Evaluating Volatility Forecasting Performance at Different Horizons

Master's thesis

Erasmus University Rotterdam

Erasmus School of Economics

Financial Economics

Supervisor: Rogier Quaedvlieg

Second assessor: Mike Mao

Xiaodi Li

422186

7 August 2017

Abstract

In this paper, I investigate the forecasting performance of stock price volatility at different horizons (1, 2, 3, 4, 5, 10, 20, 40 and 60 days). To find which model amongst GARCH, HAR-RV and HEAVY models performs the best with forecast horizon changing, I use returns and realized variances of S&P 500 and make forecasts using MATLAB. Loss functions (MSE and QLIKE) and SPA test are both implemented to measure the performance of different models. The results turns out that under different measurements, HEAVY model is consistently the best when only forecasting a few days ahead. In the long run, GARCH and HEAVY model are equally matched under the evaluation of MSE but GARCH has unbeatable advantage when using QLIKE. Moreover, I make logarithm transformation of RV and compare HAR-log(RV) model with the previous three models. The empirical result shows that when using MSE, this new model has better predicting ability than GARCH at long predicting horizon. However, GARCH is still the best when evaluating by QLIKE. Overall, my suggestion is choosing HEAVY model when forecast the near future and HAR-log(RV) or GARCH model for longer horizon.

Content

Chapter 1. Introduction

Calculating volatility is essential in both portfolio construction and in risk management. In 1952, Markowitz built the investment portfolio theory in which volatility was assumed to be constant over time and measured by historical standard deviation. It is widely used since then such as in CAPM and VAR model. However in recent years, lots of empirical researches find that standard deviation is a time-dependent random variable. Based on this assumption, a wide range of models arise to describe this feature and to predict volatility.

Due to the different modelling approaches of the models, some are able to predict volatility in the long run and some in the short run. An issue arises when forecasters with different information sets and predicting goals are going to choose among the models. For example, the heterogeneity in strategy of low frequency traders and high frequency traders reflects concentration in different trading horizons. So my main question focuses on: what is the best volatility predicting model with the time horizon changing? I will investigate GARCH, HAR-RV and a combined forecast model (HEAVY) first. After that, I will go further into HAR-log(RV) model to see if it can improve the forecast efficiency at all time horizons. To figure out these problems, I will focus on stock price volatility and do empirical research using data of S&P 500.

There is still lack of research on model comparison from multi-step-ahead forecasts performance, and some study results are contradictory. Andersen, Bollerslev, Diebold and Labys (2003) find that complicated high-frequency models are not superior to a simple long-memory Gaussian vector autoregression that uses logarithmic daily realized volatilities as input. In 2005, Koopman, Jungbacker and Hol conclude that compared to models using daily returns, models implementing realized variances produce much more accurate forecast results. Hansen and Lunde (2005) evaluate the one-step-ahead forecast performance of 330 different ARCH-type models and reach the conclusion that GARCH(1,1) is superior to other models in the investigation of exchange rate data, while it is inferior when predicting IBM returns. Shephard and Sheppard (2010) introduce HEAVY model and find that it dominates GARCH model but the advantage becomes weaker as horizon increases. Noureldin, Shephard and Sheppard (2012) also conclude that

HEAVY model outperforms GARCH model when making out-of-sample forecasts, especially at short horizons.

While there is no coherent conclusion about the predicting accuracy, there are a variety of researches on volatility forecasting models. One main category of the forecasting models is called historical volatility models which are constructed based on historical return data. The representative models - such as autoregressive conditional heteroskedasticity (ARCH) model, generalized autoregressive conditional heteroskedasticity (GARCH) model and stochastic volatility (SV) model - usually use long time scale of historical data: daily, weekly and even monthly in general. In contrast, another type of model called realized volatility (RV) model employs intraday high frequency data to measure volatility.

In 1982, Engle provides the ARCH model which is very easy to use but cannot describe the long memory and leverage effect of financial asset return rate series. In order to solve these problems, Bollerslev (1986) advises GARCH model that can better portray the clustering characteristic of volatility and remove high kurtosis effect. However, it still cannot explain leverage effect. Later, some researchers provide several asymmetric GARCH models to overcome the weakness in GARCH model. In 1991, Nelson suggests exponential GARCH (EGARCH) model that allows for asymmetric effects between positive and negative asset returns. Another volatility model widely used to solve leverage effect problem is threshold ARCH (TARCH) model, or so-called GJR model (see e.g., Glosten, Jagannathan and Runkle, 1993). However, Hansen and Lunde (2005) point out that complicated GARCH type models is not superior to GARCH(1,1) model at out-of-sample forecast.

Recent years, instead of using long time scale historical data, intraday trading data provide new methods for researching financial volatility. Andersen and Bollerslev (1998) propose that using conventional squared daily return as daily volatility has large measurement error and noise, while using intraday high-frequency returns to calculated realized volatility (RV model) can solve this problem. They also find that as the frequency of the data increases, the influence of measurement error to the underlying volatility process decreases. However, because of the market microstructure effects in practice, the highest data frequency may not be the best choice.

Based on heterogeneous market hypothesis, Corsi (2009) propose Heterogeneous autoregressive model of Realized Volatility (HAR-RV model). He regards volatility as a combined effect from high, medium and low frequency traders and describe it through three different time scales: daily, weekly and monthly. However, researchers find that though in approximately continuous time, intraday high frequency data may have big fluctuations, which is called jump. Lee and Mykland (2012) point out that jump has essential effect on describing and forecasting volatility. Andersen, Bollerslev and Diebold (2007) put up HAR-RV-J and HAR-RV-CJ models which incorporate jump as explanatory variable.

More recent papers have put forward to use forecast combination models in order to get better predictions than individual models. Shephard and Sheppard (2010) propose HEAVY model which uses high frequency data as predictors in GARCH-type models and do empirical research using stock index and exchange index. They reach the conclusion that HEAVY model is easier to estimate and is more robust than GARCH model when there is structural breaks in volatility. The forecasts are more accurate especially in the first several days.

In this paper, I select GARCH(1,1) from GARCH-type models for the reason that it is the most simple but one of the best predicting model according to Hansen and Lunde (2005). As realized volatility provides more accurate proxies for daily variance using intraday return data, one of the RV models I would like to choose is HAR-RV (Heterogeneous Autoregressions) model which has clear economic interpretation. As a combination of GARCH and realized variance, HEAVY model is applied to figure out whether it can improve the performance of the individual models. Last but not least, based on the comparison result of these three models, I find that HAR-RV tends to be the relatively worst model with large predicting errors in the long run, which I will discuss in Chapter 5. Therefore, I decide to improve HAR-RV by using logarithm of RV and put the HAR-log(RV) model into the forecast assessment.

After the model selection, the performance measurements need to be chosen. The evaluating methods have been improving over time. Hansen and Lunde (2005) advise to use six loss functions such as mean squared error (MSE), mean absolute error (MAE),

heteroskedastic adjusted MSE and MAE (HMSE and HMAE), QLIKE and R2log. However, many papers for example, Lamoureux and Lastrapes (1993), Hamilton and Susmel (1994) and Bollerslev and Ghysels (1994) use some or all of the loss functions but find that the best-performing volatility forecast model is not the same when the choice of loss function changes. Patton (2011) finds that only two of the loss functions (MSE and QLIKE) are robust to noise when used to compare matching volatility prediction models, which means that using a proxy for volatility does not influence the performance ranking as using the true conditional variance. For this reason, I use MSE and QLIKE as Patton advises instead of employ all of them. White (2000) suggests a test called reality check for data snooping (RC), but it is sensitive to the incorporation of poor and irrelevant predictions. Hansen (2005) fixes this problem by studentizing the test statistic and by putting up a null distribution depending on sample. They suggest a bootstrap procedure for SPA test which I will apply to evaluate my models in this paper. After that, Hansen, Lunde and Nason (2011) provide the model confidence set that constructs a set of models including the best model within a certain confidence interval.

My results shows that there is no dominating model at all horizons. The conclusion of SPA test for MSE and QLIKE is always consistent with using only loss functions. Despite of the fact that different loss functions may lead to diverse results, HEAVY model remains to be the best amongst all models when forecasting one- and two-day-ahead volatility, which conforms to the result of Shephard and Sheppard (2010). When losses are measured by MSE, GARCH and HEAVY by turns has the lowest errors when forecasting three days ahead and forward. This outcome changes when using QLIKE. From three to five days, HAR-RV beats the other models while GARCH is the most accurate model after five days. After putting HAR-log(RV) into the comparison, the loss of this optimized model is the lowest when forecast horizon is larger than three. However, this new adding model cannot distinct itself from the original three models under the measurement of QLIKE and therefore, the best models do not have any differences from before.

This article contributes to the existing works in three different ways. First, although many researches have compared different volatility prediction models, there is still lack of findings on horizon effects of those models. Second, as combined forecast is more

appealing in real world, putting joint forecasting models into analysis researches more innovative models that could be widely put into use in the future. Third, the use of latest data enables us to see the differences of model predicting ability between past and now.

The paper is organized as follows. I begin with an introduction of the topic including the motivation of investigating it, literature review on the models and research outcomes on their forecasting ability. In Chapter 2, I will describe the mathematical calculation of the volatility and derive the formulations of h-step-ahead forecasts of models in details. Chapter 3 defines two widely used performance evaluating methods. Data description and modelling methods are introduced in Chapter 4. After that, the model correction and calculation results of the empirical research which shows the best model over different horizons are provided in Chapter 5. I will conclude in Chapter 6 and finally, discuss the possible future improvements in Chapter 7.

Chapter 2. Theoretical Background

In this chapter, I will introduce how to calculate variance using daily asset returns and why in practice, we usually use the sum of squared daily returns as daily variance. Then, I will step into the detailed volatility models that are used in my empirical research and show the h-step-ahead forecast process, which give the theoretical background of analysis in the following chapters.

2.1 Concepts of variance and calculation

Consider the discrete series $\{p_t\}_{t=1}^{T}$, where p_t denotes the logarithm of asset price at day t. Express the return rate as the compounded daily return, so

$$r_t = p_t - p_{t-1}. \tag{2.1.1}$$

Then the daily variance is usually calculated as

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(r_t - \bar{r})^2, \tag{2.1.2}$$

Where $\bar{r}$ is the average of daily returns. When the data frequency is low, the commonly used way is to calculate the variance as above. But if high frequency data is accessible, *realized measure of variance* is better to incorporate all the useful information.

