Volatility forecasting with realized measures: HEAVY vs. HAR

Bachelor Thesis
Econometrics & Operations Research

July 27, 2017

Author
Wessel Vis
409173

Supervisor
Sander Barendse

Co-reader
Francine Gresnigt

Erasmus School of Economics
Erasmus University Rotterdam

Abstract

The use of estimators of volatility based on high-frequency data has greatly improved our ability to measure and model financial market volatility. Over the past few years there has been an abundance of research on the subject of modeling return volatility. This paper compares two of the more successful models, being the HEAVY and HAR models. The HEAVY model framework as developed in Shephard and Sheppard (2009) uses two separate equations and realized measures to model the volatility. The HAR model introduced by Corsi (2009) uses an additive cascade model of volatility components defined over different time periods. This paper also uses ideas from Patton and Sheppard (2009b) for the implementation of the HAR model. To compare the models we use a predictive ability test and compare their forecast performance at different horizons. Results of these tests show that the HAR models outperform HEAVY models when using a lot of observations to estimate. But we see the opposite when using an estimation window of four years, in that case the HEAVY models perform better at almost every horizon.

1 Introduction

Volatility forecasting has always been useful in risk management, but now that volatility is directly tradable using swaps and futures, precise volatility forecasts are more important than ever. Therefore, the main focus of the research will lie in determining which model performs best as a predictive model of volatility. We hope to establish our goal by comparing the models at different horizons and see if there is a model that consistently outperforms the other, or if we can identify different situations where different models perform the best.

The use of estimators of volatility based on high-frequency data has greatly improved our ability to measure and model financial market volatility[1]. In the past decade there has been plenty of research to develop methods that exploit the information that high-frequency data provides, Andersen et al. (2009) provides an extensive overview of this literature, categorizing the various concepts and procedures relating to the measurement and modeling of volatility. The goal of this research is to compare two model types that take completely different approaches in using realized measures to forecast future volatility.

Firstly, the HEAVY model framework as developed in Shephard and Sheppard (2009), this framework consists of two equations, one equation to model the volatility of returns using a realized measure that is modeled by the second equation. This model structure is reminiscent of the ARCH models and the authors mention using ideas from papers such as Engle (2002), Engle and Gallo (2006) and Cipollini et al. (2009). Shephard and Sheppard (2009) argue that the HEAVY models are relatively easy to estimate, compared to for example the component GARCH model by Engle and Lee (1999), since their model makes use of two sources of information to determine the long-term component of volatility. Furthermore, they show that making use of the additional information that the realized measure provides results in an increased out-of-sample performance compared to standard GARCH models. This increase in performance is magnified when using parameters that are estimated to fit a specific forecast horizon. When replicating their research our results were in line with their conclusions regarding the comparative performance of the HEAVY models against the GARCH models. The paper itself does not introduce new theories regarding the estimation of models and instead uses established results from quasi-likelihood theory.

Secondly, the Heterogeneous Autoregressive (HAR) model (see Corsi, 2009; Müller et al., 1997) as outlined in Patton and Sheppard (2009b). This model regresses the realized variance on the past 1-day, 5-day, and 22-day average realized variances and is extended to include other realized measures. The main idea behind this HAR model is that different types of agents on the market drive different components of the market volatility. There are three main volatility components that are of interest for this model, the short-term traders with daily or intra-day trading frequency, the medium-term investors that adjust their portfolios weekly, and the long-term agents who trade monthly or every few months. Patton and Sheppard (2009b) introduce a couple of extensions to this HAR model based on differentiating between the effects of positive and negative (high-frequency) returns or jump variation. We have also included their extension that decomposes the first lag of the

[1] These are some of the papers that show the positive effects of using intra-day data has on modeling volatility: Andersen et al. (2003); Liu and Maheu (2005); Fleming et al. (2003); Corsi (2009); Chen and Ghysels (2011); Visser (2008); Andersen et al. (2007).

realized variance into positive and negative realized semivariance, estimators introduced by Barndorff-Nielsen et al. (2010). They show that disentangling these effects improves forecasts of future volatility up to horizons of one month or more.

Our research shows that in-sample the HAR models perform better than the HEAVY models, particularly at longer horizons. The out-of-sample analysis shows a different picture, however, with the HEAVY models beating the HAR models at horizons ranging from one week to one month.

The remainder of the paper is set up as follows. Section 2 explains more about the methods used in the paper, including estimation techniques and employed tests. The data set used is discussed in Section 3. Section 4 shows and discusses results for both in-sample and out-of-sample estimations. Section 5 concludes.

2 Methodology

2.1 HEAVY model

Shephard and Sheppard have developed a volatility model that includes realized measures, these are called "HEAVY models" (High frEquency bAsed VolatilitY models) and consist of a system that models two different quantities:

$$\begin{Bmatrix} \mathrm{Var}(r_t|\Phi_{t-1}) \\ \mathrm{E}(y_t|\Phi_{t-1}) \end{Bmatrix}, \quad t = 2, 3, \ldots, T,$$

where r_t denotes daily financial returns, y_t denotes a realized measure[2] and Φ_{t-1} is the information set at time $t-1$ containing past values of r_t and y_t.

