range of moving averages of squared returns and moving averages of returns.

Figure 14 shows three slices $x_2 \mapsto \hat{f}(x_1, x_2)$ for (a) a low value of x_1, (b) x_1 close to zero, and (c) a large value of x_1. Previous studies have found nonsymmetric u-shaped news impact curves. The nonsymmetry is such that a moderately positive previous day return leads to a next day smaller than average volatility; see Linton and Mammen (2005, Figure 4) and Chen and Ghysels (2012, Figures 1-3). We observe a similar nonsymmetric u-shape. In addition, we can study how the previous volatility changes the shape of the news impact curve. A low previous volatility in panel (a) leads to a news impact curve which takes larger values for high returns than for small returns. A high previous volatility in panel (c) leads to a news impact curve which takes smaller values for high returns than for small returns.

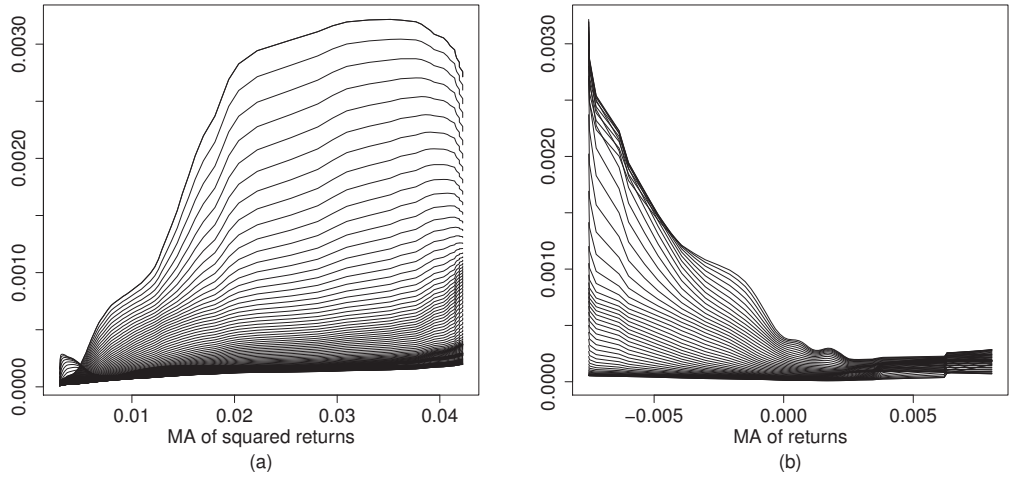

Figure 13: *Slices of the regression function estimate.* (a) Slices $x_1 \mapsto \hat{f}(x_1, x_2)$ for several values of x_2. (b) Slices $x_2 \mapsto \hat{f}(x_1, x_2)$ for several values of x_1.

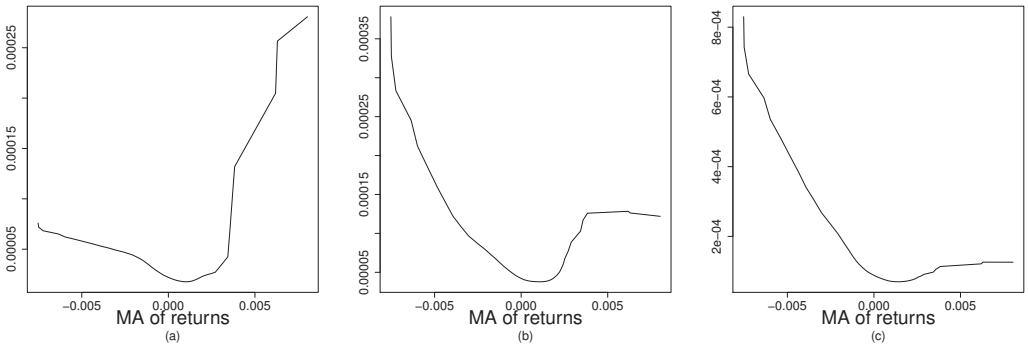

Figure 14: *Slices $x_2 \mapsto \hat{f}(x_1, x_2)$ of the regression function estimate.* (a) A low value of x_1, (b) a value of x_1 close to zero, and (c) a large value of x_1.