Let the return process be defined as standard Ito process, which is

$$\mathrm{d}p_t = \mu_t dt + \sigma_t dW_t, \tag{2.1.3}$$

where μ_t represents for the drift, σ_t is the spot volatility and W_t is a standard Wiener process. $\mu_t dt$ stands for the stable part of return rate and $\sigma_t dW_t$ is the uncertain part. When μ_t and σ_t are jointly independent of W_t,

$$r_t|\mu_t, IV_t \sim N(\mu_t, IV_t), \tag{2.1.4}$$

where $\mu_t = \int_{t-1}^{t} \mu(s)ds$ and $IV_t \equiv \int_{t-1}^{t} \sigma^2(s)ds$. IV_t is referred to as *integrated variance* and is used to measure accurate realized variance. However, the accurate $\sigma^2(s)$ is hard to get.

Given that $\mathrm{t} - 1 = \tau_0 < \tau_1 < \cdots < \tau_{N_t} = t$, the intraday returns are written as

$$r_{t,i} = p_{\tau_i} - p_{\tau_{i-1}}, \text{for i} = 1, \dots, N_t \tag{2.1.5}$$

To approximate the volatility, I write the quadratic variance QV of a stochastic process over the interval [t-1, t] as

$$QV_t \equiv \operatorname*{plim}_{N_t \to \infty} \sum_{i=1}^{N_t} (p_{\tau_i} - p_{\tau_{i-1}})^2. \tag{2.1.6}$$

When $N \to \infty$, $\max\limits_{1 \le i \le N_t} |\tau_i - \tau_{i-1}| \to 0$.

The empirical part of QV_t is regarded as *realized variance* which is the sum of the instantaneous squared returns in a continuous time of one day. Therefore, RV is defined as

$$RV_t = \sum_{i=1}^{N_t} {r_{t,i}}^2. \tag{2.1.7}$$

2.2 GARCH(1,1) model

Engle (1982) and Bollerslev (1986) put forward the generalized ARCH (GARCH) model. Given a log return series r_t, let innovation at time t be $z_t = r_t - \mu_t$. Then z_t follows a GARCH(m,s) model if

$$z_t = \sigma_t \epsilon_t,\ \sigma_t^2 = \omega + \sum_{i=1}^{m} \alpha_i z_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2, \tag{2.2.1}$$

where ϵ_t is often assumed to be a standard normal or standardized Student-*t* distribution or generalized error distribution with mean 0 and variance 1. The constant $\omega > 0$ and the constraints for parameters are $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_i) < 1$. When $i > m$ or $j > s$, $\alpha_i = 0$ or $\beta_j = 0$ respectively, which turns (2.2.1) to simpler forms. The latter constraint on $\alpha_i + \beta_j$ implies that the unconditional variance of z_t is not infinite, while the conditional variance σ_t^2 changes over time.

When m=1, s=1, this model turns into GARCH(1,1) model:

$$\sigma_t^2 = \omega + \alpha z_t^2 + \beta \sigma_t^2,\ with\ \alpha \geq\ 0, \beta \leq 1, \alpha + \beta < 1. \tag{2.2.2}$$

To make h-step-ahead forecast easier, I start from h=1:

$$\sigma_{t+1}^2 = \omega + \alpha z_t^2 + \beta \sigma_t^2, \tag{2.2.3}$$

where z_t^2 and σ_t^2 are known at time t. So

$$\sigma_t^2(1) = \omega + \alpha z_t^2 + \beta \sigma_t^2. \tag{2.2.4}$$

For multiple forward prediction, use the function $z_t^2 = \sigma_t^2 \epsilon_t^2$ and plug it into equation (2.2.2). We get

$$\sigma_t^2 = \omega + (\alpha + \beta)\sigma_t^2 + \alpha z_t^2 (\epsilon_t^2 - 1). \tag{2.2.5}$$

When h=2, the equation above becomes

$$\sigma_{t+2}^2 = \omega + (\alpha + \beta)\sigma_{t+1}^2 + \alpha z_{t+1}^2 (\epsilon_{t+1}^2 - 1). \tag{2.2.6}$$

Because $\mathrm{E}(\epsilon_{t+1}^2 - 1 | F_t) = 0$, the 2-step-ahead volatility forecast at forecast origin t is computed by

$$\sigma_t^2(2) = \omega + (\alpha + \beta)\sigma_t^2(1). \tag{2.2.7}$$

Therefore, we can derive the general formula of h-step-ahead forecast, which is

$$\sigma_t^2(h) = \omega + (\alpha + \beta)\sigma_t^2(h-1), h > 1. \quad (2.2.8)$$

2.3 HAR-RV model

To mimic the actions of various types of market participators, Corsi (2004) suggests the Heterogeneous Autoregressive model for Realized Volatility (HAR-RV), which considers volatilities over different time horizons: daily, weekly and monthly. The model is represented as the following equation. At day t, the one-step-ahead daily forecast is determined by RVs of today, past week (5 days) and past month (22 days).

$$RV_{t+1}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \omega_{t+1} \quad (2.3.1)$$

The calculation of $RV_t^{(w)}$ and $RV_t^{(m)}$ is as following:

$$RV_t^{(w)} = \frac{1}{5}(RV_{t-1}^{(d)} + RV_{t-2}^{(d)} + RV_{t-3}^{(d)} + RV_{t-4}^{(d)} + RV_{t-5}^{(d)}) \quad (2.3.2)$$

$$RV_t^{(m)} = \frac{1}{22}(RV_{t-1}^{(d)} + RV_{t-2}^{(d)} + \cdots + RV_{t-22}^{(d)}) \quad (2.3.3)$$

It is simple to deduct the formula for h-step-ahead forecast from the above equations:

$$RV_{t+h}^{(d)} = c + \beta^{(d)} RV_{t+h-1}^{(d)} + \beta^{(w)} RV_{t+h-1}^{(w)} + \beta^{(m)} RV_{t+h-1}^{(m)} + \omega_{t+h}, \quad (2.3.4)$$

2.4 HEAVY model

HEAVY model (High frEquency bAsed VolatilitY model) proposed by Shephard and Shephard (2010) is a combination of GARCH model and realized measures and is simple to estimate. They prove that using realized variance can give more accuracy both in in- and out-of-sample forecasting.

As before, $r_1, r_2, \ldots, r_t$ is the series of log return. If F_{t-1}^{LF} represents the low frequency past data, then the GARCH model as explained in equation (2.2.2) can be written as:

$$\mathrm{Var}(r_t | F_{t-1}^{LF}) = \sigma_t^2 = \omega_G + \alpha_G r_{t-1}^2 + \beta_G \sigma_{t-1}^2. \quad (2.4.1)$$

When the realized variance is used,

$$\mathrm{Var}(r_t|F_{t-1}^{HF}) = h_t = \omega + \alpha RV_{t-1} + \beta h_{t-1},\ \omega, \alpha \geq 0, \beta \in [0,1], \quad (2.4.2)$$

Where F_{t-1}^{HF} represents high frequency past data.

Assume that RV satisfies an AR (1) model which is:

$$RV_t = \omega_R + \alpha_R RV_{t-1} + \varepsilon_R. \quad (2.4.3)$$

The above equation I use in this paper that explains the development of the realized measures is different from the conditional realized variance calculated in Shephard and Sheppard (2010) for simplicity reason. The original definition is as following:

$$E(RV_t|F_{t-1}^{HF}) = \mu_t = \omega_R + \alpha_R RV_{t-1} + \beta_R \mu_{t-1},\ \omega_R, \alpha_R, \beta_R \geq 0, \alpha_R + \beta_R \in [0,1]. \quad (2.4.4)$$

The forecast when horizon equals to h can be derived from 1-step-ahead forecast formula:

$$\mathrm{Var}(r_{t+h}|F_{t-1}^{HF}) = h_{t+h} = \omega + \alpha RV_{t+h-1} + \beta h_{t+h-1},\ \omega, \alpha \geq 0, \beta \in [0,1], \quad (2.4.5)$$

$$RV_{t+h} = \omega_R + \alpha_R RV_{t+h-1} + \varepsilon_R. \quad (2.4.6)$$

Unlike other models, HEAVY is constructed by two equations where RV is also affected by past obervations.

2.5 Extension: HAR-log(RV) model

HAR-log(RV) model defined by Corsi and Reno (2009) is written as below:

$$\log\left(RV_{t+1}^{(d)}\right) = c + \beta^{(d)} \log\left(RV_t^{(d)}\right) + \beta^{(w)} \log\left(RV_t^{(w)}\right) + \beta^{(m)} \log\left(RV_t^{(m)}\right) + \omega_{t+1}, \quad (2.5.1)$$

where

$$\log(RV_t^{(w)}) = \frac{1}{5}(\log\left(RV_{t-1}^{(d)}\right) + \log\left(RV_{t-2}^{(d)}\right) + \cdots + \log\left(RV_{t-5}^{(d)}\right)) \quad (2.5.2)$$

and

$$\log(RV_t^{(m)}) = \frac{1}{22}(\log(RV_{t-1}^{(d)}) + \log(RV_{t-2}^{(d)}) + \cdots + \log(RV_{t-22}^{(d)})) \quad (2.5.3)$$

When $\omega_t \sim N(0, \sigma_\omega^2)$, $\exp(\omega_t) \sim \log N(0, \sigma_\omega^2)$. Log N is the log-normal distribution. Then the conditional realized variance is calculated as

$$RV_{t+1|t} = \exp\left(\log\left(RV_{t+1}^{(d)}\right) - \widehat{\omega} + \frac{1}{2}\hat{\sigma}_\omega^2\right), \tag{2.5.4}$$

Where $\widehat{\omega}$ is mean of the estimated value of the residuals and $\hat{\sigma}_\omega^2$ is the estimated variance.

The equation for h-step-forward forecast is the same with formula (2.3.4) except that all the RVs need to be taken logarithm.

Chapter 3. Comparison Criteria

It is difficult to evaluate the performance of volatility models for the reason that conditional variance and integrated variance are latent. As I have described in Chapter 2.1, a widely used solution is to substitute squared return as a proxy of the real variance. This is also the assumption for the comparison criteria which compare the predicted variance to the proxy.