The HEAVY model finds its roots in the ARCH literature pioneered by Engle (1982) and Bollerslev (1986) and shows similarities to the GARCH model which assumed that

$$\mathrm{Var}(r_t|\Phi_{t-1}) = \sigma_t^2 = \omega_G + \alpha_G r_{t-1}^2 + \beta_G \sigma_{t-1}^2$$

This model is easy to extend, we could for example include a realized measure, which would make it a GARCHX type model:

$$\mathrm{Var}(r_t|\Phi_{t-1}) = \sigma_t^2 = \omega_X + \alpha_X r_{t-1}^2 + \beta_X \sigma_{t-1}^2 + \gamma_X y_{t-1}$$

But when estimating this model, we find that the y_{t-1} component is the main driver of the model and α_X is very small. It seems that y_{t-1} is better at moving around the conditional average than r_{t-1}^2, so one might argue to replace the squared return with the realized measure:

$$\mathrm{Var}(r_t|\Phi_{t-1}) = h_t = \omega + \alpha y_{t-1} + \beta h_{t-1}, \tag{1}$$

with $\omega, \alpha \geq 0, \beta \in [0, 1]$. This is the first equation of the HEAVY model. However, if we want to forecast the variance of the returns multiple periods ahead, we will also need forecasts of the realized measure. Hence, we need a model for the realized measure, if we use a simple linear autoregressive model, we get the most basic example of a HEAVY model

[2]Although both HEAVY and HAR models can be used with any realized measure, this paper focuses solely on applying the models to realized variance.

$$\text{Var}(r_t|\Phi_{t-1}) = h_t = \omega + \alpha y_{t-1} + \beta h_{t-1}, \quad (2)$$
$$\text{E}(y_t|\Phi_{t-1}) = \mu_t = \omega_R + \alpha_R y_{t-1} + \beta_R \mu_{t-1}, \quad (3)$$

with $\omega_R, \alpha_R, \beta_R \geq 0, \alpha_R + \beta_R \in [0, 1]$. Equations (2) and (3) will be referred to as HEAVY-r and HEAVY-RM, respectively. Notice that by taking r_{t-1}^2 out of the GARCH equation, we have removed the feedback that GARCH models have which means that the conditional variance is solely determined by the realized measure.

Clements and Hendry (1999) explain that failure in economic forecasting might be for a large part due to long run relationships shifting, this can be solved by imposing a unit root on the model. For the HEAVY model this would mean equating $(1-\beta)(1-\alpha_R-\beta_R)$ to zero[3]. Since setting β to zero would not leave much of the HEAVY-r equation, we opt for setting $\alpha_R + \beta_R$ to one. However this means that the intercept ω_R becomes a trend slope[4], which is undesirable in a time series of volatility, so we set it to zero. Imposing these restrictions on the HEAVY model defined before we get

$$\text{Var}(r_t|\Phi_{t-1}) = h_t = \omega + \alpha y_{t-1} + \beta h_{t-1}, \quad (4)$$
$$\text{E}(y_t|\Phi_{t-1}) = \mu_t = \alpha_{IR} y_{t-1} + (1 - \alpha_{IR}) \mu_{t-1}, \quad (5)$$

with $\omega, \alpha \geq 0, \beta \in [0, 1)$ and $\alpha_{IR} \in [0, 1)$. This is referred to as the "Integrated HEAVY model". Although this model is very simple, Shephard and Sheppard show that it can produce reliable multi-period forecasts.

2.2 HAR model

Financial returns show a number of well-known stylized facts that have proven to be quite troublesome for econometric models. Strong, slowly declining autocorrelations for the absolute and squared returns, the return distributions have fat tails and show evidence of scaling and multi-scaling[5]. Standard GARCH and stochastic short-memory volatility models are unable to reproduce all of these features, hence the growing interest for modeling long-memory processes.

Long-memory is usually obtained with FIGARCH (Baillie et al., 1996) and ARFIMA (Hosking, 1981) models, or other models that make use of fractional difference operators (Granger and Joyeux, 1980), but these models are cumbersome to estimate, especially when extended. Moreover, these models lack a clear economic interpretation and the usage of the fractional difference operator requires a long build-up period, resulting in the loss of many observations. An alternative method is to view the long-memory and multi-scaling as features generated by a process, which is not actually long memory or multi-scaling. It is very difficult to distinguish between true long-memory processes and simple component models with few time scales, combined with the fact that the latter is simpler to estimate, Corsi (2009) follows this alternative view and proposes a simple

[3] See section 2.3 of Shephard and Sheppard (2009) for a derivation of this result.

[4] When we set $\alpha_R + \beta_R$ to one the first equation in section 2.3.3 of Shephard and Sheppard (2009) becomes $\Delta r_t^2 = \beta \Delta r_{t-1}^2 + \alpha\ \omega_R + \xi_t$.

[5] A process is called multi-scaling if there are different scaling exponents for different powers of the absolute returns, for a more extensive explanation of this phenomenon see Di Matteo (2007).

component model for conditional volaltility. While remaining simple and easy to estimate, this model is able to reproduce the stylized facts observed in the returns.

The standard HAR in the realized variance literature (Corsi, 2009; Müller et al., 1997) regresses realized variance on the past 1-day, 5-day, and 22-day average realized variances, these represent the short-term, medium-term, and long-term components mentioned before. As is done in Patton and Sheppard (2009b), we use a reparameterization with a clearer interpretation where the second term contains only lags 2 to 5 of the realized variance, and the third term contains only lags 6 to 22,

$$y_{t+h} = \mu + \phi_d y_t + \phi_w \left(\frac{1}{4} \sum_{i=1}^{4} y_{t-i} \right) + \phi_m \left(\frac{1}{17} \sum_{i=5}^{21} y_{t-i} \right) + \epsilon_{t+h} \tag{6}$$

where y_t is the realized variance at time t. From now on $\bar{y}_{w,t}$ denotes the average value over lags 2 to 5 and $\bar{y}_{m,t}$ denotes the average over lags 6 to 22.

In the ARCH literature it has been observed many times that negative returns have a greater impact on future volatility than positive returns. Models have been developed that exploit this relationship[6], one example of such a model is the GJR-GARCH (Glosten et al., 1993), which allows positive and negative innovations to returns to have different impacts on the conditional volatility. More recent research has found evidence that this relationship persists, even when using high-frequency returns (see Bollerslev et al., 2006; Barndorff-Nielsen et al., 2010; Visser, 2008; Chen and Ghysels, 2011). Barndorff-Nielsen et al. (2010) introduced estimators that capture the variation that is due only to positive or negative returns, realized semivariances, denoted by RS^+ and RS^- for positive and negative returns, respectively. Knowing that RV can be exactly decomposed into RS^+ and RS^- [7], Patton and Sheppard suggest an extension to (6). The next model is obtained by splitting up y_t in (6) into realized semivariances RS_t^+ and RS_t^-. The model becomes

$$y_{t+h} = \mu + \phi_d^+ RS_t^+ + \phi_d^- RS_t^- + \phi_w \bar{y}_{w,t} + \phi_m \bar{y}_{m,t} + \epsilon_{t+h} \tag{7}$$

From now on (6) will be referred to as HAR-RV and (7) will be referred to as HAR-RS. Results from Patton and Sheppard (2009b) show that the HAR-RS delivers significantly better forecasts than the HAR-RV, especially at longer horizons.