7 Estimation of Conditional Quantiles

We apply volatility prediction to quantile estimation. The pth conditional quantile is defined as

$$q_{t+1} = \inf \{x \in \mathbf{R} : F_{t+1}(x) \geq p\},$$

where $F_{t+1}(x) = P(R_{t+1} \leq x \,|\, R_t, R_{t-1}, \ldots)$ is the conditional distribution function of the return $R_{t+1} = S_{t+1}/S_t - 1$ and $0 < p < 1$. For a continuous distribution function the pth quantile x satisfies $F_{t+1}(x) = p$, and the pth conditional quantile is equal to

$$q_{t+1} = F_{t+1}^{-1}(p),$$

where $F_{t+1}^{-1} : (0,1) \to \mathbf{R}$ is the inverse of the distribution function. Note that the in the noncontinuous case the quantile is the generalized inverse of F_{t+1}, denoted often as $F_{t+1}^{\leftarrow}(p) = \inf \{x \in \mathbf{R} : F_{t+1}(x) \geq p\}$; see McNeil et al. (2005, p. 39).

Quantiles have a direct interpretation in risk management: the pth quantile is the smallest value x such that the probability that the return is smaller or equal to x is larger or equal to p. Define the loss of a portfolio as $L_{t+1} = -(S_{t+1} - S_t)$, where S_t is the value of the portfolio at time t. An upper quantile of the distribution of the loss is called the value-at-risk. The value-at-risk is the smallest value x such that the probability that the loss is larger than x has a probability less than $1 - p$.

In quantile estimation it seems to be of a lesser interest to estimate the conditional quantile of $R_{t+\eta}$ for $\eta > 1$. Instead, it is of interest to estimate the conditional quantile of $R(t,s) = S_{t+s}/S_t - 1$, where $s \geq 1$ is the return horizon. We do this by moving to a lower frequency data.

7.1 Performance Measurement in Quantile Estimation

Define the loss function for quantile estimation as

$$\rho_p(x) = x\,[p - I_{(-\infty,0)}(x))] = \begin{cases} x(p-1), & \text{if } x < 0, \\ xp, & \text{if } x \geq 0, \end{cases} \tag{26}$$

for $0 < p < 1$. For $p = 1/2$ we have $\rho_p(x) = |x|/2$, but we are interested in cases where p is close to zero. To compare two quantile estimators $\hat{q}^1$ and $\hat{q}^2$ we compute the difference of the cumulative sums of losses. Denote

$$D_t = \mathrm{SL}_t(\hat{q}^1) - \mathrm{SL}_t(\hat{q}^2),$$

where

$$\mathrm{SL}_t(\hat{q}) = \sum_{i=t_0}^{t-1} \rho_p\left(R_{i+1} - \hat{q}_{i+1}\right),$$

where $t_0 + 1 \leq t \leq T$. We begin to evaluate the performance of the estimator after t_0 observations are available, because any estimator can behave erratically when only a couple of observations are used for its construction. When $D_{t_1} > D_{t_2}$, then estimator $\hat{q}_i^1$ performs better on time period $[t_1, t_2]$ than estimator $\hat{q}_i^2$. Thus, a single time series plot of D_t reveals all time periods where the first estimator is better than the second estimator, as explained in the connection of (16). An alternative performance measure in quantile estimation is the use of the number of exceedances, which is discussed in the supplementary material.

7.2 Quantile Estimation Using Volatility Prediction

Let us write the return as

$$R_t = \mu_t + \sigma_t \epsilon_t,$$

where μ_t is the conditional mean and σ_t is the conditional standard deviation. For the financial returns the signal (the expected return) is typically of a lower order than the noise, and thus in quantile estimation the location μ_t can usually be ignored. We do not ignore μ_t but use the sample mean to estimate μ_t, instead of using any more sophisticated methods. We use the conditional quantile estimator

$$\hat{q}_{t+1} = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\, \hat{q}(\epsilon), \tag{27}$$

where $\hat{\mu}_{t+1}$ is the prediction of R_{t+1}, $\hat{\sigma}_{t+1}$ is the predicted volatility (an estimator of the conditional standard deviation of R_{t+1}), and $\hat{q}(\epsilon)$ is an estimator of the pth quantile of the distribution of $\epsilon_t = (R_t - \mu_t)/\sigma_t$.