3.1 Loss functions

Patton (2011) suggests a new family of loss functions that are robust to the noise of volatility proxy and consistent to the selection of units of measurement. This family nets MSE and QLIKE which are defined as below:

$$\mathrm{MSE}_{j,h} = \frac{1}{N}\sum_{j=1}^{N}(RV_{j,h} - \hat{\sigma}_{j,h}^2)^2 \tag{3.1.1}$$

and

$$\mathrm{QLIKE}_{j,h} = \frac{1}{N}\sum_{j=1}^{N}\left(\ln\left(\hat{\sigma}_{j,h}^2\right) + \frac{RV_{j,h}}{\hat{\sigma}_{j,h}^2}\right), \tag{3.1.2}$$

where N is the number of rolls in the modelling process that will be defined in Chapter 4.2 and $\hat{\sigma}_{j,h}^2$ is the h-step-ahead variance forecasted at the j^{th} roll. So the loss at each horizon can be computed.

From the definition, MSE measures the average of the squared error between actual realized variance and variance predictions. It is always non-negative and the closer to zero, the better. The metric QLIKE is the loss implied by a Gaussian likelihood.

3.2 SPA test

Hansen (2005) recommends a superior predictive ability (SPA) test based on bootstrap simulation to increase the robustness of comparison. The main idea of SPA test is to figure out whether other models are better than the benchmark model in term of expected losses. So he designs the test of the *null hypothesis* H_0 *that the benchmark is not inferior to any of the alternatives.*

The procedure of SPA test is explained as below.

To begin with, assume that there are K+1 categories of volatility forecast model, denoted as $M_k,\ k = 0,1 \dots K$. The h-step-ahead variance forecasts of model M_k is ${\sigma_{h,k}}^2$, where h= 1, 2,…, N. N is the total number of forecasts. Then for every prediction, the corresponding error values from the above two loss functions are calculated, which are set to be $L_{h,k,i}$, where i=1,2.

Secondly, choose one prediction model M_0 as the benchmark model, the expected loss of which is $L_{h,0,i}$. Then the relative performance between M_0 and M_{k} is defined as:

$$X_{h,k} = L_{h,0,i} - L_{h,k,i}. \tag{3.2.1}$$

The null hypothesis can be written as

$$\max_k \lambda_k = E(X_{h,k}) \leq 0. \tag{3.2.2}$$

If and only if $E(X_{h,k}) > 0$, model k is better than the base model. The statistics for the test of the hypothesis is calculated as

$$\mathrm{T} = \max \frac{\sqrt{N}\overline{X_k}}{\omega_{kk}}, k = 1,2, \dots, K, \tag{3.2.3}$$

Where

$$\overline{X_k} = N^{-1} \sum_{h=1}^{N} X_{h,k}\,, \omega_{kk} = var\left(\sqrt{N}\overline{X_k}\right). \tag{3.2.4}$$

In order to get the distribution and the p-value of the T statistics, Hansen (2005) uses the bootstrap procedure that I explain in four steps as below.

The first step is getting a new sample from $X_{h,k}$ with a length of N. To achieve this, a re-sampling process needs to be used to get a subsample from $\{X_{h,k}\}$. Choose a random integer from 1 to N, and then create a number M that follows a geometric distribution with a mean of q. q is 0.5 in this paper. Take out $\{X_{M,k}, X_{M+1,k}, \dots, X_{M+q-1,k} | k = 1,2, \dots, K\}$. When a number M+x is larger than N, it is rearranged to be the modulus by dividing N. Repeat this sampling process until the full length of the subsample is N for any K.

Reiterate the first step for B times and get a new sample which is denoted as $\{Y_{n,k}^{i}, \mathrm{i} = 1,2,\dots,\mathrm{B},\ \mathrm{n} = 1,2,\dots,\mathrm{N}, \mathrm{k} = 1,2,\dots,\mathrm{K}\}$. Compute the average value for each of the bootstrap sample:

$$\overline{Y_k^i} = N^{-1} \sum_{n=1}^{N} Y_{n,k}^{i}, i = 1,2,\dots,B, k = 1,2,\dots,K. \tag{3.2.5}$$

Calculate the variance of all B samples:

$$\widehat{\omega_{kk}} = B^{-1} \sum_{i=1}^{B} (\overline{Y_k^i} - \overline{\overline{Y}_k})^2, \overline{\overline{Y}_k} = B^{-1} \sum_{i=1}^{B} \overline{Y_k^i}, k = 1,2,\dots,K. \tag{3.2.6}$$

For the third step, define Z_k^i as:

$$\overline{Z_k^i} = \left(\overline{Y_k^i} - \overline{\overline{Y}_k}\right) \times I\{Y_k > -A_k\}, i = 1,2,\dots,B, k = 1,2,\dots,K, \tag{3.2.7}$$

Where $A_k = \frac{1}{4} N^{-4} \omega_{kk}$. $\mathrm{I}\{\cdot\}$ is an indicator function that means when the condition in the bracket is satisfied, the value of the function is 1, otherwise it is 0.

The last step is to get the empirical statistics:

$$\mathrm{T}^i = \max \frac{\sqrt{N}\overline{Z_k^i}}{\widehat{\omega_{kk}}}, \mathrm{i} = 1,2,\dots,\mathrm{B}. \tag{3.2.8}$$

Hansen shows that when null hypothesis cannot be rejected, the above statistics converges to the T statistics in equation (3.2.8). Then the p-value is:

$$\mathrm{p} = B^{-1} \sum_{i=1}^{B} I\{\mathrm{T}^i > T\}. \tag{3.2.9}$$

The p-value can be computed for every forecast model given a loss function. When it is close to 1, the null hypothesis cannot be rejected, which means the benchmark model is superior to other models.

Chapter 4. Data and Methodology

4.1 Data

The S&P 500 Index is widely regarded as the leading benchmark of the overall U.S. stock market. It is a capitalization weighted index of 500 leading companies. And because of its high market efficiency, the index provides a solid foundation for model calculation. Andersen, Bollerslev, Diebold and Labys (2001) advice to use 5-minute returns for the realized variance to avoid market microstructure and 'model free' problems. Patton (2011) also points out that 5-minute returns eliminate all the distortions. I follow their researches and use historical daily returns and 5-minute realized variances of S&P 500 from January 1, 2000 to March 31, 2017. Both return and intraday variance data are downloaded from Realized Library of Oxford-Man Institute of Quantitative Finance. I delete all the non-trading days and multiply the returns by 100 and realized variances by 10000. As the model ranking only cares about relative size of the numbers, multiplying a same factor does not affect the results but makes the numbers easier to read.

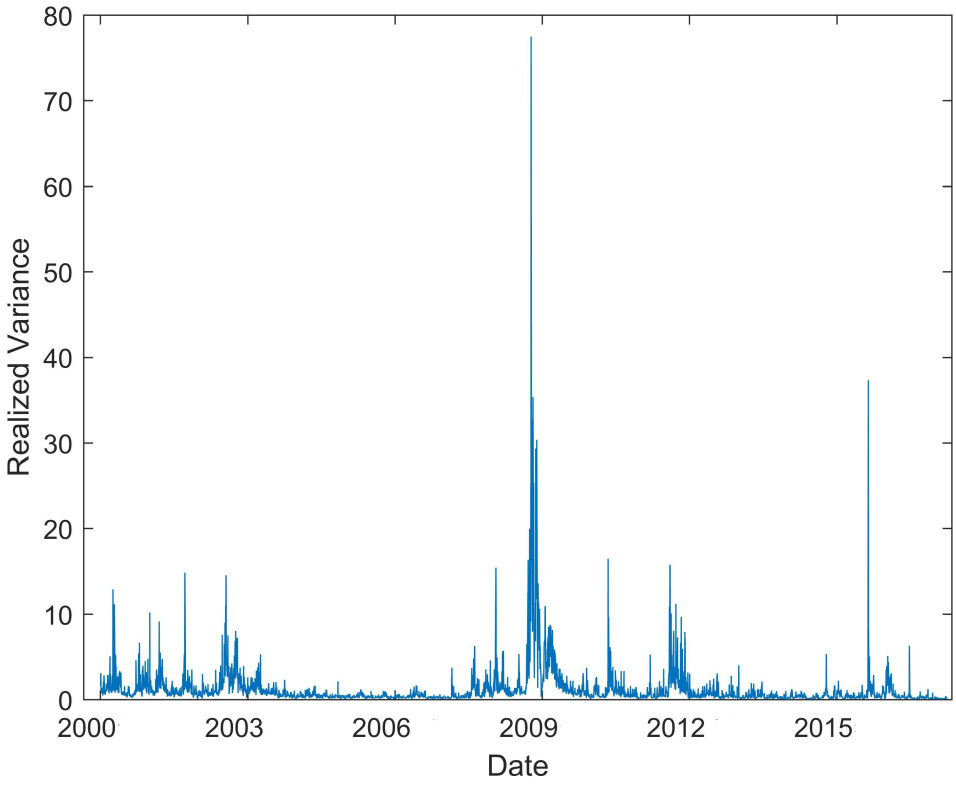

Figure 4.1.1 *Realized variances over time.* The figure plots realized variances from January 1, 2000 to March 31, 2017. All values are original realized variances time 10000.

The sample contains both relatively calm and turbulent time periods. From Figure 4.1.1, during 2000-2003, the realized variances were high because of the collapse of dot-com bubble. Stock prices fell dramatically after the speculative bubble and S&P lost around half of its value. Because RV is defined as the sum of squared returns, both consistently positive and negative returns would lead to high RV. This explains why during 2000 to 2003 there was high volatility. The calm period from 2004 to early 2007 was followed by the 2008 financial crisis. During this time periods, 56% of the value of S&P 500 vanished. In August 2011, stock market fell and S&P 500 faced a huge drop in its price again which leads to the explosive growth of RV. We can imagine that this unexpected growth may cause large forecast errors in the later modelling. After that, RV was relatively low until now.

4.2 Methodology

In this paper, I make forecasts of 1, 2, 3, 4, 5, 10, 20, 40 and 60 days forward of S&P 500 index volatility in a rolling window of 1000 days using MATLAB. In total, I have 3252 rolls. Some programming functions come from MFE Toolbox by Kevin Sheppard.

The explanations of the forecast procedures for different models are as following.