2.3 Estimation

Shephard and Sheppard (2009) mention that inference for HEAVY models is a simple application of multiplicative error models discussed by Engle (2002) who uses standard quasi-likelihood asymptotic theory. In estimating the parameters we will assume there are no links between the parameters from the HEAVY-r and HEAVY-RM models. This means we can estimate each equation separately when maximizing the quasi-likelihoods. The quasi-likelihoods are

$$\log Q_1(\omega, \alpha, \beta) = \sum_{t=2}^{T} l_t^r, \quad \text{where} \quad l_t^r = -\frac{1}{2}(\log h_t + r_t^2/h_t),$$

for the first equation, where we take $h_1 = T^{-1/2} \sum_{t=1}^{\lfloor T^{1/2} \rfloor} r_t^2$, and

[6] See for example section 1.3 in Bollerslev et al. (1994) and section 3.3 in Andersen et al. (2006).
[7] See Patton and Sheppard (2009b) for a derivation of this result.

$$\log Q_2(\omega_R, \alpha_R, \beta_R) = \sum_{t=2}^{T} l_t^{RM}, \quad \text{where} \quad l_t^{RM} = -\frac{1}{2}(\log \mu_t + RM_t/\mu_t),$$

for the second equation, where we take $\mu_1 = T^{-1/2} \sum_{t=1}^{\lfloor T^{1/2} \rfloor} RM_t$. Calculating robust standard errors is standard for the HEAVY models. Shephard and Sheppard (2009) show that the equation by equation standard errors for the HEAVY-r and HEAVY-RM are correct, even though we view the equations as a system.

For the estimation of the the HAR models we use simple Weighted Least Squares (WLS), the reason for this is that OLS would focus too much on fitting periods of high volatility, while putting less weight on less volatile periods. Since the level of variance changes significantly over our sample period and the level of the variance and the volatility of the error terms are positively related, we need to account for heteroskedasticity. We do this by using an implementation of WLS where we first estimate the model using OLS,

$$b_{\text{OLS}} = (X'X)^{-1}X'Y$$

where X is a matrix containing the regressors specific to the equation (a vector of ones, average weekly and monthly realized variance, and either realized variance or both realized semivariances) and Y is y_{t+h}. Then we construct weights as the inverse of the OLS fitted value,

$$w_t = 1/\bar{y}_t, \qquad \text{for} \quad t = 1, \ldots, T$$

where $\bar{y}_t$ are the fitted values. The next step is to perform the WLS,

$$b_{\text{WLS}} = (X'WX)^{-1}X'WY$$

where the weighting matrix W is a diagonal matrix containing the w_t. Standard errors are calculated according to the standard WLS theory.

2.4 Evaluation

Since the main objective of this paper is to compare the performance of several models, we need a means of comparing the volatility forecasts produced by the different models. We will do so by comparing the QLIK loss function, which is defined as

$$\text{Loss}\left(r_{t+s}^2, \hat{\sigma}_{t+s|t-1}^2\right) = \frac{r_{t+s}^2}{\hat{\sigma}_{t+s|t-1}^2} - \log\left(\frac{r_{t+s}^2}{\hat{\sigma}_{t+s|t-1}^2}\right) - 1, \quad s = 0, 1, \ldots, S, \tag{8}$$

where r_{t+s}^2 is the proxy we use for the variance at time $t+s$ and $\hat{\sigma}_{t+s|t-1}^2$ is some variance forecast made at time $t-1$. Patton (2011) and Patton and Sheppard (2009a) have shown that the QLIK loss function is robust to certain types of noise in the volatility proxy, this is useful as the square of returns has long been known to be a noisy proxy for the true conditional variance. Patton and Sheppard (2009a) also showed that the QLIK loss function yielded the greatest power in their Diebold-Mariono-White tests, which is very similar to the test we use here. For this reason we have chosen the QLIK as the loss function over the MSE or MAE or any of their close relatives. The test statistic is obtained by averaging the differences of the losses,

$$\widehat{L}_s = \frac{1}{T-s} \sum_{t=s+1}^{T} L_{t,s}, \quad s = 0, 1, \ldots, S, \tag{9}$$

where

$$L_{t,s} = \text{Loss}\left(r_{t+s}^2, h_{t+s|t-1}\right) - \text{Loss}\left(r_{t+s}^2, y_{t+s|t-1}\right), \quad s = 0, 1, \ldots, S. \tag{10}$$

Here $h_{t+s|t-1}$ is the forecast from the HEAVY model and $y_{t+s|t-1}$ is the corresponding forecast from a HAR model. The HEAVY model will be favored if $\widehat{L}_s$ is negative.
Then

$$\sqrt{T}\left(\hat{L}_s - L_s\right) \xrightarrow{d} N(0, V_s), \tag{11}$$

where V_s is the long-run variance of $L_{t,s}$ and can be estimated by a HAC estimator. Under the null hypothesis that the models perform equally well, L_s is equal to zero.

In the context of comparing forecasts this method is related to Diebold and Mariano (1995). This approach follows the ideas of Cox (1961b) on non-nested testing using the Vuong (1989) and Rivers and Vuong (2002) implementation which has the benefit of being valid even if neither model is correct.

2.5 Horizon tuned estimation

To increase the forecasting performance of the HEAVY models, Shephard and Sheppard (2009) implement a method of "direct forecasting", where the QLIK loss is minimized for each specific horizon. This way of tuning the parameters to produce multi-step ahead forecasts has been studied by Marcellino et al. (2006) and Ghysels et al. (2009) and dates back to Cox (1961a). In theory the iterated forecasts are more efficient if the one-period ahead model is correctly specified, but direct forecasts are more robust to model specification. Both Marcellino et al. (2006) and Ghysels et al. (2009) find that the direct forecasts are unbiased, but inefficient. The former argue that, since the single-period models are not badly misspecified, the increase in estimation variance arising from estimating the multi-period model directly outweighs the reduction in bias, which means the iterated forecasts performed better.