We consider two cases. First,

$$\hat{q}(\epsilon) = \sqrt{\frac{\nu - 2}{\nu}}\; t_\nu^{-1}(p), \tag{28}$$

where t_ν is the distribution function of the t-distribution with ν degrees of freedom, $\nu > 2$. Second, $\hat{q}(\epsilon)$ is the pth empirical quantile of the residuals. The residuals are

$$\hat{\epsilon}_1 = (R_1 - \hat{\mu}_1)/\hat{\sigma}_1, \ldots, \hat{\epsilon}_t = (R_t - \hat{\mu}_t)/\hat{\sigma}_t.$$

We define the empirical quantile as

$$\hat{q}(\epsilon) = \hat{\epsilon}_{(\lceil tp \rceil)}, \tag{29}$$

where $\hat{\epsilon}_{(1)} < \cdots < \hat{\epsilon}_{(t)}$ are the ordered residuals, and $\lceil x \rceil$ is the smallest integer $\geq x$. The empirical quantile is obtained from the generalized inverse of the empirical distribution function. Thus, the use of an empirical quantile of residuals makes sense only in the conditional quantile estimation. The method of using empirical quantiles of residuals was suggested in Fan and Gu (2003).

Chen et al. (2008) models the distribution of the residuals using the hyperbolic family of distributions, and uses as the volatility predictor the adaptive moving average of Mercurio and Spokoiny (2004).

7.3 Smoothing Parameter Selection

The smoothing parameter can be chosen similar to (14), but replacing the squared prediction error with the loss of the quantile estimation. We choose the smoothing parameter h minimizing

$$\mathrm{CV}_t(h) = \sum_{i=t_0}^{t-1} \rho_p \left(R_{i+1} - \hat{q}_{i+1}(h) \right),$$

where $t_0 + 1 \leq t \leq T$, and $\hat{q}_{i+1}(h)$ is a quantile estimator whose kernel regression predictor has smoothing parameter h.

7.4 Quantile Estimation for S&P 500

We estimate the quantile for level $p = 1\%$.

Figure 15 considers estimation of the quantiles of $R_{t+1} = S_{t+1}/S_t - 1$. Panel (a) shows time series $\mathrm{SL}_t(\hat{q}) - \mathrm{SL}_t(\hat{q}^{garch})$, where $\hat{q}$ is the conditional quantile estimator with the kernel regression predictor and $\hat{q}^{garch}$ is the conditional quantile estimator with the GARCH$(1,1)$ predictor. We consider the case (28) with Student residuals with degrees of freedom $\nu = 5$ (black), and the case (29) with the empirical quantiles (orange). Panel (b) shows the time series of cross-validated smoothing parameter h, for the case of Student residuals (black), empirical quantiles (orange), and the curve of smoothing parameters corresponding to the normal reference rule (blue). The first predictive variable of the kernel regression predictor is the moving average of squared returns with smoothing parameter $g_1 = 20$, and the second predictive variable is the moving average of returns with smoothing parameter $g_2 = 10$. We see that the kernel predictor performs better than the GARCH$(1,1)$ predictor during almost all time periods, but autumn 2008 is a time where the GARCH$(1,1)$ predictor performs better.