In GARCH(1,1) model, I use 1 to 1000 daily returns to estimate the parameters ω, α and β in the model. Then I predict h-step-ahead (h=1, 2… 60) variance using equation (2.2.8). After that, the parameters are re-estimated using daily returns from 2 to 1001 days and 60 new forecasts are calculated again.

The forecasting process of HAR-RV model is similar except that when $2 \leq h \leq 60$, the weekly RV and monthly RV are re-calculated using previous daily forecast of RV to get new predictions. For example, after getting RV_{t+1}, RV_{t+2} is defined as following:

$$RV_{t+2}^{(d)} = c + \beta^{(d)} RV_{t+1}^{(d)} + \beta^{(w)} RV_{t+1}^{(w)} + \beta^{(m)} RV_{t+1}^{(m)} + \omega_{t+1}.$$

$RV_{t+1}^{(w)}$ and $RV_{t+1}^{(m)}$ are computed from the realized variances of the past 5 days and 22 days respectively. Therefore, the newly prediction RV_{t+1} should be used for their calculation.

As to HEAVY model, it consists out of two parts (2.4.5) and (2.4.6). The difficulty is that all the parameters in these two parts should be estimated. Compared to GARCH(1,1), conditional realized variance for each step ($2 \leq h \leq 60$) need to be calculated and then put into equation (2.4.5) to get future variances.

Chapter 5. Empirical Analysis

In this chapter, GARCH(1,1), HAR-RV and HEAVY model are compared under the measurement of both loss functions and SPA test. Realized variance is used as benchmark to evaluate the forecasting performance. Erratic result comes out from MSE analysis. Therefore, "insanity filter" is used in subsequent content to amend the unrealistic large forecasts. In the end, I add HAR-log(RV) model to figure out how it performs compared to the original three models.

5.1 Model correction

By computing the corresponding MSE values at the horizons I want to investigate, I plot lines in Figure 5.1.1 with horizon as X-axis and MSE value as Y-axis. Notice that instead of plotting discrete dots, I use all 60 horizons to form smooth lines that can clearly show the tendency.

From the figure below, HAR-RV model has the largest predicting error when forecast horizon is larger than 35. The errors have exponentially tendency as horizon increases and even reach 1e+6, which is obviously weird. Moreover, because of the explosive numbers of HAR-RV, it is hard to see how GARCH and HEAVY model performs. Thus, it is vital to figure out why the crazy forecasts come out and how they should be corrected.

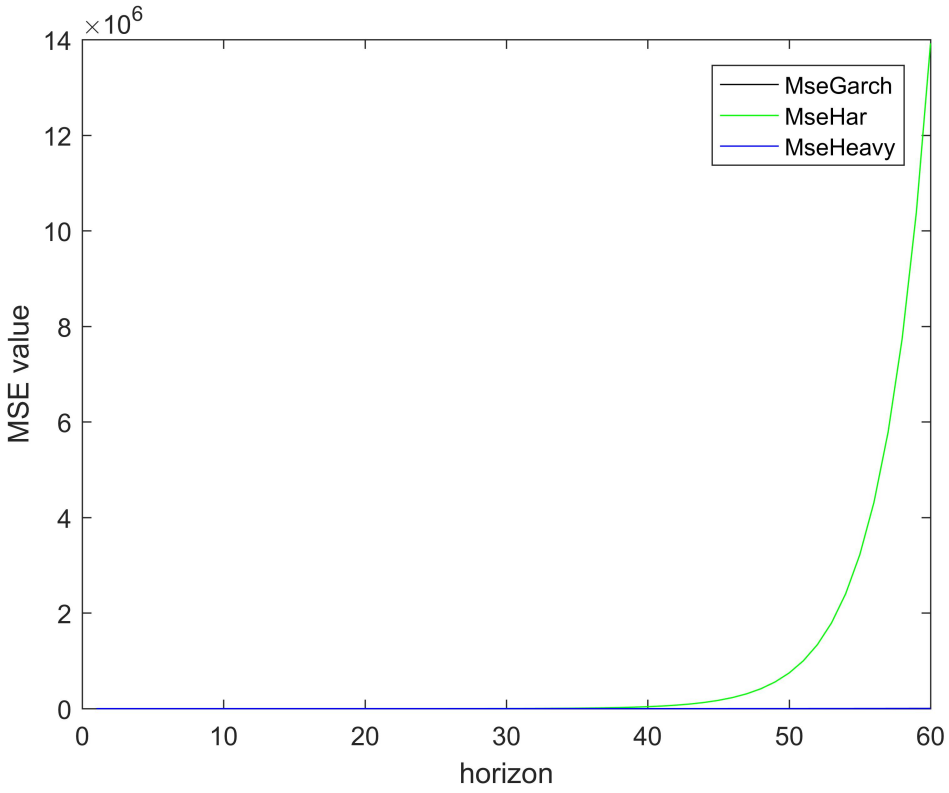

Figure 5.1.1 *Comparison using MSE.* The figure illustrates the MSE value changing with horizon. Three models (GARCH, HAR-RV and HEAVY model) are measured. Horizon is from 1 to 60.

Remember that HAR-RV model is defined as below:

$$RV_{t+1}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \omega_{t+1}, \qquad (5.1.1)$$

where the parameters $\beta^{(d)}, \beta^{(w)}$ and $\beta^{(m)}$ do not automatically meet restrictions like those in GARCH model, for example the sum of parameters should be smaller than 1. This means that without limitations, the parameters may be too large and the forecasts may deviate severely from unconditional variances when unexpected large original data appears, for example, during the 2008 financial crisis.

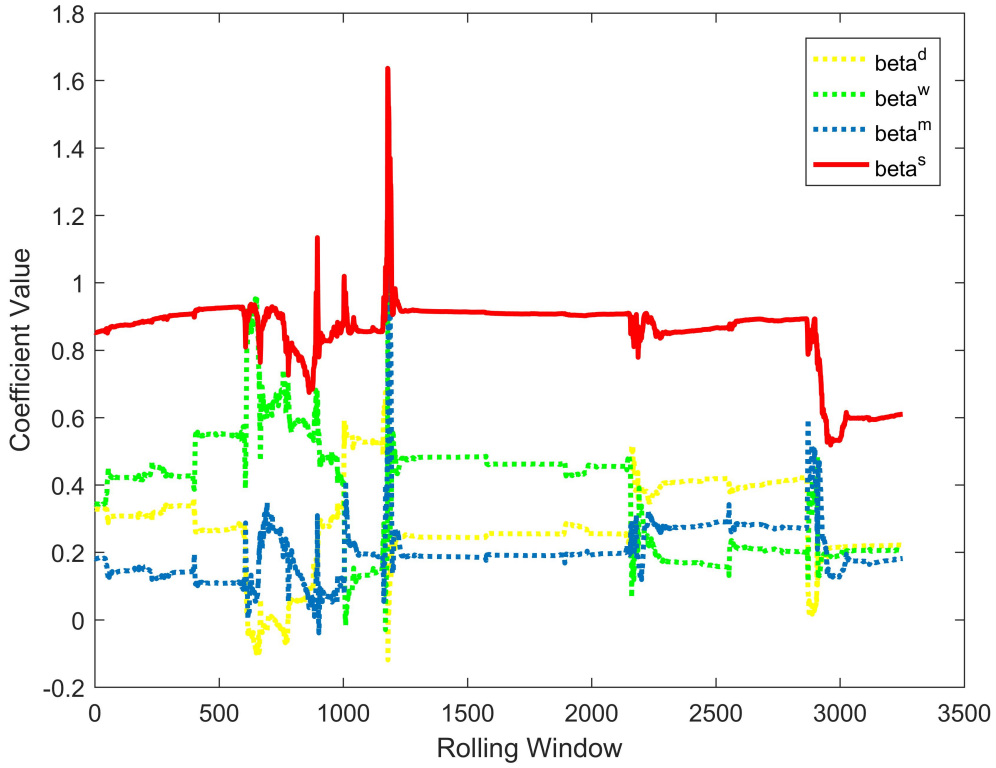

Figure 5.1.2 *HAR-RV estimated coeffecients.* This figure shows the values of $\beta^{(d)}, \beta^{(w)}, \beta^{(m)}$ and $\beta^{(s)}$ at each roll which are calculated by 1000 in-sample data. $\beta^{(s)} = \beta^{(d)} + \beta^{(w)} + \beta^{(m)}$. There are in total 3252 rolls.

According to the above analysis, in Figure 5.1.2 I show the values of $\beta^{(d)}, \beta^{(w)}, \beta^{(m)}$ and the sum of them, $\beta^{(s)}$, at each roll. We can see that around the 1200th roll, the sum of the parameters are far above 1, which results in the extreme forecasts during those days. Except from this sudden increase, all other estimated parameters are relatively steady.

To prove that only a few explosive numbers lead to the poor performance of HAR-RV when forecast horizon is long, in Figure 5.1.3, I show the percentage of forecasts that lies in different ranges. When the horizon increases (the vertical axis from bottom to up), the number of forecasts that are between 10 and 100 decreases while the number of forecasts larger than 100 increases. But even when the horizon equals to 60, the extreme forecasts (>100) are still less than 0.5% percent. Hence, if this small part of unreasonal forecasts can be corrected, the performance of HAR-RV could be improved a lot.