To obtain the horizon tuned parameters we maximize the following quasi-likelihood for each value of s

$$\log Q_{1,s}(\omega_s, \alpha_s, \beta_s) = \sum_{t=s}^{T} l_{t,s}^{r}, \quad \text{where} \quad l_{t,s}^{r} = -\frac{1}{2}\left(\log h_{t+s|t-1} + \frac{r_{t+s}^2}{h_{t+s|t-1}}\right),$$

$$\log Q_{2,s}(\omega_{R,s}, \alpha_{R,s}, \beta_{R,s}) = \sum_{t=s}^{T} l_{t,s}^{RM}, \quad \text{where} \quad l_{t,s}^{RM} = -\frac{1}{2}\left(\log \mu_{t+s|t-1} + \frac{RM_{t+s}}{\mu_{t+s|t-1}}\right),$$

Maximizing these quasi-likelihoods results in a sequence of estimators $\widehat{\omega}_s, \widehat{\alpha}_s, \widehat{\beta}_s, \widehat{\omega}_{R,s}$, $\widehat{\alpha}_{R,s}, \widehat{\beta}_{R,s}$ for each horizon s. Since the HEAVY-r equation uses the previous value from the HEAVY-RM equation, horizon tuned estimation can not be done equation by equation. One solution would be to optimize the six parameters simultaneously, but that optimization becomes very time-consuming and is liable to get stuck in a local optimum.

Therefore, when we implement the horizon tuned parameters, we use the HEAVY-r parameters optimized for one-step ahead and the HEAVY-RM parameters optimized to perform best at varying horizons.

In the context of HAR models, horizon tuned parameters are simply obtained by substituting $h = s + 1$ into (6) and (7).

3 Data

In the paper we use data on four different equity indices: S&P 500, FTSE 100, Nikkei 225 and the German DAX. The data is taken from the Oxford-Man's realized library [8], the current version of the database starts at January 3rd 2000 and is updated daily. We use data up until May 5th 2017.

Asset		r^2			RV			RS		
	T	Avol	sd	acf_1	Avol	sd	acf_1	Avol	sd	acf_1
S&P 500	4333	18.7	4.34	.213	17.2	2.58	.671	12.2	1.41	.552
FTSE 100	4354	15.0	2.25	.184	14.8	1.65	.571	10.5	0.87	.508
Nikkei 225	4200	18.6	4.92	.251	16.7	1.74	.642	12.0	1.08	.424
German DAX	4388	20.9	4.54	.210	21.2	3.05	.708	15.2	1.73	.559

Table 1: Calculations in this table use 100 times the daily data. Avol approximates the annualized volatility, $\text{Avol} = \sqrt{\frac{252}{T}\sum_{t=1}^{T} x_t}$, where x_t is either squared returns or the realized measure. The sd is the daily standard deviation and acf_1 is the first autocorrelation.

Table 1 shows some summary statistics for the data used. Avol is calculated by taking either the squared returns or the realized measure, multiplying by 252, and then taking the square root of the average over the sample period. This means that Avol is on the scale of annualized volatilty. We see that the annualized volatility lies around 15% to 20% for the squared returns and the realized variance, with Avol between 10% and 15% for the realized semivariance. Furthermore, out of the four indices, the German DAX seems to be the most volatile over the sample period.

Table 1 also shows standard deviations (sd) of percentage daily squared returns or realized measure. The sd figures show much higher standard deviations for squared returns than for the realized variance, which is to be expected.

The acf_1 figures are the first autocorrelations. They show a modest degree of serial correlation in the squared returns, higher serial correlations in the realized semivariance, and higher still in the realized variance. This is a strong indication that models with an autoregressive component will perform well. These results are in accordance with past literature on realized measures.

Also included is a plot of the realized variance for the DAX (Figure 1). There are a few noteworthy events, such as the increased volatility as a result of the internet bubble bursting in the first couple of years of the century. Also the spike following the 9/11

[8] Gerd Heber, Asger Lunde, Neil Shephard and Kevin Sheppard (2009) "Oxford-Man Institute's realized library" version 0.2, Oxford-Man Institute, University of Oxford
`http://realized.oxford-man.ox.ac.uk`

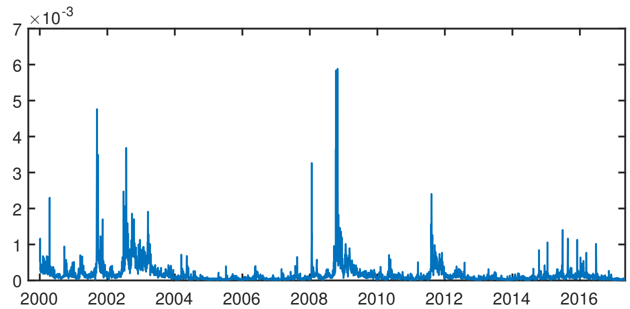

Figure 1: Plot of the realized variance of the German DAX returns.

attacks in 2001, the DAX crash of January 2008 and the credit crisis starting that same year. We will later see that the spike in January 2008 is very impactful in the parameter estimation. Fears of the European sovereign debt crisis spreading to other Mediterranean countries affected the stock market around August 2011. Spikes in more recent years can be attributed to, among other things, fears about China's economy and the United Kingdom's 2016 referendum to determine whether it would be leaving the European Union.

4 Results

4.1 Parameter Estimates

The results of the estimation of the HEAVY models are shown in Table 2, the models are estimated over the full 17-year period. Estimates for the intercepts are omitted since they are very small and almost never significantly different from zero. For the HEAVY-r equation the momentum parameter β lies around 0.7, while for the HEAVY-RM model β_R is somewhat lower with values between 0.5 and 0.6. The HEAVY-RM shows a high degree of persistence with α_R being around 0.4 and $\alpha_R+\beta_R$ very close to one. Finally, the table also shows estimates of α_{IR} for the integrated HEAVY-RM model. These do not differ much from the α_R figures and are estimated to be slightly smaller. These results coincide with the findings from Shephard and Sheppard (2009).