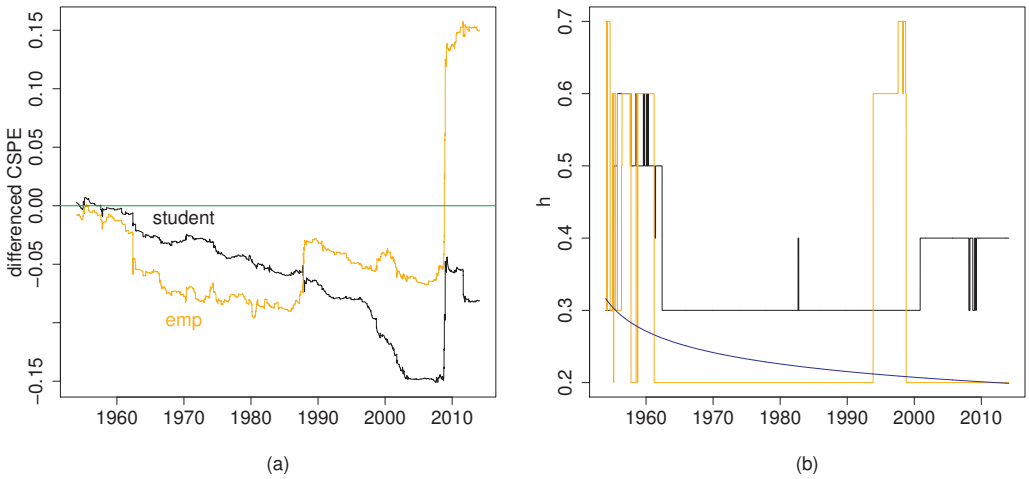

Figure 15: *One day returns.* (a) Time series $\mathrm{SL}_t(\hat{q}) - \mathrm{SL}_t(\hat{q}^{garch})$ with Student residuals (black) and empirical quantiles (orange). (b) Time series of cross-validation smoothing parameters for Student residuals (black), empirical quantiles (orange), and the smoothing parameters of the normal reference rule (blue).

Figure 16 considers estimation of the conditional quantiles of $R_{t,s} = S_{t+s}/S_t - 1$ for $s = 10$, but is otherwise similar to the setting of Figure 15. We see that with empirical quantiles the kernel predictor performs better than the GARCH$(1,1)$ predictor, during almost all time periods. With the Student residuals the GARCH$(1,1)$ predictor performs slightly better than the kernel predictor. The autumn 2008 does not have such a large impact for return horizon $s = 10$ as it has for horizon $s = 1$.

8 Conclusion

We summarize the main contributions of the article.

1. We have studied a kernel regression predictor of squared returns.
   (a) The kernel regression predictor performs well when compared to the GARCH$(1,1)$ volatility predictor, in terms of the squared prediction error and when applied in the estimation of conditional quantiles.
   (b) The predictor is semiparametric. It is not tailored for modeling returns of a particular financial asset. We conjecture that the ker-

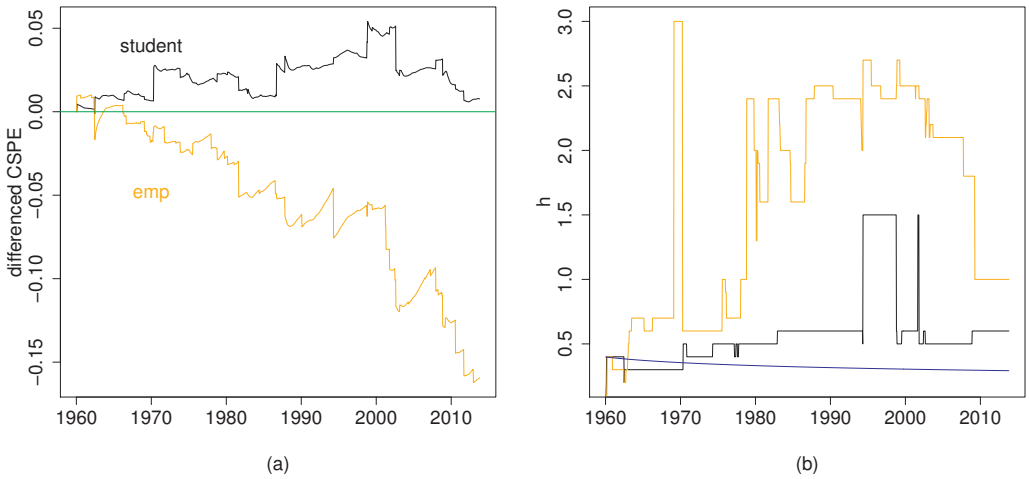

Figure 16: *Ten day returns.* (a) Time series $\mathrm{SL}_t(\hat{q}) - \mathrm{SL}_t(\hat{q}^{garch})$ with Student residuals (black) and empirical quantiles (orange). (b) Time series of cross-validation smoothing parameters for Student residuals (black), empirical quantiles (orange), and the smoothing parameters of the normal reference rule (blue).

nel regression predictor performs well in predicting the volatility of also other than S&P 500 returns.