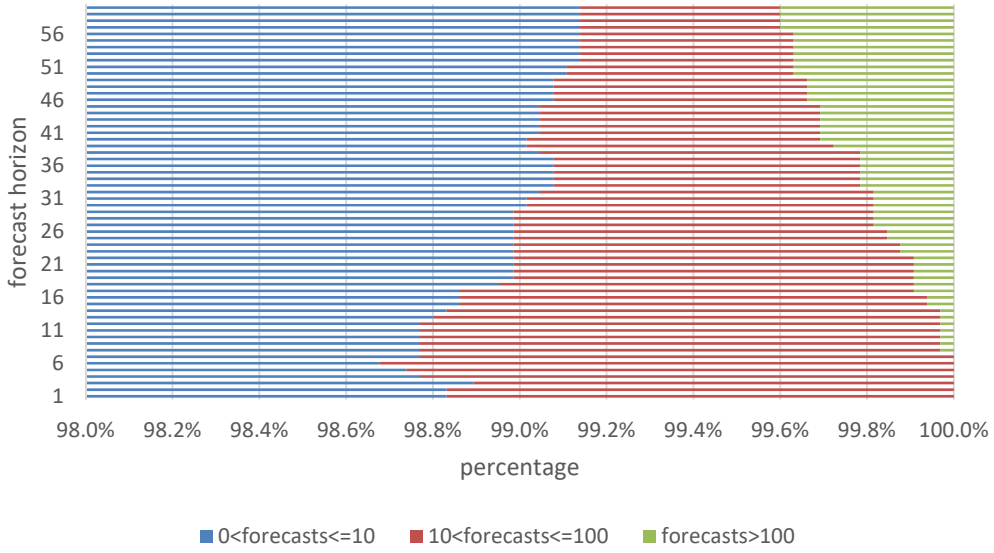

Figure 5.1.3 *Percentage of HAR-RV forecasts.* This figure describes the percentage of forecasts that lies in different ranges. The horizontal axis is percentage that starts from 98% to 100%. The vertical axis is the horizon from 1 to 60. For each horizon, the colors show the proportions of different forecasts magnitude.

In order to avoid crazy numbers, I apply "insanity filter" suggested by Swanson and White (1995,1997) to substitute the unrealistic forecasts with ones that are more conformative with observed data: "ignorance" is better than "insanity". The filter described in Clements and Hendry (2011) works as below.

Calculate the forecast difference between RV_{T+h} and RV_T, where RV_T is the most recent known volatiliy. Then compute $RV_t - RV_{t-h}$ in the estimation period. As there are 1000 data in the in-sample period, there will be 999 differences when h=1, 998 when h=2... Find the minimum and maximum value for each h. If the forecast difference is not between the minimum and maximum value, RV_{T+h} will be replaced with the last observation RV_T.

To make the comparision consistent, all models are filtered in the rest of the article.

5.2 Results from model comparison

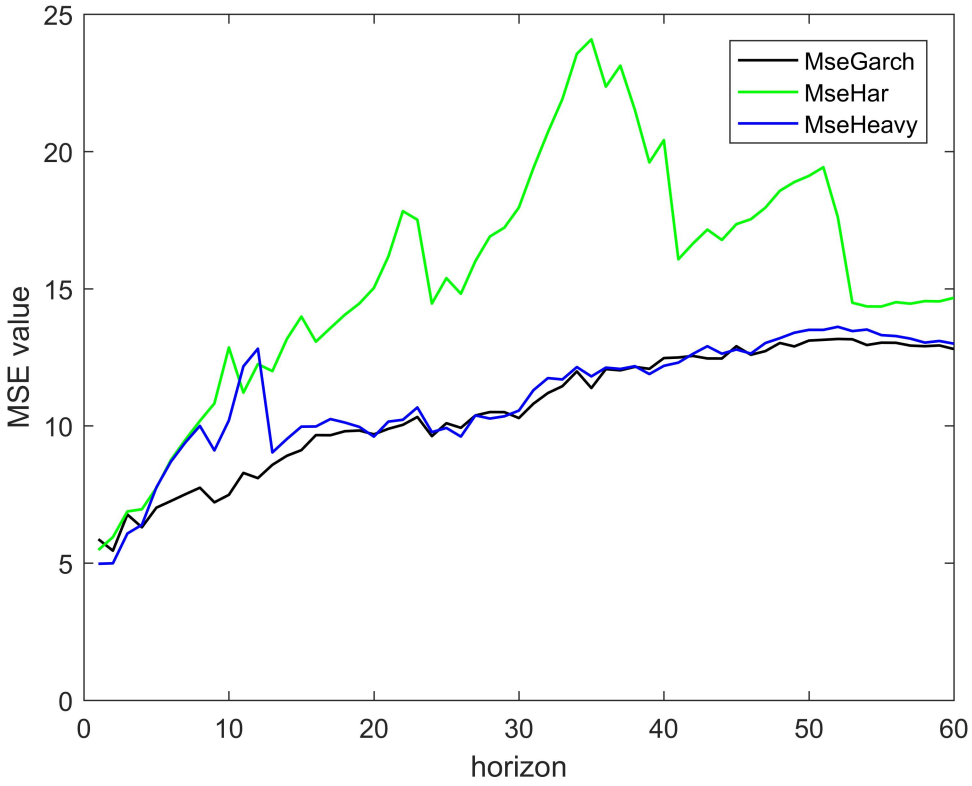

Figure 5.2.1 *MSE plot after filtering.* In this figure, all models are corrected using "insanity filter". Others are the same with Figure 5.1.1.

After the correction for all the models, I draw the MSE values with horizon changing again in Figure 5.2.1. This time the contradictory is evident. In the long run, HAR-RV has the largest errors but is obviously better than the previous model without filtering. Remember that due to the filter strategy, the extreme numbers at long horizons are substituted with the last observations in the estimated period, which largely decreases the huge forecasts. Besides, The GARCH and HEAVY model are almost on a par when forecast horizon is larger than 20.

To see which model performs the best more clearly as horizon changing, I separate all horizons into short and long horizons in Figure 5.2.2 and show the exact numbers of MSE value in Table 5.2.1.

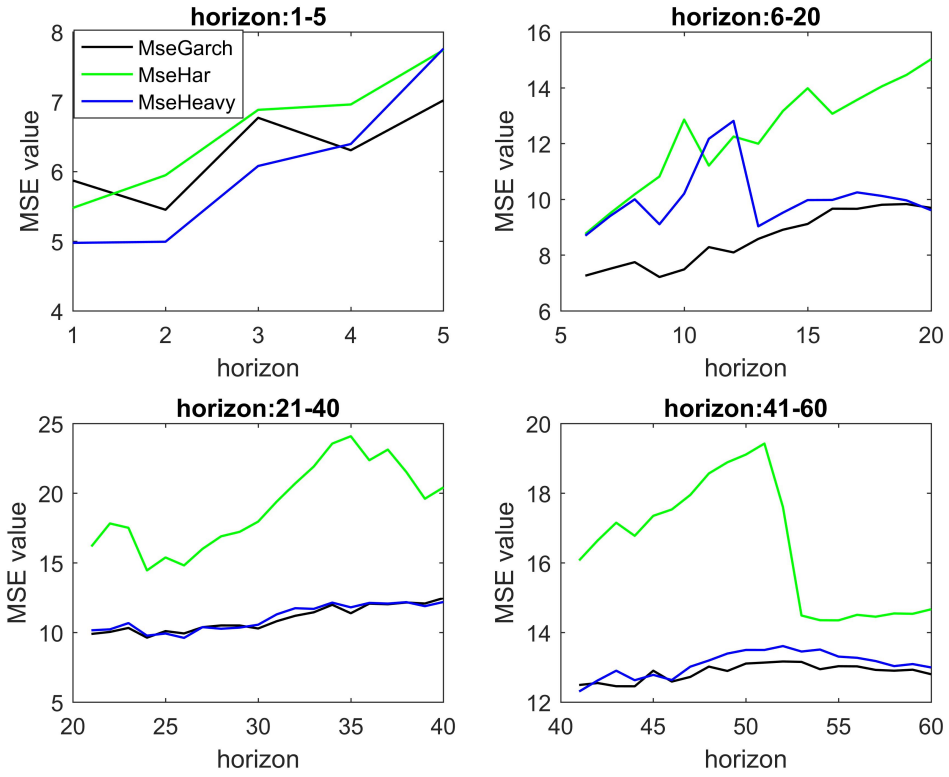

Figure 5.2.2 *Comparison using MSE at sub-periods.* This 2*2 figure separates all 60 horizons into 4 periods: 1-5, 6-20, 21-40, and 41-60. Each sub-figure zooms in corresponding periods in Figure 5.2.1.

	1	2	3	4	5	10	20	40	60
GARCH(1,1)	5.870	5.450	6.769	**6.303**	**7.019**	**7.487**	9.691	12.470	**12.800**
HAR-RV	5.477	5.945	6.883	6.960	7.740	12.857	15.024	20.411	14.665
HEAVY	**4.975**	**4.991**	**6.078**	6.393	7.759	10.200	**9.607**	**12.187**	12.995

Table 5.2.1 *MSE value at different horizons.* This table shows the MSE values at key horizons: 1, 2, 3, 4, 5, 10, 20, 40 and 60.

Apparently, when forecast horizon is no more than 3, HEAVY model has the smallest MSE value. But it is exceeded by GARCH(1,1) before horizon equals to 20. However, from 20-step-ahead and forwards, though errors of HEAVY and GARCH model have up and downs, there is no big advantage for both of them. After 40 days, GARCH model is slightly better than HEAVY. As a result from MSE, There is no winner in all situations, for example, 1-

day volatility, 5-day volatility and 60-day volatility prediction and HAR-RV cannot beat the other models at all horizons. The second loss function QLIKE is applied to see if the conclusion is about the same.

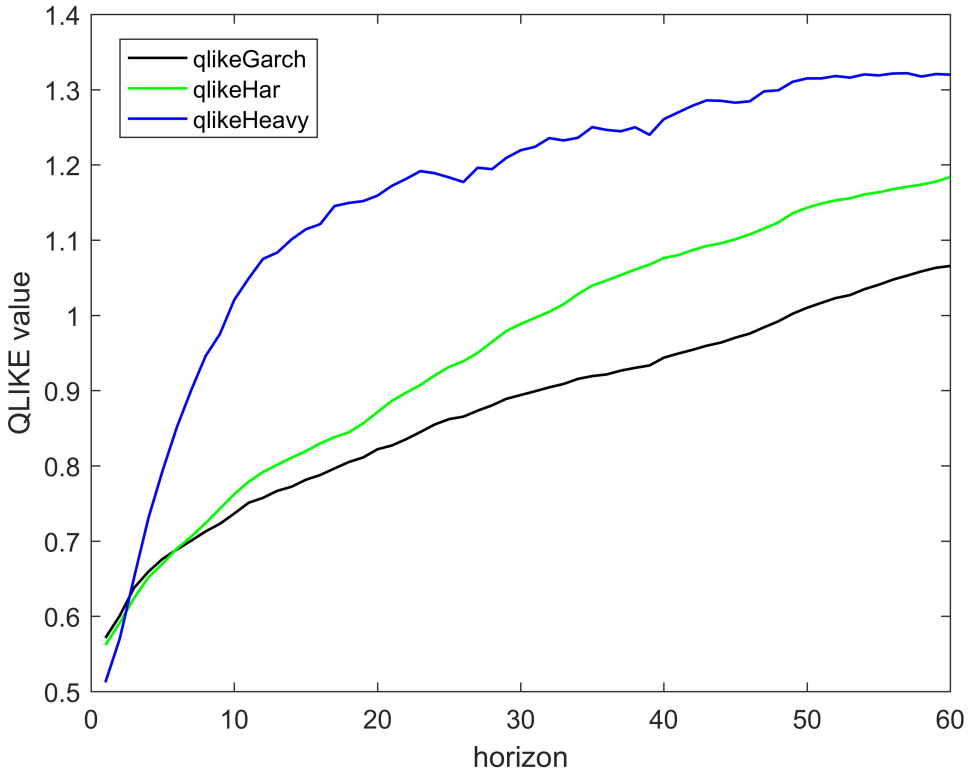

Figure 5.2.3 *QLIKE plot after filtering.* In this figure, all models are corrected after using "insanity filter". The vertical axis shows the calculated QLIKE values.