Table 2 also shows the robust standard errors of the estimated coefficients in parentheses. Most standard errors are relatively small and a great proportion of the parameters appear to be significantly different from zero, but there are a couple of striking exceptions. It seems that the HEAYV-r and HEAVY-RM equations do not form an apt specification for the DAX series, as α, α_R and β_R are not significantly different from zero. The other result that stands out is the high standard error for α_{IR} that is estimated for the Nikkei 225, which means that imposing a unit root on this particular series is a misspecification.

Parameter estimates for the HAR models at horizon one are shown in Table 3. Again, results for the intercepts are omitted on account of them being close to zero. The left panel shows results for the HAR-RV model, these figures are in line with results from previous research: a high degree of persistence, with $\phi_d + \phi_w + \phi_m$ close to one. The ϕ_d is highest for the German DAX, meaning that more weight is put on the RV of the

Table 2: Full-sample estimation results for HEAVY models, robust standard errors in parentheses.

Asset	HEAVY-r		HEAVY-RM		Int-HEAVY
	α	β	α_R	β_R	α_{IR}
S&P 500	0.385 (0.042)	0.661 (0.014)	0.441 (0.125)	0.551 (0.166)	0.350 (0.001)
FTSE 100	0.351 (0.039)	0.664 (0.035)	0.443 (0.058)	0.550 (0.054)	0.373 (0.006)
Nikkei 225	0.317 (0.033)	0.715 (0.028)	0.395 (0.025)	0.589 (0.048)	0.328 (0.237)
German DAX	0.302 (0.239)	0.663 (0.275)	0.450 (0.353)	0.531 (0.297)	0.355 (0.002)

previous day and the corresponding error, as a result the plot of the RV of the DAX returns would be more erratic than the other RV plots. Shown in parentheses are robust standard errors.

The right panel shows parameter estimates for HAR-RS, where we split the first lag of RV into RS^+ and RS^- with the intention of disentangling the effects of these two components on realized volatility. The implied effect of lagged RV in this specification is $(\phi_d^+ + \phi_d^-)/2$ and we see this is close to the coefficient found in the first specification, this indicates that models that use only lagged RV are essentially averaging the effects of positive and negative returns. Most importantly, these results show that the effects of positive and negative semivariance on future volatility are vastly different, with the effect of negative semivariance estimated to be 3 to 7 times as large as the effect of positive semivariance.

Table 3: Full-sample estimation results for HAR models, robust standard errors in parentheses. $y_{t+1} = \mu + \phi_d y_t + \phi_d^+ RS_t^+ + \phi_d^- RS_t^- + \phi_w \bar{y}_{w,t} + \phi_m \bar{y}_{m,t} + \epsilon_{t+1}$

Asset	ϕ_d	ϕ_w	ϕ_m	ϕ_d^+	ϕ_d^-	ϕ_w	ϕ_m
S&P 500	0.474 (0.021)	0.344 (0.027)	0.135 (0.023)	0.104 (0.033)	0.776 (0.034)	0.379 (0.025)	0.130 (0.021)
FTSE 100	0.477 (0.021)	0.331 (0.027)	0.148 (0.024)	0.180 (0.040)	0.741 (0.041)	0.338 (0.026)	0.154 (0.024)
Nikkei 225	0.481 (0.026)	0.276 (0.030)	0.139 (0.028)	0.225 (0.048)	0.662 (0.041)	0.307 (0.030)	0.137 (0.028)
German DAX	0.557 (0.019)	0.253 (0.022)	0.147 (0.020)	0.266 (0.039)	0.764 (0.032)	0.279 (0.022)	0.154 (0.020)

4.2 In-sample Evaluation

As explained in subsection 2.4, we compare the models by comparing their predictive accuracy, Table 4 shows t-statistics for the LR tests comparing the QLIK loss of HEAVY models to that of HAR models. Results are shown for horizons 1, 5, 10 and 22 (1-day, 1-week, 2-weeks and 1-month ahead respectively). Recall that negative values favor the HEAVY models.

The iterated multi-step ahead forecasts from the HEAVY models are made using parameters that are tuned to perform best at one-step ahead forecasting. It is therefore no surprise that both HAR models outperform both HEAVY models at longer lags, with the exception of the standard HEAVY model outperforming both HAR-RV and HAR-RS for the Nikkei 225. Something else that may contribute to the HAR models performance is the presence of the long-term volatility component in the HAR models, this term gives the HAR models long-memory which the HEAVY models lack. HEAVY models do present some degree of memory, but not for one month. We expect this to change when we use horizon tuned parameters for the HEAVY models, those results are shown in Table 5.

Table 4: t-statistics for LR tests for iterative forecasts at different horizons. Negative values favor the HEAVY model over the HAR model.
* = forecasting performance is significantly different from zero at a 95% level.

	HEAVY vs. HAR-RV				Int-HEAVY vs. HAR-RV			
Asset	1	5	10	22	1	5	10	22
S&P 500	0.94	-0.68	1.32	1.17	0.97	-0.21	2.97*	2.90*
FTSE 100	1.39	1.58	2.21*	2.14*	1.36	1.83	2.50*	2.18*
Nikkei 225	0.68	-0.31	-0.85	-0.33	0.70	0.62	0.22	1.29
German DAX	1.71	2.49*	3.17*	2.78*	1.68	2.43*	2.03*	1.64

	HEAVY vs. HAR-RS				Int-HEAVY vs. HAR-RS			
Asset	1	5	10	22	1	5	10	22
S&P 500	1.39	-0.37	1.56	1.21	1.42	0.14	3.07*	2.94*
FTSE 100	1.49	2.39*	2.65*	2.38*	1.47	2.38*	2.72*	2.28*
Nikkei 225	0.75	-0.19	-0.82	-0.32	0.77	0.73	0.27	1.30
German DAX	2.20*	2.64*	3.64*	3.17*	2.16*	2.63*	2.21*	1.73

We indeed see that the comparative performance of the HEAVY models has increased. Although the integrated HEAVY model still gets outperformed at almost every horizon and all four indices, the standard HEAVY model beats the HAR models at the 22 horizon for two of the indices. When the HEAVY model does get beaten, it is by a considerably smaller margin and the difference in performance is rarely significant. Overall it appears that the HEAVY forecasts do indeed benefit from tuning the parameters to different horizons.