(c) The kernel regression predictor estimates a two-dimensional regression function, which gives insight how the future volatility depends on the past volatility and on the past returns. In particular, "news impact functions" are obtained as slices of the regression function.

(d) The kernel regression predictor is easy to extend by adding more explanatory variables. The additional predictive variables could be macro economic variables, which contain information which is not included in the past returns.

2. We have applied a transformation of the design distribution to have standard normal marginals. This transformation is likely to be useful in many regression problems.

3. We have analyzed the predictive performance over all time periods. The visual tool of using level sets of a two-dimensional function of p-values is likely to be useful in the analysis of many prediction problems.

4. The recursive definition of the asymmetric exponentially weighted moving average in (20) might be new.

Our focus has been on prediction. The analysis of the asymptotic properties of the predictor falls out of the scope of the article. However, we conjecture that the kernel predictor of volatility could be analyzed in the semiparametric stochastic volatility model $R_t = \sigma_t \epsilon_t$, where $\{\epsilon_t\}$ is an IID$(0,1)$ process and

$$\begin{aligned} \sigma_t^2 &= f(s_t, m_t), \\ s_t^2 &= \alpha_0 + \alpha_1 R_{t-1}^2 + \beta s_{t-1}^2, \\ m_t &= (1-\gamma) R_{t-1} + \gamma m_{t-1}, \end{aligned} \tag{30}$$

where $f : \mathbf{R}^2 \to \mathbf{R}$ is a nonparametric function, $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta \geq 0$, $\alpha_1 + \beta < 1$, and $0 < \gamma < 1$. This model has some similarities with the single index model for conditional expectations, for example.

References

Andersen, T. G., Bollerslev, T., Christoffersen, P. F. and Diebold, F. X. (2006), Volatility and correlation forecasting, *in* G. Elliott, C. W. J. Granger and A. Timmerman, eds, 'Handbook of Economic Forecasting', North-Holland, Amsterdam, pp. 777–878.

Andersen, T. G., Bollerslev, T. and Diebold, F. X. (2007), 'Roughing it up: Including jump components in the measurement, modeling and forecasting of return volatility', *Rev. Economics and Statistics* **89**, 701–720.

Audrino, F. and Bühlmann, P. (2001), 'Tree-structured generalized autoregressive conditional heteroscedastic models', *J. Roy. Statist. Soc., Ser. B* **63**, 727–744.

Awartani, B. M. A. and Corradi, V. (2005), 'Predicting the volatility of the S&P-500 index via GARCH models: The role of asymmetries', *Int. J. Forecasting* **21**(1), 167–184.

Bekaert, G. and Hoerova, M. (2014), 'The VIX, the variance premium and stock market volatility', *J. Econometrics* **183**(2), 181–192.

Billingsley, P. (2005), *Probability and Measure*, Wiley, New York.

Chen, X. and Ghysels, E. (2012), 'News – good or bad – and its impact on volatility predictions over multiple horizons', *Rev. Financial Studies* **24**(1), 46–81.

Chen, Y., Härdle, W. and Jeong, S.-O. (2008), 'Nonparametric risk management with generalized hyperbolic distributions', *J. Amer. Statist. Assoc.* **103**(483), 910–923.

Corsi, F. (2009), 'A simple long memory model of realized volatility', *J. Financial Econometrics* **7**, 174–196.

Diebold, F. X. and Mariano, R. S. (1995), 'Comparing predictive accuracy', *J. Bus. Econ. Statist.* **13**, 225–263.

Eberlein, E., Kallsen, J. and Kristen, J. (2003), 'Risk management based on stochastic volatility', *J. Risk* **5**, 19–44.