Figure 5.2.3 shows the QLIKE values of the models at all 60 horizons. Surprisingly, the pattern is quite different from using MSE. The explanation would be that according to the definitions, MSE takes care of errors while QLIKE focuses on standardized errors. Patton (2011) points out that MSE is more sensitive to outliers and volatility levels. Using QLIKE, GARCH(1,1) has the best predicting accuracy after about a week and continue its advantage until 60 days. Another change is that the error of HEAVY model exceeds that of HAR-RV model, so HEAVY becomes the last model we want to employ when forecasting in the long future.

Similarly, I zoom in different periods in Figure 5.2.4. When using QLIKE as loss function, the result is a bit different from using MSE at short horizon. HEAVY model remains to have greatest forecast accuracy in the first two days. Nevertheless, in 3 to 5 days, HAR-RV model forecasts more accurate than others.

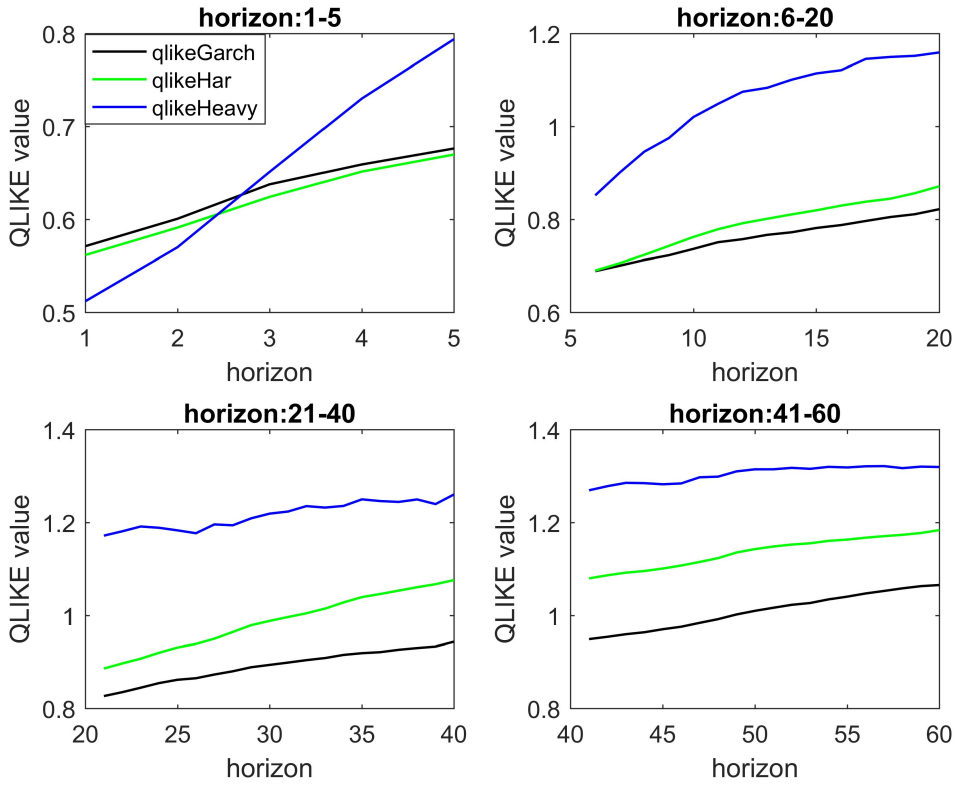

Figure 5.2.4 *Comparison using QLIKE at sub-periods.* This 2*2 figure separates all 60 horizons into 4 periods: 1-5, 6-20, 21-40, and 41-60. Each sub-figure zooms in corresponding periods in Figure 5.2.3.

	1	2	3	4	5	10	20	40	60
GARCH(1,1)	0.571	0.601	0.638	0.659	0.676	**0.737**	**0.822**	**0.944**	**1.066**
HAR-RV	0.562	0.591	**0.624**	**0.651**	**0.670**	0.762	0.871	1.076	1.184
HEAVY	**0.512**	**0.570**	0.651	0.730	0.794	1.021	1.159	1.261	1.320

Table 5.2.2 *QLIKE value at different horizons.* This table shows the QLIKE values at key horizons: 1, 2, 3, 4, 5, 10, 20, 40 and 60.

In Table 5.2.3, I compare the models by calculating p-values of SPA test using GARCH(1,1) as benchmark model. The table is divided into Panel A and Panel B using value of different loss functions as input. Remember that when the p-value is close to 1, the benchmark model is superior to the corresponding prediction model. According to Panel A, when forecasting 1-day-ahead, the p-values of HAR-RV and HEAVY model are all near 0. But the value of HEAVY model is even smaller (0.041<0.148). When forecasting the future 4, 5, 10and 60 days, p-values are all equal to 1 for both models, which indicates that the null hypothesis H_0 that the benchmark GARCH(1,1) is not inferior to any of the alternatives is harder to be rejected. This confirms the result from using MSE that GARCH (1, 1) has relatively better predicting ability. In Panel B, for the first 2 days, both models are better than GARCH(1,1), while HEAVY model predicts more precise. During 3 to 5 days, HAR-RV performs the best followed by GARCH and HEAVY model. As before, GARCH wins at long horizon. In all, the result from SPA test is exactly the same with that from loss functions, which is quite convincing.

Benchmark: GARCH(1,1)									
Bootstrap replication=1000, window size=12									
Panel A: MSE									
	1	2	3	4	5	10	20	40	60
HAR-RV	0.148	1	1	1	1	1	1	1	1
HEAVY	**0.041**	**0.182**	**0.108**	1	1	1	**0.422**	**0.309**	1
Panel B: QLIKE									
	1	2	3	4	5	10	20	40	60
HAR-RV	0.365	0.261	**0.046**	**0.235**	**0.244**	1	1	1	1
HEAVY	**0**	**0**	1	1	1	1	1	1	1

Table 5.2.3 *SPA test results.* This table illustrates the computed p-values of SPA test. In both panels, GARCH(1,1) is used as benchmark model and in my programming, I set bootstrap replication equals to 1000 and window size equals to 12. Notice that each time run the model, the numbers in this table changes but the magnitude is similar. If p-value is close to 0, GARCH(1,1) is inferior to the corresponding model. Otherwise, GARCH(1,1) performs better. Panel A is the SPA test result using MSE as input. Panel B uses QLIKE. The values are shown at key horizons: 1, 2, 3, 4, 5, 10, 20, 40 and 60.

5.3 Results with HAR-log(RV) model

From previous analysis, “insanity filter” works quite efficiently that enables the forecast losses of HAR-RV model drop a lot. However, it is still worse than the other two models at most horizons. Corsi and Reno (2009) transform RV using logarithm in case of negativity issues, which can also get approximately normal distribution of volatility measures. After that, Ma, Wei, Huang and Chen (2013) utilize high frequency data of Shanghai Composite Index as input of HAR-log(RV) model and conclude that it is the best model amongst other 22 high-frequency models based on MCS test. A pity is that in their paper, there is no GARCH and HEAVY model for assessment. So I add HAR-log(RV) in order to investigate if it can provide a better choice.

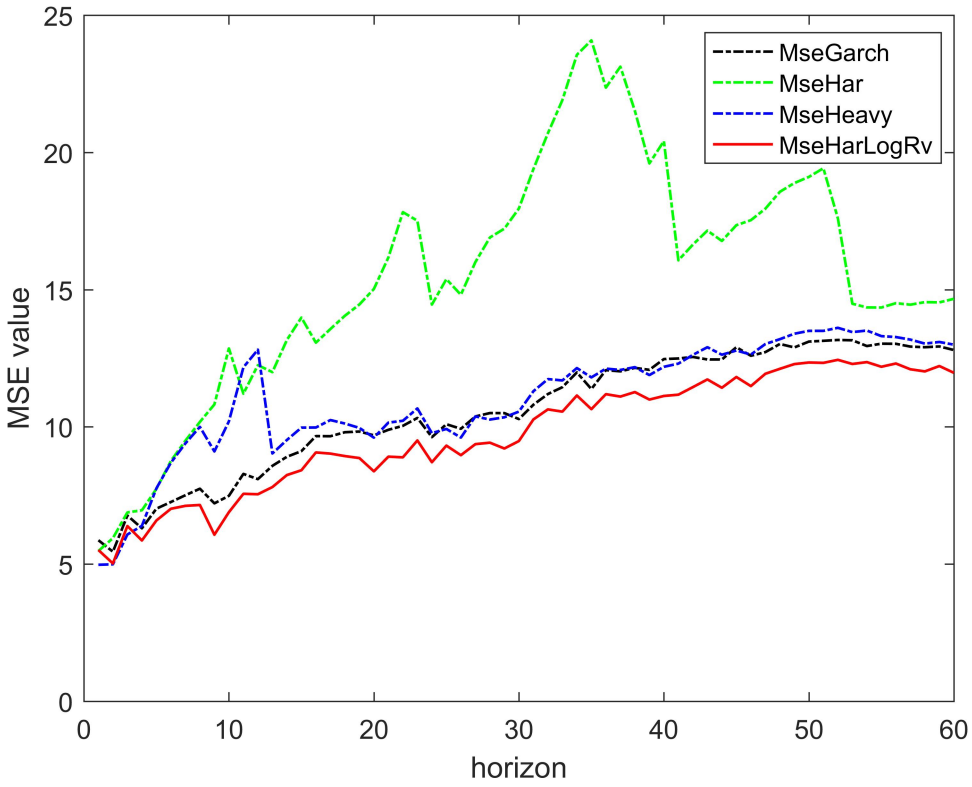

Figure 5.3.1 *MSE comparison with HAR-log(RV) model.* This figure adds HAR-log(RV) model to Figure 5.2.1.