4.3 Out-of-sample Evaluation

The out-of-sample evaluation was conducted in a more realistic setting. The models were estimated using a moving window with a width of 4 years (1008 observations) and parameters were updated daily. After that forecasts were made for horizons 1 through 22 using horizon tuned parameters. Table 6 shows the results of this exercise, making comparisons between the four pairs of models.

The figures are remarkably different from the previous results. Both the standard HEAVY and the integrated HEAVY models now consistently outperform the HAR models over all horizons and all indices. From this we can conclude that the HEAVY models do not need as much observations as the HAR models to produce accurate forecasts.

Figure 2 shows the forecasts that resulted from the moving window estimation on the data from the DAX. The shown forecasts are from the standard HEAVY model and the

Table 5: t-statistics for LR tests for iterative forecasts at different horizons using horizon tuned parameters for the HEAVY models. Negative values favor the HEAVY model over the HAR model.
* = forecasting performance is significantly different from zero at a 95% level.

	HEAVY vs. HAR-RV				Int-HEAVY vs. HAR-RV			
Asset	1	5	10	22	1	5	10	22
S&P 500	0.94	-0.76	0.79	-1.02	0.97	0.66	2.22*	-0.09
FTSE 100	1.39	1.75	1.09	-0.10	1.36	2.45*	1.47	0.51
Nikkei 225	0.68	-0.42	-1.14	-1.00	0.70	0.72	1.43	0.40
German DAX	1.71	3.03*	0.55	0.02	1.68	2.75*	1.04	1.16

	HEAVY vs. HAR-RS				Int-HEAVY vs. HAR-RS			
Asset	1	5	10	22	1	5	10	22
S&P 500	1.39	-0.36	1.09	-0.96	1.42	1.08	2.39*	-0.05
FTSE 100	1.49	2.60*	1.49	0.20	1.47	2.93*	1.67	0.66
Nikkei 225	0.75	-0.28	-1.11	-0.98	0.77	0.84	1.49	0.43
German DAX	2.20*	3.11*	1.21	0.41	2.16*	2.87*	1.29	1.26

Table 6: t-statistics for LR tests for the out-of-sample performance of iterative forecasts at different horizons. Negative values favor the HEAVY model over the HAR model.
* = forecasting performance is significantly different from zero at a 95% level.

	HEAVY vs. HAR-RV				Int-HEAVY vs. HAR-RV			
Asset	1	5	10	22	1	5	10	22
S&P 500	-1.13	-3.21*	-1.88	-1.67	-1.31	-3.46*	-1.20	-1.50
FTSE 100	-1.17	-4.49*	-2.80*	-2.45*	-1.16	-4.18*	-2.46*	-2.01*
Nikkei 225	-0.88	-1.88	-2.17*	-1.43	-0.76	-1.02	-1.98*	-0.42
German DAX	-0.23	-3.85*	-2.41*	-2.27*	0.01	-3.79*	-2.16*	-2.07*

	HEAVY vs. HAR-RS				Int-HEAVY vs. HAR-RS			
Asset	1	5	10	22	1	5	10	22
S&P 500	0.73	-3.24*	-1.65	-1.70	0.62	-3.49*	-0.80	-1.53
FTSE 100	-1.08	-4.42*	-2.97*	-2.29*	-1.09	-4.13*	-2.62*	-1.90
Nikkei 225	-0.46	-1.89	-2.18*	-1.47	-0.29	-1.03	-2.02*	-0.43
German DAX	0.23	-3.78*	-2.43*	-2.28*	0.39	-3.66*	-1.95	-2.12*

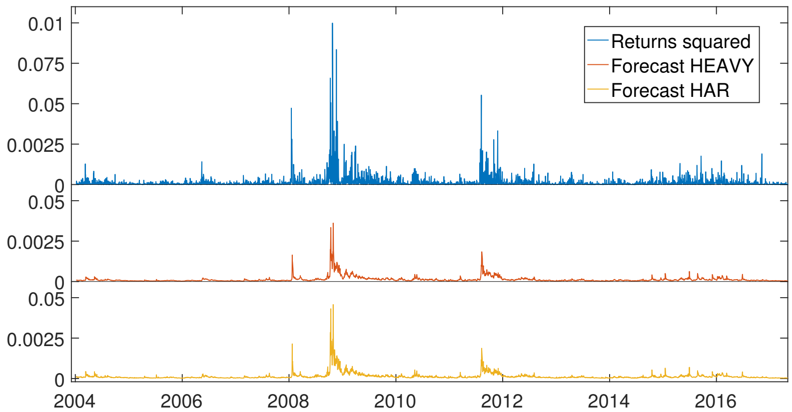

Figure 2: Plotted are the squared returns of the DAX, the conditional variance forecasts made by the HEAVY model and the forecasts from the HAR-RV model.

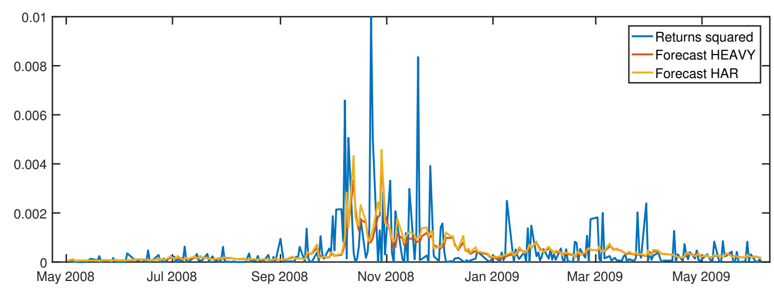

Figure 3: Squared DAX returns and variance forecasts during the credit crunch.