Engle, R. F. and Ng, V. (1993), 'Measuring and testing the impact of news on volatility', *J. Finance* **43**, 1749–1778.

Fan, J. and Gu, J. (2003), 'Semiparametric estimation of Value at Risk', *Econometrics J.* **6**, 261–290.

Fan, J. and Yao, Q. (2005), *Nonlinear Time Series*, Springer, Berlin.

Forsberg, L. and Ghysels, E. (2006), 'Why do absolute returns predict volatility so well?', *J. Financial Econometrics* **6**, 31–67.

Ghysels, E., Sinko, A. and Valkanov, R. (2006), 'MIDAS regressions: Further results and new directions', *Econometric Reviews* **26**, 53–90.

Giacomini, R. and White, H. (2006), 'Tests of conditional predictive ability', *Econometrica* **74**, 1545–1578.

Goyal, A. and Welch, I. (2003), 'Predicting the equity premium with dividend ratios', *Management Science* **49**, 639–654.

Goyal, A. and Welch, I. (2008), 'A comprehensive look at the empirical performance of equity premium prediction', *Rev. Financial Studies* **21**, 1455–1508.

Györfi, G., Lugosi, G. and Udina, F. (2006), 'Nonparametric kernel-based sequential investment strategies', *Mathematical Finance* **16**(2), 337–357.

Györfi, L., Ottucsák, G. and Walk, H. (2012), *Machine Learning for Financial Engineering*, Imperial College Press, London.

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

Härdle, W., Tsybakov, A. B. and Yang, L. (1998), 'Nonparametric vector autoregression', *J. Statistical Planning and Inference* **68**(15), 221–245.

Heston, S. L. and Nandi, S. (2000), 'A closed form GARCH option valuation model', *Rev. Financial Studies* **13**, 585–625.

Ibragimov, I. A. and Linnik, Y. V. (1971), *Independent and Stationary Sequencies of Random Variables*, Walters-Noordhoff, Gröningen.

Klemelä, J. (2018), *Nonparametric Finance*, Wiley, New York.

Linton, O. B. (2009), Semiparametric and nonparametric ARCH modeling, *in* T. G. Andersen, R. A. Davis, J.-P. Kreiss and T. Mikosch, eds, 'Handbook of Financial Time Series', Springer, New York, pp. 157–167.

Linton, O. B. and Mammen, E. (2005), 'Estimating semiparametric ARCH(∞) models by kernel smoothing methods', *Econometrica* **73**, 771–836.

Masry, E. and Tjøstheim, D. (1995), 'Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality', *Econometric Theory* **11**, 258–289.

McNeil, A. J., Frey, R. and Embrechts, P. (2005), *Quantitative Risk Management: Concepts, Techniques, and Tools*, Princeton University Press, Princeton, NJ.

Mercurio, D. and Spokoiny, V. (2004), 'Statistical inference for time inhomogeneous volatility models', *Ann. Statist.* **32**, 577–602.

Mishra, S., Su, L. and Ullah, A. (2010), 'Semiparametric estimator of time series conditional variance', *J. Bus. Econ. Statist.* **28**(2), 256–274.

Nelson, D. B. (1991), 'Conditional heteroskedasticity in asset returns: A new approach', *Econometrica* **59**, 347–370.

Newey, W. K. and West, K. D. (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**(3), 703–708.

Pagan, A. R. and Hong, Y. S. (1991), Nonparametric estimation and the risk premium, *in* 'Nonparametric and Semiparametric Methods in Econometrics and Statistics', Cambridge University Press, Cambridge, UK, pp. 51–75.

Pagan, A. R. and Schwert, W. (1990), 'Alternative models for conditional volatility', *J. Econometrics* **45**, 267–290.

Peligrad, M. (1986), Recent advances in the central limit theorems and its weak invariance principle for mixing sequencies of random variables (a survey), *in* 'Dependence in Probability and Statistics', Birkhäuser, Boston, pp. 193–223.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

West, K. D. (1996), 'Asymptotic inference about predictive ability', *Econometrica* **64**, 1067–1084.