Figure 5.3.1 includes HAR-log(RV) model in the MSE measurement. Similar with the expectation, the new added model indeed has excellent performance. It beats GARCH(1,1)

almost at any horizons. Figure 5.3.2 and Table 5.3.1 shows the difference during short predicting period more clearly. When horizon=1, 2 and 3, HEAVY model has the lowest MSE, which is the same without adding HAR-log(RV) model. Big change arises from 3 days in the future and onwards, the MSE values of HAR-log(RV) are consistently lower than that of GARCH model, which confirms the idea that using logarithm indeed has great predicting advantage.

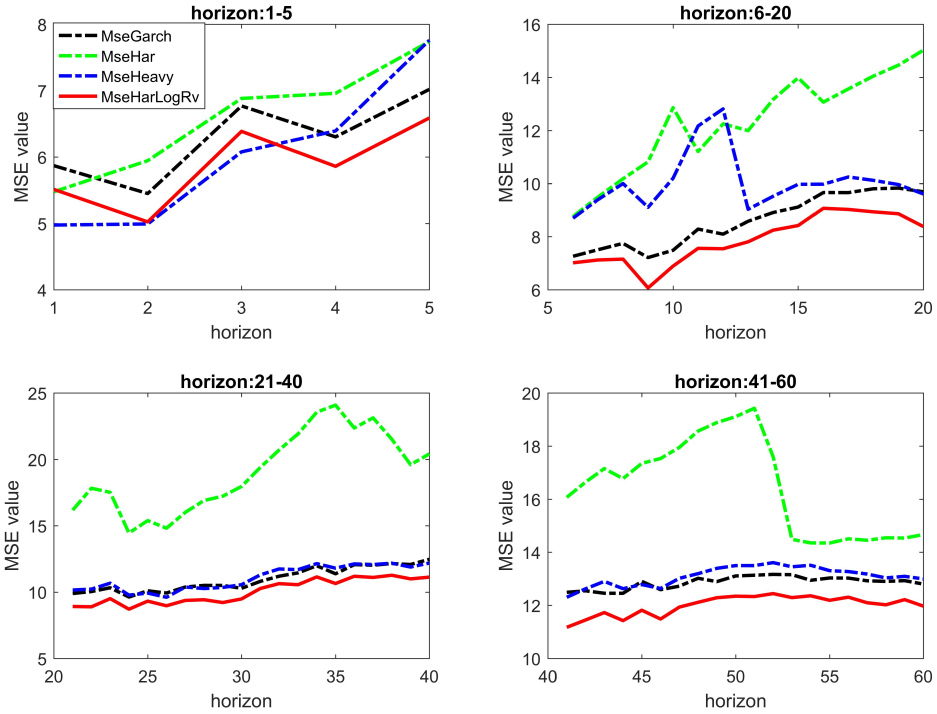

Figure 5.3.2 *MSE comparison with HAR-log(RV) at sub-periods.* This plot adds HAR-log(RV) model to Figure 5.2.2. Other illustrations refer to Figure 5.2.2.

	1	2	3	4	5	10	20	40	60
GARCH(1,1)	5.870	5.450	6.769	6.303	7.019	7.487	9.691	12.470	12.800
HAR-RV	5.477	5.945	6.883	6.960	7.740	12.857	15.024	20.411	14.665
HEAVY	**4.975**	**4.991**	**6.078**	6.393	7.759	10.200	9.607	12.187	12.995
HAR-log(RV)	5.512	5.023	6.387	**5.860**	**6.588**	**6.891**	**8.379**	**11.125**	**11.970**

Table 5.3.1 *MSE value with HAR-log(RV) at different horizons.*

Again, I test the result using QLIKE loss function in Figure 5.3.3 and Figure 5.3.4. The performance ranking is totally different from the above outcome. In the long run, GARCH beats all the other models and even HAR-log(RV) model. Besides, HAR-log(RV) has higher QLIKE value than the simple HAR-RV model, while HEAVY is still the worst to predict volatility in the far future. Because at all 60 horizons, HAR-log(RV) model is inferior to GARCH and HAR-RV model, so the preeminent models with horizon changing do not change from when logarithm is not applied.

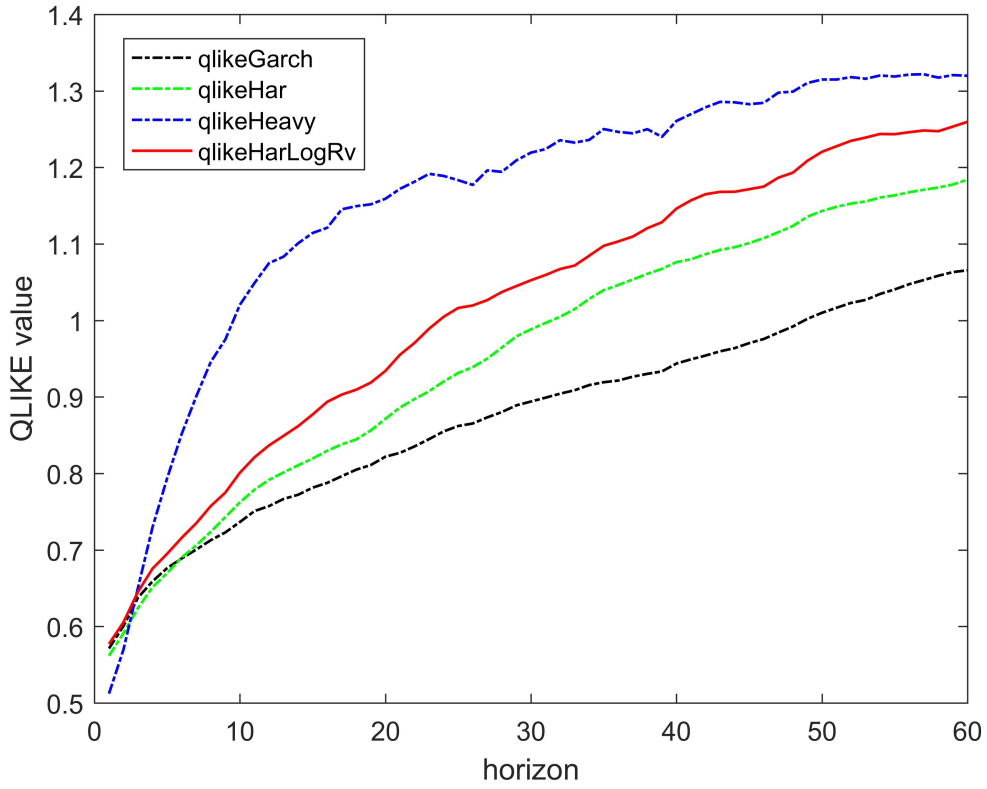

Figure 5.3.3 *QLIKE comparison with HAR-log(RV) model.* This figure adds HAR-log(RV) model to Figure 5.2.3.

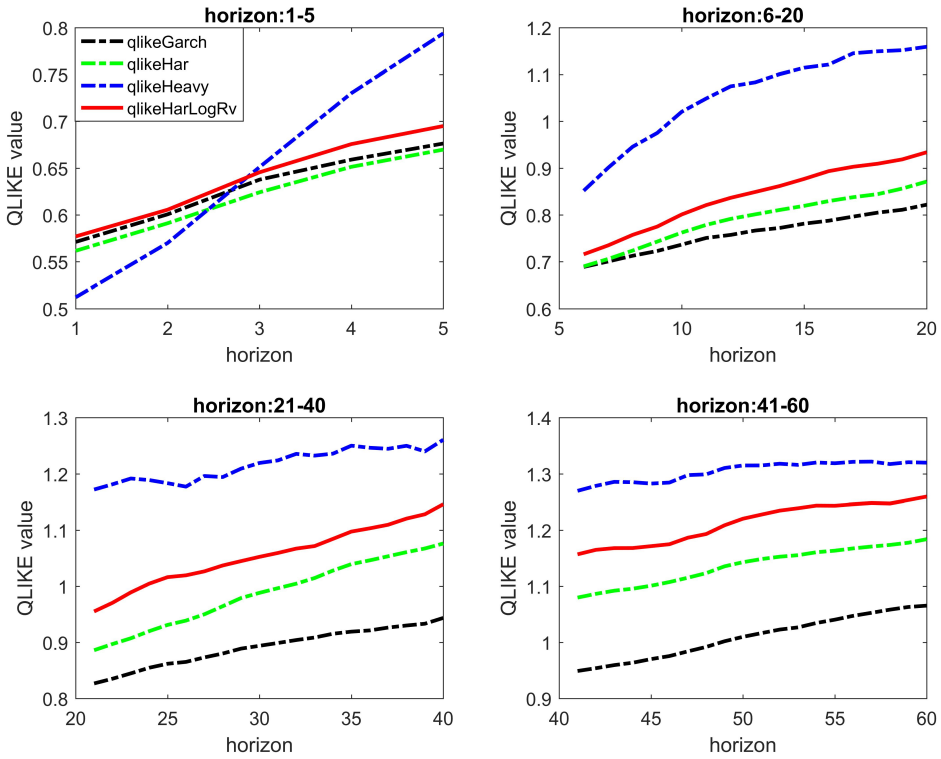

Figure 5.3.4 *QLIKE comparison with HAR-log(RV) at sub-periods.* This plot adds HAR-log(RV) model to Figure 5.2.4. Other illustrations refer to Figure 5.2.4.