HAR-RV model. The graphs are plotted below each other because the forecasts from the models lie very close together, making it very difficult to distinguish between them when plotted in the same space. Nevertheless, from this picture it becomes apparent that when the squared returns peak, the HAR forecasts are better able to follow, reaching higher values. Both models seem to need quite some time to adjust to higher levels of volatility, this becomes most apparent near the end of our sample in the period of 2015 and 2016. Here we observe short sustained periods of extreme returns, however both models display only modest increases in volatility.

For a better look at the movements of the forecast we have also included a plot of the period around the credit crunch of 2008. From Figure 3 we can see that the models show a delayed reaction to extreme returns and also take some time to return to a lower level of volatility. Again we see how close the forecasts from the models are to each other, however when the forecasts do deviate a little further from each other, it seems to be the HAR model that is at a higher level of volatility.

Figure 4 shows some of the HEAVY coefficients changing through the sample period. The first panel shows results from the HEAVY models, only the α's are plotted. Since

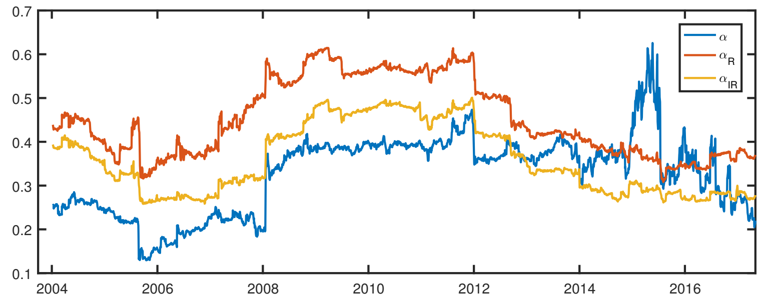

Figure 4: Plot of HEAVY coefficients as estimated using the moving window estimation on the DAX data.

$\alpha+\beta$ and $\alpha_R+\beta_R$ are always close to one, the plots of the β's are mirror images of there α counterparts. The first major thing to note is the increase in the parameters after 2008, the start of the credit crisis. In order to cope with the increased levels of volatility the models estimate these parameters to be higher than in more tranquil periods. Interesting to note is that the increase of α_R is very gradual and starts in 2006, whereas α and α_{IR} show a very abrupt increase in January 2008. Then, four years later, when the spike of January 2008 leaves the estimation sample, we observe sudden drops in the estimated values of all three parameters. Furthermore, the plot of α in the last couple of years seems very volatile, the increases are probably caused by the spikes in RV that we saw in Figure 1. Since these are just occasional spikes and not lasting increases in the level of RV, their effect on estimates of α_R and α_{IR} is limited. The increase in α, however, might be an indication that these spikes do have a longer effect on the level of the volatility of the returns as calculated by the model.

Next, Figure 5 shows the estimated coefficients for the HAR-RV model, the picture is much less volatile than that for the HEAVY parameters, but we again see a large increase at the start of 2008. It seems that the increase in ϕ_d is at the expense of ϕ_w, which means that when we enter a period of higher volatility, the model shifts weight off of the aggregate RV of the past week towards the past day volatility. The graph of ϕ_m looks quite constant over the 13 year period, staying mostly between 0.10 and 0.15. Figure 6 shows the coefficients from the HAR-RS model and looks similar to the previous figure, as the plots of ϕ_w and ϕ_m have not changed much. The shape of the plot of ϕ^- is almost identical to that of ϕ_d, but the values are up to twice as large. ϕ^+ mostly moves in the opposite direction of ϕ^-, but is not exactly a mirror image.

5 Conclusion

In this paper we have discussed and compared two different types of volatility models, both of which make use of realized measures to increase our ability to accurately predict return volatility. With full-sample estimations the HAR models produced better forecasts than the HEAVY models, indicating that in a context of measuring and analyzing long-run volatility, the HAR models might be more effective in determining the long-memory component of volatility. But in a setting more focused on practice the HEAVY models

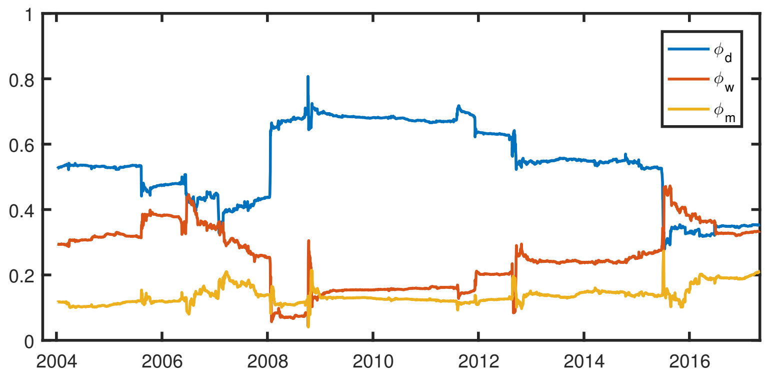

Figure 5: Plot of HAR-RV coefficients as estimated using the moving window estimation on the DAX data.

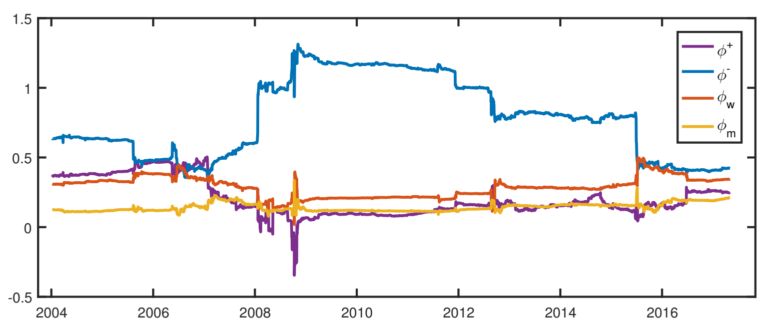

Figure 6: Plot of HAR-RS coefficients as estimated using the moving window estimation on the DAX data.

heavily outperformed both the HAR-RV and the HAR-RS specifications. One might argue that the moving window set-up that was used comes much closer to day-to-day forecasting than the other estimations performed. Therefore, the results in this paper lead us to conclude that, for forecasting purposes, the HEAVY models from Shephard and Sheppard (2009) are superior to the HAR type models.

In addition to being easy to estimate, both models have the virtue of being easily extendable, which makes them both great candidates for future research. One might figure to include the leverage effect into the HEAVY model or make use of signed jump variation as in Patton and Sheppard (2009b). Assessing whether HEAVY is successful in concrete financial applications promises to be an interesting area for further research.

References

Andersen, T. G., Bollerslev, T., Christoffersen, P. F. and Diebold, F. X. (2006). Volatility and correlation forecasting, *in* G. Elliot, C. W. J. Granger and A. Timmermann (eds), *Handbook of economic forecasting*, Vol. 1, Elsevier, 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, *The review of economics and statistics* **89**(4): 701–720.

Andersen, T. G., Bollerslev, T. and Diebold, F. X. (2009). Parametric and nonparametric volatility measurement, *in* Y. Aït-Sahalia and L. P. Hansen (eds), *Handbook of Financial Econometrics*, Amsterdam: North Holland.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2003). Modeling and forecasting realized volatility, *Econometrica* **71**(2): 579–625.

Baillie, R. T., Bollerslev, T. and Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity, *Journal of econometrics* **74**(1): 3–30.

Barndorff-Nielsen, O. E., Kinnebrock, S. and Shephard, N. (2010). Measuring downside risk-realised semivariance, *in* T. Bollerslev, J. Russel and M. Watson (eds), *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, New York: Oxford University Press.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**(3): 307–327.

Bollerslev, T., Engle, R. F. and Nelson, D. B. (1994). Arch models, *in* R. F. Engle and D. McFadden (eds), *Handbook of econometrics*, Vol. 4, Amsterdam: Elsevier North-Holland, pp. 2959–3038.

Bollerslev, T., Litvinova, J. and Tauchen, G. (2006). Leverage and volatility feedback effects in high-frequency data, *Journal of Financial Econometrics* **4**(3): 353–384.

Chen, X. and Ghysels, E. (2011). News—good or bad—and its impact on volatility predictions over multiple horizons, *The Review of Financial Studies* **24**(1): 46–81.

Cipollini, F., Engle, R. F. and Gallo, G. M. (2009). A model for multivariate non-negative valued processes in financial econometrics. Unpublished paper: Stern School of Business, New York University.

Clements, M. and Hendry, D. (1999). Forecasting non-stationary economic time series: the zeuthen lectures on economic forecasting.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility, *Journal of Financial Econometrics* pp. 174–196.

Cox, D. R. (1961a). Prediction by exponentially weighted moving averages and related methods, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 414–422.

Cox, D. R. (1961b). Tests of separate families of hypotheses, *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Vol. 1, pp. 105–123.

Di Matteo, T. (2007). Multi-scaling in finance, *Quantitative finance* **7**(1): 21–36.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy, *Journal of Business & Economic Statistics* **20**(1): 134–144.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation, *Econometrica: Journal of the Econometric Society* pp. 987–1007.

Engle, R. F. (2002). New frontiers for ARCH models, *Journal of Applied Econometrics* **17**(5): 425–446.

Engle, R. F. and Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data, *Journal of Econometrics* **131**(1): 3–27.

Engle, R. F. and Lee, G. G. J. (1999). A permanent and transitory component model of stock return volatility, *in* R. F. Engle and H. White (eds), *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive WJ Granger*, Oxford University Press Oxford, pp. 475–497.

Fleming, J., Kirby, C. and Ostdiek, B. (2003). The economic value of volatility timing using "realized" volatility, *Journal of Financial Economics* **67**(3): 473–509.

Ghysels, E., Valkanov, R. I. and Serrano, A. R. (2009). Multi-period forecasts of volatility: Direct, iterated, and mixed-data approaches. Unpublished paper: Department of Economics, UNC at Chapel Hill.

Glosten, L. R., Jagannathan, R. and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks, *The journal of finance* **48**(5): 1779–1801.

Granger, C. W. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing, *Journal of time series analysis* **1**(1): 15–29.

Hosking, J. R. (1981). Fractional differencing, *Biometrika* **68**(1): 165–176.

Liu, C. and Maheu, J. M. (2005). Modeling and forecasting realized volatility: the role of power variation, *Technical report*, University of Toronto.

Marcellino, M., Stock, J. H. and Watson, M. W. (2006). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series, *Journal of econometrics* **135**(1): 499–526.

Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., Pictet, O. V. and von Weizsäcker, J. E. (1997). Volatilities of different time resolutions—analyzing the dynamics of market components, *Journal of Empirical Finance* **4**(2): 213–239.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies, *Journal of Econometrics* **160**(1): 246–256.

Patton, A. J. and Sheppard, K. (2009a). Evaluating volatility and correlation forecasts, *in* T. G. Andersen, R. A. Davis, J. P. Kreiss and T. Mikosch (eds), *Handbook of financial time series*, Springer Verlag, pp. 801–838.

Patton, A. J. and Sheppard, K. (2009b). Good volatility, bad volatility: Signed jumps and the persistence of volatility, *Review of Economics and Statistics* **97**(3): 683–697.

Rivers, D. and Vuong, Q. (2002). Model selection tests for nonlinear dynamic models, *The Econometrics Journal* **5**(1): 1–39.

Shephard, N. and Sheppard, K. (2009). Realising the future: forecasting with high-frequency-based volatility (heavy) models, *Journal of Applied Econometrics* **25**(2): 197–231.

Visser, M. P. (2008). Forecasting s&p 500 daily volatility using a proxy for downward price pressure, *Technical report*, Korteweg-de Vries Institute for Mathematics, University of Amsterdam.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica: Journal of the Econometric Society* pp. 307–333.