	1	2	3	4	5	10	20	40	60
GARCH(1,1)	0.571	0.601	0.638	0.659	0.676	**0.737**	**0.822**	**0.944**	**1.066**
HAR-RV	0.562	0.591	**0.624**	**0.651**	**0.670**	0.762	0.871	1.076	1.184
HEAVY	**0.512**	**0.570**	0.651	0.730	0.794	1.021	1.159	1.261	1.320
HAR-log(RV)	0.577	0.606	0.646	0.676	0.695	0.801	0.934	1.146	1.260

Table 5.3.2 *QLIKE with HAR-log(RV) value at different horizons.*

The p-values of SPA test including HAR-log(RV) model are in Table 5.3.3. The values of HAR-RV and HEAVY model in this table are no doubt the same as in Table 5.2.3. The difference is that in Panel A from day 4, the p-values of HAR-log(RV) model are around 0.1 that are all the smallest among the other models though at horizon 20 and 40, HEAVY model is also superior than GARCH. The numbers show that the new model is a universal

winner at relatively long horizon. In contrast, all the p-values of HAR-log(RV) are 1 in Panel B.

Benchmark: GARCH(1,1) Bootstrap replication=1000, window size=12									
Panel A: MSE									
	1	2	3	4	5	10	20	40	60
HAR-RV	0.148	1	1	1	1	1	1	1	1
HEAVY	**0.041**	**0.182**	**0.108**	1	1	1	0.422	0.309	1
HAR-log(RV)	0.168	0.163	0.251	**0.189**	**0.207**	**0.192**	**0.097**	**0.078**	**0.049**
Panel B: QLIKE									
	1	2	3	4	5	10	20	40	60
HAR-RV	0.365	0.261	**0.046**	**0.235**	**0.244**	1	1	1	1
HEAVY	**0**	**0**	1	1	1	1	1	1	1
HAR-log(RV)	1	1	1	1	1	1	1	1	1

Table 5.3.3 *SPA test result with HAR-log(RV) model.* Detailed description see Table 5.2.3.

Chapter 6. Conclusion

I study the volatility predictability of GARCH(1,1), HAR-RV, HEAVY model and an extensive model HAR-log(RV) at both short and long horizons. Daily returns and realized variances of S&P 500 stock index is used as input of the models and the forecasts are made using a rolling window of 1000 days. The forecast accuracy in this article is measured by two loss functions (MSE and QLIKE) and SPA test, where realized variances are used as the benchmark for loss functions and GARCH(1,1) is the benchmark model for SPA test. In my research, SPA tests in terms of MSE and QLIKE provide the same result as applying only loss functions respectively.

Due to the fact that without correcting the unrealistic forecasts, the predicting errors are extremely large and harm the good forecasts. “Insanity filter” is applied to all the models to replace the “insanity” number with the most recent observed RV.

Based on my empirical analysis result, there is no universal winner. Under all the statistical criteria, HEAVY model has lowest forecast error if prediction period is short (1 to 3 days). This is consistent with what Shephard and Sheppard (2010) point out. When predicting horizon is long, the best model depends on which loss function is used. Under MSE, GARCH and HEAVY model both have well predicted ability at long horizon. However, QLIKE value indicates that HAR-RV has the most accurate 3 to 5-step-ahead forecasts while GARCH remains perfect after a week from now.

To improve the HAR-RV model, I follow Corsi and Reno (2009) and take logarithm of RV. The MSE values decreases dramatically and HAR-log(RV) is better than other models from horizon 4, but it still cannot beat HEAVY in the first 3 days. In contrast, from the QLIKE values, HAR-log(RV) is not the best model at any horizon.

To sum up, I draw Figure 6.1 to show the best model at different horizons under MSE and QLIKE. SPA test result is the same as loss functions, so it is excluded. My advice of choosing the best-performance model depends on the different characteristics of traders. For those who trade very often (speculator), HEAVY model is no doubt the best choice. Long-term investor can select from GARCH(1,1) and HAR-log(RV), as they both have their own advantages under different measurements.

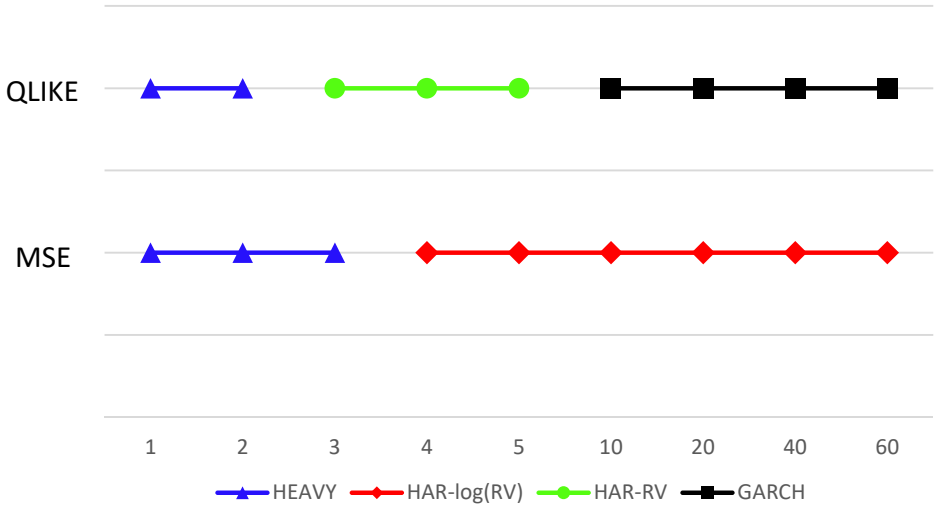

Figure 6.1 *Best model at different horizons.* This figure takes all models into account. The horizontal axis is horizon and the vertical axis indicates when using different measurements. The models shown in this figure have lowest errors at corresponding horizon.

Chapter 7. Discussion

As what I have described in Chapter 1, the comparison outcomes are not the same in previous researches and some are even contradictory. So does my result. It turns out that the predicting accuracy of GARCH(1,1) is stable while the accuracy of HEAVY and HAR-RV model decreases relatively larger with horizon increasing. Things changes after I improve HAR-RV model by applying logarithm on RV. It then has steady losses and beats the other models to be the best model. Therefore, one guess is that using extensions of the basic models to avoid unnecessary noise and increase stability may benefit the forecast precision a lot.

Another factor that may influence the forecasts may be the filtering strategy. As I have described in Chapter 5.1, the forecast is replaced if the forecast difference (RV_{T+h} - RV_T) is outside the range. However, if the observed RV_T is an abnormal number compared to the data previous and after it, the forecast different may lies in the range. Therefore, although the forecast RV_{T+h} is unrealistic, it is still regarded as a good prediction.

Last but not least, the different of data used may affect the ranking hugely. I include the variances of 2008 financial crises which increase that of the relatively calm period by almost tenfold. This strike brings extremely large predicting losses to the models and may cause distortions to the overall predictability.

References

Andersen, G. T., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International economic review*, pp. 885-905.

Andersen, G. T., Bollerslev, T., & Diebold, X. F. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The review of economics and statistics, 89*(4), pp. 701-720.

Andersen, G. T., Bollerslev, T., Diebold, X. F., & Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American statistical association, 96*(453), pp. 42-55.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica, 71*, pp. 529-626.

Barndorff-Nielsen, O. E., & Shephard, N. (2002). Estimating quadratic variation using realised variance. *Journal of Applied Econometrics, 17*(5), pp. 457-477.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics, 3*, pp. 307-327.

Bollerslev, T., & Ghysels, E. (1996). Periodic autoregressive conditional heteroscedasticity. *Journal of Business & Economic Statistics, 14*(2), pp. 139-151.

Bollerslev, T., Patton, J. A., & Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics, 192*(1), pp. 1-18.

Clements, P. M., & Hendry, F. D. (2011). *The Oxford handbook of economic forecasting.* Oxford University Press.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics, 7*(2), pp. 174-196.

Corsi, F., & Reno, R. (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *Journal of Business & Economic Statistics, 30*(3), pp. 368-380.

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

Giot, P., & Laurent, S. (2004). Modelling daily value-at-risk using realized volatility and ARCH type models. *Journal of Empirical Finance, 11*(3), pp. 379-398.

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

Hamilton, D. J., & Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of econometrics, 64*(1), pp. 307-333.

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

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

Hansen, P. R., & Lunde, A. (2011). Forecasting volatility using high frequency data. *The Oxford Handbook of Economic Forecasting*, pp. 525-556.

Hansen, R. P. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics, 23*(4), pp. 365-380.

Hansen, R. P., Lunde, A., & Nason, M. J. (2011). The model confidence set. *Econometrica, 79*(2), pp. 453-497.

Koopman, J. S., Jungbacker, B., & Hol, E. (2005). Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements. *Journal of Empirical Finance, 12*(3), pp. 445-475.

Lamoureux, G. C., & Lastrapes, D. W. (1993). Forecasting stock-return variance: Toward an understanding of stochastic implied volatilities. *The Review of Financial Studies, 6*(2), pp. 293-326.

Laurent, S., Rombouts, J. V., & Violante, F. (2013). On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics, 173*(1), pp. 1-10.

Lee, S. S., & Mykland, A. P. (2012). Jumps in equilibrium prices and market microstructure noise. *Journal of Econometrics, 168*(2), pp. 396-406.

Ma, F., Wei, Y., Huang, D., & Chen, Y. (2014). Which is the better forecasting model? A comparison between HAR-RV and multifractality volatility. *Physica A, 405*, pp. 171-180.

Manda, K. (2010). Stock market volatility during the 2008 financial crisis. *Glucksman Fellowship Program Student Research Reports*(87).

Markowitz, H. (1952). Portfolio selection. *The journal of finance, 7*(1), pp. 77-91.

Nelson, B. D. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, pp. 347-370.

Noureldin, D., Shephard, N., & Sheppard, K. (2012). Multivariate high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics, 27*(6), pp. 907-933.

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

Shephard, N., & Sheppard, K. (2010). Realising the future: forecasting with high frequency based volatility (HEAVY) models. *Journal of Applied Econometrics, 25*(2), pp. 197-231.

Swanson, R. N., & White, H. (1995). A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business & Economic Statistics, 13*(3), pp. 265-275.

Swanson, R. N., & White, H. (1997). A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *The Review of Economics and Statistics, 79*(4), pp. 540-550.

Swanson, R. N., & White, H. (1997). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International journal of Forecasting, 13*(4), pp. 439-461.

Timmermann, A. (2006). Forecast combinations. *Handbook of Economic Forecasting, 1*, pp. 135-196.

White, H. (2000). A reality check for data snooping. *Econometrica, 68*(5), pp. 1097-1126.